UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Ciências
ULisboa

Finding Biomarkers of Metastatic Potential in Colorectal Cancer

Valério da Silva Fiori

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Daniel Sobral
Lisete de Sousa

2023

Acknowledgments

I would like to express my gratitude to everyone who supported and helped me to finish this challenge. First, I would like to thank my advisers, Researcher Daniel and Professor Lisete for giving me this opportunity and also for their knowledge, guidance, support, availability and dedication which made possible the making of this project. I would also like to thank Researcher Ana Grosso for also guiding and helping me, and to all the members of the Computation Multi-Omics team for the fun times and good company. To my family, I would like to show my deepest gratitude, for always being there for me supporting in everything and believing in me. To my best friend Maria, who always reminded me that I could do this, for our times writing the thesis together as motivation, and for joining me in all kind of sports and activities. To my friends, Vasco, Ana, Joana, Catarina, Alice, Inês, Aylton, Daniela, Olga, Constança and many others that were there for me, especially due to the COVID harsh situation. And lastly, to my pets, Kiko, Leo, and Puppy, and my best friend's pets Sky and Zen, for providing me with the purest moments of love and joy.

Abstract

Colorectal cancer (CRC) is among the most common cancers and globally one of the deadliest. Treatments for this disease have been improving, increasing the survival rates for CRC patients. On the contrary, when the cancer progresses and metastasis occurs, the overall survival rates are quite low, reflecting the role of metastasis as the leading cause of death. CRC is a very heterogeneous disease that can be classified in various types depending on how it develops, and in which are involved many different pathways. Despite its complexity, CRC molecular mechanisms are each year being better understood, including events like cell proliferation, immune surveillance blockage, cell adhesion disturbances. These mechanisms may be related to cancer progression from primary to metastatic. Biomarkers are molecules found in tissue, blood, or stool samples and have been used to identify diseases like cancer. These markers have proved to be very beneficial in the aid of CRC treatment and other cancers. Despite that, the identification of reliable biomarkers for metastatic CRC remains poor, mainly biomarkers that predict metastasis development. Here we uncovered promising biomarkers in early stages that reflect the tumour potential to evolve from primary to metastasis. Bioinformatic analyses were conducted with transcriptomic data to investigate gene expression differentiation, gene sets enrichment, and to create a predictive model. Genomic data was also analysed to find correlations between mutations and metastasis occurrence, although this last step was inconclusive. The most differential expressed genes found have been also identified to be related to metastasis in other studies and the same happened with some of the enriched pathways. The predictive model had an insufficient efficacy but revealed to be promising. Here we showed the impact that gene expression analysis can have in the important field of biomarker research and the need of future studies with this type of data.

Keywords: colorectal cancer, metastasis, biomarker, transcriptomic, prediction

Resumo

O Cancro colorretal (CRC) é o terceiro cancro mais comum no mundo e o segundo mais mortífero. A sua heterogeneidade tem sido um dos maiores obstáculos na medicina de tratamento de pacientes com CRC. Este tipo de cancro pode desenvolver-se em diferentes zonas do trato gastrointestinal e crescer a partir de várias vias biológicas mostrando o quão desafiante é aprofundar o nosso conhecimento na formação e desenvolvimento deste. Apesar destes desafios, ao longo dos anos, os mecanismos por trás do CRC têm vindo a ser melhor compreendidos e, em conjunto com os avanços na tecnologia, nomeadamente na área das ómicas como genómica, transcriptómica, epigenómica e outras, o resultado dos tratamentos de pacientes com CRC tem vindo a melhorar, aumentando as taxas de sobrevivência. Isto tem acontecido graças à incorporação de biomarcadores nos procedimentos e decisões aquando do tratamento de pacientes com CRC. Biomarcadores são moléculas que podem ser detetadas em amostras de tecidos das células, no sangue e nas fezes que permitem a identificação de condições patológicos como o caso de cancros. Estas assinaturas biológicas têm vindo a ser muito utilizadas na medicina no âmbito de diagnosticar doenças, fazer prognósticos do desenvolvimento das doenças e prever a resposta, positiva ou negativa, de pacientes a tratamentos específicos. Apesar destes avanços biológicos e tecnológicos, o aparecimento de metástases continua a ser a maior causa de mortes em pacientes de CRC. A metástase, definida como a última fase de desenvolvimento do cancro, acontece quando células do tumor primário se propagam para outras zonas/órgãos onde levam à formação de um tumor secundário. Para que tal aconteça, estas células têm de sobreviver aos diversos sistemas de defesa e de controlo que o corpo humano possui, refletindo a importância em perceber os mecanismos que permitem o alastramento e sobrevivência destas. Apesar do grande esforço e foco em estudos investigarem e perceberem melhor os fatores que levam ao aparecimento de metástases, o nosso conhecimento destes continua superficial. Biomarcadores relacionados com metástases também têm vindo a ser descobertos aos poucos e a ser também incorporados na escolha de terapias na cura do cancro, mas, contrariamente aos melhores resultados que têm vindo a ser obtidos no tratamento de tumores primários, tratamentos aquando do aparecimento de metástases continuam pouco eficazes. Um dos pontos mais focados, em pacientes em fases iniciais de CRC, é descobrir fatores envolvidos no desenvolvimento de tumor primário para secundário/metastático que permitam prever o potencial de que os pacientes de CRC têm, ou não têm, de vir a sofrer desta ocorrência. Apesar de promissora, a descoberta deste tipo de marcadores tem sido escassa devido ao quão desafiante é obter dados neste contexto. Grandes quantidades de amostras têm de ser recolhidas nos estágios iniciais da doença e os pacientes têm de ser monitorizados durantes longos períodos de tempo, falando de vários anos, para controlar se ocorreu ou não o aparecimento/desenvolvimento de metástases e, só então, proceder à análise.

Neste estudo foram analisados dados de pacientes de CRC nos estágios II e III, os quais foram acompanhados durante 3 anos para se controlar se ocorreu metastização ou não. Estes dados vieram do Hospital Santa Maria e foram obtidos pela equipa médica do doutor Luís Costa. Next Generation Sequencing (NSG) foi utilizada para gerar estes dados, mais especificamente RNA-seq a qual permitiu a sequenciação do RNA para produzir dados do transcriptoma e whole exome sequencing (WES) que permitiu a sequenciação do exoma. Os dados de transcriptoma foram utilizados para fazer análises bioinformáticas de discrepâncias na expressão de genes entre amostras que metastizaram e amostras que não o fizeram. Inicialmente foram feitas análises exploratórias dos dados dos pacientes e da expressão dos genes. A análise de diferenças na expressão de genes também levou a investigar se havia enriquecimentos em vias biológicas. Para tal foram utilizadas duas bases de dados de vias biológicas, uma relacionada com o reactoma e outra com cancros. Por fim, com os dados de RNA-seq, foi desenvolvido um classificador capaz de distinguir, com base nos valores das expressões dos genes, se uma amostra metastizou ou não, com o intuito de utilizar este programa em dados de outras instituições de modo a perceber quais amostras eram identificadas com potencial risco de metastizar. Os dados do exoma permitiram descobrir mutações entre as amostras de pacientes onde houve metástase e as em que tal não aconteceu. Mutações como SNPs e CNVs foram analisadas pois tem sido comprovada a relação destes tipos de mutações com doenças como cancros. Testes estatísticos foram corridos para investigar a possível relação destes tipos de mutações com o desenvolvimento de metástases.

A análise de expressão diferencial resultou em 166 genes diferencialmente expressos (DEGs), dos quais 94 foram regulados negativamente e 72 regulados positivamente. Destes DEGs foram selecionados os 20 mais significativos para analisar individualmente a dispersão dos níveis de expressão ao longo das 149 amostras. Alguns destes genes têm sido estudados em outros artigos como potenciais biomarcadores de metástase em CRC.

A análise de vias biológicas enriquecidas foi brevemente explorada, mas no entanto, as vias resultantes mostraram-se interessantes pois estavam relacionadas com mecanismos de controlo do ciclo celular, apoptose, transporte membranar, e outras vias que têm sido conectadas com o desenvolvimento de metástases.

O classificador deparou-se com algumas barreiras devido à diferença de amostras de metástase em comparação com as de não metástase. Todavia, desempenhou a classificação com uma eficácia de valores entre 60%-65%, que apesar de não ser muito alta, mostrou ser uma abordagem promissora.

A relação entre as mutações e as metástases revelou-se sem significância a nível estatístico. No entanto, existiu uma aparente possibilidade de haver relações entre estes pois a chance de mutações em alguns genes estarem conectadas ao aparecimento da metástase foi maior que a estarem conectadas ao não aparecimento.

Este estudo deparou-se com alguns obstáculos. Primeiro, os dados vieram de uma só instituição. Dados da mesma natureza vindos de outras instituições são necessários para validar estas análises, no entanto, este tipo de dados é raro devido às dificuldades que compromete a obtenção dos mesmos, referidas anteriormente. Segundo, a discrepância entre o número de amostras de metástase e as amostras em que o cancro não metastizou dificultou as diversas análises feitas neste projeto. Terceiro, este estudo centrou o foco na análise de expressão diferencial dos genes, explorando superficialmente o campo das vias biológicas e das mutações, campos estes que se têm mostrado importantes na procura de biomarcadores no auxílio da luta contra o cancro. Quarto, a fase de desenvolvimento que se encontra o tumor revelou ter uma possível relação com uma maior ocorrência de metástases, mostrando que, futuramente, estudos focados entre CRC II e CRC III podem aprofundar o nosso entender no tópico das metástases e trazer novas visões de possíveis biomarcadores.

Concluindo, aqui foi relevada a importância que tecnologias ómicas, nomeadamente transcriptómica e genómica, têm no mundo da medicina e o impacto que podem fazer na luta contra o cancro. Alguns dos genes diferencialmente expressos revelaram-se como potenciais biomarcadores do desenvolvimento de tumor primário para secundário ou metástase. As vias biológicas enriquecidas são um campo promissor para melhor compreensão dos mecanismos por trás da proliferação e disseminação das células cancerígenas. A combinação de análises de transcriptoma com análises de genoma revela ser uma mais valia na descoberta de biomarcadores, fortalecendo o potencial de biomarcadores encontrados em ambas as duas técnicas. Futuros estudos são necessários para averiguar o potencial destas descobertas, principalmente devido ao facto de haver uma enorme escassez deste tipo de dados que, no entanto, revelam ter uma extrema importância na luta contra o cancro, principalmente o cancro colorretal.

Palavras-chave: cancro colorretal, metástase, biomarcador, transcriptómica, previsão

Index

List of Figures

List of Tables

1 Introduction

Colorectal cancer (CRC) is one of the most common cancers and one of the top causes of cancer-related deaths becoming a worldwide health emergency. Surgery to remove the primary tumour is the most relevant and curative approach to treat CRC and with the creation of more effective treatments with better prognosis boosted by biomarkers has increased the five-year survival to 80% of the patients with stages I to III CRC (Siegel et al., 2020). This shows that still 20% experience tumour recurrence highlighting the significance of creating biomarkers to identify those patients who may benefit from post-operative treatment intensification. Furthermore, patients in the stage IV have a survival rate of approximately 13%, which reveals that metastasis has a great negative impact in CRC mortality (Siegel et al., 2020). Cancer-specific biomarkers have shown to be crucial for cancer diagnosis, prognosis, treatment, and prevention. Despite the emerging number of biomarkers identified for CRC, such as PTCH1, STK31, and SPAG9, none are currently routinely employed for clinical use (Kanojia, Garg, Gupta, Gupta, & Suri, 2011; You et al., 2010; Zhong et al., 2017). Therefore, it is crucial to thoroughly investigate biomarkers and develop straightforward strategies to incorporate them into preventative, therapeutic, monitoring, and prognostic approaches. The use of genetic indicators to identify and predict CRC metastases would be very beneficial (Okita et al., 2018). Finding certain quantifiable chemicals linked to tumour aggressiveness would have significant clinical implications and open possibilities for determining CRC's early spread (Filip et al., 2020).

1.1 Colorectal Cancer

CRC represents two types of cancer, respectively colon and rectal cancers, and has revealed to be the most common and major cause of death in the gastrointestinal tract, being the third most common cancer in the world and the fourth most common in deaths related to cancer whereas lung, liver and stomach cancer are the only ones that exceed it (Siegel, Miller, & Jemal, 2019). In the intestinal mucosa, CRC can start as a polyp or as an adenoma, which is a benign tumour that has the capability to transform into a malign one over time. This progression and transformation of epithelial cells to cancer cells is caused by the progressive accumulation of genetic mutations or epigenetic alterations that affect specific pathways. Based on the CRC’s origin, it can be classified as sporadic or familial/hereditary (E. F. Fearon & Vogelstein, 1990).

Sporadic CRC accounts for 70% of CRC cases and there is a proven association with environmental and dietary factors as excessive alcohol ingestion, smoking, reduced physical activity, sedentary life, diets with too much red meat and fats and low in fibre (E. R. Fearon, 1994; Marchand, Wilkens, Hankin, Kolonel, & Lyu, 1997). One of the major risk factors in sporadic cases is considered to be the age where the incidence significantly increases over the age of 40-50 years (Levin, Lieberman, Mcfarland, Andrews, & Brooks, 2008). Familial CRC accounts for 30% of the cases, affecting people that have historical CRC in their family (Stoffel & Kastrinos, 2014). Inherited or genetic CRC categorization depends on the existence of colonic polyps (E. F. Fearon & Vogelstein, 1990; Lynch & de la Chapelle, 2003). Hereditary nonpolyposis CRC (HNPCC or Lynch syndrome) is the term used to describe diseases without polyposis. In the other hand, there are various terms for diseases with polyposis as familial adenomatous (FAP), MUTYH-associated polyposis (MAP), and others (E. R. Fearon, 1994; Umar et al., 2004; Wirtzfeld, Petrelli, & Rodriguez-bigas, 2001). Some pre-existing diseases can also be considered risk factors (Xie & Itzkowitz, 2008).

1.1.1 Molecular Pathways

The mutations that play a critical role in the formation of colorectal carcinoma appear in oncogenes, genes related to tumour suppression and genes related to DNA repair mechanisms (E. F. Fearon & Vogelstein, 1990). There are three major pathways involved in CRC. One is the chromosomal instability pathway (CIN), which is the most common one representing the majority of CRC cases. CIN is responsible for triggering critical pathways in CRC tumorigeneses caused by activation of oncogenes such as KRAS, inactivation of tumour suppressor genes (TSGs) like the adenomatous polyposis coli (APC) and TP53 and causing loss of heterozygosity (S. D. Markowitz & Bertagnolli, 2009; Pino & Chung, 2010). Another pathway is the microsatellite instability (MSI) characterized by the loss of DNA repair mechanisms usually due to mutations in mismatch repair genes (MMR). MSI is usually the indication of HNPCC cases (Geiersbach & Samowitz, 2011; Thibodeau, Bren, & Schaid, 1993; Ward et al., 2001). The last pathway is the CpG island methylator phenotype (CIMP) where the main aspect is hypermethylation of oncogene promoters leading to loss of protein expression caused by genetic silencing (Lao & Grady, 2011; Ogino et al., 2009; Weisenberger et al., 2006).

1.1.2 The Two Sides of Colorectal Cancer

One way to categorize CRC can be related to the location of the primary tumour relative to the splenic flexure site, where it can be classified as right-sided if proximal or left-sided if distant (Bufill, 2016). In the right-side category are included the caecum, the ascending colon, and the transverse colon, and in the left-side the descending colon, the sigmoid colon, and the rectum. The sidedness of the cancer contributes for differences in pathogenesis, molecular pathways, and prognosis of CRC (Weiss et al., 2011). Chromosomal instability is typically seen in left-sided CRC. Additionally, KRAS and p53 mutations have been identified as being associated with left-sided CRC. Contrarily, right-sided CRC is frequently defined as having high levels of CpG island methylator phenotype (CIMP), B-Raf proto-oncogene, serine/threonine kinase (BRAF) mutation, and microsatellite instability (MSI). Right-sided CRC patients have been reported to have bigger and advanced tumours (Glebov et al., 2003; Barry Iacopetta, 2002; Nawa et al., 2008). This reflects the importance and impact that the sidedness of the tumour can have on CRC patients' treatment.

1.1.3 Consensus Molecular Subtypes

The Consensus Molecular Subtypes are a division/characterization of CRC groups focusing on the correlation between gene expression and epigenomic, transcriptomic, microenvironmental, genetic, prognostic, and clinical characteristics. The CMS1 subtype is immunogenic and hypermutated. CMS2 tumours are activated by the WNT-β-catenin pathway. CMS3 feature a metabolic cancer phenotype and CMS4 cancers have the worst survival and have a strong stromal gene signature. The Consensus Molecular Subtypes of CRC may better inform clinicians of prognosis, therapeutic response, and potential novel therapeutic strategies (Guinney et al., 2015).

1.1.4 Stages

Cancer can be described by stages where the criteria is based on how far the cancer has grown, location and size of the tumour, if it has reached nearby tissues and whether or not it has spread to lymph nodes or other organs. The stages vary from I to IV where stage I is called the early-stage cancer where the tumour has not spread and has not grown far into nearby structures. Stage II defines cancers which the tumour has grown larger deep into the tissue but has not spread. In stage III cancer has grown larger and deeper with the possibility of having spread to other tissues or lymph nodes. The last which is stage IV represents cancers that have spread to other areas and organs, representing what is called metastasis (Gospodarowicz et al., 1998).

1.1.5 Metastasis

When a secondary tumour develops in a different tissue or organ than the one that has the primary cancer it is called metastasis and it is the final stage of cancer development (E. F. Fearon & Vogelstein, 1990; E. R. Fearon, 1994). This occurrence is the main cause of cancer mortality and why cancer treatment fails (Bozzetti, Doci, Bignami, Morabito, & Gennari, 1987; Manfredi et al., 2006; Misiakos, Karidis, & Kouraklis, 2011; Siegel et al., 2020). Although being a very focused theme in medicine studies and investigation, the knowledge about it is still poor or insufficient. For a cancer cell to go from the primary location and spread to a foreign organ or tissue it needs to survive some adversities as entering the bloodstream, circulate in blood vessels, being targeted by immune cells in the bloodstream, conditioned to blood vessels pressure, and adapt to the new environment (Massagué & Obenauf, 2016). All these factors need to be better understood to help in the advancement of cancer therapies. Metastasis is the final obstacle to create more efficient cancer treatments. However, the understanding of the dynamics behind metastasis evolution is a very hard topic which is contingent to the development of effective cancer therapies (Fares, Fares, Khachfe, Salhab, & Fares, 2020; Keum & Giovannucci, 2019; Sánchez-Gundín, Fernández-Carballido, Martínez-Valdivieso, Barreda-Hernández, & Torres-Suárez, 2018).

For this objective, studying gene expression levels and signalling pathways associated with CRC metastasis is one way to comprehend CRC evolution. Some pathways have already been identified to be related with CRC progression and metastasis, such as WNT/β-catenin, TP53, TGF-β/SMAD, Notch, VEGF, and JAKs/STAT3. Other alterations caused to the regulation of cellular mechanisms like cell cycle, transcription, apoptosis, and angiogenesis are important aspects to focus (Corvinus et al., 2005; Es et al., 2005; E. F. Fearon & Vogelstein, 1990; Guba, 2004; Klaus & Birchmeier, 2008; S. D. Markowitz & Bertagnolli, 2009; S. Markowitz et al., 2016; Pino & Chung, 2010).

Cellular growth and stem-cell differentiation are both significantly influenced by the WNT pathway (Baeg et al., 1995). Changes in this pathway may be the cause of the formation of tumours (Christie et al., 2013). Weakened tight junctions are also linked to WNT pathway changes in CRC, which promote migration and metastasis by reducing cellular adhesion (Brocardo & Henderson, 2008). Although many other changes can also target this system, APC mutations are the primary genomic aberration in CRC connected to the WNT pathway. β-catenin and c-MYC are also involved in the WNT pathway (Miller & Randall, 1996; Rennoll & Yochum, 2015; Segditsas & Tomlinson, 2006).

As the primary cell-cycle checkpoint and one of the most significant tumour suppressor genes, TP53 absence can promote the growth of tumours by enabling uncontrolled proliferation. Encoding the proteins that control the cell cycle, DNA repair, senescence, and apoptosis, is referred to as the “guardian of the genome” (Levine, 1997). Loss of p53-mediated apoptotic pathways is a crucial factor in the development of malignant tumours from adenomas (Sigal & Rotter, 2000). TP53 alterations are reported in the majority of CRC cases (Smith et al., 2002).

1.2 Next Generation Sequencing

In the last decades, Next Generation Sequencing (NGS) has been improving and providing better and faster data for genomic research. NGS creates the opportunity to have the entire genome sequenced within a day, surpassing the conventional methods used that would take a long time. NGS is a parallel sequencing technology that sequences millions of DNA fragments multiple times to provide high dept and accurate data. These fragments are mapped against a reference genome to ensemble the entire genome or parts of it. There are a lot of different technologies and areas for potential use of NGS like clinical research, microbiology, ontology and more.

Improvements in the omics areas have helped to better understand the biological processes in cancer. These areas of research include genomics, transcriptomics, and epigenomics.

RNA sequencing (RNA-seq) has been progressing over the past years. With Next Generation Sequencing, RNA-seq has expanded the understanding of many diseases providing biological information at the molecular level and allowing to profile the transcriptome. Thus, RNA-seq has been used in studies on cancer properties, on differential gene expression and biomarkers discovery, contributing to many fields of research, especially cancer research (Hematol et al., 2020; Leblanc & Marra, 2015; H. Xu, Wang, Song, Xu, & Ji, 2019).

Whole exome sequencing (WES) is another sequencing technique that focuses only on mRNA coding regions (exons). WES is a less expensive method than whole genome sequencing, that provides DNA information to investigate genetic alterations, like mutations such as singular nucleotide polymorphisms (SNPs) which can affect the gene's function when present within a gene or in a regulatory region near a gene, giving valuable insights in the knowledge of genetic disorders and diseases as cancers (Liang et al., 2012; Rabbani, Tekin, & Mahdieh, 2014).

Furthermore, combining gene expression with gene mutation data has been proven to improve prediction accuracy, revealing the benefits of incorporating multiple data types (Gerstung et al., 2015; Matos et al., 2019).

1.2.1 Biomarkers

Along the years the term "biomarker" has had different definitions in which the easier one could be to call it a biological marker. A more precise definition could be "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" ("Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.," 2001). With the advancements in technology and science, proteomic and genomic analysis have become biological measurements that can also be considered as biomarkers. Biomarkers are useful for various applications in the treatment of cancer and can be categorized as diagnostic, prognostic, or predictive biomarkers. They can be used to assess the risk of a disease, monitor and predict the progression of the disease, and predict the response to treatment. There is a wide range of biomarkers, which can include, among other things, proteins (such as an enzyme or receptor), nucleic acids (such as a microRNA or other non-coding RNA), antibodies, and peptides. Gene expression, proteomic, and metabolomic signatures are a few examples of the types of modifications that can also be considered biomarkers (Griffiths et al., 2002).

In order to be evaluated non-invasively and serially, biomarkers can be found in the circulation (whole blood, serum, or plasma), excretions or secretions (stool, urine, sputum, or nipple discharge), or they can be tissue-derived, in which case a biopsy or specialized imaging is necessary. Genetic biomarkers can be somatic and found as mutations in DNA taken from tumour tissue, or they can be inherited and found as sequence differences in germ line DNA recovered from whole blood, sputum, or buccal cells. Due to the critical role that biomarkers play at all stages of disease, it is important that they undergo rigorous evaluation, including analytical validation, clinical validation, and assessment of clinical benefits, before they are integrated into routine clinical care.

In simpler words, a biomarker is a biological component that can be used to track the presence or development of a certain disease or its therapeutic outcome/effects. High sensitivity, specificity, and safety are just a few of the crucial qualities that biomarkers must have. They also need to be simple to measure, helpful for making a precise diagnosis, and safe (Diamandis, 2010).

1.2.2 Metastasis Biomarkers

The primary cause of death in cancer is metastasis which shows the preeminent importance of discovering metastasis biomarkers to aid cancer treatment (Ferlay et al., 2010). Different sites of metastasis and different subtypes of CRC require different treatments which impact patients' survival outcomes (Hart & Fidler, 1980; Toraih et al., 2021). Advances in our understanding on the mechanisms behind cancer metastasis and in technology have allowed the finding and incorporation of metastasis related biomarkers in targeted treatment, improving the overall survival rates of late-stage patients (Brinton, Brentnall, Smith, & Kelly, 2012; Oh & Joo, 2020). One area of particular interest is the determination or prediction of the dissemination potential that a tumour has. Thus, it would be of tremendous benefit for cancer therapy to identify biomarkers capable of predicting the occurrence of metastases and even the sites of metastasis (Peixoto et al., 2023). Although promising, these types of biomarkers are poorly investigated due to the challenging and demanding procedure to obtain data for these analyses. Large cohorts need to be gathered at the initial stages of cancer before any treatments and the patients need to be followed for a long period of time, several years, to assess if there will occur metastasis.

1.3 Biomarkers in Colorectal Cancer

As time goes by, our knowledge and understanding of CRC characteristics has been improving, providing the possibility to discover more biomarkers that aid the treatment of CRC patients improving survival rates. CRC biomarkers should be simple to quantify, extremely sensitive and specific, reliable, and easy to reproduce. Despite the continuous discovery of potential new biomarkers, only a few are used clinically (Oh & Joo, 2020). Next are described recent biomarkers for diagnosis, prognosis, and prediction in CRC.

1.3.1 CRC Diagnostic Biomarkers

Tissue Biomarkers

Cytokeratins (CKs) The intracytoplasmic cytoskeleton of epithelial tissue contains keratin proteins called CKs). Staining patterns of CKs have revealed to be good for diagnosis of CRC metastasis especially the CK7-/CK20+ pattern (CK20 is detected in the normal gland cells of the colonic mucosa and in contrast CK7 isn't present) (Bayrak, Haltas, & Yenidunya, 2012; Bayrak, Yenidünya, & Haltas, 2011).

Caudal type homeobox2 (CDX2) CDX2 is responsible for coding a protein involved in the regulation of normal cell differentiation in the GI tract and tumour suppression in the colon. High expression levels of CDX2 were found correlated with the development of CRC but also with other types of cancer so it can be useful when combined with other biomarkers (Moskaluk et al., 2003; Werling, Yaziji, Bacchi, & Gown, 2003).

Special AT-rich sequence binding protein2 (SATB2) SATB2 has shown positive expression in 95% of metastatic CRC revealing it usefulness as a diagnostic marker (Moh et al., 2016; Perez Montiel et al., 2015).

Cadherin 17 (CDH17) Cadherins maintain tissue structure since they are cell-cell adhesion molecules and are reported to be expressed in 96%-100% in primary CRC and 100% in metastatic CRC (Gumbiner, 1996; Panarelli, Yantiss, Yeh, Liu, & Chen, 2012; Su, Yuan, Lin, & Jeng, 2008).

Telomerase A ribonucleoprotein that adds TTAGGG repeats onto telomeres to maintain them (Shay, Zou, Hiyama, & Wright, 2001). Upregulation of telomerase allows cancer cells to bypass pathways responsible for DNA damage response. A telomerase study reported 95% sensitivity and specificity in CRC (Roig, Wright, & Shay, 2009).

GPA33 (A33) Expressed in the stomach, colon, small intestine, and epithelial cells, A33 has shown to be expressed in practically all CRC (Garinchesa et al., 1996). A study revealed A33 to have high sensitivity as CDX2 to CRC and higher specificity which shows its potential as a diagnostic biomarker (N. A. C. S. Wong et al., 2017).

Blood Biomarkers

Circulating cell-free DNA (cfDNA) In tumour cells there is presence of larger fragments of cfDNA (Jahr et al., 2001). Measuring the ratio of DNA fragments works as quantification of cfDNAs and has shown useful to CRC diagnosis with sensitivity levels of 73%-90% and specificity of 85%-97% (El-Gayar, El-Abd, Hassan, & Ali, 2016; Hao et al., 2014).

MicroRNA (miRNA) MiRNA are small noncoding RNAs with 18-25 bp that bind to mRNA to regulate gene expression (Mitchell et al., 2008). MiRNAs are highly stable in the blood and monitorization of groups of different miRNAs (mi-21, mi-320a and mi-423-5q) have shown high levels of specificity and sensitivity for CRC (Z. Fang et al., 2015).

Long noncoding RNA (lncRNA) Involved in many biological processes including differentiation, immunological responses, chromosome dynamics, and epigenetic regulation, lncRNAs have been reported to be associated to more than 150 human diseases such as leukemia, breast cancer, and colon cancer (Gong, Tian, Qiu, & Yang, 2017). HIF1A-AS1, CRNDE-h, NEAT1, ZFAS1, and GAS5 are lncRNAs revealing promising results and potential to be used as diagnostic biomarkers of CRC (C. Fang et al., 2017; Gong et al., 2017; L. Liu et al., 2018; T. Liu et al., 2016; Peng, Wang, & Fan, 2017).

Insulin-like growth factor binding protein 2 (IGFBP-2) Malignancies of the ovary, colon, and prostate as well as other tumours have been associated with high levels of serum IGFBP-2 (Eiseman et al., 2007; el Atiq, Garrouste, Remacle-Bonnet, Sastre, & Pommier, 1994). IGFBP-2 sensitivity and specificity for early CRC is unsatisfactory but combining it with other biomarkers has shown promising value as a diagnostic biomarker (Liou et al., 2010; Renehan, Jones, Potten, Shalet, & O'Dwyer, 2000).

Stool Biomarkers

Guaiac fecal occult blood test (gFOBT) CRC mortality has been reduced by 11%-33% over the past years by using gFOBT as a screening test (Mandel et al., 1993). It has limitations in distinguishing upper gastrointestinal bleeding from lower or non-human heme from human heme (Kuipers, Rösch, & Bretthauer, 2013).

Fecal immunochemical test (FIT) By detecting human globin with a human hemoglobin-scpecific immunoassay, FIT has high specificity and sensitivity than gFOBT (Ahlquist, Harrington, Burgart, & Roche, 2000; Zou, Harrington, Klatt, & Ahlquist, 2006).

Stool DNA (sDNA) A multi-target stood DNA test for CRC called the Cologuard test revealed higher sensitivity than the last two tests but also a higher rate of false positives (Imperiale et al., 2014; Imperiale, Ransohoff, Itzkowitz, Turnbull, & Ross, 2004).

1.3.2 CRC Prognostic Biomarkers

Tissue Biomarkers

BRAF BRAF gene from the RAF family, has been associated with the development of CRC when mutations are present and has been used as a predictor of sporadic CRC (Aprile, Macerelli, Maglio, Pizzolitto, & Fasola, 2013; Fransén et al., 2004; R. Wong & Cunningham, 2008). BRAF mutations have also been related to bad prognosis and bad response to the anti-EGFR therapy (Kalady et al., 2012; Sartore-Bianchi et al., 2009).

MSI CRC patients with MSI have shown to have low aggressivity and good prognosis representing better outcome. Several studies support that MSI tumours indicate a bad response to 5-fluorouracil (5-FU) adjuvant chemotherapy but, in the other hand, a good response to irinotecan (Bacher, Flanagan, Smalley, & Nassif, 2004; Geiersbach & Samowitz, 2011; B Iacopetta & Watanabe, 2006).

CIMP Studies have shown that CRC patients with CIMP-/CIMP- high have better prognosis than CIMP+/CIMP- high (Jia, Gao, Zhang, Hoffmeister, & Brenner, 2016).

APC Most sporadic CRC cases and FAP have an association with APC (Carethers & Jung, 2015). Cytoskeletal integrity, motility, cellular proliferation, and apoptosis are associated with the WNT signalling pathway in which APC has an important role by regulating β-catenin. Rising the levels of β-catenin can influence cell proliferation since it also causes an expression increase of c-myc (Narayan & Roy, 2003). Consequently, APC mutations can be responsible for the unregulated transcription of many oncogenes and it has been reported that these mutation and high miR-21 are associated with poor survival (T.-H. Chen et al., 2013).

P53 Mutations in the TSG p53 have been associated with the majority of CRC cases (Lech, Słotwiński, Słodkowski, & Krasnodębski, 2016). Being responsible to repair damaged DNA or induce apoptosis, p53 is a key factor in the cell cycle and in suppressing tumours (Carethers & Jung, 2015). Some studies have revealed the prognostic value of p53 mutations in CRC patients (Allegra et al., 2003; Russo et al., 2005; Westra et al., 2005), but others have reported no indication of a prognostic role (Petersen, Thames, Nieder, Petersen, & Baumann, 2001; Popat et al., 2006).

SMAD4 The TGF-β superfamily signalling pathway is associated with the TSG SMAD4 and mutations in this gene are linked to cell differentiation, proliferation, cell migration, and apoptosis (Nikolic et al., 2011; Y. Xu & Pasche, 2007). SMAD4 mutations are reported in 30%-40% of CRC cases and loss of SMAD4 has been associated with poor survival revealing it's potential value as a prognostic biomarker (Riggins, Kinzler, Vogelstein, & Thiagalingam, 1997; Salovaara et al., 2002; Voorneveld, Jacobs, Kodach, & Hardwick, 2015).

Blood Biomarkers

CEA Patients with stage II or III CRC are recommended to do CEA tests every 3 months post-surgery (Locker et al., 2006). CEA levels have been reported to be strongly correlated to CRC patient outcomes (Park et al., 1999). Two studies revealed a significant association between preoperative CEA levels and prognosis in CRC patients where it occurred metastasis to the liver (Fong, Fortner, Sun, Brennan, & Blumgart, 1999; Nordlinger et al., 1996).

Neutrophil-to-lymphocyte ratio (NLR) Neutrophilia is linked to systemic inflammation, whereas lymphopenias is linked to impaired cell-mediated immunity (Grivennikov, Greten, & Karin, 2010). NLR has been investigated as a marker for immune responses to diverse stressful situations (Zahorec, 2001). Elevated levels of NLR have been reported to be associated with shorter overall survival after treatment in primary cases and also patients with liver metastasis (Tang et al., 2016; Tsai, Su, Leung, Lai, & Liu, 2016).

Circulating free DNA (cfDNA) Poor survival and high probability of recurrence have been associated with higher cfDNA concentrations. Postoperative metastasis/recurrence rates have been reported to be considerably greater when there were detected APC, KRAS and p53 mutations.

1.3.3 CRC Predictive Biomarkers

Tissue Biomarkers

KRAS, NRAS In metastatic CRC patients, mutations in the KRAS gene have been used as biomarker of bad response to the anti-EGFR (epidermal growth factor receptor) antibody-based therapy (Chang et al., 2016; De Roock et al., 2010). In addition, mutations in NRAS (a gene related to KRAS), also shows a negative response to the anti-EGFR therapy showing the usefulness of using extended RAS (KRAS and NRAS) as negative biomarkers predicting the outcome of anti-EGFR therapy (Cercek et al., 2017; Schirripa et al., 2015; Sforza et al., 2016).

BRAF BRAF has shown potential to be beneficially used in anti-EGFR antibody therapy as a predictive biomarker, but it still lacks sufficient evidence(Rowland et al., 2015).

PIK3CA Some studies have reported the potential of PIK3CA as a predictive biomarker for anti-EGFR therapy in colon cancer (Perrone et al., 2009; Sartore-Bianchi et al., 2009), while other studies have reported that it has no potential, leaving this marker with inconclusive and indecisive results, and in the need of future investigation (Karapetis et al., 2014).

Blood Biomarkers

Cell-free DNA cfDNA concentrations have been found to decrease after primary treatment (resection and chemoradiotherapy), however, when there is recurrence or no treatment response, cfDNA levels show a significantly high increase (Spindler, Pallisgaard, Andersen, Brandslund, & Jakobsen, 2015; Zitt et al., 2008).

Some other potential biomarkers that are emerging are the location or sidedness, the consensus molecular subtype (Puccini, Seeber, & Berger, 2022). Left and right-side CRC have shown result in a different tendency for the metastasis location. Right sided CRC patients tend to have peritoneal carcinomatosis while left sided tend to have lung or liver metastasis (Benedix et al., 2010). As said before, the choice of treatment plans for CRC patients can be considerably improved by the identification of suitable biomarkers. The majority of these markers can inform doctors about the disease's general prognosis and therapy outcome. Although many molecular biomarkers have been discovered lately with good and promising results, they are not yet used in medical practice. This happens because most of the studies have small sample sizes or are retrospective analysis of a unique marker, resulting in lack of resolution and reproducibility. Data interpretation and analysis continue to be difficult tasks. Data obtained frequently lack proper definition and validation, making them unreliable for use in clinical settings. Additionally, it is challenging to quantify and evaluate the obtained data due to the lack of consistent methods and standardized endogenous controls. Despite all of these real drawbacks, a lot of work is being put into this problem, and the use of biomarkers in the diagnosis and prognosis of CRC as well as in the creation of individualized and targeted therapies has a bright future (Malki et al., 2021; Mármol et al., 2017).

1.4 Thesis Objectives

Metastasis remains the main cause of cancer death revealing the importance carried by the challenging studies in this field. CRC is a very heterogeneous disease which makes its comprehension even more difficult. Progress to better understand metastatic CRC biological mechanisms is needed to find more and better biomarkers to help the treatment outcome in CRC patients. Although many studies are trying to find prognostic and predictive biomarkers, the search for factors in early-stage CRC that may influence metastasis occurrence remains poor.

In this study we aim to unveil promising biomarkers of metastatic potential and build a predictive model for CRC metastasis occurrence. Gene expression profiling, using transcriptomic data (RNA-seq) and genomic data (WES) from a Portuguese institution, will be done to discover discrepancies between primary CRC and metastatic CRC patients. RNA-seq data will be used to find differential expressed genes which will be used in gene set enrichment analysis to explore potential pathways with impact in CRC metastasis. The differential expression values will also be the data used for the predictive model. Statistical tests will also be conducted in the genomic data to search for possible relations between gene mutations and metastasis occurrence.

2 Materials and Methods

The development of the current project was performed with R and RStudio (version 4.1.1) which is a software often used for statistical computing and graphics production. R is complemented by packages with a variety of statistical and graphical resources good for data manipulation, analysis, and visualization. In this case, to accomplish the objectives said previously, I mostly used Bioconductor which is an open-source software for bioinformatics available in R and has packages for various types of analysis of biological data.

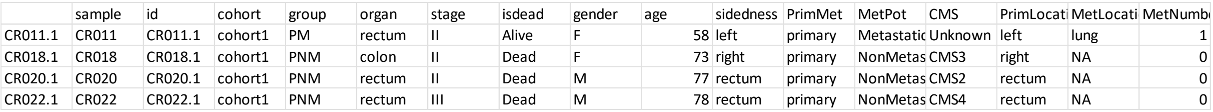
2.1 Sample Information

A total of 154 CRC samples were collected from Santa Maria Hospital by Dr. Luís Costa's team. Paired normal mucosa from each patient was taken from more than 2 cm away from the tumour. These patients did not receive any type of chemo or radiotherapy prior to sample collection and were supervised for at least 3 years to register if there were metastasis occurrence or not. For each patient there is medical information about gender, age, the stage of the cancer, the CRC consensus molecular subtype (CMS), if the patient died, if there were metastasis, the location of the first tumour, including the side, the locations of the metastases and the number of metastases. Within this 154, there were 36 patients where metastasis occurred. From all these patients there was transcriptomic (RNA-seq) and genomic (WES) data from the primary tumour (when disease was first detected and before any treatment). RNA-seq data had 5 replicates which makes a total of 31 metastatic patients in 149.

2.2 Sample Metadata

The information for each patient was imported to the R environment in table format. The table contained the information for each patient said before such as CMS, stage, metastasis occurrence, metastasis location and others, as shown for a few samples in Table 1. This data was summarized into another table using the R package gtsummary (version 1.5.0) with the objective to compare the metastatic samples versus the not metastatic. By doing this we could have a first impression if there was from the beginning a bias for the metastatic samples or even, if it was significative, a possibility to encounter promising biomarkers.

Table 1 - CRC patients information



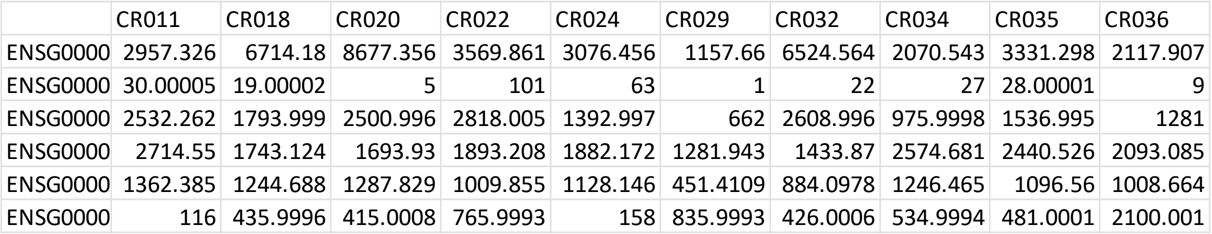
	sample	id	cohort	group	organ	stage	isdead	gender	age	sidedness	PrimMet	MetPot	CMS	PrimLocati	MetLocati	MetNumb
CR011.1	CR011	CR011.1	cohort1	PM	rectum	II	Alive	F	58	left	primary	Metastatic	Unknown	left	lung	1
CR018.1	CR018	CR018.1	cohort1	PNM	colon	II	Dead	F	73	right	primary	NonMetas	CMS3	right	NA	0
CR020.1	CR020	CR020.1	cohort1	PNM	rectum	II	Dead	M	77	rectum	primary	NonMetas	CMS2	rectum	NA	0
CR022.1	CR022	CR022.1	cohort1	PNM	rectum	III	Dead	M	78	rectum	primary	NonMetas	CMS4	rectum	NA	0

2.3 Transcriptomic

2.3.1 RNA-seq Counts Matrix

The dataset used for the majority of this project was a matrix containing the read counts representing the gene expression levels, resulted from RNA-seq. Whole transcriptome sequencing was conducted with Illumina technologies creating reads libraries. From the RNA-sequencing reads, kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) was used to estimate the transcript abundances and lastly the values were normalized to counts per millions creating the counts matrix. The counts matrix reflects for each sample, how many times a gene/transcript is read. Kallisto does transcript quantification in a faster way known as pseudoalignment which consists in skipping the alignment step by selecting the transcripts that a read is compatible with to quantify the transcript. The resulting matrix was composed by 154 columns representing the CRC patients and by 60564 rows representing genes. A sample of the table can be observed in Table 2.

Table 2 - Counts Matrix (gene expression): each row represents a gene and each column a patient.



	CR011	CR018	CR020	CR022	CR024	CR029	CR032	CR034	CR035	CR036
ENSG0000	2957.326	6714.18	8677.356	3569.861	3076.456	1157.66	6524.564	2070.543	3331.298	2117.907
ENSG0000	30.00005	19.00002	5	101	63	1	22	27	28.00001	9
ENSG0000	2532.262	1793.999	2500.996	2818.005	1392.997	662	2608.996	975.9998	1536.995	1281
ENSG0000	2714.55	1743.124	1693.93	1893.208	1882.172	1281.943	1433.87	2574.681	2440.526	2093.085
ENSG0000	1362.385	1244.688	1287.829	1009.855	1128.146	451.4109	884.0978	1246.465	1096.56	1008.664
ENSG0000	116	435.9996	415.0008	765.9993	158	835.9993	426.0006	534.9994	481.0001	2100.001

2.3.2 Principal Component Analysis

High-dimensionality data such as in gene expression makes sample visualization challenging and restricts straightforward data exploration. PCA is a statistical algorithm that reduces the dimensionality of the data while preserving the majority of its variation. It achieves this reduction by locating the principal components, or directions, along which the data's variation is greatest. Each sample can be represented by fewer numbers rather than the values for thousands of variables by utilizing a small number of components. Then, after the samples have been displayed, it will be simple to visually compare the samples to see if they can be grouped. PCA increases the interpretability of the data revealing the dispersion and relations between the variables (Everitt & Howell, 2005).

PCA was conducted for the transcriptomic data (counts matrix) resulted from the RNA-seq. The sample’s characteristics of interest to investigate their distribution was the metastatic potential and the CMS. For this purpose, it was used the R packages PCAtools (version 2.4.0) and EdgeR (version 3.34.1). First the data was normalized using the functions calcNormFactors and cpm from the EdgeR package. After normalization, the PCA is conducted with the PCAtools’ function called pca with the expression data and the parameter of metadata being the clinical data for each patient. PCA visualization was performed with PCAtools where the variables of interest, metastatic potential, and CMS groups, were selected for colour distribution. The 5 metastatic sample duplicates were also highlighted to check for batch effects.

2.3.3 Gene Expression Analysis

As described in (Y. Chen, McCarthy, Robinson, & Smyth, 2014; Love, Huber, & Anders, 2014) and at https://github.com/dsobral/ADER, a differential expression analysis usually starts with a “raw” read counts matrix for all genes of each sample. The first step is normalizing the counts due to differences in the number of reads per sample. Commonly, for a sample, each gene count is divided by the total number of millions of reads of that sample, resulting in counts per million (cpm). Applying a statistical test for differential expression comes after normalization. The majority of these widely used techniques are based on derivations of the binomial distribution because sequencing data is based on discrete counts. The negative binomial distribution appears to suit the normalized gene expression distribution the best for the majority of studies. Since low expressed genes vary more in terms of percentage of gene expression and highly expressed genes vary more in terms of absolute value, it is evident that this variation is gene dependant (fold change). Low expressed genes are more likely to have differential expression if one only considers fold change and ignores variation. As a result, we must precisely measure variation per gene. However, the number of replicates we often have is small and insufficient for precise parameter estimate. Therefore, we must devise a method for creating such estimates, and to do so, we require specific tools. The parameters of the distribution and their variance, or how much the (normalized) counts vary between the various samples, must be estimated after normalization. To do normalization, calculate variance, and run statistical tests for differential expression, there are numerous free programs available. The R packages DESeq2 and Edger are the most widely used and have shown to perform well in most situations.

The normalization considering the median of the gene expression ratio is applied by DESeq while EdgeR employs a comparable, albeit more complex, methodology using the trimmed mean of M-values (or TMM in short). The log fold change (M-value) between any sample and a reference sample is assumed by TMM to be about 0 for most genes. Genes with extreme M-values and extreme absolute expression values (A) are deleted (trimmed) from the normalization factor computation, and genes with lower variance are given a higher weight. EdgeR and DESeq2 employ the method of binning genes with comparable expression and fitting a curve to estimate variances under the assumption that genes with similar expression have similar variances. The estimated mean difference in gene expression across groups of samples is then updated using the parameter that was used to build the curve as a baseline. Finally, the variance is rescaled by making it constant at all bin levels under the presumption that the majority of genes do not express themselves differentially. Then, we check each gene for differential expression and calculate the test's likelihood. Since we examine a large number of genes, some genes may receive favourable p-values by random chance. We must therefore perform multiple test corrections. The number of tests multiplied by each p-value is one approach (a method called Bonferroni correction). However, this is excessively rigid, and we frequently end up with nothing presented differently. The most used technique is Benjamin-Hochberg, which decreases the false discovery rate by applying a correction proportionate to the ranking of the p-value. Filtering out the genes with extremely low expression across all samples is another method for reducing the number of tests improving the analysis's speed.

Firstly, three tests were conducted with three different packages, namely EdgeR, DESeq2 (version 1.32.0), and limma (version 3.48.3). This packages work with the reads count matrix from the CRC patients, and a vector with the condition variables, metastatic and not metastatic, following a typical workflow to identify DEGs with RNA-seq data performing normalization, estimating variance, and performing statistical tests for differential expression.

- EdgeR started by filtering with the filterByExpr function to remove genes with extremely low counts before normalizing the data using the trimmed mean of M-values (TMM) technique. After normalization it estimated variance for each gene (see introduction). For the differential analysis, the quasi-likelihood F-test was next applied.
- Limma-voom, the filtering and normalization steps are the same as in EdgeR. The normalized and filtered count matrix was then transformed using the voom algorithm, and the lmFit and eBayes functions were used to do the differential analysis.
- Lastly, for DESeq2, the filtering step was done manually removing the genes that didn't have at least a total of 10 reads. The variance estimation was done as explained in the introduction, the differential analysis was performed by the DESeq function, and the results are generated by the results function.

The selection of DEGs was based on the false discovery rate threshold less than 0.05 (FDR < 0.05).

The definitive differential gene expression analysis was conducted mixing the step of filtering lowly expressed genes using the filterByExpr function from EdgeR with the differential analysis performed with DESeq2. As the genes were represented by their code and not their names, a file with the respective names for each code was imported to the environment and for each gene code it was matched the corresponding name. The resulting table from the differential analysis was ordered in ascending order by the adjusted p-value, to have the most differential expressed genes in the top of the table for further analysis. The results from both gene expression analyses were cross-referenced to verify DEGs similarity.

2.3.4 Gene Expression Visualization

To increase the specificity and significance of the discovered DEGs they were filtered by the following conditions of log2FC and adjusted p-value: |log2FC| > 1 and padj < 0.05. The R package EnhancedVolcano (version 1.10.0) was used to visualize DEGs expression levels and distribution in a volcano plot, using the results from DESeq2.

Heatmaps and boxplots were used to represent the gene expression distribution of some of the most relevant DEGs by all the patients. From the DESeq2 analysis it is possible to identify the 20 most differential expressed genes. The heatmap was created with the R function heatmap.2 and the boxplot with the ggplot2 package which both receive the matrix with the gene expression levels for all patients of the desired genes.

2.3.5 Gene Set Enrichment Analysis

GSEA is a method used to evaluate expression data in the level of gene sets or pathways. It identifies the gene sets that are over-represented in a large group of genes, usually between two different conditions. These sets are created *a priori* based on biological information whereas the genes share common properties. By identifying pathways and processes contrary to single-gene methods, this technique eases the interpretation of large-scale experiments. GSEA has demonstrated to give insights in cancer-related data sets revealing its importance in biomedical research (Subramanian, Tamayo, Mootha, Mukherjee, & Ebert, 2005).

This method works with the gene expression values of each gene which represent a ranking system. GSEA can be divided in three main steps. First it starts by calculating the enrichment score (ES) to measure the extent to which a gene set is overrepresented at either the top or bottom of the entire ranked list. For each gene set, GSEA basically goes through the ranking list of genes and does a running-sum statistic increasing each time it encounters a gene that belongs to that gene set and decreasing otherwise. The final ES is then determined in relation to this null distribution. Lastly the estimated significance level is adjusted for multiple hypothesis testing. To consider the size of the set the ES is normalized for each gene set resulting in a normalized enrichment score (NES). In the end it is determined the false discovery rate (FDR) by comparing the tails of the observed and null distributions for the NES. FDR represents the estimated probability that a set constitutes a false positive discovery (Subramanian et al., 2005).

Gene set enrichment analysis was conducted with the R package fgsea (version 1.18.0). From the DESeq2 analysis a ranking system was created with the values of log2FC of each gene. Reactome and cancer hallmarks were the gene sets selected for the gsea collected from https://www.gsea-msigdb.org/gsea/index.jsp. Reactome was considered since it comprises the biological pathways for the human being. The cancer hallmarks serve as an organizational framework for explaining the complexity of neoplastic disease. They include characteristics as maintaining proliferative signalling, avoiding growth inhibitors, avoiding cell death, allowing replicative immortality, generating angiogenesis, and triggering invasion and metastasis. The ten most enriched pathways, positively and negatively, are selected for visualization. Pathways enrichment was visualized with the R standard function barplot. Gsea was executed twice, once for the reactome gene sets and another for the cancer hallmarks gene sets.

2.3.6 Prediction Analysis of Microarrays

PAM is a statistical technique that uses Nearest Shrunken Centroid Method with gene expression data for class prediction. From a set of genes, this method identifies which ones best characterize the classes. In this study, the class of interest will be metastasis occurrence. PAM creates an accurate classifier that can work as a predictive model (Hastie, Tibshirani, Narasimhan, & Chu, 2019).

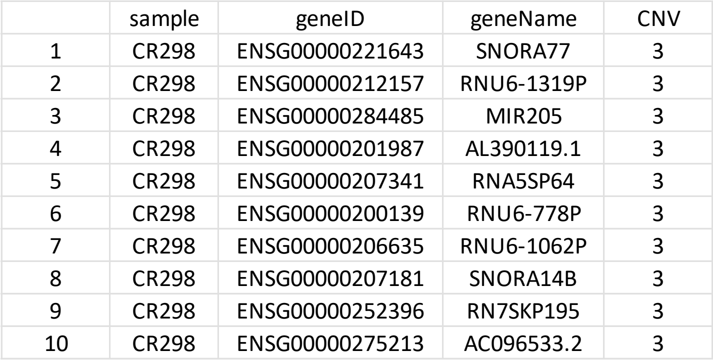
A classifier for metastatic potential was created using the pamr package (version 1.56.1). The normalized counts matrix representing the gene expression data was used by the program. A classification training for metastatic and not metastatic samples was executed with 70% of the samples. A threshold of 1.8 was manually selected by visualization of the training prediction efficacy while using a specific number of genes, shown graphically by a misclassification error plot. A test was performed for the other 30% data using the selected threshold and the information from the training step.

The genes used by the classifier and the DEGs from the DESeq2 analysis were intersected. A boxplot was created with the common genes to visualize the expression levels distribution among all samples.

2.4 Genomic

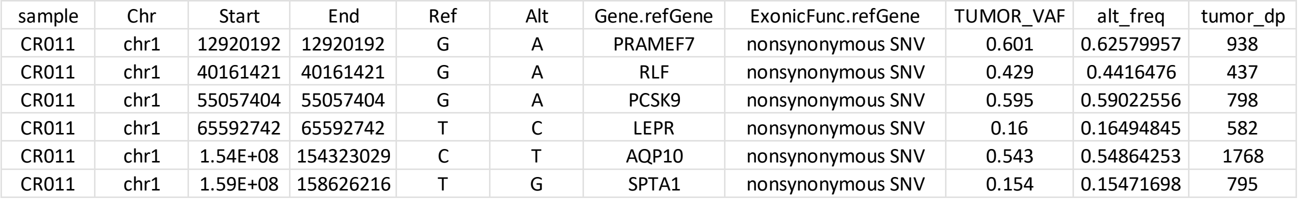
Mutations (CNVs and SNPs) were analysed in the genomic data originated from whole genome sequencing between metastatic and not metastatic samples. Two datasets were used in this analysis, one with copy number variations and another with singular nucleotide polymorphisms as partially shown in Tables 3 and 4. Statistical tests were conducted for both datasets to search for significant relations between mutations and metastatic potential. With fisher.test function from R fisher's tests were executed calculating p-values for each mutation types to validate the results. Lastly the p-values were adjusted for multiple testing using the p.adjust function from R with the false discovery rate (FDR) method.

Table 3 – Copy number variations from CRC patients.



	sample	geneID	geneName	CNV
1	CR298	ENSG00000221643	SNORA77	3
2	CR298	ENSG00000212157	RNU6-1319P	3
3	CR298	ENSG00000284485	MIR205	3
4	CR298	ENSG00000201987	AL390119.1	3
5	CR298	ENSG00000207341	RNA5SP64	3
6	CR298	ENSG00000200139	RNU6-778P	3
7	CR298	ENSG00000206635	RNU6-1062P	3
8	CR298	ENSG00000207181	SNORA14B	3
9	CR298	ENSG00000252396	RN7SKP195	3
10	CR298	ENSG00000275213	AC096533.2	3

Table 4 – Singular Nucleotide Polymorphisms from CRC patients.



sample	Chr	Start	End	Ref	Alt	Gene.refGene	ExonicFunc.refGene	TUMOR_VAF	alt_freq	tumor_dp
CR011	chr1	12920192	12920192	G	A	PRAMEF7	nonsynonymous SNV	0.601	0.62579957	938
CR011	chr1	40161421	40161421	G	A	RLF	nonsynonymous SNV	0.429	0.4416476	437
CR011	chr1	55057404	55057404	G	A	PCSK9	nonsynonymous SNV	0.595	0.59022556	798
CR011	chr1	65592742	65592742	T	C	LEPR	nonsynonymous SNV	0.16	0.16494845	582
CR011	chr1	1.54E+08	154323029	C	T	AQP10	nonsynonymous SNV	0.543	0.54864253	1768
CR011	chr1	1.59E+08	158626216	T	G	SPTA1	nonsynonymous SNV	0.154	0.15471698	795

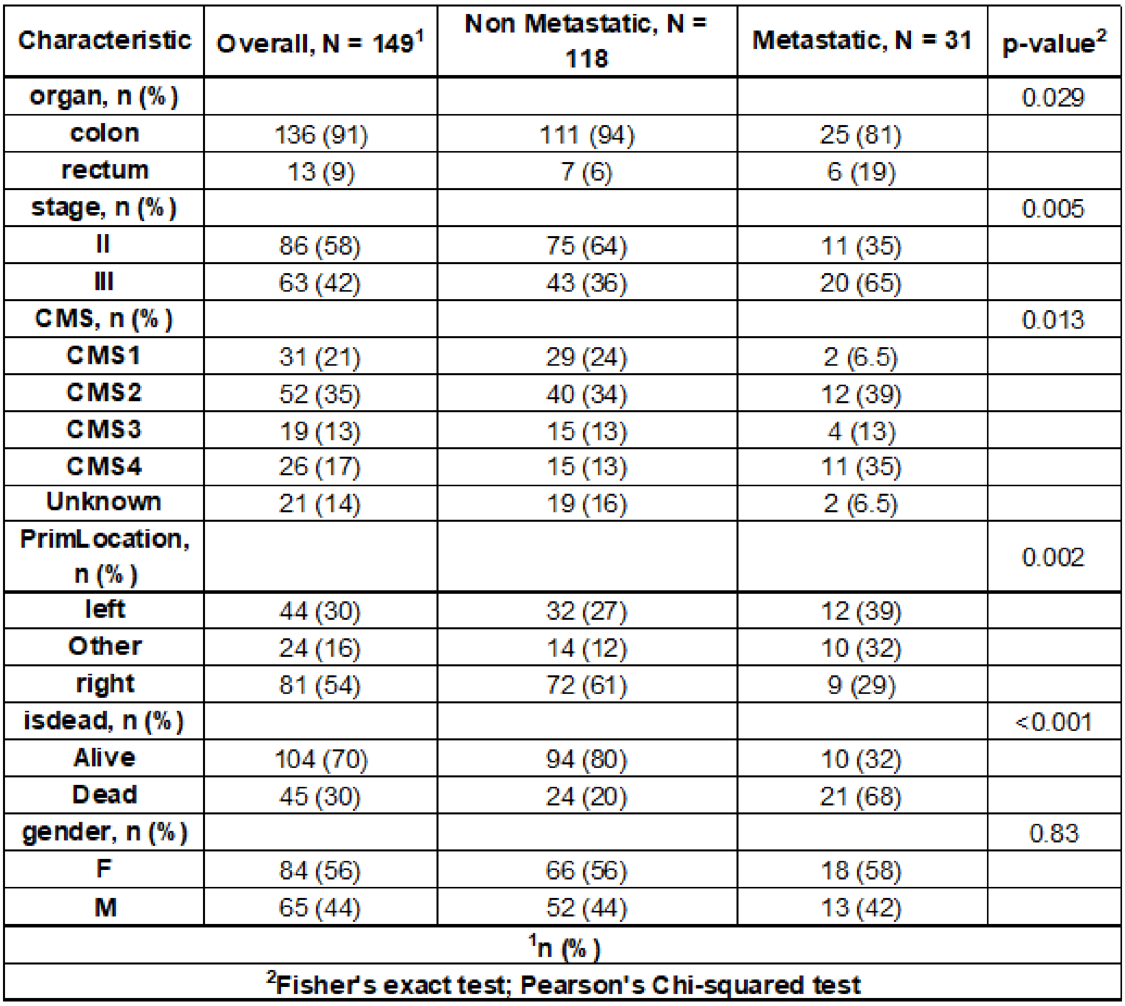
3 Results

The R code for all the upcoming results is available at https://github.com/v-fiori/CRC-Metastasis-Dissertation.

3.1 Sample Data Exploration

A total of 149 CRC patients were studied and for each one it was collected information such as age, gender, cancer stage, metastatic potential, tumour locations and others (see methods). Patient's data was explored as shown in Table 5 dividing two major groups, 31 metastatic and 118 not metastatic. Some characteristics revealed a tendency and possible relation with the presence of metastasis. Stage cancer III, with a percentage of 65% showed a higher incidence of metastasis than stage II with only 35%. The consensus molecular subtypes had two groups, CMS2 and CMS4, with higher percentage levels of metastasis, respectively 39% and 35%, and one group, CMS1, with a very low percentage of 6.5%. Another factor with opposite values was the organ of the primary tumour, whereas 81% of the metastatic samples were in the colon and only 19% in the rectum. The sidedness of the tumour revealed no differences in the metastatic individuals, but in the not metastatic ones the right side had a much bigger incidence of 61% versus the 27% in the left side.

Table 5 – CRC patient's data overview (metastatic versus not metastatic). Which test is conducted (Fisher's or Chi-squared) is unidentifiable since it is automatically done by the gtsummary package.



Characteristic	Overall, N = 149[1]	Non Metastatic, N = 118	Metastatic, N = 31	p-value[2]
organ, n (%)				0.029
colon	136 (91)	111 (94)	25 (81)	
rectum	13 (9)	7 (6)	6 (19)	
stage, n (%)				0.005
II	86 (58)	75 (64)	11 (35)	
III	63 (42)	43 (36)	20 (65)	
CMS, n (%)				0.013
CMS1	31 (21)	29 (24)	2 (6.5)	
CMS2	52 (35)	40 (34)	12 (39)	
CMS3	19 (13)	15 (13)	4 (13)	
CMS4	26 (17)	15 (13)	11 (35)	
Unknown	21 (14)	19 (16)	2 (6.5)	
PrimLocation, n (%)				0.002
left	44 (30)	32 (27)	12 (39)	
Other	24 (16)	14 (12)	10 (32)	
right	81 (54)	72 (61)	9 (29)	
isdead, n (%)				<0.001
Alive	104 (70)	94 (80)	10 (32)	
Dead	45 (30)	24 (20)	21 (68)	
gender, n (%)				0.83
F	84 (56)	66 (56)	18 (58)	
M	65 (44)	52 (44)	13 (42)	

[1] n (%)

[2] Fisher's exact test; Pearson's Chi-squared test

3.2 Transcriptomic

PCA

Clusters, outliers, and trends were explored with principal component analysis and are displayed in Figure 1. This graph demonstrates that there isn't a clear separation between the metastatic and not metastatic samples. Nevertheless, it can be observed a high concentration of the metastatic group in the upper quadrants of the plot with a higher number in the left one. The duplicates were highlighted to investigate the possibility of batch effect but as shown in the picture, almost all duplicates are proximal except for CR438. This could be to an error in one of the transcriptomic readings or procedures. On the other hand, a clear division in the CMS groups is noticeable in Figure 2, mostly of CMS1 and CMS2. Since the various subtypes are defined by expression levels of genes or pathways this division is expectable. Some samples had unknown CMS which creates a little mix in the groups.

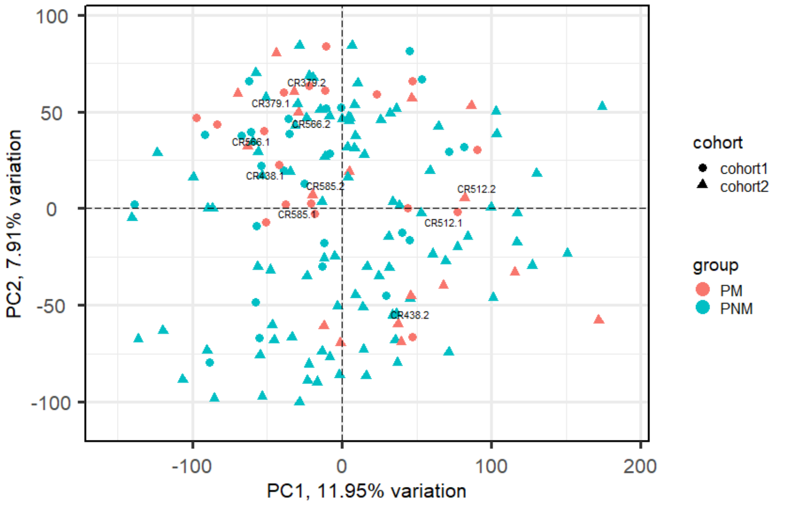


Figure 1 – Metastatic samples distribution in the PCA analysis.

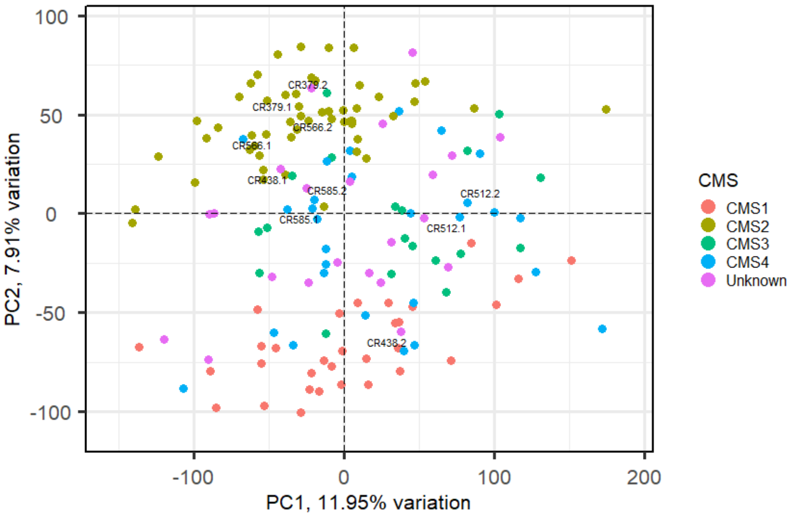


Figure 2 – CMS distribution in the PCA analysis.

DEGs Initial Tests

Next, we tested three different packages (EdgeR, DESeq2, limma-voom) to decide the best methods for differential expression analysis with this data. EdgeR was the first and, from a total of 60564 genes, FiltByExp removed the lowly expressed ones and resulted in 36962 genes. EdgeR executed the differentiation analysis in 36962 genes and resulted in 562 DEGs, being 127 down regulated and 435 up regulated as shown in Figure 3.

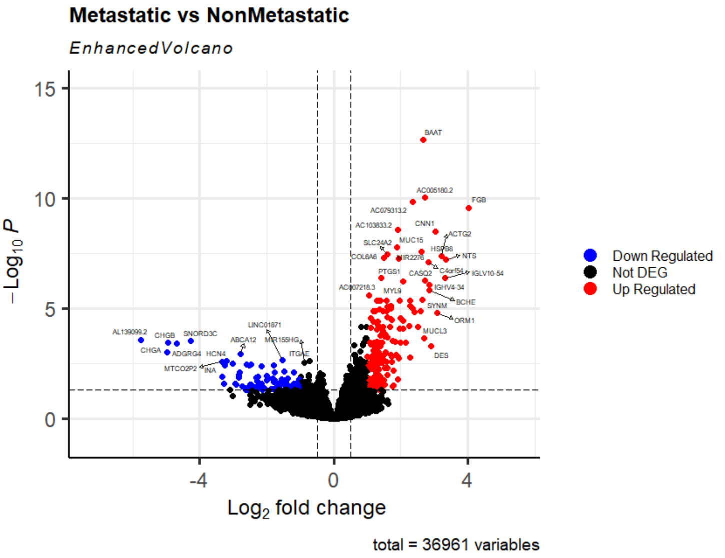


Figure 3 – Differential expressed genes distribution from EdgeR.

DESeq2 differential analysis resulted in fewer DEGs compared to EdgeR and the overall genes used was higher. The filtering step to remove the genes with less than 10 reads derived in 56710 genes for differential analysis which resulted in 319 DEGs. The differential expressed genes dispersion can be observed in Figure 4. As for the limma-voom test no DEGs were identified so that assessment is not represented graphically and limma-voom was excluded from this analysis.

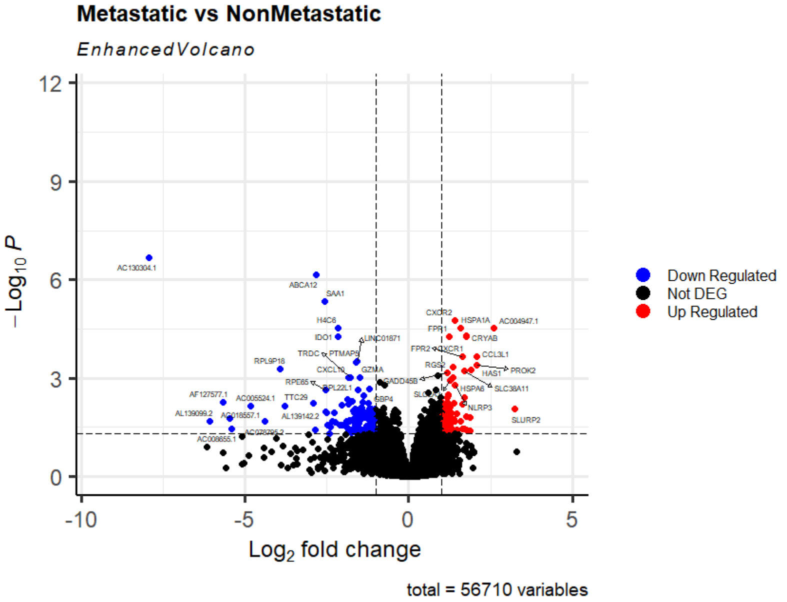


Figure 4 – Differential expressed genes distribution from DESeq2.

When the three tests were finished, we evaluated which package was best for this analysis. EdgeR had a good filtering step but ended in a great unbalance between down regulated and up regulated genes which is unusual. In the other hand, DESeq2 results were more balanced but the low expressed genes removed were very few. Removing low expressed genes is important because they can affect the outcome as they work as noise and removing them can also improve the speed of the differential analysis. Thus, our findings suggested that combining both packages using the FiltByExpr function from EdgeR for the filtering step and DESeq2 to estimate the variations and do the differential testing would contribute to optimal results.

DEGs Final Analysis

After the selection of the best method, we proceeded to do a final differential analysis. All genes were then subjected to a filtering step with the EdgeR FiltByExpr function. From a total of 60564 genes, 23549 were considered lowly expressed and for this reason removed, remaining 36962 for the differential analysis as occurred in the first trial test for EdgeR. The search for differential expression conducted with DESeq2 resulted in 438 DEGs displayed in Figure 5. The plot highlights the names of the most significant DEGs and two vertical divisions can be observed in the log2FC values of 1 and -1 and a horizontal division in p-value (-log10P) > 0.05. This delimitation was made to remove DEGs that had a very low level of differential expression causing a decrease in the number of DEGs to a final number of 166, of which 94 were down regulated and 72 up regulated.

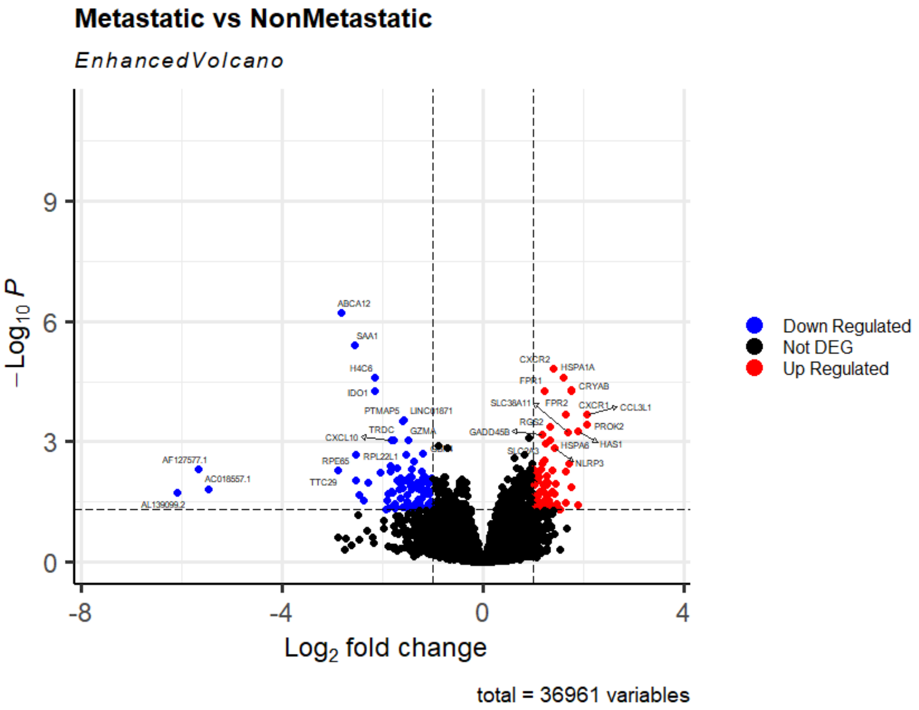


Figure 5 – Differential expressed genes distribution from EdgeR + DESeq2.

The differential analysis resulted in a table with the respective data for all the genes used. This table represents the significance of each gene in terms of differential expression between metastatic and not metastatic samples. The most relevant columns produced by DESeq2 results are “GeneID”, “Base Mean”, “log2(FC)”, “StdErr”,”P-value”, and “P-adj” which represent, respectively, the identifier of the gene, the mean normalized counts of all samples (how much the gene is expressed), the log2(FC) (when positive, more expressed in one group than the other and reverse when negative), a measure of the confidence in the true value of the estimated log2FC, a value measure of how likely it is to obtain the observed log2(FC) by chance and lastly the p-value corrected for multiple testing. The adjusted p-value was used to select the DEGs but all the other variables had to be considered to evaluate the significance and strenght of each gene in terms of expression differentiation.

Top DEGs Heatmaps

We further proceeded to explore the gene expression distribution of the top 20 DEGs with a heatmap and a boxplot. The selected genes expression levels throughout the 149 samples, defined by the colour key, can be observed in Figure 8 and 9 for the top 20. As expected, in the heatmap, although existing some inconsistency along the expression values, up regulated DEGs showed a red pattern in the metastatic area whereas the down regulated showed a blue pattern in the same area compared to the not metastatic area. Consistently, the boxplot also represented the difference in expression levels for the respective down and up regulated DEGs.

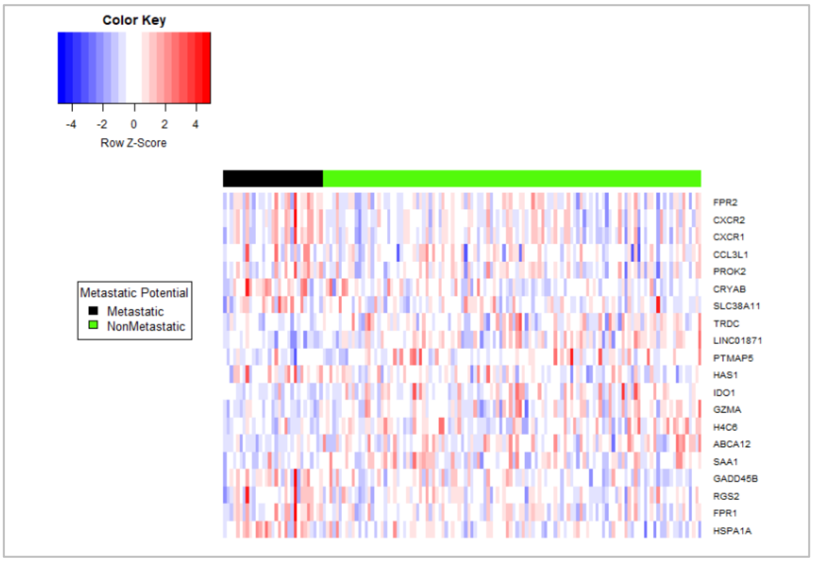


Figure 6 - Heatmap for the top 20 DEGs expression distribution throughout all patients.

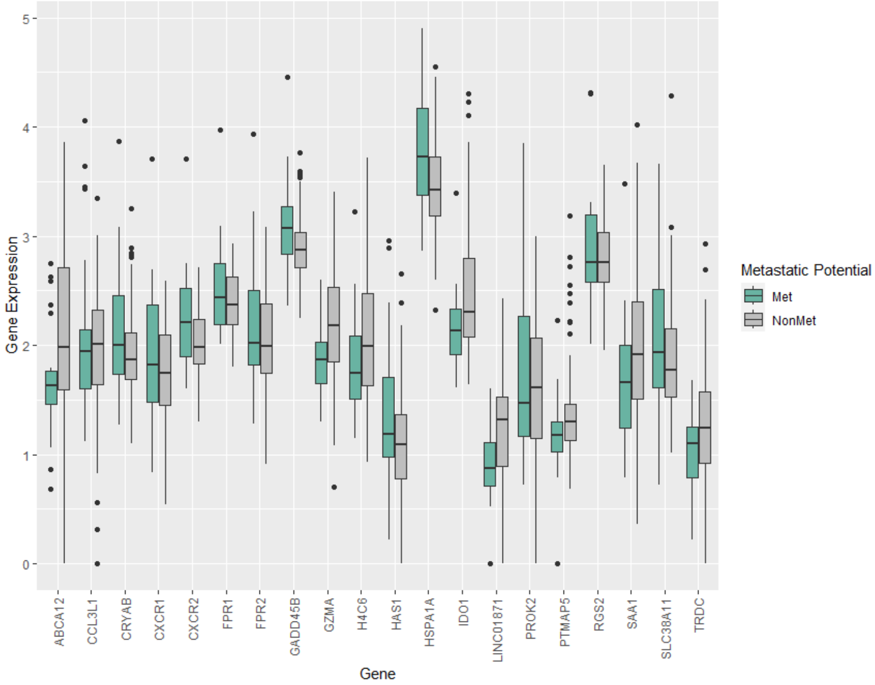


Figure 7 - Boxplot for the top 20 DEGs expression distribution throughout all patients.

GSEA

The gene set enrichment analysis revealed some degree of enrichment in the gene sets selected and the most enriched pathways in both ways, negatively and positively, are demonstrated in the Figure 8 for the Reactome pathways and in Figure 9 for the Hallmark pathways. As expected, some pathways are cancer related like cell cycle checkpoints, DNA repair, formation and activation mechanisms, KRAS, ABC transporters and MYC targeted pathways. One interesting result were that the most positively enriched pathway in the Hallmark gene set was the epithelial mesenchymal transition (EMT). This very dynamic process converts epithelial cells into a mesenchymal phenotype, involving the disruption of cell-cell adhesion and cellular polarity, changes in cell-matrix adhesion, and remodelling of the cytoskeleton. Studies have related a link between EMT and metastasis since it can enhance mobility, invasion, and cancer treatment resistance of tumour cells (Nieto, Huang, Jackson, & Thiery, 2016; Tsubakihara & Moustakas, 2018). Another positively enriched result was the hedgehog signalling pathway which affects tissue polarity, cell differentiation, and proliferation and has been shown to contribute to the spread and invasion of some types of cancer (Yao et al., 2018).

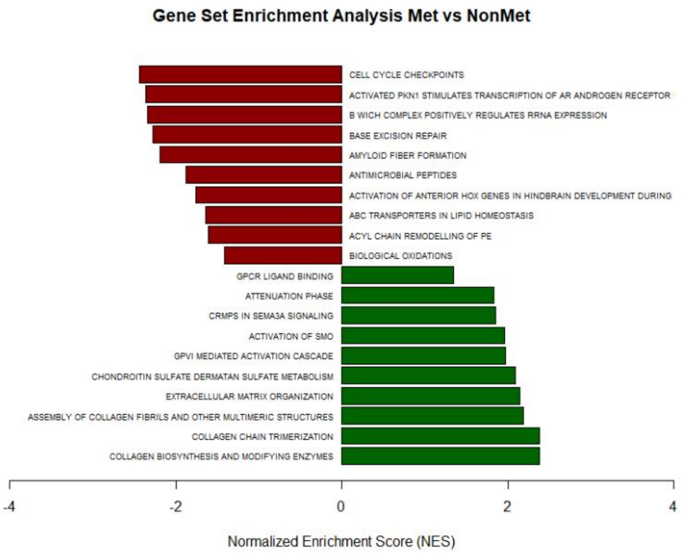


Figure 8 – GSEA with the Reactome dataset.

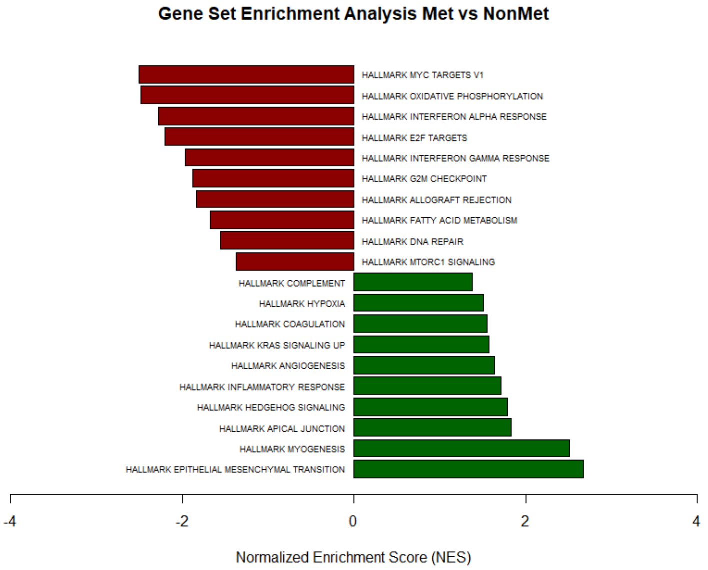


Figure 9 – GSEA with the Hallmark dataset.

The DEGs found in the enriched pathways were the following: "ABCA12", "SAA1", "HSPA1A", "CXCR2", "CXCR1", "FPR1", "HSPA6", "FPR2", "MMP12", "LTF", "CXCL11", "CYP24A1", "CXCL10", "CCL19", "PLAAT4", "LYZ", "LPAR4", "SULT1E1", "GAS1", "CALB1", "H4C6", "PRSS2", "CCL13", "PTGDR", "GPC5", "MUC16", "CXCL8", and "DEFA5" for the Reactome dataset; "HSPA1A", "IDO1", "FPR1", "HAS1", "MMP12", "GZMA", "CLDN18", "GADD45B", "LTF", "SLC2A3", "GBP4", "CXCL11", "NRXN2", "DPYSL3", "CXCL10", "CCL19", "MYH4", "HLA-DRA", "MYH1", "FABP3", "CSF3R", "PRF1", "CR2", "IFNG", "AQP9", "GAS1", "HLA-DRB1", "MT1E", "RBP4", "SOD3", "PRSS2", "IL1B", "CCL13", "PTGS2", "PRKCG", "CXCL8", "REEP1", and "NOS2" for the Hallmark dataset. The presence of these genes in the enriched gene sets, especially those that have been studied as to be related to metastasis, increases the importance on the possible impact that these DEGs may have in metastasis formation and their potential as biomarkers for CRC.

Predictive Model

Furthermore, to characterize samples with potential of metastasis occurrence, a classifier was assembled. A training step was conducted with 70% of the dataset and the errors rates in classifying the samples using different quantities of genes is shown in Figure 12. Due to a high imbalance between the number of metastatic and not metastatic samples the program could not classify any samples as metastatic from a certain number of genes used and would just classify everything as not metastatic which can be seen occurring from the 2.0 threshold mark.

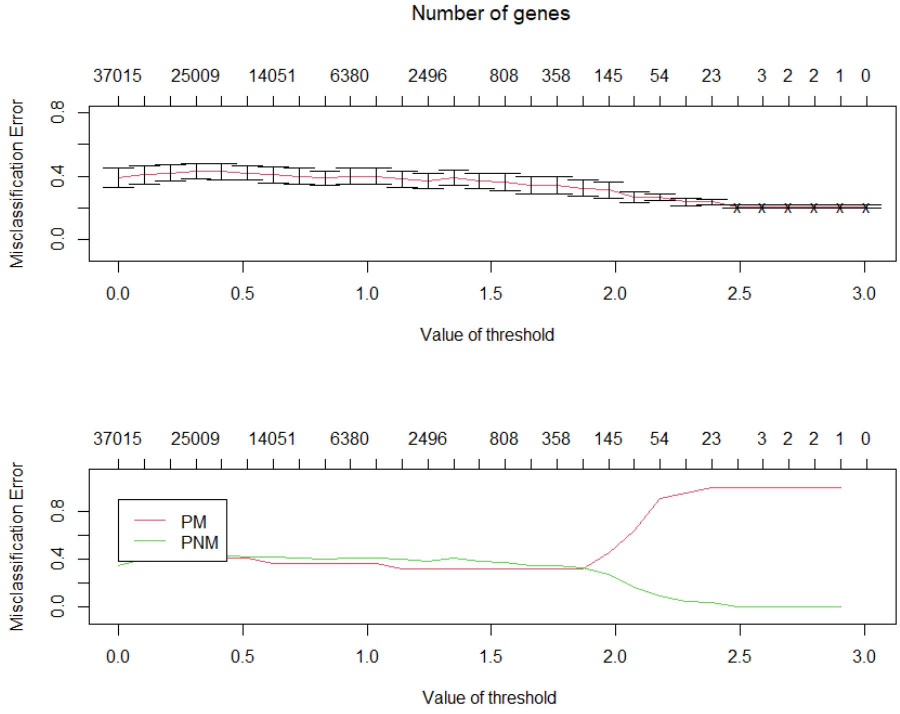


Figure 10 – PAM misclassification error graph with the training data.

The threshold for the classifier was selected according to two parameters, the less misclassification error values and the smaller number of genes used. A threshold of 1.8 showed to give the classifier the best efficiency while using 313 genes and was used for testing in the train data with 70% of the samples and in the final test with the remaining 30% of the dataset. The classifications for the train study demonstrated an error rate of 0.385 for metastatic samples and an error rate of 0.349 for not metastatic resulting in a classifier with an overall error rate of 0.355. In the test for the remaining 30%, the classifier revealed an overall error rate of 0.37.

Lastly, the 166 DEGs from the differential analysis and the 313 genes used in the classifier were compared and another boxplot (Figure 13) was generated for a total of 29 common genes discovered. The genes found were the following: "AC130304.1", "ABCA12", "HSPA1A", "CRYAB", "IDO1", "LINC01871", "GZMA", "CXCL10", "RPL22L1", "LTF", "CXCL11", "HLA-DRB5", "APOL1", "LILRP2", "IFNG", "KLRC2", "LINC00158", "GRM7-AS3", "AP001962.1" "AC012615.2", "MAP1LC3P", "TFAP2A-AS1", "DRGX", "WFDC21P", "AC068658.1", "NOS2", "PPP2R2C", "SLC10A5P1", and "AC100814.2". The expression distribution of these 29 genes was analysed in Figure 13 where the values kept in accordance with the genes being up or down regulated.

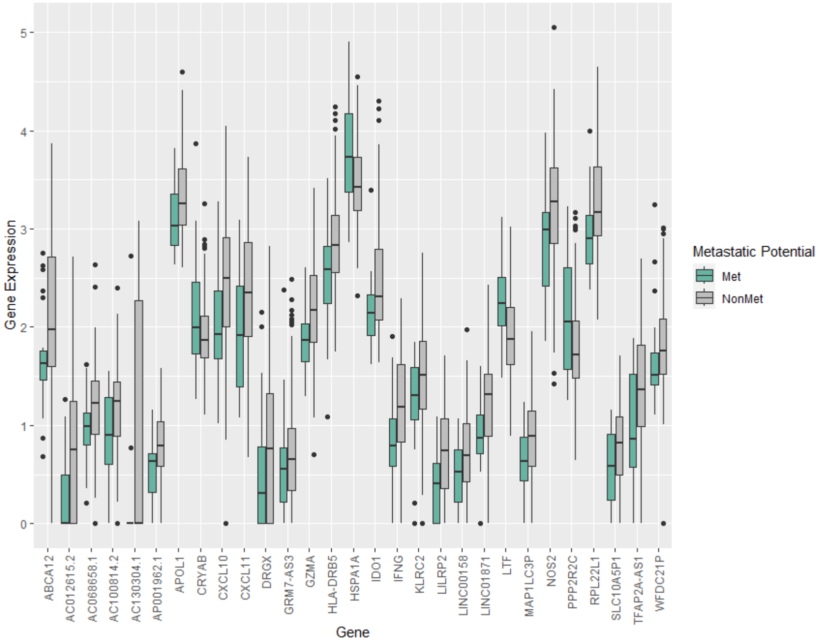
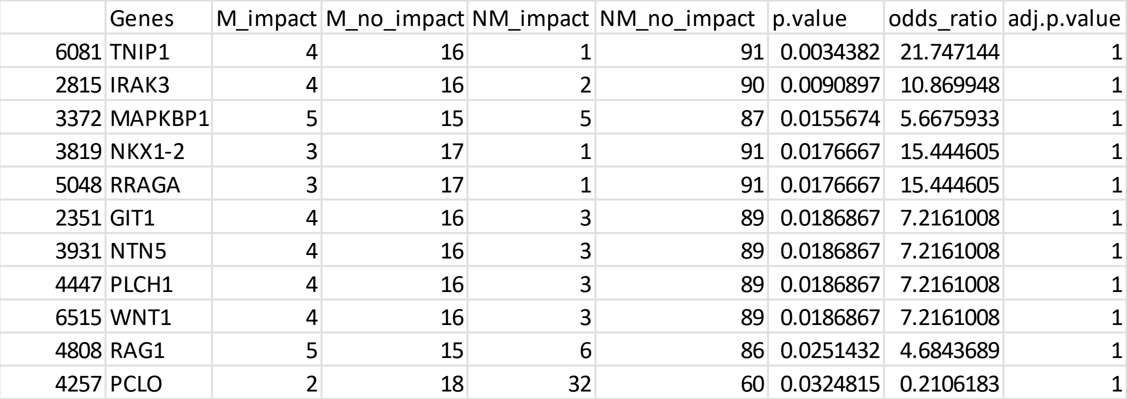


Figure 11 - Boxplot for the 29 common genes (PAMR + DESeq2) expression distribution throughout all patients.

3.3 Genomic Mutations

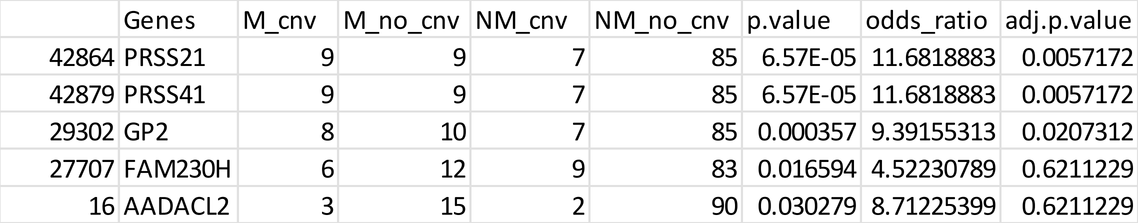
Due to the high difference in the samples and the large number of genes being analysed, when the p-value adjustment was executed practically all values would go up to 1 or close by, becoming statistically irrelevant as observed in Table 6. In Table 7 there were only 3 genes, PRSS21, PRSS41 and GP2, where the p-value after adjustment was still significative (<0.05). Despite the very low number of adjusted p-values with relevance, the raw p-values showed some potential of this type of analysis. The odds ratio (OR) represents the comparison between the chance of the mutated gene having impact in the metastatic samples and the chance of the mutated gene having impact in the samples without metastasis. An OR > 4 means that the chance of the mutated gene having impact in the metastatic samples is more than four times the chance of the mutated gene having impact in the samples without metastasis, reflecting the importance of genes with OR > 4. In the other hand, the PCLO gene has a OR < 0.25 which also reveals its importance, as it means that the chance of the mutated gene having impact in the metastatic samples is less than four times the chance of the mutated gene having impact in the samples without metastasis.

Table 6 – Statistical tests of the relation between metastasis occurrence and SNPs. "M_impact" represents how many samples with metastasis the mutated gene had impact; "M_no_impact" represents how many samples with metastasis the mutated gene had no impact; "NM_impact" represents how many samples without metastasis the mutated gene had impact; "NM_no_impact" represents how many samples without metastasis the mutated gene had no impact.



	Genes	M_impact	M_no_impact	NM_impact	NM_no_impact	p.value	odds_ratio	adj.p.value
6081	TNIP1	4	16	1	91	0.0034382	21.747144	1
2815	IRAK3	4	16	2	90	0.0090897	10.869948	1
3372	MAPKBP1	5	15	5	87	0.0155674	5.6675933	1
3819	NKX1-2	3	17	1	91	0.0176667	15.444605	1
5048	RRAGA	3	17	1	91	0.0176667	15.444605	1
2351	GIT1	4	16	3	89	0.0186867	7.2161008	1
3931	NTN5	4	16	3	89	0.0186867	7.2161008	1
4447	PLCH1	4	16	3	89	0.0186867	7.2161008	1
6515	WNT1	4	16	3	89	0.0186867	7.2161008	1
4808	RAG1	5	15	6	86	0.0251432	4.6843689	1
4257	PCLO	2	18	32	60	0.0324815	0.2106183	1

Table 7 - Statistical tests of the relation between metastasis occurrence and CNVs.



	Genes	M_cnv	M_no_cnv	NM_cnv	NM_no_cnv	p.value	odds_ratio	adj.p.value
42864	PRSS21	9	9	7	85	6.57E-05	11.6818883	0.0057172
42879	PRSS41	9	9	7	85	6.57E-05	11.6818883	0.0057172
29302	GP2	8	10	7	85	0.000357	9.39155313	0.0207312
27707	FAM230H	6	12	9	83	0.016594	4.52230789	0.6211229
16	AADACL2	3	15	2	90	0.030279	8.71225399	0.6211229

4 Discussion

A comprehensive exploration and analysis at genomic and transcriptomic levels was performed to study the relations between metastatic occurrence and genetic variants. Understanding these alterations and finding a correlation with the metastasis phenomenon can produce markers for clinical practice which brings interesting insights into the metastasis topic in human cancer disease for therapeutic methods, diagnostic procedures, and prevention tactics.

Exploring the clinical data of the 149 CRC patients we found some tendencies for metastatic occurrence. Metastasis occurred more in stage III than in stage II which makes sense since the higher the stage the more advanced is the tumour. The stage represents an important category as a potential confounding factor, which can influence the outcome of the analysis. Gene expression profiling between metastatic and not metastatic samples using stage as a confounding factor of the analysis and/or gene expression profiling between only stage III and stage II could bring new insights of which factors influence only the advance of the stages or the metastatic occurrence. In the CMS, two groups, 2 and 4, showed a higher and similar percentage of metastasis incidence compared with the rest. Although having the highest number of metastatic samples (12), CMS2 was less significative for also having the highest number of not metastatic samples (40), while CMS4 had 15 not metastatic and 11 metastatic. As said before, the CMS are divided by gene expression levels in different pathways, so these results can be related by some of these groups having pathways associated to metastasis and cancer evolution. On the contrary, CMS1 had the lowest percentage of metastatic frequency showing a negative relation with metastatic potential. The biggest difference was in the tumour primary location where the colon had 81% of the metastatic cases and the rectum only 19%. These differences in metastatic occurrence could be seen as potential biomarkers for CRC metastasis but they can only be considered as trends due to the low number of samples which result in low viability in a statistic level.

We primarily explored the RNA-seq data by a PCA. When we highlighted the metastatic and not metastatic samples there wasn’t a clear division between the two which showed that at a genetic level there would be some difficulties in getting solid results. Highlighting the CMS groups revealed more clear divisions. Although this was a more beneficial result it was more expected as the CMS groups represent different levels of gene expression within CRC. Despite the unclear division and distribution of metastatic and not metastatic samples we performed further analysis to discover, even so, possible relations in transcriptomic level and metastatic occurrence.

Various tests were conducted to do differential gene expression analysis evaluating which methods and R packages would work better to this type of analysis and data. EdgeR and DESeq2 were the only two considered since limma-voom resulted in no differential expressed genes. EdgeR produced the higher number of DEGs but with a big discrepancy between up and down regulated. DESeq2 resulted in a more solid up/down DEGs percentage but the filtering step was almost inexistent since it used almost all genes. Taking these two tests in consideration we decided that the optimal procedure would be combining the filtering step of EdgeR and the differential analysis of DESeq2. The final deferential expression analysis produced 438 DEGs that were also filtered to increase viability by excluding the very lowly differential expressed genes which resulted in a final value of 166 DEGs, with 94 under expressed and 72 over expressed.

The 20 more significant DEGs were further analysed with heatmaps and boxplots to explore their expression distribution within all samples. These plots supported their selection as DEGs since their expression showed the respective difference for the high and low expressed genes.

CRYAB, FPR1, FPR2, HSPA1A, GADD45B and PROK2 were among the up regulated top DEGs. These genes could be potential markers for CRC metastasis. When cells are harmed by heat shock, radiation, oxidative stress, and other insults, CRYAB predominantly acts as a molecular chaperone to stop the aggregation and destruction of damaged unfolded proteins. This promotes cell survival and prevents apoptosis thus indicating that high expression of these gene could increase tumour cells resistance (Ness et al., 2021). G protein-coupled chemoattractant receptors called formyl peptide receptors (FPRs) are mostly expressed in phagocytic leukocytes. High expression of FPR1 and FPR2 has been associated with tumour progression, migration, and invasion in CRC and other cancers (Shu-qin Li et al., 2017; Xiang et al., 2016). HSPA1A works as a chaperone protein of LASP1 in the cytoplasm and is a member of the heat shock protein family A. It controls the folding of freshly translated proteins and keeps stabilized proteins from aggregating. High expression of HSPA1A and its interaction with LASP1 has been related to CRC proliferation, invasion and metastasis (Q. Chen, Wu, Qin, Yu, & Wang, 2020). GADD45B belongs to the growth arrest DNA damage-inducible gene family and has been shown to be essential for cell growth, DNA repair, and apoptosis. High expression of this gene has been related to tumour cell proliferation (Colorectal et al., 2018). PROK2 is a secreted protein rich in cysteine that is expressed in the testis and small intestine. PROK2 has been related to be involved in vital physiological processes such as inflammation, tissue development, neurogenesis, angiogenesis, and nociception. Studies have associated PROK2 expression with tumour invasion and metastasis in CRC (Kurebayashi, Goi, Shimada, & Tagai, 2015; Yoshida, Goi, Kurebayashi, & Morikawa, 2018).

GZMA, CXCL10, IDO1, SAA1, and ABCA12 were among the down regulated top DEGs. These genes could also be potential markers for CRC metastasis. Granzymes (GZMs) are proteins released by lymphocytes which have an important role in protecting our body from viral infections. Low expression of GMZA has been found to be related to metastasis occurrence due to its function in inflammatory processes suggesting that these proteins have anti-tumour properties and can be associated with good prognosis (Łukaszewicz-zaj & Sara, 2022). CXCL10 is an interferon-inducible protein that has been reported to decrease angiogenesis and boost cell-mediated immunity. Low expression of these gene has been linked to metastasis occurrence since it has been considered a good inhibitor of growth and spread of tumour cells (Jiang, Xu, & Cai, 2010; Kanegane et al., 1998; Sato et al., 2007; Sgadari et al., 1996). IDO1 produces an enzyme responsible in pathophysiological processes such as immunoregulation, antimicrobial, and antitumour defence which suggests that could be a relation with metastasis occurrence when this gene is lowly expressed (Id, Yamashita, Morine, Yoshikawa, & Tokunaga, 2021; Z. Liu et al., 2021). SAA1 encodes a protein member of the serum amyloid A family which is important in the response to tissue injury and inflammation and other processes. Upregulation of SAA1 in CRC has been related to promoting cell migration and invasion. Although in this study these gene is downregulated it should not be discarded as a potential biomarker since there is still uncertainty about the molecular process through which SAA1 promotes cancer cell invasion and migration (Sen Li, Cheng, Cheng, Xu, & Ye, 2021; Sudo et al., 2021). ABCA12 is a transmembrane transporter from the ABC transporter family that can activate the AKT pathway. This pathway contributes to tumour cells proliferation, invasion, and metastasis suggesting that high expression of the ABCA12 gene is connected to cancer metastasis (Rascio et al., 2021; Zheng et al., 2022). In our case, ABCA12 lower expression could have been caused by possible errors in the sequencing reading, differences in the types and conditions that our samples were collected or could be a specific case to CRC that would need further investigation.

GSEA had interesting results revealing enriched pathways related to cancer such as cell cycle checkpoints, DNA repair, and pathways related to metastasis like the activation of oncogenes such as KRAS, ABC transporters, MYC, EMT and Hedgehog. Several benefits distinguish GSEA from single-gene approaches. Through the identification of pathways and processes, it makes it simpler to interpret a large-scale experiment. Researchers can concentrate on gene sets, which tend to be easier to understand and more reproducible, rather than high scoring genes (which can sometimes be poorly annotated and may not be replicable). Deeper analysis and better understanding of enriched gene sets is a promising approach in the future.

A classifier for metastasis occurrence was developed in this study with an overall efficiency of 63%. There were some interesting genes commonly used by the classifier and found in the differential expression analysis such as "ABCA12", "HSPA1A", "CRYAB", "IDO1", "GZMA", and "CXCL10", which were also described earlier as potential biomarkers by other studies revealing that would be interesting to better investigate these genes in the future. The huge unbalance between metastatic and not metastatic samples made it difficult for the program to execute the classification with lower numbers of genes which would be preferential for a more accurate performance. Nonetheless, with a larger dataset and further improvements, the classifier showed potential to be an interesting tool that could bring new insights in the medical field by identifying patients in danger of developing metastasis.

The analysis conducted between metastasis occurrence and genetic mutations (CNV and SNP) revealed to be promising but inconclusive. Although the mutation analysis had some lack of statistical significance because of the high discrepancy in the types of samples, it revealed some grade of interesting results as the raw p-values showed a possibility to exist a correlation between the occurrence of metastasis and the existence of both CNV and SNP mutations. Genes that would show a high level of differential expression and mutations with statistical significance related to metastasis appearance would be strong potential biomarkers for the cancer therapy field showing the importance in combining both genomic and transcriptomic analysis.

This study had some limitations. First, samples were from a single institution, meaning that further validation from multiples establishments are needed. Second, the number of samples was relatively small which compromised the efficacy of the various analysis conducted. Third, the study focused on gene expression levels and not so much on pathways enrichment. Thus, further analysis and exploration of this field could bring good insights. Lastly, the patients were followed for a duration of 3 years which is not enough time to guarantee that there would not occur future metastasis in the patients that it did not happen. In conclusion, even with these obstacles, this study revealed some potential biomarkers for CRC metastasis prediction with transcriptomic data approaches and the possible value increase that can be brought by combining with genomic data. Therefore, future studies involving larger cohorts from different institutions, controlled for longer periods of time, are needed to verify and increase the robustness of these discoveries.

References

Ahlquist, D. A., Harrington, J. J., Burgart, L. J., & Roche, P. C. (2000). Morphometric analysis of the “mucocellular layer” overlying colorectal cancer and normal mucosa: relevance to exfoliation and stool screening. *Human Pathology*, *31*(1), 51–57. https://doi.org/10.1016/s0046-8177(00)80198-7

Allegra, C. J., Paik, S., Colangelo, L. H., Parr, A. L., Kirsch, I., Kim, G., … Wieand, H. S. (2003). Prognostic value of thymidylate synthase, Ki-67, and p53 in patients with Dukes’ B and C colon cancer: a National Cancer Institute-National Surgical Adjuvant Breast and Bowel Project collaborative study. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *21*(2), 241–250. https://doi.org/10.1200/JCO.2003.05.044

Aprile, G., Macerelli, M., Maglio, G. De, Pizzolitto, S., & Fasola, G. (2013). *Relevance of BRAF and extended RAS mutational analyses for metastatic colorectal cancer patients*. *1*, 1–8.

Bacher, J. W., Flanagan, L. A., Smalley, R. L., & Nassif, N. A. (2004). *Development of a fluorescent multiplex assay for detection of MSI-High tumors*. *20*, 237–250.

Baeg, G. H., Matsumine, A., Kuroda, T., Bhattacharjee, R. N., Miyashiro, I., Toyoshima, K., & Akiyama, T. (1995). The tumour suppressor gene product APC blocks cell cycle progression from G0/G1 to S phase. *The EMBO Journal*, *14*(22), 5618–5625. https://doi.org/10.1002/j.1460-2075.1995.tb00249.x

Bayrak, R., Haltas, H., & Yenidunya, S. (2012). The value of CDX2 and cytokeratins 7 and 20 expression in differentiating colorectal adenocarcinomas from extraintestinal gastrointestinal adenocarcinomas: cytokeratin 7-/20+ phenotype is more specific than CDX2 antibody. *Diagnostic Pathology*, *7*, 9. https://doi.org/10.1186/1746-1596-7-9

Bayrak, R., Yenidünya, S., & Haltas, H. (2011). Cytokeratin 7 and cytokeratin 20 expression in colorectal adenocarcinomas. *Pathology, Research and Practice*, *207*(3), 156–160. https://doi.org/10.1016/j.prp.2010.12.005

Benedix, F., Kube, R., Meyer, F., Schmidt, U., Gastinger, I., & Lippert, H. (2010). Comparison of 17,641 patients with right- and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival. *Diseases of the Colon and Rectum*, *53*(1), 57–64. https://doi.org/10.1007/DCR.0b013e3181c703a4

Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. (2001). *Clinical Pharmacology and Therapeutics*, *69*(3), 89–95. https://doi.org/10.1067/mcp.2001.113989

Bozzetti, F., Doci, R., Bignami, P., Morabito, A., & Gennari, L. (1987). Patterns of failure following surgical resection of colorectal cancer liver metastases. Rationale for a multimodal approach. *Annals of Surgery*, *205*(3), 264–270. https://doi.org/10.1097/00000658-198703000-00008

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Brinton, L. T., Brentnall, T. A., Smith, J. A., & Kelly, K. A. (2012). *Metastatic Biomarker Discovery Through Proteomics*. *356*, 345–355.

Brocardo, M., & Henderson, B. R. (2008). APC shuttling to the membrane, nucleus and beyond. *Trends in Cell Biology*, *18*(12), 587–596. https://doi.org/10.1016/j.tcb.2008.09.002

Bufill, J. A. (2016). *Colorectal Cancer : Evidence for Distinct Genetic Categories Based on Proximal or Distal Tumor Location*. 779–788.

Carethers, J. M., & Jung, B. H. (2015). Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. *Gastroenterology*, *149*(5), 1177-1190.e3. https://doi.org/10.1053/j.gastro.2015.06.047

Cercek, A., Braghiroli, M. I., Chou, J. F., Hechtman, J. F., Kemeny, N., Saltz, L., … Yaeger, R. (2017). Clinical Features and Outcomes of Patients with Colorectal Cancers Harboring NRAS Mutations. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *23*(16), 4753–4760. https://doi.org/10.1158/1078-0432.CCR-17-0400

Chang, Y.-Y., Lin, P.-C., Lin, H.-H., Lin, J.-K., Chen, W.-S., Jiang, J.-K., … Chang, S.-C. (2016). Mutation spectra of RAS gene family in colorectal cancer. *American Journal of Surgery*, *212*(3), 537-544.e3. https://doi.org/10.1016/j.amjsurg.2016.02.013

Chen, Q., Wu, K., Qin, X., Yu, Y., & Wang, X. (2020). *LASP1 promotes proliferation , metastasis , invasion in head and neck squamous cell carcinoma and through direct interaction with HSPA1A*. (November 2019), 1626–1639. https://doi.org/10.1111/jcmm.14854

Chen, T.-H., Chang, S.-W., Huang, C.-C., Wang, K.-L., Yeh, K.-T., Liu, C.-N., … Cheng, Y.-W. (2013). The prognostic significance of APC gene mutation and miR-21 expression in advanced-stage colorectal cancer. *Colorectal Disease : The Official Journal of the Association of Coloproctology of Great Britain and Ireland*, *15*(11), 1367–1374. https://doi.org/10.1111/codi.12318

Chen, Y., McCarthy, D., Robinson, M., & Smyth, G. K. (2014). edgeR: differential expression analysis of digital gene expression data User's Guide. *Bioconductor User's Guide*.

Christie, M., Jorissen, R. N., Mouradov, D., Sakthianandeswaren, A., Li, S., Day, F., … Jones, I. T. (2013). *Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT / b -catenin signalling thresholds for tumourigenesis*. (August 2012), 4675–4682. https://doi.org/10.1038/onc.2012.486

Colorectal, I. I., Zhao, Z., Gao, Y., Guan, X., Liu, Z., Jiang, Z., … Wang, X. (2018). *GADD45B as a Prognostic and Predictive Biomarker in Stage II Colorectal Cancer*. (Ci). https://doi.org/10.3390/genes9070361

Corvinus, F. M., Orth, C., Moriggl, R., Tsareva, S. A., Wagner, S., Pfitzner, E. B., … Friedrich, K. (2005). Persistent STAT3 activation in colon cancer is associated with enhanced cell proliferation and tumor growth. *Neoplasia (New York, N.Y.)*, *7*(6), 545–555. https://doi.org/10.1593/neo.04571

De Roock, W., Claes, B., Bernasconi, D., De Schutter, J., Biesmans, B., Fountzilas, G., … Tejpar, S. (2010). Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *The Lancet. Oncology*, *11*(8), 753–762. https://doi.org/10.1016/S1470-2045(10)70130-3

Diamandis, E. P. (2010). Cancer biomarkers: can we turn recent failures into success? *Journal of the National Cancer Institute*, *102*(19), 1462–1467. https://doi.org/10.1093/jnci/djq306

Eiseman, J. L., Guo, J., Ramanathan, R. K., Belani, C. P., Solit, D. B., Scher, H. I., … Egorin, M. J. (2007). Evaluation of plasma insulin-like growth factor binding protein 2 and Her-2 extracellular domain as biomarkers for 17-allylamino-17-demethoxygeldanamycin treatment of adult patients with advanced solid tumors. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *13*(7), 2121–2127. https://doi.org/10.1158/1078-0432.CCR-06-2286

El-Gayar, D., El-Abd, N., Hassan, N., & Ali, R. (2016). Increased Free Circulating DNA Integrity Index as a Serum Biomarker in Patients with Colorectal Carcinoma. *Asian Pacific Journal of Cancer Prevention : APJCP*, *17*(3), 939–944. https://doi.org/10.7314/apjcp.2016.17.3.939

el Atiq, F., Garrouste, F., Remacle-Bonnet, M., Sastre, B., & Pommier, G. (1994). Alterations in serum levels of insulin-like growth factors and insulin-like growth-factor-binding proteins in patients with colorectal cancer. *International Journal of Cancer*, *57*(4), 491–497. https://doi.org/10.1002/ijc.2910570409

Es, J. H. Van, Gijn, M. E. Van, Riccio, O., Born, M. Van Den, Vooijs, M., Begthel, H., … Clevers, H. (2005). *Notch / g -secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells*. *435*(June), 959–963. https://doi.org/10.1038/nature03659

Everitt, B. S., & Howell, D. C. (2005). *Principal Component Analysis*. *3*, 1580–1584.

Fang, C., Zan, J., Yue, B., Liu, C., He, C., & Yan, D. (2017). Long non-coding ribonucleic acid zinc finger antisense 1 promotes the progression of colonic cancer by modulating ZEB1 expression. *Journal of Gastroenterology and Hepatology*, *32*(6), 1204–1211. https://doi.org/10.1111/jgh.13646

Fang, Z., Tang, J., Bai, Y., Lin, H., You, H., Jin, H., … Zhang, Z.-Y. (2015). Plasma levels of microRNA-24, microRNA-320a, and microRNA-423-5p are potential biomarkers for colorectal carcinoma. *Journal of Experimental & Clinical Cancer Research : CR*, *34*(1), 86. https://doi.org/10.1186/s13046-015-0198-6

Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A., & Fares, Y. (2020). Molecular principles of metastasis : a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy*, (October 2019). https://doi.org/10.1038/s41392-020-0134-x

Fearon, E. F., & Vogelstein, B. (1990). *for Colorectal Tumorigenesis*. *61*, 759–767.

Fearon, E. R. (1994). *Molecular Genetics of Colorectal Cancer*. (ii), 101–110.

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, *127*(12), 2893–2917. https://doi.org/https://doi.org/10.1002/ijc.25516

Filip, S., Vymetalkova, V., Petera, J., Vodickova, L., Cervena, K., & Vodicka, P. (2020). *Distant Metastasis in Colorectal Cancer Patients — Do We Have New Predicting Clinicopathological and Molecular Biomarkers ? A Comprehensive Review*. 1–24.

Fong, Y., Fortner, J., Sun, R. L., Brennan, M. F., & Blumgart, L. H. (1999). Clinical score for predicting recurrence after hepatic resection for metastatic colorectal cancer: analysis of 1001 consecutive cases. *Annals of Surgery*, *230*(3), 309–321. https://doi.org/10.1097/00000658-199909000-00004

Fransén, K., Klintenäs, M., Osterström, A., Dimberg, J., Monstein, H.-J., & Söderkvist, P. (2004). Mutation analysis of the BRAF, ARAF and RAF-1 genes in human colorectal adenocarcinomas. *Carcinogenesis*, *25*(4), 527–533. https://doi.org/10.1093/carcin/bgh049

Garinchesa, P., Sakamoto, J., Welt, S., Real, F., Rettig, W., & Old, L. (1996). Organ-specific expression of the colon cancer antigen A33, a cell surface target for antibody-based therapy. *International Journal of Oncology*, *9*(3), 465–471. https://doi.org/10.3892/ijo.9.3.465

Geiersbach, K. B., & Samowitz, W. S. (2011). Microsatellite instability and colorectal cancer. *Archives of Pathology & Laboratory Medicine*, *135*(10), 1269–1277. https://doi.org/10.5858/arpa.2011-0035-RA

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M. G. Della, Dolatshad, H., … Boultwood, J. (2015). *in myelodysplastic syndromes*. https://doi.org/10.1038/ncomms6901

Glebov, O. K., Rodriguez, L. M., Nakahara, K., Jenkins, J., Cliatt, J., Humbyrd, C.-J., … Kirsch, I. R. (2003). Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *12*(8), 755–762.

Gong, W., Tian, M., Qiu, H., & Yang, Z. (2017). Elevated serum level of lncRNA-HIF1A-AS1 as a novel diagnostic predictor for worse prognosis in colorectal carcinoma. *Cancer Biomarkers : Section A of Disease Markers*, *20*(4), 417–424. https://doi.org/10.3233/CBM-170179

Gospodarowicz, M., Benedet, L., Hutter, R. V, Fleming, I., Henson, D. E., & Sobin, L. H. (1998). History and international developments in cancer staging. *Cancer Prevention & Control : CPC = Prevention & Controle En Cancerologie : PCC*, *2*(6), 262–268.

Griffiths, H. R., Møller, L., Bartosz, G., Bast, A., Bertoni-freddari, C., Collins, A., … Verhagen, H. (2002). *Biomarkers*. *23*, 101–208.

Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, inflammation, and cancer. *Cell*, *140*(6), 883–899. https://doi.org/10.1016/j.cell.2010.01.025

Guba, M. (2004). *Vascular endothelial growth factor in colorectal cancer*. 510–517. https://doi.org/10.1007/s00384-003-0576-y

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., … Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350–1356. https://doi.org/10.1038/nm.3967

Gumbiner, B. M. (1996). Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell*, *84*(3), 345–357. https://doi.org/10.1016/s0092-8674(00)81279-9

Hao, T. B., Shi, W., Shen, X. J., Qi, J., Wu, X. H., Wu, Y., … Ju, S. Q. (2014). Circulating cell-free DNA in serum as a biomarker for diagnosis and prognostic prediction of colorectal cancer.

British Journal of Cancer, *111*(8), 1482–1489. https://doi.org/10.1038/bjc.2014.470

Hart, I. R., & Fidler, I. J. (1980). Role of organ selectivity in the determination of metastatic patterns of B16 melanoma. *Cancer Research*, *40*(7), 2281–2287.

Hastie, A. T., Tibshirani, R., Narasimhan, B., & Chu, G. (2019). *Package ' pamr .'*

Hematol, J., Hong, M., Tao, S., Zhang, L., Diao, L. T., Huang, X., … Xie, S. J. (2020). RNA sequencing : new technologies and applications in cancer research. *Journal of Hematology & Oncology*, 1–16. https://doi.org/10.1186/s13045-020-01005-x

Iacopetta, B, & Watanabe, T. (2006, November). Predictive value of microsatellite instability for benefit from adjuvant fluorouracil chemotherapy in colorectal cancer. *Gut*, Vol. 55, pp. 1671–1672.

Iacopetta, Barry. (2002). Are there two sides to colorectal cancer? *International Journal of Cancer*, *101*(5), 403–408. https://doi.org/10.1002/ijc.10635

Id, C. T., Yamashita, S., Morine, Y., Yoshikawa, K., & Tokunaga, T. (2021). *The role of the immunoescape in colorectal cancer liver metastasis*. 1–19. https://doi.org/10.1371/journal.pone.0259940

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Levin, T. R., Lavin, P., Lidgard, G. P., … Berger, B. M. (2014). Multitarget stool DNA testing for colorectal-cancer screening. *The New England Journal of Medicine*, *370*(14), 1287–1297. https://doi.org/10.1056/NEJMoa1311194

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Turnbull, B. A., & Ross, M. E. (2004). Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *The New England Journal of Medicine*, *351*(26), 2704–2714. https://doi.org/10.1056/NEJMoa033403

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F. O., Hesch, R. D., & Knippers, R. (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Research*, *61*(4), 1659–1665.

Jia, M., Gao, X., Zhang, Y., Hoffmeister, M., & Brenner, H. (2016). Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. *Clinical Epigenetics*, *8*, 25. https://doi.org/10.1186/s13148-016-0191-8

Jiang, Z., Xu, Y., & Cai, S. (2010). *CXCL10 expression and prognostic significance in stage II and III colorectal cancer*. 3029–3036. https://doi.org/10.1007/s11033-009-9873-z

Kalady, M. F., Dejulius, K. L., Sanchez, J. A., Jarrar, A., Liu, X., Manilich, E., … Church, J. M. (2012). BRAF mutations in colorectal cancer are associated with distinct clinical characteristics and worse prognosis. *Diseases of the Colon and Rectum*, *55*(2), 128–133. https://doi.org/10.1097/DCR.0b013e31823c08b3

Kanegane, C., Sgadari, C., Kanegane, H., Teruya-Feldstein, J., Yao, L., Gupta, G., … Tosato, G. (1998). Contribution of the CXC chemokines IP-10 and Mig to the antitumor effects of IL-12. *Journal of Leukocyte Biology*, *64*(3), 384–392. https://doi.org/10.1002/jlb.64.3.384

Kanojia, D., Garg, M., Gupta, S., Gupta, A., & Suri, A. (2011). Sperm-associated antigen 9 is a novel biomarker for colorectal cancer and is involved in tumor growth and tumorigenicity. *The American Journal of Pathology*, *178*(3), 1009–1020. https://doi.org/10.1016/j.ajpath.2010.11.047

Karapetis, C. S., Jonker, D., Daneshmand, M., Hanson, J. E., O'Callaghan, C. J., Marginean, C., … Lorimer, I. A. J. (2014). PIK3CA, BRAF, and PTEN status and benefit from cetuximab in the treatment of advanced colorectal cancer--results from NCIC CTG/AGITG CO.17. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *20*(3), 744–753. https://doi.org/10.1158/1078-0432.CCR-13-0606

Keum, N., & Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews. Gastroenterology & Hepatology*, *16*(12), 713–732. https://doi.org/10.1038/s41575-019-0189-8

Klaus, A., & Birchmeier, W. (2008, May). Wnt signalling and its impact on development and cancer. *Nature Reviews. Cancer*, Vol. 8, pp. 387–398. https://doi.org/10.1038/nrc2389

Kuipers, E. J., Rösch, T., & Bretthauer, M. (2013). Colorectal cancer screening--optimizing current strategies and new directions. *Nature Reviews. Clinical Oncology*, *10*(3), 130–142. https://doi.org/10.1038/nrclinonc.2013.12

Kurebayashi, H., Goi, T., Shimada, M., & Tagai, N. (2015). *Prokineticin 2 (PROK2) is an important factor for angiogenesis in colorectal cancer*. *6*(28).

Lao, V. V., & Grady, W. M. (2011). Epigenetics and colorectal cancer. *Nature Reviews. Gastroenterology & Hepatology*, *8*(12), 686–700. https://doi.org/10.1038/nrgastro.2011.173

Leblanc, V. G., & Marra, M. A. (2015). *Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us?* 1925–1958. https://doi.org/10.3390/cancers7030869

Lech, G., Słotwiński, R., Słodkowski, M., & Krasnodębski, I. W. (2016). Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. *World Journal of Gastroenterology*, *22*(5), 1745–1755. https://doi.org/10.3748/wjg.v22.i5.1745

Levin, B., Lieberman, D. A., Mcfarland, B., Andrews, K. S., & Brooks, D. (2008). *DigitalCommons @ University of Nebraska - Lincoln Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps , 2008 : A Joint Guideline From the American Cancer Society , the US Multi-Society Task Force on Colorectal Cancer , and the American College of Radiology*.

Levine, A. J. (1997). *p53 , the Cellular Gatekeeper for Growth and Division*. *88*, 323–331.

Li, Sen, Cheng, Y., Cheng, G., Xu, T., & Ye, Y. (2021). *High SAA1 Expression Predicts Advanced Tumors in Renal Cancer*. *11*(May), 1–11. https://doi.org/10.3389/fonc.2021.649761

Li, Shu-qin, Su, N., Gong, P., Zhang, H., Liu, J., Wang, D., & Sun, Y. (2017). The Expression of Formyl Peptide Receptor 1 is Correlated with Tumor Invasion of Human Colorectal Cancer. *Scientific Reports*, (December 2016), 1–10. https://doi.org/10.1038/s41598-017-06368-9

Liang, H., Cheung, L. W. T., Li, J., Ju, Z., Yu, S., Stemke-hale, K., … Mills, G. B. (2012). *Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer*. 2120–2129. https://doi.org/10.1101/gr.137596.112.10

Liou, J.-M., Shun, C.-T., Liang, J.-T., Chiu, H.-M., Chen, M.-J., Chen, C.-C., … Lin, J.-T. (2010). Plasma insulin-like growth factor-binding protein-2 levels as diagnostic and prognostic biomarker of colorectal cancer. *The Journal of Clinical Endocrinology and Metabolism*, *95*(4), 1717–1725. https://doi.org/10.1210/jc.2009-2668

Liu, L., Meng, T., Yang, X.-H., Sayim, P., Lei, C., Jin, B., … Wang, H.-J. (2018). Prognostic and predictive value of long non-coding RNA GAS5 and mircoRNA-221 in colorectal cancer and their effects on colorectal cancer cell proliferation, migration and invasion. *Cancer Biomarkers : Section A of Disease Markers*, *22*(2), 283–299. https://doi.org/10.3233/CBM-171011

Liu, T., Zhang, X., Gao, S., Jing, F., Yang, Y., Du, L., … Wang, C. (2016). Exosomal long noncoding RNA CRNDE-h as a novel serum-based biomarker for diagnosis and prognosis of colorectal cancer. *Oncotarget*, *7*(51), 85551–85563. https://doi.org/10.18632/oncotarget.13465

Liu, Z., Zhang, Y., Dang, Q., Wu, K., Jiao, D., & Li, Z. (2021). *Genomic Alteration Characterization in Colorectal Cancer Identifies a Prognostic and Metastasis Biomarker : FAM83A | IDO1*. *11*(April), 1–19. https://doi.org/10.3389/fonc.2021.632430

Locker, G. Y., Hamilton, S., Harris, J., Jessup, J. M., Kemeny, N., Macdonald, J. S., … Bast, R. C. J. (2006). ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *24*(33), 5313–5327. https://doi.org/10.1200/JCO.2006.08.2644

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Łukaszewicz-zaj, M., & Sara, P. (2022). *Granzymes — Their Role in Colorectal Cancer*.

Lynch, H. T., & de la Chapelle, A. (2003). Hereditary colorectal cancer. *The New England Journal of Medicine*, *348*(10), 919–932. https://doi.org/10.1056/NEJMra012242

Malki, A., Elruz, R. A., Gupta, I., Allouch, A., Vranic, S., & Moustafa, A. Al. (2021). *Molecular Mechanisms of Colon Cancer Progression and Metastasis : Recent Insights and Advancements*.

Mandel, J. S., Bond, J. H., Church, T. R., Snover, D. C., Bradley, G. M., Schuman, L. M., & Ederer, F. (1993). Reducing mortality from colorectal cancer by screening for fecal occult blood.

Minnesota Colon Cancer Control Study. *The New England Journal of Medicine*, *328*(19), 1365–1371. https://doi.org/10.1056/NEJM199305133281901

Manfredi, S., Bouvier, A. M., Lepage, C., Hatem, C., Dancourt, V., & Faivre, J. (2006). Incidence and patterns of recurrence after resection for cure of colonic cancer in a well defined population. *The British Journal of Surgery*, *93*(9), 1115–1122. https://doi.org/10.1002/bjs.5349

Marchand, L. Le, Wilkens, L. R., Hankin, J. H., Kolonel, L. N., & Lyu, L. (1997). *A case-control study of diet and colorectal cancer in a multiethnic population in Hawaii (United States): lipids and foods of animal origin*. *8*, 637–648.

Markowitz, S. D., & Bertagnolli, M. M. (2009). *Molecular Basis of Colorectal Cancer*.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., … Myeroff, L. (2016). *Inactivation of the Type II TGF-β Receptor in Colon Cancer Cells with Microsatellite Instability Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2888776 Inactivation of the Type II TGF-p Receptor in Colon Cancer Cells with Microsatellite Instabil*. 18–21.

Mármol, I., Sánchez-de-diego, C., Dieste, A. P., Cerrada, E., Jesús, M., & Yoldi, R. (2017). *Colorectal Carcinoma : A General Overview and Future Perspectives in Colorectal Cancer*. https://doi.org/10.3390/ijms18010197

Massagué, J., & Obenauf, A. C. (2016). Metastatic colonization by circulating tumour cells. *Nature*, *529*(7586), 298–306. https://doi.org/10.1038/nature17038

Matos, M. R. De, Posa, I., Carvalho, F. S., Morais, V. A., Grosso, A. R., & Almeida, F. De. (2019). *A Systematic Pan-Cancer Analysis of Genetic Heterogeneity Reveals Associations with Epigenetic Modifiers*. 7–9. https://doi.org/10.3390/cancers11030391

Miller, J. R., & Randall, T. M. (1996). *Signal transducti through [-catenin and speclflcatlon ° o cell fate during embryogenesis*. 2527–2539.

Misiakos, E. P., Karidis, N. P., & Kouraklis, G. (2011). Current treatment for colorectal liver metastases. *World Journal of Gastroenterology*, *17*(36), 4067–4075. https://doi.org/10.3748/wjg.v17.i36.4067

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., … Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10513–10518. https://doi.org/10.1073/pnas.0804549105

Moh, M., Krings, G., Ates, D., Aysal, A., Kim, G. E., & Rabban, J. T. (2016). SATB2 Expression Distinguishes Ovarian Metastases of Colorectal and Appendiceal Origin From Primary Ovarian Tumors of Mucinous or Endometrioid Type. *The American Journal of Surgical Pathology*, *40*(3), 419–432. https://doi.org/10.1097/PAS.0000000000000553

Moskaluk, C. A., Zhang, H., Powell, S. M., Cerilli, L. A., Hampton, G. M., & Frierson, H. F. J. (2003). Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *16*(9), 913–919. https://doi.org/10.1097/01.MP.0000086073.92773.55

Narayan, S., & Roy, D. (2003). Role of APC and DNA mismatch repair genes in the development of colorectal cancers. *Molecular Cancer*, *2*, 41. https://doi.org/10.1186/1476-4598-2-41

Nawa, T., Kato, J., Kawamoto, H., Okada, H., Yamamoto, H., Kohno, H., … Shiratori, Y. (2008). Differences between right- and left-sided colon cancer in patient characteristics, cancer morphology and histology. *Journal of Gastroenterology and Hepatology*, *23*(3), 418–423. https://doi.org/10.1111/j.1440-1746.2007.04923.x

Ness, C., Katta, K., Garred, Ø., Kumar, T., Kristoffer, O., Petrovski, G., … Noer, A. (2021). *Integrated differential DNA methylation and gene expression of formalin-fixed paraffin-embedded uveal melanoma specimens identifies genes associated with early metastasis and poor prognosis*. *203*(December 2020). https://doi.org/10.1016/j.exer.2020.108426

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., & Thiery, J. P. (2016). EMT: 2016. *Cell*, *166*(1), 21–45. https://doi.org/10.1016/j.cell.2016.06.028

Nikolic, A., Kojic, S., Knezevic, S., Krivokapic, Z., Ristanovic, M., & Radojkovic, D. (2011). Structural and functional analysis of SMAD4 gene promoter in malignant pancreatic and colorectal tissues: detection of two novel polymorphic nucleotide repeats. *Cancer Epidemiology*, *35*(3), 265–271. https://doi.org/10.1016/j.canep.2010.10.002

Nordlinger, B., Guiguet, M., Vaillant, J. C., Balladur, P., Boudjema, K., Bachellier, P., & Jaeck, D. (1996). Surgical resection of colorectal carcinoma metastases to the liver. A prognostic scoring system to improve case selection, based on 1568 patients. Association Française de Chirurgie. *Cancer*, *77*(7), 1254–1262.

Ogino, S., Nosho, K., Kirkner, G. J., Kawasaki, T., Meyerhardt, J. A., Loda, M., … Fuchs, C. S. (2009). CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut*, *58*(1), 90–96. https://doi.org/10.1136/gut.2008.155473

Oh, H.-H., & Joo, Y.-E. (2020). Novel biomarkers for the diagnosis and prognosis of colorectal cancer. *Intestinal Research*, *18*(2), 168–183. https://doi.org/10.5217/ir.2019.00080

Okita, A., Takahashi, S., Ouchi, K., Inoue, M., Watanabe, M., Endo, M., … Ishioka, C. (2018). Consensus molecular subtypes classification of colorectal cancer as a predictive factor for chemotherapeutic efficacy against metastatic colorectal cancer. *Oncotarget*, *9*(27), 18698–18711. https://doi.org/10.18632/oncotarget.24617

Panarelli, N. C., Yantiss, R. K., Yeh, M. M., Liu, Y., & Chen, Y.-T. (2012). Tissue-specific cadherin CDH17 is a useful marker of gastrointestinal adenocarcinomas with higher sensitivity than CDX2. *American Journal of Clinical Pathology*, *138*(2), 211–222. https://doi.org/10.1309/AJCPKSHXI3XEHW1J

Park, Y. J., Park, K. J., Park, J. G., Lee, K. U., Choe, K. J., & Kim, J. P. (1999). Prognostic factors in 2230 Korean colorectal cancer patients: analysis of consecutively operated cases. *World Journal of Surgery*, *23*(7), 721–726. https://doi.org/10.1007/pl00012376

Peixoto, C., Lopes, M. B., Martins, M., Casimiro, S., Sobral, D., Grosso, A. R., … Costa, L. (2023). Identification of biomarkers predictive of metastasis development in early - stage colorectal cancer using network - based regularization. *BMC Bioinformatics*, 1–23. https://doi.org/10.1186/s12859-022-05104-z

Peng, W., Wang, Z., & Fan, H. (2017). LncRNA NEAT1 Impacts Cell Proliferation and Apoptosis of Colorectal Cancer via Regulation of Akt Signaling. *Pathology Oncology Research : POR*, *23*(3), 651–656. https://doi.org/10.1007/s12253-016-0172-4

Perez Montiel, D., Arispe Angulo, K., Cantú-de León, D., Bornstein Quevedo, L., Chanona Vilchis, J., & Herrera Montalvo, L. (2015). The value of SATB2 in the differential diagnosis of intestinal-type mucinous tumors of the ovary: primary vs metastatic. *Annals of Diagnostic Pathology*, *19*(4), 249–252. https://doi.org/10.1016/j.anndiagpath.2015.05.004

Perrone, F., Lampis, A., Orsenigo, M., Di Bartolomeo, M., Gevorgyan, A., Losa, M., … Pilotti, S. (2009). PI3KCA/PTEN deregulation contributes to impaired responses to cetuximab in metastatic colorectal cancer patients. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *20*(1), 84–90. https://doi.org/10.1093/annonc/mdn541

Petersen, S., Thames, H. D., Nieder, C., Petersen, C., & Baumann, M. (2001). The results of colorectal cancer treatment by p53 status: treatment-specific overview. *Diseases of the Colon and Rectum*, *44*(3), 322–324. https://doi.org/10.1007/BF02234727

Pino, M. S., & Chung, D. C. (2010). The chromosomal instability pathway in colon cancer. *Gastroenterology*, *138*(6), 2059–2072. https://doi.org/10.1053/j.gastro.2009.12.065

Popat, S., Chen, Z., Zhao, D., Pan, H., Hearle, N., Chandler, I., … Houlston, R. (2006). A prospective, blinded analysis of thymidylate synthase and p53 expression as prognostic markers in the adjuvant treatment of colorectal cancer. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *17*(12), 1810–1817. https://doi.org/10.1093/annonc/mdl301

Puccini, A., Seeber, A., & Berger, M. D. (2022). *Biomarkers in Metastatic Colorectal Cancer : Status Quo and Future Perspective*. 1–25.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). *The promise of whole-exome sequencing in medical genetics*. (November 2013), 5–15. https://doi.org/10.1038/jhg.2013.114

Rascio, F., Spadaccino, F., Rocchetti, M. T., Castellano, G., Stallone, G., Netti, G. S., & Ranieri, E. (2021). *The Pathogenic Role of PI3K / AKT Pathway in Cancer Onset and Drug Resistance : An Updated Review*.

Renehan, A. G., Jones, J., Potten, C. S., Shalet, S. M., & O'Dwyer, S. T. (2000). Elevated serum insulin-like growth factor (IGF)-II and IGF binding protein-2 in patients with colorectal cancer. *British Journal of Cancer, 83*(10), 1344–1350. https://doi.org/10.1054/bjoc.2000.1462

Rennoll, S., & Yochum, G. (2015). Regulation of MYC gene expression by aberrant Wnt/β-catenin signaling in colorectal cancer. *World Journal of Biological Chemistry*, *6*(4), 290–300. https://doi.org/10.4331/wjbc.v6.i4.290

Riggins, G. J., Kinzler, K. W., Vogelstein, B., & Thiagalingam, S. (1997). Frequency of Smad gene mutations in human cancers. *Cancer Research*, *57*(13), 2578–2580.

Roig, A. I., Wright, W. E., & Shay, J. W. (2009). *Is Telomerase a Novel Target for Metastatic Colon Cancer ? 1*.

Rowland, A., Dias, M. M., Wiese, M. D., Kichenadasse, G., McKinnon, R. A., Karapetis, C. S., & Sorich, M. J. (2015). Meta-analysis of BRAF mutation as a predictive biomarker of benefit from anti-EGFR monoclonal antibody therapy for RAS wild-type metastatic colorectal cancer. *British Journal of Cancer*, *112*(12), 1888–1894. https://doi.org/10.1038/bjc.2015.173

Russo, A., Bazan, V., Iacopetta, B., Kerr, D., Soussi, T., & Gebbia, N. (2005). The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *23*(30), 7518–7528. https://doi.org/10.1200/JCO.2005.00.471

Salovaara, R., Roth, S., Loukola, A., Launonen, V., Sistonen, P., Avizienyte, E., … Aaltonen, L. A. (2002). Frequent loss of SMAD4/DPC4 protein in colorectal cancers. *Gut*, *51*(1), 56–59. https://doi.org/10.1136/gut.51.1.56

Sánchez-Gundín, J., Fernández-Carballido, A. M., Martínez-Valdivieso, L., Barreda-Hernández, D., & Torres-Suárez, A. I. (2018). New Trends in the Therapeutic Approach to Metastatic Colorectal Cancer. *International Journal of Medical Sciences*, *15*(7), 659–665. https://doi.org/10.7150/ijms.24453

Sartore-Bianchi, A., Martini, M., Molinari, F., Veronese, S., Nichelatti, M., Artale, S., … Bardelli, A. (2009). PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer Research*, *69*(5), 1851–1857. https://doi.org/10.1158/0008-5472.CAN-08-2466

Sato, E., Fujimoto, J., Toyoki, H., Sakaguchi, H., Alam, S. M., Jahan, I., & Tamaya, T. (2007). Expression of IP-10 related to angiogenesis in uterine cervical cancers. *British Journal of Cancer*, *96*(11), 1735–1739. https://doi.org/10.1038/sj.bjc.6603790

Schirripa, M., Cremolini, C., Loupakis, F., Morvillo, M., Bergamo, F., Zoratto, F., … Falcone, A. (2015). Role of NRAS mutations as prognostic and predictive markers in metastatic colorectal cancer. *International Journal of Cancer*, *136*(1), 83–90. https://doi.org/10.1002/ijc.28955

Segditsas, S., & Tomlinson, I. (2006). *Colorectal cancer and genetic alterations in the Wnt pathway*. 7531–7537. https://doi.org/10.1038/sj.onc.1210059

Sforza, V., Martinelli, E., Ciardiello, F., Gambardella, V., Napolitano, S., Martini, G., … Troiani, T. (2016). Mechanisms of resistance to anti-epidermal growth factor receptor inhibitors in metastatic colorectal cancer. *World Journal of Gastroenterology*, *22*(28), 6345–6361. https://doi.org/10.3748/wjg.v22.i28.6345

Sgadari, C., Angiolillo, A. L., Cherney, B. W., Pike, S. E., Farber, J. M., Koniaris, L. G., … Tosato, G. (1996). Interferon-inducible protein-10 identified as a mediator of tumor necrosis in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13791–13796. https://doi.org/10.1073/pnas.93.24.13791

Shay, J. W., Zou, Y., Hiyama, E., & Wright, W. E. (2001). Telomerase and cancer. *Human Molecular Genetics*, *10*(7), 677–685. https://doi.org/10.1093/hmg/10.7.677

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., …

Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164. https://doi.org/10.3322/caac.21601

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, *69*(1), 7–34. https://doi.org/10.3322/caac.21551

Sigal, A., & Rotter, V. (2000). *Oncogenic Mutations of the p53 Tumor Suppressor : The Demons of the Guardian of the Genome*. *53*(27), 6788–6793.

Smith, G., Carey, F. A., Beattie, J., Wilkie, M. J. V, Lightfoot, T. J., Coxhead, J., … Wolf, C. R. (2002). *genetic pathways to colorectal cancer*. (13). https://doi.org/10.1073/pnas.122612899

Spindler, K. L. G., Pallisgaard, N., Andersen, R. F., Brandslund, I., & Jakobsen, A. (2015). Circulating free DNA as biomarker and source for mutation detection in metastatic colorectal cancer. *PloS One*, *10*(4), e0108247. https://doi.org/10.1371/journal.pone.0108247

Stoffel, E. M., & Kastrinos, F. (2014). Familial colorectal cancer, beyond Lynch syndrome. *Clinical Gastroenterology and Hepatology : The Official Clinical Practice Journal of the American Gastroenterological Association*, *12*(7), 1059–1068. https://doi.org/10.1016/j.cgh.2013.08.015

Su, M.-C., Yuan, R.-H., Lin, C.-Y., & Jeng, Y.-M. (2008). Cadherin-17 is a useful diagnostic marker for adenocarcinomas of the digestive system. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *21*(11), 1379–1386. https://doi.org/10.1038/modpathol.2008.107

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., & Ebert, B. L. (2005). *Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide*.

Sudo, G., Aoki, H., Yamamoto, E., Takasawa, A., Niinuma, T., Yoshido, A., … Nakase, H. (2021). *Activated macrophages promote invasion by early colorectal cancer via an interleukin 1 β - - serum amyloid A1 axis*. (April), 4151–4165. https://doi.org/10.1111/cas.15080

Tang, H., Li, B., Zhang, A., Lu, W., Xiang, C., & Dong, J. (2016). Prognostic Significance of Neutrophil-to-Lymphocyte Ratio in Colorectal Liver Metastasis: A Systematic Review and Meta-Analysis. *PloS One*, *11*(7), e0159447. https://doi.org/10.1371/journal.pone.0159447

Thibodeau, S. N., Bren, G., & Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science (New York, N.Y.)*, *260*(5109), 816–819. https://doi.org/10.1126/science.8484122

Toraih, E. A., Hussein, M. H., Zerfaoui, M., Attia, A. S., Marzouk Ellythy, A., Mostafa, A., … Kandil, E. (2021). Site-Specific Metastasis and Survival in Papillary Thyroid Cancer: The Importance of Brain and Multi-Organ Disease. *Cancers*, *13*(7). https://doi.org/10.3390/cancers13071625

Tsai, P.-L., Su, W.-J., Leung, W.-H., Lai, C.-T., & Liu, C.-K. (2016). Neutrophil-lymphocyte ratio and CEA level as prognostic and predictive factors in colorectal cancer: A systematic review and meta-analysis. *Journal of Cancer Research and Therapeutics*, *12*(2), 582–589. https://doi.org/10.4103/0973-1482.144356

Tsubakihara, Y., & Moustakas, A. (2018). *Epithelial-Mesenchymal Transition and Metastasis under the Control of Transforming Growth Factor β*. (Figure 1), 1–30. https://doi.org/10.3390/ijms19113672

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., … Srivastava, S. (2004, February). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, Vol. 96, pp. 261–268. https://doi.org/10.1093/jnci/djh034

Voorneveld, P. W., Jacobs, R. J., Kodach, L. L., & Hardwick, J. C. H. (2015). A Meta-Analysis of SMAD4 Immunohistochemistry as a Prognostic Marker in Colorectal Cancer. *Translational Oncology*, *8*(1), 18–24. https://doi.org/10.1016/j.tranon.2014.11.003

Ward, R., Meagher, A., Tomlinson, I., O'Connor, T., Norrie, M., Wu, R., & Hawkins, N. (2001). Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. *Gut*, *48*(6), 821–829. https://doi.org/10.1136/gut.48.6.821

Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., … Laird, P. W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature Genetics*, *38*(7), 787–793. https://doi.org/10.1038/ng1834

Weiss, J. M., Pfau, P. R., O'Connor, E. S., King, J., LoConte, N., Kennedy, G., & Smith, M. A. (2011). Mortality by stage for right- versus left-sided colon cancer: analysis of surveillance, epidemiology, and end results--Medicare data. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *29*(33), 4401–4409. https://doi.org/10.1200/JCO.2011.36.4414

Werling, R. W., Yaziji, H., Bacchi, C. E., & Gown, A. M. (2003). CDX2, a highly sensitive and specific marker of adenocarcinomas of intestinal origin: an immunohistochemical survey of 476 primary and metastatic carcinomas. *The American Journal of Surgical Pathology*, *27*(3), 303–310. https://doi.org/10.1097/00000478-200303000-00003

Westra, J. L., Schaapveld, M., Hollema, H., de Boer, J. P., Kraak, M. M. J., de Jong, D., … Plukker, J. T. M. (2005). Determination of TP53 mutation is more relevant than microsatellite instability status for the prediction of disease-free survival in adjuvant-treated stage III colon cancer patients. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *23*(24), 5635–5643. https://doi.org/10.1200/JCO.2005.04.096

Wirtzfeld, D. A., Petrelli, N. J., & Rodriguez-bigas, M. A. (2001). *Hamartomatous Polyposis Syndromes : Molecular Genetics , Neoplastic Risk , and Surveillance Recommendations*. *8*(4), 319–327.

Wong, N. A. C. S., Adamczyk, L. A., Evans, S., Cullen, J., Oniscu, A., & Oien, K. A. (2017). A33 shows similar sensitivity to but is more specific than CDX2 as an immunomarker of colorectal carcinoma. *Histopathology*, *71*(1), 34–41. https://doi.org/10.1111/his.13194

Wong, R., & Cunningham, D. (2008, December). Using predictive biomarkers to select patients with advanced colorectal cancer for treatment with epidermal growth factor receptor antibodies. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 26, pp. 5668–5670. https://doi.org/10.1200/JCO.2008.19.5024

Xiang, Y., Yao, X., Chen, K., Wang, X., Zhou, J., & Gong, W. (2016). *The G-protein coupled chemoattractant receptor FPR2 promotes malignant phenotype of human colon cancer cells*. *6*(11), 2599–2610.

Xie, J., & Itzkowitz, S. H. (2008). Cancer in inflammatory bowel disease. *World Journal of Gastroenterology*, *14*(3), 378–389. https://doi.org/10.3748/wjg.14.378

Xu, H., Wang, C., Song, H., Xu, Y., & Ji, G. (2019). *RNA-Seq profiling of circular RNAs in human colorectal Cancer liver metastasis and the potential biomarkers*. 4–9.

Xu, Y., & Pasche, B. (2007). TGF-beta signaling alterations and susceptibility to colorectal cancer. *Human Molecular Genetics*, *16 Spec No 1*(SPEC), R14-20. https://doi.org/10.1093/hmg/ddl486

Yao, Z., Han, L., Chen, Y., He, F., Sun, B., Zhang, Y., … Yang, Z. (2018). Hedgehog signalling in the tumourigenesis and metastasis of osteosarcoma , and its potential value in the clinical therapy of osteosarcoma. *Cell Death and Disease*, 1–12. https://doi.org/10.1038/s41419-018-0647-1

Yoshida, Y., Goi, T., Kurebayashi, H., & Morikawa, M. (2018). *Prokineticin 2 expression as a novel prognostic biomarker for human colorectal cancer*. *9*(53), 30079–30091.

You, S., Zhou, J., Chen, S., Zhou, P., Lv, J., Han, X., & Sun, Y. (2010). PTCH1, a receptor of Hedgehog signaling pathway, is correlated with metastatic potential of colorectal cancer. *Upsala Journal of Medical Sciences*, *115*(3), 169–175. https://doi.org/10.3109/03009731003668316

Zahorec, R. (2001). Ratio of neutrophil to lymphocyte counts--rapid and simple parameter of systemic inflammation and stress in critically ill. *Bratislavske Lekarske Listy*, *102*(1), 5–14.

Zheng, S., Liu, D., Wang, F., Jin, Y., Zhao, S., & Sun, S. (2022). *ABCA12 Promotes Proliferation and Migration and Inhibits Apoptosis of Pancreatic Cancer Cells Through the AKT Signaling Pathway*. *13*(June), 1–12. https://doi.org/10.3389/fgene.2022.906326

Zhong, L., Liu, J., Hu, Y., Wang, W., Xu, F., Xu, W., … Biskup, E. (2017). STK31 as novel biomarker of metastatic potential and tumorigenicity of colorectal cancer. *Oncotarget*, *8*(15), 24354–24361. https://doi.org/10.18632/oncotarget.15396

Zitt, M., Müller, H. M., Rochel, M., Schwendinger, V., Zitt, M., Goebel, G., … Ofner, D. (2008). Circulating cell-free DNA in plasma of locally advanced rectal cancer patients undergoing preoperative chemoradiation: a potential diagnostic tool for therapy monitoring. *Disease*

Markers, *25*(3), 159–165. https://doi.org/10.1155/2008/598071

Zou, H., Harrington, J. J., Klatt, K. K., & Ahlquist, D. A. (2006). A sensitive method to quantify human long DNA in stool: relevance to colorectal cancer screening. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *15*(6), 1115–1119. https://doi.org/10.1158/1055-9965.EPI-05-0992