

Gene Expression Data Analysis for Stomach Cancer

Anita Vibhuti¹, U.M. Muddapur², V.G. Shanmuga Priya³, Preeti Honagudi⁴

P.G. Student, Department of Biotechnology, KLE Dr M. S. Sheshgiri College of Engineering and Technology,
Belgaum, Karnataka, India^{1,4}

Professor, Department of Biotechnology, KLE Dr M. S. Sheshgiri College of Engineering and Technology, Belgaum,
Karnataka, India^{2,3}

ABSTRACT: Gene expression indicates the present state of the cell. Samples associated with the normal and cancer stomach tissues in *Homo sapiens* are collected by providing different stomach tissue query terms from GEO and ArrayExpress database. Manual curation is carried using Standard Operating Procedure (SOP) for both normal and cancerous condition. These samples are submitted to the novel algorithm developed at IBAB, Bangalore, to validate samples and identify stomach cancer specific genes. The study retrieved 1336 samples for stomach cancer condition. From 1336 samples it is known that genes expressed in stomach cancer were 12309; genes not expressed in stomach cancer were 7531. The study also retrieved 434 samples for normal stomach condition, by analyzing it is known that genes expressed in Normal stomach were 10477 and genes not expressed in Normal stomach were 9327. By comparison, 2080 genes were identified which were expressed in stomach cancer and not detected in Normal stomach tissue.

KEYWORDS: Gene expression, Meta analysis, stomach cancer, biocuration (manual curation)

I. INTRODUCTION

Meta analysis is a method that focuses on contrasting and combining results from different studies, in the hope of identifying patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. It can help to investigate the relationship between study features and study outcomes. Stomach cancer is the fifth most common cancer in the world, with 952,000 new cases diagnosed in 2012. The bacterium *Helicobacter pylori* [1] are an important cause of stomach cancers.

Paper is organized as follows. Section II describes materials and methods, Section III displays the obtained results. Finally, Section IV presents conclusion.

II. MATERIALS AND METHODS

1. Identification of tissue of interest:

Stomach tissue of species *Homo sapiens* is selected for the study. The study is carried out for cancerous and normal condition.

2. Identification of data sets from repositories:

The results of microarray experiments are deposited in public repositories. GEO (Gene Expression Omnibus) and ArrayExpress are two such public repositories. The data sets are collected from these databases using different specific query terms as given below:

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

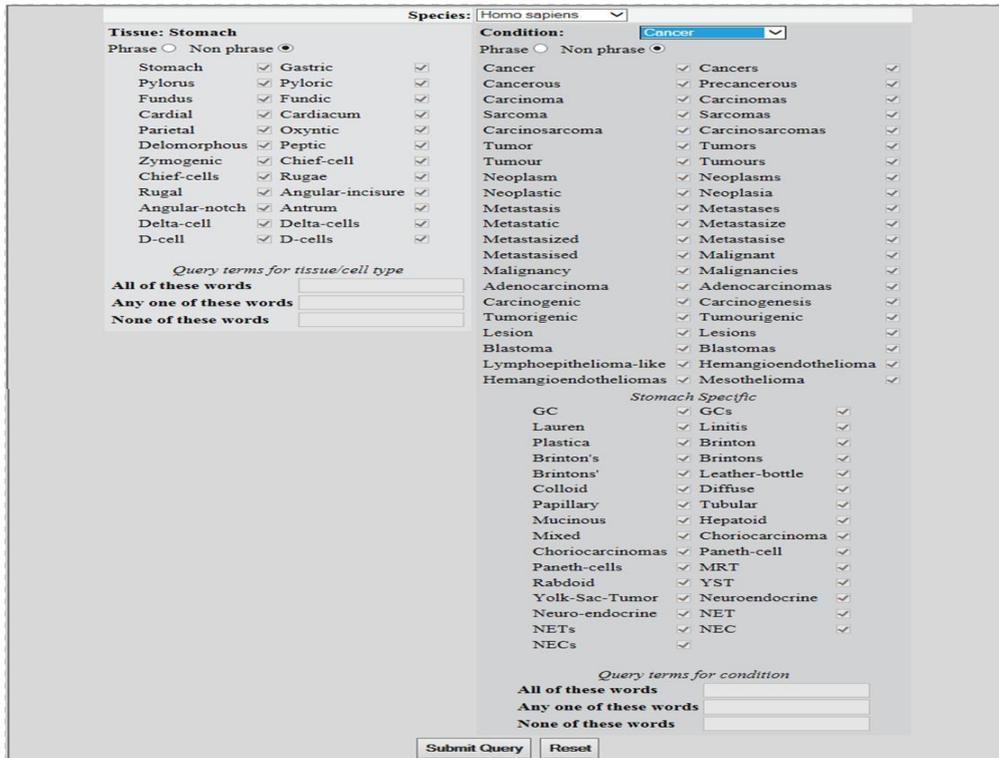


Fig. 1: Stomach cancer tissue synonyms and stomach related terms used to query public repositories (GEO and ArrayExpress)

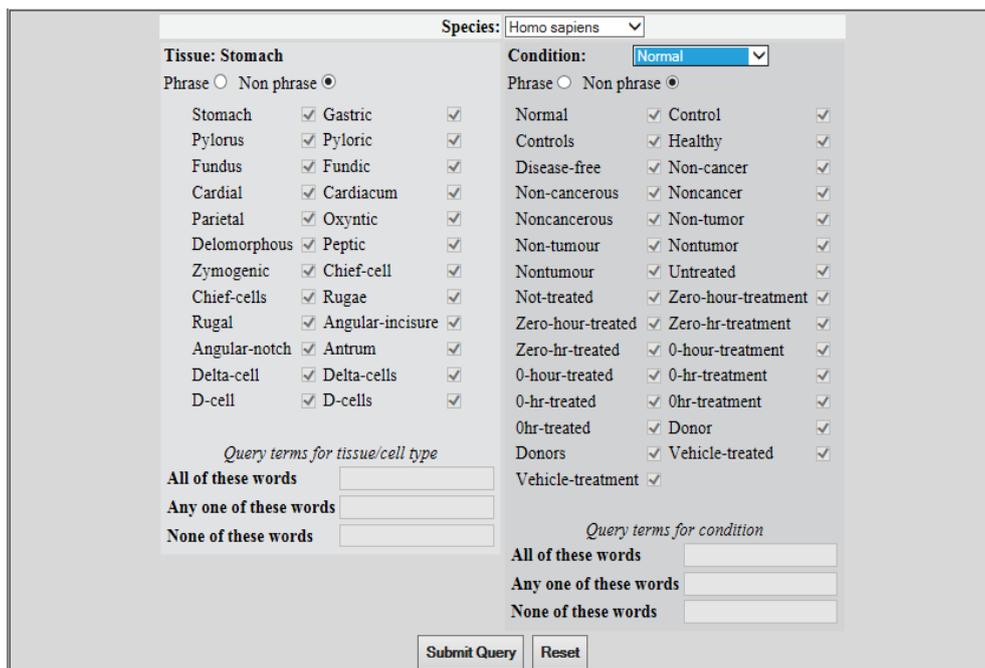


Fig. 2: Normal stomach tissue synonyms and stomach related terms used to query public repositories (GEO and ArrayExpress)

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

3. Manual curation:

Data validation demands precision. Validation of samples requires a set of rules and regulations to be framed and followed strictly, to avoid unhealthy data which might bias the results. For the current study SOP (Standard Operating Procedure) was provided, carefully framed by a team of researchers at IBAB, Bangalore to validate the samples. The SOP as shown in the table 1 describes the different conditions for selection of valid datasets for stomach tissue and also categorizing samples into different categories for normal and cancer condition.

Sl. No	Condition	Description
1	Normal	Tissue/cells/commercial-sample/cell line/cultured cells, from an adult subject with no evidence of abnormality. It should NOT BE: tissue/cells/commercial-sample/cell line/cultured cells, with no age specification; any treated tissue/cells/commercial-sample/cell line/cultured cells [e.g., hormone, chemical, radiation, knockout, surgical (not the surgery done for collection of sample itself), vehicle, placebo, alterations in diet/ habits/ environment/ lifestyle/ experiment]; normal cells/tissue adjacent to tumor; pooled sample/RNA with multiple tissues including the tissue of interest (e.g., “a pool of 10 tissues including lung” should not be included under lung or any other single tissue).
2	Maybe adult	Tissue/cells/commercial-sample/ cell line/cultured cells, from a subject with no age specification, and no evidence of abnormality. It should NOT BE: any treated tissue/cells/commercial-sample/cell line/cultured cells [e.g., hormone, chemical, radiation, knockout, surgical (not the surgery done for collection of sample itself), vehicle, placebo, alterations in diet/ habits/ environment/ lifestyle/ experiment], normal cells/tissue adjacent to tumor; tissue with known developmental stages; pooled sample/RNA with multiple tissues including the tissue of interest (e.g., “a pool of 10 tissues including lung” should not be included under lung or any other single tissue).
3	Maybe normal	Tissue/cells/commercial-sample/cells/cell line/cultured cells, from an adult subject with an evidence of suspicion of variation to normal state [vehicle/placebo/surgical (not the surgery done for collection of sample itself) treatment, normal tissue adjacent to tumor tissue/cells, subject with other tissue disorders (e.g., normal lung tissue from subject with prostate cancer)]. It should NOT BE: tissue/cells/commercial-sample/cell line/cultured cells, with no age specification; any treated tissue/commercial-sample/cell line/cultured cells [e.g., hormone, chemical, radiation, knockout, alterations in diet/ habits/ environment/ lifestyle/ experiment]; pooled sample/RNA with multiple tissues including the tissue of interest (e.g., “a pool of 10 tissues including lung” should not be included under lung or any other single tissue).
4	Cancer	Cancer tissue/cells/commercial-sample/cell line/ cultured cells. It should NOT BE: benign tumor; tissue/cells/commercial-sample/cell line/ cultured cells with no evidence of malignancy (samples with no indication of either benign or malignant status); any treated tissue [e.g., hormone, chemical, radiation, surgical (not the surgery done for collection of sample itself), vehicle, placebo, alterations in diet/ habitsnote3/ environment/ lifestyle/ experiment]; metastatic tissue/cells from a different origin, other than tissue of interest (e.g., metastasized tissue/cells originating from breast cancer/prostate cancer); cancer tissue/cells metastasized to any other tissue (e.g., metastasized lung cancer tissue/cells taken from prostate); pooled sample/RNA with multiple tissues including the tissue of interest (e.g., “a pool of 10 cancer/normal/cancer & normal tissues including lung cancer” should not be included under lung cancer or any other cancer or normal tissue).

Table 1: Conditions in SOP (Standard operating procedure) used for validating the datasets

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

4. Identification of crucial genes:

A tool with novel algorithm [2] at IBAB, accepts the tissue name, condition for which the datasets are to be Meta analysed, and the species for which the proposed work should be carried out. The crucial genes associated with the cancer and normal condition of the stomach tissue are derived using this algorithm. Based on the scoring method, the algorithm gives reliability scores for all the genes transcribed, and not transcribed during the study. The algorithm also gives the number of samples, number of studies, EST count, from which the score has been derived. The algorithm gives two sets of genes, first a list of transcribed genes, with reliability score and second list of dormant genes, which are basically not expressed in the tissue along with the condition (Normal and cancerous).

III. RESULTS

1. Identification of data sets from public repositories (GEO and ArrayExpress):

The query set prepared from stomach cancer synonyms and related terms using relational expression, was used to query GEO and ArrayExpress. This yielded 1115 samples (hits), from 36 studies in GEO and 221 samples (hits) from 10 studies in ArrayExpress, with Affymetrix platform. The same procedure was followed for normal stomach tissue. This yielded 222 samples (hits) from 21 studies in GEO and 212 samples (hits) from 6 studies in ArrayExpress, with Affymetrix platform. Fig 3 displays the result of data sets obtained for the query terms. Each hit contains samples, sample description which is further used for manual curation.

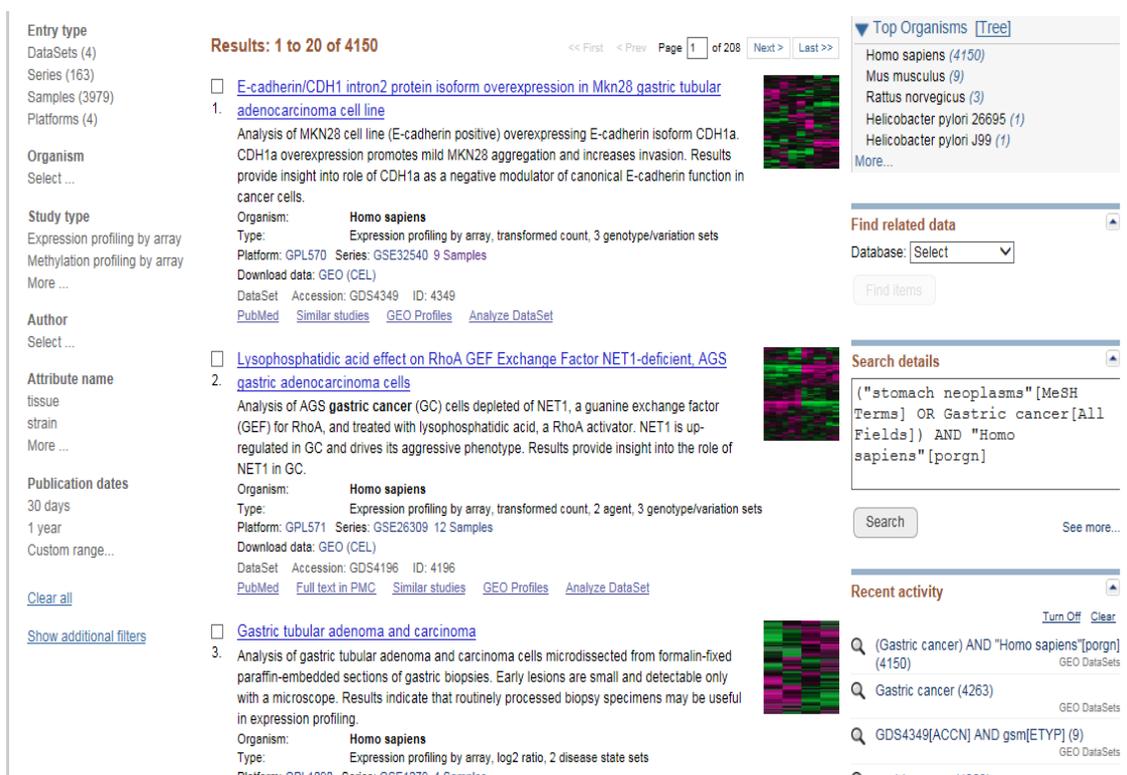


Fig 3: List of data sets obtained for stomach cancer

2. Manual curation:

The validation of microarray data is carried out manually following the SOP as given in the table 1. Fig 4 shows the framework of manual curation. For each sample the screening status is mentioned as either relevant (For example, if the curating condition is cancer and the sample fulfils the condition as per the SOP), completely relevant (For example, if

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

3. Identification of crucial genes:

Using the tool with novel algorithm, crucial genes associated with cancer and normal condition of the stomach tissue are derived. Based on the scoring system, reliability score is given for all the genes transcribed or not transcribed during the study. Following are the genes which are transcribed with high reliability score in cancer condition.

1	Gene list			
2	FAM149B1	NUDT6	TFAP2C	CTLA4
3	MLN	TRIM49	FCN1	SMO
4	KIFC3	KBTBD11	LEPREL4	TMCC2
5	TUBAL3	C21orf2	IL32	GMIP
6	GINS2	SFT2D2	KRI1	CD82
7	POU5F1B	SLC25A10	PAQR4	EHD3
8	TF	KCNAB2	GUCA2B	ABCD1
9	CABYR	ACSM5	IBA57	SPN
10	TACR1	CLDN15	ADRA1D	SERGEF
11	PLD2	KLK5	EFNB1	TRO
12	FPR2	TIAM2	CNGA1	TAF1C
13	ANXA2P3	ZSCAN2	ENTPD3	
14	SPAG6	FABP7	KLK7	
15	CYP27B1	MUC2	CHI3L2	
16	MAPK12	MED25	PTCD1	
17	INA	SLC24A2	NPY1R	
18	MYLK3	FLVCR2	GLB1L2	
19	HAND2-AS1	KIR2DL4	TIMM50	
20	CDT1	ZNF787	PARP11	
21	FAM110B	RIMS1	CDH17	
22	E2F5	TBX2	PTHLH	
23	MMP3	XCL1	GREB1L	
24	APOA4	BIRC7	UBE2NL	
25	GRB7	MYO7B	FGFBP1	
26	PTGDR	GRIK5	SLC4A5	
27	PF4V1	PPIP5K1	IHH	
28	DOT1L	TAT	FAAH	
29	ARID3A	MYL10	TMPRSS6	
30	ESM1	NOX4	VPREB3	

Fig 6: Genes involved in stomach tissue (cancer condition)

IV. CONCLUSION

Gene expression data from public repositories (GEO and ArrayExpress) was compiled to apply a meta-analysis algorithm to compare expression of the genes across studies, identify differentially expressed genes across normal stomach tissue and stomach cancer. Genes which show differential expression with high reliability score in cancer condition were identified. FAM149B1, MLN, KIFC3 etc are the genes which show high variation in transcription in cancer condition. Further analysis of the set of differentially expressed genes in cancer condition can lead to identification of drug targets and biomarkers for stomach cancer condition.

V. ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Dr. Basavaraj G Katageri, Dr. S C Mali, whose kind consent and guidance helped me to successfully complete my research work.

REFERENCES

1. Zuraihan Zakaria "The role of interleukin-10 in Helicobacter pylori infection" July 2010
2. Kshitish K Acharya, Darshan S Chandrashekar, Neelima Chitturi, Hardik Shah, Varun Malhotra, Sreelakshmi KS, Deepti H, Akhilesh Bajpai, Sravanthi Davuluri, Pranami Bora, Leena Rao "A novel tissue-specific meta-analysis approach for gene expression predictions, initiated with a mammalian gene expression testis database."
3. Magic Z, Radulovic S, Brankovic-Magic M "cDNA microarrays: identification of gene signatures and their application in clinical practice". *J BUON*. 12 Suppl 1: S39-44, 2007
4. Fangxin Hong, and Rainer Breitling."A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments". Vol. 24 no. 3 2008
5. "Statistics and outlook for stomach cancer". Cancer Research UK. Retrieved 19 February 2014.
6. Jakszyn P, González CA, "Nitrosamine and related food intake and gastric and oesophageal cancer risk: A systematic review of the epidemiological evidence". *World J Gastroenterol* 12 (27): 4296-4303,2006

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2014

7. "Detailed Guide: Stomach Cancer Treatment Choices by Type and Stage of Stomach Cancer". American Cancer Society.
8. Ron Edgar, Michael Domrchev and Alex E Lash," Gene Expression Omnibus: NCBI gene expression and hybridization array data repository."
9. Nucleic acid research, 2002, vol. 30, No-1
10. Jonathan D Pollock, "Gene expression profiling: methodological challenges, results, and prospects for addiction research". Chemistry and Physics of Lipids, VOL. 121, pages 241-256, 31 December 2002.

BIOGRAPHY



Anita Vibhuti from Belgaum, Karnataka, is pursuing M.Tech in Bioinformatics from KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum, Karnataka, India. She completed her B.E (Information Science) in 2012 from KLE College of Engineering and Technology, Chikodi, Karnataka, India.



Dr.U.M.Muddapur from Belgaum, Karnataka, has done B.Sc, M.Sc(1999) and PhD(2004) –all in Biochemistry, from Karnataka University, Dharwad, Karnataka. Is Presently working as Professor in Dept of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum, Karnataka, India. Is a life member of ISTE and SBC societies.



V.G.Shanmuga Priya from Vellore, Tamil Nadu, India has done B.Sc (zoology) at Auxilium college, Vellore under Madras University and secured University First Rank (1995) and M.Sc(Bioinformatics) at Auxilium College under Thiruvalluvar University and secured University Third Rank(2005) and M.Phil(Bioinformatics) at Bharathidasan University, Tamil Nadu(2007). Is Presently working as Assistant Professor in Dept of Biotechnology, KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum, Karnataka, India.



Preeti Honagudi from Gadag, Karnataka, is pursuing M.Tech in Bioinformatics from KLE Dr.M.S.Sheshgiri College of Engineering and Technology, Belgaum, Karnataka, India. She completed her B.E (Information Science) in 2009 from Smt Kamala and Sri Venkappa M Agadi College of Engineering and Technology, Laxmeshwar, Karnataka, India. She worked as Lecturer at R.L.JALAPPA Institute of Technology, Doddaballapur, Karnataka, India (2010-2012)