Background

The primary objective of this project is to identify gene biomarkers that distinguish peritoneal metastases from primary colorectal tumors. This identification is crucial as peritoneal metastasis significantly alters the prognosis and treatment options for colorectal cancer patients. Early detection and a deeper understanding of the molecular characteristics of peritoneal metastases could greatly enhance patient management.

Colorectal cancer (CRC) is the third most common type of cancer and the second leading cause of cancer-related deaths globally [1]. Metastases, especially in the liver, peritoneal cavity, and lungs, are major contributors to the high mortality rate in CRC patients. Peritoneal metastases (PM) are particularly notorious, associated with the poorest outcomes [2]. They occur in 17 to 40% of patients with concurrent primary cancer and in 44 to 50% of recurrence cases [3]. Despite the formulation of consensus guidelines for PM treatment from CRC, PM presents a particularly aggressive prognosis, resulting in poor overall survival compared to other metastatic sites. Early detection of PM is challenging, primarily due to the absence of typical symptoms and the limited accuracy of current imaging modalities [4]. Surgical removal of PM is often complicated as the cancer cells tend to spread extensively across the peritoneum, and conventional anticancer or immunotherapeutic agents have displayed limited efficacy against these metastases [5].

Unlike liver or lung metastases, which often spread through the lymphatic system and blood vessels, peritoneal metastases primarily result from direct seeding into the peritoneal cavity [6]. This occurs when primary tumor cells breach the intestinal wall, enter the abdominal cavity, evade apoptosis, and attach to the peritoneal surface. Despite its significance, the molecular pathways facilitating this process remain poorly understood, underlining the need for focused research in this area.

The intended application of the discovered biomarkers is to develop diagnostic tools that can be used in clinical settings to improve the early detection of peritoneal metastases in the future. Additionally, these biomarkers could be instrumental in tailoring personalized treatment strategies, thereby enhancing patient management and outcomes.

Dataset

The dataset chosen for this biomarker discovery project is highly relevant as it specifically focuses on colorectal cancer and peritoneal metastases (GSE225182). This RNA-Seq dataset, used in the study "CILP2 is a potential biomarker for the prediction and therapeutic target of peritoneal metastases in colorectal cancer" by Ha et al. [7], was sourced from a peer-reviewed scientific article, underscoring its reliability and the credibility of the source. The dataset I decided to focus on includes 7 samples (fastq files): 3 from primary colorectal tumors and 4 from peritoneal metastases. Importantly, no patients included in the dataset underwent preoperative

treatment, peritonectomy, or hyperthermic intraperitoneal chemotherapy, ensuring that the samples reflect the natural state of the tumor and metastases.

The dataset's quality is assured by the use of Illumina NovaSeq 6000, a cutting-edge sequencing technology known for high-throughput and high-fidelity results. However, to ensure the dataset meets the high standards required for accurate biomarker discovery, I will conduct a comprehensive quality control assessment and include these results in my further analysis. Although the dataset comprises only 7 samples, which is relatively small, it is appropriate for preliminary exploratory research to identify potential biomarkers. Expanding the dataset in future studies will be important to validate the findings and enhance statistical power.

In terms of ethical considerations, the study I mentioned above was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of the Asan Medical Center (Approval No.: 2020-0287). Written informed consent was obtained from all patients and/or their legal guardians.

Methods

For this project, I decided to focus on differential gene expression analysis and further Gene Ontology enrichment analysis, driven by the need to identify biomarkers that distinguish peritoneal metastases from primary colorectal tumors. This decision aligns with the urgent need for early detection and a deeper understanding of the molecular characteristics of peritoneal metastases, which could significantly enhance patient management by improving prognosis and personalizing treatment options.

Differential gene expression analysis is pivotal in this study because it allows us to identify specific genes that are upregulated or downregulated in peritoneal metastases compared to primary colorectal tumors. Understanding these differences is crucial for uncovering the biological drivers of metastasis. By pinpointing which genes are differentially expressed, we can not only gain insights into the pathophysiological processes that facilitate the spread of cancer but also identify potential targets for therapeutic intervention.

Gene ontology enrichment analysis further extends the value of our findings from the differential gene expression analysis. By categorizing the identified genes into groups associated with specific biological processes, cellular components, and molecular functions, GO analysis helps in interpreting large data sets from a biological perspective. This step is essential for understanding how groups of co-regulated genes contribute to the underlying mechanisms of metastasis [8].

The initial step involved a quality control assessment utilizing FastQC (v0.12.1) and MultiQC (v1.22.2). Following these assessments, I conducted the alignment of the reads with HISAT2 (v2.2.1), an efficient and accurate aligner tailored for RNA-Seq data which utilizes an indexing scheme to allow for fast and memory-efficient alignment of RNA sequences to a reference

genome. In terms of a reference genome, I chose the Homo sapiens GRCh38 genome assembly, subsetted to chromosome 17 due to memory constraints. Next, in order to convert .sam files to sorted and indexed .bam files I used samtools (v1.19.2). Finally, to retrieve raw counts, I utilized featureCounts (v2.0.6), a part of the Subread package, which is highly efficient in gene-level quantification of RNA-Seq reads mapped to the genome. This setup facilitated a streamlined analysis pipeline that was manageable within the available computational resources.

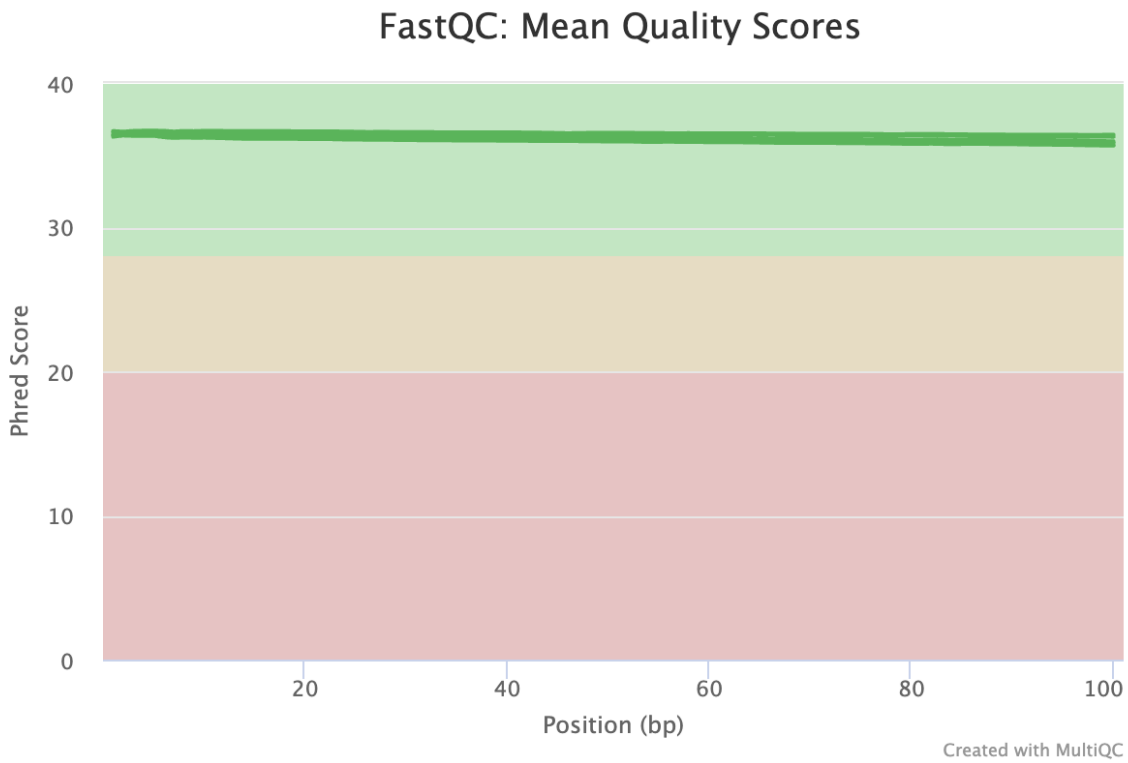
Subsequently, the count data obtained were analyzed in R (v4.2.1) using the DESeq2 (v1.38.3) package for differential expression analysis, which aids in identifying genes with statistically significant changes in expression across different conditions. DESeq2 employs the Wald test to compute p-values for each gene, determining statistically significant differential expression based on the Wald statistic's distribution. This distribution assesses the probability of observing such an extreme test statistic under the null hypothesis.

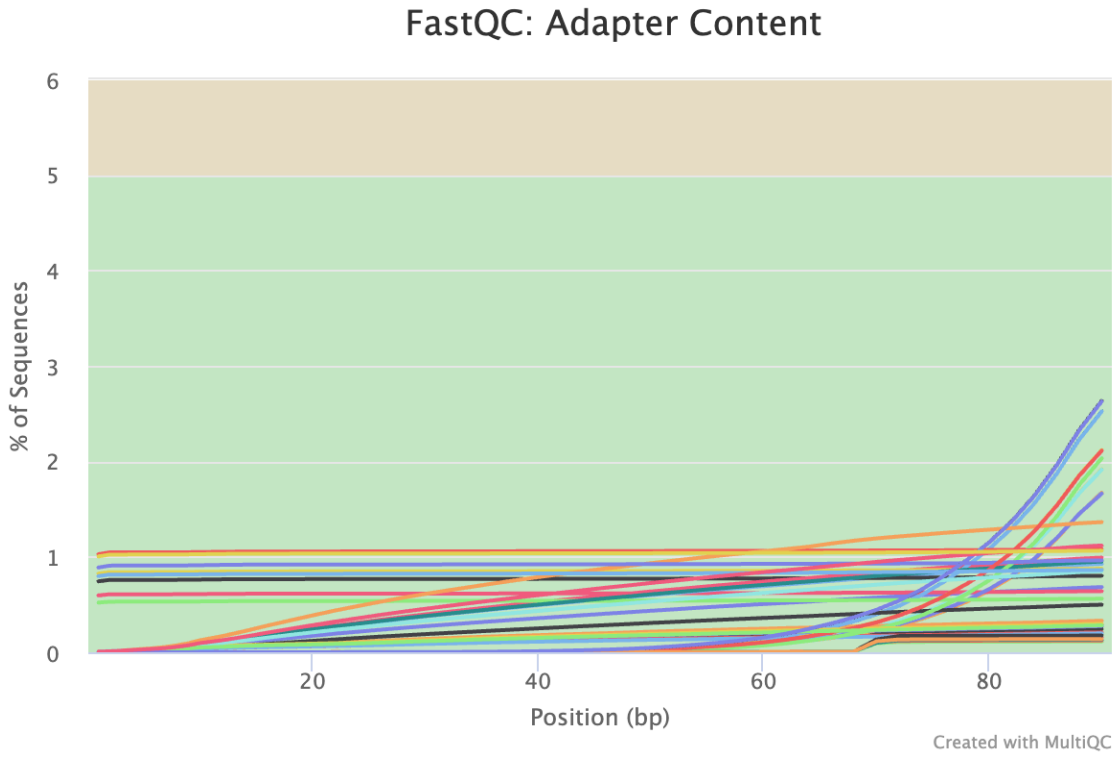
Next, I visualized the normalized counts for differentially expressed genes (padj < 0.05 and abs(log2FoldChange) > 2) using the ComplexHeatmap (v2.14.0) package. To specifically focus on upregulated genes in metastasized tumor (MT) samples compared to primary tumor (PT) samples, I performed k-means clustering. After this, I selected a cluster that showed noticeable upregulation of gene expression in MT samples.

Following the identification of upregulated genes in the metastatic samples, I retrieved the list of these genes and submitted it to the DAVID (Database for Annotation, Visualization, and Integrated Discovery) web tool for Gene Ontology enrichment analysis.

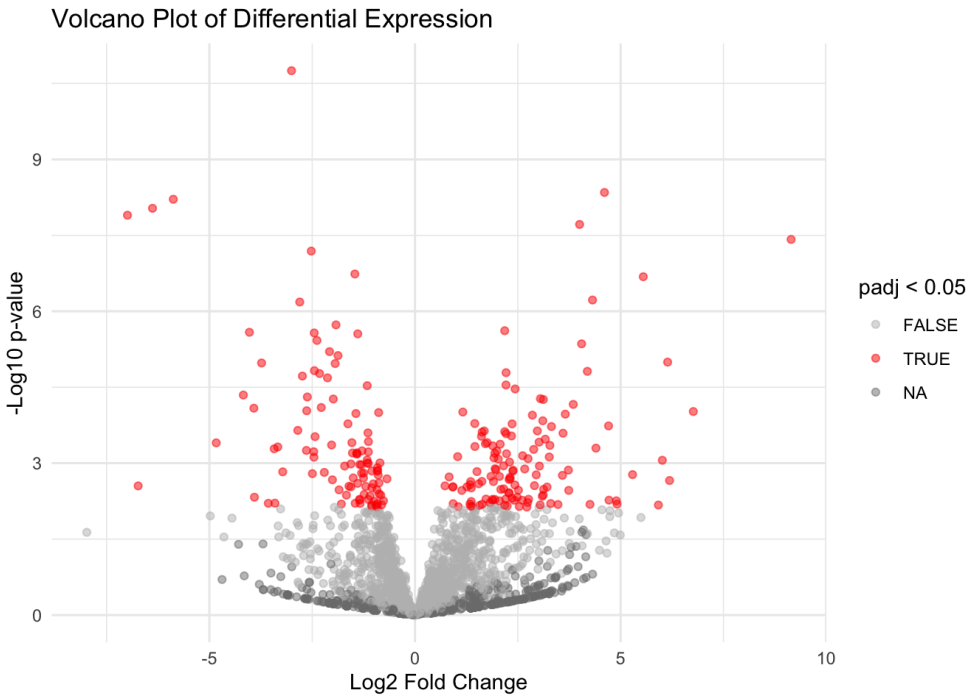
Results

The quality control assessment confirmed the high quality of the data, evidenced by elevated mean quality scores and the absence of adapter content. This indicates that the sequencing outputs are both reliable and clean, making them suitable for further analysis without the need for quality and/or adapter trimming:



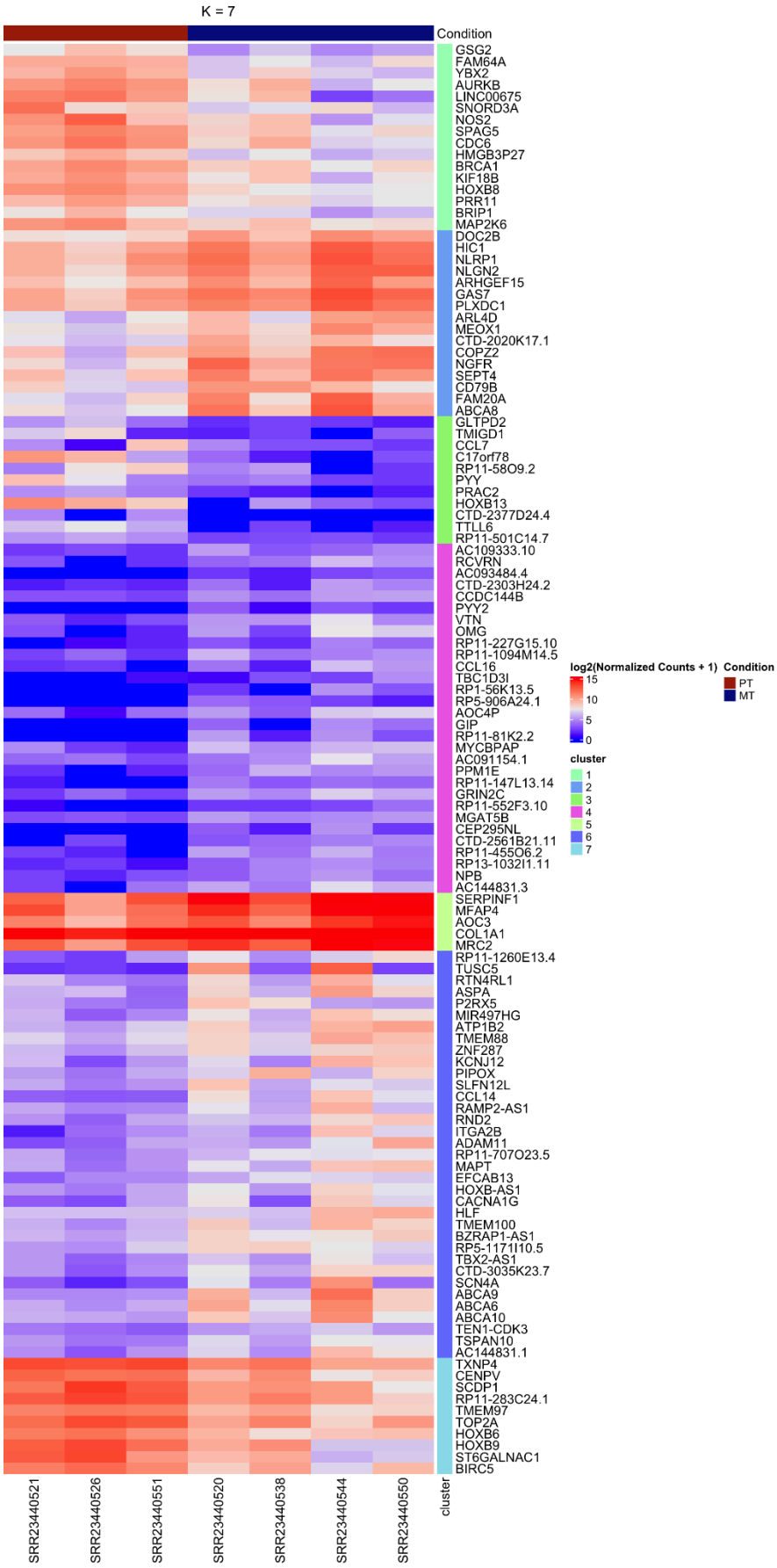


To visualize the distribution of p-values and fold changes among differentially expressed genes (padj < 0.05 and abs(log2FoldChange) > 2), I created a volcano plot:



As evident from the plot, a substantial number of genes are significantly differentially expressed, both upregulated and downregulated.

To specifically identify upregulated genes in metastasized tumor (MT) samples compared to primary tumor (PT) samples, I employed k-means clustering using values of K ranging from 2 to 8. The analysis was visualized through heatmaps for each K value. Among these, the heatmap for K = 7 most effectively represented the data, indicating distinct gene expression patterns within the clusters. This optimal selection was based on observing the unique characteristics and clear separations of gene expression profiles in each cluster at K = 7, suggesting significant biological differences captured by this clustering approach:


K = 7
Condition
GSG2
FAM64A
YBX2
AURKB
LINC00675
SNORD3A
NOS2
SPAG5
CDC6
HMGB3P27
BRCA1
KIF18B
HOXB8
PRR11
BRIP1
MAP2K6
DOC2B
HIC1
NLRP1
NLGN2
ARHGEF15
GAS7
PLXDC1
ARL4D
MEOX1
CTD-2020K17.1
COPZ2
NGFR
SEPT4
CD79B
FAM20A
ABCA8
GLTPD2
TMIGD1
CCL7
C17orf78
RP11-58O9.2
PYY
PRAC2
HOXB13
CTD-2377D24.4
TTLL6
RP11-501C14.7
AC109333.10
RCVRN
AC093484.4
CTD-2303H24.2
CCDC144B
PYY2
VTN
OMG
RP11-227G15.10
RP11-1094M14.5
CCL16
TBC1D3I
RP1-56K13.5
RP5-906A24.1
AOC4P
GIP
RP11-81K2.2
MYCBPAP
AC091154.1
PPM1E
RP11-147L13.14
GRIN2C
RP11-552F3.10
MGAT5B
CEP295NL
CTD-2561B21.11
RP11-455O6.2
RP13-1032I1.11
NPB
AC144831.3
SERPINF1
MFAP4
AOC3
COL1A1
MRC2
RP11-1260E13.4
TUSC5
RTN4RL1
ASPA
P2RX5
MIR497HG
ATP1B2
TMEM88
ZNF287
KCNJ12
PIPOX
SLFN12L
CCL14
RAMP2-AS1
RND2
ITGA2B
ADAM11
RP11-707O23.5
MAPT
EFCAB13
HOXB-AS1
CACNA1G
HLF
TMEM100
BZRAP1-AS1
RP5-1171I10.5
TBX2-AS1
CTD-3035K23.7
SCN4A
ABCA9
ABCA6
ABCA10
TEN1-CDK3
TSPAN10
AC144831.1
TXNP4
CENPV
SCDP1
RP11-283C24.1
TMEM97
TOP2A
HOXB6
HOXB9
ST6GALNAC1
BIRC5
log2(Normalized Counts + 1)
15
10
5
0
Condition
PT
MT
cluster
1
2
3
4
5
6
7
SRR23440521
SRR23440526
SRR23440551
SRR23440520
SRR23440538
SRR23440544
SRR23440550
cluster

As we can see from the heatmap above, cluster 2 might be the one worth investigating further. This cluster includes a distinct set of genes that are upregulated, which might play a significant role in the peritoneal metastasis formation. The genes in cluster 2 are as follows:

- ***DOC2B***
- ***HIC1***
- ***NLRP1***
- ***NLGN2***
- ***ARHGEF15***
- ***GAS7***
- ***PLXDC1***
- ***ARL4D***
- ***MEOX1***
- ***CTD-2020K17.1***
- ***COPZ2***
- ***NGFR***
- ***SEPT4***
- ***CD79B***
- ***FAM20A***
- ***ABCA8***

These genes represent potential biomarkers for distinguishing between metastasized tumor (MT) samples and primary tumor (PT) samples.

Following the identification of upregulated genes in the metastatic samples, I submitted the list of the genes to the DAVID (Database for Annotation, Visualization, and Integrated Discovery) web tool for Gene Ontology enrichment analysis:

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_DIRECT	intracellular protein transport	RT		3	21.4	1.5E-2	1.0E0
☐	UP_KW_MOLECULAR_FUNCTION	Developmental protein	RT		4	28.6	1.9E-2	1.0E-1
☐	UP_KW_MOLECULAR_FUNCTION	Developmental protein	RT		4	28.6	1.9E-2	1.0E-1
☐	GOTERM_CC_DIRECT	plasma membrane	RT		8	57.1	3.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	positive regulation of insulin secretion	RT		2	14.3	3.9E-2	1.0E0
☐	GOTERM_CC_DIRECT	cytoplasm	RT		8	57.1	4.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	activation of cysteine-type endopeptidase activity involved in apoptotic process	RT		2	14.3	4.8E-2	1.0E0
☐	UP_KW_CELLULAR_COMPONENT	Cell membrane	RT		7	50.0	5.0E-2	7.1E-1
☐	GOTERM_BP_DIRECT	neuron apoptotic process	RT		2	14.3	5.4E-2	1.0E0
☐	INTERPRO	DEATH-like_dom_sf	RT		2	14.3	6.4E-2	1.0E0

The results from our GO enrichment analysis have identified several biological processes and cellular components that may be crucial in understanding the progression from primary colorectal tumors to peritoneal metastases. The identified genes provide a foundation for hypothesizing their roles in cancer pathology:

- **Intracellular Protein Transport**: The involvement in protein transport suggests that these genes might be critical in managing the cellular environment and could influence cancer cell survival and proliferation by regulating the availability and functionality of proteins essential for tumor growth and metastasis.
- **Developmental Protein**: The association with developmental processes hints at these genes' roles in cellular differentiation and growth, potentially contributing to the aggressive nature of peritoneal metastases.
- **Apoptotic Processes**: The involvement in apoptosis, particularly through the activation of cysteine-type endopeptidase activity, is particularly relevant. These genes could either promote or inhibit programmed cell death, offering a dual role where they might help cancer cells evade apoptosis or could be leveraged therapeutically to induce death in cancer cells.
- **Cell Membrane Localization**: The significant presence of genes at the cellular boundary, such as those associated with the cell membrane, could implicate these genes in interactions with the tumor microenvironment or in the metastatic process, where cell migration and adhesion are critical.

These findings collectively suggest that these differentially expressed genes play diverse roles in the cellular mechanisms underlying colorectal cancer progression, highlighting their potential as biomarkers for detecting and understanding peritoneal metastases.

Validating the potential biomarkers identified in this study is essential for confirming their relevance and utility in clinical settings. The validation process is comprehensive and includes several critical steps to ensure the robustness and applicability of the findings.

The first step in the validation process is reproducibility. It is crucial to assess the expression of these genes in a larger, independent cohort of colorectal cancer samples. This step verifies that the differential expression observed is consistent across different sample sets and is not an artifact of the specific dataset initially used.

Following reproducibility, functional studies are important. Functional assays, such as gene knockdown or overexpression studies, are conducted to test the roles of these genes directly in critical cancer behaviors, including cell survival, proliferation, migration, and invasion. These experiments help to establish relationships between the gene's expression levels and specific cancer phenotypes, providing insight into their functional significance.

Another vital aspect of validation is the clinical correlation. It's important to correlate the levels of gene expression with clinical outcomes, such as survival rates, response to therapy, and the likelihood of recurrence. This correlation helps to determine whether the biomarkers can predict clinical outcomes and thus their prognostic and predictive values.

In order to assess whether the identified biomarkers are druggable in terms of cancer treatment, we need to consider their involvement in critical biological processes and pathways that can be potential targets for therapeutic intervention. Key processes such as signaling pathways, cell cycle regulation, apoptosis, and cell migration are areas where intervention could substantially impact cancer progression and metastasis.

Among the genes we previously analyzed, certain ones stand out due to their roles in significant biological functions that make them good candidates for drug targeting. These include:

1. NLRP1: Involved in inflammatory responses and apoptosis, making it a potential target for drugs aimed at modulating immune responses and cell death in cancer.
2. ARHGEF15: Regulates cell motility and could be targeted to inhibit cancer cell invasion and metastasis.
3. PLXDC1: Plays a role in angiogenesis, providing a target for therapies aimed at disrupting the blood supply to tumors, thus inhibiting their growth and spread.

The involvement of the NLRP1 pathway in skin cancer development, highlighted by monogenic syndromes predisposing to squamous cell carcinoma (SCC) due to gain-of-function mutations in NLRP1, suggests that targeting this pathway could be a viable therapeutic strategy in cutaneous oncology. Clinical findings further support this, as inhibiting the IL-1β pathway with **canakinumab** has been shown to potentially reduce the incidence and mortality of lung cancer [9]. This indicates that interventions aimed at modulating the NLRP1 pathway could hold significant promise for not only managing but also potentially reversing the progression of cancers, offering new avenues for therapeutic development.

Additionally, ARHGEF15 plays a critical role in cancer progression through its influence on the Rho-family proteins, which significantly affect cell motility and proliferation in pancreatic cancer cells. The activation of RhoA, Cdc42, and Rac by ARHGEF15 leads to enhanced motility by promoting the phosphorylation of myosin light chain, organization of actin stress fibers, and formation of focal adhesions, essential for cancer cell migration and invasion. Furthermore, ARHGEF15's impact on cell proliferation aligns with findings that Rho-family proteins, particularly RhoA, are involved in cancer cell proliferation through pathways like PI3K. Small-molecule inhibitors such as **Rhosin, Y16, and Y-27632** have shown effectiveness in suppressing motility and proliferative activities by targeting the RhoA pathway, indicating that ARHGEF15 may be a promising target for therapeutic intervention, offering a valuable foundation for developing targeted therapies aimed at modulating its activity in cancer treatment strategies [10].

PLXDC1, identified as overexpressed in tumor endothelial cells, is crucial for angiogenesis, metastasis, migration, and invasion. This makes PLXDC1 a viable target for anti-angiogenic therapy, aimed at disrupting the blood supply necessary for tumor growth and spread. The specific role of PLXDC1 in promoting these processes highlights its potential as a target for therapies designed to inhibit angiogenesis and thereby help manage tumor progression and dissemination. Advances in **siRNA delivery systems targeting PLXDC1** further demonstrate

the potential to specifically silence this gene, reinforcing the promise of this approach in clinical settings to address aggressive cancer types effectively [11].

The therapeutic potential of targeting biomarkers such as NLRP1, ARHGEF15, and PLXDC1 is significant, given their critical roles in cancer-associated processes. By focusing on these specific genes, we can tailor therapies that address fundamental mechanisms of cancer progression such as inflammation, cell motility, angiogenesis, and metastasis. The ability to modulate immune responses with drugs targeting NLRP1, inhibit cancer cell invasion and metastasis through ARHGEF15, and disrupt tumor angiogenesis via PLXDC1, provides a multifaceted approach to cancer therapy. Each biomarker offers a unique angle for intervention that could lead to more effective treatments with potentially fewer side effects due to their specific actions in key pathways.

Discussion

One of the primary limitations of this study is the small sample size included in the dataset, which consists of only 7 samples. This limits the statistical power of the analysis and may affect the generalizability of the findings. Additionally, while the RNA-Seq data provide comprehensive insights into gene expression, it doesn't account for post-transcriptional modifications that might influence protein activity and function. Therefore, the biological implications derived from gene expression levels alone should be interpreted carefully.

To overcome the limitations mentioned, future research should aim to include a larger cohort of samples to validate the findings and enhance the statistical power of the analyses. Additionally, integrating proteomic and/or metabolomic data could provide a more comprehensive view of the biological processes at play, helping to confirm the roles of identified biomarkers at the protein and metabolic levels. Longitudinal studies could also be beneficial to track the progression of peritoneal metastases and evaluate the prognostic value of these biomarkers over time.

The identification of biomarkers that distinguish between primary colorectal tumors and peritoneal metastases has significant clinical implications. If validated, these biomarkers could lead to the development of diagnostic tests that improve the early detection of peritoneal metastases, potentially enabling timely and targeted therapeutic interventions. Furthermore, understanding the molecular mechanisms underlying peritoneal metastases could aid in the discovery of new therapeutic targets, contributing to the development of more effective treatments for colorectal cancer patients with advanced disease stages. This could enhance patient outcomes and reduce mortality associated with this aggressive form of cancer.

Supplementary Table

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NGFR	1899.31559642311	4.00237878547826	0.712336702149204	5.61866147483712	1.92442490471388E-08	5.00029737741491E-06
ABCA8	1785.0228355765	4.05146408760227	0.882325412863866	4.59180255780223	4.3943400113892E-06	0.000360567161987145
DOC2B	821.327164961182	2.2142842911147	0.513771798020507	4.30985955174269	1.63358260442314E-05	0.000943242696405803
FAM20A	1392.95676875412	3.59921293154975	0.984418676195296	3.65618107273262	0.000256000558434225	0.006542702796704 21
CTD-2020K17.1	374.70852393312	2.21486265364651	0.606709001449838	3.65061775637696	0.000261610333399446	0.006578234028544 14
HIC1	2344.39053853805	2.22518030361245	0.6526975561117	3.40920581481638	0.000651523057367228	0.0112069284313978
GAS7	3373.60788591979	2.29824995581883	0.695496631012429	3.30447316829311	0.000951550838509667	0.014543801541535
PLXDC1	2599.24032079906	2.15247216881863	0.660944766033066	3.25665967783902	0.001127315040413 81	0.0160249181986843
ARL4D	542.3753030295	2.667929394565 18	0.829764257672579	3.21528599225098	0.001303146213987 66	0.0178210960316383
ARHGEF15	1470.10283428295	2.3851476764448	0.745419537542366	3.199738612042	0.001375522741103 05	0.0183285466100825
SEPT4	1357.42097680999	2.37991065101706	0.746940943748212	3.18620992855803	0.001441499585332 4	0.0185727095333323
NLGN2	2573.63988153082	2.31417512422262	0.75034229410493	3.08415924625861	0.002041282997704 52	0.0228946776505132
MEOX1	649.427054866124	2.28586474230707	0.743151020685991	3.07590877046368	0.002098620455329 14	0.023369637784701
NLRP1	2926.26598763754	2.08714310601452	0.710307621031501	2.93836507481589	0.003299482400924 63	0.0309873076086837
CD79B	659.039977472793	2.28952321144306	0.793294063467983	2.88609648915576	0.003900525549335 33	0.0351498227249351
COPZ2	1449.98625084496	2.60908526918283	0.937683860273987	2.78247859403325	0.005394542380383 94	0.0435354578807618

References

1. Lenos KJ, et al. Molecular characterization of colorectal cancer related peritoneal metastatic disease. Nat. Commun. 2022;13:4443. doi: 10.1038/s41467-022-32198-z.
2. Bang K, et al. Clinical outcomes of curative surgical resection of peritoneal metastasis in patients with colorectal cancer: A long-term follow-up study. Cancer Med. 2023;12:2861–2868. doi: 10.1002/cam4.5195.
3. Razenberg LG, et al. Trends in cytoreductive surgery and hyperthermic intraperitoneal chemotherapy for the treatment of synchronous peritoneal carcinomatosis of colorectal

origin in the Netherlands. Eur. J. Surg. Oncol. 2015;41:466–471. doi: 10.1016/j.ejso.2015.01.018.
4. Koh JL, Yan TD, Glenn D, Morris DL. Evaluation of preoperative computed tomography in estimating peritoneal cancer index in colorectal peritoneal carcinomatosis. Ann. Surg. Oncol. 2009;16:327–333. doi: 10.1245/s10434-008-0234-2.
5. Sato H, et al. Prognoses and treatment strategies for synchronous peritoneal dissemination of colorectal carcinoma. Surg. Today. 2016;46:860–871. doi: 10.1007/s00595-015-1254-8.
6. de Cuba EM, et al. Understanding molecular mechanisms in peritoneal dissemination of colorectal cancer: Future possibilities for personalised treatment by use of biomarkers. Virchows Arch. 2012;461:231–243. doi: 10.1007/s00428-012-1287-y.
7. Ha YJ, Park SH, Tak KH, Lee JL, Kim CW, Kim JH, Kim SY, Kim SK, Yoon YS. CILP2 is a potential biomarker for the prediction and therapeutic target of peritoneal metastases in colorectal cancer. Sci Rep. 2024 May 31;14(1):12487. doi: 10.1038/s41598-024-63366-4.
8. Tomczak, A., Mortensen, J.M., Winnenburg, R. et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci Rep 8, 5115 (2018). https://doi.org/10.1038/s41598-018-23395-2.
9. Laura Calabrese, Zeno Fiocco, Mark Mellett, Rui Aoki, Pietro Rubegni, Lars E French, Takashi K Satoh, Role of the NLRP1 inflammasome in skin cancer and inflammatory skin diseases, British Journal of Dermatology, Volume 190, Issue 3, March 2024, Pages 305–315, https://doi.org/10.1093/bjd/ljad421.
10. Fukushima H, Yasumoto M, Ogasawara S, Akiba J, Kitasato Y, Nakayama M, Naito Y, Ishida Y, Okabe Y, Yasunaga M, Horiuchi H, Sakamoto E, Itadani H, Mizuarai S, Oie S, Yano H. ARHGEF15 overexpression worsens the prognosis in patients with pancreatic ductal adenocarcinoma through enhancing the motility and proliferative activity of the cancer cells. Mol Cancer. 2016 May 4;15(1):32. doi: 10.1186/s12943-016-0516-4.
11. Kim, G. H., Won, J. E., Byeon, Y., Kim, M. G., Wi, T. I., Lee, J. M., … Park, Y. M. (2018). Selective delivery of PLXDC1 small interfering RNA to endothelial cells for anti-angiogenesis tumor therapy using CD44-targeted chitosan nanoparticles for epithelial ovarian cancer. Drug Delivery, 25(1), 1394–1402. https://doi.org/10.1080/10717544.2018.1480672.