Identification of Differentially Expressed Biomarkers in Neoplasm Colorectal Cancer

Farah M.Salih Al-qurashi[1], [1,2] Department of Medical Laboratories Techniques/ Al-Yarmouk university collage.

Asad hameed alnajar[2] [1,2] Department of Medical Laboratories Techniques/ Al-Yarmouk university collage.

ABSTRACT

Colorectal cancer is the third most common cancer worldwide, in which cells in the colon or rectum grow out of control. There are estimated 1.93 million colorectal cancer cases and 0.94 million deaths due to colorectal cancer. The aim of this study was to identify differentially expressed genes (DEGs) and associated biological processes between colorectal neoplasm and adjacent normal tissues using a high-throughput bioinformatics approach to elucidate their potential pathogenesis. The raw reads of gene expression profiles of the E-MTAB-8448 dataset, originally produced through use of the high-throughput RNA sequencing technique, were downloaded from the ArrayExpress database, excluding cell-cultured samples. The E-MTAB-8448 dataset contains information from 124 samples, including 58 normal samples and 66 neoplasm samples. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analyses were performed to identify functional categories and associated molecular and biochemical pathways, respectively, for the identified DEGs. GO and KEGG results suggested that several biological pathways PI3K-Akt signaling pathway, MAPK signaling pathway, JAK-STAT signaling are commonly involved in the development of various cancers including colorectal cancer. This study provides further insights into the underlying pathogenesis of colorectal cancer through identification of key biomarkers, which may facilitate the diagnosis and treatment of these diseases.

Keywords: Differentially Expressed Genes, Colorectal Cancer, Neoplasm, Underlying Pathogenesis

Introduction

Colorectal cancer occurs when the cells that line the colon or the rectum become abnormal and grow out of control [1]. Patients with colorectal cancer do not show symptoms until the cancer has reached its advanced stages [2]. Approximately, 10% of the global cancer incidence and 9.4% of all cancer caused deaths are due to colorectal cancer [1, 3]. Although colorectal cancer is one of the most studied cancers, identification of key biomarkers is a necessity of the time. An abnormal mass of tissue that forms when cells grow uncontrollably and divide without control and lack apoptotic behavior is called neoplasm, it may be benign or malignant [4]. Benign neoplasm may grow large but may not spread into other tissues, whereas the malignant form of neoplasm can invade other tissues such as lymph nodes [4]. Through previous studies, distinct molecular and clinicopathological features for prognosis and diagnosis of the colorectal patients have been observed [3, 5-9]. The metastatic progression and development in colorectal cancer has been reported to be mainly due to disturbance cellular processes, epigenetic modifications, and genomic alterations [2, 8]. There is no proper cure available for colorectal cancer and aside from conventional management and treatment, there are several new targeted agents available for metastatic colorectal cancer [3, 10]. Chemotherapy (5-Fluorouracil, Oxaliplatin, Irinotecan and Capecitabine), including vascular endothelial growth factor (VEGF)-targeted therapy (Bevacizumab) and anti-epidermal growth factor receptor (EGFR)-targeted therapy (Cetuximab and Panitumumab) are some of the few available options [10]. Nevertheless, it remains a challenge to treat CRC which could be attributed to intratumoral heterogeneity (ITH) and the presence of circulating tumor cells (CTCs) [6, 11]. Despite significant advances toward an understanding of the pathophysiology of colorectal cancer, early diagnosis, therapeutic intervention, and underlying pathogenesis remain challenging. Therefore, elucidating the unique molecular characteristics belonging to neoplasm region and adjacent tissues in the colorectal cancer is paramount in developing therapies to improve patient outcome. Recently, numerous research strategies have explored the molecular characteristics of colorectal cancer [4, 7, 8, 12-14]. Among these research strategies, high-throughput sequencing methodologies have received extensive attention and produced a significant progression in the molecular oncology field. In addition, multiple gene expression profiling studies on colorectal cancer have been performed using high-throughput RNA-sequencing and to identify the biomarkers involved in several diseases, including the profiling of hundreds of differentially expressed genes (DEGs) involved in different pathways, biological processes, or molecular functions. Therefore, in the present study, we downloaded the original data (E-MTAB-8448), from the publicly available ArrayExpress database to identify DEGs and the associated biological processes between neoplasm and adjacent normal tissues in colorectal patients using comprehensive bioinformatics RNA-seq analyses. The DEGs were subjected to functional enrichment and pathway analysis. Analysis of the biological functions and pathways may shed light on further insights regarding colorectal cancer development at the molecular level and pave the way toward understanding potential disease pathogenesis mechanisms to facilitate diagnosis, prognosis, and the identification of drug targets.

Methodology

3.1 RNA-Seq Data Collection

We selected colorectal cancer high-throughput RNA sequencing raw datasets from available on ArrayExpress [15] database (accession: *E-MTAB-8448*). Metadata for the study to identify the distinct samples according to phenotypes was retrieved. The datasets from cultured cells were not retrieved and excluded for further analysis. The raw RNA-seq reads were downloaded for each sample corresponding to normal tissues adjacent to neoplasm and tumor neoplasm.

3.2 Data Cleaning and Preprocessing (Quality Control)

Quality control is the process in which bad quality reads, adapters and other unnecessary read data is removed or trimmed down from the raw sequenced sample datasets stored in FASTQ files [16]. Without quality control, downstream analysis including read alignment can be very cumbersome. Reads were first checked for the original quality with the FASTQC tool to assess the quality. FASTQC requires FASTQ files as input and the results are in-depth statistical informative graphical output. For trimming out the bad quality reads and adapters, we employed fastp and cutadapt. Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequences from the high-throughput sequencing reads, all of this is done in an error-tolerant way. Default parameters for cutadapt were employed to trim out the adapters and reads with length of less than 20 base pairs [16].

Reads-Genome Alignment

Once the quality control has been executed on the reads and the bad quality reads have been trimmed, filtered, or deleted, the next step is to align the remaining reads to a reference genome or assemble the reads into a genome [17]. Splice-aware aligner HISAT2 was used for aligning the reads against a reference genome, hg38 (GCA_000001405.15) Homo sapiens genome assembly was used as a reference genome. HISAT2 uses a hierarchical graph FM (HGFM) index which represents the entire genome and variants, covering the genome with global index whereas overlaps with local indexes [17]. Default parameters were used for the alignment where trimmed reads data was provided as input along with the reference genome index produced by HISAT2 and the aligned reads were stored in (Sequence Alignment Map) SAM format. SAMtools was used to sort the reads stored in the SAM files according to their genomic coordinates to make them easier to be read by various downstream analysis tools [18]. The sorted SAM files were then converted using the SAMtools convert module into BAM files.

Reads Quantification

After reads are mapped to a reference genome, the mapped reads must be quantified to count number of reads per gene to account for expression level. Genomic annotation (GTF), if available, can be used to match the reads location in the BAM files with the positions in annotation file. This allows us to quantify gene expression by counting reads per gene, transcripts, and exons, as well as provide new quality control options [19]. Htseq-count is part of the HTSeq Python package for NGS data processing. Htseq-count uses a GTF file to store genome annotation and aligned reads in SAM/BAM format [19]. Htseq-count locates the exons with which the reads overlap and organizes the exon-level counts according to the exons' gene ID in the GTF file [19]. This necessitates that all a gene's exons have the same gene ID. Default parameters were used and a tabular output was gained as a result.

Differential Gene Expression (DGE) Analysis

DGE analysis is performed to identify genes that are expressed differently in a treated or disease state as compared to a normal or control state of the organism, tissue, or sample [20, 21]. Samples from both phenotypes were taken and compared using this technique to account for variability in expression of levels of the genes among the samples with respect to the phenotypes. DESEQ2 is a complex R package for dealing with RNA-seq data and performing differential gene expression (DGE) analysis on two or more conditions [20, 21]. DESEQ2 performs statistical analysis of the count matrices for identifying the genetic profile changes between conditions. It takes read count files from different samples, combines them into a big table with genes in the rows and samples in the columns and then applies normalization for sequencing depth and library composition based on negative binomial distributions including empirical bayes estimation and generalized linear models [20, 21]. F test was used to assess the variability among the samples. A positive log2 fold-change (FC) value indicates a gene that is upregulated whereas a negative log2 FC value represents a gene that is downregulated. DESEQ2 also takes into account false-discovery rate using Benjamini-Hochberg method and therefore produces adjusted-p-value which is then used as a parameter to assess the genetic expression profile of a gene, list of genes or the entire sample [20, 21]. To select the candidate genes for further downstream analysis, logFC>1.25 and padj-value<0.05 threshold was used.

Functional Enrichment Analysis

One of the most important steps of downstream transcriptomic study is functional enrichment analysis which analyzes the DEGs to test for their enrichment in terms of cellular functions, biological processes, and pathways [22, 23]. By this step, the genes that are differentially expressed among two or more conditions are filtered out. This step helps in determination of categorical biological functions that might be impacted due to the change in the gene expression [22, 23]. The genes that met the threshold of logFC>1.25 and padj-value<0.05 in the previous step were further used for downstream analysis such as gene ontology and KEGG pathways analysis using the enrichR R language package [22]. The genes were tested for enrichment in biological process, molecular function and cellular component in Gene Ontology Resource on enrichR database [22, 23].

Results

4.1 Alignment of the Reads

Splice-aware alignment using HISAT2 was performed for aligning the reads against a reference genome, hg38 (GCA_000001405.15) Homo sapiens genome assembly with an average of 90% alignment percentage. Alignment percentage greater than 80% is regarded as a good alignment. Hence, the average alignment score indicates that the samples were of good quality reads. No sample was identified to have an alignment percentage less than 80%.

4.2 Identification of Differentially Expressed Genes (DEGs)

Differential gene expression analysis was performed between the normal tissues adjacent to neoplasm and tumor neoplasm samples using DeSEQ2 to account for variability in expression of levels of the genes among the samples with respect to their phenotypes.

By calculating the p-value, logFC and FDR values through DeSEQ2 statistical analysis of the read counts a 786 of DEGs were obtained between normal tissues and tumor neoplasm samples. Through volcano plot visualization of the differentially expressed genes which plots the statistically significant (P-value) versus fold change (logFC) values for a quick identification of genes with logFC and p-values crossing the set threshold. The genes that exceed p-value<0.05 and logFC>1.25 are regarded as biologically and statistically upregulated genes which can be seen on the right side of the plot, whereas the genes that exceed p-value<0.05 and logFC<1.25 are regarded as downregulated genes which can be seen on the left side of the plot **Figure 1**. Through DEGs analysis, a total of 23 genes were found to be upregulated whereas a total of 141 were found to be downregulated. The top 10 upregulated differentially expressed genes were identified to be ITGAL, BIRC3, CETP, LYZ, CALB1, ALDH1A2, PHLDA1, DUOX2, ALMS1P1 and IRS1 **Table 1**. The top 10 down regulated differentially expressed genes were identified to be RNF10, CHPF2, FAM168A, OPHN1, LYRM2, COL9A3, CBY1, C14ORF93, RIMS4 and TBL1X **Table 2.**

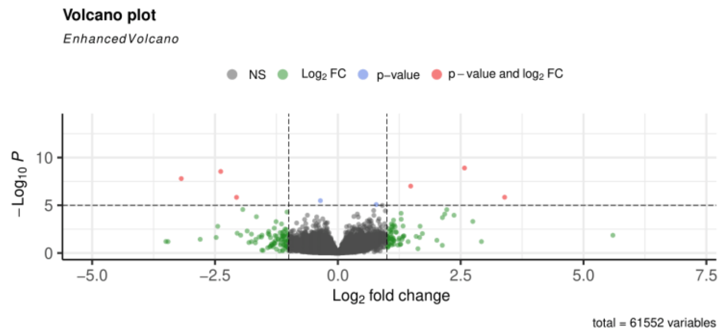

Figure 1(Volcano plot for differential gene expression values between neoplasm and normal tissues)

Symbols	log2FoldChange	pvalue	Expression
ITGAL	2.746446924	0.0004917682747	UP
BIRC3	1.331171717	0.01675540909	UP
CETP	1.346181364	0.01605610566	UP
LYZ	1.593556822	0.0182339292	UP
CALB1	2.364065736	0.0001110489432	UP
ALDH1A2	1.720588258	0.009203559482	UP
PHLDA1	1.663521264	0.02372225176	UP
DUOX2	1.317250027	0.0219864261	UP

Table 2(Top 10 upregulated differentially expressed genes identified between neoplasm vs. normal adjacent tissues)

Symbol	log2FoldChange	pvalue	Expression
RNF10	-0.7274431211	0.02499849205	DOWN
CHPF2	-1.121676957	0.01354855327	DOWN
FAM168A	-0.7244360895	0.007825950011	DOWN
OPHN1	-0.9604375629	0.04782673924	DOWN
LYRM2	-1.08794182	0.01883815697	DOWN
COL9A3	-0.697275184	0.02607253304	DOWN
CBY1	-0.7867990563	0.04862034068	DOWN
C14orf93	-0.6372756721	0.04267384366	DOWN
RIMS4	-1.370305146	0.0250338973	DOWN
TBL1X	-0.935337355	0.04465172635	DOWN

Table 2(Top 10 downregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

4.3 Gene Ontology Analysis of the Upregulated and Downregulated Genes

The genes that are differentially expressed among neoplasm and normal adjacent tissues of the neoplasm conditions are filtered out based on threshold of logFC>1.25 and padj-value<0.05 in the previous step. This step helps in determination of categorical biological functions that might be impacted due to the change in the gene expression. Through gene ontology analysis, it was identified that the upregulated DEGs are enriched in biological processes of alcohol metabolic process, response to fatty acids, protein tetramerization, homotetramerization, neutrophil activation and other metabolic processes which indicates that the genes are involved in the inflammatory responses, neutrophil mediated immunity response and binding of multiple protein complexes to give an immune response against the neoplasm region. Through molecular function enrichment, it was found that the proteins have transmembrane receptor tyrosine-kinase adaptor activity, NAD(P)H oxidase activity, oxidoreductase activity and similar NAD(P)+ activities which indicates that these genes have cancer development and progression-related functions. Through cellular component enrichment, it was identified that the DEGs are being expressed inside tertiary granule lumen, NADPH oxidase complex and other cytoplasmic vesicle lumen which indicates that these genes play a crucial role in the development of the neoplasm as it has been previously reported in several studies cancer grows into the lumen. Whereas the downregulated DEGs were identified to be enriched in biological processes such as proton transmembrane transport, positive regulation of response to DNA damage stimulus, positive regulation of double-strand break repair via homologous recombination, positive regulation of DNA repair which indicates that these genes play a crucial role in positive regulation and repair of the DNA when the damage is caused to the DNA. However, due to the downregulation of these genes it can be concluded that a significant amount of DNA damage may be left unrepaired

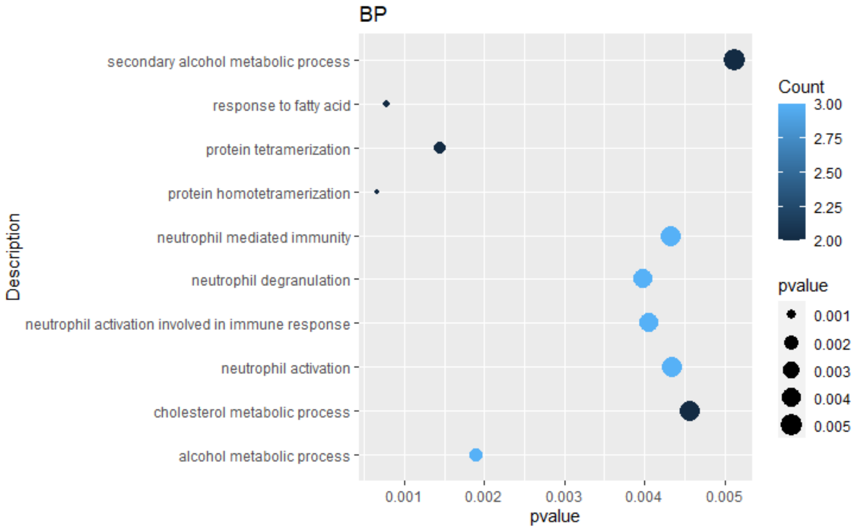

Figure 2(Biological processes that are enriched by upregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

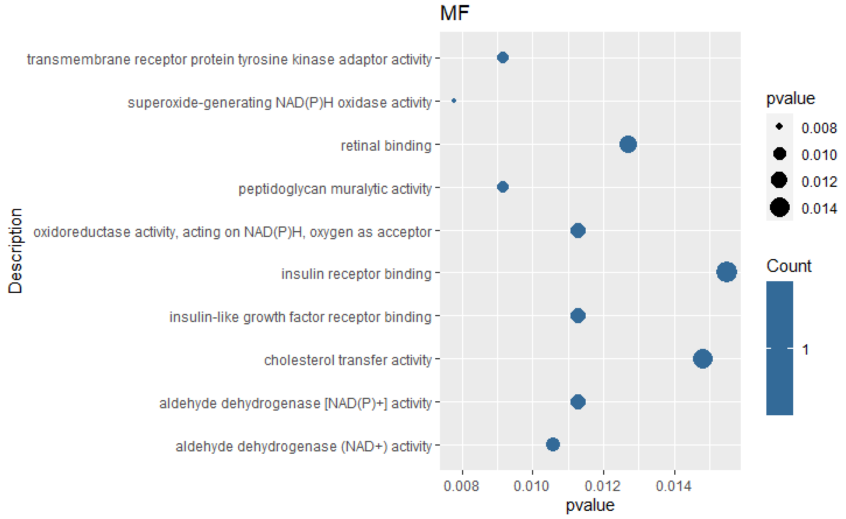

Figure 2(Molecular functions that are enriched by upregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

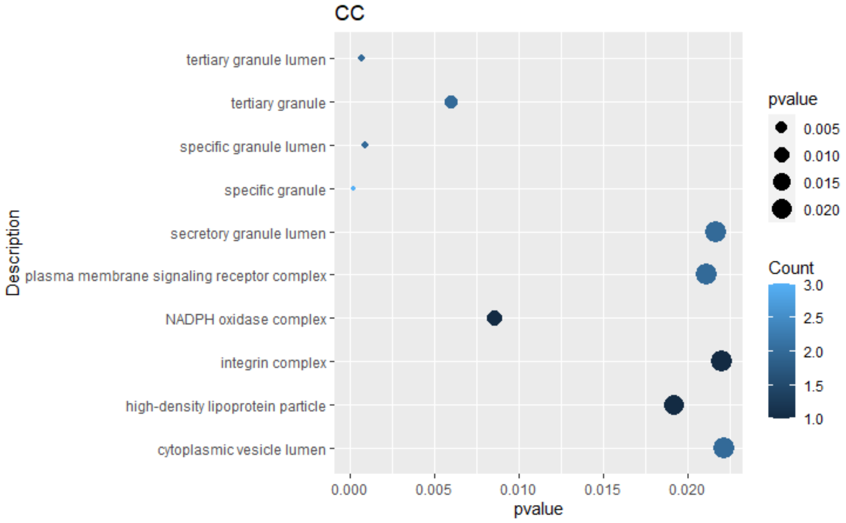

Figure 3(Cellular component that are enriched by upregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

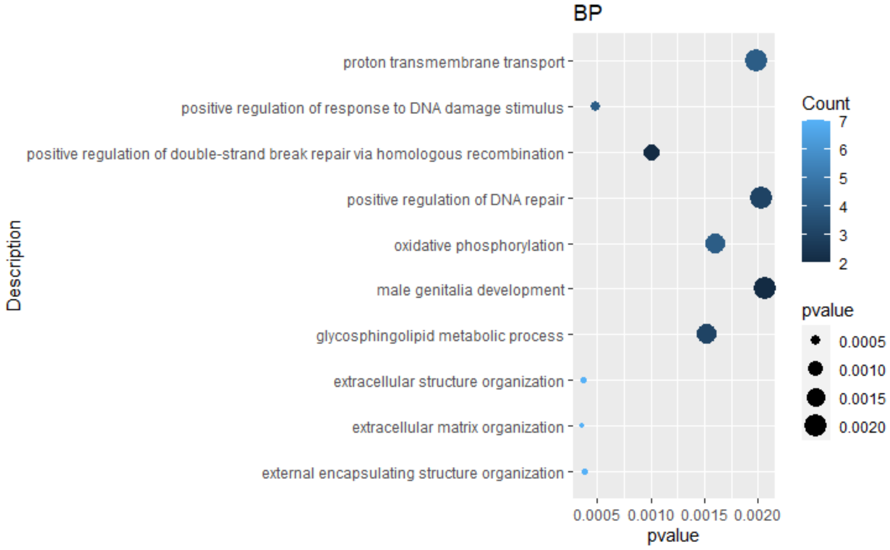

Figure 4(Biological processes that are enriched by downregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

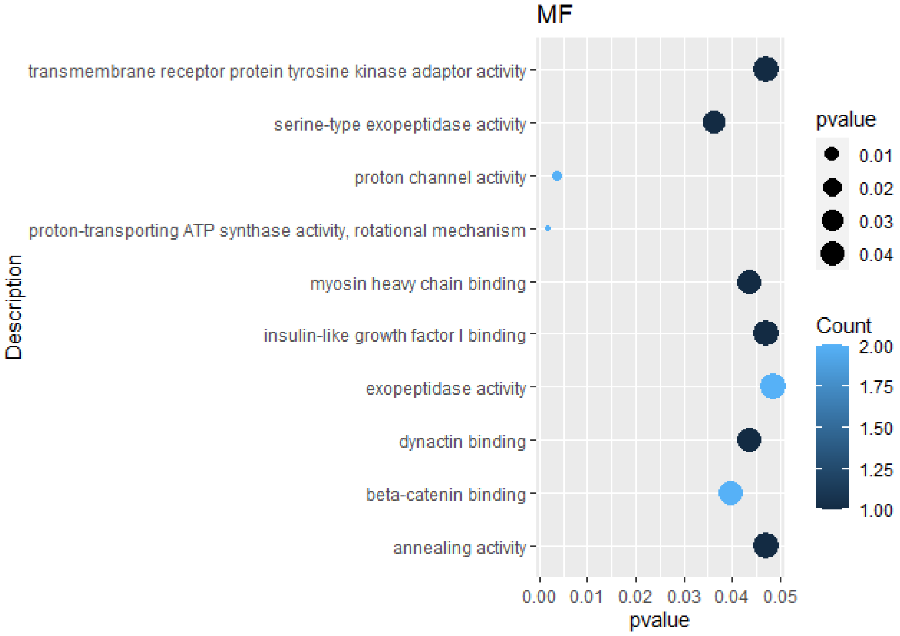

Figure 5(Molecular functions that are enriched by downregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

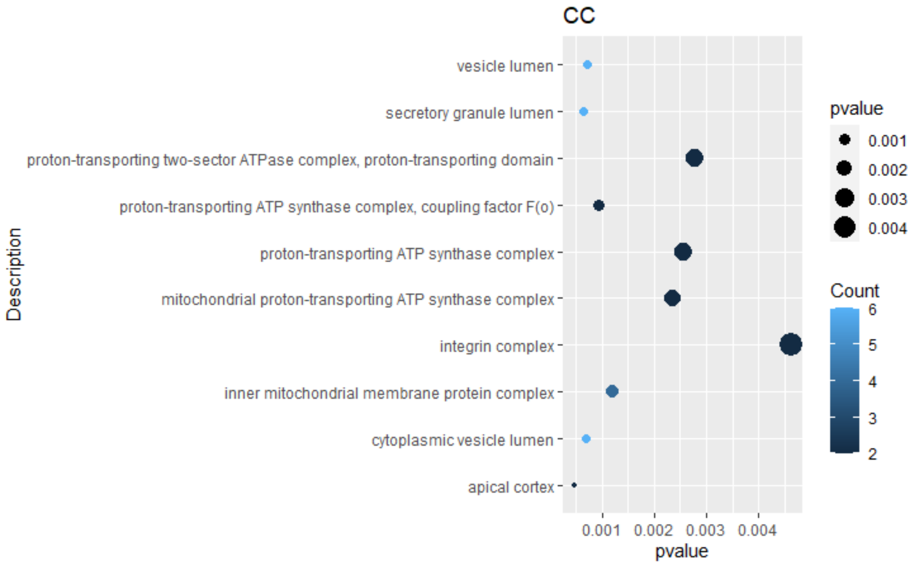

Figure 6(Cellular component that are enriched by downregulated differentially expressed genes between neoplasm vs. adjacent normal tissues)

4.4 KEGG Pathways Analysis of the Upregulated and Downregulated Genes

Through KEGG pathways analysis, the dysregulated biological pathways can be identified that are affected by the up- and down-regulated genes. The up-regulated and down-regualated genes were found to be statistically significant in proteoglycans in cancer, PI3K-Akt signaling pathway, MAPK signaling pathway, JAK-STAT signaling pathway, alcohol liver disease and similar pathways which indicates that those biological pathways that have been reported previously in various different cancer types and disease are dysregulated to the up and downregulation of various different genes due to neoplasm condition as compared to the normal adjacent tissue of the neoplasm. This indicates that the genes are directly and significantly deregulating cancer-inducing biological pathways.

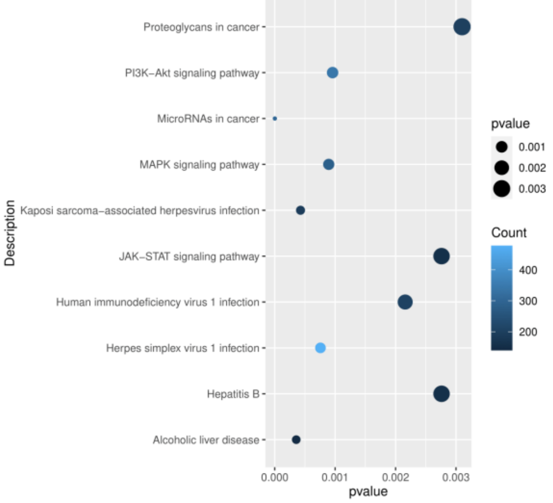

Figure 7(KEGG pathways analysis that are dysregulated by the differentially expressed genes, PI3K-Akt, MAPK signaling pathway, JAK-STAT pathway are crucial pathways identified in neoplasm vs normal tissues)

5.1 Discussion

Colorectal remains one of the most common cancers constituting chronic painful conditions that affects an individual's quality of life because rectum of the patient is severely affected [4]. However, the potential molecular alterations of colorectal cancer remain uncertain [1, 3, 14]. Understanding the underlying pathogenesis of colorectal cancer is of critical importance for diagnosis, prognosis, and identifying drug targets [3]. As high-throughput RNA sequencing can provide information regarding the expression levels of thousands of genes in the human genome simultaneously, this methodology has been widely used to predict the potential diagnostic and therapeutic targets for various other diseases and cancers [19-21, 24]. In the present study, we extracted the data from E-MTAB-8448, which includes 58 normal samples and 66 neoplasm samples. We identified 23 up-regulated including ITGAL, BIRC3, CETP, LYZ, CALB1, ALDH1A2, PHLDA1, DUOX2, ALMS1P1 and IRS1, and 141 down-regulated including RNF10, CHPF2, FAM168A, OPHN1, LYRM2, COL9A3, CBY1, C14ORF93, RIMS4 and TBL1X DEGs between colorectal cancer samples using bioinformatics analysis. As cumulative evidence has shown that co-expressed genes normally represent those with similar expression profiles that also frequently participate in similar biological processes [20]. To better understand the interactions of the identified DEGs, we further performed GO and KEGG functional annotation and pathway enrichment analyses. These analyses suggested that the identified DEGs were mainly involved in alcohol metabolic process, response to fatty acids, protein tetramerization, homotetramerization, neutrophil activation, transmembrane receptor tyrosine-kinase adaptor activity, NAD(P)H oxidase activity, oxidoreductase activity, positive regulation of response to DNA damage stimulus, positive regulation of double-strand break repair via homologous recombination, positive regulation of DNA repair and key cancer-related pathways such as PI3K-Akt signaling pathway, MAPK signaling pathway, JAK-STAT signaling pathway, alcohol liver disease. The Janus Kinase/Signal Transducer and Activator of Transcription (JAK/STAT) pathway was found to be dysregulated which indicates that the pathway has been affected due to neoplasm behavior of the tissue [9, 24]. JAK/STAT pathway is considered as a key mediator of important biological signaling pathways such as cell proliferation, cell survival and the immune response [9, 11, 13, 24]. JAK/STAT pathway helps HPV in manipulating JAK/STAT signaling to evade the immune system and promote cell proliferation, enabling viral persistence and driving cancer development. Therefore, deregulated JAK/STAT signaling contributes to cancer progression and metastatic development [9, 11, 13, 24]. Several proteins involved in this pathway play an essential role in the development of colorectal cancer. The PI3k-Akt pathway has been reported previously to be involved in cell growth, tumor proliferation and plays a crucial role in colorectal cancer [5, 7]. PI3k-Akt cancer is a major intracellular signaling pathway that works by responding to the availability of nutrients, hormones, and growth factor stimulation. It plays a crucial role in the tumor cell growth and proliferation [5, 7]. Therefore, aberrations in its activation and functioning may lead to colorectal cancer development and progression. Thus, this study provides key insights into the upregulated and downregulated genes due to colorectal cancer. Further research is required to study these genes in-depth to find their interacting proteins and functions and target them through drugs such as inhibitors.

5. Conclusions

In conclusion, ITGAL, BIRC3, CETP may be key genes for colorectal cancer. Furthermore, the results of the present study suggested that several biological pathways PI3K-Akt signaling pathway, MAPK signaling pathway, JAK-STAT signaling pathway are commonly involved in the development of colorectal cancer and progression. This study thus provides further insights into the underlying pathogenesis of colorectal, which may facilitate the diagnosis and treatment of these diseases.

References

1. Wrobel, P. and S. Ahmed, *Current status of immunotherapy in metastatic colorectal cancer.* International Journal of Colorectal Disease, 2019. **34**(1): p. 13-25.

2. Yachida, S., et al., *Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer.* Nature Medicine, 2019. **25**(6): p. 968-976.
3. Kishore, C. and P. Bhadra, *Current advancements and future perspectives of immunotherapy in colorectal cancer research.* European Journal of Pharmacology, 2021. **893**: p. 173819.
4. Buccafusca, G., et al., *Early colorectal cancer: diagnosis, treatment and survivorship care.* Critical Reviews in Oncology/Hematology, 2019. **136**: p. 20-30.
5. Ma, Z., S. Lou, and Z. Jiang, *PHLDA2 regulates EMT and autophagy in colorectal cancer via the PI3K/AKT signaling pathway.* Aging, 2020. **12**(9): p. 7985-8000.
6. Marcuello, M., et al., *Circulating biomarkers for early detection and clinical management of colorectal cancer.* Molecular Aspects of Medicine, 2019. **69**: p. 107-122.
7. Narayanankutty, A., *PI3K/ Akt/ mTOR Pathway as a Therapeutic Target for Colorectal Cancer: A Review of Preclinical and Clinical Evidence.* Current Drug Targets, 2019. **20**(12): p. 1217-1226.
8. Okugawa, Y., W.M. Grady, and A. Goel, *Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers.* Gastroenterology, 2015. **149**(5): p. 1204-1225.e12.
9. Wang, J., et al., *The circular RNA circSPARC enhances the migration and proliferation of colorectal cancer by regulating the JAK/STAT pathway.* Molecular Cancer, 2021. **20**(1): p. 81.
10. Kim, J.H., *Chemotherapy for colorectal cancer in the elderly.* World Journal of Gastroenterology, 2015. **21**(17): p. 5158-5166.
11. Wei, C., et al., *Crosstalk between cancer cells and tumor associated macrophages is required for mesenchymal circulating tumor cell-mediated colorectal cancer metastasis.* Molecular Cancer, 2019. **18**(1): p. 64.
12. Bhullar, D.S., et al., *Biomarker concordance between primary colorectal cancer and its metastases.* EBioMedicine, 2019. **40**: p. 363-374.
13. Fang, Y., et al., *CPEB3 functions as a tumor suppressor in colorectal cancer via JAK/STAT signaling.* Aging, 2020. **12**(21): p. 21404-21422.
14. Heinimann, K., *[Hereditary Colorectal Cancer: Clinics, Diagnostics and Management].* Therapeutische Umschau. Revue Therapeutique, 2018. **75**(10): p. 601-606.
15. Sarkans, U., et al., *From ArrayExpress to BioStudies.* Nucleic Acids Research, 2021. **49**(D1): p. D1502-D1506.
16. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.
17. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.
18. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics (Oxford, England), 2009. **25**(16): p. 2078-2079.
19. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics (Oxford, England), 2015. **31**(2): p. 166-169.
20. Liu, S., et al., *Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2.* Journal of Visualized Experiments: JoVE, 2021(175).
21. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
22. Consortium, G.O., *Gene Ontology Consortium: going forward.* Nucleic Acids Research, 2015. **43**(Database issue): p. D1049-1056.
23. Webber, C., *Functional enrichment analysis with structural variants: pitfalls and strategies.* Cytogenetic and Genome Research, 2011. **135**(3-4): p. 277-285.

24. Jiang, L., et al., *Long non-coding RNA RP11-468E2.5 curtails colorectal cancer cell proliferation and stimulates apoptosis via the JAK/STAT signaling pathway by targeting STAT5 and STAT6.* Journal of experimental & clinical cancer research: CR, 2019. **38**(1): p. 465.