

Prognostic Biomarker Identification and Classification of Colorectal Cancer Patients: A Dual Gene-Based and Sample-Based Approach

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MOLECULAR BIOLOGY AND GENETICS

By
Ronak Naeemae
August 2023

Prognostic Biomarker Identification and Classification of Colorectal Cancer Patients: A Dual Gene-Based and Sample-Based Approach

By Ronak Naeemae

August 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ali Osmay Güre (Advisor)

Özlen Konu Karakayalı

Seçil Demirkol Canlı

Approved for the Graduate School of Engineering and Science:

ORHAN ARIKAN

Director of the Graduate School

ABSTRACT

Prognostic Biomarker Identification and Classification of Colorectal Cancer Patients: A Dual Gene-Based and Sample-Based Approach

Ronak Naeemae

M.Sc. in Molecular Biology and Genetics

Advisor: Ali Osmay Güre

August 2023

Colorectal cancer (CRC) is one of the most heterogeneous cancer types, with high mortality rates making it the one of the deadliest cancer among men and women. The heterogeneity of CRC comes from the numerous clinicopathological characteristics of these tumors, including; KRAS/BRAF mutation, Microsatellite Instability (MSI), and stage. Another essential factor recently emphasized is the tumor location (proximal or distal). Consequently, many studies have focused on finding prognostic biomarkers for CRC patients to increase the efficiency of their treatment plans. However, despite the attempts, these biomarkers fail in clinical transition as they lack robustness and consistent results in their validation studies. Moreover, understanding the mechanism behind CRC progression can significantly help the personalization of treatments. Recently, the cancer neuroscience field has been focusing on elucidating neuropeptides' role in cancer and CRC as they have been proven to be involved in cancer progression. Accordingly, the thesis was divided into two approaches. The first approach was to further examine the role of neuropeptides by finding a subset of neuropeptides for the classification of the CRC samples and following functional analysis to understand the mechanism of their involvement. Moreover, the second approach attempted the

determination of robust prognostic biomarkers in a specific sample group (Proximal, Stages 2 and 3) while controlling for the inconsistencies. In the first approach, a subset of 9 neuropeptide genes was found through Principle Component Analysis (PCA) with the ability to stratify the CRC samples into high and low expression groups. Functional analyses of these groups identified an association between the up-regulation of these neuropeptides and Hedgehog's (HHG) signaling pathway, and these activities are hypothesized to be primarily specific to the stroma of the tumor. Up-regulation of these neuropeptides was also linked with other pathways involved in cancer progression, such as; EMT, angiogenesis, and TGF β activities. The second approach utilized a new methodology pipeline that aimed to ensure the selection of genes with no discrepancies among their probesets and across different technologies. Following the pipeline, 3 genes were identified, associated with favorable and non-favorable prognoses for Proximal, Stage 2, and 3 samples. However, although a very stringent methodology was used and various clinicopathological parameters such as the stage and location were considered, the prognostic associations observed were not as consistent, indicating the importance of the sample's molecular characteristics. This study also pointed out potential implications of neuropeptides in CRC progression and development. More elaborative studies are required for the clarification of the interactions of neuropeptides with the HHG signaling pathway. Furthermore, the identified prognostic biomarkers need to be validated through comprehensive validation studies in their associating subgroups of samples, as they are robust biomarkers with the potential to be used in clinics.

Keywords: Colorectal Cancer, Prognosis, Survival, Biomarkers, Neuropeptides, Hedgehog Signaling Pathway

ÖZET

Kolorektal Kanser Hastalarında Prognostik Biyomarkör Belirleme ve Sınıflandırma: İkili Gen-Temelli ve Örnek-Temelli Yaklaşım

Ronak Naeemae

M.Sc. in Molecular Biology and Genetics

Advisor: Ali Osmay Güre

August 2023

Kolorektal kanser (KRK), diğer kanserler arasında ikinci en ölümcül kanser olup moleküler çeşitlilik gösteren bir kanserdir. KRK'in moleküler heterojenitesi, KRAS/BRAF mutasyonları, Mikrosatellit Instabilitesi (MSI) ve tümör evresi gibi çeşitli klinikopatolojik özelliklerden kaynaklanmaktadır. Son zamanlarda, tümörün yerleşiminin (proksimal veya distal) KRK ilerlemesini etkileyen diğer önemli bir faktör olduğuna dikkat çekilmiştir. Bu nedenlerle, tedavi stratejilerinin etkinliğini artırmak amacıyla prognostik biyobelirteçlerin belirlenmesi için çaba sarf edilmiştir. Ancak, bu biyobelirteçlerin çoğunun klinikte uygulanmaları, güvenilirlik ve tutarlığı saptamakla ilgili araştırma sonuçlarının eksik oluşu nedeniyle başarılı olmamıştır. Bu çalışma, KRK biyolojisini daha detaylı anlamak ve elde edilecek sonuçların klinik ile olabilecek ilişkilerini tespit etmek üzere iki yaklaşım içermektedir. İlk yaklaşım, nöropeptidlerin KRK prognozundaki rollerini aydınlatmayı amaçlamıştır; KRK örneklerini sınıflandırmak için bir nöropeptid gen alt kümesi belirleyip ve mekanistik etkilerini anlamak için fonksiyonel analizler gerçekleştirilmiştir. Ana komponent Analizi (PCA) ile KRK örneklerini yüksek ve düşük ekspresyon gruplarına ayırabilen dokuz nöropeptid gen ile tümör alt kümeleri belirledik. Fonksiyonel analizler (Gen-set Zenginleştirme analizi (GSEA)), bu nöropeptidlerin tümör stromasına özgü olan Hedgehog (HHG)

sinyal yolları ile ilişkisini ortaya koydu. Bu nöropeptidlerin düzeylerindeki artış, kanser ilerlemesi için kritik olan diğer yollarla, özellikle epitel-mezensimal geçiş (EMT), anjiyogenez ve TGF β aktiviteleri ile de ilişkili bulundu. İkinci yaklaşımın hedefi, biyobelirteç doğrulamasındaki tutarsızlıkları ele alırken, belirli bir örnek grup; Proksimal, Evre 2 ve 3 KRK'de robust prognostik biyobelirteçler belirlemektir. Farklı teknolojiler ve probsetleri arasında doğru gen seçimini sağlamak için yeni bir metodoloji geliştirildi. Tüm örnekler seçildiğinde (Proksimal, Evre 2 ve 3), prognoz ile doğrudan veya ters ilişkili üç gen, prognostik belirteçler olarak belirlendi. Ancak, değişik doğrulama çalışmaları ve klinikopatolojik veriler dikkate alındığı durumlarda dahi, örneklerin klinik alt gruplara ayrıldıklarında gözlenen prognostik ilişkiler tamamen tutarlı değildi. Bu sonuçlar, biyobelirteç çalışmaları için örneklerin moleküler özelliklerinin dikkate alınmasının önemini daha da vurgulamaktadır. Bu çalışma, özellikle HHG sinyal yolu ile ilişkili olarak, CRC'nin ilerlemesi ve gelişiminde nöropeptidlerin rolünü doğrulamıştır. Bununla birlikte, nöropeptidler ve HHG sinyal yolağı arasındaki etkileşimleri açıklığa kavuşturmak için daha kapsamlı çalışmalar gereklidir. Ayrıca, ikinci yaklaşımdan belirlenen prognostik biyobelirteçler, klinik uygulama için potansiyel prognostik biyobelirteçlerdir, ve fakat geçerlilikleri KRK örneklerinin belirli alt gruplarında ayrıntılı doğrulama gerektirir. Genel olarak, bu araştırma, KRK'in heterojenitesinin daha iyi anlaşılabilmesine katkıda bulunmuştur ve kişiselleştirilmiş tedavi stratejileri için potansiyel olasılıklar sunabileceği ihtimali kuvvetlidir.

Anahtar Kelimeler: Kolorektal Kanser, Prognoz, sağkalım, Biyobelirteçler, Nöropeptidler, Hedgehog Sinyal Yolağı



To Nadia and Lenin

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my supervisor Prof. Dr. Ali Osmay Gure for giving me the opportunity to study under his guidance. More importantly, I would like to thank him for always encouraging me to only pursue my passions and supporting my curiosity, even when it might not have been the most conventional. My interest in joining his research lab was sparked during the department introduction on my first day of Bilkent as an undergraduate as he talked about his research. Taking on this journey as a freshman student in his lab has been an immense privilege. I am confident that his advices will continue to be valuable to me, regardless of my future path. Thank you hocam for your support and mentorship!

Secondly, I wish to extend my sincere appreciation to my jury members, Assoc. Prof. Özlen Konu Karakayalı and Assist. Prof. Seçil Demirkol Canlı. Their willingness to allocate their invaluable time for my thesis, combined with their insightful and expert feedback, significantly enriches my work. I genuinely appreciate their professional contributions.

I would also like to thank Secil. It is challenging to describe in words how much I appreciated her help and support throughout these years. In moments of frustration, her presence and guidance motivated me and kept me going. Despite her demanding schedule, she has supported me without any hesitation. I would like to thank her for believing in me, it truly has been an amazing experience working alongside her!

I would like to thank Mehdi Ghasemi, for all his help when I first started working. I would also like to thank all the AOG lab member Waqas, Baris, Ege, and Farid.

I would also like to thank Dr. Meral Üner for the insights regarding the interpretation of immunohistochemistry stainings.

And of course my friends from AOG lab. Marzana, Isli and Shila. Sadly, I did not have enough time to enjoy all of Marzana's positive and lovely energy, but even the smallest amount was absolute pleasure. Isli and Shila who truly are the best Albanian duo ever. Isli was the reliable and sweet friend with whom I could talk about all my frustrations. And Shila with her tough love and

support kept me going. I would like to thank them being there for me and making my first work trip the best I could have asked for. They have set the bar for the best colleagues pretty high!

And my Bilkent friends; Said, for being the peace I needed during all these hectic years. Alim, whose presence in the past year has helped pushing me forward. It took us a while to actually meet, but I am truly happy we finally did.

And I would like to thank the friends who became my family; Noor and Osama. Their friendship is one of my most precious achievements from Bilkent. I do not think Bilkent's journey would have been as perfect as it was without them. Thank you for letting me enjoy, laugh, and get extremely excited about all the smallest things.

Lastly, my family. I have been blessed with a great family. They have been my biggest rock and support. They truly are the best. I would like to thank my Mom and Dad, for being my best friends rather than my parents. Rojin, who is my best friend, older sister, supporter, and many more. Thank you for always being there by my side. You make going through life much easier for me.

Table of Contents

Chapter 1: INTRODUCTION	1
1.1. Colorectal Cancer	1
1.2. Colorectal Cancer Risk Factors	2
1.2.1. Molecular characteristics	4
1.2.1.1. Microsatellite Stability	4
1.2.1.2. Mutations	5
1.3. Colorectal Cancer Staging	7
1.4. Laterality in Colorectal Cancer	10
1.5. Common Colorectal Cancer classifications	12
1.6. Prognostic biomarker studies in Colorectal Cancer	13
1.7. Neuropeptides	16
1.7.1. Neuropeptides in cancer	17
1.7.2. Neuropeptides in Colorectal Cancer	18
Chapter 2: AIMS	20
Chapter 3: METHODOLOGY	21
3.1. Approach 1	21
3.1.1. Cohorts and Datasets	21
3.1.2. Principle Component Analysis	22
3.1.3. Hierarchical Clustering Analysis	22
3.1.4. Immunohistochemistry (IHC) Staining	22
3.1.5. Gene Set Enrichment Analysis (GSEA)	22
3.1.6. Functional Analysis	23
3.1.7. Statistical Analyses	23
3.2. Approach 2	24
3.2.1. Cohorts and Datasets	24
3.2.2. Cox Regression Analysis	24
3.2.3. Log Rank with Multiple Cut-offs (LRMC)	25
3.2.4. Kaplan Meier (KM) plots	25

3.2.5. KRAS/BRAF Mutant and Wildtype group definition.....	25
3.2.6. Univariate, Multivariate analysis.....	26
Chapter 4: RESULTS.....	27
4.1. Approach 1	27
4.1.1. Classification of CRC samples based on Neuropeptide genes.....	27
4.1.2. Determination of subset of Neuropeptides classifying the CRC samples	28
4.1.3. Stratification of the samples into Up and Down-regulated groups	32
4.1.4. Differential expression between Tumor and Normal sample groups	33
4.1.5. Immunohistochemistry (IHC) staining.....	36
4.1.6. Clinicopathological data distributions in the Up and Down-regulated groups	38
4.1.7. Biological and functional characteristics of the Up and Down-regulated groups.....	41
4.1.8. Expression profile in Tumor, Normal, and Metastatic groups	46
4.1.9. Prognostic associations of the 9 neuropeptide genes with Overall survival	48
4.2. Approach 2	50
4.2.1. Gene selection	51
4.2.2. Survival associations of the identified genes.....	54
4.2.2.1. Prognostic associations in All samples (Proximal, Stage 2 and 3)	54
4.2.2.2. Prognostic associations in Molecular subgroups	57
4.2.3. Univariate and Multivariate analysis of the Clinicopathological data	58
Chapter 5: DISCUSSION.....	60
5.1. Approach 1	60
5.1.1. Higher expression observed in Normal versus Tumor samples	62
5.1.2. Hedgehog signaling pathway in CRC development.....	63
5.1.3. CMS4 subtype molecular characteristics	64
5.1.4. HHG, CMS4 Subtype associations with Neuropeptides' overexpression.....	64
5.2. Approach 2	65
5.2.1. Advantageous aspects of the utilized methodology	66
5.2.2. The prognostic association of the 3 genes in All samples versus Subgroups	66
5.3. Limitations and Future Studies	68
Chapter 6: BIBLIOGRAPHY	72
Chapter 7: APPENDIX	78

LIST OF FIGURES

Figure 1.1 - Right-sided versus Left-sided CRC and their corresponding characteristics	10
Figure 1.2 - CMS subtyping and their corresponding molecular characteristics	13
Figure 1.3. The bi-directional interaction of neuropeptides and cancer cells in TME	18
Figure 4.1. The PCA results of A. Proximal, and B. Distal samples	28
Figure 4.2. The common genes between Distal and Proximal samples	30
Figure 4.3. Heatmaps of Distal and Proximal samples based on the 9 PC1 genes.....	32
Figure 4.4. Expression Profiles of the 9 genes in Normal versus tumor and up-regulated/ down-regulated samples in Distal and Proximal	34
Figure 4.5. Vasoactive Intestinal Peptide (VIP) Expression in Normal Colon vs. Colon Adenocarcinoma	37
Figure 4.6. Pituitary Adenyl Cyclase Activated Peptide 1 (ADCYAP1) Expression in Normal Colon vs. Colon Adenocarcinoma	37
Figure 4.7. Chi-Square results depicting the significantly distributed clinical parameters among the Up and Down-regulated samples for CMS and Gender	38
Figure 4.8. Chi-Square results depicting the non-significantly distributed clinical parameters among the Up and Down-regulated samples for MSI, KRAS/BRAF mutation, and stage	39
Figure 4.9. The common core enriched genes in Tumor-associated gene sets, and Abnormal morphology-associated gene sets between Distal and Proximal samples	42
Figure 4.10. The Venn diagram of the common genes between the common abnormal morphology associated and tumor-associated genes sets' core enriched genes	42
Figure 4.11. HHG signaling pathway networks	43
Figure 4.12. Immunosuppression-associated pathways network	43
Figure 4.13. Digestive and reproductive system development associated pathways network	44
Figure 4.14. The Venn diagram of the common Gene Ontology (GO) Biological Pathways (BP) gene sets between distal and proximal	45

Figure 4.15. Expression profiles of the 9 genes in normal, non-metastatic tumor, and metastatic tumor samples	47
Figure 4.16. LRMC graphs of Proximal, Distal, and combination of Proximal + Distal colon cancer patients based on different neuropeptide genes' expressions in TCGA	48
Figure 4.17. The workflow for the selection of the genes	53
Figure 4.18. Overall Survival (OS) association graphs of Proximal, Stage 2 and 3 CRC samples stratified based on GAA, OXNAD1, SEMA4C	55
Figure 5.1. Schematic demonstration of the association of the identified neuropeptide genes' up-regulation with the Hedgehog signaling pathway, CMS4 subtype, and TGF β response pathway and their impact on CRC progression	65

LIST OF TABLES

Table 3.1. The details of the dataset (TCGA) used in this study	21
Table 4.1. The top 20 PC1 and PC2 genes in Proximal and Distal samples	29
Table 4.2. The number of samples with expression values higher than 0	31
Table 4.3. The list of final 9 neuropeptide genes.....	31
Table 4.4. List of the 22 common gene sets between Proximal and Distal from Figure 4.14	45
Table 4.5. The datasets used and their details.....	51
Table 4.6. Examples of the selected and eliminated genes in GSE39582	52
Table 4.7. The list of the final 3 genes and their corresponding Cox P and HR values	53
Table 4.8. The common significantly associated subgroup for respective genes	58
Table 4.9. Univariate Cox P regression analysis of the Clinicopathological data in GSE39582 (OS), (RFS), and TCGA (OS)	59
Table 5.1. List of identified neuropeptide genes and their associations in CRC	62

LIST OF ABBREVIATIONS

CRC	Colorectal Cancer
MSI	Microsatellite Instable
MSS	Microsatellite Stable
OS	Overall Survival
RFS	Relapse-Free Survival
DFS	Disease-Free Survival
CMS	Consensus Molecular Subtypes
NPP	Neuropeptide
PCA	Principal Component Analysis
GSEA	Gene Set Enrichment Analysis
LEA	Leading Edge Analysis
GO	Gene Ontology
IHC	Immunohistochemistry
KB	Kras/Braf
MMR	Mismatch Repair
TME	Tumor Microenvironment
HHG	Hedgehog Signaling Pathway

CHAPTER 1

INTRODUCTION

1.1. Colorectal Cancer

Colorectal cancer is the third-most commonly diagnosed cancer, with an estimated 1,880,725 cases diagnosed in 2020, and the fourth major cause of cancer mortality, in men and women, globally. According to Turkish Health Ministry statistics, colorectal cancer ranks third in both men and women, with a frequency of 13.8 and 22.8 per hundred thousand in men and women, respectively. Colorectal cancer survival rates can vary depending on a number of factors, and it should be emphasized that colon and rectal cancer survival rates are presented separately. The overall 5-year survival rate for individuals with localized and non-metastatic colorectal cancer can be as high as 91%. In patients with metastasis to the neighboring tissues or lymph nodes, it is estimated to be 70%. Lastly, if the disease has progressed to other organs, the 5-year survival rate might drop as low as 15%. [1, 4]

According to the global distribution of colorectal cancer burden, Asia has the largest mortality rate for colorectal cancer. China and Japan have the highest number of deaths per year, 280,000 and 60,000, respectively. Further research by the *International Agency for Research on Cancer* indicates that colorectal cancer death rates are expected to rise by 56% over the next 20 years. [2] It has been shown that colorectal cancer is more common in industrialized nations. However, its prevalence is rising in middle and low-income countries. The high number of annual cases, along with an increasing prevalence of early-onset colorectal cancer, is posing an alarming global public health challenge. [3]

1.2. Colorectal Cancer Risk Factors

Due to the rising risk of colorectal cancer incidence and mortality rates, it has now become even more crucial to increase public awareness regarding the risk factors that can lead to an increase in colorectal cancer development. Understanding these risk factors can be beneficial in the prevention or early diagnosis of the disease, resulting in a better prognosis.

Genetic background

Colorectal cancer can develop spontaneously or hereditarily. Nevertheless, most identified cases, around 70-80%, are sporadic, meaning that disease development is caused by changes in genetic composition, such as somatic mutations, rather than the genetic composition at birth. The condition cannot be handed down to future generations in these cases. In hereditary cases, which account for only 20% of all cases, the illness is caused by the genetic profile at birth, such as gene mutations that are handed down from generation to generation. Moreover, people with a family history of colorectal cancer have an increased risk of developing the illness than those without a hereditary factor. These patients account for around 10-15% of all colorectal cases.

Underlying Diseases

In addition to the genetic basis, the presence and history of other diseases can also lead to an increased risk of colorectal cancer development in patients. The main diseases include:

- Inflammatory bowel disease (IBD): IBD causes chronic inflammation in the large intestine and increases the likelihood of colorectal cancer development.
- Adenomatous polyps (adenomas): Adenomas are non-cancerous polyps that can develop into malignant ones over time. That is why patients are advised to have these polyps

removed through a colonoscopy, to lower the chance of colorectal cancer development and undergo regular checkups with their doctors, as polyps can reoccur.

- Other cancer types: Patients previously diagnosed with uterus and ovarian cancer are at a higher risk of developing colorectal cancer.
- Some other disorders, including Lynch syndrome, are also associated with a higher chance of colorectal cancer development.

Age and Race

Some other factors that can increase colorectal cancer are age, gender, and race:

As the age of the patients increases, the chance of colorectal cancer also increases, with the majority of patients being above 50 years old. Furthermore, as the patient's age increases, they also face other challenges, such as the deteriorating ability of their bodies to recover from the effects of the treatment. However, it is important to note that cancer development patterns are changing, as the rate of colorectal cancer in adults over the age of 55 is declining by 3.6%, while the rate in adults under the age of 55 is increasing. Different populations with distinct races and ethnicities have different genetic compositions and environmental factors impacting their risk of cancer development. As such, some populations and ethnicities are more prone to colorectal cancer. For example, populations in East Asia, such as China and Japan, have a higher incidence rate than those in the rest of the world. Even within countries, certain subsets of populations are more to the disease than others, like the African American population in the USA.

Lifestyle Habits

Colorectal cancer is also greatly impacted by the lifestyle choices of the individual. The three main lifestyle patterns that can significantly increase the chance of colorectal cancer development are as follows:

- Physical activities: Lack of physical activities increases the risk of colorectal cancer development as lower physical activities, such as regular exercise routines and prolonged sitting, can lead to obesity.
- Diet: a positive correlation between an increased risk of colorectal cancer development and a diet rich in red and processed meat has been observed.
- Smoking: it has been shown that smokers are more likely to develop colorectal cancer compared to non-smokers. [2]

1.2.1. Molecular characteristics

Colorectal cancer is one of the most molecularly diverse malignancies, with several subtypes and classifications. This level of tumor heterogeneity stems from the unique mix of genetic alterations and molecular characteristics in each tumor and individual. Among these characteristics, two distinct known molecular features increase the likelihood of developing colorectal cancer: microsatellite instability and mutation in multiple genes, the most important of which are BRAF, KRAS, and TP53. Since all colorectal cancer tumors and individuals differ drastically, treatment becomes considerably more challenging as it must be personalized for each patient. Thus, it is of great importance to address these molecular characteristics and their link to colorectal cancer development. [5]

1.2.1.1. Microsatellite Stability

Microsatellite sequences are the repeating short DNA sequences. Microsatellite instability (MSI) is an epigenetic phenomenon that implies an alteration in the number of these microsatellite sequences, as compared to the individual's germ-line DNA. It can be caused by mutations in mismatch repair-related genes, such as MLH1, MSH2, MSH6, and PMS2, or by a deficiency of DNA mismatch repair proteins, a condition which is commonly referred to as

deficient Mismatch repair (dMMR). As a result of these mutations, protein deficiencies, and changed repetition counts of microsatellites, the cells become hypermutable, which can lead to colorectal cancer development. [6, 8]

As such, colorectal cancer tumors are categorized based on their mismatch repair status into three subgroups; Microsatellite stable (MSS), Microsatellite instability-low (MSI-L), and Microsatellite instability-high (MSI-H). However, for clinical reasons, researchers consider MSS and MSI-L as a single group; MSS, and MSI-H as a separate group; MSI. [5]

The MSI phenotype has been used as an important prognostic biomarker in colorectal cancer, but its predictive value varies depending on the stage of cancer. [7, 8] Clinical studies have shown that the distribution of MSI frequency rates decreases as cancer progresses through the stages, with Stage 2 having the highest MSI rate of 22%, Stage 3 with 12%, and Stage 4 with only 2%. [9]

The MSI phenotype is characterized by the accumulation of insertions and deletions in microsatellites caused by the increase in the frequency of different mutations because of the deficiency of mismatch repair proteins (dMMR). There are different genes and several epigenetic, somatic, or germline gene inactivation mechanisms involved in tumors with dMMR. The most common inactivated gene in these tumors is MLH1 (with a hypermethylated promoter), accounting for 95% of the cases, compared to other genes, such as MSH2 and MSH6, that account for the remaining 5%. [10, 18]

1.2.1.2. Mutations

As previously stated, an insufficient repair pathway might result in mutations that drive cancer growth. Colorectal cancer is caused by the aggregation of all the newly acquired and pre-

existing mutations in the genome. Comprehensive genomic analysis has demonstrated that some genes are particularly important and more frequently mutated in colorectal cancer. These driver mutations in colorectal cancer are; APC, KRAS, BRAF, TGFBR2, and TP53. [11, 12] BRAF and KRAS mutations happen in a mutually exclusive manner in CRC, which could mean that there is functional redundancy. Additionally, a double mutant model (Kras and Braf) could be beneficial in the stratification of CRC samples. [13, 14]

BRAF

BRAF is a proto-oncogene that belongs to the RAF serine/threonine protein kinases family, which regulates the MAPK/ERK signaling pathways. A valine-to-glutamate change at position 600 (V600E) of this gene is present in 10% of colorectal cancer patients. These patients are generally older and are more likely to have proximal tumors. Furthermore, female patients with this mutation have a worse relapse-free survival prognosis than male patients. On the other hand, individuals with an overlap between BRAF mutation and microsatellite instability have a good prognosis. Among the BRAF-mutated CRC patients, a subset also has hypermethylated promoter CpG island sites and is classified as CIMP (CpG island methylator phenotype) positive patients. Around 25% of all BRAF-mutated CRC patients are CIMP-positive, and almost half of these patients are also MSI which accounts for around 10% of all CRC patients. [11]

Although the BRAF V600E mutation is less common in patients with metastatic CRC, it is linked with a worse prognosis in these patients. On the other hand, the prognostic role of this mutation in non-metastatic cancer patients is more complicated given that other molecular characteristics of the patients, such as their microsatellite stability and stage of their disease, are also heavily involved in the prognosis. [10, 14]

KRAS

KRAS is a G protein from the Ras family of proteins, which is involved in intracellular signaling pathways and are among the most researched genes in cancer genesis and progression. KRAS is the most frequently mutated isoform among all isoforms in various cancer types. Mutations in this gene lead to its own activation, and as a result, the activation of its downstream signaling pathways. Some of these series of events results in resistance to surface kinase receptor inhibitions.

KRAS has been identified as a prognostic biomarker in CRC as a consequence of these mutations and their impacts. According to several studies, the prevalence of KRAS mutations is the highest in the early stages of CRC. [15] Moreover, although the majority of research suggests an unfavorable prognosis linked with this gene's mutated phenotype, there are conflicting data that show the opposite pattern. Additionally, some studies suggest a higher recurrence frequency for stage 2 and 3 CRC patients with KRAS mutation, while some other studies consider this marker specific to stage 1 and 2 CRC patients only. These conflicting results might be attributed to the heterogeneous nature of CRC tumors and the need for a more stratified study. [16, 17] Lastly, for a better understanding of the prognostic significance of all these markers, a more thorough study connecting multiple driver gene mutations and microsatellite stability/instability, location, and stage should be done. Due to the limited data available, only KRAS and BRAF mutations are examined in this study.

1.3.Colorectal Cancer Staging

The oncologist's initial step in the clinic is to establish the stage of the patient's cancer, as this determines what type of therapy should be considered for the patient. The degree of the tumor's

growth and the tissue into which it may have protruded, as well as its metastasis to other regions of the body, are all factors when deciding the stage.

TNM staging is the most used colorectal cancer staging system. The cancer stage is defined by three factors in this system: tumor (T), node (N), and metastasis (M). The term "tumor" refers to the protrusion of the primary tumor into the colon lining and goes from T0 to T4b, with no tumor detected in T0 and the spread of the tumor into all layers of the colon and possibly extending to other organs in T4b. The term "node" in this staging system refers to the affected lymph nodes. The range of this variable is from N0 to N2b, where N0 implies that no tumor/ cancer cell was identified in the patient's lymph nodes near the colon or rectum, and N2b indicates that tumor/ cancer cells were found in at least 7 of the lymph nodes near the colon or rectum. Lastly, the term "metastasis" refers to the tumor's spread to various body organs. This variable goes from M0 to M2c, with no tumor spread observed in other organs in M0 and the detection of tumor spread in other organs, such as the liver, in M2c.

The TNM staging system is formed by combining these three components. Patients are classified into Stage 0 through Stage 4 based on these combinations. (Based on the 7th edition)

- Stage 0: Cancer in these individuals is confined within the inner lining structure of the colon and has not spread to the lymph nodes or other organs.
- Stage 1 (T1 or T2, N0, M0): Cancer has spread from the inner lining of the colon/rectum but has not spread to surrounding lymph nodes or other organs.
- Stage 2 (T3 - T4b, N0, and M0): cancer has spread from the colon/rectum lining into the surrounding structure, but there has been no metastasis to lymph nodes or other organs.

- Stage 3 (T1 - T4b, N1 - N2b, M0): cancer has spread to at least 1-3 surrounding lymph nodes, independent of how far it went through the colon/rectum lining, but there has been no metastasis to other organs.
- Stage 4 (T4a - T4b, N1 - N2b, M1 - M2c): cancer has metastasized to at least one other organ, in addition to at least 1-3 surrounding lymph nodes. [1]

Importance of Biomarker Studies for Stages 2 and 3

As mentioned earlier, the stage of the disease influences the treatment plan's course of action. In patients with Stage 0, only tumor surveillance is recommended because the tumor is confined. On the other hand, oncologists recommend a radical hemicolectomy or lymph node resection for Stage 1 colon cancer patients. These individuals have a 5-year survival rate of 93%, which declines as the disease proceeds through the stages, from 85% in the early stages of Stage 2 to 44% in the later stages of Stage 3. In terms of treatment choices for Stages 2 and 3, individuals frequently undergo surgery followed by adjuvant chemotherapy. Unfortunately, the survival rate declines even more dramatically from Stage 3 to 4, with individuals with Stage 4 CRC only having an 8% chance of survival. For Stage 4 patients, prognostic biomarkers are not effective since cancer has advanced and spread in these patients, so surgery and a combination of various treatments are recommended. There is also a high risk of tumor recurrence at a distant site. [20, 21]

Overall, individuals in Stages 2 and 3 have the most heterogeneous tumors. [22, 23] Several studies have shown that the outcomes of treatments for these patients vary greatly from one individual to the next, as one patient responds to a chemotherapy treatment very efficiently while another patient shows resistance to the same treatment [21] Consequently, it is critical to focus specifically on Stages 2 and 3 patients, especially considering that the prognosis for Stage 1 patients after the aforementioned treatments is very favorable, and the prognosis for Stage 4 patients is extremely

poor. Therefore, the identification of Stage 2 and 3 specific biomarkers may be used to further classify patients with various molecular characteristics, which can assist in improving the efficacy of treatments for individuals without the use of toxic treatments patients are resistant to. [33]

1.4. Laterality in colon cancer

One of the shortcomings in CRC biomarker studies is considering the colon as a single organ and presuming that the tissue lining the colon is uniform throughout. However, studies revealed that the colon splits into two embryonically distinct tissues with very different molecular properties. The colon is divided into two sections: the proximal (right-sided) colon and the distal (left-sided) colon. The proximal colon consists of the cecum, appendix, ascending colon, hepatic flexure, and 2/3 of the transverse colon, whereas the distal colon consists of 1/3 of the transverse colon, splenic flexure, sigmoid, and descending colon. (The transverse colon is considered proximal in this study). [25, 27]

Proximal Colon		Distal Colorectum
Incidence		Incidence
Lower than Distal		Higher than Proximal
Higher in females		Higher in males
Increasing number of cases		Decreasing number of cases
Presentation		Presentation
Higher TNM stage		Lower TNM stage
Larger tumors		Smaller tumors
More mucinous type		
Genetics		Genetics
Common site for CRC in MUTYH-associated polyposis		Common site for CRC in familial adenomatous polyposis
Immunology		Immunology
More active immune cells promoting immunogenicity		Immunologically less active, promoting tolerogenesis
Molecular Pathway		Molecular Pathway
CIMP/MSI/BRAF positive tumor predominate		Chromosomal instability tumors predominate
Survival		Survival
Worse overall Survival		Better overall Survival

Figure 1.1 - Right-sided versus Left-sided CRC and their corresponding characteristics.

(Modified from [25])

Molecular Characteristics Differences

The main reason for separating the colon into two separate entities is due to their different embryologic origins. The cell migration and differentiation phases of development require specific and distinctive developmental gene expression patterns. Similar to the other parts of the body, such distinct patterns also exist in the colon. The fundamental causes for the significant variation in the proximal and distal colon later in life are the combination of these two elements; different embryologic origins and unique gene expression patterns. In addition to the molecular distinctions seen in Figure 1.1, the proximal and distal colons exhibit very different morphologies. The proximal colon is characterized by its flat and mucinous morphology, which has a higher metastasis rate compared to the distal colon, which is more polypoid-like. [29]

Prognosis

The above-mentioned biological differences, combined with the differences in molecular characteristics between distal and proximal CRC, lead to different overall prognoses. Proximal CRC has been shown to have a 5% decrease in survival compared to distal tumors. [24, 25] There are various different hypotheses as to why that might be the case. One reason could be the location of the tumor; since the proximal tumor is further away from the rectum, clinical symptoms, such as bleeding, are observed when it has progressed to the later stages of the disease. Several studies confirm this, as they show that the chances of proximal CRC patients presenting a more advanced disease are higher. [28, 30] Moreover, there is a difference in the incidence rate of distal and proximal CRC, with distal tumors having a higher incidence rate than proximal tumors. [26, 30] Lastly, the prognostic significance of a biomarker and the tumor's location are very closely associated, as evidenced by a previous study from our group that combined ULBP2/SEMA5 expression as a prognostic gene signature in distal CRC. [31, 32] This strong association also

indicates the importance of considering the location of the CRC tumors as an influential factor while performing biomarker studies and in the determination of treatment plans. [29]

1.5. Common molecular and risk classifications in colorectal cancer

Oncotype DX

One of the most prevalent prognostic classifications of CRC is based on the Oncotype DX scoring system. For a more successful and effective therapeutic plan and course of action in treating patients, a precise assessment of recurrence risk is essential. Marisa L. et al. created the 12-gene Oncotype DX Colon Cancer Recurrence Score Assay to address this requirement. To predict recurrence risk, the model was developed using gene expression data from 1,851 stage 2 and 3 CRC patients from four large, independent research studies. The risk score calculation uses 12 genes consisting of 5 reference genes and 7 recurrence-associated genes. The Risk score ranges from 0 - 100, where samples with a score equal to or higher than 41 are considered the “high risk” group, and samples with a score less than 41 are considered the “low risk” group. [33, 34]

Consensus Molecular Subtypes (CMS)

Unfortunately, the attempts at classification of these tumors only based on their gene expression profile cannot provide consistent classifications, despite many studies trying to do so. That is because all the above-mentioned molecular characteristic differences in CRC tumors make them one of the most heterogeneous tumors. Thus, for a more consistent classification, these molecular characteristics need to be considered. An international consortium utilized large-scale shared data and created the four *Consensus Molecular Subtypes* referred to as CMS1-4. Each of these CMS subtypes consists of sample subgroups with specific molecular features; CMS1 is the MSI group which has an activated immune system, CMS2 is the epithelial subgroup with WNT signaling

activation as its marker, CMS3 is also an epithelial subgroup with prominent metabolic dysregulation, and lastly, CMS4, with the worst prognosis, is the mesenchymal subgroup that also has an activated stroma. Further details of the characteristics of each subtype are depicted in Figure 1.2. Since this classification is based on molecular characteristics, it has been shown to be more robust than other available classifications, and its utilization in biomarker studies can increase the success rate of clinical transition of the diagnostic/ prognostic biomarkers. [35]

CMS1	CMS2	CMS3	CMS4
MSI immune	Canonical	Metabolic	Mesenchymal
MSI CIMP high Hypermethylation	SCNA high	Mixed MSI status SCNA low CIMP low	SCNA high
BRAF mutations			
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	stromal infiltration TGFB activation Angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

Figure 1.2 - CMS subtyping and their corresponding molecular characteristics. (Modified from [35])

1.6. Prognostic Biomarker Studies in Colorectal Cancer

As CRC is one of the most aggressive tumors and a leading contributor to cancer deaths, it is crucial to have effective monitoring methods to quickly diagnose the disease at earlier stages and find prognostic markers to determine the precise outcome of the disease. In many countries, such as Eastern Europe and the USA, newly implemented improved treatment plans that utilize some of the clinically used biomarkers have shown a decline in cancer deaths to some extent. [36] However, to further improve the overall survival of patients, more personalized prognostic markers

need to transition into clinics. One of the main reasons for the hesitance of physicians in utilizing the prognostic/ predictive biomarkers in clinics is the lack of consistent and robust results across different studies. [37, 38]

KRAS, BRAF mutations, and microsatellite stability of the samples are some of the main factors in the progress and prognosis of CRC. These two factors are also widely known to physicians and are part of the routine testing done in the clinics. [38, 39] Additionally, the CMS subgroups have greatly improved the specificity of these biomarkers to one molecularly distinct group of patients. However, CMS subgroups do not integrate the stage of the disease, and as mentioned previously, the survival rate drastically drops from stage 1 to stage 4. Consequently, there should be a further classification considering all the CMS subtypes as well as the key clinical/ molecular characteristics of patients. Only then can a more adequate treatment plan, which can avoid excessive toxicity, be developed. [40, 42]

Issues in clinical translation

The emphasis on the determination of robust and consistent biomarkers is to solve the issue of translating them into clinics. Regardless of the vast amount of biomarkers available, only a small fraction of those are usable in the clinics. [35] While this stage is very challenging, it is also one of the most important ones for ensuring the best outcome of the treatment plan for each patient group. There are various factors affecting the inconsistencies between clinically used and prognostic biomarkers. These reasons are as follows; lack of large datasets for biomarker studies and thus the consequent lack of systematic validation studies, no suggested stratification threshold, the scoring system offered by these studies, discrepancies between the technologies [40, 43] used in each study, and the presence of various patient populations. Koncina E. et al. suggest a set of

criteria for assessing the strength of prognostic biomarkers in regard to these issues. These criteria are listed below;

1. The prognostic significance of the biomarkers must be observed in large datasets
2. Homogenous expression within and between tumor tissues should be preferred;
 - a. Multi-gene panels are better than single-gene biomarkers
 - b. Platform/ technology differences should be considered when comparing different datasets
3. A good scoring system should be offered with the biomarkers for easier clinical translation, and/or a defined threshold for categorizing patients based on their qPCR gene expression levels should accompany the biomarkers for easier clinical translation.
4. The validation step must be done very thoroughly and systematically.
5. All prognostic biomarkers should be validated through clinical studies. [40]

In addition to the issues stated above, other factors can explain the inconsistencies observed between the discovery and validation studies done for one biomarker or in clinical validation. The molecular differences between the distal and proximal colon have recently been integrated into a few biomarker studies. There is also evidence that these biological and molecular differences also extend to differences in the gene expression profile of distal and proximal tumors. Thus, considering the primary tumor's location is an important parameter. [41]

Genes with multiple Probesets in Microarray datasets

Another important factor that could explain the observed inconsistencies is the presence of multiple probesets for one gene in microarray datasets but not in RNA-seq dataset. These different

probesets may refer to different isoforms or different splice forms of one gene. However in the vast majority of transcriptomic studies, the differences between multiple probesets of a gene are overlooked and the studies are performed at a gene level rather than a probeset level. Recent studies have shown that all of these different transcripts and splice forms need to be considered separately, as they can have differential expression and prognosis. A study in 2013 showed that the different isoforms of a gene might be functionally distinct. [44] This suggests that the different probesets/isoforms might be biologically different, which could explain why many genes have shown differential expression levels at the isoform level, with some of the genes' transcripts even having opposite expression profile patterns. The importance of considering the differential isoform level expression is to minimize the risk of cancellation or domination of one isoform over another, which can lead to over or under-representation of one isoform. Therefore, to eliminate potential inconsistencies, there should be a probeset/ transcript level analysis performed at the beginning of any transcriptomic analysis using Microarray datasets. [45]

1.7. Neuropeptides

Neuropeptides are small amino acid chains released from neurons that act as regulatory molecules and chemical messengers. They are synthesized from prohormones as larger peptides that undergo posttranslational cleavage for activation into their neuropeptide forms. These are then packaged into vesicles which are stored in the synaptic terminals until they are released and transported throughout the body. They are mostly found in the central, peripheral nervous system in addition to the gastrointestinal and cardiovascular system. They induce the activation of diverse downstream signaling as they bind to their G-Protein Coupled Receptors (GPCRs). However, their effect depends on various factors, such as the location and the amount of the neuropeptide release. [61, 63]

1.7.1. Neuropeptides in Cancer

Cancer cells that have abnormal growth require support from their surrounding cells to promote and sustain their growth. The tumor microenvironment (TME) consists of a variety of cells, and it is an essential element in sustaining tumor growth. It has been shown that not only is there local neuropeptide release in the TME aiding the maintenance of solid tumors through paracrine and autocrine signaling, but the changes in their expression are also associated with the progression of tumor cells. Neuropeptide release is dosage dependent, and as the firing rate increases, the neuropeptide release also increases. The interaction between the tumor and the stromal cells in the TME leads to firing rate changes that lead to higher neuronal activity. As a result of this increase, the local neurons' excitability patterns change and release more neuropeptides which induce cancer growth, metastasis, and progression. Recent studies have described this interaction between the cancer cells and the neurons as bi-directional. As a result of the dosage dependency release of neuropeptides, the production and release of the lowly expressed neuropeptides in healthy conditions change due to inflammation or tumorigenesis and, in return, increase the local release of more neuropeptides. Hence, cancer cells induce the release of neuropeptides which help the advancement of the disease. (Figure 1.3) [62, 63]

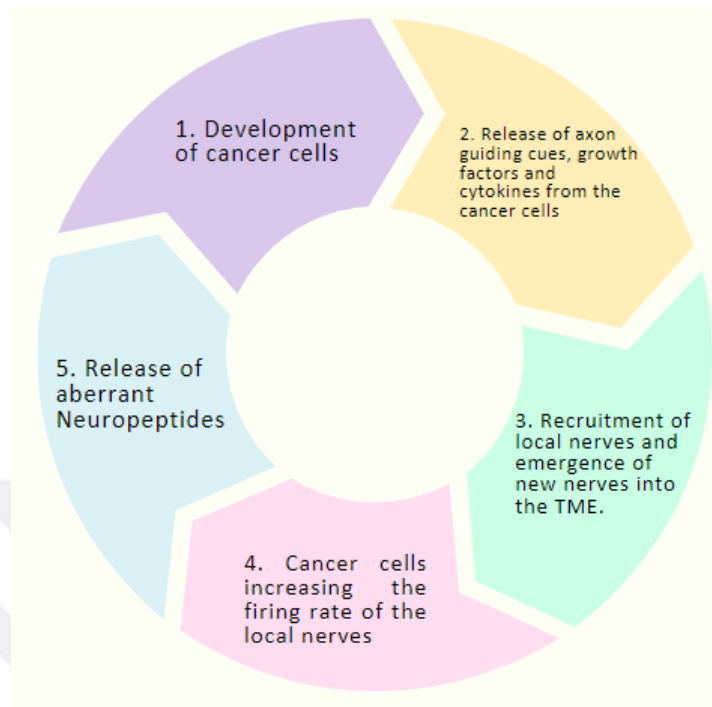


Figure 1.3. The bi-directional interaction of neuropeptides and cancer cells in TME.

(Modified from [61])

1.7.2. Neuropeptides in Colorectal Cancer

The studies focusing on the role and involvement of neuropeptides have become a new interest for researchers as they have been shown to have abnormal behaviour in cancer cells and their association with multiple cancer types (such as; CRC, breast, lung, glioblastoma) and their progression has been confirmed. Additionally, an increase in cancer cases has been related to an increase in stress in the population. Under stressful situations, various neuropeptides like serotonin and dopamine are released which can inhibit or activate cell cycle or proliferation pathways. CRC is one of the most common cancer types worldwide, and as neuropeptides are involved in the regulation of normal gastrointestinal system function, the dysregulation in the expression of these neuropeptides can explain the observed symptoms in CRC patients. Among the studied neuropeptides, the up-regulation of several has been linked to poor cancer prognosis and shown to

be potential drug targets in CRC. Some of these include Bombesin (BB), leading to cancer progression by acting as a morphogenetic molecule, and Neuropeptide Y (NPY), by modulating the proliferation of cancer cells. However, deeper insight into the involvement of neuropeptides and the nervous system in the development and progression of CRC can be beneficial in the development of better and more effective treatment options. [64]



CHAPTER 2

AIMS

The studies in this thesis were conducted through two different approaches with different focus points;

1. **APPROACH 1 – Neuropeptide-based classification of Colorectal Cancer (CRC) patients:**

In light of recent discoveries about the role of nervous neuropeptides in cancer progression, the aim of this part of the study was to evaluate the potential of neuropeptides as biomarkers to be used for classification of colorectal cancer (CRC) patients, and to conduct a functional analysis of neuropeptides to gain a deeper understanding of their biological role in CRC progression. The focus of this approach was the utilization of a limited gene list.

2. **APPROACH 2 – Identification of robust prognostic biomarkers in Proximal, Stage 2 and 3 Colorectal Cancer (CRC) patients:**

The aim of this approach was to determine robust prognostic biomarkers in Proximal CRC patients. Additionally, we attempted to consider the factors contributing to the inconsistencies observed between research to clinical transitions, such as; location, stage, probeset discrepancies, and technological differences. The focus of this approach was the utilization of a limited sample list.

CHAPTER 3

METHODOLOGY

3.1. Approach 1: Neuropeptide-based classification of CRC patients

3.1.1. Cohort and Dataset

TCGA COAD was used as the main dataset for this part of the study. The gene expression and clinical data were downloaded through the GDC portal website, accessible at <https://portal.gdc.cancer.gov/>. The normalization was done via the Deseq2 Package available for R (version 3.1.4). The complete detail of the sample numbers is presented in Table 3.1. The samples selected for this study were COAD primary tumors and their associated normal samples. TCGA provides the specific sections of the colon. Thus all the samples were classified into proximal and distal, as indicated below; [50]

Proximal: cecum, appendix, ascending colon, hepatic flexure, and the transverse colon

Distal: splenic flexure, sigmoid, and descending colon

Table 3.1. The details of the dataset (TCGA) used in this study

Dataset	Dat Type	Platform	Samples Number			
			Distal		Proximal	
			Primary Tumor	Tumor Adjacent Normal	Primary Tumor	Tumor Adjacent Normal
TCGA	RNA Seq	Illumina	147	10	226	18

3.1.2. Principal Component Analysis (PCA)

PCA was performed on R (version 3.1.4) with the prcomp function from the stats package, and the data was scaled and centered. The input data was transported so that the samples were as rows and genes as the columns. Additionally, the loadings of the genes in each PC were extracted by using the rotation value of the prcomp results. [65]

3.1.3. Hierarchical clustering analysis

The samples were standardized and hierarchically clustered using Euclidean distance and complete linkage as the method on Cluster 3.0 software, and the .cdt result files were plotted using Java TreeView. Based on the clusters observed in each figure, the samples were divided into the “Up” and “Down” regulated groups.

3.1.4. Immunohistochemistry (IHC) staining

The immunohistochemistry staining data was evaluated with the aid and supervision of Dr. Meral Üner. For the analysis, the Normal Colon (tissue) and Colon Adenocarcinoma (Pathology) IHC staining of each of the genes were examined from the Human Protein Atlas website (<https://www.proteinatlas.org/>), and the presence or absence of staining was noted in each of the samples. [60]

3.1.5. Gene Set Enrichment Analysis (GSEA)

The GSEA software was downloaded from the Broad Institute website. (<https://www.gsea-msigdb.org/gsea/index.jsp>) GSEA was performed using the gene symbols, and Up and Down-regulated samples were used as the two different phenotype groups defined on the software. Additionally, the Gene Ontology (GO), c5_all, was used as the gene set database for the analysis. The .edb output files were then used to run Leading Edge Analysis (LEA) using GSEA software

with the gene sets of interest. The LEA barplots were used for the extraction of the Core enriched genes in these gene sets. The extracted core genes in different gene sets were plotted in Venny diagrams, and the common genes were found. [54, 55] The common genes from the distal and proximal analysis were also plotted in Venny diagrams. [52]

3.1.6. Functional analysis

The freely available EnrichNet website (<http://www.enrichnet.org>) was used for the network-based functional analysis of the core enriched genes. The core enriched genes were used as the input, and STRING as the molecular network type. [56] The developed networks were downloaded and visualized on Cytoscape version 3.9.1 for Windows with the default styling. (<https://cytoscape.org/index.html>) [57]

3.1.7. Statistical analyses

The expressions of all the genes in the paired normal samples, up-regulated and down-regulated samples, were plotted in bar plots on GraphPad Prism version 8 for Windows (www.graphpad.com). For the difference significance calculation between the different sample groups, student t-tests were performed with P values of lower or equal to 0.05 being considered statistically significant. [53] The Chi-square test was performed on the Chi-square calculator website with the significance level set to 0.05. (<https://www.socscistatistics.com/tests/chisquare2/default2.aspx>) [58] The normal, tumor, and metastatic plots were plotted on the TNMplot website (<https://tnmplot.com/analysis/>) using the TNM plot from the gene expression comparison tab. [59] The Venny diagrams were plotted on the freely available Bioinfogp website (<https://bioinfogp.cnb.csic.es/tools/venny/index.html>) [52]. The single cell RNA sequencing analysis was performed using “Seurat” R package with the

available cell types for CRC in the GSE178318 dataset. This analysis was performed by Dr. Seçil Demirkol Canlı. [68]

3.2. Approach 2: Identification of robust prognostic biomarkers in Proximal, Stage 2 and 3 CRC patients

3.2.1. Cohorts and Datasets

In this part of the study, three different datasets were used, among which there were two separate platforms (Microarray and RNA-Seq). The GEO microarray datasets were downloaded from the GEO website, which can be accessed through <https://www.ncbi.nlm.nih.gov/geo/>. TCGA, the RNA-Seq dataset, was also downloaded through the GDC website, accessible at <https://portal.gdc.cancer.gov/>. The clinical data of each dataset were obtained from their corresponding links mentioned above. The microarray datasets were RMA normalized via BRB ArrayTools available in Microsoft Excel, and TCGA was normalized with the Deseq2 Package available for R (version 3.1.4). [49, 50]

3.2.2. Cox Regression analysis

The Cox regression analysis was performed with the continuous gene expression of all the genes in each dataset and survival data of the selected samples for the determination of the linear relationship between these two variables, using the maxstat and survival packages on R (version 3.1.4). The same script was later used to determine whether there was any association between the clinical variables of the samples and the survival rate in a univariate and multivariate analysis. [66, 67]

3.2.3. Log Rank with Multiple Cut-offs (LRMC) analysis

The LRMC graphs are plotted using an R in-house script. The graphs are generated by setting each of the gene expression values of each patient as a threshold for separating the patients into two high and low expressing groups, performing Log-Rank tests between these two defined groups at each threshold, and plotting the calculated p-values for each corresponding threshold. The colors blue and red colors are the indicator of the gene expression association with good or poor prognosis, respectively. The horizontal dotted line on the graph represents the significance threshold of 0.05, while the vertical lines are the first, second, and third quantiles (25th, 50th, and 75th percentile expression values). [18]

3.2.4. Kaplan Meier (KM) plots

The cut-offs for defining the high and low sample groups were defined based on the LRMC results. To ensure close to equal sample numbers between up and down-regulated groups of samples, we chose a significant cut-off closest to the median with the lowest P-value. The KM plots were plotted on GraphPad Prism.

3.2.5. KRAS/BRAF Mutant and Wildtype group definition

Since KRAS and BRAF mutations are mutually exclusive, the mutant groups were defined as mutant if one of the genes were mutated, and wild-type groups if none of the genes were mutated. Additionally, if one gene was unknown and the other gene mutated, the sample was assigned to the mutant group. However, the sample was assigned to the unknown group in the case of one gene being unknown and the other gene being wild-type.

3.2.6. Univariate, Multivariate analysis

The Univariate and Multivariate Cox regression analyses were performed on the SPSS Statistics 25 software. The Multivariate model was generated by using the ‘backward: Wald’ method of the Cox Regression function on SPSS.



CHAPTER 4

RESULTS

4.1. Approach1: Neuropeptide-based classification of CRC Patients

This approach aimed to classify colon cancer patients based on neuropeptide (NPP) genes. After the normalization of the samples with all the available genes in TCGA, the genes were filtered with the neuropeptide gene list extracted from NeuroPep's neuropeptide database; (<http://www.neuropeptides.nl/>). There are a total of 100 neuropeptide genes. However, only 85 were available in the TCGA dataset. All 100 neuropeptide genes are listed in Appendix A. [51]

* As the importance of separation of proximal and distal has been discussed previously, all the analyses performed in this part of the study are performed for distal and proximal samples separately.

4.1.1. Classification of CRC samples based on NPP genes

In order to find a subset of NPP genes with the ability to classify the samples into two groups, we used PCA. This unsupervised learning method could confirm the presence of different clusters within our sample based on the expression of the NPP genes in addition to the determination of the subset of the genes responsible for the observed clusters. The PCA results indicated the presence of two different clusters, as depicted in Figure 4.1.

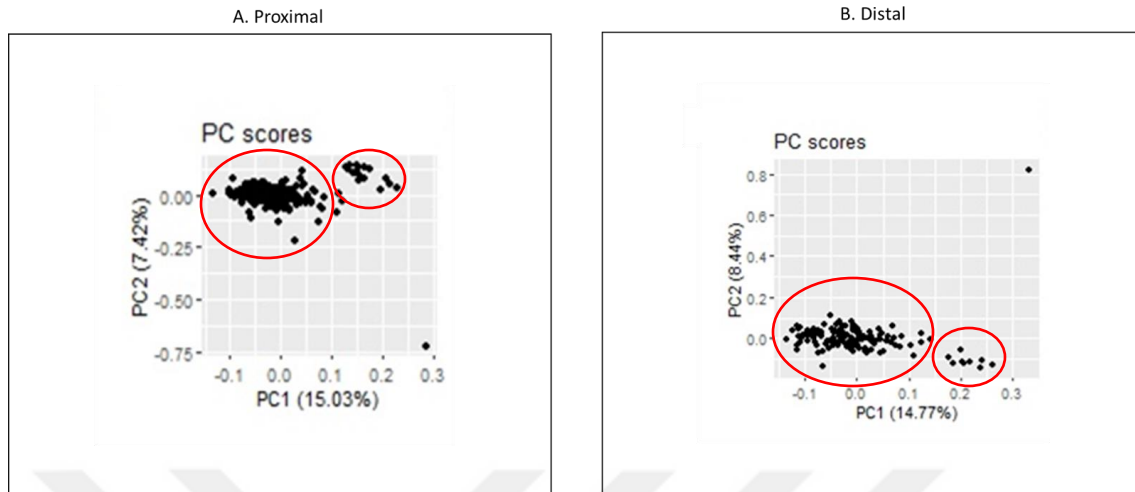


Figure 4.1. The PCA results of A. Proximal, and B. Distal samples. The dots on PCA plots represent the samples. These plots show the presence of two groups in both proximal and distal samples. The two groups are circled on the diagram.

4.1.2. Determination of the subset NPPs that classify the CRC samples

For the selection of the Principle Components (PCs) to be used for the further analyses, we chose the top 2 PCs in terms of their variance. After the determination of the PCs, the loadings of the genes, which are the weights of the genes in each PC, were ranked from the highest to lowest, and the top 10 genes of each component were selected. These genes are listed in Table 4.1.

Table 4.1. The top 20 PC1 and PC2 genes in Proximal and Distal samples

PROXIMAL	DISTAL
SST	IGF1
SCG3	SST
VIP	SCG2
SCG2	SCG3
IGF1	CHGB
CHGA	VIP
CHGB	ADCYAP1
NPY	PENK
PYY	NPY
ADCYAP1	ADIPOQ
NPS	NPS
AVP	NMS
NMS	CRH
GIP	IAPP
NPPB	GALP
CRH	GIP
VGF	NPPB
NMB	RLN3
GHRH	PDYN
NPW	PYY

To be able to explain a more general pattern of the involvement of neuropeptides in the progression of colon cancer, the top 20 gene lists of distal and proximal were compared via the online Venn diagram tool, and the common genes were used for further analysis. There were 14 common genes between distal and proximal's gene lists. (Figure 4.2)

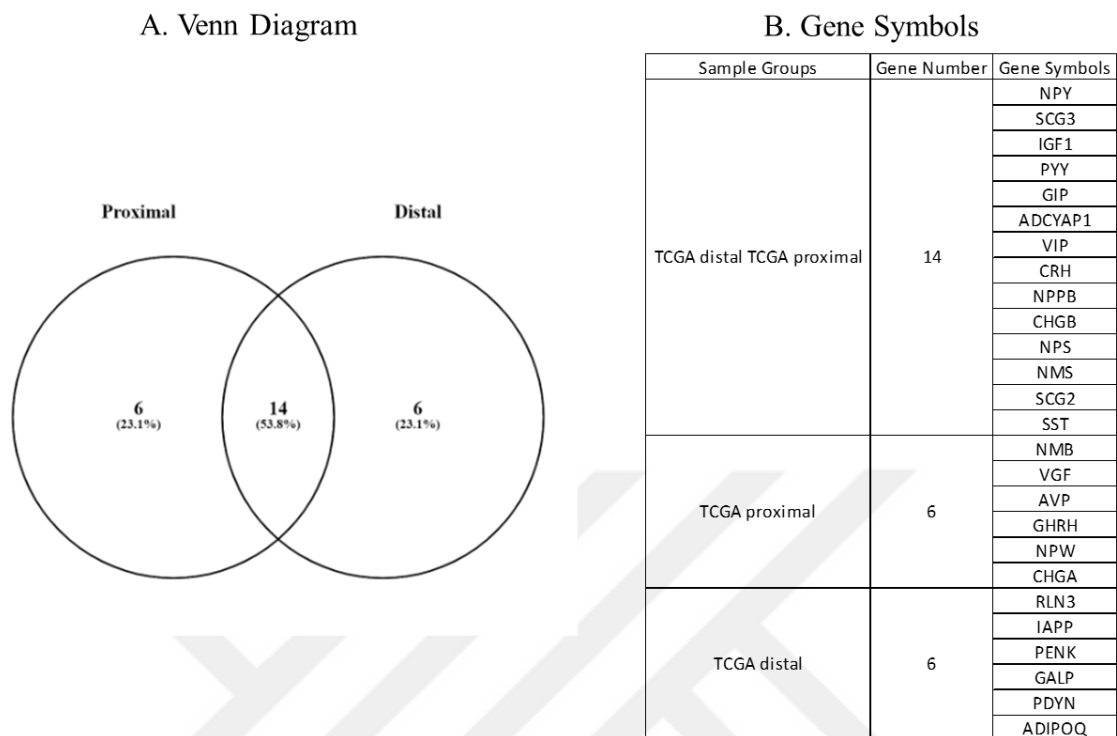


Figure 4.2. The common genes between Distal and Proximal samples. A. The Venn diagram shows the number of proximal and distal's top 20 PC1 and PC2 genes, and B. is the list of the gene symbols of these common genes.

The expression profile of the 14 genes showed that several of the genes had 0 expression values in the majority of the sample. Accordingly, the number of samples with an expression value of higher than 0 was calculated for all 14 genes, and among them, the PC2 genes had to be eliminated as they had 0 expressions in the vast majority of the samples. (Table 4.2) Consequently, the analysis was continued with the remaining PC1 genes. (Table 4.3)

Table 4.2. The number of samples with expression values higher than 0. The Sample # in the table represents the number of samples with an expression value of higher than 0 over the total number of samples present in each sample group. (Proximal: 190 total samples, Distal: 141 samples)

PROXIMAL		DISTAL	
GENE	SAMPLE #/190	GENE	SAMPLE #/141
NPY	106	NPY	92
SCG3	166	SCG3	120
IGF1	185	IGF1	136
PYY	167	PYY	131
GIP	22	GIP	17
ADCYAP1	178	ADCYAP1	131
VIP	188	VIP	139
CRH	10	CRH	8
NPPB	27	NPPB	22
CHGB	184	CHGB	135
NPS	1	NPS	1
NMS	3	NMS	5
SCG2	190	SCG2	139
SST	139	SST	105

■ PC1 genes
■ PC2 genes

Table 4.3. The list of final 9 neuropeptide genes

Gene Symbol	Gene Name
NPY	Neuropeptide Y
SCG2	Secretogranin III
SCG3	Secretogranin II
IGF1	Insulin Like Growth Factor I
PYY	Peptide YY
ADCYAP1	Adenylate Cyclase Activating Polypeptide 1
VIP	Vasoactive Intestinal Peptide
CHGB	Chromogranin B
SST	Somatostatin

4.1.3. Stratification of samples into Up and Down-regulated groups

In order to confirm the presence of the two clusters in PCA results, we generated heatmaps by performing hierarchical clustering. The results of hierarchical clustering also showed two distinct groups of samples based on the expression profile of the 9 PC1 genes. The two clusters of the samples were divided into Up and Down-regulated expressed groups. (Figure 4.3)

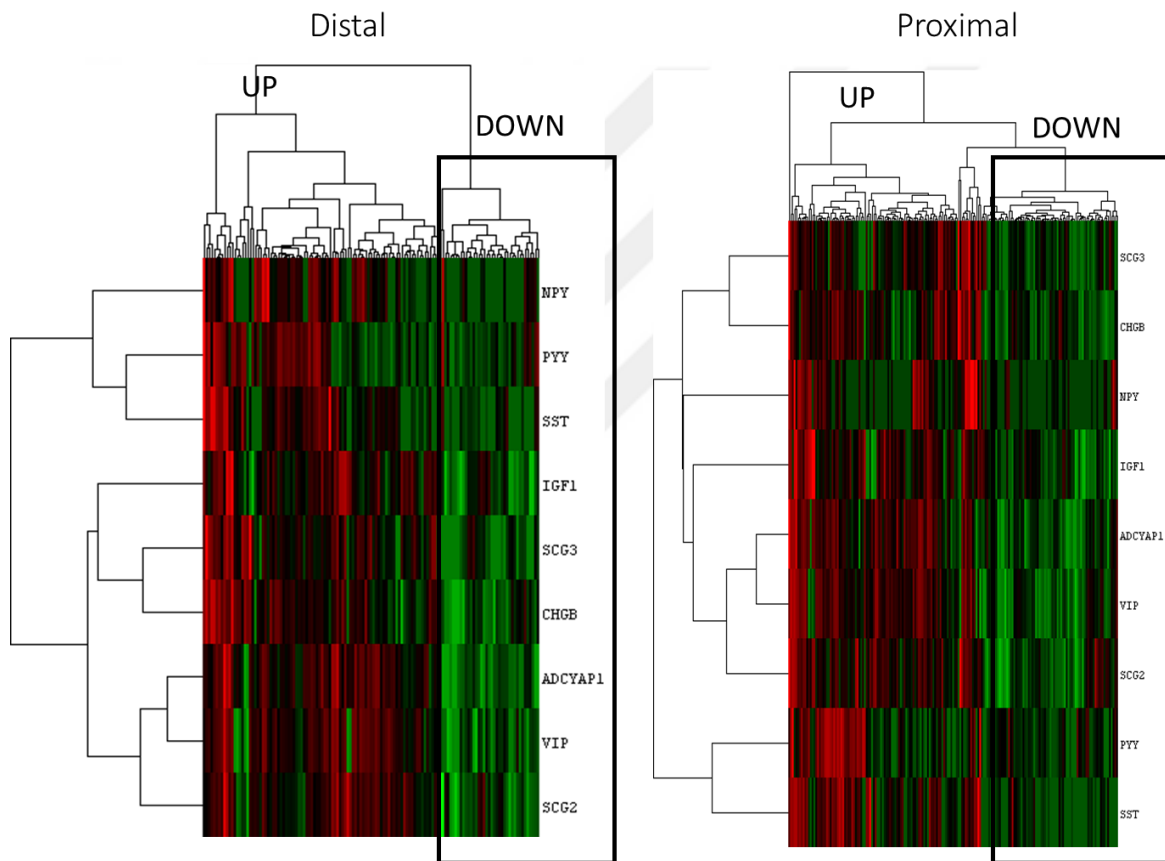
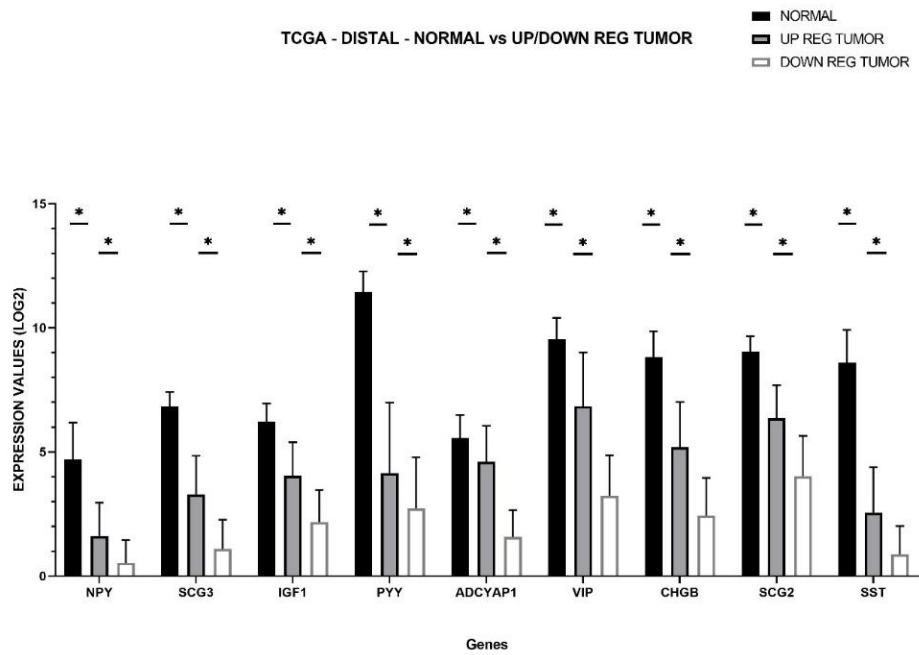
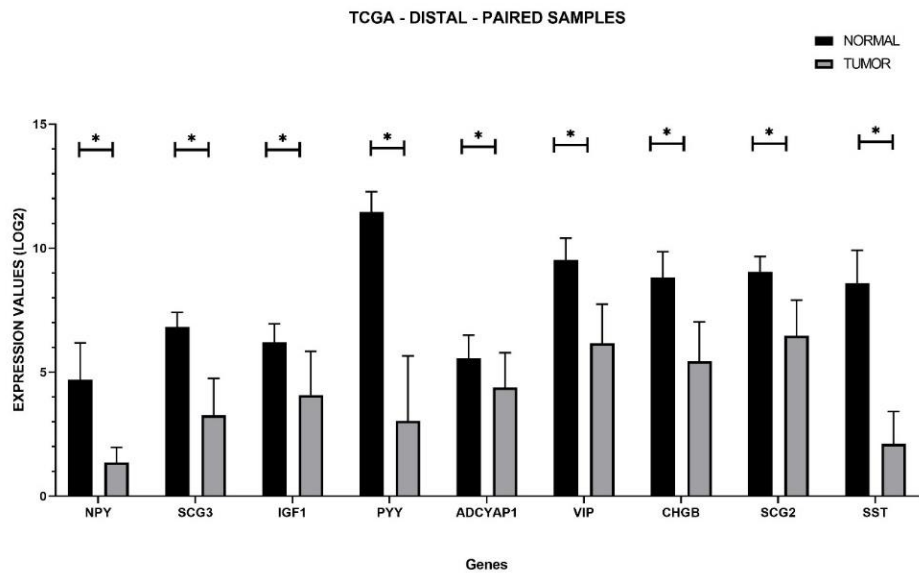


Figure 4.3. Heatmaps of Distal and Proximal samples based on the 9 PC1 genes. The Hierarchical clusters are plotted with Euclidean distance and complete linkage. The samples (x-axis) are divided into two expressional distinct groups in both Distal and Proximal samples based on the expression of the 9 neuropeptide genes (y-axis). The Up and Down-regulated samples are labeled on the figures.

4.1.4. Differential expression between Tumor and Normal samples groups

Next, we wanted to assess whether the expression difference observed in the hierarchical clusters was statistically significant. Additionally, since neuropeptide genes are involved in the regulation of various systems and mechanisms in the body under normal conditions, we were interested to see how these genes' expression compares in the tumor versus normal samples. Thus, we used the tumor-adjacent normal samples available in TCGA, plotted the expression bar plots of the 9 genes in normal and tumor sample groups, and performed paired student t-tests for each gene. The results showed these genes are significantly differentially expressed between normal and tumor and also between up and down-regulated samples. Interestingly, the expression of these neuropeptides was lower in the tumor compared to the normal samples. (Figure 4.4)

A. Distal



B. Proximal

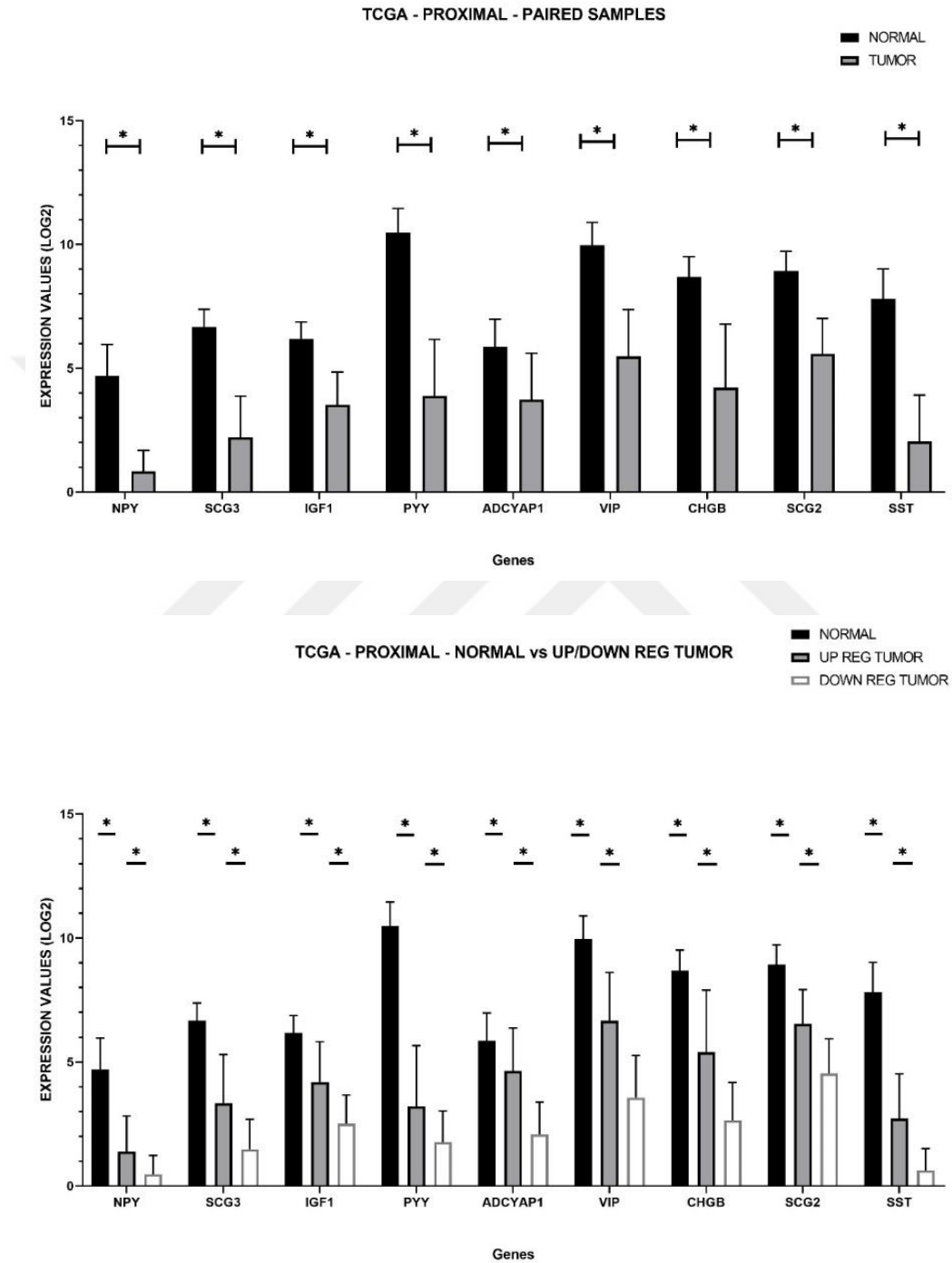


Figure 4.4. Expression Profiles of the 9 genes in Normal versus tumor and up-regulated/down-regulated samples; in A. Distal, B. Proximal. The bar plots on the left include the Normal, Up, and Down-regulated Tumor samples, while the bar plots on the right show the Normal and overall Tumor samples. (The asterisks indicate the significance of the t-test at $p\text{-value} \leq 0.05$)

4.1.5. Immunohistochemistry (IHC) staining

To confirm the expression pattern observed for Normal and Tumor samples, we utilized the immunohistochemistry staining results from Human Protein Atlas. The ‘Tissue’ (normal colon) samples were used for the assessment of the normal samples, and ‘Pathology’ (Colon adenocarcinoma) samples for the tumors. Additionally, for consistency of the staining quality and conditions, we ensured we compared the staining results of the same antibody clone for both normal and tumor samples. These IHC staining results also confirmed the higher expression of neuropeptides in normal colon tissues when compared to colon adenocarcinoma tissues, as we observed staining in the normal but not tumor samples. As examples, VIP and ADCYAP1 results are depicted in Figures 4.5 and 4.6, and the remaining neuropeptides’ results are presented in Appendix B. (The dark brown spots indicate the expression of the specified gene since the antibodies in Human Protein Atlas are labeled with DAB, which results in dark brown staining when the antibody binds to its antigen.)

Antibody Colony: HPA017324

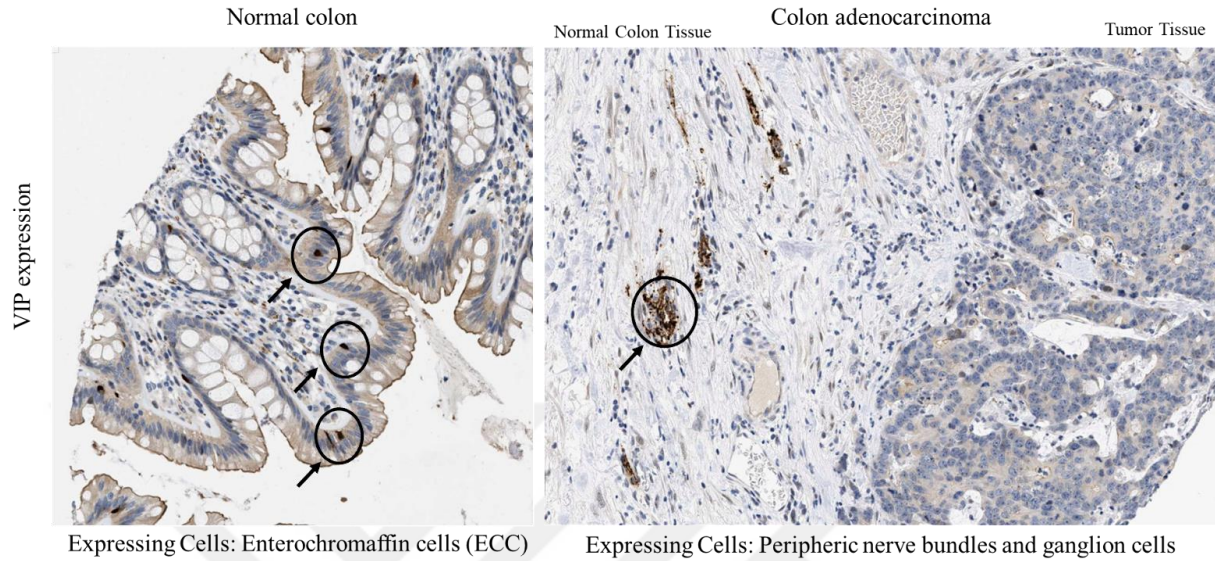


Figure 4.5. Vasoactive Intestinal Peptide (VIP) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of VIP in the cell type mentioned. As depicted in the figures, VIP is expressed in the normal tissues' ECC cells, peripheral nerve bundles, and ganglion cells, while no VIP expression is observed in the tumor tissue.

Antibody Colony: HPA065887

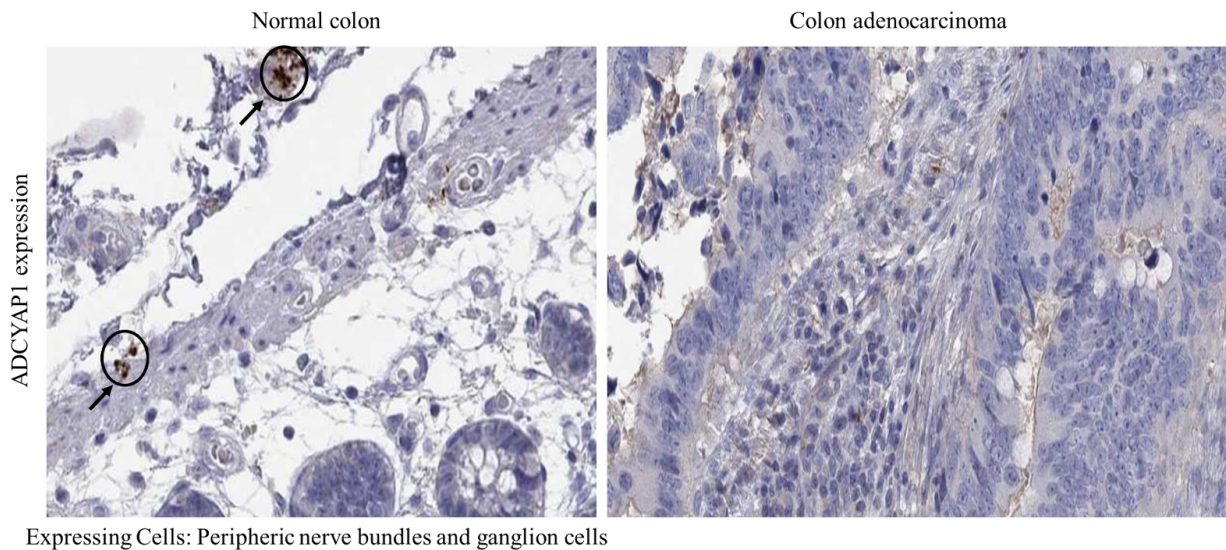


Figure 4.6. Pituitary Adenyl Cyclase Activated Peptide 1 (ADCYAP1) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of ADCYAP1 in the cell type mentioned. As depicted in the figures, ADCYAP1 is expressed in the normal tissues' peripheral nerve bundles and ganglion cells, while no ADCYAP1 expression is observed in the tumor tissue.

4.1.6. Clinicopathological characteristics distributions in the Up and Down-Regulated groups

Upon confirming a significant expression difference between the Up and Down-regulated tumor samples, we wanted to determine whether there was a significant differential distribution of the clinicopathological characteristics among these two groups that could explain the expression patterns observed. Thus, we performed Chi-square tests with the available clinical data for TCGA data. These clinical data included; stage, Kras/Braf (KB) mutation, Microsatellite instability (MSI), and gender, in addition to CMS subgroups. We observed a significant differential distribution of the CMS4 subgroup, with the majority of this subtype in the Up-regulated group in both distal and proximal samples. Moreover, there was a differential distribution of the male individuals, with the majority being in the Down-regulated groups. However, this was only observed for distal samples, and the other clinical data (MSS, Kras/Braf mutation, and stage) did not show any significant results. (Figures 4.7, 4.8)

A. CMS Subtypes

Distal						
Results						
	high	low				Row Totals
CMS1	5 (4.27) [0.12]	1 (1.73) [0.31]				6
CMS2	41 (46.26) [0.00]	24 (18.74) [1.48]				65
CMS3	6 (7.12) [0.18]	4 (2.88) [0.43]				10
CMS4	27 (21.35) [1.49]	3 (6.65) [3.69]				30
Column Totals	79	32				111 (Grand Total)

The chi-square statistic is 8.3. The p -value is .040202. The result is significant at $p < .05$.

Proximal						
Results						
	high	low				Row Totals
CMS1	22 (22.09) [0.00]	15 (14.91) [0.00]				37
CMS2	22 (23.88) [0.15]	18 (16.12) [0.22]				40
CMS3	9 (14.92) [2.35]	16 (10.08) [3.48]				25
CMS4	24 (16.12) [3.86]	3 (10.88) [5.71]				27
Column Totals	77	52				129 (Grand Total)

The chi-square statistic is 15.7648. The p -value is .001267. The result is significant at $p < .05$.

B. Gender

Distal

Results						
	High	Low				Row Totals
Female	48 (28.88) [12.67]	18 (37.12) [9.85]				66
Male	36 (55.12) [5.64]	90 (70.88) [5.16]				126
Column Totals	84	108				192 (Grand Total)

The chi-square statistic is 34.3154. The p -value is $< .00001$. The result is significant at $p < .05$.

Proximal

Results						
	High	Low				Row Totals
Female	49 (48.22) [0.17]	27 (29.78) [0.26]				76
Male	100 (102.78) [0.08]	99 (96.22) [0.12]				199
Column Totals	149	96				245 (Grand Total)

The chi-square statistic is 0.6184. The p -value is .431625. The result is *not* significant at $p < .05$.

Figure 4.7. Chi-Square results depicting the significantly distributed clinical parameters among the Up and Down-regulated samples; A. CMS subtypes distribution indicates CMS4 subtypes have a significant differential distribution, with mostly being observed in up-regulated vs. down-regulated groups in both distal and proximal samples. B. The genders are significantly differentially distributed in the distal samples but not in the proximal ones. (P-values significance at ≤ 0.05)

A. Microsatellite Instability

Distal

Results						
	HIGH	LOW				Row Totals
MSI	5 (5.90) [0.14]	3 (2.10) [0.39]				8
MSS	85 (84.10) [0.01]	29 (29.90) [0.03]				114
Column Totals	90	32				122 (Grand Total)

The chi-square statistic is 0.562. The p -value is .453444. The result is *not* significant at $p < .05$.

Proximal

Results						
	HIGH	LOW				Row Totals
MSI	23 (25.71) [0.29]	21 (18.29) [0.40]				44
MSS	74 (71.29) [0.10]	48 (50.71) [0.14]				122
Column Totals	97	69				166 (Grand Total)

The chi-square statistic is 0.9356. The p -value is .333407. The result is *not* significant at $p < .05$.

B. Kras/Braf Mutation

Distal

	Results					
	HIGH	LOW				Row Totals
KB MUTANT	17 (18.87) [0.00]	6 (6.13) [0.00]				23
KB WILD TYPE	38 (38.13) [0.00]	14 (13.87) [0.00]				52
Column Totals	55	20				75 (Grand Total)

The chi-square statistic is 0.0057. The *p*-value is .939814. The result is *not* significant at $p < .05$.

Proximal

	Results					
	HIGH	LOW				Row Totals
KB MUTANT	30 (30.68) [0.02]	22 (21.32) [0.02]				52
KB WILD TYPE	29 (28.32) [0.02]	19 (19.68) [0.02]				48
Column Totals	59	41				100 (Grand Total)

The chi-square statistic is 0.0766. The *p*-value is .781981. The result is *not* significant at $p < .05$.

C. Stage

Distal

	Results					
	HIGH	LOW				Row Totals
STAGE 1	10 (13.48) [0.00]	9 (5.52) [2.20]				19
STAGE 2	30 (28.39) [0.00]	10 (11.61) [0.22]				40
STAGE 3	31 (30.52) [0.01]	12 (12.48) [0.02]				43
STAGE 4	17 (15.61) [0.12]	5 (6.39) [0.30]				22
Column Totals	88	36				124 (Grand Total)

The chi-square statistic is 3.867. The *p*-value is .276185. The result is *not* significant at $p < .05$.

Proximal

	Results					
	HIGH	LOW				Row Totals
STAGE 1	2 (2.59) [0.14]	2 (1.41) [0.25]				4
STAGE 2	21 (20.73) [0.00]	11 (11.27) [0.01]				32
STAGE 3	82 (83.59) [0.03]	47 (45.41) [0.08]				129
STAGE 4	22 (20.09) [0.18]	9 (10.91) [0.34]				31
Column Totals	127	69				196 (Grand Total)

The chi-square statistic is 0.9968. The *p*-value is .802037. The result is *not* significant at $p < .05$.

Figure 4.8. Chi-Square results depicting the non-significantly distributed clinical parameters among the Up and Down-regulated samples; A. Microsatellite instable samples, B. Kras/Braf mutant samples, and C. different staged samples do not have differential distribution in up and down-regulated sample groups. (P-values significance at ≤ 0.05)

4.1.7. Biological and functional characteristics of the Up and Down-Regulated groups

For understanding the mechanism of the expression profile pattern observed, we performed functional analysis through GSEA. We used the Up and Down-regulated sample lists as the input for the two different phenotypes of GSEA with Gene ontology gene sets. After running the analysis, we used FDR values smaller than 0.25 for our significance threshold, as suggested by GSEA guidelines. This resulted in 7190 gene sets in the up-regulated samples, with 273 gene sets in the down-regulated samples for distal and 6405 gene sets in the up-regulated samples, with 0 gene sets in the down-regulated samples for proximal. (Appendix D)

Interestingly, we noticed a large amount of gene sets belonging to abnormal morphology pathways. Therefore, we decided to examine whether there was an association between abnormal morphology and cancer progression and development by finding the core enriched genes of abnormal morphology and tumor associated gene sets. Core-enriched genes are the main genes contributing to the enrichment of a specific gene set in a sample subgroup. The leading edge analysis (LEA) performed by GSEA uses the core enriched genes of the specified gene sets. Therefore, we performed LEA with these gene sets and extracted their core enriched genes. For the abnormal morphology gene-sets, we only extracted the genes which were present in at least 10 of the gene-sets. Complete lists of the number of selected gene sets, FDR values, and their core enriched gene numbers are available in Appendices E-J. Additionally, the bar plots of the core-enriched genes and the number of gene sets, including these genes, are shown in Appendix K.

The core enriched genes between abnormal morphology and tumor-associated pathways were compared to find the common genes between them. Since all the analyses were performed in distal and proximal samples separately, we first found the common genes for each gene set in distal and proximal (Figure 4.9) and then found their common gene lists (Figure 4.10). This resulted in 77

common genes between abnormal morphology and tumor-associated pathways in both distal and proximal samples. (Appendix L)

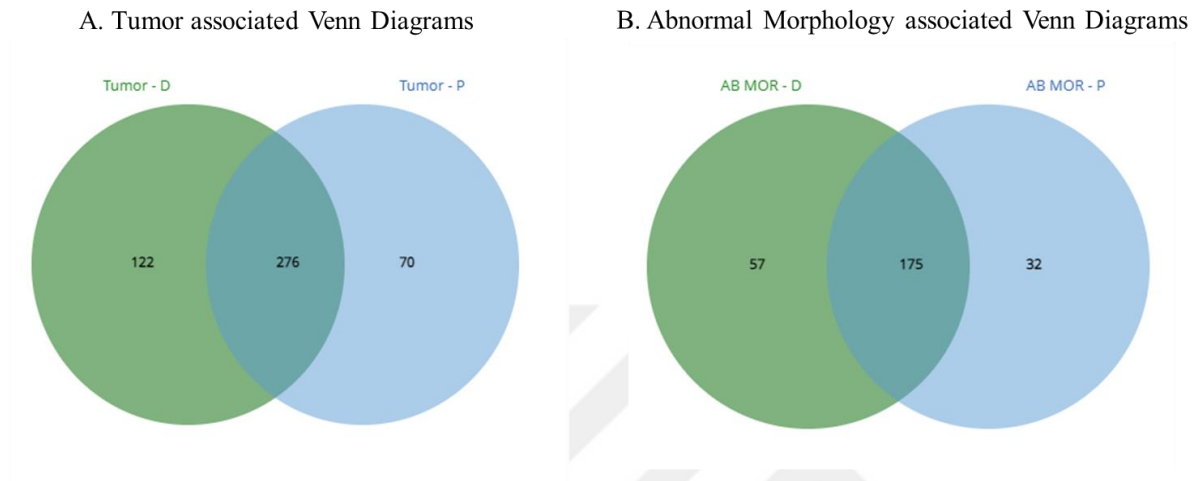


Figure 4.9. The common core enriched genes in; A. Tumor-associated gene sets, and B. Abnormal morphology-associated gene sets between Distal and Proximal samples. (AB MOR: Abnormal morphology, D: Distal, P: Proximal)

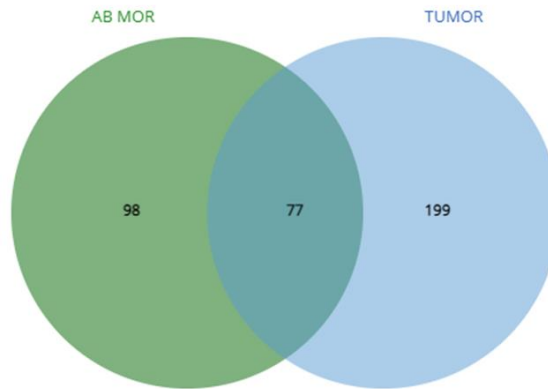


Figure 4.10. The Venn diagram of the common genes between the common abnormal morphology associated and tumor-associated genes sets' core enriched genes. There were 77 common enriched gene-sets between the two. (AB MOR: Abnormal morphology)

Following the determination of the 77 gene list, we performed a functional analysis in which networks are generated based on the pathways enriched in the input gene list via the EnrichNet website. These networks were later visualized in Cytoscape, and the networks with at least 3 nodes

were selected. These selected pathway networks consist of; hedgehog signaling pathways, immunosuppression, and digestive and reproductive system development, as depicted in Figures 4.11–4.13, respectively. (The lists of the genes involved in each network are available in Appendix M)

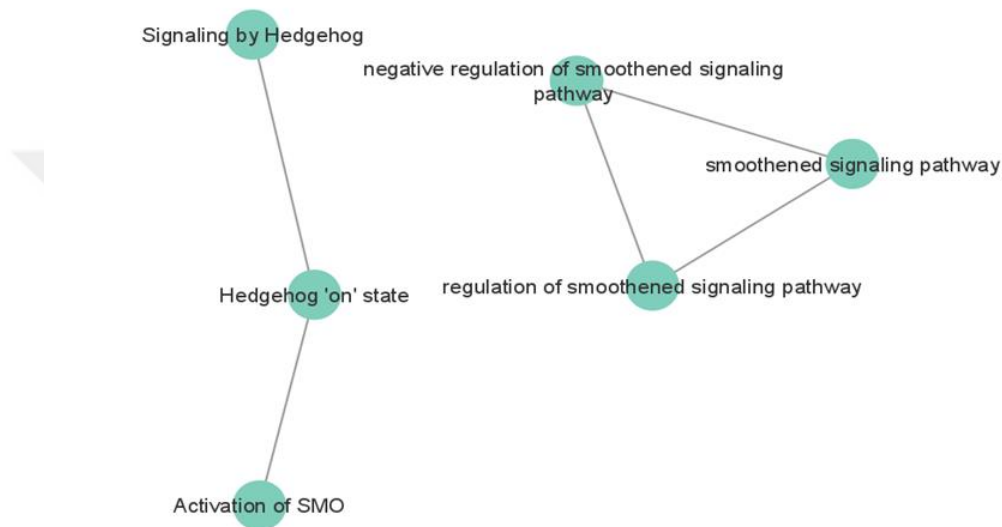


Figure 4.11. HHG signaling pathway networks. The nodes represent the enriched pathways

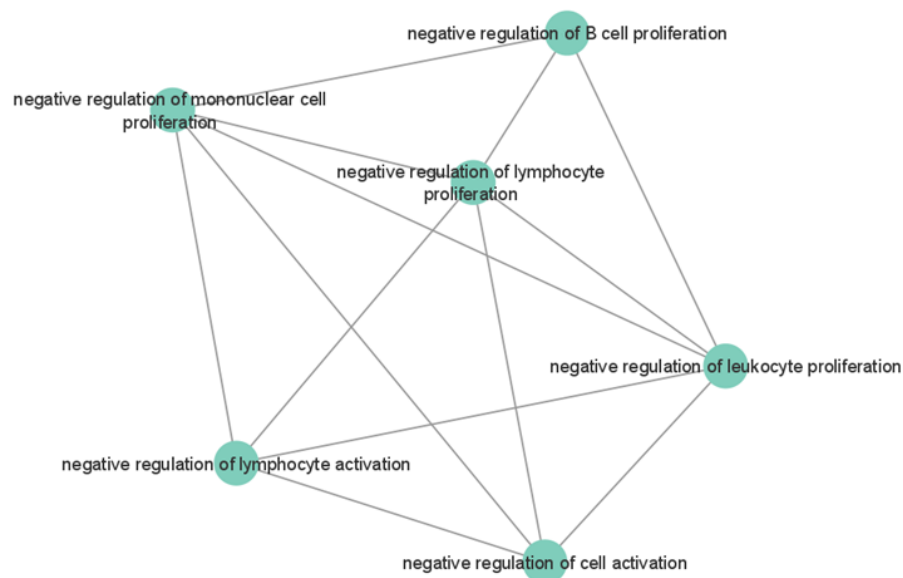


Figure 4.12. Immunosuppression-associated pathways network. The nodes represent the enriched pathways.

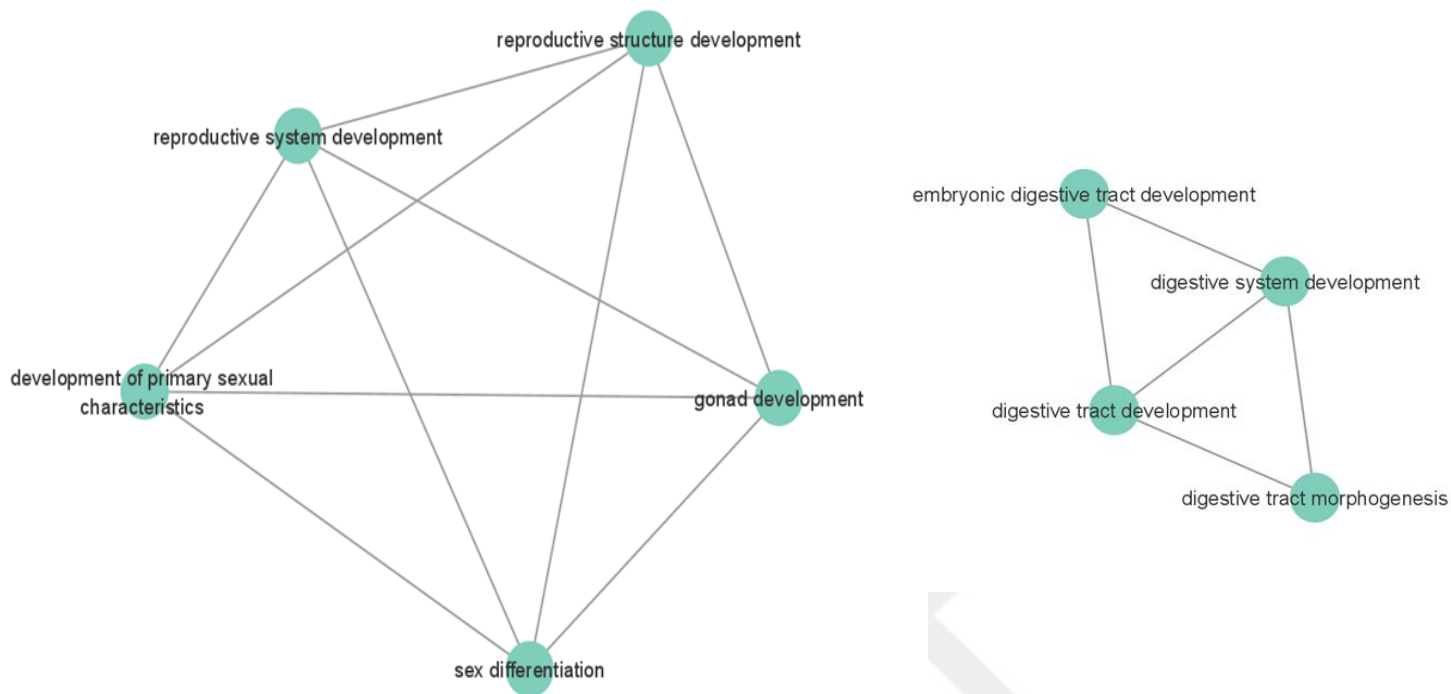


Figure 4.13. Digestive and reproductive system development associated pathways network. The nodes represent the enriched pathways.

Furthermore, we assessed the most significant gene sets in the up-regulated samples to gain more insight into the mechanism of neuropeptides' association in the progression of CRC. Hence, we filtered the significant biological pathway (GOBP) gene sets with a more strict FDR value criterion ($FDR < 0.1$) and later ranked them based on their Normalized Enrichment Scores (NES) and selected the common genes between the top 100 gene sets in proximal and distal samples. We found 22 common gene sets, as depicted in Figure 4.14, and the names of the gene sets are presented in Table 4.4.

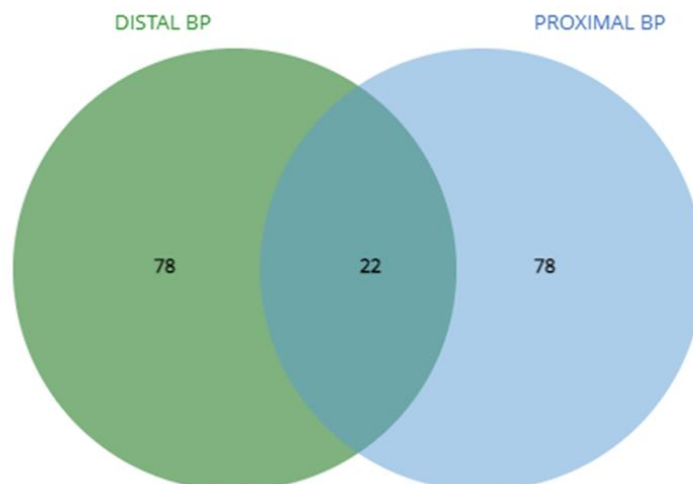


Figure 4.14. The Venn diagram of the common Gene Ontology (GO) Biological Pathways (BP) gene sets between distal and proximal. 22 gene sets were common between the two lists of the top 100 gene sets based on their Normalized Enrichment Scores (NES)

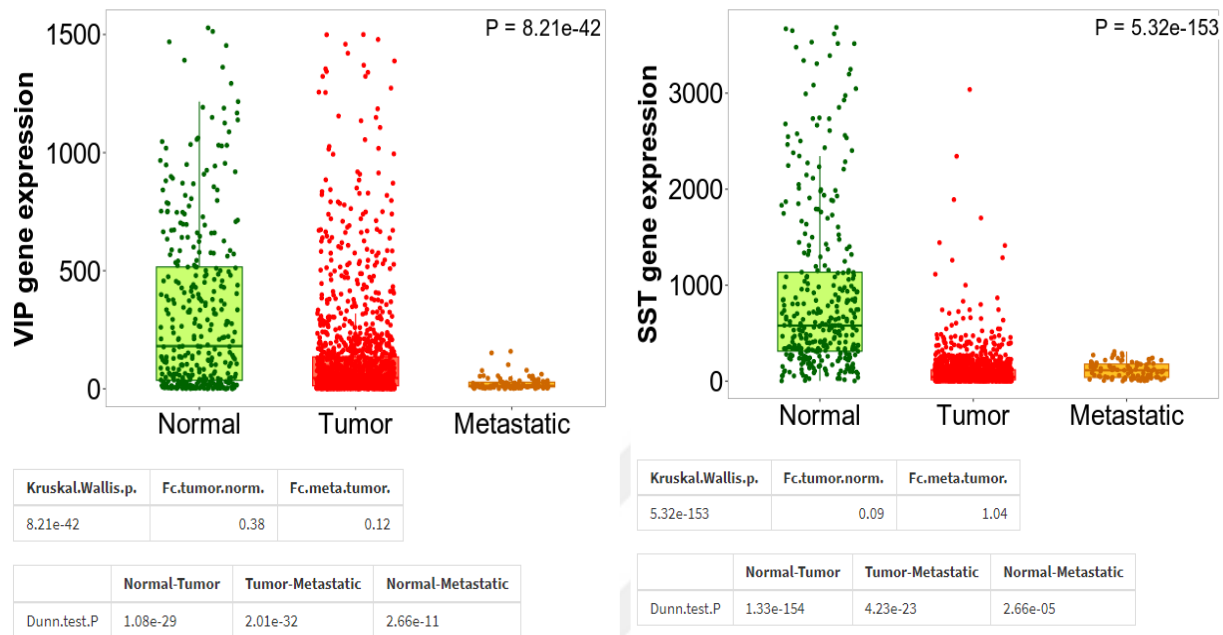
Table 4.4. List of the 22 common gene sets between Proximal and Distal from Figure 4.14

Common Gene sets in DISTAL vs PROXIMAL:
GOBP_POSITIVE_REGULATION_OF_PROTEIN_BINDING
GOBP_REGULATION_OF_CELL_MORPHOGENESIS
GOBP_REGULATION_OF_PROTEIN_LOCALIZATION_TO_NUCLEUS
GOBP_REGULATION_OF_CELL_SHAPE
GOBP_GLYCOPROTEIN_METABOLIC_PROCESS
GOBP_PROTEIN_LOCALIZATION_TO_CELL_SURFACE
GOBP_RHYTHMIC_PROCESS
GOBP_POSITIVE_REGULATION_OF_CELL_PROJECTION_ORGANIZATION
GOBP_ACTIN_FILAMENT_BUNDLE_ORGANIZATION
GOBP_CELL_SUBSTRATE_ADHESION
GOBP_SULFUR_COMPOUND_BIOSYNTHETIC_PROCESS
GOBP_REGULATION_OF_PROTEIN_IMPORT_INTO_NUCLEUS
GOBP_HOMEOSTASIS_OF_NUMBER_OF_CELLS
GOBP_REGULATION_OF_FILOPODIUM_ASSEMBLY
GOBP_REGULATION_OF_CELL_SUBSTRATE_ADHESION
GOBP_PEPTIDYL_TYROSINE_MODIFICATION
GOBP_POSITIVE_REGULATION_OF_CYTOSKELETON_ORGANIZATION
GOBP_RESPONSE_TO_TRANSFORMING_GROWTH_FACTOR_BETA
GOBP_EPITHELIAL_CELL_PROLIFERATION
GOBP_MYELOID_CELL_DIFFERENTIATION
GOBP_REGULATION_OF_ENDOCYTOSIS
GOBP_AMEBOIDAL_TYPE_CELL_MIGRATION

4.1.8. Expression profile in Tumor, Normal, and Metastatic groups

Besides our designated Up and Down-regulated samples, we were interested in the expression profile of these neuropeptides in different subgroups of tumors, such as the metastatic and non-metastatic samples. This would not only provide us with more insight into the mechanisms of the involvement of these genes in CRC progression, but also, they could explain the association between the increase in the expression of these genes, cancer progression, and poor prognosis. Therefore, through the TNM plot website, which includes normal samples from GTex (a healthy RNA-seq dataset), TCGA's tumor-adjacent normal tissues, in addition to TCGA's tumor samples. The overall pattern observed was that all of the neuropeptides had a significantly higher expression in normal/ healthy samples compared to tumor samples (both metastatic and non-metastatic), confirming our previous results. However, within tumor samples, the genes were divided into two groups in terms of their expressions in metastatic and non-metastatic samples. One group of genes showed a significantly lower expression in the metastatic group than in the non-metastatic group, while the other group had the opposite pattern. Some of the examples of each group are depicted in Figure 4.15. The remaining neuropeptide genes' plots can be found in Appendix N.

A. Lower expression in metastatic samples



B. Higher expression in metastatic samples

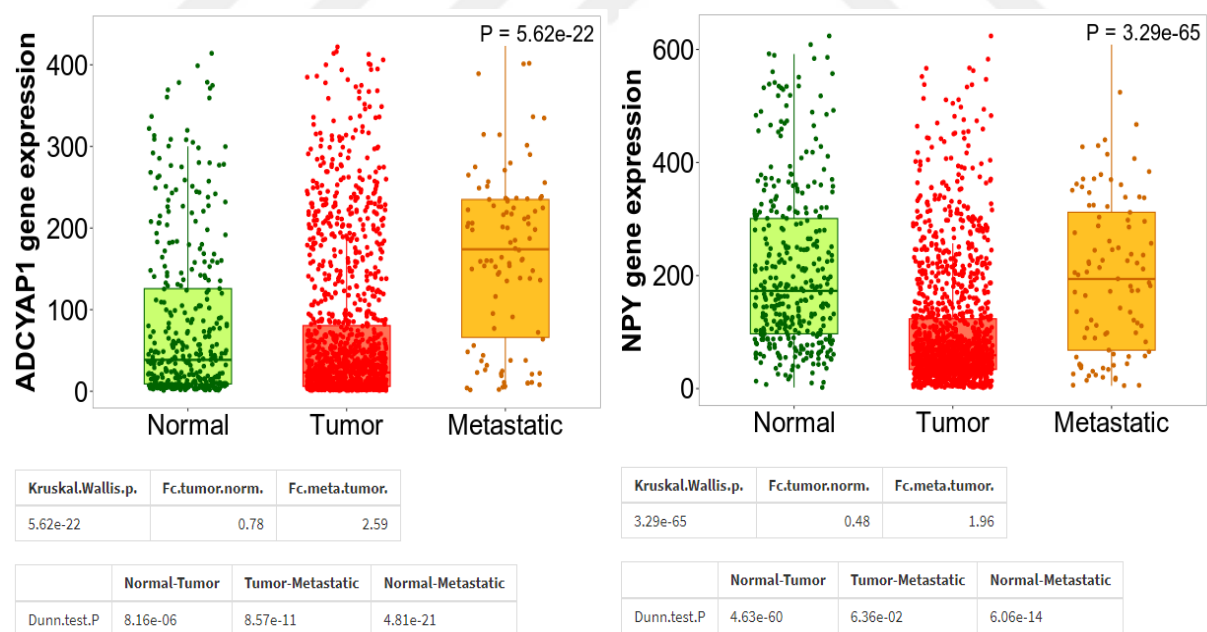
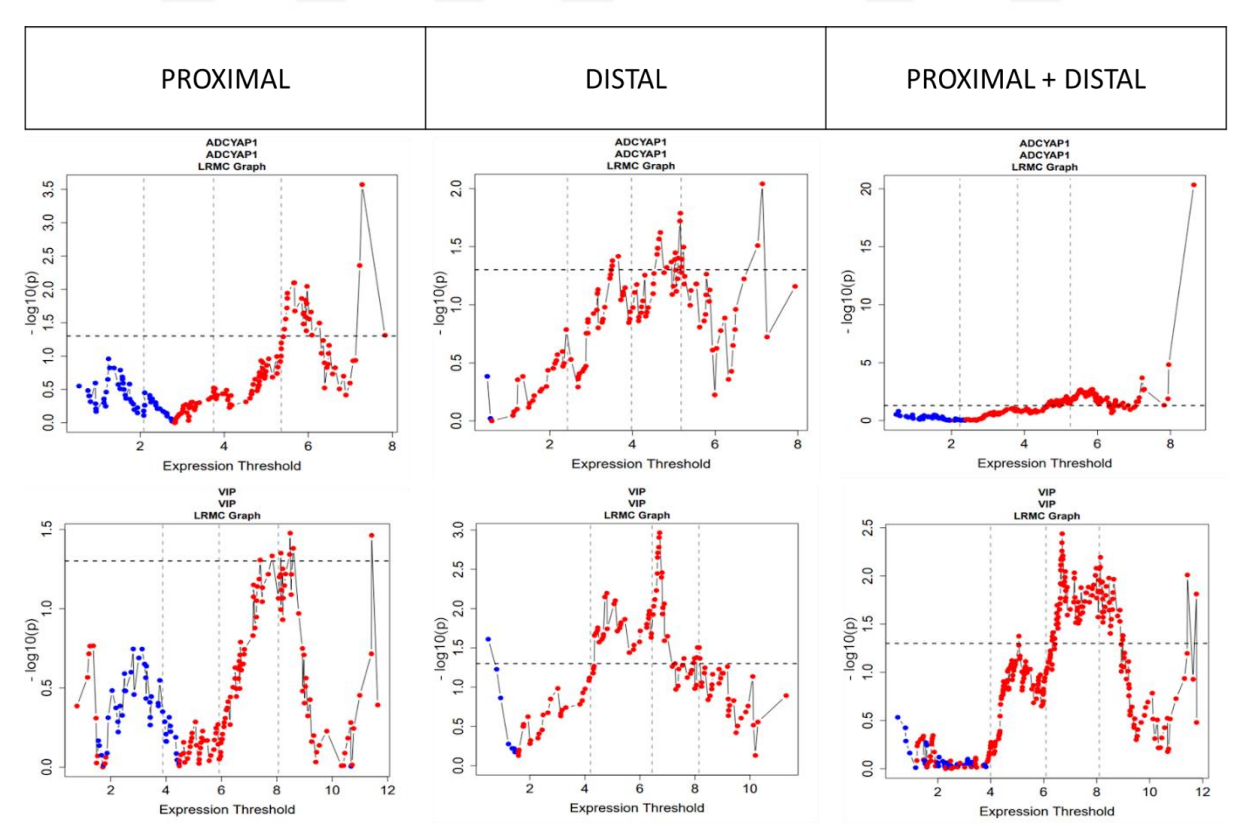


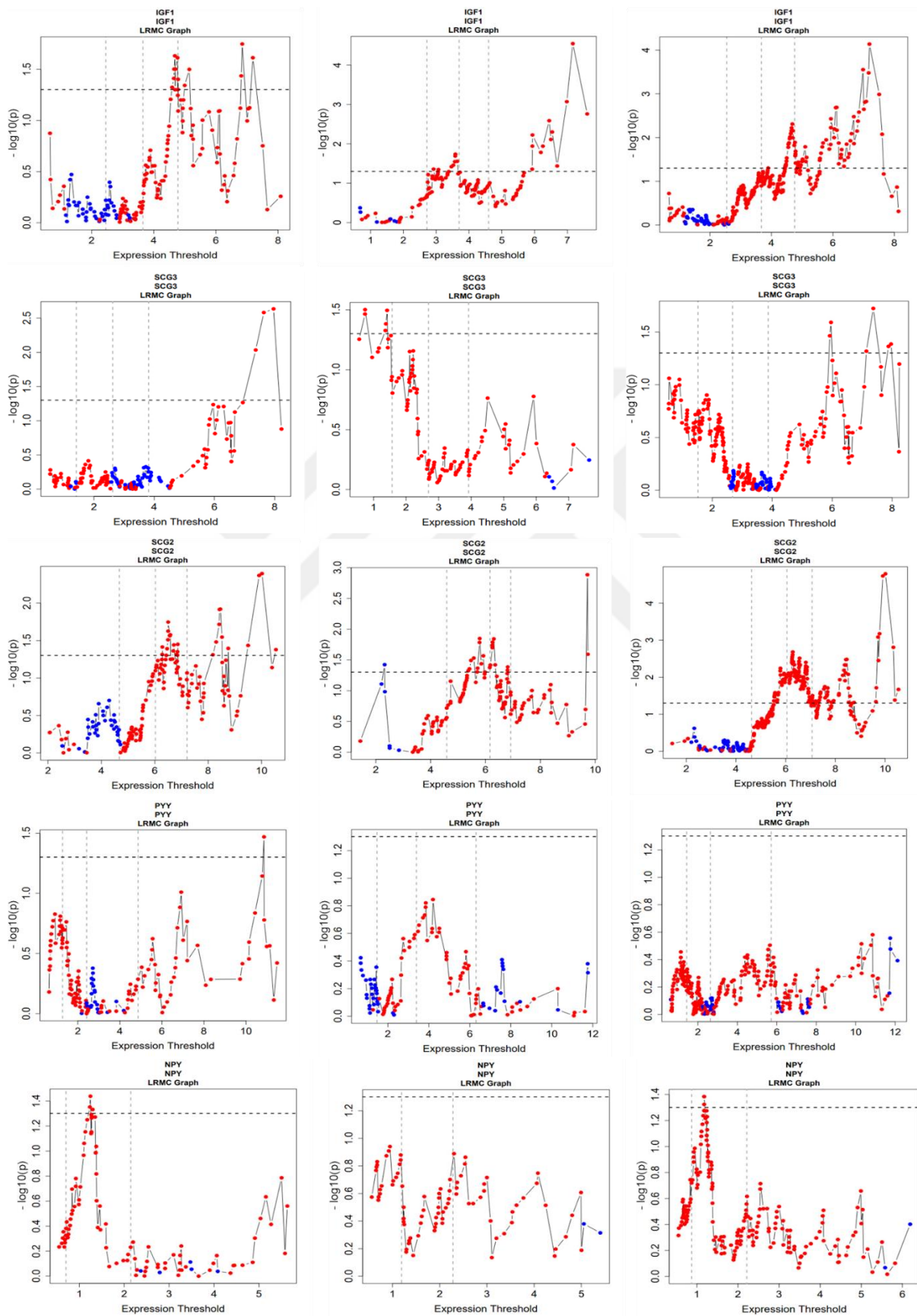
Figure 4.15. Expression profiles of the 4 genes in normal, non-metastatic tumor, and metastatic tumor samples. The expressions of these genes are significantly lower in the tumor samples compared to the normal samples. However, within tumors, their expression profile is divided into two groups; A. genes which have a significantly higher expression in the metastatic vs. non-metastatic samples, and B. genes with an opposite expression pattern. (Non-parametric

Kruskal–Wallis and Dunn’s test was performed among all three groups, and all the P values are significant at $P \leq 0.05$).

4.1.9. Prognostic associations of the 9 neuropeptide genes

Lastly, to determine the prognostic association of our neuropeptide genes, we plotted LRMG graphs in distal, proximal, and proximal combined with distal samples. The results of these graphs are presented in Figure 4.16. Out of the 9 neuropeptide genes, 5 genes (VIP, ADCYAP1, SCG2, SCG3, IGF-1) showed significant prognostic associations at several cut-offs in both proximal and distal samples. The remaining 4 genes either had no significant prognostic association or the association was only observed in one location. Interestingly, VIP, and its paralogue ADCYAP1, results indicated a good prognosis at lower expression levels of the gene in addition to poor prognosis association with its overexpression.





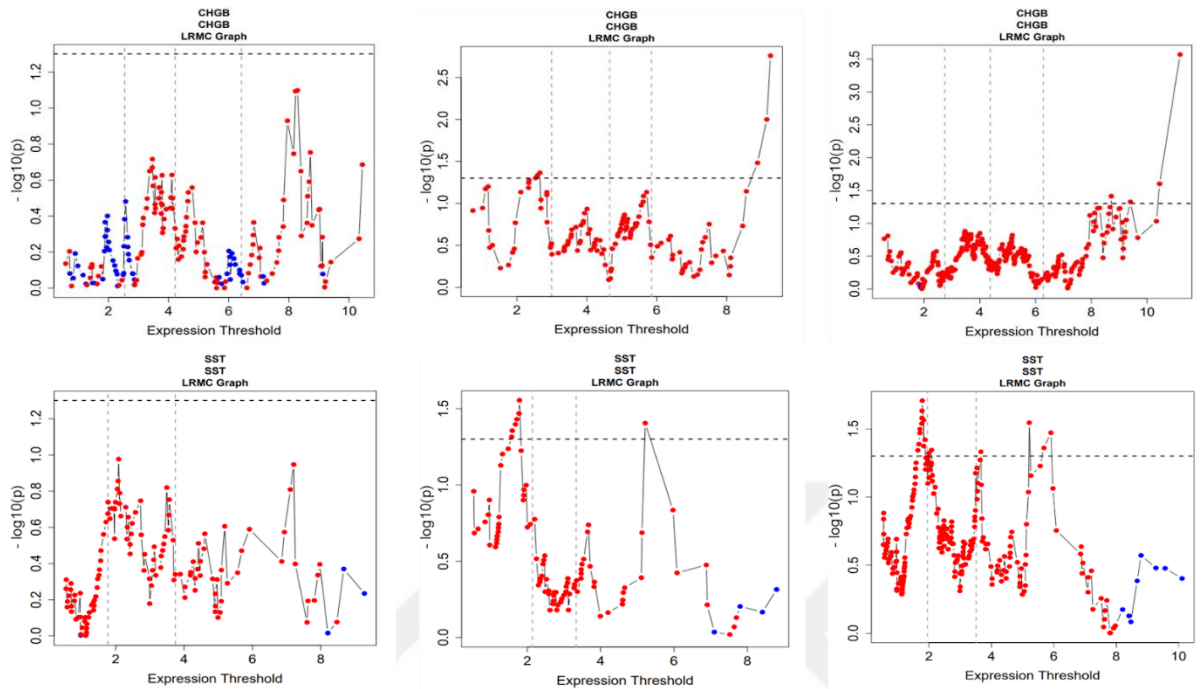


Figure 4.16. LRMG graphs of Proximal, Distal, and combination of Proximal + Distal colon cancer patients based on different neuropeptide genes' expressions in TCGA. The dots on the plots indicate all the cut-offs. The red color represents a bad prognosis (>1 HR value), and the blue color indicates a good prognosis (<1 HR value). The horizontal dotted line is the indication of log-rank test P value = 0.05, and the vertical lines represent the quantiles (1st - 3rd, left-right)

4.2. Approach 2: Identification of robust prognostic biomarker in Proximal, Stage 2 and 3 CRC patients

The aim of this part of the study was identification of robust prognostic biomarkers in a specific group of samples. To decrease the inconsistencies as a result of technological differences between different platforms. Similarly, two different cohorts from the same platform were utilized for the elimination of cohort-specific biases in the later analyses, as the molecular and clinical characteristics of each cohort can be very specific and drastically different. The details of these datasets can be found in Table 4.5. The three different survival data used were Overall Survival (the time from initial diagnosis to recorded death), Relapse-Free and Disease-Free Survival (the time of surgery to the first relapse/ recurrence).

As mentioned previously, the advantage of this study is the specification of a sample group to increase the power of the prognostic marker determined. Thus, the analyses in this study only included Proximal, Stage 2, and 3 samples (In the further steps, ‘All Samples’ refer to Proximal, Stage 2, and 3 samples). Table 4.5 only includes the number of samples used in this study.

Table 4.5. All the datasets used and their details

Datasets	Platform	Sample Number	Survival Data	Reference
GSE39582	HG-U133_Plus_2	190	* OS, ** RFS	34
TCGA	ILUMINA	107	OS	46, 47
GSE14333	HG-U133_Plus_3	85	*** DFS	48

* OS: Overall Survival

** RFS: Relapse-Free Survival

*** DFS: Disease-Free Survival

4.2.1. Gene selection

After the normalization of the datasets, we ran the Cox Regression Analysis for all three datasets. As the importance of considering all the probesets of a gene in microarray and the inconsistencies resulting from not doing so has been discussed previously in the introduction, we attempted to factor in these inconsistencies. Hence, the analysis for the microarray dataset was divided into genes with 1 probeset and genes with 2 or more probesets. There were 2 main criteria for the selection of the genes in TCGA and genes with 1 probeset in GSE39582;

- Cox P value ≤ 0.05)
- HR value of ≥ 1.5 or HR value of ≤ 0.67

We had similar criteria for genes with 2 or more probesets in GSE39582. For such genes, all the probesets of one gene had to pass the criteria, meaning we eliminated genes with a least one probeset not fulfilling the criteria. Moreover, we had an additional criterion for these genes.

- All the probesets of the gene had to have the same prognostic value, meaning all the probesets need to be associated with either favorable or not favorable prognosis.

Table 4.6 below further clarifies the criteria with some examples from selected and eliminated genes.

Table 4.6. Examples of the selected and eliminated genes in GSE39582

(For genes with 2 or more than 2 probesets)

Examples of Eliminated Genes					
Gene Name	Probset	HR	Cox P		Fullfilling Criteria?
A1BG	229819_at	1.470464	0.4142		NO
A2M	217757_at	1.336746	0.0305		NO
A2M	1558450_at	1.5412	0.3856		
MS4A3	210254_at	6.1576	0.0076		NO
MS4A3	1554892_a_at	0.06501	0.0081		
Examples of Selected Genes					
Gene Name	Probset	HR	Cox P		Fullfilling Criteria?
A2MP1	236394_at	3.1814	0.0301		YES
ACSM1	215432_at	3.1886	0.0288		YES
ACSM1	234197_at	3.3264	0.02439		
GRSF1	201501_s_at	0.3293	1.19E-05		YES
GRSF1	201520_s_at	0.44889	0.01358		
GRSF1	201530_at	0.4861	0.0002		

After the gene selection process was completed for TCGA and GSE39582, the gene lists between these two datasets were compared, and a final of 43 common genes was found. We performed the analysis for GSE39582 with overall survival to keep the survival data consistent across TCGA and GSE39582. Later, for validation, we utilized another microarray dataset (GSE14333) with Disease Free Survival data. The HR and Cox P values of the 43 common genes were extracted, and the genes that met the above-mentioned criteria were selected. The list of the 3 final genes and details in corresponding datasets is presented in Table 4.7. Among these 3 genes, there are 2 genes with 1 probeset and 1 gene with 2 probesets. Additionally, two genes (GAA and SEMA4C) are associated with a ‘not favorable’ prognosis, while 1 gene (OXNAD1) is associated with a

‘favorable’ prognosis. Figure 4.17 demonstrates the workflow and the number of the genes at each step.

Table 4.7. The list of the final 3 genes and their corresponding Cox P and HR values

		Datasets						
		GSE39582 OS		TCGA OS		GSE14333 DFS		
Gene Name	Probset Number	Cox P	* HR	Cox P	HR	Cox P	HR	** Prognosis
GAA	1	0.004302	2.020762	0.018369	1.932148	0.009173	4.768178	Not Favorable
OXNAD1	1	0.000788	0.457167	0.014347	0.245395	0.006885	0.141569	Favorable
SEMA4C_46665_at	2	0.013716	1.817876	0.01828	2.353941	0.01548	5.61277	Not Favorable
SEMA4C_219039_at		0.049656	1.952973			0.017102	8.87426	

* Cox proportional hazards regression

** Prognosis: HR>1, not favorable, HR<1, favorable

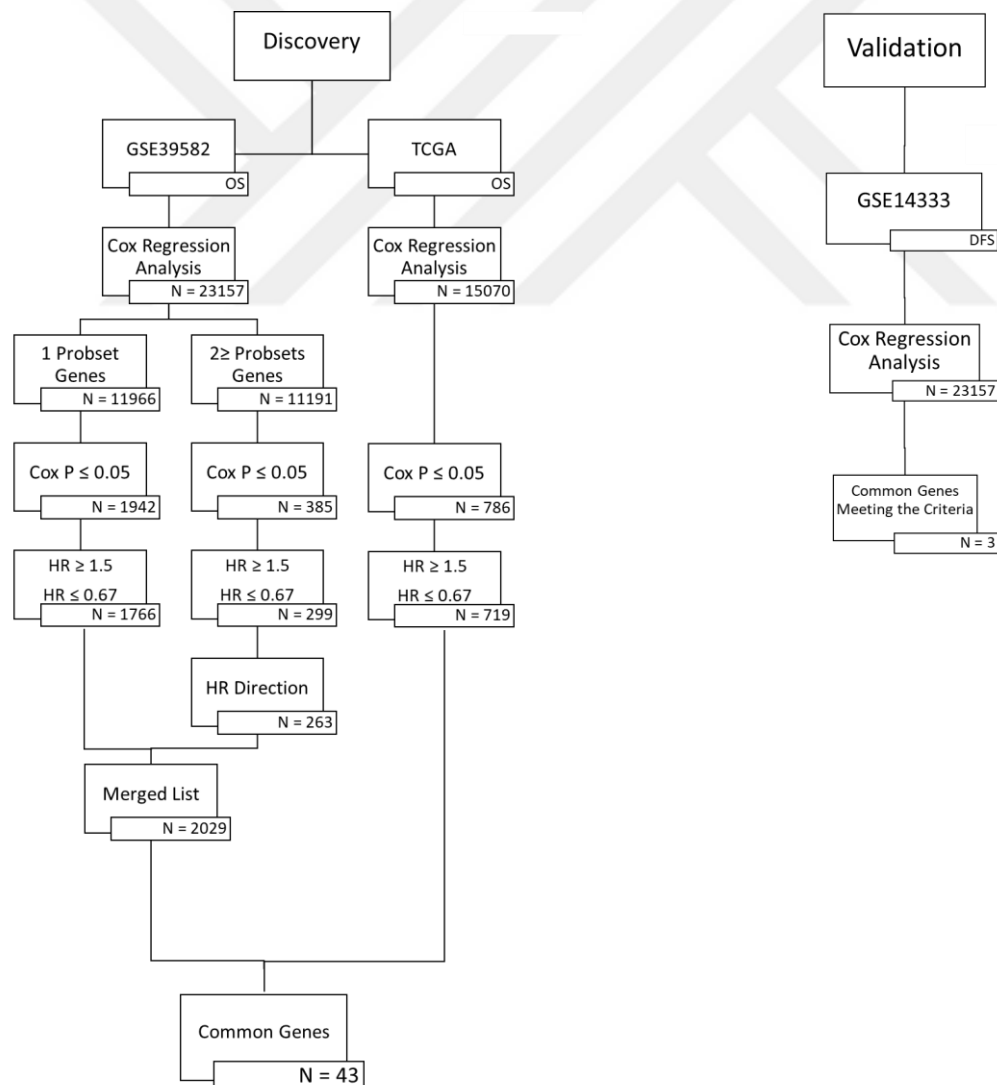


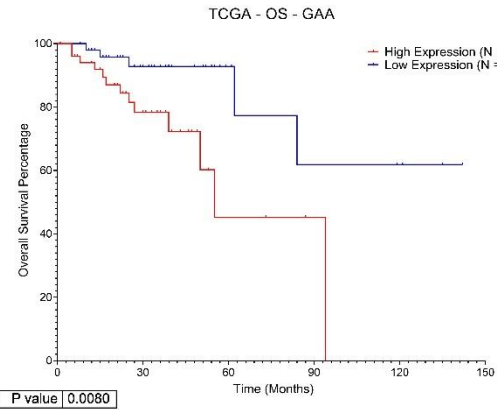
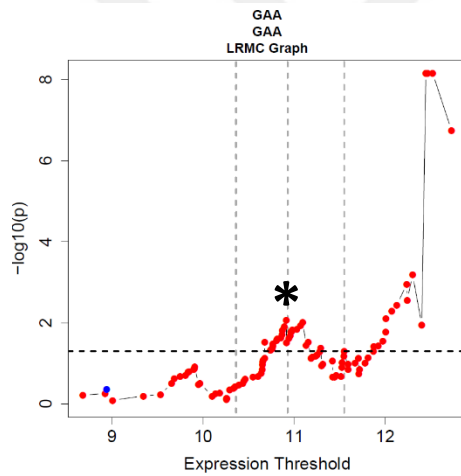
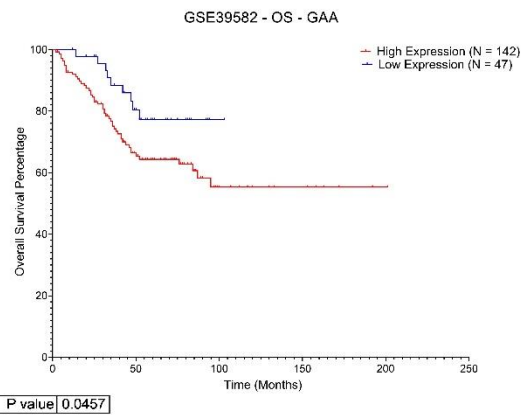
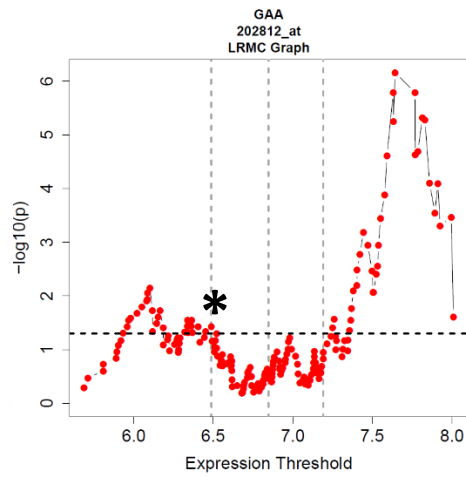
Figure 4.17. The workflow for the selection of the genes. (N; number of genes at each step)

4.2.2. Survival Association of the identified genes

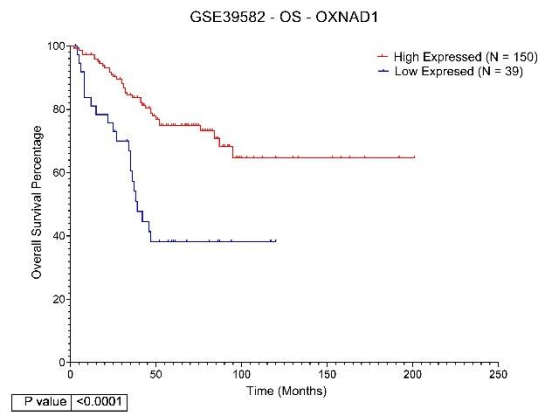
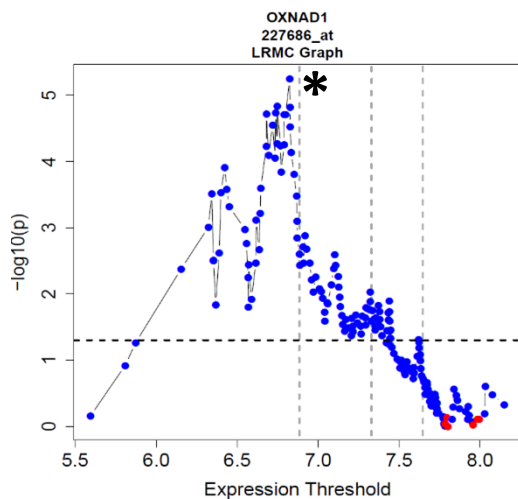
4.2.2.1. Prognostic associations in all samples (Proximal, Stage 2 and 3)

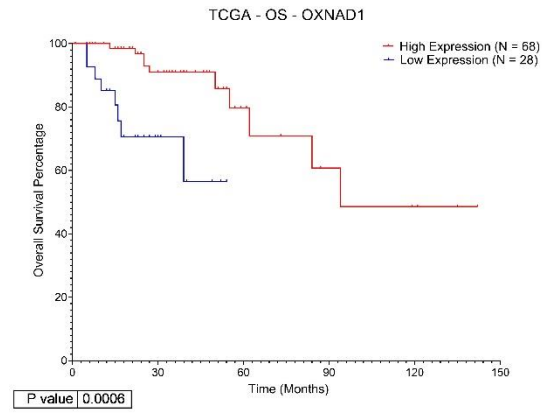
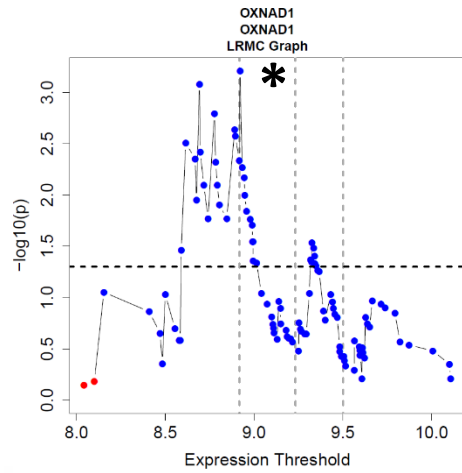
We performed further survival analysis for the assessment of the prognostic association robustness of our 3 genes. The survival analysis used was LRMC graphs which enabled us to test various cut-offs to help us define a threshold/ cut-off for the stratification of our samples (Proximal, stage 2 and 3) into high and low-expressed groups. The criteria for choosing the cut-off was the closest to the median with the smallest p-value. These LRMC graphs also verified a not favorable prognosis (shorter overall survival) association with the up-regulation of GAA and SEMA4C and a favorable prognosis (longer overall survival) with OXNAD1 up-regulation. We later plotted Kaplan Meier plots to confirm the samples divided at each defined cut-off for each gene were significantly associated with shorter and longer overall survivals. (Figure 4.18) The Relapse-Free Survival (RFS) and Disease-Free Survival (DFS) plots can be found in Appendix O.

A. GAA – Overall Survival

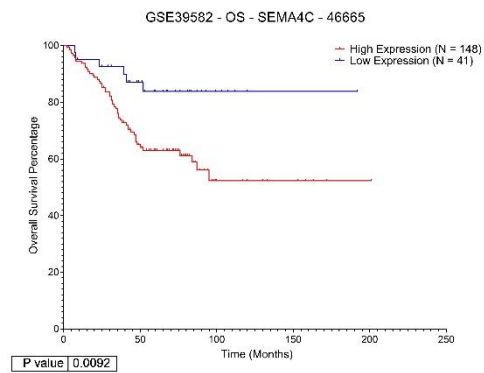
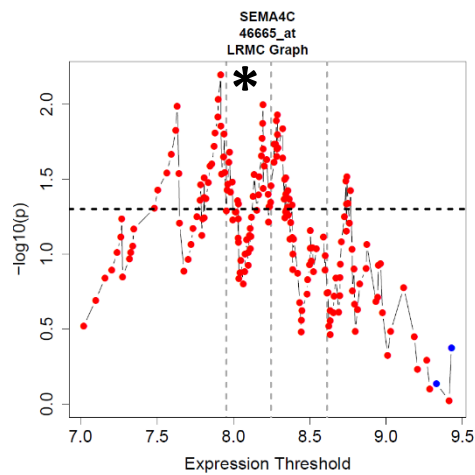
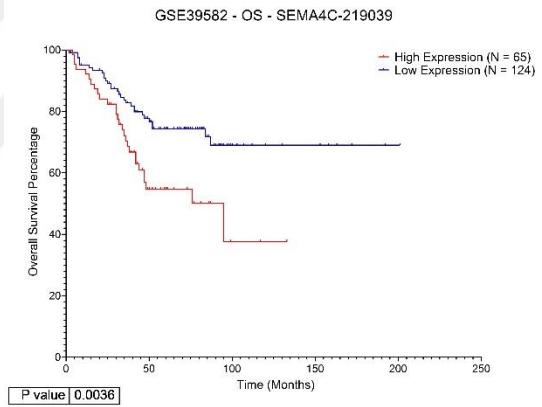
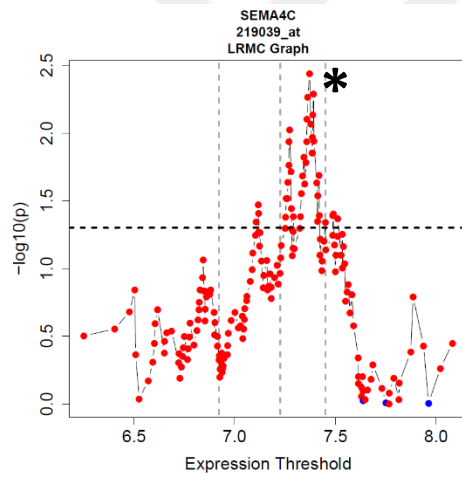


B. OXNAD1 – Overall Survival





C. SEMA4C – Overall Survival



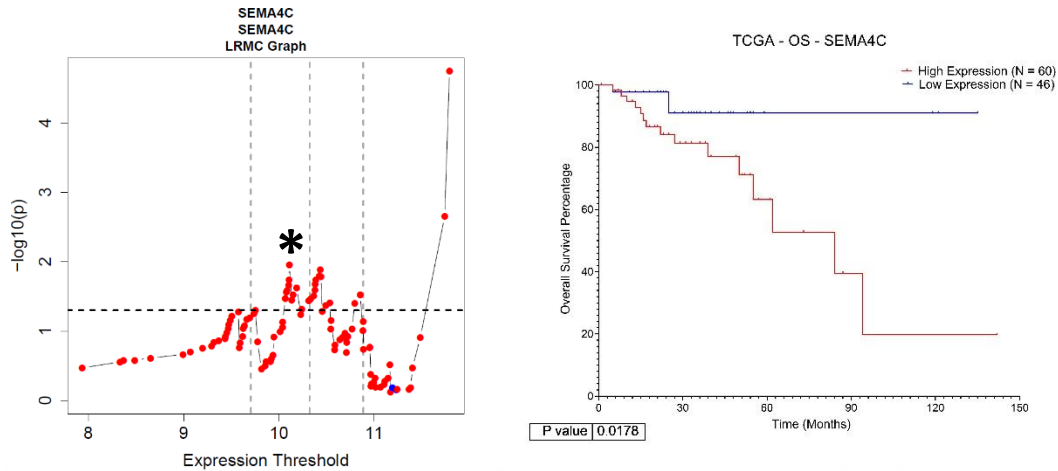


Figure 4.18. Overall Survival (OS) association graphs of Proximal, Stage 2 and 3 CRC samples stratified based on; A.GAA, B.OXNAD1, and C.SEMA4C expressions individually. The Kaplan Meier plots compare the overall survival of the high and low expression groups based on the cut-offs from the LPMC graph on the left. The asterisk indicates the chosen cut-off.

4.2.2.2. Prognostic associations in Molecular Subgroups

Our gene signature expressions showed significant association with prognosis within proximal and stage 2 and 3 samples. However, an issue in the translation of the gene expression-based biomarkers is a lack of a specified patient population; one of the suggestions for improving the strength of prognostic biomarkers is the combination of clinical and molecular characteristics. [35, 38] In order to improve our gene signatures, we used the available clinicopathological characteristics of the sample to stratify them into various subgroups and tested the prognosis of our genes in the newly stratified sample groups. We continued the analysis with GSE39582 and TCGA since GSE14333 did not have clinical data available. The clinical data used for the stratification consisted of; cancer stage, KRAS/BRAF mutation, and microsatellite stability. Additionally, we used the common CRC classification, such as; CMS and Oncotype DX risk groups. LPMC and KM plots were used for the assessment of any prognostic association of our

genes in each of the subgroups. The analysis was done for each of these groups separately and their combinations. Interestingly, the expressions of our gene signatures were not associated with the survival of all the subgroups. Moreover, we only found the Oncotype DX Low-Risk group to be common between GAA and OXNAD1. The subgroups with significant survival associations in both TCGA and GSE39582 OS were compared, and the common subgroups were found, as summarized in Table 4.8. (The LRMC and KM plots of the subgroups in Table 4.8. can be found in Appendix P)

Table 4.8. The common significantly associated subgroup for respective genes
(Common: between TCGA and GSE39582 OS)

Gene	Sample Group
GAA	Stage 2 + Kras/Braf Mutatnt
	CMS2 Subtype
	Oncotype DX Low Risk
OXNAD1	MSS + Kras/Braf Mutant
	Stage 3 + Kras/Braf Mutant
	Stage 3 + MSS
	Oncotype DX Low Risk
	Oncotype DX High Risk
SEMA4C	CMS4 Subtype

4.2.3. Univariate, Multivariate analysis of the Clinicopathological parameters

In the last step of the study, we performed univariate and multivariate Cox proportional-hazards regression analyses to assess the subgroups' clinical data association with overall and recurrence free survival data. The univariate analysis was performed with all the available clinical variables independently, and the clinical variable with a significant data was included in the multivariate analysis. Among all the available data, we found that stage and the expression of GAA, OXNAD1, and SEMA4C were linearly associated with overall and but not with relapse-free cancer survival. When the patients were stratified into groups with high and low-expression of the specified genes,

statistically significant ($p < 0.05$) relationships were noted with both overall and relapse free survival. (Table 4.9). Accordingly, only the linear and categorical gene expression values with stage were included as the confounding factors in the multivariate analysis to test whether the three genes were independent prognostic markers. The multivariate results demonstrated, when the gene expression values are considered as continuous variables, OXNAD1, was independent of stage as a confounding factor. However, the remaining two genes did not show consistent results regarding their stage independency in GSE39582 and TCGA datasets. Similarly, the multivariate results for all three genes were not consistent across the two datasets for the conclusion of stage independency of the genes. (Appendix Q)

Table 4.9. Univariate Cox P regression analysis of the clinicopathological data in; GSE39582 (OS), (RFS), and TCGA (OS)

Characteristics	GSE39582 OS		TCGA OS		GSE39582 RFS	
	Cox P	HR	Cox P	HR	Cox P	HR
* GAA	0.004	2.018	0.018	1.932	0.000	2.614
* OXNAD1	0.001	0.458	0.014	0.245	0.000	0.337
* SEMA4C_219039_at	0.05	1.95	0.018	2.354	0.077	1.909
* SEMA4C_46665_at	0.014	1.816			0.094	1.562
** GAA	0.05	2.033	0.0141	3.75	0.002	2.608
** OXNAD1	0.000	0.329	0.002	0.187	0.000	0.267
** SEMA4C_219039_at	0.005	2.115	0.029	4.01	0.011	2.061
** SEMA4C_46665_at	0.013	2.913			0.026	1.908
Stage	0.036	1.751	0.03	2.885	0.022	1.949
MSI	0.37	1.295	0.808	1.151	0.901	1.038
KRAS/BRAF	0.84	0.926	0.911	0.937	0.122	0.564

* Gene expression: continuous values

** High and low (based on LRMC and KM plot cut-offs): categorical values

CHAPTER 5

DISCUSSIONS

5.1. Neuropeptide-based classification of CRC patients

Understanding the mechanism of the role of neuropeptides in various cancer type, their progression, and the association of neuropeptides with prognosis has been a research focus in recent years. Many studies have shed light on some of these mechanisms, such as the bi-directional interaction between the neuropeptide release and changes in the TME that are suggested to be required for the maintenance of solid tumors. [62, 63] Regardless of these new discoveries in this field, many unknowns exist about neuropeptides' involvement in cancer. In this thesis, we utilized these neuropeptides as biomarkers to gain more insight into their functions in CRC development and progression.

Our analysis identified 9 neuropeptide genes with two distinct expression patterns, dividing our samples into Up and Down-regulated groups. Table 5.1 below describes the neuropeptides' association with CRC, their identified prognostic associations, and possible mechanisms. These genes later stratified our samples based on their expression profiles. The functional analysis of these groups highlighted several pathways that lead to cancer progression.

All the 9 genes were associated with CRC, however, only 6 had been investigated in CRC in regards to their association with either good or poor prognosis of CRC patients. Two of these genes (VIP and ADCYAP1) were the two most extensively studied in CRC compared to the other genes. VIP, a member of the glucagon family, is responsible for lowering the blood pressure and

increasing glycogenesis. VIP has been shown to be associated with both good and poor prognosis of CRC in various studies. [72] This was consistent with our results. As it can be seen from Figure 4.16, lower expression of the gene showed good prognostic associations, while higher expression was associated with worse prognosis. ADCYAP1, a cMAP stimulator and the main paralogue of VIP, has been linked with poor prognosis in CRC. [73] Similar to VIP, our results indicated a poor prognosis associated with the higher expression of this gene, however, in addition to poor prognosis, we also showed a good prognosis at the lower expression levels of the gene.

PYY, an essential protein in the maintenance of the normal function of the gut such as its motility and energy homeostasis, was linked to good prognosis in CRC in one study. [69] But, our results did not show a coherent significant association with the changes in the expression of the gene and poor or good prognosis. Moreover, IGF1, a member of the growth and development mediator family group of genes [70], SCG2, one of the chromogranin proteins with endocrine secretory functions involved in the packaging of hormones [74], and NPY, expressed widely in the central nervous system with influence on various physiological functions [75] were the three other genes associated with poor prognosis in CRC in the literature. This was consistent with the prognostic association pattern observed in our results. As for NPY, the association with the higher expression of the gene and poor prognosis was not significant in distal but followed the same pattern.

The 3 remaining genes were the ones with less extensive prognostic studies compared to the other genes. SST, an important regulator of the endocrine system with the ability to interact with and influence the hormones of the gastrointestinal system, showed an association with a more poorly differentiated tumor at lower expression levels which could suggest a good prognosis association with higher expression of SST in CRC. [71] This was not in line with the results of our analyses which showed a poor prognosis when SST was up-regulated. Lastly, SCG3, with an unknown

function could act as a helper in prohormone's sorting and proteolytic processes, and CHGB, a regulator protein precursor, were only shown to have lower expression in the tumor samples compared to the normal samples which was what we had also observed in our data, and no link with the prognosis of CRC. [77, 78] Our prognostic studies indicated a significant poor prognosis linked with the higher expression of SCG3 but no coherent significant results for CHGB.

In conclusion, the prognostic analysis (LRMC graph of Figure 4.16) done in this study showed similar patterns with what was observed in the literature for these genes. Our study also provided a significant poor prognosis with up-regulation of SCG3 in both distal and proximal samples. However, more extensive prognostic studies are required for a more clear and comprehensive conclusion of the associations of these genes in CRC. Further information on the functions of the genes and potential mechanisms of their associations in CRC are depicted in Appendix R and Table 5.1, respectively.

Table 5.1. List of identified neuropeptide genes and their associations in CRC

Gene	Association in CRC	Prognosis	Mechanism	Reference
VIP	YES	Bad	Higher expression observed in aggressive and poorly differentiated tumors	72
		Good	Tumor cell migration inhibited by VIP	73
ADCYAP1	YES	Bad	Dysregulation of Circadian clock genes such as ADCYAP1 lead to CRC development. Up-regulation of ADCYAP1 is linked to shorter Disease Free survival.	76
IGF1	YES	Bad	Overexpression promotes HOXA13 facilitated metastasis	70
SST	YES	Not investigated	One study showed low expression levels correlated with grading/ poorly differentiated tumors	71
SCG2	YES	Bad	Influences tumor immunity by regulating infiltration of immune cells and macrophage polarization	74
SCG3	YES	Not investigated	Lower expression in tumor vs normal	77
CHGB	YES	Not investigated	Lower expression in tumor vs normal	78
NPY	YES	Bad	High serum level is correlated with relapse and metastasis	75
PYY	YES	Good	Overexpression induces apoptosis, migration, and invasion	69

5.1.1. Higher expression observed in Normal versus Tumor samples

The up-regulation of the majority of the identified neuropeptides leads to the progression of colorectal cancer and signifies a bad prognostic association. Thus, we expected to see a lower expression profile in normal samples. However, contrary to our hypothesis, the expression profiles

in TCGA (Figure 4.4), IHC staining results, and Tumor, Normal, and Metastatic plots (Figures 4.5, 4.6, and 4.15, respectively) showed significantly lower expression level of these neuropeptides in the tumors compared to normal samples. Several studies have shown that no nerve fibers are detected within the CRC tumor masses, and they have only seen nerve fibers and some signs of neurogenesis in the submucosa layer. “Silent clinical presentation” of colorectal cancer is due to the lack of any clinical symptoms until the tumor has metastasized through and out of the colonic wall. [71, 72] Additionally, another study focusing on the innervation in tumors stated that the blood vessels in colorectal cancer tumors are not innervated, meaning the nerve supply is not in the tumors, and the support must be through other mechanisms. [73] Since neuropeptides are expressed from the neurons, denervation could be a plausible explanation for the observed pattern based on these results.

5.1.2. Hedgehog signaling pathway in CRC development

Hedgehog (HHG) signaling pathway plays a crucial role during the embryonic development phase of vertebrates by regulating the patterning. Dysregulation of it has been linked to diseases, mainly cancer. HHG was one of the main enriched pathways of our identified genes from the GSEA enriched gene-sets in the up-regulated sample group. HHG signaling pathway has been linked with the progression of various cancer types as well as CRC. [74, 75] However, its association with CRC has been controversial. [76] Some studies have shown an association of CRC with the down-regulation of HHG, while others show the opposite. This controversy is also present in our results, as there are two enriched networks of HHG signaling pathways, one referring to its activation and one referring to its suppression. (Figure 4.11) Moreover, HHG’s involvement and impact on CRC depends on the individuals’ clinicopathological characteristics. Despite these inconsistencies, several studies have identified some of the mechanisms of HHG in cancer progression. These

include; helping the de-differentiation of the tumors at the Epithelial-Mesenchymal Transitions' initial steps, angiogenesis, and gene mutation. Additionally, HHG activities in the TME provide immune tolerance. [74-76]

5.1.3. CMS4 subtype molecular characteristics

The CMS subtyping is based on the samples' molecular characteristics and activated pathways. The features of the CMS4 subtype include; stroma infiltration and activation, mesenchymal associating genes' up-regulation, angiogenesis, and elevated TGF β levels. Moreover, CMS4 patients are associated with the highest resistance to therapies and the lowest survival rate consequently. [35] Our Chi-square results showed that among all 4 CMS subtypes, only CMS4 showed significant results, with most up-regulated samples being CMS4. (Figure 4.7) The identified neuropeptides are over-expressed in the CMS4 samples; thus, the poor prognosis observed with the up-regulation of the majority of the neuropeptide genes (5 out of 9) could be due to this subtype's metastatic and aggressive nature. (Figure 4.16)

5.1.4. HHG, CMS4 Subtype associations with Neuropeptides' overexpression

As mentioned above, the up-regulated samples were mostly the CMS4 subtype. Moreover, the HHG signaling pathway was overrepresented in the genes involved in the enriched gene-sets of the up-regulated samples. Lastly, the TGF β response pathway was significantly enriched in the up-regulated samples. (Table 4.4) Therefore, as depicted in Figure 5.1 below, there are commonalities between the discussed results, which could suggest how HHG might be leading to CRC progression. The CMS4 subtype patients have activated stroma, and several studies indicate that HHG's tumorigenic activities are detected in the stroma of the tumor rather than the epithelial tissue. [77] HHG also leads to CRC progression through EMT, and the CMS4 subtype is known as the mesenchymal subtype. These common features observed with the up-regulation of the

neuropeptides lead to the question of the role of neuropeptides in the HHG signaling pathway in the stroma of CRC tumors.

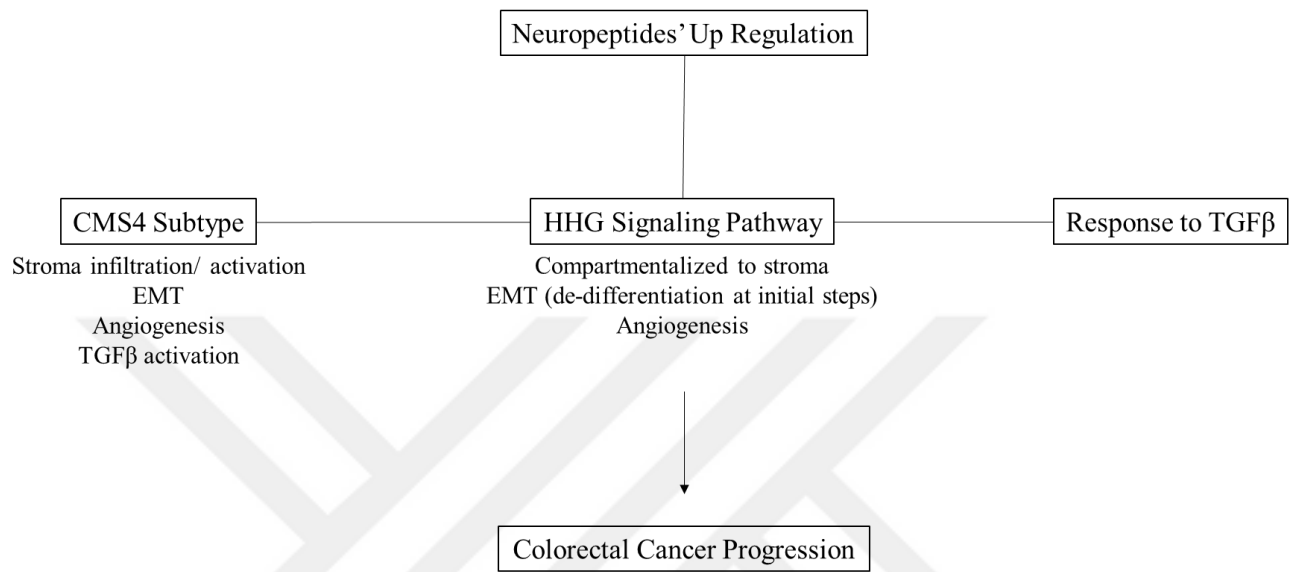


Figure 5.1. Schematic demonstration of the association of the identified neuropeptide genes' up-regulation with the Hedgehog signaling pathway, CMS4 subtype, and TGFβ response pathway and their impact on CRC progression. (EMT: Epithelial-Mesenchymal Transition)

5.2. Identification of robust prognostic biomarkers in Proximal, Stage 2 and 3 CRC patients

As summarized previously and stated by Koncina E. et al, some of the factors causing inconsistencies between the biomarkers in literature and their applications in the clinics are the variations and differences among different platforms, technologies, and cohorts used. [38] Besides these differences, the samples used for the study and their clinicopathological data (such as; location, stage, MSI status, and mutations) also play an important role in the robustness and consistency of a biomarker. For instance, a gene signature/ biomarker might have prognostic significance in distal but not proximal tumors. Similar to the gene signature identified by Demirkol S. et al, that showed significant prognostic association in distal but not proximal disease. [86]

Additionally, another very crucial aspect overlooked in many biomarker studies using microarray datasets is the discrepancies among the probesets of one gene. It has been shown that in genes with multiple probesets, each might be associated with functionally distinct isoforms of the gene. Thus, encouraging the researchers to perform transcriptomic biomarker studies at a probeset level rather than a gene level. [44, 45]

5.2.1. Advantageous aspects of the utilized methodology

In this study, we tried to include all these parameters in our analyses to ensure the identification of robust prognostic biomarkers. We started the analyses by utilizing three different datasets from two different technologies; 2 microarray datasets with the same platform, and 1 RNA-Seq dataset. The advantage of the 3 datasets was not only the elimination of technology and cohort biases, but also it enabled us to test the prognostic association of our biomarker with different survival data types (OS, RFS/ DFS). Moreover, our study only focused on proximal, Stage 2 and 3 samples. Thus, providing a specific sample group in terms of their molecular characteristics that would allow for a more definite biomarker. Lastly, one other main focus of this study was the development of a pipeline for the microarray dataset. In this pipeline, we only included genes with no discrepancies in their probesets including their significance and prognostic values. We also tested our biomarkers in the commonly clinically used classification methods; CMS and Oncotype Dx risk groups.

5.2.2. The prognostic association of the 3 genes in all samples versus subgroups

After determining a gene list in GSE39582 and TCGA, and further validation in GSE14333, we identified 3 genes with significant association with favorable and non-favorable prognosis. These 3 genes included GAA, SEMA4C, and OXNAD1. (Table 4.7) These genes all satisfy our criteria; Cox P value ≤ 0.05 , HR value ≤ 0.67 or HR value ≥ 1.5 , and consistent HR directions across all

probesets of the genes in 3 different datasets. The LRMC graphs followed by their Kaplan Meier plots at their defined cut-offs verified the significant associations observed previously since all the graphs indicated significant prognostic association with higher expression of these genes. (Figure 4.18) Although the tumor location and stage were taken into account, we did not observe significance for all the clinicopathologically determined subgroups. Contrary to our results with all the samples (proximal, Stage 2 and 3), only a limited subgroup of samples, such as stage 2 with Kras/Braf mutant samples for GAA or stage 3 with microsatellite stable samples, showed significant prognostic associations. (Table 4.8) These results further highlighted the importance of the selection of the samples, and the role of the molecular characteristics in the progression and prognosis of CRC and further emphasizes the molecular heterogeneity observed in CRC.

Literature research on the identified genes

GAA, Alpha Glucosidase, gene is a crucial element in the degradation of glycogen to glucose as it encodes the lysosomal acid alpha glucosidase enzyme. Consistent with its association with non-favorable prognosis from our analyses, a systematic review done in 2017 concentrating on Diabetic patients demonstrated that in patients with Diabetes 2, alpha glucosidase (AG) inhibitors lowered the risk of CRC development. Indicating a relationship with non-favorable prognosis for GAA expression in CRC patients with Diabetes 2. [78, 79] A more recent study in 2021 also identified an 11 poor prognostic gene panel, including GAA. [80] The suggested prognosis from the literature of this gene is in line with our findings.

OXNAD1, Oxidoreductase NAD-binding domain-containing protein 1, is a protein coding gene predicted to be an oxidoreductase activator. Studies researching its impact on CRC have not tested the prognosis of this gene individually. In one study, it was involved in the assay developed for the validation of a prognostic assay for stage 2 CRC patients. However, specific details on the

prognostic value of this gene individually are not mentioned. The other studies proposing prognostic associations for OXNAD1 are not CRC specific. One pan-cancer study, including CRC, suggested a favorable prognosis association with increased expression of OXNAD1, similar to the favorable prognosis found from our analyses. [81, 82]

SEMA4C, Semaphorin 4 C, is a chemorepellent and semaphorin binding activator. It is involved in stress induced activation of MAPK positive regulation. One study, including all the samples in the TCGA-COAD dataset, linked the overexpression of SEMA4C with increased EMT and poor prognosis in CRC. [83] Similarly, another study explained that the FOXD2-AS1 leads to overexpression of SEMA4C that increases invasion and emigration, causing EMT and a poor prognosis. [84] These results, like the other 2 genes, are consistent with our suggested non-favorable prognosis of SEMA4C in the specified CRC subgroups.

5.3. Limitations and Future Studies

This study consisted of two different approaches for classification of CRC patients. In approach one, the focus was on the classification of all colon cancer patients based on a specific gene list, while in approach 2, the focus was on the classification of a specific sample group.

APPROACH 1: In this part of the study, we considered PC1 genes, which only explain 15% of the variance within the samples. Thus, the other PC components should also be studied, as those other genes can provide more insight into the mechanisms of neuropeptides' involvement in colorectal cancer. Furthermore, we focused on the common genes between distal and proximal samples. However, the distinct differences between these two differentially located tumors indicate colorectal cancer development and progression via neuropeptides can have a different mechanism in distal compared to proximal and more comprehensive studies on those would be valuable in

providing information about the location specific mechanisms involved. Besides expanding the gene list studied, there can be further studies with the current gene list. The up-regulation of these genes showed a decrease in the hedgehog's downstream signaling pathways in the stroma of colorectal cancer. Therefore, an elaborate study on the role of neuropeptides in the stroma of colorectal cancer and their associations with the hedgehog pathway will provide more mechanistic information. Lastly, we observed a significant drop in the expression of the neuropeptides in tumor samples compared to normal samples. One reason behind such a pattern was proposed to be the nerve loss and a lack of nerve supplies and nerves within the tumors. In order to confirm the nerve loss, there can be immunohistochemistry staining studies where the antibodies/ markers utilized are specific to nerve cells, and healthy samples are compared to tumor-adjacent normal and tumor samples. Moreover, the single-cell analysis with the RNAseq dataset from CRC tumors (GSE178318) showed no expression of our genes from the available cell types' clusters available, except for IGF1, which is expressed in Cancer-Associated Fibroblasts (CAFs) and plasma, meaning the neuropeptides are not expressed from the cells included in this analysis. (Appendix C) So the IHC staining specific to neurons can also further provide information on their expression sources.

APPROACH 2: One of the main limitations faced during this approach was the sample number. Since the aim of the study was the inclusion of the clinicopathological data mentioned, the number of datasets with these clinical data were limited. The two selected datasets for the discovery of the gene list were GSE39582 and TCGA with 190 and 100 samples, respectively. Even though these two datasets had the highest sample number compared to other available datasets at the time of the study, further classification of the samples into stages 2 and 3 or CMS subgroups would have resulted in statistical issues as the sample number would have been low. Furthermore, the third

dataset used for the gene selection (GSE14333) could not be used at later stages for the validation of the prognostic significance of the genes in the stratified subgroups as this dataset lacked the required clinical data.

In this approach, we aimed to develop a pipeline for biomarker studies that factors in the issues discussed by the literature for the inconsistencies across prognostic biomarkers. For doing so, the sample list was narrowed down to only proximal and limiting the stage to stages 2 and 3 only. The next step was the inclusion of three different datasets from two different technologies and platforms for eliminating any biases as a result of the cohort or technology/ platform specificities. As for the microarray datasets, the discrepancies across one gene were factored in during the selection process. This advantage of this workflow was the consistency of the results across microarray and RNA-Seq datasets and in both OS and RFS/ DFS survival data. However, the workflow also has drawbacks due to its stringency. Since the selection criteria is significantly more stringent for genes with more than one probeset, the selection process favors genes with a single probeset. Additionally, a large number of genes with more than one probeset are lost as any gene with opposing and inconsistent probeset results are not included. As seen in Figure 4.17, the initial step of the analysis (Cox Regression analysis) is run by more than 23000 genes in GSE39582, with similar distribution in one probeset genes and more than one probeset genes. (around 12000 genes in each group). The elimination of genes with at least one non-significant probeset, the gene list is reduced to as low as 385. Additionally, the GSE39582 final gene list (N = 2029) is further narrowed down, as it is compared with TCGA gene list (N = 719) resulting in only 43 genes, with only 3 being validated in GSE14333. However, some of the eliminated genes could be of importance and interest as there might be genes with functionally distinct isoforms and

different role in CRC progression. Further studies on these genes can provide valuable information on the involvement of these genes on the advancement of the disease.

In the future, larger datasets with comprehensive clinical data is required for further validation of the identified genes in their specified clinical subgroups. Also these biomarkers need to be validated in in-vitro in patients with known clinicopathological data. Additional functional analysis would also be beneficial in understanding the biological mechanisms which can aid in the personalization of the treatment for each patient subgroup followed by drug sensitivity tests for the selection of the best and most efficient treatment plan. Lastly, the workflow could be adjusted according to the points discussed above for future implication in prognostic biomarker studies.

CHAPTER 6

BIBLIOGRAPHY

1. Colorectal cancer - statistics. Cancer.Net. (2023, April 4). <https://www.cancer.net/cancer-types/colorectal-cancer/statistics>
2. World Health Organization. (n.d.). World Health Organization. <https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/>
3. Xi, Y., & Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology*, 14(10), 101174. <https://doi.org/10.1016/j.tranon.2021.101174>
4. *Kanser istatistikleri*. Halk Sağlığı Genel Müdürlüğü. (n.d.). <https://hsgm.saglik.gov.tr/tr/kanser-istatistikleri>
5. Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*, 138(6), 2073-2087.
6. Breakstone, R. (2021). Colon cancer and immunotherapy—can we go beyond microsatellite instability?. *Translational Gastroenterology and Hepatology*, 6.
7. Laghi, L., Negri, F., Gaiani, F., Cavalleri, T., Grizzi, F., de'Angelis, G. L., & Malesci, A. (2020). Prognostic and predictive cross-roads of microsatellite instability and immune response to colon cancer. *International Journal of Molecular Sciences*, 21(24), 9680.
8. Saridaki, Z., Souglakos, J., & Georgoulis, V. (2014). Prognostic and predictive significance of MSI in stages II/III colon cancer. *World Journal of Gastroenterology: WJG*, 20(22), 6809.
9. Tejpar, S., Bertagnolli, M., Bosman, F., Lenz, H. J., Garraway, L., Waldman, F., ... & Roth, A. (2010). Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *The oncologist*, 15(4), 390-404.
10. Sargent, D. J., Marsoni, S., Monges, G., Thibodeau, S. N., Labianca, R., Hamilton, S. R., ... & Gallinger, S. (2010). Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *Journal of clinical oncology*, 28(20), 3219.
11. Ahmed, M. (2020). Colon cancer: a clinician's perspective in 2019. *Gastroenterology research*, 13(1), 1.
12. Nakayama, M., & Oshima, M. (2019). Mutant p53 in colon cancer. *Journal of molecular cell biology*, 11(4), 267-276.
13. Taieb, J., Le Malicot, K., Shi, Q., Penault-Llorca, F., Bouché, O., Tabernero, J., ... & Sinicrope, F. A. (2017). Prognostic value of BRAF and KRAS mutations in MSI and MSS stage III colon cancer. *Journal of the National Cancer Institute*, 109(5), djw272.
14. Popovici, V., Budinska, E., Tejpar, S., Weinrich, S., Estrella, H., Hodgson, G., ... & Delorenzi, M. (2012). Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J Clin Oncol*, 30(12), 1288-1295.
15. Siddiqui, A. D., & Piperdi, B. (2010). KRAS mutation in colon cancer: a marker of resistance to EGFR-I therapy. *Annals of surgical oncology*, 17, 1168-1176.

16. Franklin, W. A., Haney, J., Sugita, M., Bemis, L., Jimeno, A., & Messersmith, W. A. (2010). KRAS mutation: comparison of testing methods and tissue sampling techniques in colon cancer. *The Journal of Molecular Diagnostics*, 12(1), 43-50.
17. Arrington, A. K., Heinrich, E. L., Lee, W., Duldulao, M., Patel, S., Sanchez, J., ... & Kim, J. (2012). Prognostic and predictive roles of KRAS mutation in colorectal cancer. *International journal of molecular sciences*, 13(10), 12153-12168.
18. Demirkol, S., Gomceli, I., Isbilen, M., Dayanc, B. E., Tez, M., Bostanci, E. B., ... & Gure, A. O. (2017). A combined ULBP2 and SEMA5A expression signature as a prognostic and predictive biomarker for colon cancer. *Journal of Cancer*, 8(7), 1113.
19. Mojarad, E. N., Kuppen, P. J., Aghdaei, H. A., & Zali, M. R. (2013). The CpG island methylator phenotype (CIMP) in colorectal cancer. *Gastroenterology and hepatology from bed to bench*, 6(3), 120.
20. O'Connell, J. B., Maggard, M. A., & Ko, C. Y. (2004). Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *Journal of the National Cancer Institute*, 96(19), 1420-1425.
21. Stintzing, S. (2014). Management of colorectal cancer. *F1000prime reports*, 6.
22. Barrier, A., Boelle, P. Y., Roser, F., Gregg, J., Tse, C., Brault, D., ... & Dudoit, S. (2006). Stage II colon cancer prognosis prediction by tumor gene expression profiling. *Journal of Clinical Oncology*, 24(29), 4685-4691.
23. Jiang, D., Liao, J., Duan, H., Wu, Q., Owen, G., Shu, C., ... & Wang, Z. (2020). A machine learning-based prognostic predictor for stage III colon cancer. *Scientific reports*, 10(1), 1-9.
24. Shen H, Yang J, Huang Q, Jiang MJ, Tan YN, Fu JF, Zhu LZ, Fang XF, Yuan Y. Different treatment strategies and molecular features between right-sided and left-sided colon cancers. *World J Gastroenterol*. 2015 Jun 7;21(21):6470-8. doi: 10.3748/wjg.v21.i21.6470. PMID: 26074686; PMCID: PMC4458758.
25. Ghidini, M., Petrelli, F. & Tomasello, G. Right Versus Left Colon Cancer: Resectable and Metastatic Disease. *Curr. Treat. Options in Oncol.* 19, 31 (2018). <https://doi.org/10.1007/s11864-018-0544-y>
26. Barras, D. (2015). BRAF mutation in colorectal cancer: an update: supplementary issue: biomarkers for colon cancer. *Biomarkers in cancer*, 7, BIC-S25248.
27. Lee, M. S., Menter, D. G., & Kopetz, S. (2017). Right versus left colon cancer biology: integrating the consensus molecular subtypes. *Journal of the National Comprehensive Cancer Network*, 15(3), 411-419.
28. Meguid, R. A., Slidell, M. B., Wolfgang, C. L., Chang, D. C., & Ahuja, N. (2008). Is there a difference in survival between right-versus left-sided colon cancers?. *Annals of surgical oncology*, 15, 2388-2394.
29. De Renzi, G., Gaballo, G., Gazzaniga, P., & Nicolazzo, C. (2021). Molecular biomarkers according to primary tumor location in colorectal cancer: current standard and new insights. *Oncology*, 99(3), 135-143.
30. Wray, C. M., Ziogas, A., Hinojosa, M. W., Le, H., Stamos, M. J., & Zell, J. A. (2009). Tumor subsite location within the colon is prognostic for survival after colon cancer diagnosis. *Diseases of the Colon & Rectum*, 52(8), 1359-1366.
31. Demirkol, S., Gomceli, I., Isbilen, M., Dayanc, B. E., Tez, M., Bostanci, E. B., ... & Gure, A. O. (2017). A combined ULBP2 and SEMA5A expression signature as a prognostic and predictive biomarker for colon cancer. *Journal of Cancer*, 8(7), 1113.

32. Demirkol Canlı, S., Seza, E. G., Sheraj, I., Gömçeli, I., Turhan, N., Carberry, S., ... & Banerjee, S. (2020). Evaluation of an aldo-keto reductase gene signature with prognostic significance in colon cancer via activation of epithelial to mesenchymal transition and the p70S6K pathway. *Carcinogenesis*, 41(9), 1219-1228.
33. Yothers, G., O'Connell, M. J., Lee, M., Lopatin, M., Clark-Langone, K. M., Millward, C., ... & Wolmark, N. (2013). Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *Journal of clinical oncology*, 31(36), 4512.
34. Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M. C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J. F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., ... Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, 10(5), e1001453. <https://doi.org/10.1371/journal.pmed.1001453>
35. Guinney, J., Dienstmann, R., Wang, X. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 21, 1350–1356 (2015). <https://doi.org/10.1038/nm.3967>
36. Filip S, Vymetalkova V, Petera J, Vodickova L, Kubecek O, John S, Cecka F, Krupova M, Manethova M, Cervena K, Vodicka P. Distant Metastasis in Colorectal Cancer Patients-Do We Have New Predicting Clinicopathological and Molecular Biomarkers? A Comprehensive Review. *Int J Mol Sci.* 2020 Jul 24;21(15):5255. doi: 10.3390/ijms21155255. PMID: 32722130; PMCID: PMC7432613.
37. Sagaert, X. Prognostic biomarkers in colorectal cancer: where do we stand?. *Virchows Arch* 464, 379–391 (2014). <https://doi.org/10.1007/s00428-013-1532-z>
38. Koncina, E., Haan, S., Rauh, S., & Letellier, E. (2020). Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. *Cancers*, 12(2), 319.
39. Sinicrope, F. A., Shi, Q., Smyrk, T. C., Thibodeau, S. N., Dienstmann, R., Guinney, J., ... & Alberts, S. R. (2015). Molecular markers identify subtypes of stage III colon cancer associated with patient outcomes. *Gastroenterology*, 148(1), 88-99.
40. Fontana, E., Eason, K., Cervantes, A., Salazar, R., & Sadanandam, A. (2019). Context matters—consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals of Oncology*, 30(4), 520-527.
41. Dienstmann, R., Mason, M. J., Sinicrope, F. A., Phipps, A. I., Tejpar, S., Nesbakken, A., ... & Guinney, J. (2017). Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. *Annals of Oncology*, 28(5), 1023-1031.
42. Alwers, E., Jia, M., Kloor, M., Bläker, H., Brenner, H., & Hoffmeister, M. (2019). Associations between molecular classifications of colorectal cancer and patient survival: a systematic review. *Clinical Gastroenterology and Hepatology*, 17(3), 402-410.
43. Huang, Z., Yang, Q., & Huang, Z. (2018). Identification of Critical Genes and Five Prognostic Biomarkers Associated with Colorectal Cancer. *Medical science monitor : international medical journal of experimental and clinical research*, 24, 4625–4633. <https://doi.org/10.12659/MSM.907224>
44. Zhang, Z., Pal, S., Bi, Y. *et al.* Isoform level expression profiles provide better cancer signatures than gene level expression profiles. *Genome Med* 5, 33 (2013). <https://doi.org/10.1186/gm437>

45. Yi, L., Pimentel, H., Bray, N.L. et al. Gene-level differential analysis at transcript-level resolution. *Genome Biol* 19, 53 (2018). <https://doi.org/10.1186/s13059-018-1419-z>
46. The Cancer Genome Atlas Research Network., Weinstein, J., Collisson, E. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120 (2013). <https://doi.org/10.1038/ng.2764>
47. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337 (2012). <https://doi.org/10.1038/nature11252>
48. Jorissen, R. N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., ... & Sieber, O. M. (2009). Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clinical Cancer Research*, 15(24), 7642-7651.
49. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, et al. (2006) Design and Analysis of DNA Microarray Investigations: Springer New York.
50. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, 15, 550. [doi:10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
51. Yan Wang and others, NeuroPep: a comprehensive resource of neuropeptides, *Database*, Volume 2015, 2015, bav038, <https://doi.org/10.1093/database/bav038>
52. Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. <https://bioinfogp.cnb.csic.es/tools/venny/index.html>
53. Graphpad Software Inc., San Diego, CA, USA
54. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545-15550.
55. Mootha, V., Lindgren, C., Eriksson, KF. et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267–273 (2003). <https://doi.org/10.1038/ng1180>
56. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*. 2012 Sep 15;28(18):i451-i457. doi: 10.1093/bioinformatics/bts389. PMID: 22962466; PMCID: PMC3436816.
57. Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
58. Chi-Square Test Calculator. (2023, June 10). Retrieved from <https://www.socscistatistics.com/tests/chisquare2/default2.aspx>.
59. Bartha Á, Györfy B. TNMplot.com: A Web Tool for the Comparison of Gene Expression in Normal, Tumor and Metastatic Tissues. *International Journal of Molecular Sciences*. 2021; 22(5):2622. <https://doi.org/10.3390/ijms22052622>
60. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Al-Khalili Szigartyo, C., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., . . . Pontén, F. (2015). Tissue-based map of the human proteome. *Science*. <https://doi.org/1260419>
61. Wu, Y., Berisha, A., Borniger, J.C., 2022. Neuropeptides in Cancer: Friend and Foe?. *Advanced Biology* 6, 2200111.. <https://doi.org/10.1002/adbi.202200111>
62. Wang, H., Zheng, Q., Lu, Z., Wang, L., Ding, L., Xia, L., ... & Li, G. (2021). Role of the nervous system in cancers: a review. *Cell Death Discovery*, 7(1), 76.

63. David L. Felten, M. Kerry O'Banion, Mary Summo Maida, "1-Neurons and Their Properties", Editor(s): David L. Felten, M. Kerry O'Banion, Mary Summo Maida, Netter's Atlas of Neuroscience (Third Edition), Elsevier, 2016, Pages 1-42, ISBN 9780323265119, <https://doi.org/10.1016/B978-0-323-26511-9.00001-1>
64. Srivastava, A., Rikhari, D., Pradhan, B., Bharadwaj, K. K., Gaballo, A., Quarta, A., ... & Ragusa, A. (2022). An Insight into Neuropeptides Inhibitors in the Biology of Colorectal Cancer: Opportunity and Translational Perspectives. *Applied Sciences*, 12(18), 8990.
65. Venables, W. N. and B. D. Ripley (2002) *Modern Applied Statistics with S*, Springer-Verlag.
66. Terry M. Therneau, Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN 0-387-98784-3.
67. Lausen, B., Hothorn, T., Bretz, F. and Schumacher, M. (2004). Assessment of Optimally Selected Prognostic Factors. *Biometrical Journal*, 46(3), 364–374.
68. Che, L. H., Liu, J. W., Huo, J. P., Luo, R., Xu, R. M., He, C., ... & Li, J. M. (2021). A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. *Cell discovery*, 7(1), 80.
69. Jing, F., Liu, G., Zhang, R., Xue, W., Lin, J., Zhu, H., ... & Bao, M. (2022). PYY modulates the tumorigenesis and progression of colorectal cancer unveiled by proteomics. *American Journal of Cancer Research*, 12(12), 5500.
70. Qiao, C., Huang, W., Chen, J. et al. IGF1-mediated HOXA13 overexpression promotes colorectal cancer metastasis through upregulating ACLY and IGF1R. *Cell Death Dis* 12, 564 (2021). <https://doi.org/10.1038/s41419-021-03833-2>
71. Baraldi, J. H., Martyn, G. V., Shurin, G. V., & Shurin, M. R. (2022). Tumor innervation: History, methodologies, and significance. *Cancers*, 14(8), 1979.
72. Tomita, T. Localization of Nerve Fibers in Colonic Polyps, Adenomas, and Adenocarcinomas by Immunocytochemical Staining for PGP 9.5. *Dig Dis Sci* 57, 364–370 (2012). <https://doi.org/10.1007/s10620-011-1876-7>
73. Mitchell, B. S., Schumacher, U., & Kaiserling, E. (1994). Are tumours innervated? Immunohistological investigations using antibodies against the neuronal marker protein gene product 9.5 (PGP 9.5) in benign, malignant and experimental tumours. *Tumor biology*, 15(5), 269-274.
74. Bian, Y. H., Huang, S. H., Yang, L., Ma, X. L., Xie, J. W., & Zhang, H. W. (2007). Sonic hedgehog-Gli1 pathway in colorectal adenocarcinomas. *World journal of gastroenterology: WJG*, 13(11), 1659.
75. Chung, J. H., & Bunz, F. (2013). A loss-of-function mutation in PTCH1 suggests a role for autocrine hedgehog signaling in colorectal tumorigenesis. *Oncotarget* 4 (12): 2208–2211.
76. Papadopoulos V, Tsapakidis K, Riobo Del Galdo NA, Papandreou CN, Del Galdo F, Anthony A, Sakellaridis N, Dimas K, Kamposioras K. The Prognostic Significance of the Hedgehog Signaling Pathway in Colorectal Cancer. *Clin Colorectal Cancer*. 2016 Jun;15(2):116-27. doi: 10.1016/j.clcc.2016.02.010. Epub 2016 Feb 16. PMID: 27032873.
77. Geyer N, Gerling M. Hedgehog Signaling in Colorectal Cancer: All in the Stroma? *Int J Mol Sci*. 2021 Jan 20;22(3):1025. doi: 10.3390/ijms22031025. PMID: 33498528; PMCID: PMC7864206.
78. Horibe, Y., Adachi, S., Ohno, T., Goto, N., Okuno, M., Iwama, M., Yamauchi, O., Kojima, T., Saito, K., Ibuka, T., Yasuda, I., Araki, H., Moriwaki, H., & Shimizu, M. (2017). Alpha-glucosidase inhibitor use is associated with decreased colorectal neoplasia risk in patients with

- type 2 diabetes mellitus receiving colonoscopy: a retrospective study. *Oncotarget*, 8(58), 97862–97870. <https://doi.org/10.18632/oncotarget.18416>
79. Zhao, Y., Wang, Y., Lou, H., & Shan, L. (2017). Alpha-glucosidase inhibitors and risk of cancer in patients with diabetes mellitus: a systematic review and meta-analysis. *Oncotarget*, 8(46), 81027–81039. <https://doi.org/10.18632/oncotarget.17515>
 80. Chen S, Wang Y, Wang B, Zhang L, Su Y, Xu M, Zhang M. A signature based on 11 autophagy genes for prognosis prediction of colorectal cancer. *PLoS One*. 2021 Oct 26;16(10):e0258741. doi: 10.1371/journal.pone.0258741. PMID: 34699544; PMCID: PMC8547631.
 81. Barrera, G., Cucci, M. A., Grattarola, M., & Pizzimenti, S. (2021). Nrf2, YAP, antioxidant potential, and cancer. In *Cancer* (pp. 159-170). Academic Press.
 82. Sinnott-Armstrong, N., Seoane, J. A., Sallari, R., Pritchard, J. K., Curtis, C., & Snyder, M. P. (2020). Convergent mutations in tissue-specific regulatory regions reveal novel cancer drivers. *bioRxiv*.
 83. Hou, Y., Wang, W., Zeng, Z., Gan, W., Lv, S., Li, T., Yan, Z., Zhang, R., & Yang, M. (2020). High SEMA4C expression promotes the epithelial-mesenchymal transition and predicts poor prognosis in colorectal carcinoma. *Aging*, 12(21), 21992–22018. <https://doi.org/10.18632/aging.104038>
 84. Zhang, M., Jiang, X., Jiang, S., Guo, Z., Zhou, Q., & He, J. (2019). LncRNA FOXD2-AS1 Regulates miR-25-3p/Sema4c Axis To Promote The Invasion And Migration Of Colorectal Cancer Cells. *Cancer management and research*, 11, 10633–10639. <https://doi.org/10.2147/CMAR.S228628>
 85. Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D [Current Protocols in Bioinformatics\(2016\), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5 \[PDF\]](https://doi.org/10.1002/cpbi.5)
 86. Demirkol, S. (2017). Prediction of prognosis and chemosensitivity in gastrointestinal cancers [Doctoral Dissertation, Bilkent University]. <http://hdl.handle.net/11693/35716>

APPENDIX

Appendix A

List of 100 Neuropeptide genes

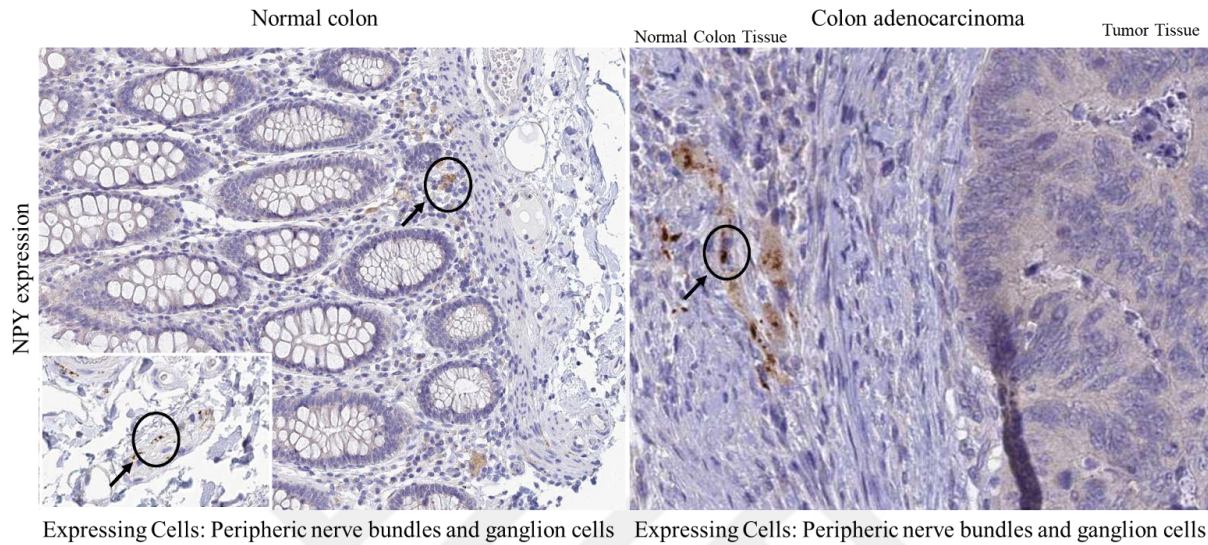
Gene Names	Gene Symbol
Pro-enkephalin gene	PENK
Pro-opiomelanocortin gene	POMC
Pro-dynorphin gene	PDYN
Orphanin gene, pre-pronociceptin gene	PNOC
Vasopressin gene	AVP
Oxytocin gene	OXT
Gastrin gene	GAST
Cholecystokinin gene	CCK
Somastostatin gene	SST
Cortistatin gene	CST
Gonadotropin inhibitory hormone gene, RF-amide related peptide gene	RFRP
Neuropeptide FF gene	NPFF
Neuropeptide Y gene	NPY
Pancreatic polypeptide gene	PPY
Peptide YY gene	PYY
Prolactin-releasing peptide	PRLH
Calcitonin I gene	CALCA
Calcitonin II gene	CALCB
Islet amyloid polypeptide gene	IAPP
Adrenomedullin gene	ADM
Adrenomedullin-2 gene	ADM2
Atrial natriuretic factor gene	NPPA
Brain natriuretic factor gene	NPPB
Natriuretic peptide precursor C gene	NPPC
Gastrin releasing peptide gene	GRP
Neuromedin B gene	NMB
Endothelin 1 gene	EDN1
Endothelin 2 gene	EDN2
Endothelin 3 gene	EDN3
Glucagon gene	CGC
Secretin gene	SCT
Vasoactive intestinal peptide gene	VIP
Pituitary adenylcyclase-activated peptide gene	ADCYAP1
Growth hormone releasing hormone gene	GHRH
Gastric inhibitory peptide gene	GIP
Corticotropin-releasing hormone gene	CRH
Urocortin gene	UCN

Urocortin II gene	UCN2
Urocortin III gene	UCN3
Urotensin-II	UTS2
Urotensin-II domain containing	UTS2D
Preprotachykinin A gene	TAC1
Preprotachykinin B gene	TAC3
Neuromedins	
Neuromedin S gene	NMS
Neuromedin U gene	NMU
Tensins and Kinins	
Kininogen-1 gene	KNG1
Angiotensin gene	AGT
Neurotensin gene	NTS
Granins	Granins
Chromogranin A gene	CHGA
Chromogranin B gene	CHGB
Secretogranin II gene	SCG2
Secretogranin III gene	SCG3
Secretory granule neuroendocrine protein 1, 7B2 gene	SGNE1
VGF nerve growth factor inducible protein	VGF
Motilin gene	MLN
Ghrelin gene	GHRL
Galanin gene	GAL
Galanin-like peptide precursor gene	GALP
Gonadotropin-releasing hormone gene	GnRH1
Gonadotropin-releasing hormone gene	GnRH2
Neuropeptide B gene	NPB
Neuropeptide W gene	NPW
Neuropeptides S gene	NPS
Neurexophilins	
Neurexophilin-1	NXPH1
Neurexophilin-2	NXPH2
Neurexophilin-3	NXPH3
Neurexophilin-4	NXPH4
Insulin gene	INS
IGF-1 gene	IGF1
IGF-2 gene	IGF2
Relaxin-1 gene	RLN1

Relaxin-2 gene	RLN2
Relaxin-3 gene	RLN3
Thyrotropin-releasing hormone gene	TRH
Parathyroid hormone-like hormone gene	PTH LH
Melanin-concentrating hormone gene	PMCH
Hypocretin gene	HCRT
Cocaine- and amphetamine-regulated transcript gene	CART
Agouti-related protein homolog gene	AGRP
Prolactin	PRL
Apelin gene	APLN
Metastasis-suppressor KISS	KISS1
Diazepam-binding inhibitor	DBI
Cerebellins	
Cerebellin-1 gene	CBLN1
Cerebellin-2 gene	CBLN2
Cerebellin-3 gene	CBLN3
Cerebellin-4 gene	CBLN4
Adipose neuropeptides	
Leptin/ob gene	LEP
Adiponectin gene	ADIPOQ
Visfatin gene	PBEF1
Resistin gene	RETN
Resistin-like alpha gene	RETLA
Resistin-like beta gene	RETLB
Resistin-like gamma gene	RETLG
Nucleobindin-2/NEFA gene	NUCB2
Ubiquitin-like 5	UBL5

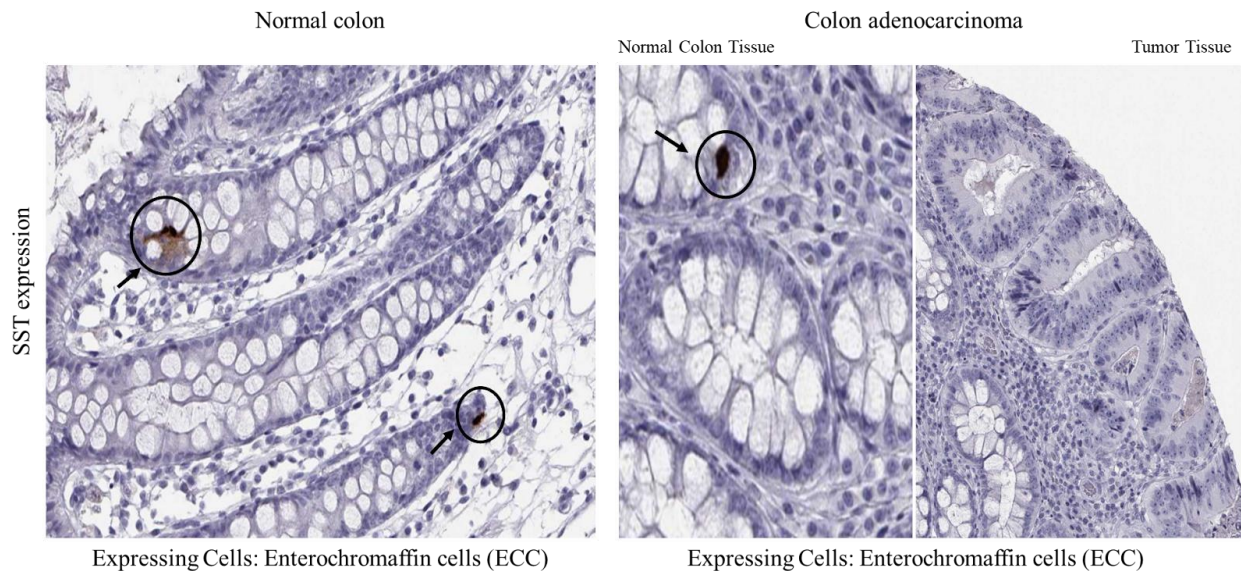
Appendix B.

Antibody Colony: CAB034368



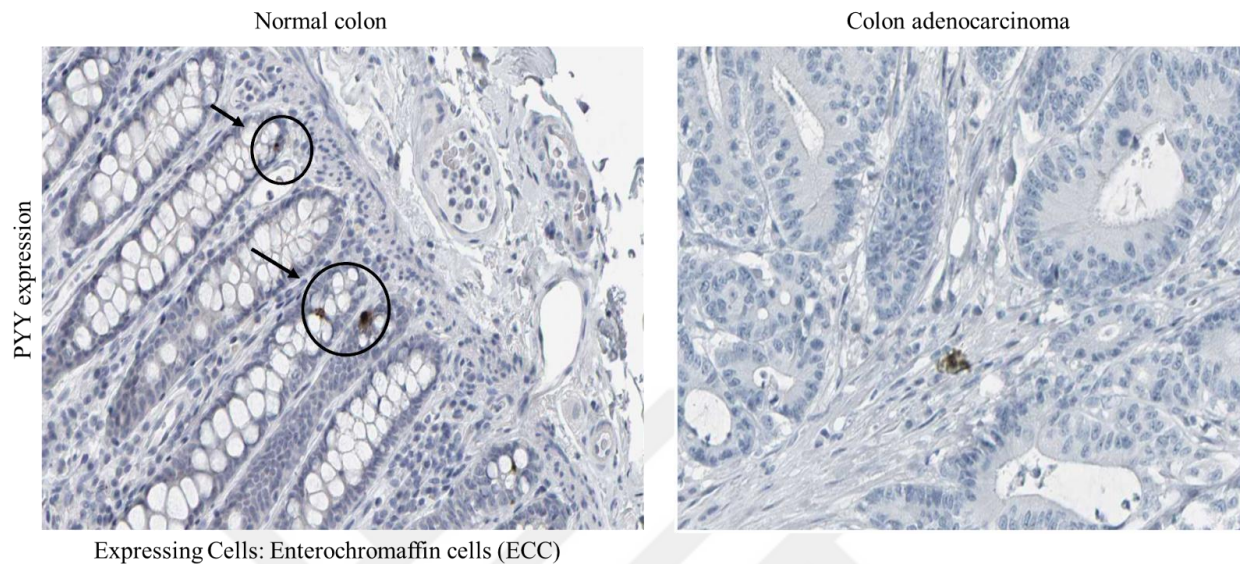
Neuropeptide Y (NPY) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of NPY in the cell type mentioned. As depicted in the figures, NPY is expressed in the normal tissues' peripheral nerve bundles and ganglion cells, while no NPY expression is observed in the tumor tissue.

Antibody Colony: HPA019472



Somatostatin (SST) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of SST in the cell type mentioned. As depicted in the figures, SST is expressed in the normal tissues' ECC cells, while no SST expression is observed in the tumor tissue.

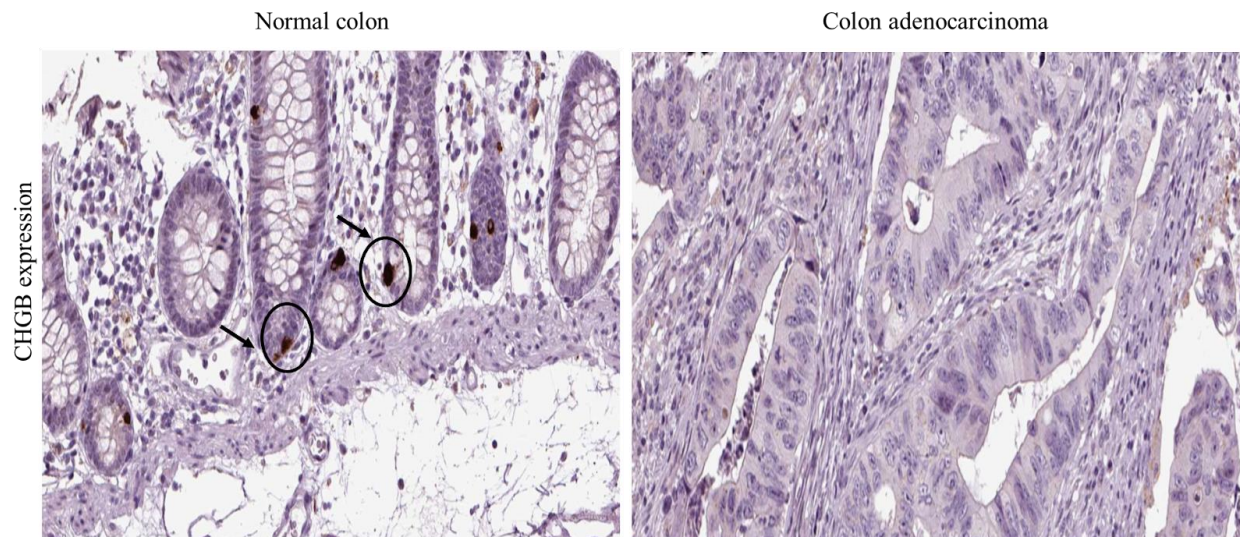
Antibody Colony: HPA010973



Expressing Cells: Enterochromaffin cells (ECC)

Peptide YY (PYY) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of PYY in the cell type mentioned. As depicted in the figures, PYY is expressed in the normal tissues' ECC cells, while no PYY expression is observed in the tumor tissue.

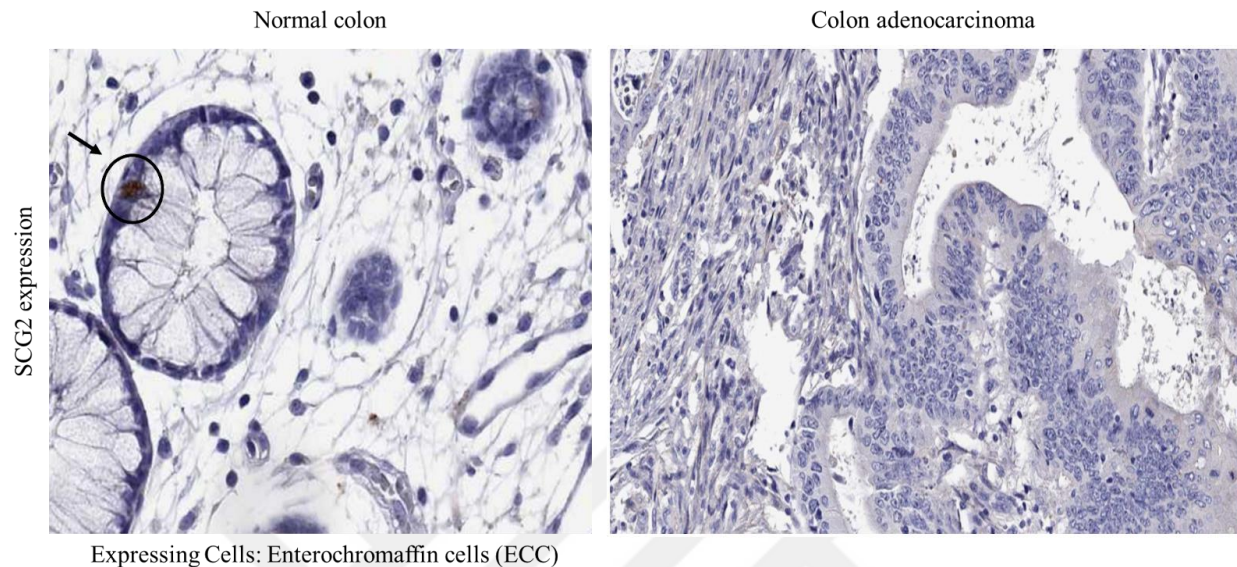
Antibody Colony: HPA012602



Expressing Cells: Enterochromaffin cells (ECC)

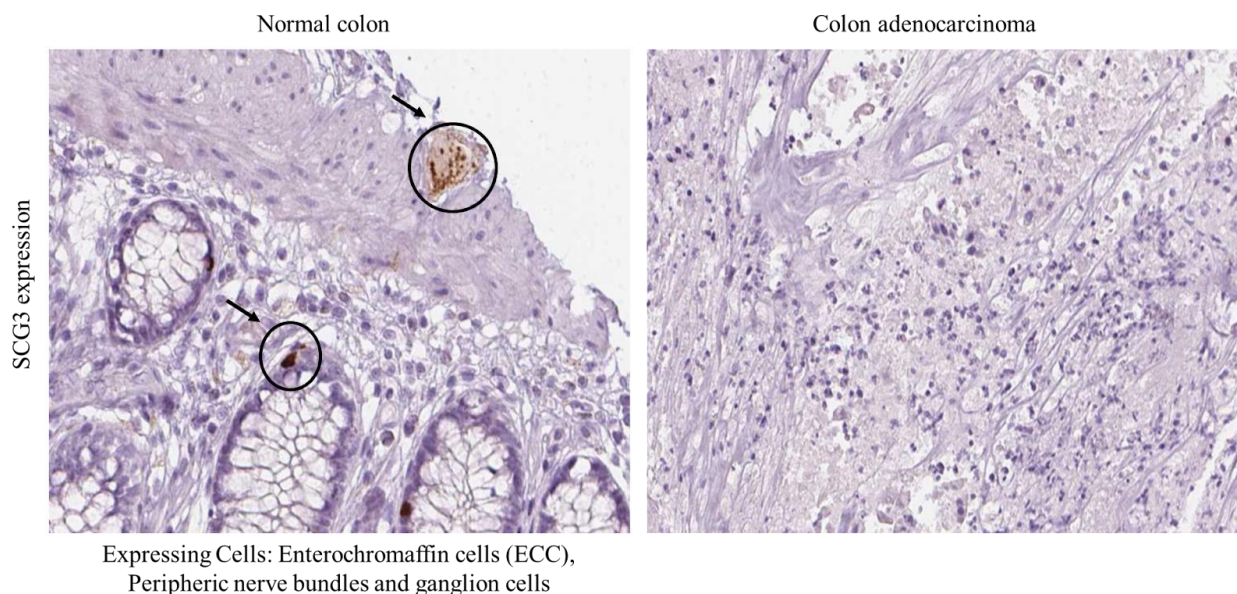
Chromogranin B (CHGB) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of CHGB in the cell type mentioned. As depicted in the figures, CHGB is expressed in the normal tissues' ECC cells, while no CHGB expression is observed in the tumor tissue.

Antibody Colony: HPA011893

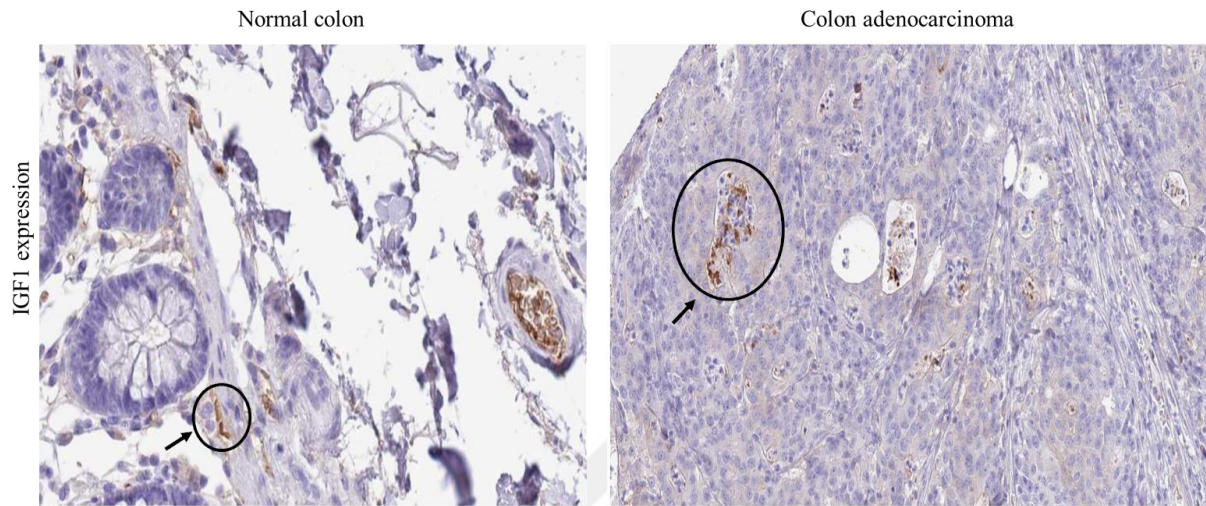


Secretogranin 2 (SCG2) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of SCG2 in the cell type mentioned. As depicted in the figures, SCG2 is expressed in the normal tissues' ECC cells, while no SCG2 expression is observed in the tumor tissue.

Antibody Colony: HPA006880



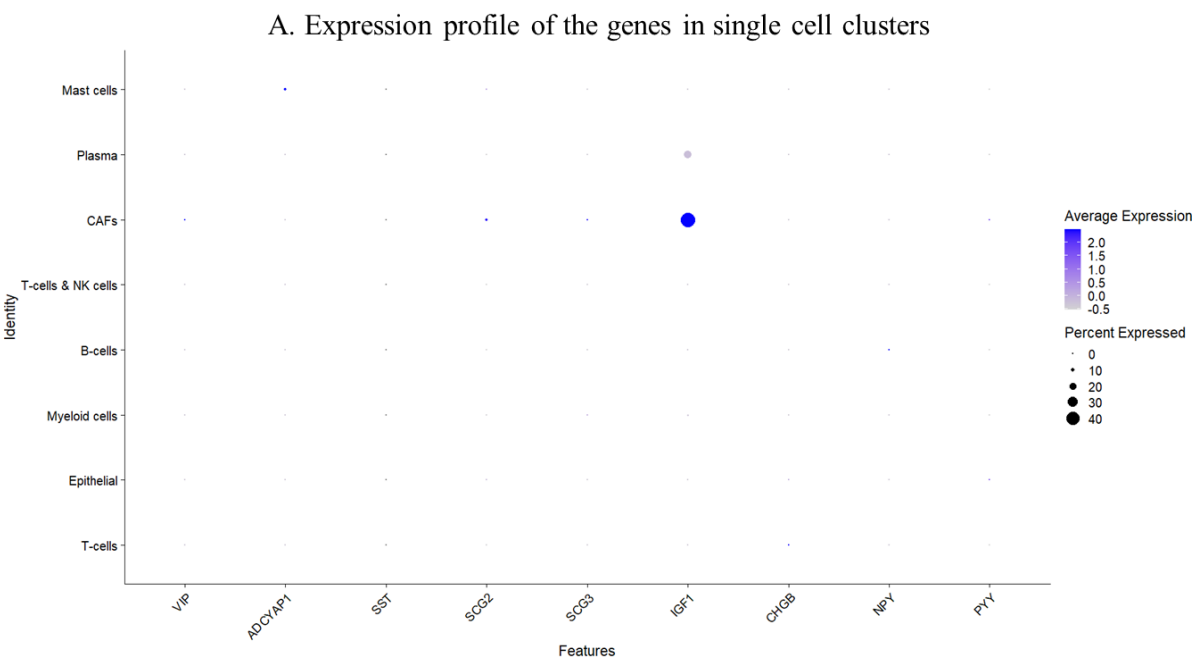
Secretogranin 3 (SCG3) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of SCG3 in the cell type mentioned. As depicted in the figures, SCG3 is expressed in the normal tissues' ECC cells, peripheral nerve bundles, and ganglion cells, while no SCG3 expression is observed in the tumor tissue.



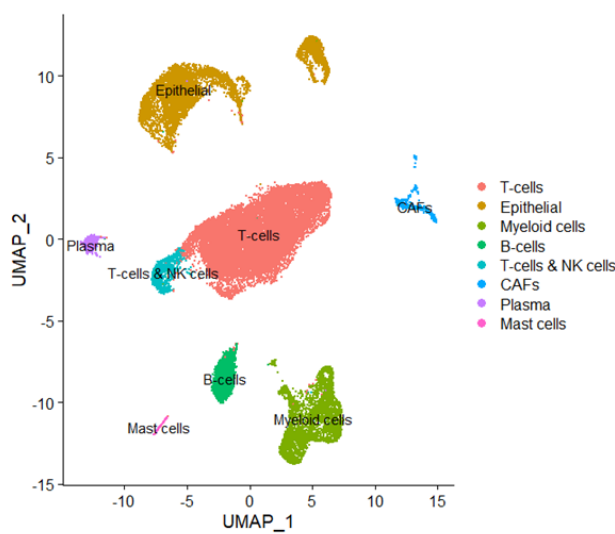
Expressing Cells: Peripheral nerve bundles and ganglion cells, Expressing Cells: Normal Epithelial cells (such as blood vessels) and normal epithelial cells

Insulin-like Growth Factor 1 (IGF1) Expression in Normal Colon vs. Colon Adenocarcinoma. The arrows indicate the expression of IGF1 in the cell type mentioned. As depicted in the figures, IGF1 is expressed in the normal tissues' peripheral nerve bundles and ganglion cells. There is additional IGF1 staining in various epithelial tissues found in normal and tumor samples.

Appendix C.



B. The cell types' clusters present in the dataset



The expression profile of the 9 genes in the indicated single cell types in CRC. A. The demonstration of the expression profiles of the genes in each cell type cluster is presented in B. The size of the dot represents the expression percentage, and the intensity of the dot indicates the average expression level. As it can be seen from the graph, the only gene with significant expression in any of these cell types is IGF1, whereas the other genes are not expressed by any of these cell types.

Appendix D. GSEA Sample numbers and the number of significantly enriched Gene Ontology (GO) Gene sets in Up and Down-regulated samples

Sample Location	Sample Group	Sample Number	Number of Significantly Enriched Gene Sets
Distal	Up Regulated	85	7190
	Down regulated	35	273
Proximal	Up Regulated	90	6405
	Down regulated	36	0

Appendix E. Tumor-associated significantly enriched gene sets in Proximal up-regulated samples.

Proximal				
NAME	SIZE	ES	NES	FDR q-val
HP_BENIGN_NEOPLOSM_OF_THE_CENTRAL_NERVOUS_SYSTEM	95	0.443752	2.398115	0.003289
HP_NEOPLOSM_BY_HISTOLOGY	351	0.431271	2.389789	0.003416
HP_NEOPLOSM_OF_THE_GENITOURINARY_TRACT	226	0.340579	2.184255	0.006898
GOBP_RESPONSE_TO_TUMOR_NECROSIS_FACTOR	250	0.531966	2.170651	0.007117
HP_GENITAL_NEOPLOSM	154	0.393398	2.166646	0.007275
HP_HEMATOLOGICAL_NEOPLOSM	273	0.433047	2.163404	0.007356
HP_NEOPLOSM_OF_THE_NERVOUS_SYSTEM	185	0.392654	2.156554	0.007562
HP_NEOPLOSM_OF_THE_GASTROINTESTINAL_TRACT	304	0.383799	2.14366	0.00783
GOBP_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	187	0.605023	2.115407	0.008727
HP_NEOPLOSM_OF_THE_STOMACH	77	0.409342	2.095523	0.009292
HP_NEOPLOSM_OF_THE_LARGE_INTESTINE	110	0.352837	2.073756	0.010013
HP_VASCULAR_NEOPLOSM	146	0.380084	2.056505	0.010424
HP_NEOPLOSM_OF_THE_SMALL_INTESTINE	18	0.622118	2.050376	0.010686
GOBP_NEGATIVE_REGULATION_OF_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	78	0.620907	2.049904	0.010694
HP_FEMALE_REPRODUCTIVE_SYSTEM_NEOPLOSM	62	0.42741	2.048924	0.010712
HP_NEOPLOSM_OF_THE_SKIN	196	0.426837	2.04157	0.011034
HP_FIBROUS_TISSUE_NEOPLOSM	76	0.456157	2.016709	0.012347
GOBP_POSITIVE_REGULATION_OF_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	108	0.622138	2.008947	0.012631
HP_MALIGNANT_NEOPLOSM_OF_THE_CENTRAL_NERVOUS_SYSTEM	87	0.38692	2.008147	0.012653
HP_ESOPHAGEAL_NEOPLOSM	29	0.603296	1.978325	0.014173
HP_NEOPLOSM_OF_THE_PITUITARY_GLAND	60	0.36316	1.96804	0.014833
HP_BLADDER_NEOPLOSM	35	0.435746	1.891954	0.019949
HP_OVARIAN_NEOPLOSM	72	0.374692	1.873524	0.021373
HP_URINARY_TRACT_NEOPLOSM	138	0.312939	1.860308	0.022648
GOMF_TUMOR_NECROSIS_FACTOR_RECEPTOR_SUPERFAMILY_BINDING	48	0.577127	1.854648	0.023248
HP_PROSTATE_NEOPLOSM	34	0.450947	1.845624	0.024141
HP_BILIARY_TRACT_NEOPLOSM	15	0.591509	1.842239	0.024433

Appendix F. Tumor-associated significantly enriched gene sets in Distal up-regulated samples

Distal				
NAME	SIZE	ES	NES	FDR q-val
HP_NEOPLASM_OF_THE_GASTROINTESTINAL_TRACT	304	0.473387	2.484856	0.00066
HP_NEOPLASM_BY_HISTOLOGY	351	0.455286	2.502589	0.000667
HP_NEOPLASM_OF_THE_NERVOUS_SYSTEM	185	0.434729	2.392963	0.000772
HP_FIBROUS_TISSUE_NEOPLASM	76	0.521062	2.333941	0.00099
HP_HEMATOLOGICAL_NEOPLASM	273	0.558049	2.629739	0.001022
GOBP_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	187	0.685122	2.315632	0.001093
GOBP_RESPONSE_TO_TUMOR_CELL	41	0.720501	2.305837	0.001118
GOBP_POSITIVE_REGULATION_OF_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	108	0.716273	2.273598	0.001313
GOBP_RESPONSE_TO_TUMOR_NECROSIS_FACTOR	250	0.550471	2.27028	0.001329
HP_GENITAL_NEOPLASM	154	0.418038	2.253254	0.00145
HP_BENIGN_NEOPLASM_OF_THE_CENTRAL_NERVOUS_SYSTEM	95	0.431709	2.222642	0.001724
HP_MALIGNANT_NEOPLASM_OF_THE_CENTRAL_NERVOUS_SYSTEM	87	0.433863	2.193978	0.002053
HP_NEOPLASM_OF_THE_LIVER	138	0.449468	2.182799	0.002186
GOBP_NEGATIVE_REGULATION_OF_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION	78	0.663057	2.120688	0.003075
HP_NEOPLASM_OF_THE_GENITOURINARY_TRACT	226	0.34338	2.074175	0.004055
HP_BILIARY_TRACT_NEOPLASM	15	0.670375	2.042435	0.004874
HP_NEOPLASM_OF_THE_SKIN	196	0.450116	2.031087	0.005172
HP_NEOPLASM_OF_THE_ENDOCRINE_SYSTEM	150	0.365877	2.022102	0.005424
GOBP_IMMUNE_RESPONSE_TO_TUMOR_CELL	25	0.749701	2.012827	0.005741
GOBP_REGULATION_OF_RESPONSE_TO_TUMOR_CELL	21	0.768214	1.971308	0.007155
HP_GONADAL_NEOPLASM	104	0.39454	1.964397	0.007385
HP_NEOPLASM_OF_THE_THYROID_GLAND	84	0.395773	1.945598	0.00818
HP_NEOPLASM_OF_THE_BREAST	91	0.378878	1.941029	0.008372
HP_VASCULAR_NEOPLASM	146	0.378909	1.935034	0.008638
GOBP_TUMOR_NECROSIS_FACTOR_MEDIATED_SIGNALING_PATHWAY	107	0.427446	1.926739	0.009011
HP_NEOPLASM_OF_THE_RESPIRATORY_SYSTEM	77	0.40807	1.925124	0.009067
HP_PRIMITIVE_NEUROECTODERMAL_TUMOR	59	0.411776	1.902112	0.010276
HP_NEOPLASM_OF_THE_HEART	16	0.793709	1.895875	0.01058
HP_FEMALE_REPRODUCTIVE_SYSTEM_NEOPLASM	62	0.422117	1.886936	0.011133
HP_GASTROINTESTINAL_STROMA_TUMOR	19	0.73426	1.876397	0.011694
HP_OVARIAN_NEOPLASM	72	0.388296	1.867557	0.012226
HP_NEOPLASM_OF_THE_PITUITARY_GLAND	60	0.361055	1.862755	0.012547
HP_NEOPLASM_OF_THE_STOMACH	77	0.393999	1.858724	0.012798
HP_MALE_REPRODUCTIVE_SYSTEM_NEOPLASM	67	0.408655	1.847797	0.01354
HP_ESOPHAGEAL_NEOPLASM	29	0.548043	1.828414	0.015024
GOMF_TUMOR_NECROSIS_FACTOR_RECEPTOR_SUPERFAMILY_BINDING	48	0.583271	1.81565	0.016028
HP_NEOPLASM_OF_THE_LARGE_INTESTINE	110	0.338807	1.782461	0.018963
HP_NEOPLASM_OF_THE_PARATHYROID_GLAND	18	0.586727	1.769522	0.020171
GOMF_TUMOR_NECROSIS_FACTOR_RECEPTOR_BINDING	30	0.658594	1.759645	0.021201

Appendix G. Abnormal morphology associated with significantly enriched gene sets in Proximal up-regulated samples

Proximal				
NAME	SIZE	ES	NES	FDR q-val
HP_ABNORMAL_METAPHYSIS_MORPHOLOGY	223	0.510956	2.758799	0.000849
HP_ABNORMAL_FEMORAL_NECK_MORPHOLOGY	146	0.57015	2.691655	0.001945
HP_ABNORMAL_EPIPHYSIS_MORPHOLOGY	270	0.48268	2.620441	0.002077
HP_ABNORMAL_STERNUM_MORPHOLOGY	446	0.47646	2.594059	0.002601
HP_ABNORMAL_THUMB_MORPHOLOGY	382	0.472095	2.586267	0.002648
HP_ABNORMAL_FRONTAL_BONE_MORPHOLOGY	391	0.433354	2.431285	0.002769
HP_ABNORMAL_PLEURA_MORPHOLOGY	170	0.626155	2.422728	0.002885
HP_ABNORMAL_ACETABULUM_MORPHOLOGY	87	0.546098	2.436092	0.002922
HP_ABNORMAL_UPPER_LIMB_BONE_MORPHOLOGY	443	0.510358	2.602542	0.002948
HP_ABNORMAL_AORTIC_MORPHOLOGY	363	0.43701	2.4363	0.002952
HP_ABNORMAL_NASAL_TIP_MORPHOLOGY	337	0.419653	2.469358	0.002995
HP_ABNORMAL_COLUMELLA_MORPHOLOGY	85	0.513204	2.477538	0.003019
HP_ABNORMAL_AORTIC_VALVE_MORPHOLOGY	168	0.470517	2.469923	0.003023
HP_ABNORMAL_CORTICAL_BONE_MORPHOLOGY	80	0.519019	2.407974	0.003062
HP_ABNORMAL_STOMACH_MORPHOLOGY	265	0.453034	2.409471	0.003062
HP_ABNORMAL_FEMORAL_NECK_HEAD_MORPHOLOGY	187	0.507612	2.577316	0.003067
HP_ABNORMAL_RENAL_PELVIS_MORPHOLOGY	240	0.498168	2.453033	0.003068
HP_ABNORMAL_DISTAL_PHALANX_MORPHOLOGY_OF_FINGER	161	0.571175	2.470264	0.003069
HP_ABNORMAL_CEREBRAL_ARTERY_MORPHOLOGY	73	0.642114	2.445023	0.003086
HP_ABNORMAL_SCAPULA_MORPHOLOGY	167	0.556473	2.445378	0.003122
HP_ABNORMALITY_OF_FEMUR_MORPHOLOGY	279	0.469046	2.529359	0.003135
HP_ABNORMAL_LIMB_EPIPHYSIS_MORPHOLOGY	129	0.533811	2.403584	0.003144
HP_ABNORMAL_ADIPOSE_TISSUE_MORPHOLOGY	231	0.483411	2.56991	0.003241
HP_ABNORMAL_SYSTEMIC_ARTERIAL_MORPHOLOGY	493	0.432862	2.499389	0.00329
HP_ABNORMAL_NASAL_DORSUM_MORPHOLOGY	238	0.453678	2.395747	0.003317
HP_ABNORMAL_HEART_VALVE_MORPHOLOGY	340	0.48078	2.39602	0.003328
HP_ABNORMAL_ILIUM_MORPHOLOGY	101	0.56041	2.383898	0.003363
HP_ABNORMAL_FOREARM_MORPHOLOGY	234	0.513185	2.382412	0.003368
HP_ABNORMAL_5TH_FINGER_MORPHOLOGY	468	0.394367	2.382921	0.003383
HP_ABNORMAL_MORPHOLOGY_OF_ULNA	158	0.494399	2.380684	0.003387
HP_ABNORMAL_METATARSAL_MORPHOLOGY	149	0.547843	2.386598	0.003389
HP_ABNORMAL_FETAL_MORPHOLOGY	269	0.390356	2.39113	0.003428
HP_ABNORMAL_AORTIC_VALVE_CUSP_MORPHOLOGY	111	0.542858	2.372873	0.003509
HP_ABNORMAL_SOFT_PALATE_MORPHOLOGY	243	0.429182	2.370569	0.003547
HP_ABNORMAL_CAROTID_ARTERY_MORPHOLOGY	89	0.50994	2.353651	0.003695
HP_ABNORMAL_DIAPHYSIS_MORPHOLOGY	315	0.450696	2.33391	0.003824
HP_ABNORMAL_CEREBRAL_VASCULAR_MORPHOLOGY	329	0.438183	2.334538	0.003825
HP_ABNORMAL_CEREBELLAR_VERMIS_MORPHOLOGY	334	0.379432	2.339875	0.003827
HP_ABNORMAL_METACARPAL_MORPHOLOGY	218	0.491352	2.319529	0.004019
HP_ABNORMAL_EARLOBE_MORPHOLOGY	115	0.443624	2.316099	0.004053
HP_ABNORMAL_ATRIOVENTRICULAR_VALVE_MORPHOLOGY	166	0.513903	2.31651	0.004062
HP_ABNORMAL_EYELASH_MORPHOLOGY	265	0.44377	2.306438	0.004241
HP_ABNORMAL_ORAL_MUCOSA_MORPHOLOGY	340	0.528788	2.303169	0.004273
HP_ABNORMAL_RIB_CAGE_MORPHOLOGY	377	0.47046	2.297945	0.00435
HP_ABNORMAL_CHEEK_MORPHOLOGY	250	0.423494	2.29837	0.004351
HP_ABNORMAL_ANTIHELIX_MORPHOLOGY	40	0.637244	2.298734	0.004353
HP_ABNORMALITY_IRIS_MORPHOLOGY	426	0.451038	2.292754	0.004464
HP_ABNORMAL_DUODENUM_MORPHOLOGY	86	0.517025	2.291697	0.004466

HP_ABNORMAL_PHALANGEAL_JOINT_MORPHOLOGY_OF_THE_HAND	211	0.52158	2.291099	0.004475
HP_ABNORMAL_SUBARACHNOID_SPACE_MORPHOLOGY	84	0.541307	2.286809	0.004478
HP_ABNORMAL_SCALP_MORPHOLOGY	411	0.357752	2.288022	0.004501
HP_ABNORMAL_ZYGOMATIC_BONE_MORPHOLOGY	242	0.457006	2.287141	0.004505
HP_ABNORMAL_UVULA_MORPHOLOGY	184	0.432191	2.281076	0.004618
HP_ABNORMAL_TRICUSPID_VALVE_MORPHOLOGY	43	0.611451	2.276565	0.004744
HP_ABNORMALITY_OF_DENTAL_MORPHOLOGY	244	0.437583	2.267421	0.005016
HP_ABNORMAL_LOWER_LIP_MORPHOLOGY	235	0.463405	2.260655	0.005118
HP_ABNORMAL_ANTERIOR_CHAMBER_MORPHOLOGY	56	0.665862	2.242322	0.005597
HP_ABNORMAL_CLAVICLE_MORPHOLOGY	107	0.512127	2.22657	0.006127
HP_ABNORMAL_RIB_MORPHOLOGY	311	0.474219	2.224097	0.006128
HP_ABNORMAL_LARGE_INTESTINE_MORPHOLOGY	291	0.442155	2.226737	0.006135
HP_ABNORMAL_MENINGEAL_MORPHOLOGY	188	0.46721	2.224797	0.006137
HP_ABNORMAL_VERTEBRAL_MORPHOLOGY	468	0.433904	2.220657	0.006203
HP_ABNORMAL_HELIX_MORPHOLOGY	168	0.424899	2.221561	0.00625
HP_ABNORMAL_CORNEAL_STROMA_MORPHOLOGY	98	0.544795	2.215179	0.006315
HP_ABNORMAL_MYELOID_LEUKOCYTE_MORPHOLOGY	356	0.466895	2.211515	0.006453
HP_ABNORMALITY_OF_UPPER_LIMB_EPIPHYSIS_MORPHOLOGY	57	0.613977	2.208625	0.006524
HP_ABNORMAL_LABIA_MORPHOLOGY	84	0.529276	2.199677	0.006709
HP_ABNORMAL_CONJUNCTIVA_MORPHOLOGY	154	0.519407	2.193256	0.006717
HP_ABNORMAL_MIDDLE_PHALANX_MORPHOLOGY_OF_THE_HAND	63	0.551872	2.182237	0.006902
HP_ABNORMAL_SCLERA_MORPHOLOGY	125	0.482566	2.183543	0.006904
HP_ABNORMAL_EPIPHYSIS_MORPHOLOGY_OF_THE_PHALANGES_OF_THE_HAND	39	0.651481	2.184833	0.006934
HP_ABNORMALITY_OF_ENTERIC_NERVOUS_SYSTEM_MORPHOLOGY	124	0.437996	2.176685	0.007013
HP_ABNORMAL_ORAL_FRENULUM_MORPHOLOGY	61	0.555703	2.17027	0.007132
HP_ABNORMAL_UMBILICUS_MORPHOLOGY	295	0.427375	2.164554	0.007327
HP_ABNORMAL_RECTUM_MORPHOLOGY	135	0.444898	2.16168	0.007407
HP_ABNORMAL_UTERUS_MORPHOLOGY	134	0.435874	2.159538	0.007473
HP_ABNORMAL_ILIAC_WING_MORPHOLOGY	67	0.569971	2.151985	0.007616
HP_ABNORMAL_SACRUM_MORPHOLOGY	154	0.480693	2.150274	0.007688
HP_ABNORMAL_URETER_MORPHOLOGY	148	0.416906	2.147856	0.007789
HP_ABNORMAL_MORPHOLOGY_OF_THE_MUSCULATURE_OF_THE_NECK	74	0.580888	2.144935	0.007816
HP_ABNORMAL_ESOPHAGUS_MORPHOLOGY	224	0.434131	2.140497	0.007918
HP_ABNORMAL_ASCENDING_AORTA_MORPHOLOGY	26	0.772611	2.138179	0.007964
HP_ABNORMAL_SKIN_MORPHOLOGY_OF_THE_PALM	422	0.418597	2.124762	0.008436
HP_ABNORMAL_CARPAL_MORPHOLOGY	85	0.549958	2.121251	0.008556
HP_ABNORMAL_SHOULDER_MORPHOLOGY	97	0.53516	2.112963	0.008776
HP_ABNORMAL_LYMPHOCYTE_MORPHOLOGY	219	0.560026	2.109021	0.008941
HP_ABNORMAL_LATERAL_VENTRICLE_MORPHOLOGY	101	0.336749	2.090306	0.009374
HP_ABNORMAL_CORONARY_ARTERY_MORPHOLOGY	67	0.555491	2.086909	0.009482
HP_ABNORMAL_LOWER_LIMB_METAPHYSIS_MORPHOLOGY	21	0.625625	2.07656	0.009861
HP_ABNORMAL_MALE_INTERNAL_GENITALIA_MORPHOLOGY	77	0.501272	2.076595	0.009871
HP_ABNORMAL_LABIA_MAJORA_MORPHOLOGY	44	0.537865	2.065229	0.010222
HP_ABNORMAL_LABIA_MINORA_MORPHOLOGY	35	0.61681	2.055958	0.010442
HP_ABNORMAL_THORACIC_SPINE_MORPHOLOGY	102	0.436598	2.05385	0.010512
HP_ABNORMAL_CRANIAL_NERVE_MORPHOLOGY	246	0.442925	2.050031	0.010694
HP_ABNORMAL_FEMORAL_HEAD_MORPHOLOGY	90	0.443391	2.041886	0.01103
HP_ABNORMAL_SELLA_TURCICA_MORPHOLOGY	38	0.53925	2.03318	0.011562
HP_ABNORMALITY_OF_BLADDER_MORPHOLOGY	75	0.533468	2.032279	0.011612
HP_ABNORMAL_PUPIL_MORPHOLOGY	64	0.545919	2.029919	0.011674

HP_ABNORMAL_NIPPLE_MORPHOLOGY	237	0.299792	2.011725	0.012474
HP_ABNORMAL_MALE_URETHRAL_MEATUS_MORPHOLOGY	376	0.327254	2.005242	0.012803
HP_ABNORMAL_CORNEAL_EPITHELIUM_MORPHOLOGY	109	0.540713	2.003716	0.012882
HP_ABNORMAL_UPPER_LIMB_METAPHYSIS_MORPHOLOGY	17	0.791464	2.001598	0.012995
HP_ABNORMAL_SPINAL_MENINGEAL_MORPHOLOGY	23	0.786078	2.001207	0.013009
HP_ABNORMAL_HALLUX_PHALANX_MORPHOLOGY	40	0.644931	1.998927	0.013097
HP_ABNORMAL_TEMPORAL_LOBE_MORPHOLOGY	140	0.391743	1.996804	0.013207
HP_ABNORMAL_PULMONARY_VALVE_MORPHOLOGY	56	0.476946	1.9927	0.01336
HP_ABNORMAL_LARYNX_MORPHOLOGY	142	0.426301	1.991561	0.013379
HP_ABNORMAL_MORPHOLOGY_OF_THE ABDOMINAL MUSCULATURE	59	0.480355	1.990717	0.013412
HP_ABNORMAL_THYROID_MORPHOLOGY	169	0.42592	1.98556	0.013794
HP_ABNORMAL_BASAL_GANGLIA_MORPHOLOGY	197	0.362302	1.984331	0.013845
HP_ABNORMAL_PHARYNX_MORPHOLOGY	213	0.451616	1.981837	0.013952
HP_ABNORMAL_RENAL_CORTX_MORPHOLOGY	223	0.4035	1.976731	0.01425
HP_ABNORMAL_RENAL_CORPUSCLE_MORPHOLOGY	184	0.43243	1.976524	0.01425
HP_ABNORMAL_FORAMEN_MAGNUM_MORPHOLOGY	17	0.608405	1.964187	0.015052
HP_ABNORMAL_RIGHT_VENTRICLE_MORPHOLOGY	65	0.515714	1.961173	0.015109
HP_ABNORMAL_PERICARDIUM_MORPHOLOGY	76	0.580468	1.957694	0.015342
HP_ABNORMAL_NEPHRON_MORPHOLOGY	251	0.390636	1.954066	0.015546
HP_ABNORMAL_CHORIORETINAL_MORPHOLOGY	155	0.399114	1.952681	0.015665
HP_ABNORMAL_CHOROID_MORPHOLOGY	191	0.41918	1.951762	0.015738
HP_ABNORMALITY_OF_LOWER_LIMB_EPIPHYSIS_MORPHOLOGY	91	0.446242	1.951613	0.015741
HP_ABNORMAL_TOENAIL_MORPHOLOGY	165	0.3806	1.947401	0.016004
HP_ABNORMAL_GLIAL_CELL_MORPHOLOGY	150	0.2903	1.940975	0.016417
HP_ABNORMAL_TONGUE_MORPHOLOGY	330	0.350453	1.937467	0.016606
HP_ABNORMAL_FINGERTIP_MORPHOLOGY	82	0.451756	1.93437	0.016772
HP_ABNORMAL_PROSTATE_MORPHOLOGY	42	0.483285	1.933935	0.016778
HP_ABNORMAL_LYMPH_NODE_MORPHOLOGY	222	0.519245	1.93289	0.016847
HP_ABNORMAL_DURA_MATER_MORPHOLOGY	26	0.608857	1.929924	0.017068
HP_ABNORMAL_PERIAURICULAR_REGION_MORPHOLOGY	106	0.44022	1.918039	0.017812
HP_ABNORMAL_TRACHEAL_MORPHOLOGY	162	0.400009	1.911352	0.018326
HP_ABNORMAL_INCISOR_MORPHOLOGY	55	0.563251	1.907424	0.018593
HP_ABNORMALITY_OF_TIBIA_MORPHOLOGY	97	0.50192	1.904038	0.018891
HP_ABNORMAL_MORPHOLOGY_OF_THE CHOROIDAL VASCULATURE	25	0.68547	1.902777	0.018971
HP_ABNORMALITY_OF_FIBULA_MORPHOLOGY	55	0.524301	1.899032	0.019336
HP_ABNORMAL_PARANASAL_SINUS_MORPHOLOGY	179	0.483754	1.89365	0.019795
HP_ABNORMAL_B_CELL_MORPHOLOGY	74	0.568812	1.890632	0.020031
HP_ABNORMAL_FINGERNAIL_MORPHOLOGY	191	0.399996	1.888791	0.020143
HP_ABNORMALITY_OF_BONE_MARROW_CELL_MORPHOLOGY	205	0.397802	1.884947	0.020431
HP_ABNORMAL_PROXIMAL_PHALANX_MORPHOLOGY_OF_THE HAND	28	0.608721	1.881051	0.02074
HP_ABNORMAL_TRACHEOBRONCHIAL_MORPHOLOGY	355	0.38922	1.875419	0.021288
HP_ABNORMAL_GREATER_SCIATIC_NOTCH_MORPHOLOGY	20	0.601472	1.875014	0.021314
HP_ABNORMAL_EPIGLOTTIS_MORPHOLOGY	22	0.665135	1.859556	0.022695
HP_ABNORMAL_METACARPOPHALANGEAL_JOINT_MORPHOLOGY	25	0.644528	1.848744	0.02384
HP_ABNORMAL_NEURAL_TUBE_MORPHOLOGY	309	0.337176	1.844397	0.024237
HP_ABNORMAL_RENAL_VASCULAR_MORPHOLOGY	29	0.508655	1.844462	0.024238
HP_ABNORMAL_NATURAL_KILLER_CELL_MORPHOLOGY	30	0.569017	1.843655	0.024257
HP_ABNORMAL_CAUDATE_NUCLEUS_MORPHOLOGY	27	0.565257	1.835896	0.024993
HP_ABNORMAL_PATELLA_MORPHOLOGY	103	0.3952	1.834155	0.025216
HP_ABNORMAL_T_CELL_MORPHOLOGY	90	0.552807	1.827132	0.025969

HP_ABNORMALITY_OF_ADRENAL_MORPHOLOGY	106	0.384931	1.8194	0.026892
HP_ABNORMAL_CORPUS_STRIATUM_MORPHOLOGY	43	0.474685	1.815528	0.027257
HP_ABNORMAL_4TH_METACARPAL_MORPHOLOGY	22	0.599168	1.810811	0.027865
HP_ABNORMAL_MIDBRAIN_MORPHOLOGY	84	0.356729	1.807818	0.02823
HP_ABNORMAL_MUSCLE_FIBER_MORPHOLOGY	252	0.376397	1.798675	0.029209
HP_ABNORMAL_OCCIPITAL_BONE_MORPHOLOGY	114	0.323427	1.79174	0.030064
HP_ABNORMAL_ARACHNOID_MATER_MORPHOLOGY	51	0.365847	1.785162	0.030904
HP_ABNORMAL_5TH_FINGER_PHALANX_MORPHOLOGY	28	0.577881	1.784689	0.030999
HP_ABNORMAL_MACULAR_MORPHOLOGY	207	0.389626	1.782066	0.031206
HP_ABNORMAL_1ST_METACARPAL_MORPHOLOGY	37	0.506748	1.775467	0.032081
HP_ABNORMALITY_OF_THALAMUS_MORPHOLOGY	52	0.424354	1.765429	0.033503
HP_ABNORMAL_CORNEAL_ENDOTHELIUM_MORPHOLOGY	18	0.666115	1.75254	0.035376
HP_ABNORMAL_HYPOTHALAMUS_MORPHOLOGY	21	0.606747	1.74923	0.0357
HP_ABNORMAL_EOSINOPHIL_MORPHOLOGY	67	0.517044	1.737867	0.037272
HP_ABNORMAL_FETAL_CARDIOVASCULAR_MORPHOLOGY	38	0.397248	1.722801	0.03932
HP_ABNORMAL_LEFT_VENTRICLE_MORPHOLOGY	125	0.37831	1.719394	0.039759
HP_ABNORMAL_OVARIAN_MORPHOLOGY	165	0.334777	1.706964	0.041916
HP_ABNORMAL_SEPTUM_PELLUCIDUM_MORPHOLOGY	103	0.359605	1.69873	0.043246
HP_ABNORMAL_ODONTOID_TISSUE_MORPHOLOGY	220	0.325593	1.695914	0.043832
HP_ABNORMAL_CARTILAGE_MORPHOLOGY	46	0.414384	1.690314	0.044742
HP_ABNORMAL_MACROPHAGE_MORPHOLOGY	36	0.657443	1.671867	0.04823
HP_ABNORMAL_AORTIC_ARCH_MORPHOLOGY	62	0.422183	1.661016	0.050465
HP_ABNORMAL_PLACENTA_MORPHOLOGY	37	0.492183	1.65066	0.052535
HP_ABNORMAL_GLOMERULAR_MESANGIUM_MORPHOLOGY	47	0.383046	1.638501	0.055362
HP_ABNORMAL_BRONCHUS_MORPHOLOGY	198	0.398124	1.630023	0.057336
HP_ABNORMAL_HEAD_BLOOD_VESSEL_MORPHOLOGY	77	0.37688	1.618416	0.05997
HP_ABNORMAL_HARD_PALATE_MORPHOLOGY	16	0.532909	1.615135	0.06091
HP_ABNORMAL_SUBSTANTIA_NIGRA_MORPHOLOGY	25	0.498472	1.61403	0.061121
HP_ABNORMAL_VAS_DEFERENS_MORPHOLOGY	21	0.584259	1.608688	0.062562
HP_ABNORMAL_MONOCYTE_MORPHOLOGY	18	0.590032	1.598016	0.065169
HP_ABNORMALITY_OF_NEUTROPHIL_MORPHOLOGY	16	0.510046	1.597694	0.06522
HP_ABNORMAL_BRAINSTEM_MORPHOLOGY	262	0.249417	1.59343	0.066382
HP_ABNORMAL_MORPHOLOGY_OF_THE_OLFACTORY_BULB	31	0.488378	1.574335	0.071283
HP_ABNORMAL_VAGINA_MORPHOLOGY	103	0.383894	1.56977	0.072571
HP_ABNORMAL_INTERVERTEBRAL_DISK_MORPHOLOGY	43	0.418011	1.550671	0.078029
HP_ABNORMAL_TALUS_MORPHOLOGY	67	0.368229	1.548325	0.078707
HP_ABNORMAL_ENDOCARDIUM_MORPHOLOGY	64	0.399132	1.53833	0.081816
HP_ABNORMAL_NEURON_MORPHOLOGY	85	0.275449	1.531847	0.083932
HP_ABNORMALITY_OF_THE_PARATHYROID_MORPHOLOGY	22	0.516584	1.53052	0.084301
HP_ABNORMAL_RENAL_COLLECTING_SYSTEM_MORPHOLOGY	28	0.391009	1.530021	0.084513
HP_ABNORMAL_PANCREAS_MORPHOLOGY	105	0.33381	1.513714	0.08969
HP_ABNORMAL_NASOLACRIMAL_SYSTEM_MORPHOLOGY	67	0.349289	1.50407	0.092993
HP_ABNORMAL_PERIVENTRICULAR_WHITE_MATTER_MORPHOLOGY	123	0.267225	1.496106	0.095868
HP_ABNORMAL_MORPHOLOGY_OF_THE_LIMBIC_SYSTEM	43	0.456809	1.494752	0.096379
HP_ABNORMAL_5TH_METACARPAL_MORPHOLOGY	20	0.480739	1.49027	0.098008
HP_ABNORMAL_DENTAL_PULP_MORPHOLOGY	52	0.382941	1.4901	0.098071
HP_ABNORMAL_SHARPEY_FIBER_MORPHOLOGY	33	0.497981	1.489144	0.098455
HP_ABNORMAL_3RD_FINGER_MORPHOLOGY	23	0.504536	1.480891	0.101544
HP_ABNORMAL_RETINAL_ARTERY_MORPHOLOGY	64	0.357758	1.480075	0.101837
HP_ABNORMAL_PONS_MORPHOLOGY	105	0.262113	1.477907	0.10263

HP_ABNORMAL_4TH_FINGER_MORPHOLOGY	20	0.519395	1.456607	0.111122
HP_ABNORMAL_LUMBAR_SPINE_MORPHOLOGY	32	0.365126	1.449828	0.113882
HP_ABNORMAL_PELVIS_BONE_MORPHOLOGY	22	0.423844	1.438192	0.118868
HP_ABNORMAL_LEFT_VENTRICULAR_OUTFLOW_TRACT_MORPHOLOGY	26	0.453488	1.426333	0.123943
HP_ABNORMAL_EPIDERMAL_MORPHOLOGY	88	0.404634	1.424917	0.124555
HP_ABNORMAL_RENAL_INSTERSTITIAL_MORPHOLOGY	63	0.309803	1.422766	0.125362
HP_ABNORMAL_RETINAL_VASCULAR_MORPHOLOGY	241	0.25765	1.407436	0.13242
HP_ABNORMAL_VENOUS_MORPHOLOGY	151	0.287103	1.399966	0.135957
HP_ABNORMAL_RENAL_TUBULE_MORPHOLOGY	101	0.299339	1.3968	0.137305
HP_ABNORMAL_FOVEAL_MORPHOLOGY	56	0.426915	1.376377	0.147441
HP_ABNORMALITY_OF_SMALL_INTESTINAL_VILLUS_MORPHOLOGY	28	0.377643	1.36083	0.15521
HP_ABNORMAL_COCHLEA_MORPHOLOGY	30	0.445775	1.349853	0.16091
HP_ABNORMAL_ENDOMETRIUM_MORPHOLOGY	24	0.34859	1.349358	0.161227
HP_ABNORMAL_SALIVARY_GLAND_MORPHOLOGY	19	0.442216	1.346853	0.162415
HP_ABNORMAL_FALLOPIAN_TUBE_MORPHOLOGY	27	0.367038	1.346229	0.1627
HP_ABNORMAL_METACARPAL_EPIPHYSIS_MORPHOLOGY	16	0.496779	1.31686	0.178879
HP_ABNORMAL_RIGHT_ATRIUM_MORPHOLOGY	21	0.463645	1.312237	0.181615
HP_ABNORMAL_PARIETAL_BONE_MORPHOLOGY	27	0.325106	1.304387	0.186437
HP_ABNORMAL_VITREOUS_HUMOR_MORPHOLOGY	46	0.330828	1.301246	0.188473
HP_ABNORMAL_ILEUM_MORPHOLOGY	37	0.306936	1.289357	0.195834
HP_ABNORMAL_LOWER_MOTOR_NEURON_MORPHOLOGY	26	0.321508	1.283468	0.199375
HP_ABNORMAL_RENAL_MEDULLA_MORPHOLOGY	59	0.237719	1.278377	0.202673
HP_ABNORMAL_LACRIMAL_DUCT_MORPHOLOGY	42	0.351717	1.277814	0.203015
HP_ABNORMAL_CERVICAL_SPINE_MORPHOLOGY	20	0.411186	1.272893	0.206076
HP_ABNORMAL_2ND_FINGER_MORPHOLOGY	38	0.338993	1.263547	0.21206
HP_ABNORMAL_BRONCHOALVEOLAR_LAVAGE_FLUID_MORPHOLOGY	16	0.552104	1.259743	0.214643
HP_ABNORMAL_PULMONARY_INTERSTITIAL_MORPHOLOGY	150	0.306921	1.239959	0.228171
HP_ABNORMAL_GASTRIC_MUCOSA_MORPHOLOGY	50	0.326881	1.211126	0.248525

Appendix H. Abnormal morphology associated with significantly enriched gene sets in Distal up-regulated samples

Distal				
NAME	SIZE	ES	NES	FDR q-val
HP_ABNORMAL_NEPHRON_MORPHOLOGY	251	0.484333	2.472956	0.000624
HP_ABNORMALITY_OF_FEMUR_MORPHOLOGY	279	0.480936	2.473909	0.000634
HP_ABNORMAL_DIAPHYSIS_MORPHOLOGY	315	0.500036	2.469871	0.000634
HP_ABNORMAL_STOMACH_MORPHOLOGY	265	0.488379	2.481374	0.000644
HP_ABNORMAL_NASAL_DORSUM_MORPHOLOGY	238	0.465373	2.463238	0.000651
HP_ABNORMALITY_OF_BONE_MARROW_CELL_MORPHOLOGY	205	0.52543	2.463756	0.000655
HP_ABNORMAL_LARGE_INTESTINE_MORPHOLOGY	291	0.510657	2.445282	0.000661
HP_ABNORMAL_TRACHEOBRONCHIAL_MORPHOLOGY	355	0.52306	2.455879	0.000662
HP_ABNORMAL_ADIPOSE_TISSUE_MORPHOLOGY	231	0.499212	2.493473	0.000663
HP_ABNORMAL_METATARSAL_MORPHOLOGY	149	0.574576	2.45669	0.000663
HP_ABNORMAL_PLEURA_MORPHOLOGY	170	0.63635	2.491053	0.000664
HP_ABNORMAL_AORTIC_MORPHOLOGY	363	0.446913	2.448742	0.000665
HP_ABNORMAL_CORNEAL_EPITHELIUM_MORPHOLOGY	109	0.651374	2.466241	0.000665
HP_ABNORMAL_UVULA_MORPHOLOGY	184	0.455182	2.456895	0.000665
HP_ABNORMAL_SYSTEMIC_ARTERIAL_MORPHOLOGY	493	0.454222	2.498566	0.000667
HP_ABNORMAL_AORTIC_VALVE_MORPHOLOGY	168	0.476178	2.492128	0.000669
HP_ABNORMAL_LYMPHOCYTE_MORPHOLOGY	219	0.661762	2.494743	0.00067
HP_ABNORMAL_PHALANGEAL_JOINT_MORPHOLOGY_OF_THE_HAND	211	0.575599	2.440809	0.000674
HP_ABNORMAL_DISTAL_PHALANX_MORPHOLOGY_OF_FINGER	161	0.588729	2.586134	0.000675
HP_ABNORMAL_NASAL_TIP_MORPHOLOGY	337	0.424359	2.496855	0.000678
HP_ABNORMAL_CORTICAL_BONE_MORPHOLOGY	80	0.562991	2.579772	0.000685
HP_ABNORMAL_CRANIAL_NERVE_MORPHOLOGY	246	0.529698	2.42061	0.00069
HP_ABNORMAL_PHARYNX_MORPHOLOGY	213	0.604673	2.591805	0.000692
HP_ABNORMAL_EPIPHYSIS_MORPHOLOGY	270	0.481386	2.431163	0.000697
HP_ABNORMAL_ORAL_MUCOSA_MORPHOLOGY	340	0.610362	2.592229	0.000702
HP_ABNORMAL_URETER_MORPHOLOGY	148	0.482799	2.51009	0.000707
HP_ABNORMAL_METAPHYSIS_MORPHOLOGY	223	0.485328	2.425556	0.000711
HP_ABNORMAL_UPPER_LIMB_BONE_MORPHOLOGY	443	0.514015	2.513941	0.000714
HP_ABNORMAL_METACARPAL_MORPHOLOGY	218	0.536244	2.41844	0.000718
HP_ABNORMAL_HEART_VALVE_MORPHOLOGY	340	0.498426	2.514529	0.00072
HP_ABNORMALITY_IRIS_MORPHOLOGY	426	0.497166	2.545667	0.000734
HP_ABNORMAL_RENAL_CORTEX_MORPHOLOGY	223	0.507385	2.524412	0.000736
HP_ABNORMAL_RENAL_CORPUSCLE_MORPHOLOGY	184	0.553586	2.534249	0.000738
HP_ABNORMAL_THUMB_MORPHOLOGY	382	0.46805	2.406726	0.000741
HP_ABNORMAL_PARANASAL_SINUS_MORPHOLOGY	179	0.609476	2.413279	0.000746
HP_ABNORMAL_COLUMELLA_MORPHOLOGY	85	0.529315	2.540035	0.000748
HP_ABNORMAL_AORTIC_VALVE_CUSP_MORPHOLOGY	111	0.551642	2.409412	0.000749
HP_ABNORMAL_ESOPHAGUS_MORPHOLOGY	224	0.488789	2.408465	0.000754
HP_ABNORMAL_SUBARACHNOID_SPACE_MORPHOLOGY	84	0.58186	2.59483	0.000754
HP_ABNORMAL_CHOROID_MORPHOLOGY	191	0.538516	2.54027	0.000758
HP_ABNORMAL_RENAL_PELVIS_MORPHOLOGY	240	0.530695	2.550176	0.000759
HP_ABNORMAL_CONJUNCTIVA_MORPHOLOGY	154	0.615137	2.552909	0.000765
HP_ABNORMAL_STERNUM_MORPHOLOGY	446	0.471254	2.399365	0.000773
HP_ABNORMAL_LOWER_LIP_MORPHOLOGY	235	0.489268	2.389688	0.000785
HP_ABNORMAL_CHORIORETINAL_MORPHOLOGY	155	0.501339	2.564004	0.000788
HP_ABNORMAL_FRONTAL_BONE_MORPHOLOGY	391	0.439799	2.386957	0.00079
HP_ABNORMAL_ORAL_FRENULUM_MORPHOLOGY	61	0.60067	2.379933	0.000811
HP_ABNORMAL_RIB_MORPHOLOGY	311	0.533209	2.379792	0.000814

HP_ABNORMAL_PERICARDIUM_MORPHOLOGY	76	0.715946	2.378115	0.000821
HP_ABNORMAL_THYROID_MORPHOLOGY	169	0.517225	2.374718	0.000839
HP_ABNORMAL_UMBILICUS_MORPHOLOGY	295	0.465722	2.369016	0.00085
HP_ABNORMAL_SKIN_MORPHOLOGY_OF_THE_PALM	422	0.458781	2.362804	0.000892
HP_ABNORMAL_TEMPORAL_LOBE_MORPHOLOGY	140	0.46625	2.359767	0.000907
HP_ABNORMAL_MACULAR_MORPHOLOGY	207	0.538493	2.356844	0.000911
HP_ABNORMAL_ANTERIOR_CHAMBER_MORPHOLOGY	56	0.713656	2.358166	0.000919
HP_ABNORMAL_MYELOID_LEUKOCYTE_MORPHOLOGY	356	0.573468	2.617266	0.000921
HP_ABNORMAL_LYMPH_NODE_MORPHOLOGY	222	0.623315	2.355299	0.000924
HP_ABNORMAL_B_CELL_MORPHOLOGY	74	0.721428	2.34994	0.000955
HP_ABNORMAL_RIB_CAGE_MORPHOLOGY	377	0.501056	2.345497	0.000957
HP_ABNORMAL_BRONCHUS_MORPHOLOGY	198	0.562525	2.343482	0.000961
HP_ABNORMAL_FEMORAL_NECK_HEAD_MORPHOLOGY	187	0.493353	2.346248	0.000961
HP_ABNORMAL_SOFT_PALATE_MORPHOLOGY	243	0.421677	2.332019	0.000992
HP_ABNORMAL_MENINGEAL_MORPHOLOGY	188	0.47647	2.33205	0.000993
HP_ABNORMAL_FINGERTIP_MORPHOLOGY	82	0.53163	2.330567	0.001002
HP_ABNORMAL_5TH_FINGER_MORPHOLOGY	468	0.399385	2.318434	0.001077
HP_ABNORMAL_CHEEK_MORPHOLOGY	250	0.425355	2.308825	0.001116
HP_ABNORMAL_CEREBELLAR_VERMIS_MORPHOLOGY	334	0.424909	2.643799	0.00113
HP_ABNORMAL_GLOMERULAR_MESANGIUM_MORPHOLOGY	47	0.542295	2.300975	0.001158
HP_ABNORMAL_FETAL_MORPHOLOGY	269	0.419674	2.299184	0.001161
HP_ABNORMAL_MUSCLE_FIBER_MORPHOLOGY	252	0.467603	2.29814	0.001169
HP_ABNORMAL_FEMORAL_NECK_MORPHOLOGY	146	0.528665	2.296695	0.001182
HP_ABNORMAL_EYELASH_MORPHOLOGY	265	0.451935	2.293994	0.001202
HP_ABNORMAL_CORNEAL_STROMA_MORPHOLOGY	98	0.594619	2.289442	0.001207
HP_ABNORMAL_LIMB_EPIPHYSIS_MORPHOLOGY	129	0.524298	2.284642	0.001238
HP_ABNORMAL_T_CELL_MORPHOLOGY	90	0.678641	2.285763	0.001247
HP_ABNORMAL_PUPIL_MORPHOLOGY	64	0.612098	2.28347	0.00125
HP_ABNORMAL_FOREARM_MORPHOLOGY	234	0.503654	2.27348	0.001311
HP_ABNORMAL_SCAPULA_MORPHOLOGY	167	0.527038	2.269177	0.001342
HP_ABNORMAL_NIPPLE_MORPHOLOGY	237	0.354669	2.265771	0.001363
HP_ABNORMAL_ZYGOMATIC_BONE_MORPHOLOGY	242	0.468814	2.263053	0.001375
HP_ABNORMAL_CEREBRAL_ARTERY_MORPHOLOGY	73	0.57374	2.259471	0.001403
HP_ABNORMALITY_OF_ENTERIC_NERVOUS_SYSTEM_MORPHOLOGY	124	0.444303	2.25138	0.001466
HP_ABNORMAL_ILIUM_MORPHOLOGY	101	0.542387	2.248137	0.001498
HP_ABNORMAL_ATRIOVENTRICULAR_VALVE_MORPHOLOGY	166	0.518536	2.24286	0.001525
HP_ABNORMAL_ACETABULUM_MORPHOLOGY	87	0.514518	2.236806	0.001575
HP_ABNORMAL_PULMONARY_INTERSTITIAL_MORPHOLOGY	150	0.523791	2.231109	0.001646
HP_ABNORMAL_VERTEBRAL_MORPHOLOGY	468	0.455091	2.230347	0.001657
HP_ABNORMAL_THORACIC_SPINE_MORPHOLOGY	102	0.483517	2.21897	0.001762
HP_ABNORMAL_EARLOBE_MORPHOLOGY	115	0.457919	2.217807	0.001778
HP_ABNORMAL_DUODENUM_MORPHOLOGY	86	0.47221	2.217006	0.00179
HP_ABNORMALITY_OF_DENTAL_MORPHOLOGY	244	0.450239	2.213514	0.001828
HP_ABNORMAL_NATURAL_KILLER_CELL_MORPHOLOGY	30	0.691175	2.211225	0.001866
HP_ABNORMAL_ENDOCARDIUM_MORPHOLOGY	64	0.577264	2.205881	0.001916
HP_ABNORMAL_TONGUE_MORPHOLOGY	330	0.400642	2.205403	0.001919
HP_ABNORMAL_HELIX_MORPHOLOGY	168	0.426323	2.20307	0.001934
HP_ABNORMAL_GLIAL_CELL_MORPHOLOGY	150	0.373786	2.193101	0.002057
HP_ABNORMAL_CEREBRAL_VASCULAR_MORPHOLOGY	329	0.417849	2.192943	0.002061
HP_ABNORMALITY_OF_TIBIA_MORPHOLOGY	97	0.585692	2.192968	0.002061

HP_ABNORMAL_FEMORAL_HEAD_MORPHOLOGY	90	0.47653	2.187689	0.00213
HP_ABNORMAL_SACRUM_MORPHOLOGY	154	0.526085	2.187179	0.002132
HP_ABNORMAL_ARACHNOID_MATER_MORPHOLOGY	51	0.444361	2.181163	0.002199
HP_ABNORMAL_PONS_MORPHOLOGY	105	0.393249	2.177389	0.002258
HP_ABNORMAL_EOSINOPHIL_MORPHOLOGY	67	0.645536	2.171259	0.002323
HP_ABNORMAL_VAGINA_MORPHOLOGY	103	0.513797	2.167573	0.002367
HP_ABNORMAL_MORPHOLOGY_OF_ULNA	158	0.464674	2.165164	0.002399
HP_ABNORMAL_DURA_MATER_MORPHOLOGY	26	0.700428	2.163769	0.002418
HP_ABNORMAL_ILIAC_WING_MORPHOLOGY	67	0.585351	2.158136	0.002509
HP_ABNORMAL_UTERUS_MORPHOLOGY	134	0.436569	2.151646	0.002613
HP_ABNORMALITY_OF_UPPER_LIMB_EPIPHYSIS_MORPHOLOGY	57	0.606695	2.150237	0.002618
HP_ABNORMAL_MIDDLE_PHALANX_MORPHOLOGY_OF_THE_HAND	63	0.574977	2.144183	0.002682
HP_ABNORMAL_TRACHEAL_MORPHOLOGY	162	0.454908	2.135993	0.00281
HP_ABNORMAL_NEURON_MORPHOLOGY	85	0.395412	2.133249	0.00285
HP_ABNORMAL_LARYNX_MORPHOLOGY	142	0.455774	2.132457	0.002863
HP_ABNORMAL_SHARPEY_FIBER_MORPHOLOGY	33	0.726166	2.132002	0.00287
HP_ABNORMAL_SEPTUM_PELLUCIDUM_MORPHOLOGY	103	0.442604	2.128029	0.002934
HP_ABNORMAL_HALLUX_PHALANX_MORPHOLOGY	40	0.69	2.123715	0.00302
HP_ABNORMAL_CAROTID_ARTERY_MORPHOLOGY	89	0.457885	2.123576	0.003023
HP_ABNORMALITY_OF_BLADDER_MORPHOLOGY	75	0.565809	2.116761	0.003135
HP_ABNORMAL_PROSTATE_MORPHOLOGY	42	0.563666	2.115457	0.003155
HP_ABNORMAL_BRAINSTEM_WHITE_MATTER_MORPHOLOGY	16	0.598028	2.114999	0.003162
HP_ABNORMAL_EPIPHYSIS_MORPHOLOGY_OF_THE_PHALANGES_OF_THE_HAND	39	0.648931	2.111232	0.003242
HP_ABNORMAL_TOENAIL_MORPHOLOGY	165	0.435994	2.104309	0.003377
HP_ABNORMAL_VENOUS_MORPHOLOGY	151	0.422117	2.10403	0.003385
HP_ABNORMAL_CLAVICLE_MORPHOLOGY	107	0.473886	2.103551	0.003386
HP_ABNORMAL_SCLERA_MORPHOLOGY	125	0.4788	2.086949	0.003771
HP_ABNORMAL_PROXIMAL_PHALANX_MORPHOLOGY_OF_THE_HAND	28	0.689205	2.086949	0.003773
HP_ABNORMAL_RIGHT_VENTRICLE_MORPHOLOGY	65	0.554648	2.085456	0.003809
HP_ABNORMAL_INCISOR_MORPHOLOGY	55	0.604485	2.078286	0.003961
HP_ABNORMAL_1ST_METACARPAL_MORPHOLOGY	37	0.580197	2.077904	0.003961
HP_ABNORMAL_MALE_URETHRAL_MEATUS_MORPHOLOGY	376	0.372607	2.070989	0.004131
HP_ABNORMAL_MALE_INTERNAL_GENITALIA_MORPHOLOGY	77	0.509222	2.069707	0.004164
HP_ABNORMAL_PERIAURICULAR_REGION_MORPHOLOGY	106	0.477207	2.066856	0.004232
HP_ABNORMAL_METACARPOPHALANGEAL_JOINT_MORPHOLOGY	25	0.732442	2.045773	0.004793
HP_ABNORMAL_CORONARY_ARTERY_MORPHOLOGY	67	0.566478	2.039752	0.004939
HP_ABNORMAL_SCALP_MORPHOLOGY	411	0.351627	2.035041	0.005062
HP_ABNORMALITY_OF_LOWER_LIMB_EPIPHYSIS_MORPHOLOGY	91	0.474484	2.031359	0.005165
HP_ABNORMAL_MORPHOLOGY_OF_THE_LIMBIC_SYSTEM	43	0.61931	2.029244	0.005234
HP_ABNORMAL_NEURAL_TUBE_MORPHOLOGY	309	0.382258	2.024804	0.005332
HP_ABNORMAL_LABIA_MORPHOLOGY	84	0.492133	2.024466	0.005346
HP_ABNORMAL_CARPAL_MORPHOLOGY	85	0.546398	2.018754	0.00555
HP_ABNORMAL_MORPHOLOGY_OF_THE_MUSCULATURE_OF_THE_NECK	74	0.560469	2.009517	0.005843
HP_ABNORMALITY_OF_SMALL_INTESTINAL_VILLUS_MORPHOLOGY	28	0.575649	2.006955	0.005926
HP_ABNORMAL_RETINAL_VASCULAR_MORPHOLOGY	241	0.372427	2.001334	0.006077
HP_ABNORMAL_FINGERNAIL_MORPHOLOGY	191	0.424145	2.001044	0.006087
HP_ABNORMAL_NASOLACRIMAL_SYSTEM_MORPHOLOGY	67	0.462895	1.999004	0.006136
HP_ABNORMAL_LEFT_VENTRICLE_MORPHOLOGY	125	0.443685	1.998194	0.006164
HP_ABNORMAL_EPIGLOTTIS_MORPHOLOGY	22	0.695356	1.993929	0.006303
HP_ABNORMAL_ASCENDING_AORTA_MORPHOLOGY	26	0.7082	1.988316	0.006526

HP_ABNORMAL_TRICUSPID_VALVE_MORPHOLOGY	43	0.533852	1.983187	0.006695
HP_ABNORMAL_MACROPHAGE_MORPHOLOGY	36	0.785139	1.977715	0.006908
HP_ABNORMAL_FOVEAL_MORPHOLOGY	56	0.644991	1.976046	0.006977
HP_ABNORMAL_FETAL_CARDIOVASCULAR_MORPHOLOGY	38	0.504891	1.973782	0.007055
HP_ABNORMALITY_OF_THE_PARATHYROID_MORPHOLOGY	22	0.673726	1.973423	0.007065
HP_ABNORMALITY_OF_FIBULA_MORPHOLOGY	55	0.563051	1.966816	0.0073
HP_ABNORMAL_GASTRIC_MUCOSA_MORPHOLOGY	50	0.538733	1.956711	0.007719
HP_ABNORMAL_CARTILAGE_MORPHOLOGY	46	0.460373	1.947359	0.008098
HP_ABNORMAL_RECTUM_MORPHOLOGY	135	0.417557	1.936538	0.008573
HP_ABNORMAL_SHOULDER_MORPHOLOGY	97	0.513755	1.928816	0.00892
HP_ABNORMAL_4TH_FINGER_MORPHOLOGY	20	0.686536	1.926086	0.009031
HP_ABNORMAL_SPINAL_MENINGEAL_MORPHOLOGY	23	0.729277	1.923145	0.00916
HP_ABNORMAL_2ND_FINGER_MORPHOLOGY	38	0.546836	1.920704	0.009294
HP_ABNORMAL_ANTIHELIX_MORPHOLOGY	40	0.531881	1.919584	0.009358
HP_ABNORMAL_HYPOTHALAMUS_MORPHOLOGY	21	0.647756	1.911987	0.009724
HP_ABNORMAL_PULMONARY_VALVE_MORPHOLOGY	56	0.473947	1.898706	0.010451
HP_ABNORMAL_FORAMEN_MAGNUM_MORPHOLOGY	17	0.630923	1.893861	0.010695
HP_ABNORMAL_CORPUS_STRIATUM_MORPHOLOGY	43	0.502141	1.880915	0.011451
HP_ABNORMAL_UPPER_LIMB_METAPHYSIS_MORPHOLOGY	17	0.761237	1.879231	0.011533
HP_ABNORMAL_OVARIAN_MORPHOLOGY	165	0.370865	1.866509	0.012276
HP_ABNORMAL_MORPHOLOGY_OF_THE_CHOROIDAL_VASCULATURE	25	0.642927	1.864056	0.01246
HP_ABNORMAL_INTRAHEPATIC_BILE_DUCT_MORPHOLOGY	29	0.521234	1.844877	0.013757
HP_ABNORMAL_INTERVERTEBRAL_DISK_MORPHOLOGY	43	0.504881	1.837861	0.014269
HP_ABNORMAL_5TH_FINGER_PHALANX_MORPHOLOGY	28	0.611431	1.83692	0.014355
HP_ABNORMAL_BASAL_GANGLIA_MORPHOLOGY	197	0.354599	1.826814	0.015156
HP_ABNORMAL_5TH_METACARPAL_MORPHOLOGY	20	0.612876	1.822701	0.01549
HP_ABNORMAL_LABIA_MINORA_MORPHOLOGY	35	0.524906	1.820556	0.015658
HP_ABNORMAL_TALUS_MORPHOLOGY	67	0.460172	1.816584	0.015957
HP_ABNORMAL_LACRIMAL_DUCT_MORPHOLOGY	42	0.494287	1.81505	0.016067
HP_ABNORMAL_VITREOUS_HUMOR_MORPHOLOGY	46	0.486166	1.814553	0.016108
HP_ABNORMAL_CORNEAL_ENDOTHELIUM_MORPHOLOGY	18	0.699475	1.807211	0.016714
HP_ABNORMAL_MORPHOLOGY_OF_THE ABDOMINAL MUSCULATURE	59	0.418725	1.804998	0.0169
HP_ABNORMAL_MONOCYTE_MORPHOLOGY	18	0.666415	1.802783	0.017102
HP_ABNORMALITY_OF_ADRENAL_MORPHOLOGY	106	0.386014	1.793942	0.017859
HP_ABNORMAL_LABIA_MAJORA_MORPHOLOGY	44	0.479255	1.7923	0.01801
HP_ABNORMAL_LOWER_LIMB_METAPHYSIS_MORPHOLOGY	21	0.587851	1.779646	0.019195
HP_ABNORMAL_GREATER_SCIATIC_NOTCH_MORPHOLOGY	20	0.576705	1.757258	0.021473
HP_ABNORMAL_AORTIC_ARCH_MORPHOLOGY	62	0.430661	1.75683	0.021505
HP_ABNORMAL_BRAINSTEM_MORPHOLOGY	262	0.297124	1.754189	0.021786
HP_ABNORMAL_NASAL_MUCOSA_MORPHOLOGY	78	0.457621	1.72918	0.024807
HP_ABNORMAL_VAS_DEFERENS_MORPHOLOGY	21	0.598058	1.728361	0.024892
HP_ABNORMAL_EPIDERMAL_MORPHOLOGY	88	0.450806	1.720362	0.025912
HP_ABNORMAL_ODONTOID_TISSUE_MORPHOLOGY	220	0.34536	1.71804	0.026216
HP_ABNORMAL_SALIVARY_GLAND_MORPHOLOGY	19	0.56655	1.712481	0.027021
HP_ABNORMAL_MORPHOLOGY_OF_THE_OLFACTORY_BULB	31	0.521719	1.691205	0.030108
HP_ABNORMAL_COCHLEA_MORPHOLOGY	30	0.545999	1.69087	0.030146
HP_ABNORMAL_HARD_PALATE_MORPHOLOGY	16	0.564271	1.682226	0.031452
HP_ABNORMAL_PATELLA_MORPHOLOGY	103	0.377261	1.678038	0.032097
HP_ABNORMAL_MIDBRAIN_MORPHOLOGY	84	0.317971	1.667413	0.033743
HP_ABNORMAL_4TH_METACARPAL_MORPHOLOGY	22	0.542171	1.660927	0.034801

HP_ABNORMAL_PANCREAS_MORPHOLOGY	105	0.372581	1.655053	0.035811
HP_ABNORMAL_RENAL_INSTERSTITIAL_MORPHOLOGY	63	0.363071	1.655017	0.035812
HP_ABNORMAL_FALLOPIAN_TUBE_MORPHOLOGY	27	0.450682	1.64146	0.038285
HP_ABNORMAL_LEFT_ATRIUM_MORPHOLOGY	22	0.565637	1.629052	0.040618
HP_ABNORMAL_CAUDATE_NUCLEUS_MORPHOLOGY	27	0.51716	1.625425	0.041316
HP_ABNORMAL_SELLA_TURCICA_MORPHOLOGY	38	0.446011	1.619116	0.042589
HP_ABNORMAL_3RD_FINGER_MORPHOLOGY	23	0.559122	1.616709	0.043064
HP_ABNORMAL_ASTROCYTE_MORPHOLOGY	43	0.360108	1.611382	0.044107
HP_ABNORMAL_VENA_CAVA_MORPHOLOGY	81	0.40556	1.596166	0.047368
HP_ABNORMAL_ANTERIOR_HORN_CELL_MORPHOLOGY	22	0.467573	1.593589	0.047974
HP_ABNORMAL_RENAL_COLLECTING_SYSTEM_MORPHOLOGY	28	0.42694	1.593475	0.047995
HP_ABNORMAL_RETICULOCYTE_MORPHOLOGY	80	0.409795	1.579515	0.05121
HP_ABNORMAL_NASAL_SEPTUM_MORPHOLOGY	22	0.498203	1.564738	0.054793
HP_ABNORMAL_RENAL_TUBULE_MORPHOLOGY	101	0.317179	1.56364	0.055062
HP_ABNORMAL_DENTAL_PULP_MORPHOLOGY	52	0.394727	1.552764	0.057868
HP_ABNORMAL_CERVICAL_SPINE_MORPHOLOGY	20	0.508838	1.532677	0.06336
HP_ABNORMALITY_OF_NEUTROPHIL_MORPHOLOGY	16	0.502115	1.513213	0.068975
HP_ABNORMAL_MORPHOLOGY_OF_MYOCARDIAL TRABECULAE	20	0.552467	1.504215	0.071623
HP_ABNORMAL_OCCIPITAL_BONE_MORPHOLOGY	114	0.271725	1.497346	0.073841
HP_ABNORMAL_PLACENTA_MORPHOLOGY	37	0.440738	1.493609	0.074992
HP_ABNORMAL_LUMBAR_SPINE_MORPHOLOGY	32	0.384374	1.493157	0.075131
HP_ABNORMAL_MEGAKARYOCYTE_MORPHOLOGY	22	0.438128	1.461109	0.086276
HP_ABNORMAL_CELL_MORPHOLOGY	37	0.38947	1.449717	0.090499
HP_ABNORMAL_HUMERUS_MORPHOLOGY	15	0.486033	1.437926	0.095058
HP_ABNORMAL_HEAD_BLOOD_VESSEL_MORPHOLOGY	77	0.340909	1.437119	0.095373
HP_ABNORMAL_BRONCHOALVEOLAR_LAVAGE_FLUID_MORPHOLOGY	16	0.584696	1.433244	0.096955
HP_ABNORMAL_RENAL_VASCULAR_MORPHOLOGY	29	0.381321	1.432436	0.097258
HP_ABNORMAL_RETINAL_ARTERY_MORPHOLOGY	64	0.356628	1.431062	0.09773
HP_ABNORMAL_SUBSTANTIA_NIGRA_MORPHOLOGY	25	0.454225	1.430835	0.097807
HP_ABNORMAL_GALLBLADDER_MORPHOLOGY	93	0.342032	1.41898	0.10262
HP_ABNORMAL_RESPIRATORY_MOTILE_CILIUM_MORPHOLOGY	26	0.479059	1.396017	0.112451
HP_ABNORMAL_RIGHT_ATRIUM_MORPHOLOGY	21	0.504661	1.386012	0.116837
HP_ABNORMAL_LATERAL_VENTRICLE_MORPHOLOGY	101	0.233501	1.287036	0.169285
HP_ABNORMALITY_OF_THALAMUS_MORPHOLOGY	52	0.320158	1.278988	0.174429
HP_ABNORMAL_PARIETAL_BONE_MORPHOLOGY	27	0.351454	1.267692	0.181348
HP_ABNORMAL_HAIRSHAFT_MORPHOLOGY	34	0.360298	1.265153	0.182989
HP_ABNORMAL_GERM_CELL_MORPHOLOGY	50	0.415609	1.257207	0.188137
HP_ABNORMAL_MORPHOLOGY_OF_THE_CONJUNCTIVAL_VASCULATURE	24	0.318756	1.249918	0.193113
HP_ABNORMAL_PREPUTIUM_MORPHOLOGY	34	0.334191	1.228484	0.207401
HP_ABNORMAL_PELVIS_BONE_MORPHOLOGY	22	0.373117	1.226361	0.208802
HP_ABNORMAL_RENAL_MEDULLA_MORPHOLOGY	59	0.231804	1.205765	0.223247
HP_ABNORMAL_LEFT_VENTRICULAR_OUTFLOW_TRACT_MORPHOLOGY	26	0.385745	1.189584	0.235656

Appendix I.

The number of significantly enriched GO gene sets in Up regulated samples. (There was no malignancy pathway in the Proximal list)

Sample Location	Tumor, Neoplastic, Malignancy	Abnormla Morphology
Distal	39	240
Proximal	26	226

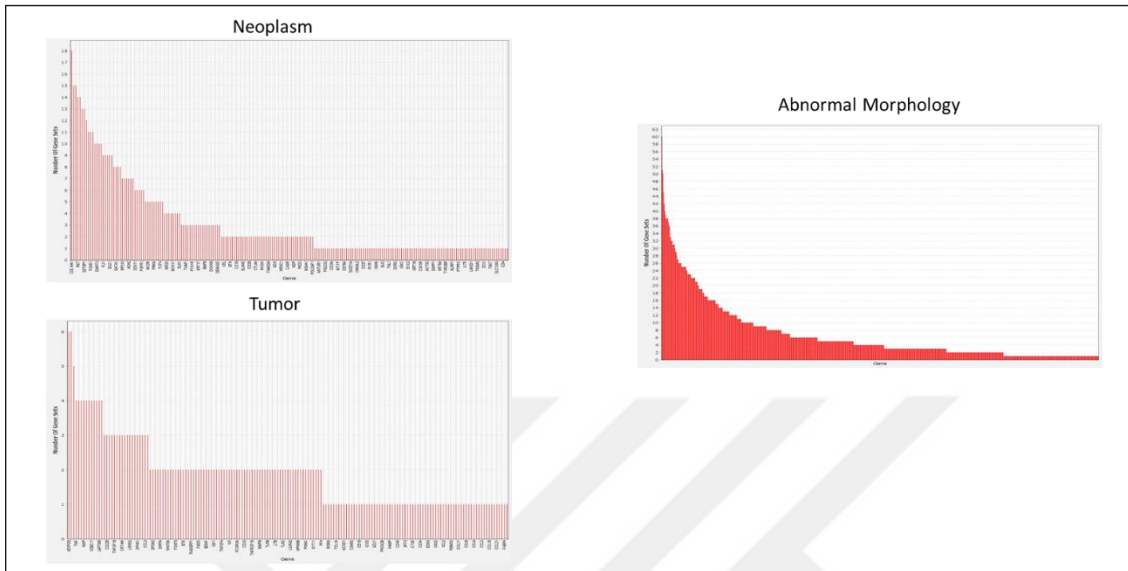
Appendix J.

Number of core enriched genes in the selected gene sets

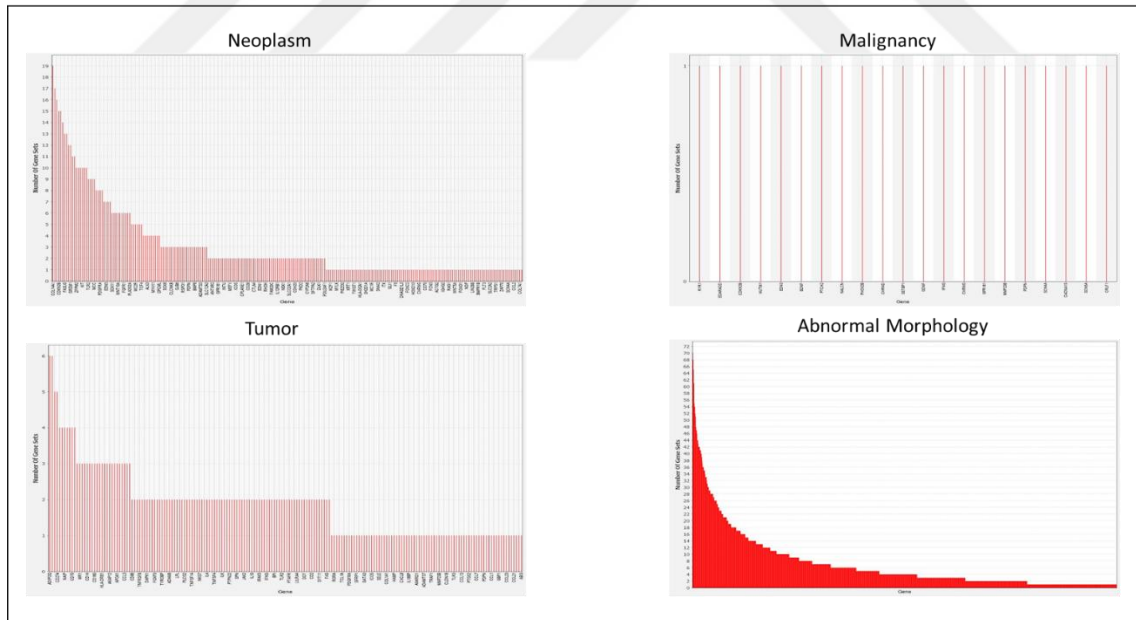
Sample Location	Tumor, Neoplastic, Malignancy	Abnormla Morphology
Distal	398	230
Proximal	346	208

Appendix K.

A. Proximal



B. Distal



Leading Edge Analysis Bar plots of A. Proximal, B. Distal of the Significantly Enriched Gene Sets demonstrating the number of gene sets each Core Enriched Gene is Present in.

Appendix L.

List of 77 common genes between tumor and abnormal morphology associated gene sets core enriched genes

Gene Symbol	Gene Name
IGHM	immunoglobulin heavy constant mu(IGHM)
WIPF1	WAS/WASL interacting protein family member 1(WIPF1)
TREM2	triggering receptor expressed on myeloid cells 2(TREM2)
PTH1R	parathyroid hormone 1 receptor(PTH1R)
IKZF1	IKAROS family zinc finger 1(IKZF1)
GLI1	GLI family zinc finger 1(GLI1)
NR3C1	nuclear receptor subfamily 3 group C member 1(NR3C1)
AFF3	ALF transcription elongation factor 3(AFF3)
ANTXR1	ANTXR cell adhesion molecule 1(ANTXR1)
ACTG2	actin gamma 2, smooth muscle(ACTG2)
GLI3	GLI family zinc finger 3(GLI3)
GLI2	GLI family zinc finger 2(GLI2)
MYLK	myosin light chain kinase(MYLK)
SIX6	SIX homeobox 6(SIX6)
GJA1	gap junction protein alpha 1(GJA1)
TUBA1A	tubulin alpha 1a(TUBA1A)
TRPS1	transcriptional repressor GATA binding 1(TRPS1)
CTSK	cathepsin K(CTSK)
SCN9A	sodium voltage-gated channel alpha subunit 9(SCN9A)
PIEZO2	piezo type mechanosensitive ion channel component 2(PIEZO2)
CTLA4	cytotoxic T-lymphocyte associated protein 4(CTLA4)
TNFSF11	TNF superfamily member 11(TNFSF11)
KCNH1	potassium voltage-gated channel subfamily H member 1(KCNH1)
IL10	interleukin 10(IL10)
PDGFRB	platelet derived growth factor receptor beta(PDGFRB)
CCBE1	collagen and calcium binding EGF domains 1(CCBE1)
PDGFRA	platelet derived growth factor receptor alpha(PDGFRB)
MMP1	matrix metalloproteinase 1(MMP1)
WNT5A	Wnt family member 5A(WNT5A)
CYBB	cytochrome b-245 beta chain(CYBB)
SOX11	SRY-box transcription factor 11(SOX11)
FOXP3	forkhead box P3(FOXP3)
AR	androgen receptor(AR)
EVC	EvC ciliary complex subunit 1(EVC)
TYROBP	transmembrane immune signaling adaptor TYROBP(TYROBP)
SETBP1	SET binding protein 1(SETBP1)
BTK	Bruton tyrosine kinase(BTK)
FAM20C	FAM20C golgi associated secretory pathway kinase(FAM20C)
GAS1	growth arrest specific 1(GAS1)
DDR2	discoidin domain receptor tyrosine kinase 2(DDR2)
MEG3	maternally expressed 3(MEG3)
SASH3	SAM and SH3 domain containing 3(SASH3)
NOTCH3	notch receptor 3(NOTCH3)
FOXC2	forkhead box C2(FOXC2)
WAS	WASP actin nucleation promoting factor(WAS)
Twist1	twist family bHLH transcription factor 1(TWIST1)
CXCR4	C-X-C motif chemokine receptor 4(CXCR4)
MEFV	MEFV innate immunity regulator, pyrin(MEFV)
GATA1	GATA binding protein 1(GATA1)

RASGRP1	RAS guanyl releasing protein 1(RASGRP1)
PDPN	podoplanin(PDPN)
CDH23	cadherin related 23(CDH23)
FLNA	filamin A(FLNA)
MYH11	myosin heavy chain 11(MYH11)
NDP	norrin cystine knot growth factor NDP(NDP)
KCNN3	potassium calcium-activated channel subfamily N member 3(KCNN3)
ICOS	inducible T cell costimulator(ICOS)
GPC6	glypican 6(GPC6)
ELANE	elastase, neutrophil expressed(ELANE)
EVC2	EvC ciliary complex subunit 2(EVC2)
CDKN2B	cyclin dependent kinase inhibitor 2B(CDKN2B)
FZD2	frizzled class receptor 2(FZD2)
TGFB1	transforming growth factor beta 1(TGFB1)
PTCH2	patched 2(PTCH2)
ADA2	adenosine deaminase 2(ADA2)
GDF5	growth differentiation factor 5(GDF5)
BICC1	BicC family RNA binding protein 1(BICC1)
FLI1	Fli-1 proto-oncogene, ETS transcription factor(FLI1)
DLK1	delta like non-canonical Notch ligand 1(DLK1)
COL1A1	collagen type I alpha 1 chain(COL1A1)
WT1	WT1 transcription factor(WT1)
TCF4	transcription factor 4(TCF4)
FAT4	FAT atypical cadherin 4(FAT4)
BMPR1B	bone morphogenetic protein receptor type 1B(BMPRI1B)
IL7R	interleukin 7 receptor(IL7R)
HLA-DRB1	major histocompatibility complex, class II, DR beta 1(HLA-DRB1)
FGFR1	fibroblast growth factor receptor 1(FGFR1)

Appendix M

Hedgehog signaling pathway networks' gene lists

regulation of smoothened signaling pathway	smoothened signaling pathway	Signaling by Hedgehog	Hedgehog 'on' state	negative regulation of smoothened signaling pathway	Activation of SMO
EVC	EVC	EVC	EVC	GLI3	EVC
GLI3	GLI2	GLI2	GLI2	PTCH2	EVC2
GLI1	GLI3	GLI3	GLI3		GAS1
PTCH2	GLI1	GLI1	GLI1		
GAS1	PTCH2	TUBA1A	EVC2		
	EVC2	EVC2	GAS1		
	GAS1	GAS1			

Immunosuppression-associated pathways networks' gene lists

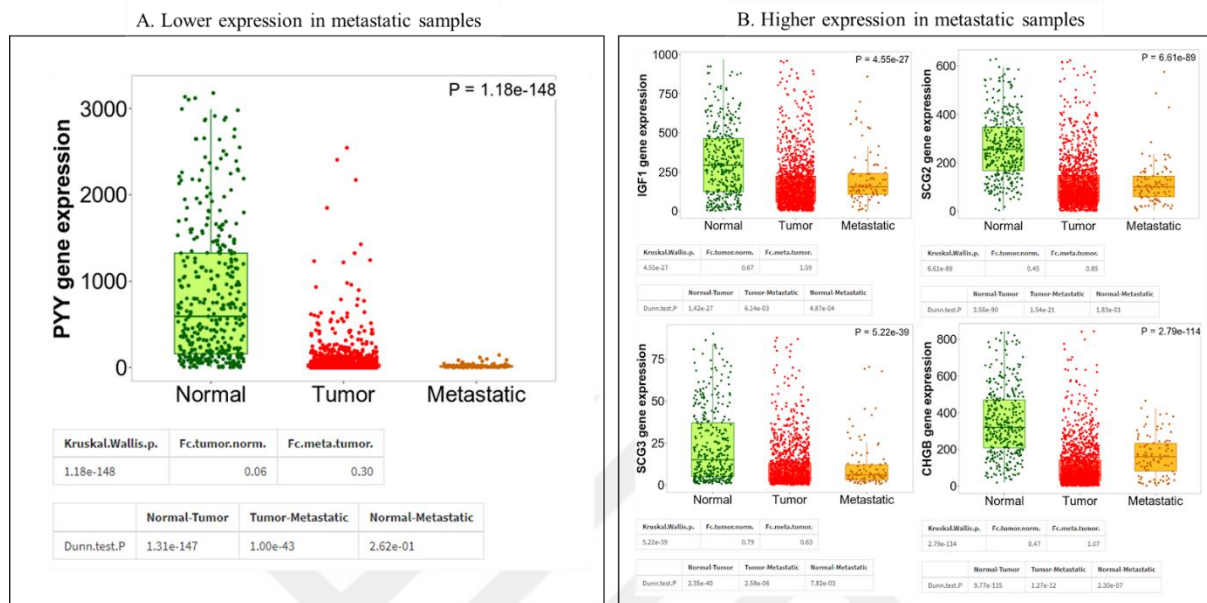
negative regulation of B cell proliferation	negative regulation of lymphocyte activation	negative regulation of mononuclear cell proliferation	negative regulation of lymphocyte proliferation	negative regulation of cell activation	negative regulation of leukocyte proliferation
BTX	BTX	BTX	BTX	BTX	BTX
TYROBP	TYROBP	TYROBP	TYROBP	TYROBP	TYROBP
IL10	FOXP3	FOXP3	FOXP3	FOXP3	FOXP3
CTLA4	TGFB1	TGFB1	TGFB1	TGFB1	TGFB1
	GLI3	IL10	IL10	GLI3	IL10
	IL10	CTLA4	CTLA4	PDGFRA	CTLA4
	CTLA4	SOX11	SOX11	IL10	SOX11
	SOX11	HLA-DRB1	HLA-DRB1	CTLA4	HLA-DRB1
	HLA-DRB1			SOX11	
				HLA-DRB1	

Digestive and reproductive system development associated pathways networks' gene lists

reproductive system development	reproductive structure development	development of primary sexual characteristics	sex differentiation	gonad development
GLI2	GLI2	GATA1	GATA1	GATA1
GATA1	GATA1	PDGFRB	PDGFRB	PDGFRB
GLI3	GLI3	WNT5A	WNT5A	WNT5A
GLI1	GLI1	PDGFRA	PDGFRA	PDGFRA
PDGFRB	PDGFRB	BMPR1B	BMPR1B	BMPR1B
WNT5A	WNT5A	AR	AR	AR
NDP	NDP	WT1	WT1	WT1
PDGFRA	PDGFRA	FLNA	FLNA	FLNA
IL10	IL10			
BMPR1B	BMPR1B			
GJA1	GJA1			
AR	AR			
WT1	WT1			
FLNA	FLNA			

embryonic digestive tract development	digestive system development	digestive tract development	digestive tract morphogenesis
GLI2	GLI2	GLI2	GLI2
GLI3	TGFB1	TGFB1	GLI3
PDGFRA	GLI3	GLI3	GLI1
SOX11	GLI1	GLI1	WNT5A
	WNT5A	WNT5A	PDGFRA
	PDGFRA	PDGFRA	SOX11
	SOX11	SOX11	
	FAT4	FAT4	

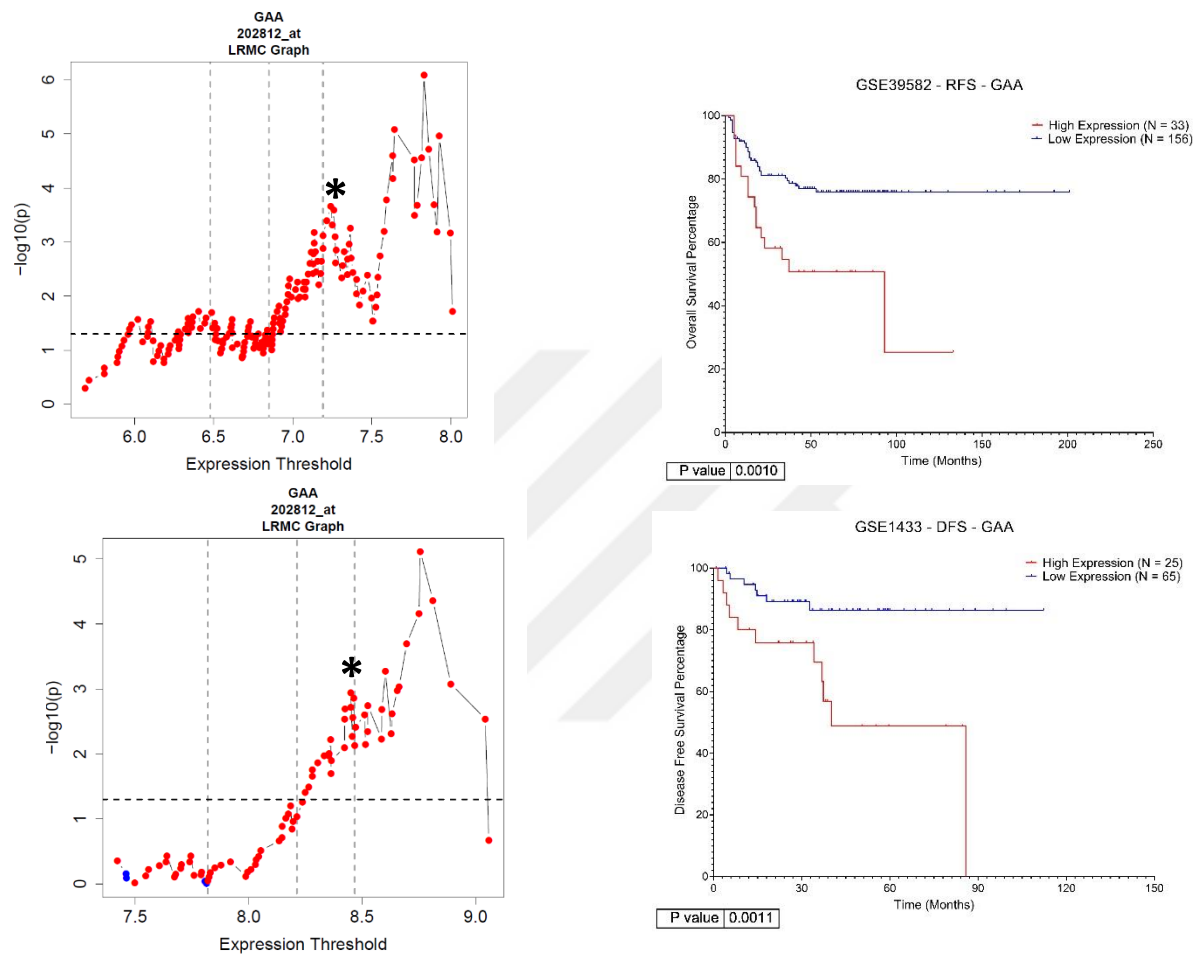
Appendix N



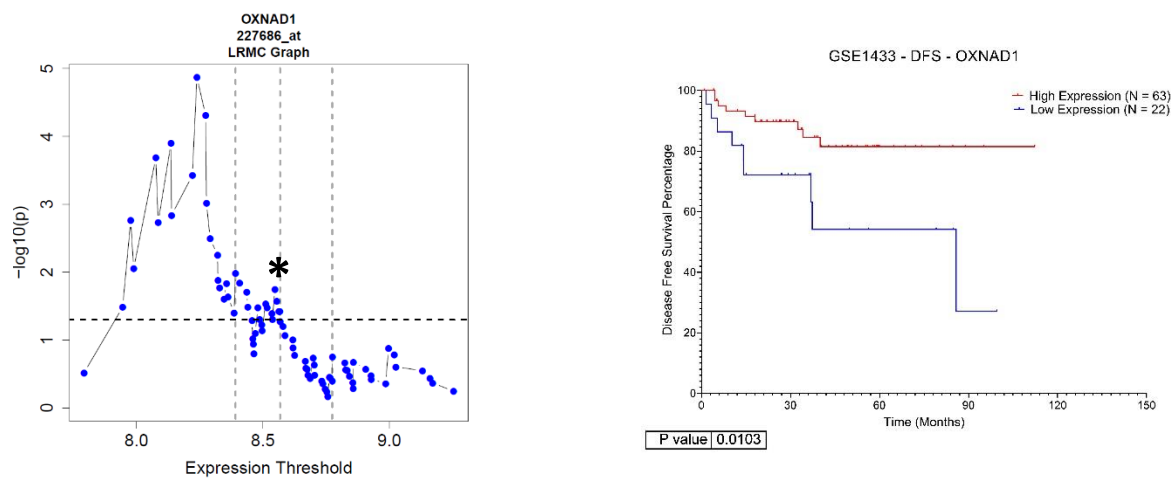
Expression profiles of the remaining 5 genes in normal, non-metastatic tumor, and metastatic tumor samples. The expressions of these genes are significantly lower in the tumor samples compared to the normal samples. However, within tumors, their expression profile is divided into two groups; A. genes which have a significantly higher expression in the metastatic vs. non-metastatic samples, and B. genes with an opposite expression pattern. (Non-parametric Kruskal–Wallis and Dunno test was performed among all three groups, and all the P values are significant at $P \leq 0.05$).

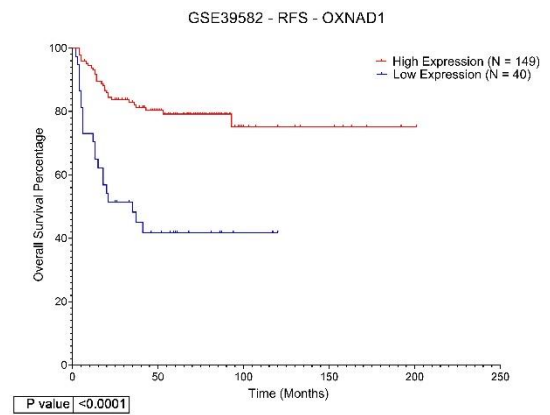
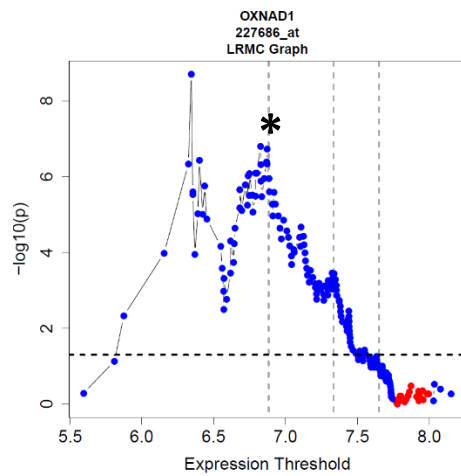
Appendix O

A. GAA – Relapse/ Disease Free Survival

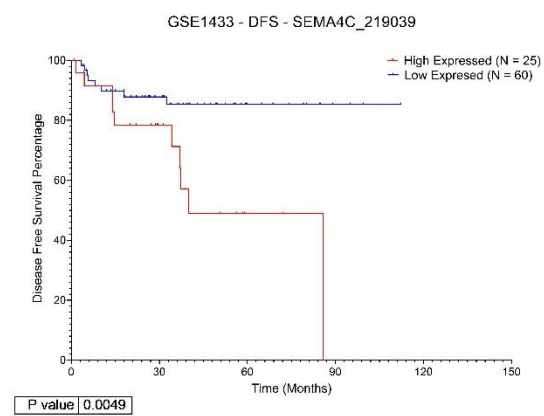
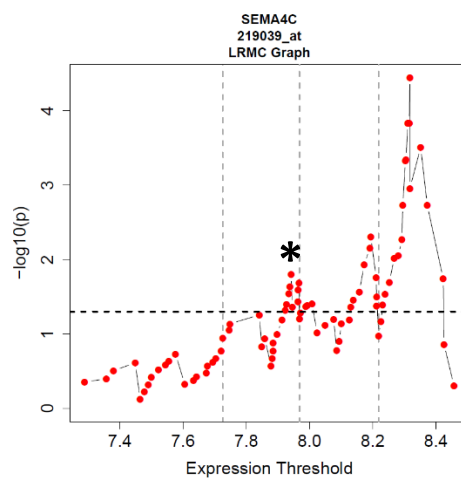
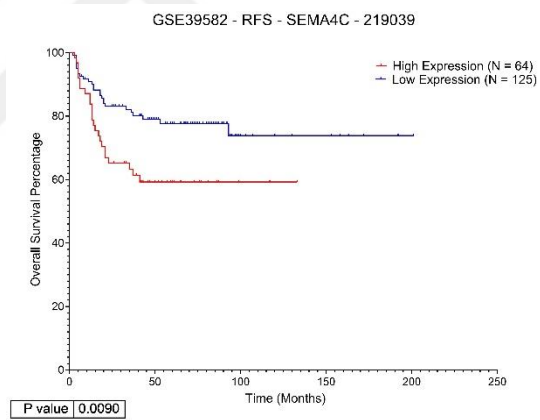
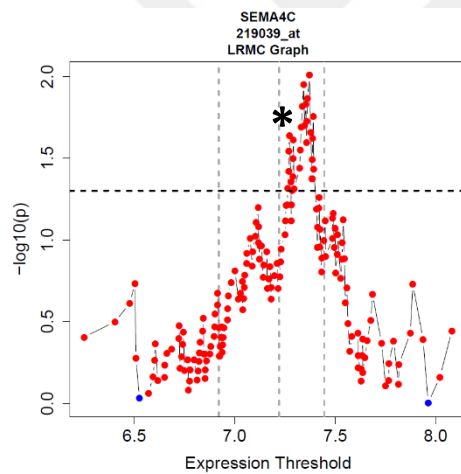


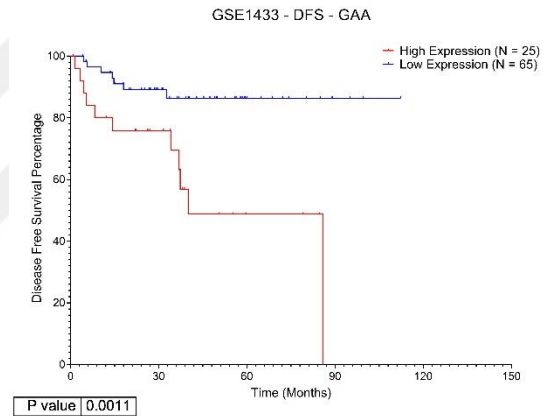
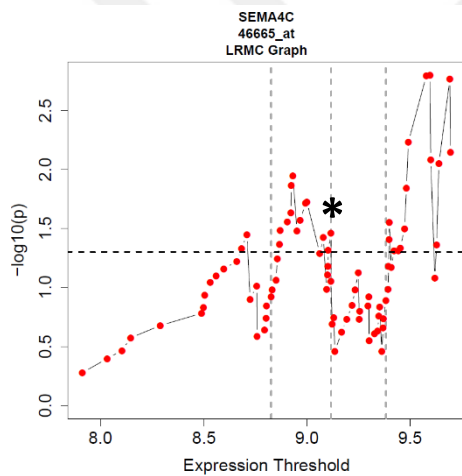
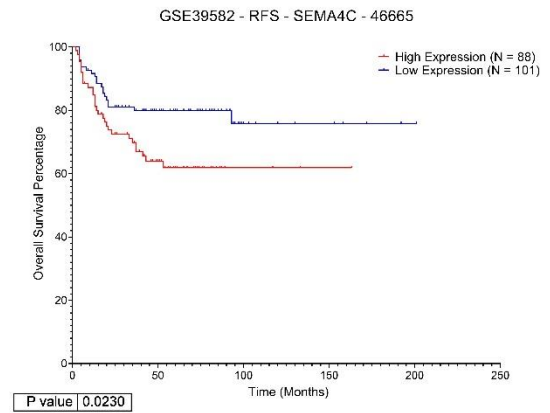
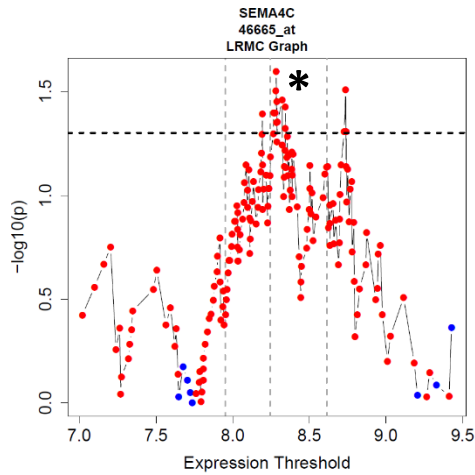
B. OXNAD1 - Relapse/ Disease Free Survival





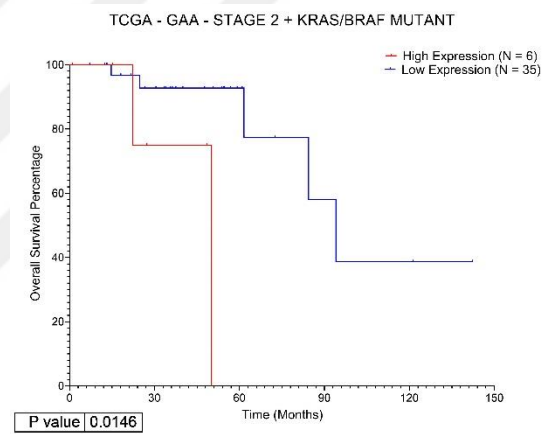
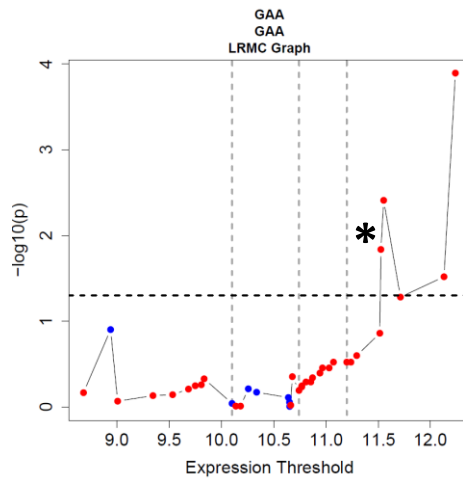
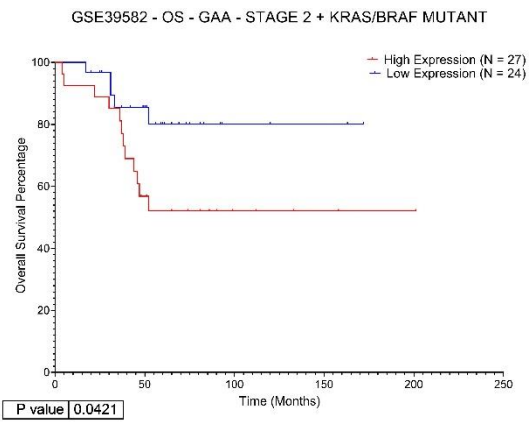
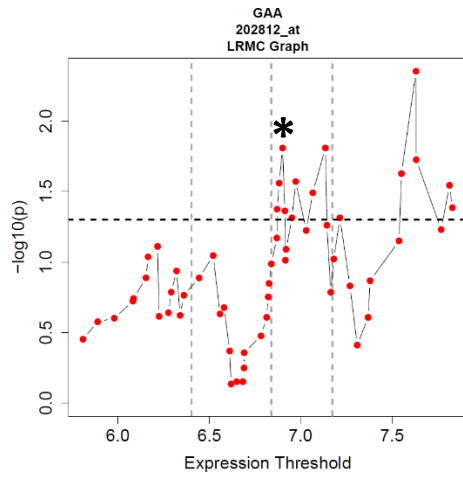
C. SEMA4C - Relapse/ Disease Free Survival



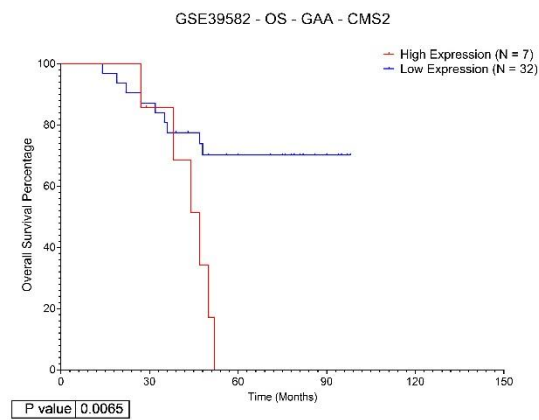
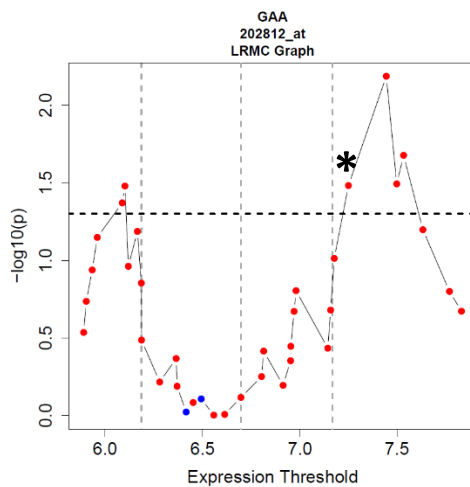


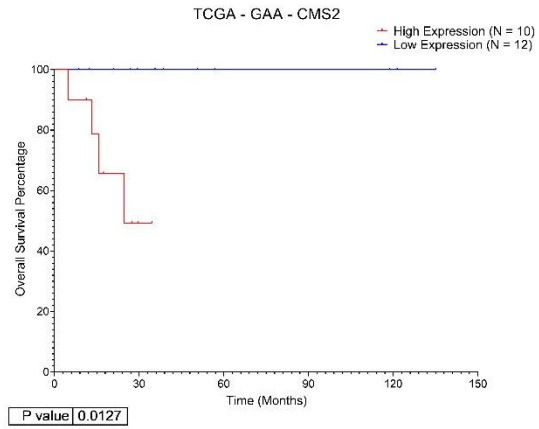
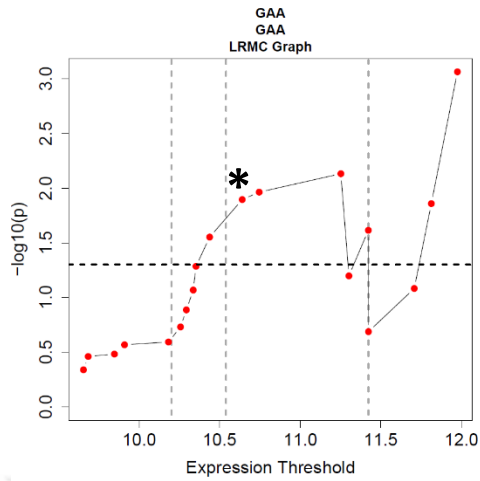
Relapse/ Disease Free Survival (RFS/DFS) plots. The Kaplan Meier plots demonstrate the comparison between the high and low expression groups stratified based on the LRMCs on the left. The asterisk indicate the chosen cut-offs.

Stage 2 + KRAS/BRAF Mutant Samples

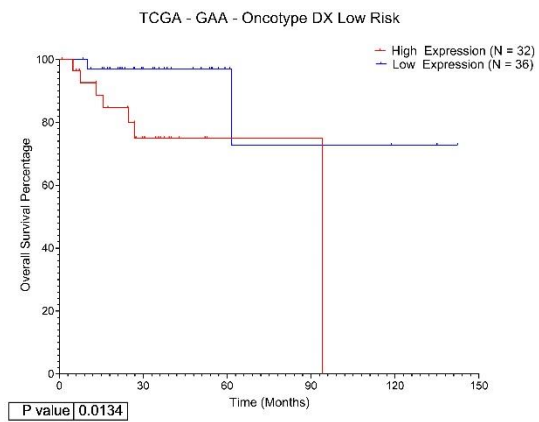
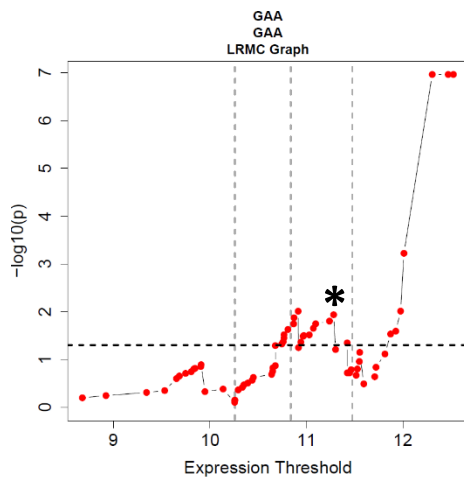
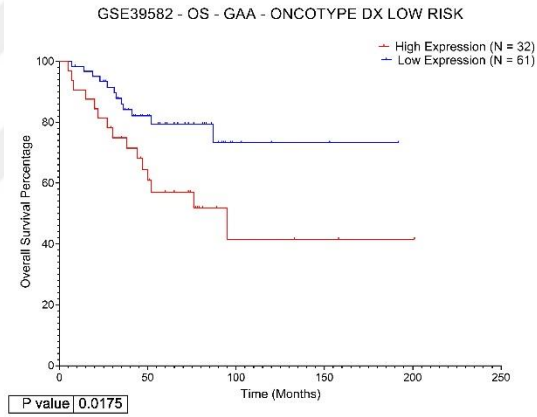
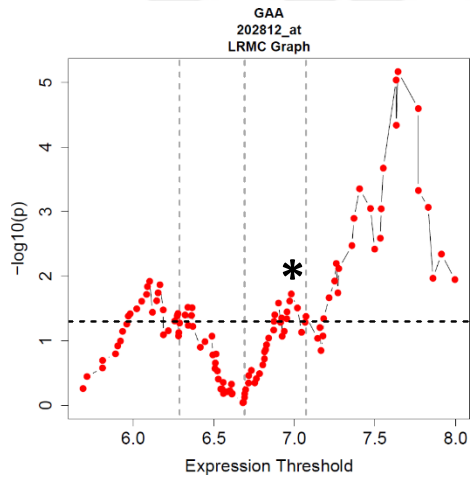


CMS2 Samples



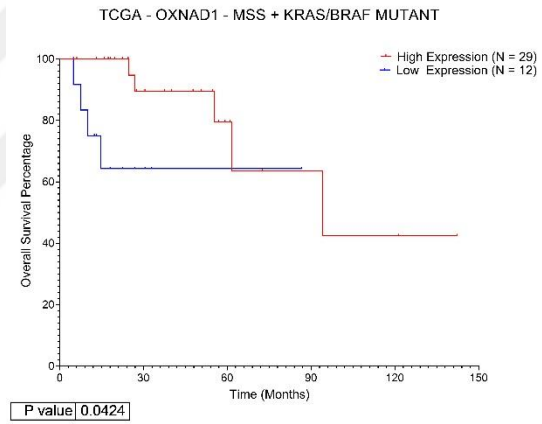
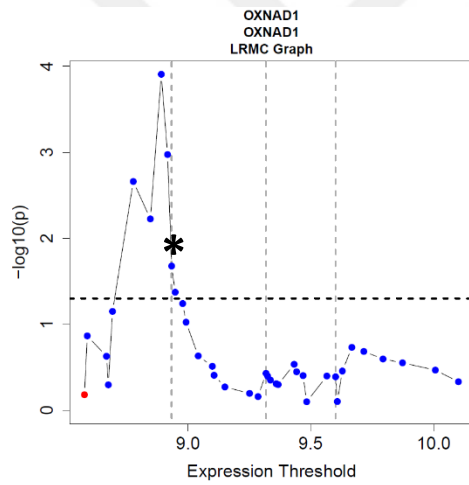
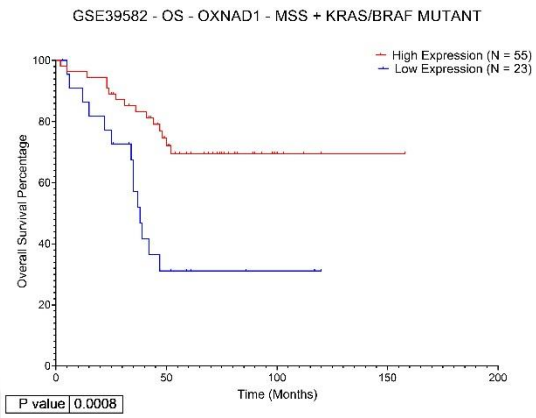
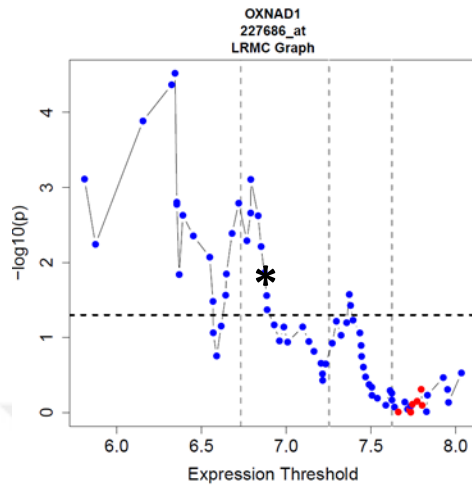


Oncotype DX Low Risk Samples

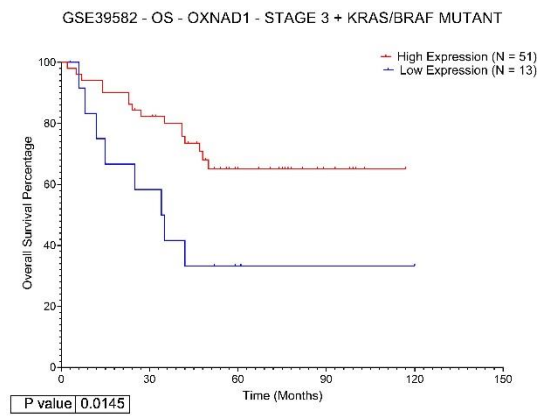
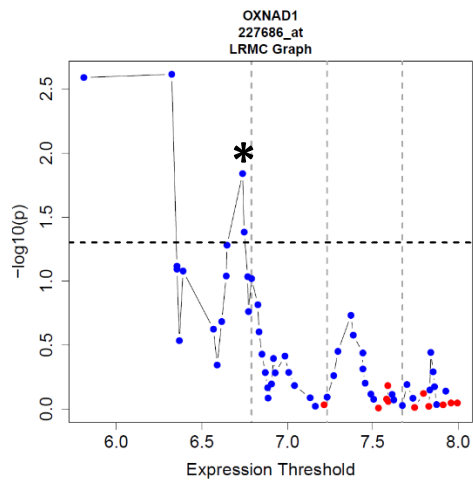


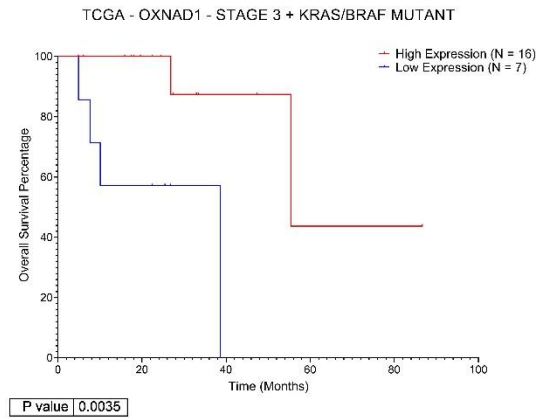
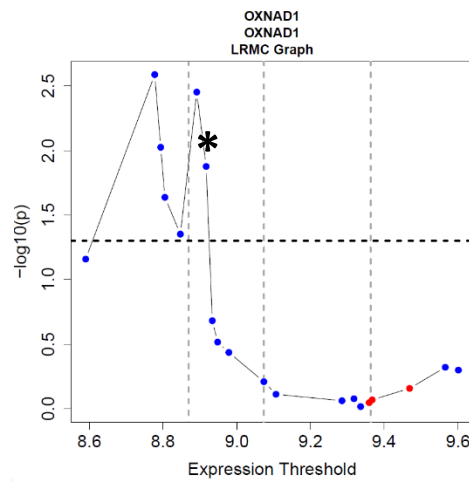
B. OXNAD1 Subgroups

Microsatellite Stable + KRAS/BRAF Mutant Samples

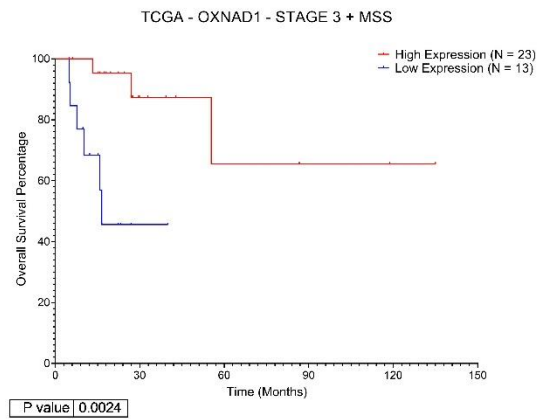
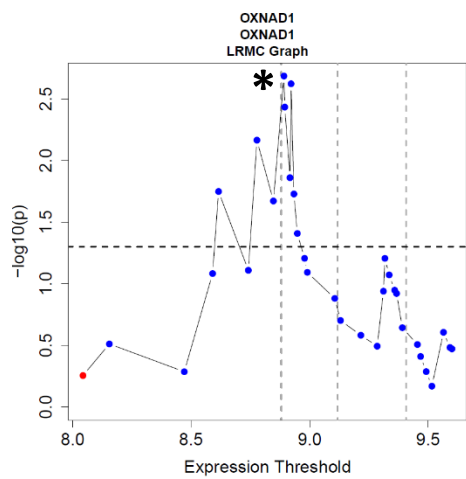
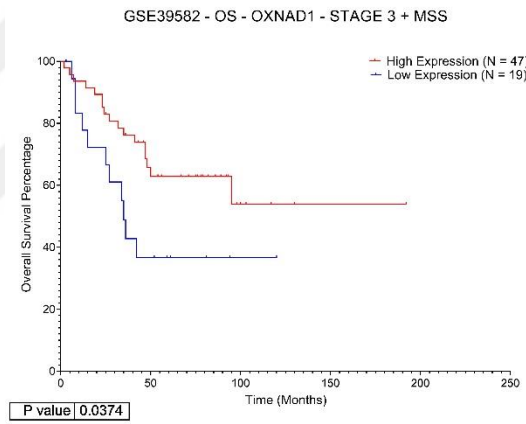
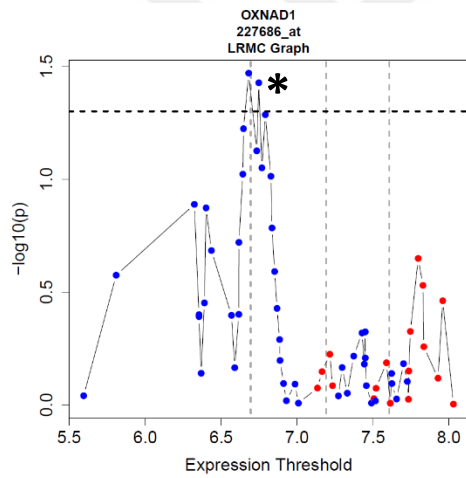


Stage 3 + KRAS/BRAF Mutant Samples

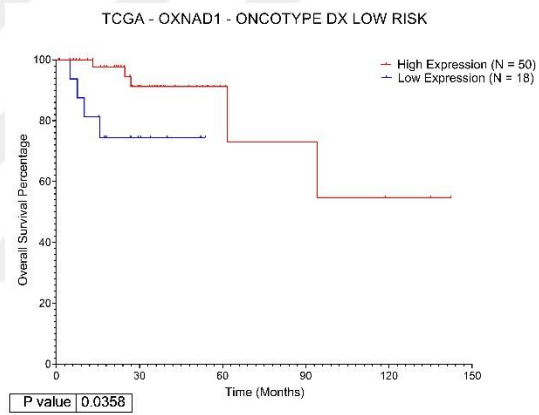
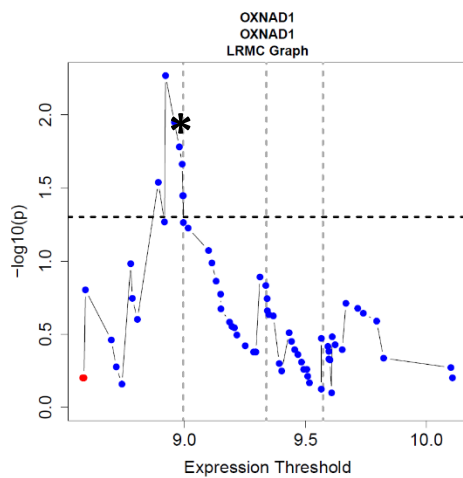
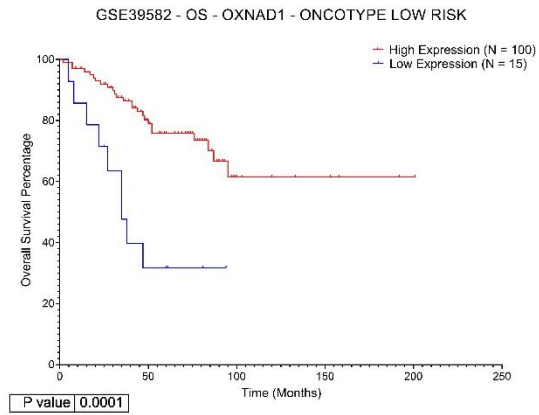
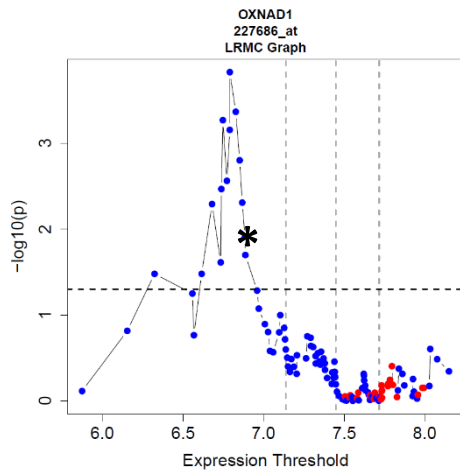




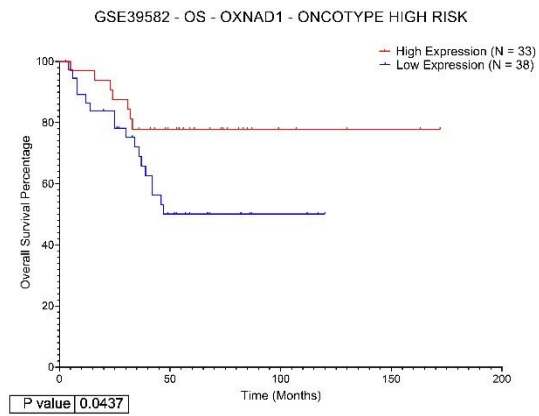
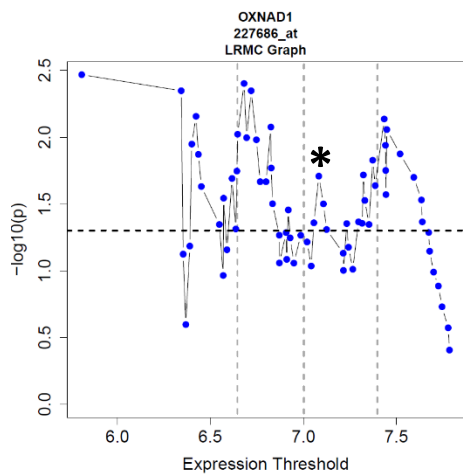
Stage 3 + Microsatellite Stable Samples

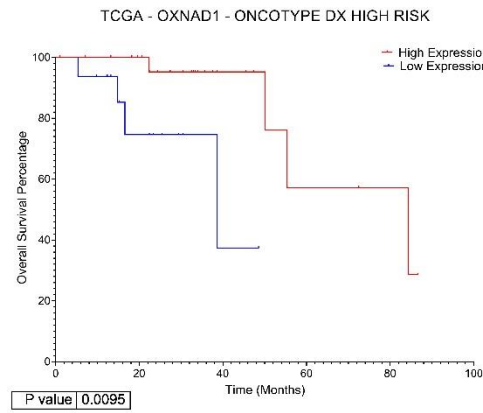
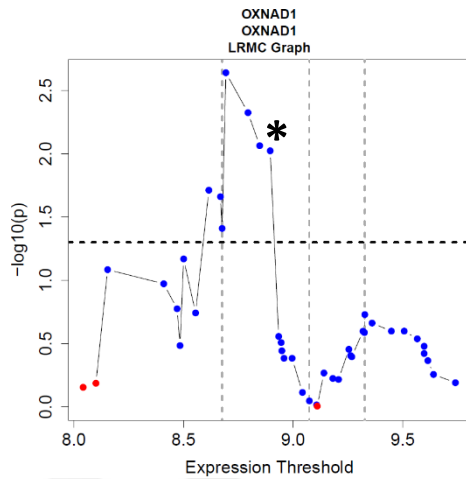


Oncotype DX Low Risk Samples



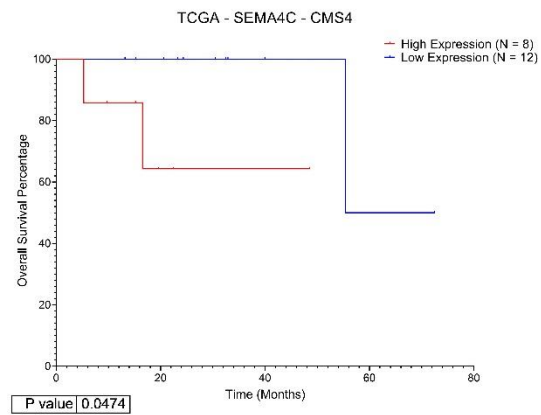
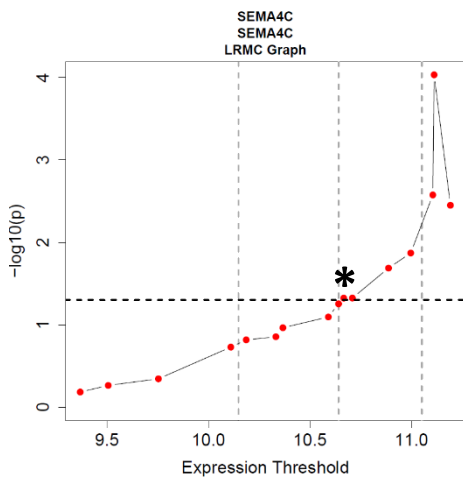
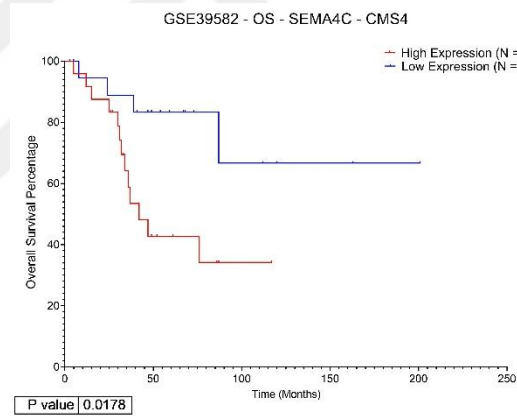
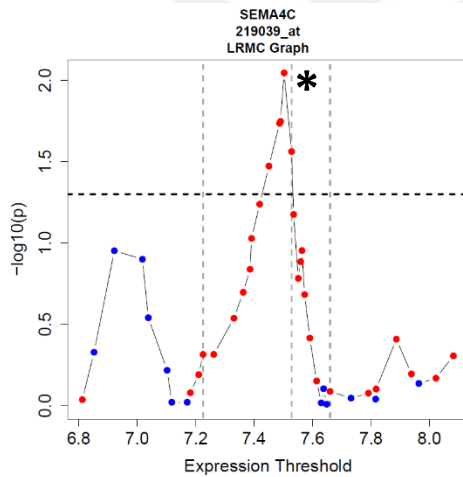
Oncotype DX High Risk Samples





C. SEMA4C Subgroups

CMS4 Subtype Samples



Overall Survival (OS) and molecular subgroups' prognostic association graphs. The Kaplan Meier plots compare the overall survival of the high and low expression groups based on the cut-offs from the LROC graph on the left. The asterisk indicates the chosen cut-off.

Appendix Q

A. GSE39582 OS MVA tables (Continuous gene expression data)

GSE39582 - OS - GENE EXPRESSIONS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA_202812_at	0.525	0.252	4.336	1	0.037	1.691	1.031	2.772
	OXNAD1_227686_at	-0.537	0.256	4.407	1	0.036	0.585	0.354	0.965
	SEMA4C_46665_at	0.296	0.265	1.250	1	0.264	1.344	0.800	2.259
	STAGE	0.437	0.270	2.607	1	0.106	1.548	0.911	2.630
Step 2	GAA_202812_at	0.586	0.247	5.650	1	0.017	1.797	1.108	2.915
	OXNAD1_227686_at	-0.624	0.242	6.646	1	0.010	0.536	0.333	0.861
	STAGE	0.423	0.270	2.446	1	0.118	1.526	0.898	2.593
Step 3	GAA_202812_at	0.577	0.245	5.533	1	0.019	1.780	1.101	2.879
	OXNAD1_227686_at	-0.690	0.241	8.194	1	0.004	0.502	0.313	0.805

B. GSE39582 OS MVA tables (Categorical gene expression data)

GSE39582 - OS - GENE EXPRESSIONS' GROUPS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA_202812_at_OS_group	0.573	0.366	2.450	1	0.118	1.774	0.865	3.639
	OXNAD1_OS_25_GROUP	-0.860	0.291	8.715	1	0.003	0.423	0.239	0.749
	SEMA4C_219039_OS_50_75_GROUP	0.446	0.278	2.560	1	0.110	1.561	0.905	2.695
	STAGE	0.385	0.271	2.024	1	0.155	1.470	0.865	2.501
Step 2	GAA_202812_at_OS_group	0.573	0.367	2.445	1	0.118	1.774	0.865	3.640
	OXNAD1_OS_25_GROUP	-0.927	0.290	10.228	1	0.001	0.396	0.224	0.698
	SEMA4C_219039_OS_50_75_GROUP	0.456	0.280	2.650	1	0.104	1.577	0.911	2.731
Step 3	OXNAD1_OS_25_GROUP	-0.949	0.289	10.806	1	0.001	0.387	0.220	0.682
	SEMA4C_219039_OS_50_75_GROUP	0.512	0.278	3.407	1	0.065	1.669	0.969	2.875

C. GSE39582 RFS MVA tables (Continuous gene expression data)

GSE39582 - RFS - GENE EXPRESSIONS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA_202812_at	0.816	0.277	8.690	1	0.003	2.261	1.315	3.890
	OXNAD1_227686_at	-0.919	0.278	10.958	1	0.001	0.399	0.231	0.687
	SEMA4C_46665_at	-0.029	0.286	0.010	1	0.919	0.971	0.555	1.701
	STAGE	0.414	0.301	1.900	1	0.168	1.514	0.840	2.729
Step 2	GAA_202812_at	0.811	0.272	8.886	1	0.003	2.250	1.320	3.834
	OXNAD1_227686_at	-0.910	0.262	12.016	1	0.001	0.403	0.241	0.673
	STAGE	0.417	0.299	1.940	1	0.164	1.518	0.844	2.729
Step 3	GAA_202812_at	0.809	0.269	9.019	1	0.003	2.245	1.324	3.805
	OXNAD1_227686_at	-0.999	0.258	14.948	1	0.000	0.368	0.222	0.611

D. GSE39582 RFS MVA tables (Categorical gene expression data)

GSE39582 - RFS - GENE EXPRESSIONS' GROUPS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA_RFS_75_GROUP	0.782	0.323	5.865	1	0.015	2.185	1.161	4.113
	OXNAD1_RFS_25_GROUP	-1.065	0.311	11.731	1	0.001	0.345	0.187	0.634
	SEMA4C_219039_RFS_50_75_GROUP	0.197	0.314	0.395	1	0.530	1.218	0.658	2.256
	STAGE	0.499	0.298	2.806	1	0.094	1.648	0.919	2.955
Step 2	GAA_RFS_75_GROUP	0.839	0.310	7.329	1	0.007	2.314	1.261	4.247
	OXNAD1_RFS_25_GROUP	-1.123	0.297	14.261	1	0.000	0.325	0.182	0.583
	STAGE	0.504	0.298	2.862	1	0.091	1.656	0.923	2.971

E. TCGA OS MVA tables (Continuous gene expression data)

TCGA - OS - GENE EXPRESSIONS' GROUPS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA_MEDIAN_GROUP	0.929	0.561	2.746	1	0.097	2.532	0.844	7.596
	OXNAD1_25_GROUP	-1.273	0.580	4.823	1	0.028	0.280	0.090	0.872
	SEMA4C_25_50_GROUP	0.908	0.669	1.841	1	0.175	2.478	0.668	9.196
	STAGE	0.699	0.517	1.827	1	0.177	2.011	0.730	5.537
Step 2	GAA_MEDIAN_GROUP	1.079	0.542	3.969	1	0.046	2.942	1.018	8.507
	OXNAD1_25_GROUP	-1.366	0.572	5.704	1	0.017	0.255	0.083	0.783
	SEMA4C_25_50_GROUP	0.921	0.659	1.957	1	0.162	2.513	0.691	9.137
Step 3	GAA_MEDIAN_GROUP	1.232	0.545	5.113	1	0.024	3.427	1.178	9.966
	OXNAD1_25_GROUP	-1.580	0.555	8.097	1	0.004	0.206	0.069	0.611

F. TCGA OS MVA tables (Categorical gene expression data)

TCGA - OS - GENE EXPRESSIONS + STAGE - MULTIVARIATE RESULTS									
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	GAA	0.420	0.287	2.150	1	0.143	1.522	0.868	2.669
	OXNAD1	-0.736	0.643	1.312	1	0.252	0.479	0.136	1.688
	SEMA4C	0.684	0.401	2.908	1	0.088	1.982	0.903	4.352
	STAGE	0.778	0.520	2.241	1	0.134	2.178	0.786	6.035
Step 2	GAA	0.445	0.289	2.369	1	0.124	1.560	0.885	2.749
	SEMA4C	0.843	0.391	4.649	1	0.031	2.323	1.080	4.996
	STAGE	0.878	0.523	2.815	1	0.093	2.406	0.863	6.707
Step 3	SEMA4C	0.875	0.367	5.684	1	0.017	2.400	1.169	4.929
	STAGE	1.135	0.497	5.220	1	0.022	3.112	1.175	8.240

Appendix R. List of 9 neuropeptide genes and their functions. (Extracted from Gene Cards) [85]

Gene Symbol	Description	Function	Related Pathways
NPY	Neuropeptide Y	widely expressed in the central nervous system and influences many physiological processes, including cortical excitability, stress response, food intake, circadian rhythms, and cardiovascular function.	GPCR downstream signalling and FOXO-mediated transcription
SCG2	Secretogranin III	The protein encoded by this gene is a member of the chromogranin/secretogranin family of neuroendocrine secretory proteins. Studies in rodents suggest that the full-length protein, secretogranin II, is involved in the packaging or sorting of peptide hormones and neuropeptides into secretory vesicles.	Metabolism of proteins and Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
SCG3	Secretogranin II	Granins may serve as precursors for biologically active peptides. Some granins have been shown to function as helper proteins in sorting and proteolytic processing of prohormones; however, the function of this protein is unknown.	Response to elevated platelet cytosolic Ca2+ and Metabolism of proteins.
IGF1	Insulin Like Growth Factor I	The protein encoded by this gene is similar to insulin in function and structure and is a member of a family of proteins involved in mediating growth and development.	NFAT and Cardiac Hypertrophy and PI3K-Akt signalling pathway.
PYY	Peptide YY	This gene encodes a member of the neuropeptide Y (NPY) family of peptides. The encoded preproprotein is proteolytically processed to generate two alternative peptide products that differ in length by three amino acids. These peptides, secreted by endocrine cells in the gut, exhibit different binding affinities for each of the neuropeptide Y receptors. Binding of the encoded peptides to these receptors mediates regulation of pancreatic secretion, gut mobility and energy homeostasis. Rare variations in this gene could increase susceptibility to obesity and elevated serum levels of the encoded peptides may be associated with anorexia nervosa.	GPCR downstream signalling and Signal Transduction.
ADCYAP1	Adenylate Cyclase Activating Polypeptide 1	This gene encodes a secreted proprotein that is further processed into multiple mature peptides. These peptides stimulate adenylate cyclase and increase cyclic adenosine monophosphate (cAMP) levels, resulting in the transcriptional activation of target genes. The products of this gene are key mediators of neuroendocrine stress responses.	GPCR downstream signalling and Disease.
VIP	Vasoactive Intestinal Peptide	The protein encoded by this gene belongs to the glucagon family. It stimulates myocardial contractility, causes vasodilation, increases glycogenolysis, lowers arterial blood pressure and relaxes the smooth muscle of trachea, stomach and gall bladder. The protein also acts as an antimicrobial peptide with antibacterial and antifungal activity.	GPCR downstream signalling and Disease.
CHGB	Chromogranin B	This gene encodes a tyrosine-sulfated secretory protein abundant in peptidergic endocrine cells and neurons. This protein may serve as a precursor for regulatory peptides.	Metabolism of proteins and Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs).
SST	Somatostatin	Somatostatin is expressed throughout the body and inhibits the release of numerous secondary hormones by binding to high-affinity G-protein-coupled somatostatin receptors. This hormone is an important regulator of the endocrine system through its interactions with pituitary growth hormone, thyroid stimulating hormone, and most hormones of the gastrointestinal tract. Somatostatin also affects rates of neurotransmission in the central nervous system and proliferation of both normal and tumorigenic cells.	GPCR downstream signalling and Transcriptional Regulation by MECP2