DIFFERENTIAL EXPRESSION ANALYSIS OF LONG NON-CODING RNAs IN THE CANCER OF THE BILE DUCT.

ABSTRACT

Cancer is a heterogenous disease that can begin at the cellular level and metastasize to any number of organs. Till date the root cause of its origin was quiet ambiguous but with the birth of high throughput technologies, we notice a paradigm shift in the way we think of cancer.
In this study, I discovered that protein coding genes and non-coding genes create a network of interactions in the prognosis of cancer. To have a deeper understanding of my study and analysis, I worked with cholangiocarcinoma(cancer of the bile duct) from TCGA which I got from GDC portal to find differentially expressed genes of Non-Coding RNAs and Protein Coding RNAs that played major role in the formation of the cancer type due to the lncRNAs acting as CeRNAs (Competing Endogenous RNAs).

INTRODUCTION

According to the central dogma of molecular biology, a gene is transcribed into RNA which gets translated into proteins and proteins in-turn perform various complex activities in the body of an organism through various pathways. Though this is the general scenario, there is a twist to this. Present research using high throughput technologies shows that most of the genes (>98%) though transcribed, are not actually translated to proteins and are tagged as 'non-coding'(not coding – for proteins) and do not possess functional Open Reading Frames(ORFs). There are two groups of non-coding RNAs. Small non-coding RNAs (<200 nucleotides in length) and long non-coding RNAs (>200 nucleotides in length). Since a lot of research is already done and much is already known about miRNAs (small non-coding RNAs), focus in my project is laid on long non-coding RNAs. LncRNAs play a very important role in many mechanisms like chromatin re-modeling and transcriptional and post-transcriptional regulations. This is because they are capable of functioning as decoys, scaffolds and enhancers. Cancer is a disease characterized by genetic alterations. (National cancer Institute, NIH). Due to this versatility of lncRNAs, they play a vital role in mediating many human diseases in general, cancer in particular mainly by interacting with mRNAs and miRNAs and act as hallmarks of cancer[1].

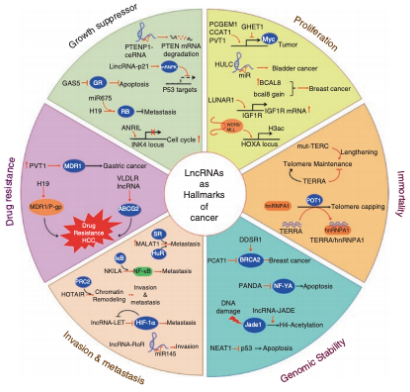


Fig:1:LcRNAs as hallmarks of cancer[2]

DATA DOWNLOAD AND METHODS:

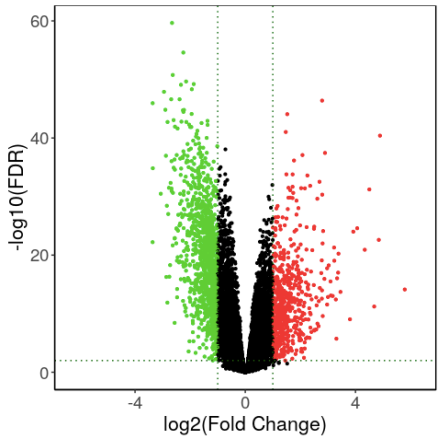
Cancer profiling has been attributed and circled around protein coding genes under the presumption that it's the proteins which are responsible for all the activities in the human body. Recent studies reveal that lncRNAs are associated with metastasis and invasion of many types of cancer.[1] To know how lncRNA profiling categorizes biliary cancer, focus is laid on studying the association of lncRNAs and cancer, which could be achieved by doing a differential gene expression analysis. To, understand the molecular mechanisms of these long non coding RNAs, it is mandatory to study their networking with other protein coding RNAs.[2]
The data was downloaded using command-line toolkit version of gdc-client, a software package that aids in downloading data from the gdc portal. A Bioconductor package called the GDCRNATools was also used to automatically download data into R. This package also has ggplot2,limma, EdgeR, DESeq2 as added libraries which facilitated analysis and visualization.[3]

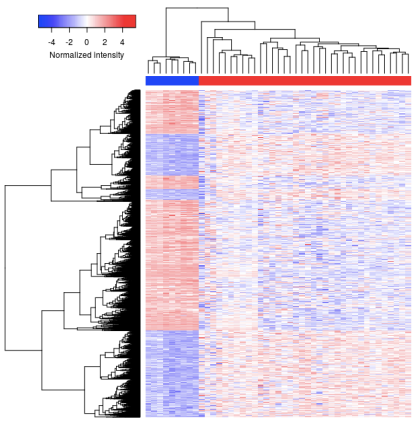
RESULTS

Enrichment Analysis:
Functional Enrichment as done to see how many genes were upregulated and how many were down regulated. The figure shows top ten upregulated genes of CeRNAs .
Top 10 Enriched genes

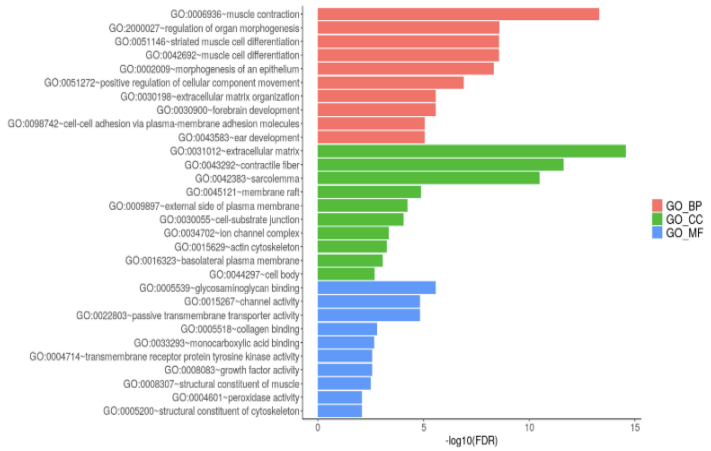
Counts	GeneRatio	BgRatio	pValue	FDR	foldEnrichment	geneID	geneSymbol
77	77/1353	326/16447	9.881094e-18	4.882249e-14	2.871191	ENSG00000087258/ENSG00000095303/ENSG00000...	GNAO1/PTGS1/ACTC1/PRKG1/SLMAP/ADRA1A/MYO...
47	47/1353	185/16447	1.527429e-12	2.515676e-09	3.088268	ENSG00000105707/ENSG00000128714/ENSG00000...	HPN/HOXD13/PRICKLE2/GATA3/SNAI2/FGF7/FGFR2/...
56	56/1353	249/16447	2.588835e-12	2.741716e-09	2.733868	ENSG00000105971/ENSG00000159251/ENSG00000...	CAV2/ACTC1/RARB/MYOCD/KIAA1161/CFL2/SCGB3...
71	71/1353	360/16447	2.774455e-12	2.741716e-09	2.397423	ENSG00000105971/ENSG00000159251/ENSG00000...	CAV2/ACTC1/SOX15/RARB/MYOCD/KIAA1161/CFL2...
87	87/1353	493/16447	5.909254e-12	4.866271e-09	2.145168	ENSG00000128714/ENSG00000163637/ENSG00000...	HOXD13/PRICKLE2/NDRG4/GATA3/SNAI2/FOXA1/PG...
73	73/1353	412/16447	2.580487e-10	1.275019e-07	2.153845	ENSG00000154188/ENSG00000133067/ENSG00000...	ANGPT1/LGR6/SEMA6D/GATA3/SNAI2/S100A14/KIT...
66	66/1353	390/16447	1.259644e-08	2.689102e-06	2.057161	ENSG00000105707/ENSG00000197565/ENSG00000...	HPN/COL4A6/ADAMTS5/NCAM1/SERPINB5/SH3PXD2...
65	65/1353	382/16447	1.306182e-08	2.689102e-06	2.068421	ENSG00000109819/ENSG00000087258/ENSG00000...	PPARGC1A/GNAO1/CRTAC1/PRKG1/DUOX2/ID4/DLX...
43	43/1353	219/16447	6.470292e-08	8.640463e-06	2.386785	ENSG00000165323/ENSG00000169851/ENSG00000...	FAT3/PCDH7/FAT2/AJUBA/TRO/CD200/DAB1/CDHR...
42	42/1353	212/16447	7.055669e-08	8.938990e-06	2.408254	ENSG00000105707/ENSG00000184564/ENSG00000...	HPN/SLITRK6/GATA3/DUOX2/LRIG1/CDH23/PLS3/FG...
50	50/1353	286/16447	2.649137e-07	2.469695e-05	2.125166	ENSG00000105707/ENSG00000105971/ENSG00000...	HPN/CAV2/GATA3/SNAI2/PGF/RAP1GAP/MCC/TGFBR...

Heatmap (Hirarchical clustering) and volcano plot give the visual interpretation of all DEGs





GO ontology was done to predict three biological concepts
Biological Processes(BP), Cellular Component(CC), and Molecular Function(MF)



Top 10 DE Protein Coding genes

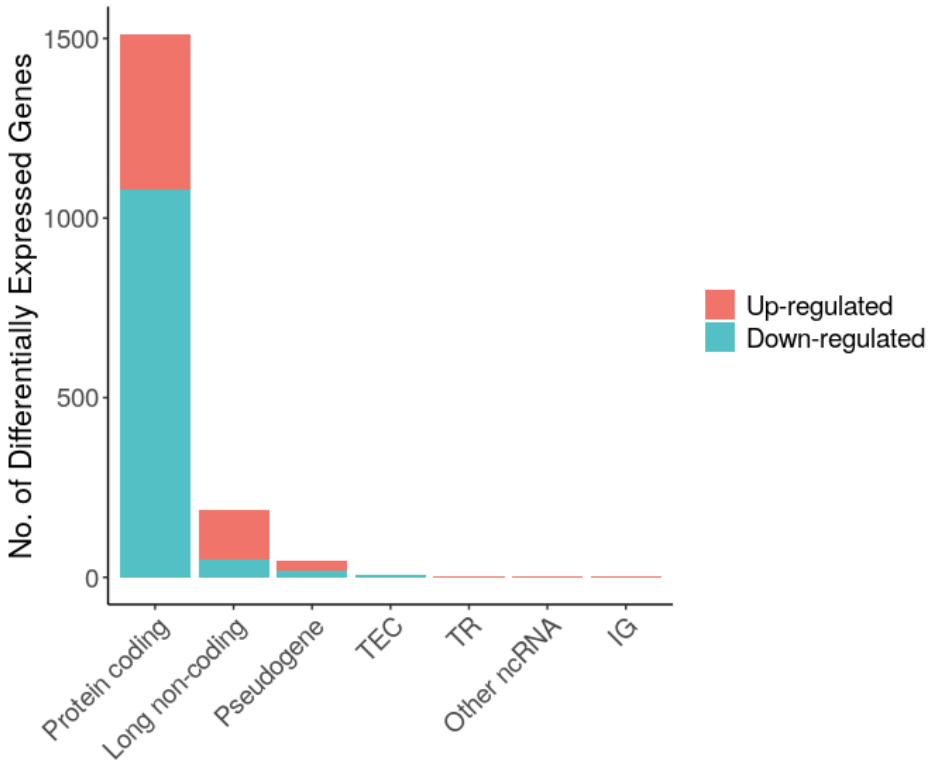
	symbol	group	logFC	AveExpr	t	PValue	FDR	B
ENSG00000143257	NR1I3	protein_coding	-6.916825	7.023129	-17.290861	4.244355e-22	2.419282e-19	40.042879
ENSG00000205707	ETFRF1	protein_coding	-2.492182	9.515997	-16.067528	8.353256e-21	2.380678e-18	37.197506
ENSG00000134532	SOX5	protein_coding	-4.871118	6.228227	-15.035891	1.168746e-19	2.220617e-17	34.498280
ENSG00000141338	ABCA8	protein_coding	-5.653794	7.520581	-14.860685	1.851519e-19	2.638414e-17	34.115811
ENSG00000066583	ISOC1	protein_coding	-2.370131	10.466194	-14.565324	4.053959e-19	4.621513e-17	33.356400
ENSG00000164188	RANBP3L	protein_coding	-5.624376	4.356284	-14.224770	1.013592e-18	9.629125e-17	32.256592
ENSG00000163959	SLC51A	protein_coding	-4.669273	8.154014	-13.870245	2.670535e-18	2.174579e-16	31.489207
ENSG00000181192	DHTKD1	protein_coding	-3.337642	11.438587	-13.757214	3.648663e-18	2.363508e-16	31.179869
ENSG00000175336	APOF	protein_coding	-10.855558	5.693465	-13.749071	3.731855e-18	2.363508e-16	31.121750
ENSG00000186480	INSIG1	protein_coding	-4.914546	11.666362	-13.518547	7.088579e-18	4.040490e-16	30.523653

Showing 1 to 11 of 222 entries

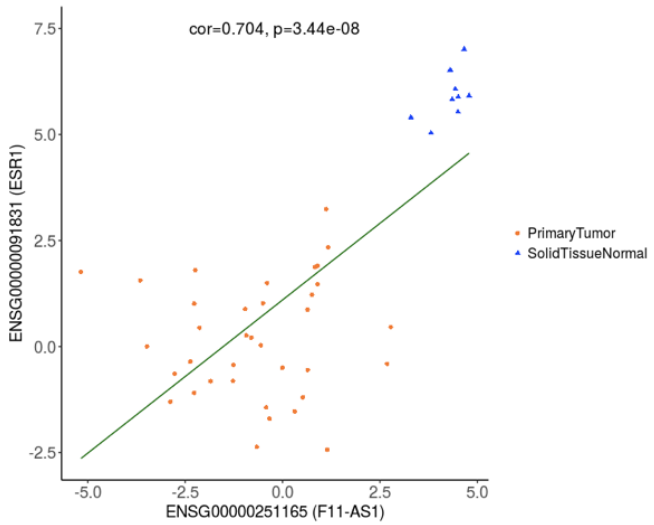
Differentially expressed LNCRNAs

	symbol	group	logFC	AveExpr	t	PValue	FDR	B
ENSG00000234456	MAGI2-AS3	long_non_coding	-2.940855	8.4856698	-10.492654	6.018950e-14	1.270667e-12	21.5507440
ENSG00000204949	FAM83A-AS1	long_non_coding	-6.391414	2.9784616	-8.674627	2.394010e-11	2.966491e-10	15.7410423
ENSG00000239799	ITIH4-AS1	long_non_coding	-4.079870	2.0936658	-8.462870	4.921591e-11	5.610614e-10	15.0288578
ENSG00000234741	GAS5	long_non_coding	2.208649	12.6178085	7.667423	7.633207e-10	6.215611e-09	12.1128965
ENSG00000259319	AF111167.2	long_non_coding	-2.288354	5.5240038	-6.602015	3.174907e-08	1.740093e-07	8.7192311
ENSG00000261437	AC108860.2	long_non_coding	4.506530	5.4697803	6.462983	5.172781e-08	2.755594e-07	8.3123604
ENSG00000255717	SNHG1	long_non_coding	2.035828	10.5197646	6.441780	5.572476e-08	2.887556e-07	8.0384268
ENSG00000229852	AC019205.2	long_non_coding	2.760694	5.4710695	6.048427	2.212462e-07	9.411218e-07	6.9010860
ENSG00000231074	HCG18	long_non_coding	1.401800	9.8376495	5.853688	4.368269e-07	1.753460e-06	6.0321099
ENSG00000260852	FBXL19-AS1	long_non_coding	2.388592	6.8967395	5.533193	1.330449e-06	4.892618e-06	5.1379778

Showing 1 to 11 of 47 entries



Correlation of CeRNAs



Hypergenometric test is performed to test whether a lncRNA and mRNA share many miRNAs significantly. The HyperPValues give the statistical significance of enrichment among the genes.

lncRNAs	Genes	Counts	listTotal	popHits	popTotal	foldEnrichment	hyperPValue
ENSG00000234456	ENSG00000162688	2	2	2	277	138.5	2.6160205096008e-05
ENSG00000234456	ENSG00000107864	2	2	95	277	2.91578947368421	0.116805315753675
ENSG00000234456	ENSG00000158825	2	2	2	277	138.5	2.6160205096008e-05
ENSG00000234456	ENSG00000148943	2	2	38	277	7.28947368421053	0.0183906241824936
ENSG00000234456	ENSG00000106049	2	2	2	277	138.5	2.6160205096008e-05
ENSG00000234456	ENSG00000100664	1	2	18	277	7.69444444444444	0.125961387537278
ENSG00000234456	ENSG00000135111	2	2	24	277	11.5416666666667	0.0072202166064982
ENSG00000234456	ENSG00000198947	2	2	38	277	7.28947368421053	0.0183906241824936
ENSG00000234456	ENSG00000165672	2	2	8	277	34.625	0.000732485742688222
ENSG00000234456	ENSG00000198252	2	2	61	277	4.54098360655738	0.0478731753256946

DISCUSSION AND LIMITATIONS

Long non-coding RNAs can act as sponges (CeRNAs) to regulate the functions of miRNAs and mRNAs in various cells of the body. The prediction and prognosis of biliary cancer largely depends on this feature of the lncRNAs. More research and analysis of the interaction and networking of protein coding and non coding genes hold the plethora of information about future of medicine and pharmacology.

The limitations of my analysis are as follows:

1. Since I preferred to work in R, it was challenging enough to handle the huge data.
2. The code ran only once and refused to get executed then on.

CONCLUSION

The above analysis gives a clear idea that lncRNAs act as scaffolds and decoys in the prognosis of cancer and probably show a path to prognostic and precision medicine for various cancer types.

FUTURE SCOPE

1. To use better computational methods, software packages to do better analysis.
2. To rely more on command line interface to understand high throughput data that is associated with my data.

REFERENCES

1. Zhou, S., He, Y., Yang, S., Hu, J., Zhang, Q., Chen, W., … Tang, J. (2018). The regulatory roles of lncRNAs in the process of breast cancer invasion and metastasis. *Bioscience reports*, *38*(5), BSR20180772. doi:10.1042/BSR20180772

2. The hallmarks of cancer, *A long non-coding RNA point of view* Tony Gutshner & Diedfichs

Pages 703-719 | Published online: 01 Jun 2012

3.Yousefi M., Nosrati R., Salmaninejad A., Dehghani S., Shahryari A. and Saberi A. (2018) Organ-specific metastasis of breast cancer: molecular and cellular mechanisms underlying lung metastasis. Cell. Oncol. (Dordrecht) 41, 123–140, 10.1007/s13402-018-0376-6 [PubMed] [CrossRef] [Google Scholar]

4. Li, R., Qu, H., Wang, S., Wei, J., Zhang, L., Ma, R., Lu, J., Zhu, J., Zhong, W., and Jia, Z. (2018). GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics 34, 2515-2517. https://doi.org/10.1093/bioinformatics/bty124.

5. Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., … Feng, F. Y. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nature communications*, *7*, 12791. doi:10.1038/ncomms12791

7. Kim, M., Yu, Y., Moon, J. H., Koh, I., & Lee, J. H. (2018). Differential Expression Profiling of Long Noncoding RNA and mRNA during Osteoblast Differentiation in Mouse. *International journal of genomics*, *2018*, 7691794. doi:10.1155/2018/7691794

8. Furi'o-Tar'i, Pedro, Sonia Tarazona, Toni Gabald'on, Anton J. Enright, and Ana Conesa. 2016. "spongeScan: A Web for Detecting microRNA Binding Elements in lncRNA Sequences." *Nucleic Acids Research* 44 (Web Server issue):W176–W180. https://doi.org/10.1093/nar/gkw443.