Available online at http://www.ijcrls.com

International Journal of Current Research in Life Sciences
Vol. 07, No. 03, pp.1302-1307, March, 2018

ISSN: 2319-9490

RESEARCH ARTICLE

INTEGRATION OF CANCER ASSOCIATED GENES FROM HUMAN GENOME BY MANUAL CURATION OF BIOLOGICAL DATABASES

[1]Mumtaj, P. and [2,*]Vijaya, P.P.

[1]Post Graduate and Research Department of Biotechnology, Mohamed Sathak College of Arts and Science, Sholinganallur, Chennai – 600 119, Tamilnadu, India
[2]Associate Professor, Department of Nanoscience and Technology, Bharathiar University, Coimbatore, Tamilnadu, India

Received 25th January, 2018; Accepted 24th February, 2018; Published Online 30th March, 2018

ABSTRACT

Cancer has remarkable molecular circuitry connections. Cancer treatment is complex and depends on a number of factors, including genetic, transcriptomic, epigenetic and environmental factors. Advanced Cancer therapeutics research requires molecular data integration and network and pathway analysis. It is essential to know the genes and proteins involved in cancer networks and pathways for the better treatment of cancer. We conducted an extensive manual text mining in our study to compile the genes involved in cancer as study of interest. We used various public biological databases to compile the cancer associated genes in the human genome. Total of 869 cancer associated genes in human genome was extracted. In this study, an extensive public biological databases text mining was performed and extracted the cancer associated genes in the human genome.

Key words: Data mining, Cancer, Genes, Proteins, Human Genome, Public Databases.

Citation: Mumtaj, P. and Vijaya, P.P. 2018. "Integration of cancer associated genes from human genome by manual curation of biological databases" *International Journal of Current Research in Life Sciences*, 7, (03), 1302-1307.

INTRODUCTION

Biological functions are enforced through the interactions among genes, proteins and other intracellular molecules. The cellular networks that mediate the signals from the intracellular and extracellular environments inside cells, are being actively investigated to better understand normal and dysregulated processes. Recent years have significantly advanced our understanding of the genetic and molecular events underlying the metabolic functional phenotype of normal and altered cells. This has been achieved due to the advancement of molecular biology technologies, accumulating data of gene sequences and gene methylation patterns, gene, protein and micro RNA expression measurements, as well as metabolites levels, enabling the characterization of complex cellular processes. Highly complex molecular networks, are known to be dysregulated in a number of diseases, most notably in cancer. Cancer is known as highly complex disease that has caused millions of human deaths. Its study has a long history of well over 100 years. There have been an enormous number of publications on cancer research in literature and multiomics data. This integrated but unstructured data is of great value for cancer diagnostics, treatment, and prevention. Retrieving biological knowledge of genes and proteins involved in carcinogenesis and their relationship is highly important for further research analysis. Extraction of these information rely heavily on expert curation manually or automated. Text mining has contributed to their recognition in the unstructured data (Leaman *et al.,* 2008; Raja *et al.,* 2014). The biomedical data mining task is listed as follows:

Data Retrieval*:* The process of extracting relevant documents from a large collection is called document retrieval or information retrieval (Natarajan *et al.,* 2005). The query-based and document-based retrieval two basic strategies are applied .

Data Prioritization: The data prioritization usually performed to get the most relevant document and it is achieved based on certain parameters (Lin *et al.,* 2007).

Information Extraction: To extract and present the information in a structured format. Concept extraction and relation/event extraction are the two major components of information extraction (Leaman *et al.,* 2008; Raja *et al.,* 2014).

Knowledge Discovery: It is a conceptual process to discover novel and potentially useful biological information from the structured text obtained from information extraction. Knowledge discovery uses techniques from a wide range of disciplines such as artificial intelligence, machine learning, pattern recognition, data mining, and statistics (Lin *et al.,*

Corresponding author: **Vijaya, P.P.,**
Associate Professor, Department of Nanoscience and Technology, Bharathiar University, Coimbatore, Tamilnadu, India.

2007). Both information extraction and knowledge discovery find their application in database curation (Leaman, 2008) and pathway construction (Raja *et al.,* 2014; Natarajan *et al.,* 2015).

***Knowledge Summarization*:** From one or many documents the information is generated for a given topic in knowledge summarization.

Hypothesis Generation: Unknown biomedical facts are predicted from biomedical articles is an important task of text mining These hypotheses are useful in designing experiments or explaining existing experimental results (Zhu *et al.* 2013). Today compiling valuable data are critical in biomedical research by being a firsthand tool for researchers to investigate their hypothesis or research results. In this study we performed manual curation of cancer associated genes in human genome as a data set from various biological databases.

METHODOLOGY

We prepared a list of cancer associated genes which implicated 1% of human genome by manual curation of biomedical literature and various biological databases.

The following biological databases are used in our study:

1. **HGNC:** All approved symbols including protein coding genes, non coding RNA genes and pseudo genes, are stored in the Human Genome Nomenclature database (http://www. genenames.org/).
2. **DAVID:** DAVID bioinformatics resources consist of an integrated biological knowledgebase and analytic tools for systematically extracting large gene/protein lists https://david. ncifcrf.gov/).
3. **Cancer Gene Census at Sanger's Insitute:** The Cancer Gene Census (CGC) is an ongoing effort to catalogue those genes for which mutations have been casually implicated in cancer (http://cancer.sanger.ac.uk/census).
4. **DISEASE:** A DISEASE is a frequently updated web resource that integrates evidence on disease-gene associations from automatic text mining, manually curated literature, cancer mutation data, and genome-wide association studies http://diseases.jensenlab.org).

In these databases manual text data mining was done by using the query term cancer genes. Finally, we integrated these datasets to a compiled list of cancer-associated genes .

RESULTS AND DISCUSSION

Total of 869 cancer associated genes in human genome was extracted in this study and provided in Tables 2, 3, 4 &5.

Table 1. List of biological databases used in the study and number of genes manually curated

Biological Databases Used	No.of Genes Curated
HGNC	137
David	148
Cancer Census	584
Disease	72
Total of Genes Integrated	869

Different text mining implementations for exploring the findings of genome research have been developed in the past decade.

Table 2. 137 Cancer genes extracted from HGNC Database

Gene Symbol			
BCAR3	HEATR6	BLID	MACC1
BRMS1	ICE2	CCAR2	MAGEC2
BRMS1L	MAGED2	DEPDC1B	NDC80 NDC80
GREB1	SNCG	FAM84B	RTL6
PBOV1	SYTL2	DOLPP1	BPHL
AMPH	CAGE1	TMEM173	CHMP2A
BRCA1	CASC1	AGO2	CXCL14
BRCA2 BRCA2	CASC3	CADM1	SLC45A3
BCAR1	CASC4	HIC1	BCAS2 BCAS2
ERGIC3	CASC10	HIC2	BCAS2P1 BCAS2
TFF1	NTPCR	KNL1	CHMP2A
TRERF1	SCAI	LDOC1	PRICKLE4
CREG1	VOPP1	BAGE4	TOMM6
CREG2	SLC45A3	BAGE5	DDX53
BCAS1	PRAC1	BRDT	DSCR8
BCAS4	BCAS2	C1orf74	PAGE5
KIAA0100	BCAS3	C2orf40	SPAG9
AGR3	EPSTI1	C4orf46	XAGE3
ANKRD30A	ST18	ODF3	BAGE3
C8orf4	FAM168A	ODF2	
C14orf93	FATE1	OVCA2	MAGEA1
C18orf8	FGF4	ZNF165	MAGEA2
CRISP2	FMR1NB	XAGE5	MAGEA3
CABYR	FTHL17	XAGE1D	MAGEA4
CALR3	GAGE1	XAGE2	MAGEA6
CCDC33	HID1	XAGE5	MAGEA8
CCDC36	HSPB9	TBC1	MAGEA1
CCDC62	LETMD1	SYCP1	MAGEA1
CCDC110	LIPI	SYCE1	OIP5
CDK2AP2	LSM1	SPINK7	ODF4
CENPW	LUZP4	CNOT9	ELOVL4
CEP55	LYPD6B	CTNNA2	EPPIN
CEP290	LY6K	CTCFL	ERG
CHD1L	LZTS1	DCAF12	FAM46D
DDX43	FAM133A		

Table 3. 148 Cancer genes extracted from DAVID Database

Gene Symbol			
NCAM1	5SLC22A5	EPM2A	SF1
TNFRSF25	1FARP1	SLC17A3	IGHV4-31
NTRK2	ERVH-1	CYP2B7P	CRCP
TNFAIP6	HCFH	HTR4	AUTS2
CYP4A11	2GAD2	OR7E12P	POM121L1P
CCL5	RPS14	4DEFA4	uncharacterized LOC101927057LOC101927057
2NOS2	1CPDE1C	MAGEA9	INHBC
ABCB1	ATP4B	ATF7	PSG4
KISS1	CEACAM3	IHH	4A1EIF4A1
MERTK	TNIK	DEFA5	CORO2A
FGF3	1NEURL1	TP53I11	SHBG
15GDF15	4AGAP4	ALOX15	1AUPK1A
2ERBB2	RAP1GAP	1TPTEP1	2LECT2
BCL2	MPO	HHLA1	8GAS8
p53TP53	FAIM2	BNFIB	25CCL25
D2CCND2	RAB31	DNAH17	58TRIM58
6WNT6	HBD	MYBPC3	RNASE2
1DEFA1	CEACAM8	PRAMEF1	1CHIT1
IGKC	WNT2B	2NRG2	ELANE
1HBA1	ATP8B3	NRG2	SLC17A1
ELN	PRDM2	6SLC6A6	GCTSG
HBB	azurocidin 1AZU1	COL16A1	HAAO
testis expressed 28TEX28	RNASE3	WBP4	B2CPB2
DEFA1	LOC100653049	LGALS3	mannose receptor C type 2MRC2
4AKDM4A	NKTR	PIK3R1	CCL13
MEF2C	1BCAT1	MTF1	WBSCR22
DEAD-box helicase 51DDX51	CYP3A5	PPP1R3D	9DUSP9
TFF3	AGRIN2A	4BRD4	NUP93
putativeGNL1	claudin 7CLDN7	RUNX2	ABCB9
1ASIC1	TNXA	TRAPPC12	2ASIC2
A8ANXA8	NPY	8A2ATP8A2	ADAM23
NCR3	1 alphaREG1A	AFDN	HAL
1AGTR1	KLHL9	HOPX	e40BHLHE40
GIGYF2	SFTPC	TAF6L	PLEKHG3
NAT8B	2TACC2	PDLIM5	ZNF500
1MAPK8IP1	RND2	GNA12	CRADD
TLE1	GCTSG	1NSG1	RAD54L

Table 4. 584 Cancer genes extracted from Cancer Gene Census Database

Gene Symbol			
ABI1	ARNT	BCORL1	CBFA2T3
ABL1	ASPSCR1	BCR	CBFB
ABL2	ASXL1	BIRC3	CBL
ACKR3	ATF1	BLM	CBLB
ACSL3	ATIC	BMPR1A	CBLC
ACSL6	ATM	BRAF	CCDC6
ACVR1	ATP1A1	BRCA1	CCNB1IP1
ACVR2A	ATP2B3	BRCA2	CCND1
AFF1	ATR	BRD3	CCND2
AFF3	ATRX	BRD4	CCND3
AFF4	AXIN1	BRIP1	CCNE1
AKAP9	AXIN2	BTG1	CD274
AKT1	B2M	BTK	CD74
AKT2	BAP1	BUB1B	CD79A
ALDH2	BCL10	C12orf9	CD79B
ALK	BCL11A	C15orf65	CDC73
AMER1	BCL11B	C2orf44	CDH1
APC	BCL2	CACNA1D	CDH11
APOBEC3B	BCL3	CALR	CDK12
AR	BCL5	CAMTA1	CDK4
ARHGAP26	BCL6	CANT1	CDK6
ARHGEF12	BCL7A	CARD11	CDKN1B
ARID1A	BCL9	CARS	CDKN2A
ARID1B	BCL9L	CASC5	CDKN2A(p14)
ARID2	BCOR	CASP8	CDKN2C
MEF2C	1BCAT1	MTF1	WBSCR22
51DDX51	CYP3A5	PPP1R3D	9DUSP9
TFF3	AGRIN2A	4BRD4	NUP93
GNL1	claudin 7CLDN7	RUNX2	ABCB9
aASIC1	TNXA	TRAPPC12	2ASIC2
A8ANXA8	NPY	A2ATP8A2	ADAM23
NCR3	REG1A	AFDN	HAL
1AGTR1	KLHL9	HOPX	e40BHLHE40
GIGYF2	SFTPC	TATA-box binding protein associated factor 6 likeTAF6L	PLEKHG3

…… *Continue*

NAT8B	2TACC2	PDLIM5	ZNF500
1MAPK8IP1	RND2	GNA12	CRADD
TLE1	GCTSG	1NSG1	RAD54L
CDX2	EGFR	FGFR2	HOXA13
CEBPA	EIF3E	FGFR3	HOXA9
CEP89	EIF4A2	FGFR4	HOXC11
CHCHD7	ELF4	FH	HOXC13
CHD4	ELK4	FHIT	HOXD11
CHEK2	ELL	FIP1L1	HOXD13
CHIC2	ELN	FLCN	HRAS
CIC	EML4	FLI1	HSP90AA1
CIITA	EP300	FLT3	HSP90AB1
CLIP1	EPAS1	FLT4	IDH1
CLP1	EPS15	FNBP1	IDH2
CLTC	ERBB2	FOXA1	IGH
CLTCL1	ERBB3	FOXL2	IGK
CNBP	ERBB4	FOXO1	IGL
CNOT3	ERC1	FOXO3	IKBKB
CNTRL	ERCC2	FOXO4	IKZF1
COL1A1	ERCC3	FOXP1	IL2
COL2A1	ERCC4	FSTL3	IL21R
COX6C	ERCC5	FUBP1	IL6ST
CREB1	ERG	FUS	IL7R
CREB3L1	ESR1	GAS7	IRF4
CREB3L2	ETNK1	GATA1	ITK
CREBBP	ETV1	GATA2	JAK1
CRLF2	ETV4	GATA3	JAK2
CRTC1	ETV5	GMPS	JAK3
CRTC3	ETV6	GNA11	JAZF1
CSF3R	EWSR1	GNAQ	JUN
CTCF	EXT1	GNAS	KAT6A
CTNNB1	EXT2	GOLGA5	KAT6B
CUX1	EZH2	GOPC	KCNJ5
CXCR4	EZR	GPC3	KDM5A
CYLD	FAM131B	GPHN	KDM5C
DAXX	FAM46C	GRIN2A	KDM6A
DCTN1	FANCA	H3F3A	KDR
DDB2	FANCC	H3F3B	KDSR
DDIT3	FANCD2	HERPUD1	KEAP1
DDR2	FANCE	HEY1	KIAA1549
DDX10	FANCF	HIF1A	KIAA1598
DDX3X	FANCG	HIP1	KIF5B
DDX5	FAS	HIST1H3B	KIT
DDX6	FAT1	HIST1H4I	KLF4
DEK	FAT4	HLA-A	KLF6
DICER1	FBXO11	HLF	KLK2
DNAJB1	FBXW7	HMGA1	KMT2A
DNM2	FCGR2B	HMGA2	KMT2C
DNMT3A	FCRL4	HMGN2P46	KMT2D
DROSHA	FES	HNF1A	KRAS
DUX4L1	FEV	HNRNPA2B1	KTN1
EBF1	FGFR1	HOOK3	LASP1
ECT2L	FGFR1OP	HOXA11	LCK
LCP1	MAX	MTOR	P2RY8
LEF1	MDM2	MUC1	PAFAH1B2
LHFP	MDM4	MUTYH	PALB2
LIFR	MDS2	MYB	PAX3
LMNA	MECOM	MYC	PAX5
LMO1	MED12	MYCL	PAX7
LMO2	MEN1	MYCN	PAX8
LPP	MET	MYD88	PBRM1
LRIG3	MITF	MYH11	PBX1
LRP1B	MKL1	MYH9	PCM1
LSM14A	MLF1	MYO5A	PCSK7
LYL1	MLH1	MYOD1	PDCD1LG2
LZTR1	MLLT1	NAB2	PDE4DIP
MAF	MLLT10	NACA	PDGFB
MAFB	MLLT11	NBN	PDGFRA
MALAT1	MLLT3	NCKIPSD	PDGFRB
MALT1	MLLT4	NCOA1	PER1
MAML2	MLLT6	NCOA2	PHF6
MAP2K1	MN1	NCOA4	PHOX2B
MAP2K2	MNX1	NCOR1	PICALM
MAP2K4	MPL	NCOR2	PIK3CA
MAP3K1	MSH2	NDRG1	PIK3R1
MAP3K13	MSH6	NF1	PIM1
MAPK1	MSI2	NF2	PLAG1
MAX	MSN	NFATC2	PLCG1
PML	MTCP1	NFE2L2	RMI2

…… Continue

PMS1	PTPN13	NFIB	RNF213
PMS2	PTPRB	NFKB2	RNF217-AS1
POLE	PTPRC	NFKBIE	RNF43
POT1	PTPRK	NIN	ROS1
POU2AF1	PTPRT	NKX2-1	RPL10
POU5F1	PWWP2A	NONO	RPL22
PPARG	QKI	NOTCH1	RPL5
PPFIBP1	RABEP1	NOTCH2	RPN1
PPM1D	RAC1	NPM1	RSPO2
PPP2R1A	RAD21	NR4A3	RSPO3
PPP6C	RAD51B	NRAS	RUNDC2A
PRCC	RAF1	NRG1	RUNX1
PRDM1	RALGDS	NSD1	RUNX1T1
PRDM16	SET	NT5C2	SALL4
PREX2	SETBP1	NTRK1	SBDS
PRF1	SETD2	NTRK3	SDC4
PRKACA	SF3B1	NUMA1	SDHA
PRKAR1A	SFPQ	NUP214	SDHAF2
PRRX1	SFRP4	NUP98	SDHB
PSIP1	SH2B3	NUTM1	SDHC
PTCH1	SH3GL1	NUTM2A	SDHD
PTEN	SLC34A2	NUTM2B	05-Sep
PTK6	SLC45A3	OLIG2	06-Sep
PTPN11	SMAD2	OMD	09-Sep
RANBP17	SMAD3	SRGAP3	STRN
RANBP2	SMAD4	SRSF2	SUFU
RAP1GDS1	SMARCA4	SRSF3	SUZ12
RARA	SMARCB1	SS18	SYK
RB1	SMARCD1	SS18L1	TAF15
RBM10	SMARCE1	SSX1	TAL1
RBM15	SMO	SSX2	TAL2
RECQL4	SND1	SSX4	TBL1XR1
REL	SOCS1	STAG2	TBX3
RET	SOX2	STAT3	TCEA1
RHOA	SPECC1	STAT5B	TCF12
RHOH	SPEN	STAT6	TFEB
TCF3	SPOP	STIL	TFG
TCF7L2	SRC	STK11	TFPT
TCL1A	TNFRSF14	TP63	TFRC
TCL6	TNFRSF17	TPM3	TGFBR2
TERT	TOP1	TPM4	THRAP3
TET1	TP53	TPR	TLX1
TET2	TNFAIP3	TRA	TLX3
TFE3	TMPRSS2	TRAF7	TMEM127
TRD	TRIM27	TRB	TRIM33
TRIM24			

Table 5. 72 Cancer genes extracted from DISEASE Database

Gene Symbols		
TP53	TP53	HRAS
TMPRSS2	ERBB2	EML4
PTEN	EGFR	TP53
MAPK8IP1	AKT1	ROS1
AKT1	KLK3	SMUG1
C6orf15	HRAS	AKT1
ENSG00000214921	MYC	ERCC1
GAST	ESR1	NUP62
MAPK8IP1	VEGFA	CALB2
TAAR6	CCND1	ATP10B
INSM1	ARHGEF5	RAB11B
MAPK8IP2	STYK1	CAGE1
CDKN2A	BCL3	TRG-GCC2
CEACAM5	FSTL3	CEACAM5
TP53	FGF4	LEPROTL1
TMCC1	SIRT4	RAB1A
PTPRN	XBP1	DKKL1
KRT7	PNN	SCRT1
TAF12	AURKA	SLC14A2
SMARCB1	SALL4	BRCA1
NFYC	KLK3	BRCA2
RAD54L	AR	FOLH1
EGFR	PTEN	KRAS
ALK	TMPRSS2	AKT1

More sophisticated approach integrate gene expressions from microarray experiments, biomedical data extracted by text mining, and gene interaction data to predict gene-based drug indications A similar approach (Faro *et al.,* 2012) attempt to support manual curation of links between biological databases such as Gene Expression Omnibus (GEO) and PubMed database. Text mining data with microarray data for discovering disease-gene association by using unsupervised clustering. The gene-drug interaction information extracted by text mining is used to predict the drug-drug interaction (Percha *et al.,* 2012). Above all, the researchers have attempted to use text mining for annotating genome function with gene ontology (Daley *et al.,* 2015). Thus, text mining and genomics together reveals much biomedical information that was previously unknown.

Conclusion

The cancer associated genes in human genomes as a data set prepared in this study from various bioinformatics resources would be useful to understand these genes, their coding protein products, interactions and the ways in which they function is very much important in basic cancer genome research.

Acknowledgements: We acknowledge the Management, Dean, Academic Director, Principal and Faculties of Department of Biotechnoloy of Mohamed Sathak College of Arts and Science, Chennai – 600119

REFERENCES

Daley J. M., Niu H., Miller A. S., Sung P. 2015. Biochemical mechanism of DSB end resection and its regulation.*DNA Repair*, 32:66–74.

Faro A., Giordano D., Spampinato C. 2012. Combining literature text mining with microarray data: advances for system biology modeling. *Briefings in Bioinformatics*, 13(1): 61–82.

Leaman R, Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput, 652-6*

Leaman R., Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. Proceedings of the 13th Pacific Symposium on Biocomputing (PSB '08); Kohala Coast, Hawaii, USA. pp. 652–663.

Lin Y, Li W, Chen K, Liu Y. 2007. A document clustering and ranking system for exploring MEDLINE citations. *J Am Med Inform Assoc., 14(5):651-61.*

Natarajan J., Berrar D., Hack C. J., Dubitzky W. 2005. Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology*, 25(1-2):31–52.

Percha B., Garten Y., Altman R. B. 2012. Discovery and explanation of drug-drug interactions via text mining. Proceedings of the 17th Pacific Symposium on Biocomputing (PSB '12); Kohala Coast, Hawaii, USA. pp. 410–421.

Raja K, Subramani S, Natarajan J. 2014. A hybrid named entity tagger for tagging human proteins/genes . *Int J Data Min Bioinform, 10(3):315-28.*

Raja K., Subramani S., Natarajan J. 2014. A hybrid named entity tagger for tagging human proteins/genes. *International Journal of Data Mining and Bioinformatics*, 10(3):315–328.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B. 2013. Biomedical text mining and its applications in cancer research . *J Biomed Inform,* 46(2):200-11.

http://www.genenames.org/

https://david.ncifcrf.gov/

http://cancer.sanger.ac.uk/census/

http://diseases.jensenlab.org

****