Exploring potential solutions to optimize cancer therapy with cell reprogramming using gene network analysis: Inspired by KAIST's work on colon cells

Ka Lam Tam, Jenny Hua, Aditya Verma, Alok Kumar Singh and Safwan Ahmad Saffi

Abstract

This project aims to computationally evaluate the feasibility of cancer cell reprogramming by identifying key genes within malignant networks. This data-driven approach provides a proof of concept that can guide future wet-lab validation. A gene expression profile (GSE44076) was downloaded from the Gene Expression Omnibus database (GEO). Differentially expressed genes (DEGs) were screened using the GEO2R tools. Moreover, a protein–protein interaction (PPI) network of the DEGs was constructed, functional enrichment analysis was performed and hub genes from the PPI were explored on STRING and with Microsoft Excel calculations. A total of 500 DEGs are screened, including 299 upregulated genes and 201 downregulated genes. DEGs were enriched in several biological processes, cellular components and molecular functions. For each dataset, we picked out the top 10 nodes with the most degree (edges) which we identified as hub genes. GTPBP4, RPF2, GRWD1, RRS1, WDR36, CEBPZ, DDX52, KRR1, MPHOSPH10 and PUM3 are picked out in GSE44067(Fig.10). In GSE21510, NOP56, GTPBP4, NOP58, RPF2, RRS1, GRWD1, NIFK, WDR12, BRIX1 and BYSL are selected(Fig.11). Among the two datasets, 4 genes: GTPBP4, RPF2, GRWD1 and RRS1 are shared which converge on ribosome biogenesis. These findings promote the understanding and provide a proof of concept of the molecular mechanism of molecular targets for cancer reprogramming.

Introduction

Cancer is one of the major threats to human life and health worldwide. Colorectal cancer (CRC) is one of the most common malignant tumors and ranks as the third most common cancer in the United States. It holds the second-highest mortality rate among cancer types, following lung cancer[1]. To date, surgery remains one of the primary and most effective strategies for early-stage cancers. However, the feasibility and outcomes of surgery highly depend on patient-specific circumstances, including cancer stages and physiological status. More than 50% of patients in stage III and IV will receive conventional chemo- and radio-therapy. However, most of them quickly develop acquired resistance. Although immunotherapy and targeted therapy have emerged as effective strategies in the past few years, their effects have been partially impeded due to cancer heterogeneity and the existence of cancer stem cells. Therefore, finding potential treatments that can globally manage cancer remains a crucial task[2].

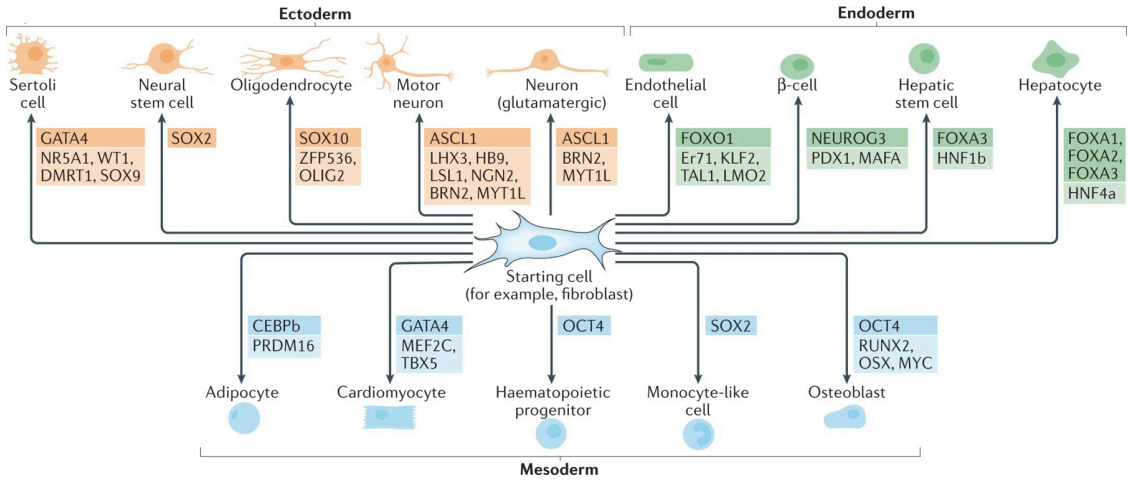
What is cell reprogramming?

Direct cell reprogramming (also known as transdifferentiation) refers to cell fate conversion without transitioning through an intermediary pluripotent state[3]. The idea of cancer cell reprogramming was suggested when the concept of cellular plasticity (the ability of a cell to

reprogram and change its phenotype identity[4]) was first proposed by Gurdon et al., which confirmed that terminally differentiated somatic cells could be reprogrammed into other lineages. Given that cancer cells are also genetically and epigenetically plastic, it has been suggested that they have the potential to regain benign cell functions by re-expressing lineage-specific genes[2].

Cell reprogramming is a complex and dynamic process that involves widespread changes in gene expression, as well as alterations in epigenetic states. Several approaches have been explored for inducing cell reprogramming, including the forced expression of lineage-specific transcription factors, chemical modulation of epigenetic regulators, and the use of small molecules to influence signaling pathways[2].

After the first report of the conversion of mouse embryonic fibroblasts (MEFs) into myoblasts by forced expression of MyoD, the so-called transcription factors were found to be capable of converting one cell type to another. Transcription factors or even a combination of them often play a crucial role in determining and maintaining cell function. For example, a combination of Gata4, Mef2c, and Tbx5 was found to be essential for heart development[3]. The image below shows examples of transition factors for different conversions across germ layers.



[1]

Given the central role of transcription factors in maintaining cellular identity, their dysregulation is particularly relevant in cancer, where abnormal gene expression drives malignant transformation. Hence, in cancer cells, transcription factors are seen as transcriptional regulators that modulate gene expression in the intricate layers of gene regulation. Subsequent studies have demonstrated that benign and malignant cells show distinct patterns of gene expression, highlighting key transcriptional differences that may underlie the malignant phenotype. This discovery provided the foundation for identifying molecular targets that could be manipulated to revert cancer cells toward a more normal state[4]. A recent study from KAIST (Korea Advanced Institute of Science and Technology) exemplifies this approach by building a Boolean network model (BENEIN) to analyze gene regulatory interactions in colon cancer cells. This model identified three master regulators: MYB, HDAC2, and FOXA2, whose simultaneous inhibition prompted colon cancer cells in vitro to revert toward a normal-like intestinal phenotype and significantly suppressed malignancy, as evidenced by reduced tumor growth in mouse models[6].

Given that a wet-lab approach requires time, resources, and lab facilities, we have chosen to use a data-driven prototype based on real cancer gene expression data to explore the

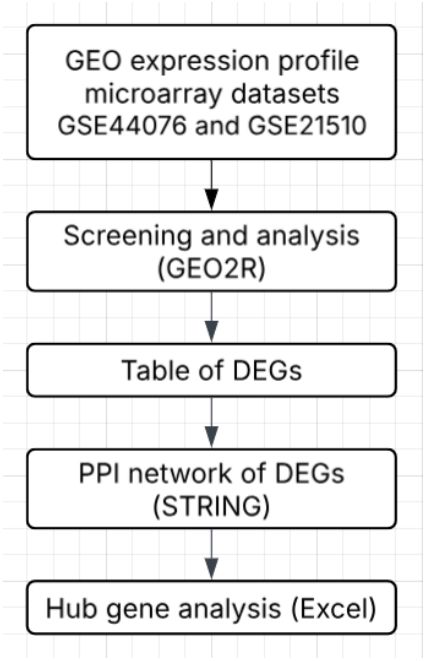
identification stage of cell reprogramming practically. By using existing data sets and online analytical tools, GEO2R and STRING, we are able to see how gene networks behave during malignancy[7]. A data-driven prototype also allowed us to test different conditions and large data samples much more efficiently than in a lab setting.

This project aims to computationally evaluate the feasibility of cancer cell reprogramming by identifying key genes within malignant networks. Using publicly available gene expression datasets and analytical tools (GEO2R and STRING), we identify differentially expressed genes, map their interactions and highlight potential genes as reprogramming targets. This data-driven approach provides a proof of concept that can guide future wet-lab validation.

Method

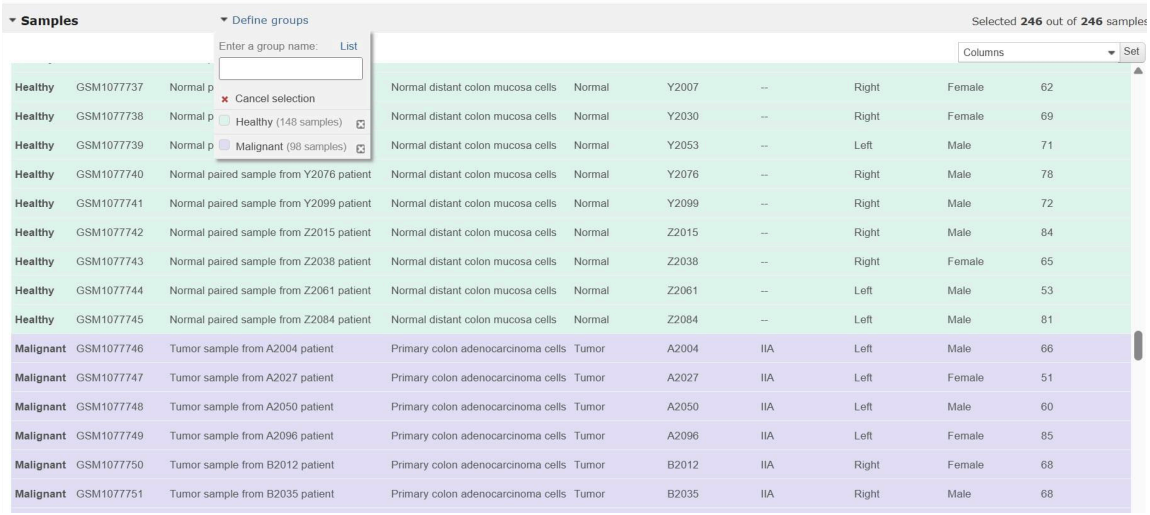
Dataset selection

Our methods are inspired by a 2021 study on the progression of cervical cancer[7]. The gene expression profile related to cancer progression was retrieved and downloaded from the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI). We have chosen GSE44076 and GSE21510 because they have been used to analyse hub genes by another group of scientists[8][9]. We have also decided to standardize our selection by only using colon cells gene expression dataset, as inspired by KAIST's work. The gene expression profile of GSE44076 includes 98 primary colon cancers and 98 normal distant colon mucosa which were selected from a series of cases with a new diagnosis of colorectal adenocarcinoma histologically confirmed. Additionally, samples of colon mucosa from 50 healthy donors without colonic lesions were obtained during colonoscopy[10]. The gene expression profile of GSE21510 includes a total of 148 microarray datasets obtained from LCM[11]. Below is a flowchart of this project.



Analysis of the dataset

GEO2R tool was used to analyse the two datasets where we grouped the samples according to the information provided in the dataset (normal vs. cancer), and compare gene expressions to identify differentially expressed genes.



Healthy	GSM1077737	Normal p	Normal distant colon mucosa cells	Normal	Y2007	--	Right	Female	62
Healthy	GSM1077738	Normal p	Normal distant colon mucosa cells	Normal	Y2030	--	Right	Female	69
Healthy	GSM1077739	Normal p	Normal distant colon mucosa cells	Normal	Y2053	--	Left	Male	71
Healthy	GSM1077740	Normal paired sample from Y2076 patient	Normal distant colon mucosa cells	Normal	Y2076	--	Right	Male	78
Healthy	GSM1077741	Normal paired sample from Y2099 patient	Normal distant colon mucosa cells	Normal	Y2099	--	Right	Male	72
Healthy	GSM1077742	Normal paired sample from Z2015 patient	Normal distant colon mucosa cells	Normal	Z2015	--	Right	Male	84
Healthy	GSM1077743	Normal paired sample from Z2038 patient	Normal distant colon mucosa cells	Normal	Z2038	--	Right	Female	65
Healthy	GSM1077744	Normal paired sample from Z2061 patient	Normal distant colon mucosa cells	Normal	Z2061	--	Left	Male	53
Healthy	GSM1077745	Normal paired sample from Z2084 patient	Normal distant colon mucosa cells	Normal	Z2084	--	Left	Male	81
Malignant	GSM1077746	Tumor sample from A2004 patient	Primary colon adenocarcinoma cells	Tumor	A2004	IIA	Left	Male	66
Malignant	GSM1077747	Tumor sample from A2027 patient	Primary colon adenocarcinoma cells	Tumor	A2027	IIA	Left	Female	51
Malignant	GSM1077748	Tumor sample from A2050 patient	Primary colon adenocarcinoma cells	Tumor	A2050	IIA	Left	Male	60
Malignant	GSM1077749	Tumor sample from A2096 patient	Primary colon adenocarcinoma cells	Tumor	A2096	IIA	Left	Female	85
Malignant	GSM1077750	Tumor sample from B2012 patient	Primary colon adenocarcinoma cells	Tumor	B2012	IIA	Right	Female	68
Malignant	GSM1077751	Tumor sample from B2035 patient	Primary colon adenocarcinoma cells	Tumor	B2035	IIA	Right	Male	68

GEO2R applies the limma (Linear Models for Microarray Data) package in R to calculate fold changes and adjusted p-values, correcting for multiple testing using the Benjamini–Hochberg false discovery rate (FDR). Genes with an adjusted p-value < 0.05 and |log2 fold change| ≥ 1 were considered significantly differentially expressed. The resulting DEG list was then exported for network analysis in STRING.

Construction of the PPI network

The DEGs identified from GEO2R were entered into the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database to explore their potential protein-level interactions. We set the species to *Homo sapiens* and applied a medium confidence score cut-off of 0.4 to ensure reliable interactions while still capturing relevant connections. STRING generated a network of nodes (proteins) and edges (interactions), which was then exported as a table of interactions. This table was imported into an Excel file to allow for further analysis and identification of hub genes by examining the degree of connectivity for each node.

Results

Identification of hub genes

The dataset was then imported into Microsoft Excel, where the degree of connectivity for each protein was calculated using the COUNTIF formula by counting the number of interactions (edges) associated with each node. Proteins with the highest number of connections were considered hub genes as their high degree of interaction suggests an important regulatory role within the malignant gene network.

Identification of DEGs

By analysing both GSE44076 and GSE21510, the top 250 DEGs are found for each dataset.

For GSE44076, 139 upregulated genes (log2 fold change >0) and 111 downregulated genes (log2 fold change <0) were identified, and for GSE21510, 160 upregulated genes and 90 downregulated genes were identified. Among the 2 datasets, 52 genes were shared: *ABCA8 ABCG2, ACADS, APPL2, AQP8, ATP11A, BEST4, C11orf86, C20rf88, CA1, CA4, CA7, CBFB, CBX3, CDKN2B, CEACAM7, CITED2, CLDN1, CSE1L, DDX21, FOXQ1, GCNT2, GLTP, GNA11, GTPBP4, GUCA2A, GUCA2B, HIGD1A, HS2ST1, IL6R, LDHD, MMP28, NFE2L3, NUFIP1, OSBPL3, PLCD1, POLR1B, PPM1H, PPP2R3A, SCARA5, SCIN, SLC4A4, SLC6A6, TEX11, TGFBI, TP53INP2, UGP2, USP2, WDR75, XPOT, ZNF575 and ZZEF1*. A section of the tables is shown in Fig.1 and Fig.2. By analysing the mean-difference plot, we realised that upregulated genes have a log2 fold change of >0, whereas downregulated genes have a log2 fold change of <0. This is further supported as the mean-different plots(Fig. 3) for both datasets show the same results. Volcano plots (Fig.4) helped us identify genes that are strongly differentially expressed and statistically significant as they combine both the log2 fold change and the -log10(p-value). Hub genes are therefore outliers which reinforced their relevance when we later constructed the PPI network. Furthermore, the UMAP (Uniform Manifold Approximation and Projection) plot (Fig.5) allowed us to visualise the overall expression patterns between the malignant and normal samples. The distinct separation between the groups in both datasets suggested that they capture biologically relevant differences, providing confidence in the downstream differential expression analysis[12]. Any overlaps between clusters could indicate heterogeneity within the cancer samples, which is consistent with the complexity of tumour biology.

ID	adj.P.Val	P.Value	t	B	logFC	Gene.Symbol	GB_LIST
11728232_a	9.75E-119	2.93E-123	-46.113183	271.03686	-4.94	CLDN1	NM_021101
11735833_a	9.75E-119	3.95E-123	-46.051222	270.74225	-4.77	KIAA1199	NM_018689
11719434_a	2.62E-115	1.59E-119	-44.347503	262.51295	-3.28	ETV4	NM_001079675,NM_001986
11728234_a	1.08E-114	8.77E-119	-44.003048	260.81843	-4.28	CLDN1	NM_021101
11739128_a	8.76E-114	8.87E-118	-43.539986	258.52375	-3.59	CDH3	NM_001793
11732838_at	9.71E-114	1.18E-117	43.483185	258.24095	6.74	GUCA2B	NM_007102
11721993_at	4.06E-113	5.75E-117	-43.168324	256.66801	-3.39	SLC6A6	NM_001134367,NM_001134368,NM_003043
11715637_a	6.51E-112	1.05E-115	42.595227	253.78176	1.71	UGP2	NM_001001521,NM_006759
11737294_a	1.56E-106	2.85E-110	40.196139	241.36319	7	TMIGD1	NM_206832
11724538_a	2.51E-106	5.08E-110	40.087494	240.78759	5.82	ABCG2	NM_004827
11733581_a	3.21E-105	7.37E-109	39.588574	238.1292	4.52	CA7	NM_001014435,NM_005182
11726764_at	3.21E-105	7.80E-109	39.578166	238.07348	6.95	AQP8	NM_001169
11750604_a	1.19E-104	3.27E-108	-39.312756	236.64883	-2.34	GTF2IRD1	NM_005685,NM_016328
11722783_at	1.19E-104	3.38E-108	-39.306555	236.61546	-4.62	FOXQ1	NM_033260
11759464_at	9.50E-104	2.89E-107	38.911982	234.48402	5.05	OTOP2	NM_178160
11742188_a	2.23E-102	7.21E-106	38.325189	231.28466	4.5	SLC4A4	NM_001098484,NM_001134742,NM_003759
11747996_a	1.97E-101	6.78E-105	-37.920446	229.05702	-3.07	ETV4	NM_001079675,NM_001986
11721557_a	1.01E-100	3.70E-104	37.615871	227.36933	4.14	ABCA8	NM_007168
11758134_s	1.13E-99	4.35E-103	-37.176517	224.91745	-2.65	PPM1H	NM_020700
11746142_a	2.23E-98	9.02E-102	36.641199	221.90202	3.24	ZNF611	NM_001161499,NM_001161500,NM_001161501,NM_030972
11758028_s	2.54E-98	1.08E-101	-36.609337	221.72157	-5.46	FOXQ1	NM_033260
11717822_a	2.44E-97	1.09E-100	36.205476	219.42462	4.15	SLC4A4	NM_001098484,NM_001134742,NM_003759
11719811_a	3.01E-97	1.40E-100	-36.161303	219.1723	-3.42	TRIB3	NM_021158
11729582_s	3.70E-97	1.80E-100	36.117981	218.92463	7.07	CA1	NM_001128829,NM_001128830,NM_001128831,NM_001164830,NM_001738
11729583_x	2.18E-96	1.11E-99	35.803036	217.11791	7.81	CA1	NM_001128829,NM_001128830,NM_001128831,NM_001164830,NM_001738
11719591_s	2.86E-95	1.50E-98	35.353409	214.51946	1.61	GLTP	NM_016433
11715813_at	1.12E-94	6.14E-98	35.112806	213.11968	2.59	HIGD1A	NM_001099668,NM_001099669,NM_014056
11723826_a	2.02E-94	1.15E-97	35.006121	212.49693	4.57	C2orf88	NM_001042519,NM_001042520,NM_001042521,NM_032321
11734322_at	3.29E-94	1.93E-97	34.917553	211.97894	3.49	BMP3	NM_001201
11742938_at	5.79E-94	3.52E-97	-34.815623	211.38171	-3.64	ASCL2	NM_005170
11734320_a	5.87E-93	3.69E-96	34.418447	209.04328	3.98	SLC17A4	NM_005495

GSE44076.top.table

Fig.1 Top 250 DEGs from GSE44076

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
205200_at	1.48E-56	3.31E-61	28.0609347	128.897454	3.08192017	EXOSC7	///CLEC exosome component 7///C-type lectin domain family 3 member B
1559977_a_a	1.48E-56	5.41E-61	27.9504433	128.410616	2.12828942	SLC25A34	solute carrier family 25 member 34
209612_s_at	1.96E-55	1.22E-59	27.259222	125.333764	4.93503471	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide
219669_at	1.96E-55	1.54E-59	27.2069574	125.096907	5.14802107	CD177	CD177 molecule
209613_s_at	1.96E-55	1.80E-59	27.1737937	124.949719	4.46240989	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide
241765_at	3.42E-55	3.76E-59	27.0119569	124.219879	2.82673013	CPM	carboxypeptidase M
207502_at	3.98E-55	5.09E-59	26.9454233	123.918955	5.72215882	GUCA2B	guanylate cyclase activator 2B
243403_x_at	8.67E-55	1.27E-58	26.746678	123.017	2.29443394	CPM	carboxypeptidase M
230788_at	1.01E-54	1.66E-58	26.6884632	122.751939	4.58670084	GCNT2	glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group)
211494_s_at	1.83E-53	3.34E-57	26.0417573	119.780693	3.97960422	SLC4A4	solute carrier family 4 member 4
207003_at	2.01E-53	4.05E-57	26.0006816	119.590307	5.19619482	GUCA2A	guanylate cyclase activator 2A
222549_at	3.53E-53	7.74E-57	-25.862732	118.949446	-4.9442311	CLDN1	claudin 1
219909_at	6.33E-53	1.51E-56	25.7213438	118.290257	4.54784086	MMP28	matrix metallopeptidase 28
1554522_at	8.78E-53	2.25E-56	25.6365038	117.893563	2.82335616	CNNM2	cyclin and CBS domain divalent metal cation transport mediator 2
219267_at	1.10E-52	3.02E-56	25.5743121	117.602221	2.92563755	GLTP	glycolipid transfer protein
207530_s_at	1.61E-52	4.71E-56	25.4804568	117.161667	1.65878324	CDKN2B	cyclin dependent kinase inhibitor 2B
205861_at	1.99E-52	6.42E-56	25.4150419	116.853985	2.94474128	SPIB	Spi-B transcription factor
1552296_at	1.99E-52	6.55E-56	25.4111122	116.835485	4.82155704	BEST4	bestrophin 4
229337_at	6.79E-52	2.36E-55	25.1422976	115.565551	3.52030191	USP2	ubiquitin specific peptidase 2
201479_at	7.79E-52	2.85E-55	-25.102935	115.378858	-2.4966837	MIR664B	///SNOI microRNA 664b///small nucleolar RNA, H/ACA box 56///dyskerin pseudouridine synthase 1
204700_x_at	1.17E-51	4.51E-55	-25.007441	114.92516	-2.6296692	DIEXF	digestive organ expansion factor homolog (zebrafish)
219230_at	4.26E-51	1.71E-54	24.7300765	113.601081	3.32338459	TMEM100	transmembrane protein 100
205945_at	5.98E-51	2.52E-54	24.6509502	113.221622	4.08868504	IL6R	interleukin 6 receptor
204699_s_at	6.34E-51	2.78E-54	-24.630073	113.121375	-2.0047169	DIEXF	digestive organ expansion factor homolog (zebrafish)
212160_at	7.56E-51	3.46E-54	-24.585482	112.90708	-2.3338146	XPOT	exportin for tRNA
205125_at	8.30E-51	3.95E-54	24.5581075	112.775405	2.20990122	PLCD1	phospholipase C delta 1
219177_at	1.15E-50	5.81E-54	-24.478508	112.391992	-2.3320209	BRIX1	BRX1, biogenesis of ribosomes
226177_at	1.15E-50	5.90E-54	24.4754159	112.377083	2.39487403	GLTP	glycolipid transfer protein
225806_at	3.33E-50	1.77E-53	-24.250701	111.290378	-4.0879294	AJUBA	ajuba LIM protein
205464_at	5.92E-50	3.25E-53	24.126492	110.68703	4.38883581	SCNN1B	sodium channel epithelial 1 beta subunit
207504_at	6.44E-50	3.65E-53	24.1025764	110.57064	3.01536868	CA7	carbonic anhydrase 7
209434_s_at	1.06E-49	6.23E-53	-23.994049	110.041577	-3.1562269	PPAT	phosphoribosyl pyrophosphate amidotransferase
232245_at	1.12E-49	6.76E-53	23.9773865	109.960217	2.90841341	SLC25A34	solute carrier family 25 member 34
212601_at	1.66E-49	1.04E-52	23.889311	109.529593	2.89401283	ZZEF1	zinc finger ZZ-type and EF-hand domain containing 1
209420_s_at	1.66E-49	1.06E-52	23.8859976	109.513374	3.051617	SMPD1	sphingomyelin phosphodiesterase 1

GSE21510.top.table

Fig.2 Top 250 DEGs from GSE21510

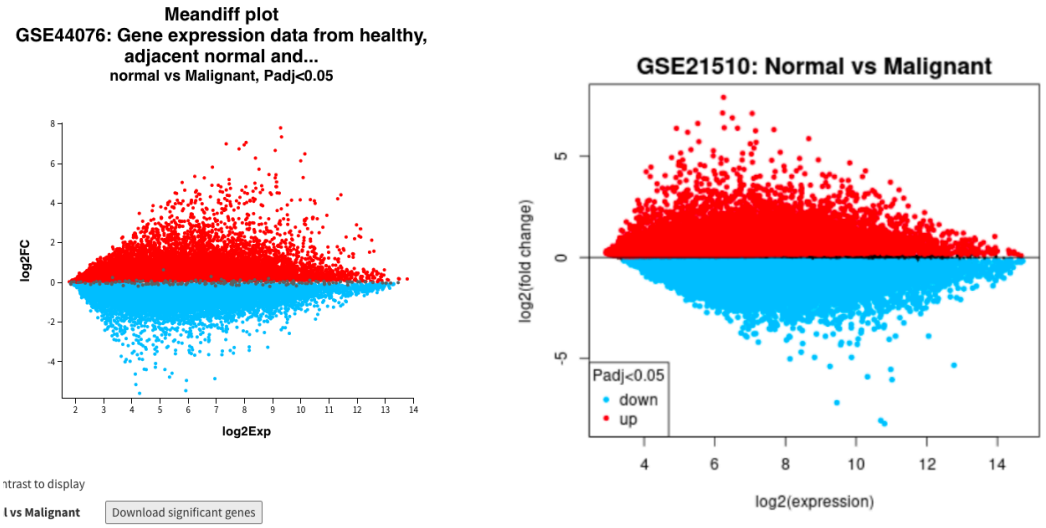


Fig.3 Mean-different plots of Top 250 DEGs from GSE44076 and GSE21510

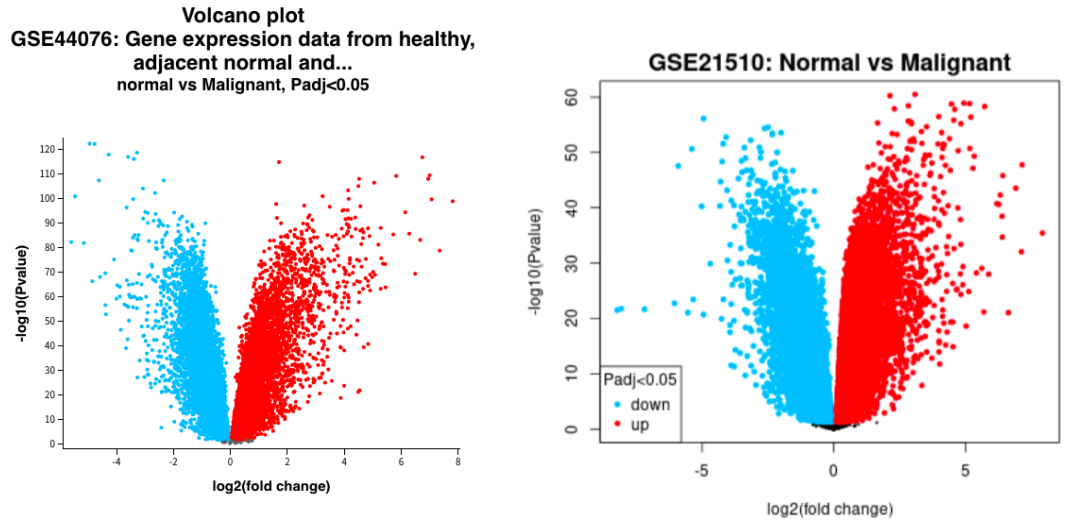


Fig.4 Volcano plots of Top 250 DEGs from GSE44076 and GSE21510

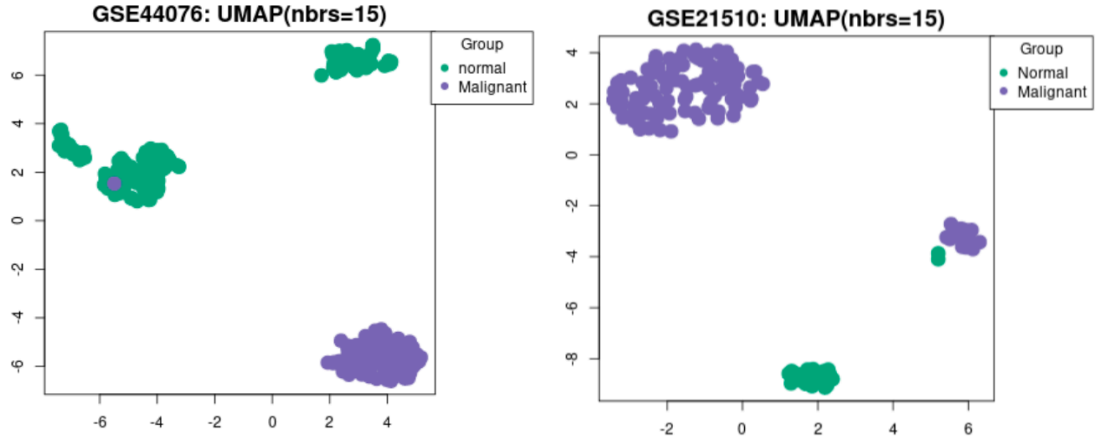


Fig.5 UMAP plots of Top 250 DEGs from GSE44076 and GSE21510

PPI network construction

A total of 500 genes were uploaded to STRING database. The PPI network of GSE44076 is shown in Fig.6 which includes 243 nodes and 4296 edges. The functional enrichment analysis in this PPI network included 37 clusters, 84 GO terms, 1 KEGG pathway, 4

Reactome pathways and 14 protein domains. According to Fig.7, it also revealed that most of the genes were associated with broad biological processes like cellular processes and metabolism. Specifically, many were enriched in categories such as organic substance metabolic process, cellular metabolic process and primary metabolic process, indicating that the network is strongly involved in fundamental metabolic pathways which are essential for cancer cell survival and proliferation. Moreover, beyond broad categories, the clusters found to be enriched in more significant ones which are cellular biogenesis, RNA processing and maturation. This could suggest that tumor cells require elevated biogenesis to sustain rapid proliferation and exploit RNA processing pathways to alter gene expression in their favor[13]. Other top categories in GO terms are also shown in Fig.7.

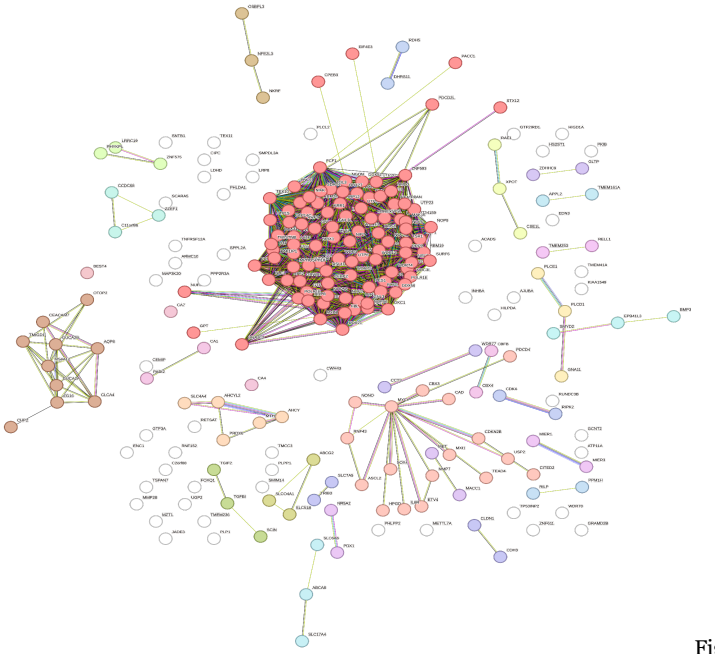
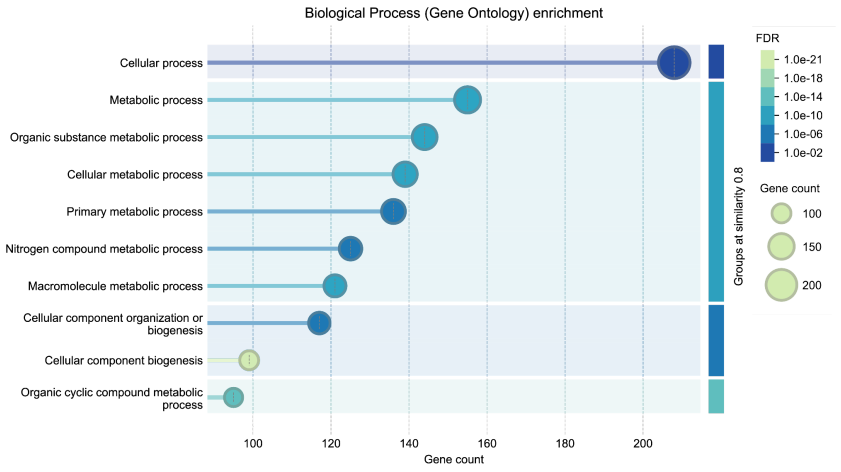
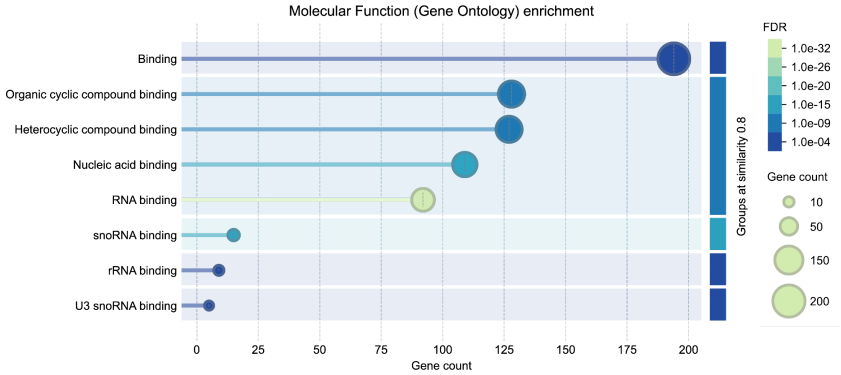


Fig.6 PPI network of DEGs from GSE44076





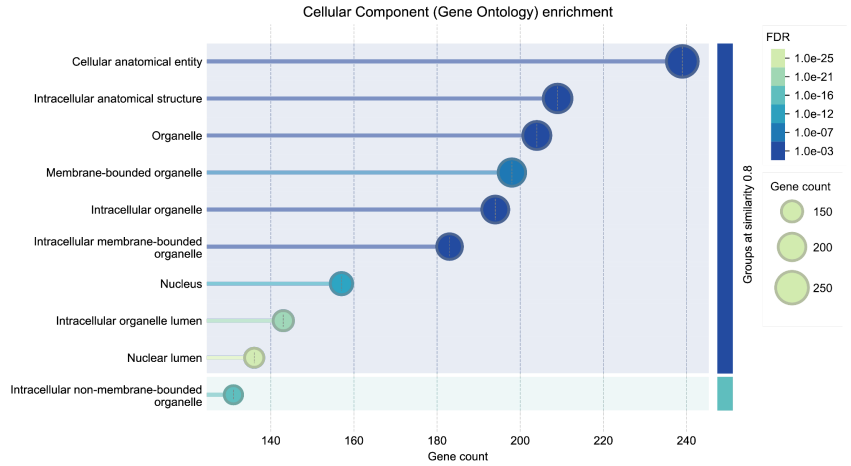


Fig.7 gene count and FDR tables of GO terms from GSE44076

The PPI network of GSE21510 is shown in Fig.8, which includes 250 nodes and 4518 edges. The functional enrichment analysis in this PPI network included 30 clusters, 93 GO terms, 2 KEGG pathways, 5 Reactome pathways and 14 protein domains. The analysis (Fig.9) also revealed that most of the genes were associated with cellular processes, metabolism, RNA processing and maturation, along with other categories in molecular function and cellular component, which also showed the same results as GSE44076. These genes express proteins and then interact functionally in both PPI networks, revealing their role in the progression of colon cancer.

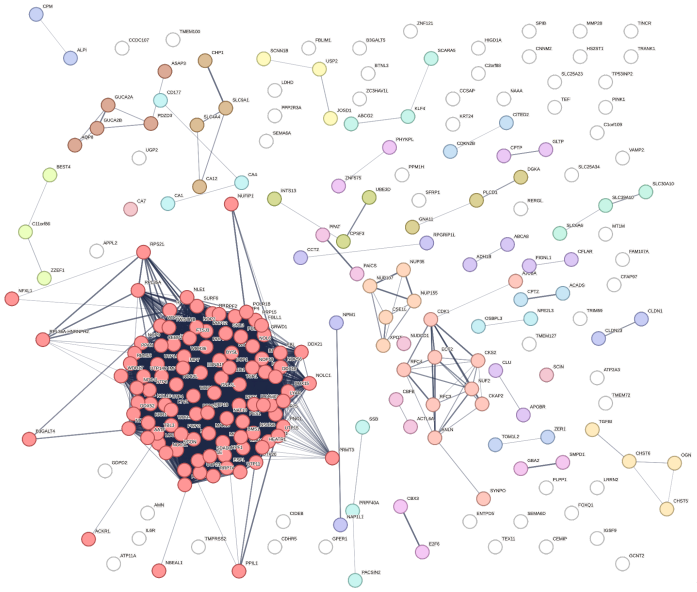
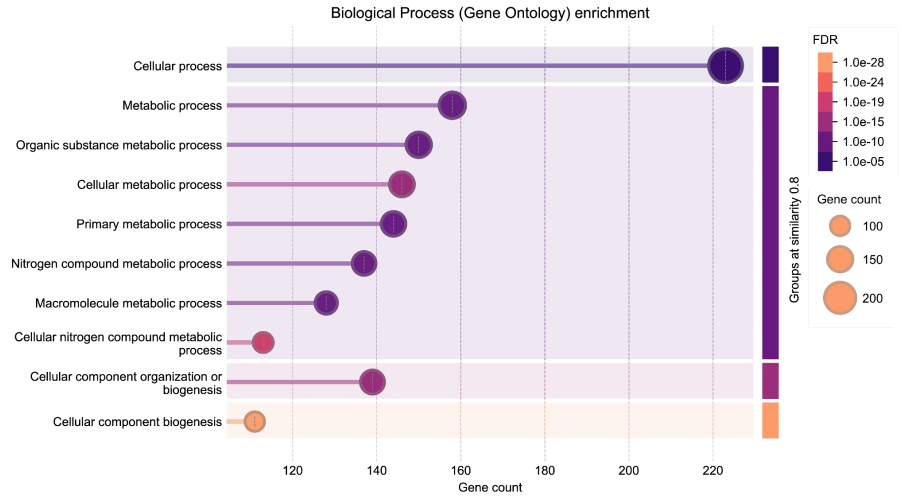


Fig.8 PPI network of DEGs from GSE21510



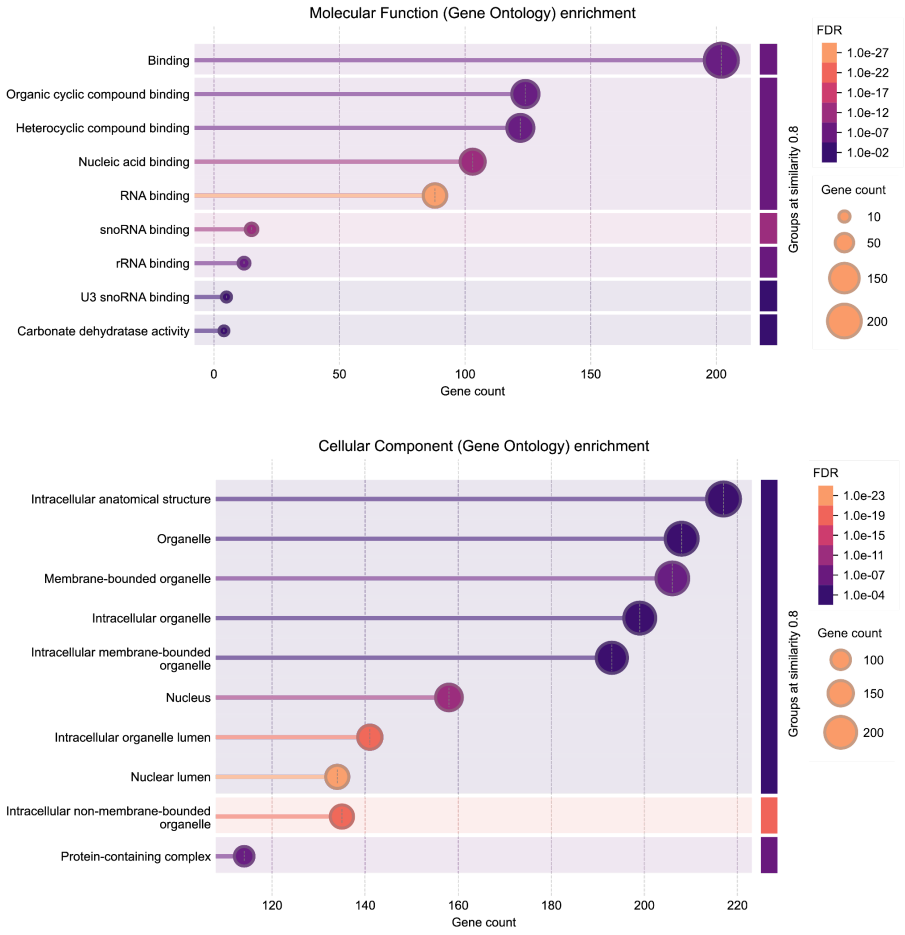


Fig.9 gene count and FDR tables of GO terms from GSE21510

Identification of hub genes

The interaction tables for each dataset were downloaded from STRING and exported as EXCEL spreadsheets for further analysis. For each dataset, we picked out the top 10 nodes with the most degree (edges) which we identified as hub genes. GTPBP4, RPF2, GRWD1, RRS1, WDR36, CEBPZ, DDX52, KRR1, MPHOSPH10 and PUM3 are picked out in GSE44067(Fig.10). In GSE21510, NOP56, GTPBP4, NOP58, RPF2, RRS1, GRWD1, NIFK, WDR12, BRIX1 and BYSL are selected(Fig.11). Among the two datasets, 4 genes: GTPBP4, RPF2, GRWD1 and RRS1 are shared, suggesting these genes are promising or potential targets for reprogramming.

	Nodes	Appearances N1	Appearances N2	Total Appearances
2	GTPBP4	11	9	20
3	RPF2	5	15	20
4	GRWD1	11	8	19
5	RRS1	4	15	19
6	WDR36	2	16	18
7	CEBPZ	16	1	17
8	DDX52	15	2	17
9	KRR1	9	8	17
10	MPHOSPH10	8	9	17
11	PUM3	7	10	17
12	BYSL	16	0	16
13	HEATR1	10	6	16
14	PWP2	6	10	16
15	UTP4	2	14	16
16	OKC1	12	3	15
17	TEX10	3	11	14
18	GUCA2B	7	5	12
19	SLC26A3	3	8	11
20	CLCA4	7	3	10
21	GUCA2A	7	3	10
22	MS4A12	4	6	10
23	AQP8	9	0	9
24	CA7	6	3	9
25	TMIGD1	1	8	9
26	CA4	6	2	8
27	CLCA1	7	1	8
28	CA1	7	0	7
29	DNAJC2	6	1	7
30	OTOP2	2	4	6
31	BEST4	5	0	5

Fig.10 Top 30 most connected genes from GSE44067

	Node	Appearances N1	Appearances N2	Total Appearances
2	NOP56	44	28	72
3	GTPBP4	56	14	70
4	NOP58	42	27	69
5	RPF2	26	43	69
6	RRS1	18	51	69
7	GRWD1	55	13	68
8	NIFK	49	19	68
9	WDR12	6	62	68
10	BRIX1	66	1	67
11	BYSL	65	2	67
12	RSL1D1	19	48	67
13	EBNA1BP2	59	7	66
14	FTSJ3	58	8	66
15	PES1	36	30	66
16	WDR36	5	61	66
17	WDR43	3	63	66
18	BOP1	65	0	65
19	MAK16	51	14	65
20	NOB1	47	18	65
21	RRP9	21	44	65
22	TSR1	18	47	65
23	UTP3	9	56	65
24	WDR3	6	59	65
25	WDR46	2	63	65
26	LTV1	51	13	64
27	MPHOSPH10	49	15	64
28	NOP2	40	24	64
29	RRP12	23	41	64
30	UTP4	8	56	64
31	DCAF13	59	4	63

Fig. 11 Top 30 most connected genes from GSE21510

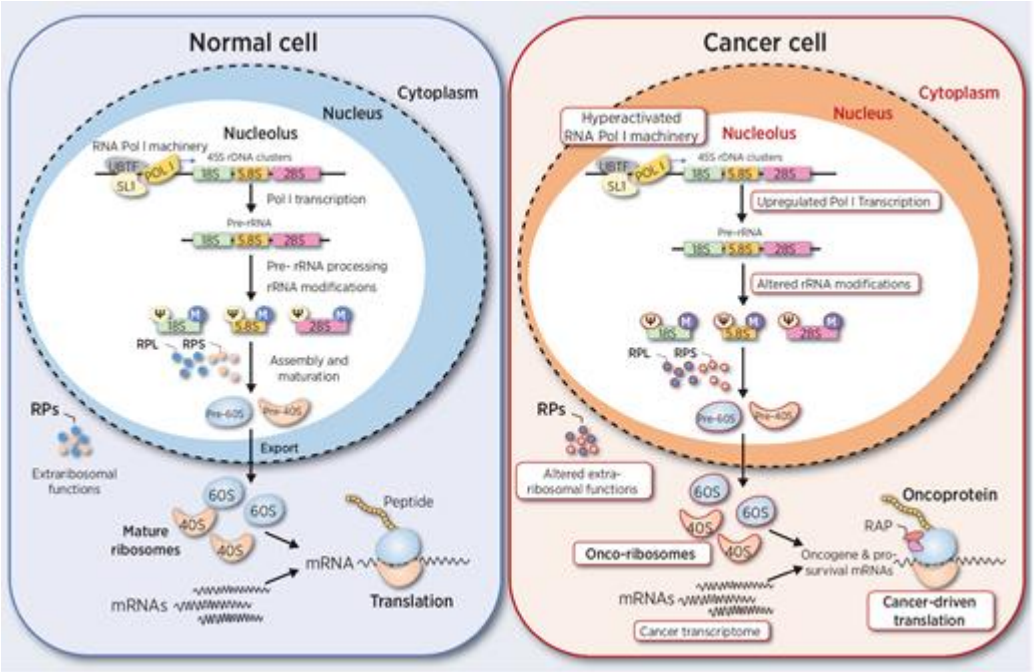
Discussion

Based on the analysis of the two datasets, this project deepened our understanding of the molecular mechanism of colon cancer and identified key hub genes. The hub genes GTPBP4, RPF2, GRWD1, and RRS1, which we identified in both PPI networks, serve as central regulators of gene interaction in colon cancer cells in different ways. GTBP4 (GTP binding protein 4) is a GTPase and functions as a molecular switch that can flip between two states: active(the molecule acts as a signal to trigger other events in the cell), when GTP is bound, and inactive, when GDP is bound[14]. It is said to be closely related to tumor metastasis, promotes cell motility and is detected in CRC metastatic tissues. GTPBP4 promotes CRC metastasis by primarily disrupting the actin cytoskeleton [15]. RPF2 (ribosome production factor 2 homolog) is a gene that enables 5S rRNA binding activity and is involved in protein localization to the nucleolus[16]. An elevated expression of RPF2 was observed in cancerous cells compared to normal colorectal cells which served as an indication that RPF2 may be involved in the activation process of Epithelial-Mesenchymal Transition(EMT) (a cellular program in which epithelial cells acquire a mesenchymal phenotype, resulting in increased invasiveness, enhanced stemness, and heightened resistance to therapeutic agents and immune responses in epithelial tumor cells), therefore enhancing the invasive and migratory

capabilities of CRC cells[1]. Additionally, GRWD1 (glutamate-rich WD repeat containing 1) encodes a glutamate-rich protein that contains five WD-repeat motifs which plays a critical role in ribosome biogenesis[17]. Moreover, GRWD1 was found to stimulate cell migration, induce EMT and promote colony formation; hence, it is positively correlated with tumour size. Interestingly, this glutamate-rich gene also activates the Notch signaling pathway which is involved in development, differentiation, cell proliferation and apoptosis. Some studies have shown that it also plays a regulatory role in malignant tumors[18]. Lastly, RRS1 (regulator of ribosome synthesis 1) enables 5S rRNA binding activity. It is involved in several processes, including mitotic metaphase chromosome alignment, protein localization to the nucleolus and ribosomal large subunit assembly[19]. Recent studies have shown that RRS1 interacts with RPF2 to form a complex that regulates the maturation of the 60S ribosomal subunit. In this way, it plays an important role in ribosome biogenesis. RRS1 is highly expressed in colorectal cancer (CRC) tissues, and its expression is inversely correlated with the survival of CRC patients[20].

Because all these hub genes converge on ribosome biogenesis, they represent attractive reprogramming targets. Aberrant cell growth and proliferation depend on hyperactive, in other words, dysregulated ribosome biogenesis, meaning increased protein synthesis and overactive translation. This is enabled by cellular regulatory pathways that are hijacked to tune transcription and translation. This is consistent with the acquisition of genetic and epigenetic alterations by cancer cells and changes in the regulatory layers of translation such as microRNAs and RNA-binding proteins that play significant roles during tumor progression and metastasis[21].

Thereby, modulating the expression of those hub genes could essentially reduce translational output, weaken metastatic potential and oppose excessive changes in ribosome biosynthesis and halt cell growth. Ultimately, pushing cells toward a less proliferative, more benign phenotype.[21]. Additionally, reprogramming hub genes could trigger a wider network effect and possibly shut down multiple malignant pathways in one go while sparing normal cells due to non-oncogene addiction[22], enhancing therapeutic effects. The diagram below outlines the process of ribosome biogenesis.



[21]

How this project informs future research

This project provides a validated bioinformatics pipeline as it demonstrates a clear, replicable and accessible workflow using public tools (GEO2R, STRING, basic Excel analysis) to move from raw genomic data to a list of high-value therapeutic targets. Excel is used over Cytoscape, considering that Excel is a common tool used in daily life. This essentially serves as a guideline for other young researchers to apply similar analysis to other cancer types. Moreover, successfully identifying known central players in colorectal cancer like GTPBP4 and RPF2, provides strong evidence that analyzing PPI networks built from DEGs is a valid strategy for uncovering key regulatory genes. This justifies further investment in more complex network medicine approaches in the future. The precise reduction of gene targets from thousands of DEGs to a handful of hub genes directly informs wet-lab research by providing a strong, data-driven hypothesis to test, which saves time, resources and funding.

Hub genes also assist in the discovery of more biomarkers. For example, receiver operating characteristic (ROC) analysis can be used to further evaluate their diagnostic value for targeted therapies[23].

Limitations

While informative, this study is a prototype and has several important limitations. The size of the GEO datasets used inherently limits the analysis, as it may not capture the full genetic diversity of cancer patients or account for the tumor microenvironment's influence on gene expressions, which play a crucial role in regulating pathways like EMT and ribosome biogenesis[24].

The PPI network from STRING represents a composite of interactions from various cell types and conditions. It is a static model that does not capture the dynamic, context-specific nature of gene regulatory networks within a living tumor. Moreover, STRING integrates predicted as well as experimental interactions, so some connections between genes may not actually occur in vivo, causing false positives. In terms of hub genes, identifying them based solely on

their degree is a useful first step, but it is too simplistic as it does not incorporate other important network metrics, such as "betweenness centrality" (how crucial a node is to connecting others[25]) or the direction of regulation (activation vs. inhibition). From our results, most of our hub genes are involved in ribosome biogenesis. Therefore, even if those hub genes are essential for cancer progression, targeting ribosome biogenesis can also harm normal proliferating cells, limiting therapeutic use[21].

Most importantly, this project is based on computational predictions, not functional validation as the entire project is in silico. The role of these hub genes in functionally maintaining the cancerous state and their reprogrammability remains a prediction until validated experimentally in cell and animal models. Due to this lack of clinical validation, we cannot guarantee that these genes can be safely targeted in humans. The lack of patient specificity should also be acknowledged, as it does not account for inter-patient heterogeneity and does not constitute a personalized medicine approach without further patient-specific data integration.

Potential next steps

To build upon this prototype and overcome its limitations, future directions can include both experimental and computational strategies. CRISPR/Cas9 can be used in cancer research to edit genomes for the exploration of tumorigenesis and development. More specifically, CRISPR activation (CRISPRa) can be used to epigenetically upregulate tumor-suppressor genes and CRISPR interference (CRISPRi) to silence oncogenes by providing a valid measure for deletion, thereby inhibiting tumour growth[26]. CRISPR techniques can be explored in lab models such as Patient-Derived Organoids, which preserve tumor biology, heterogeneity and show the advantages for editable genes. Orthotopic xenograft models are also used to test efficacy in vivo to assess the impact on tumor growth and metastasis[27]. Moreover, using single-cell RNA-seq can track transcriptomic changes at individual cell level and determine if a stable, reprogrammed state is achieved[28].

More recently, the use of AI has been studied in terms of its contribution to clinical science. Training AI models (e.g., Graph Neural Networks) on larger, multi-omics datasets to identify hubs that are consistently central across a large population can separate core CRC drivers from context-specific ones. AI can also be used to discover novel interactions by mining these networks for synthetic lethal interactions, so non-obvious secondary targets become essential only when a primary hub is perturbed.

From a personalised medicine approach, creating personalized gene expression profiles for individual patients can identify which hubs are most dominant in their specific cancer. This can be achieved by integrating genomic and transcriptomic data from their tumor biopsies[29].

By addressing these next steps, the promising predictions of this prototype can be rigorously tested, refined, and translated into a tangible strategy for overcoming cancer through network reprogramming.

Conclusion

In conclusion, a comprehensive analysis of DEGs and pathways involved in the occurrence and development of colorectal cancer was performed. We explored and obtained key regulatory genes and pathways contributing to the progression of colorectal cancer which promote the understanding of molecular mechanisms and clinically related molecular targets for reprogramming from malignant cells to their benign states. This prototype, although preliminary, mirrors the strategy pioneered by KAIST, where gene network analysis was used to identify master regulators in colon cancer while providing a proof of concept for cell reprogramming.

Reference list

1. Cheng, C., Zhang, K., Lu, M., Zhang, Y., Wang, T., & Zhang, Y. (2024). RPF2 and CARM1 cooperate to enhance colorectal cancer metastasis via the AKT/GSK-3β signaling pathway. *Experimental Cell Research*, *444*(2), 114374. https://doi.org/10.1016/j.yexcr.2024.114374
2. Gong, L., Yan, Q., Zhang, Y., Fang, X., Liu, B., & Guan, X. (2019). Cancer cell reprogramming: a promising therapy converting malignancy to benignity. *Cancer Communications*, *39*(1), 48. https://doi.org/10.1186/s40880-019-0393-5
3. Wang, H., Yang, Y., Liu, J., & Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. *Nature Reviews Molecular Cell Biology*, *22*. https://doi.org/10.1038/s41580-021-00335-z
4. Wu, X., Huang, S., He, W., & Song, M. (2023). Emerging insights into mechanisms of trastuzumab resistance in HER2-positive cancers. *International Immunopharmacology*, *122*, 110602–110602. https://doi.org/10.1016/j.intimp.2023.110602
5. Danielsson, F., Skogs, M., Huss, M., Rexhepaj, E., O'Hurley, G., Klevebring, D., Ponten, F., Gad, A. K. B., Uhlen, M., & Lundberg, E. (2013). Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. *Proceedings of the National Academy of Sciences*, *110*(17), 6853–6858. https://doi.org/10.1073/pnas.1216436110
6. Gong, J., Lee, C., Kim, H., Kim, J., Jeon, J., Park, S., & Cho, K. (2024). Control of Cellular Differentiation Trajectories for Cancer Reversion. *Advanced Science*. https://doi.org/10.1002/advs.202402132
7. Wu, B., & Xi, S. (2021). Bioinformatics analysis of differentially expressed genes and pathways in the development of cervical cancer. *BMC Cancer*, *21*(1). https://doi.org/10.1186/s12885-021-08412-4
8. Yang, W., Ma, J., Zhou, W., Li, Z., Zhou, X., Cao, B., Zhang, Y., Liu, J., Yang, Z., Zhang, H., Zhao, Q., Hong, L., & Fan, D. (2018). Identification of hub genes and outcome in colon cancer based on bioinformatics analysis. *Cancer Management and Research*, *Volume 11*, 323–338. https://doi.org/10.2147/cmar.s173240
9. Wen, Q., O'Reilly, P., Dunne, P. D., Lawler, M., Van Schaeybroeck, S., Salto-Tellez, M., Hamilton, P., & Zhang, S.-D. (2015). Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC Systems Biology*, *9*(S5). https://doi.org/10.1186/1752-0509-9-s5-s4
10. *GEO Accession viewer*. (2024). Nih.gov. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44076
11. *GEO Accession viewer*. (2024). Nih.gov. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21510
12. Tjoonk, N. (2023, January 20). *What is a UMAP plot?* Single Cell Discoveries. https://www.scdiscoveries.com/blog/knowledge/what-is-a-umap-plot/
13. Stępiński, D. (2018). The nucleolus, an ally, and an enemy of cancer cells. *Histochemistry and Cell Biology*, *150*(6), 607–629. https://doi.org/10.1007/s00418-018-1706-5
14. *GTPBP4 GTP binding protein 4 [Homo sapiens (human)] - Gene - NCBI*. (2025). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/23560

15. Yu, H., Jin, S., Zhang, N., & Xu, Q. (2016). Up-regulation of GTPBP4 in colorectal carcinoma is responsible for tumor metastasis. *Biochemical and Biophysical Research Communications*, *480*(1), 48–54. https://doi.org/10.1016/j.bbrc.2016.10.010
16. *RPF2 ribosome production factor 2 homolog [Homo sapiens (human)] - Gene - NCBI*. (2025). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/84154
17. *GRWD1 glutamate rich WD repeat containing 1 [Homo sapiens (human)] - Gene - NCBI*. (2025). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/83743
18. Wang, Q., Ren, H., Xu, Y., Jiang, J., Wudu, M., Liu, Z., Su, H., Jiang, X., Zhang, Y., Zhang, B., & Qiu, X. (2019). GRWD1 promotes cell proliferation and migration in non-small cell lung cancer by activating the Notch pathway. *Experimental Cell Research*, *387*(2), 111806. https://doi.org/10.1016/j.yexcr.2019.111806
19. *RRS1 regulator of ribosome synthesis 1 [Homo sapiens (human)] - Gene - NCBI*. (2025). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/23212
20. Yan, X., Wu, S., Liu, Q., & Zhang, J. (2020). RRS1 Promotes Retinoblastoma Cell Proliferation and Invasion via Activating the AKT/mTOR Signaling Pathway. *BioMed Research International*, *2020*, 1–10. https://doi.org/10.1155/2020/2420437
21. Elhamamsy, A. R., Metge, B. J., Alsheikh, H. A., Shevde, L. A., & Samant, R. S. (2022). Ribosome Biogenesis: A Central Player in Cancer Metastasis and Therapeutic Resistance. *Cancer Research*, *82*(13), 2344–2353. https://doi.org/10.1158/0008-5472.can-21-4087
22. Chang, H. R., Jung, E., Cho, S., Jeon, Y.-J., & Kim, Y. (2021). Targeting Non-Oncogene Addiction for Cancer Therapy. *Biomolecules*, *11*(2), 129. https://doi.org/10.3390/biom11020129
23. Lei, X., Zhang, M., Guan, B., Chen, Q., Dong, Z., & Wang, C. (2021). Identification of hub genes associated with prognosis, diagnosis, immune infiltration and therapeutic drug in liver cancer by integrated analysis. *Human Genomics*, *15*(1). https://doi.org/10.1186/s40246-021-00341-4
24. Baghban, R., Roshangar, L., Jahanban-Esfahlan, R., Seidi, K., Ebrahimi-Kalan, A., Jaymand, M., Kolahian, S., Javaheri, T., & Zare, P. (2020). Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signaling*, *18*(1). https://doi.org/10.1186/s12964-020-0530-4
25. Zaoli, S., Mazzarisi, P., & Lillo, F. (2021). Betweenness centrality for temporal multiplexes. *Scientific Reports*, *11*(1), 4919. https://doi.org/10.1038/s41598-021-84418-z
26. Zhang, H., Qin, C., An, C., Zheng, X., Wen, S., Chen, W., Liu, X., Lv, Z., Yang, P., Xu, W., Gao, W., & Wu, Y. (2021). Application of the CRISPR/Cas9-based Gene Editing Technique in Basic research, diagnosis, and Therapy of Cancer. *Molecular Cancer*, *20*(1), 126.
27. Yoshida, G. J. (2020). Applications of patient-derived tumor xenograft models and tumor organoids. *Journal of Hematology & Oncology*, *13*(1). https://doi.org/10.1186/s13045-019-0829-z
28. Pascal Grobecker, Sakoparnig, T., & Nimwegen, E. van. (2024). Identifying cell states in single-cell RNA-seq data at statistically maximal resolution. *PLoS Computational Biology*, *20*(7), e1012224–e1012224. https://doi.org/10.1371/journal.pcbi.1012224
29. *Brain tumour biopsy - Macmillan Cancer Support*. (n.d.). Www.macmillan.org.uk. https://www.macmillan.org.uk/cancer-information-and-support/diagnostic-tests/brain-tumour-biopsy