A transcriptomic analysis on the differentially expressed genes in oral squamous cell carcinoma

Agnik Haldar
Central University of South Bihar
Ajay Kumar Singh (ajaysingh@cusb.ac.in)
Central University of South Bihar

Research Article

Keywords: Oral Cancer, Cancer Genetics, NGS Data analysis, Bioinformatics

Posted Date: August 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1941558/v1

Abstract

Oral cancer has plagued the majority of the world as one of the most prevalent cancers. The main reason often highlighted is the high usage of tobacco, which has been reported to be the main cause of oral cancer, and the lack of proper health and sexual h+ygiene. This often leads to HPV infection which has been seen as one of the leading causes of oral cancer. Numerous reports have identified the dysregulation of genes as one of the major causes at play in the mechanisms of cancer. A detailed investigation of the dysregulated gene expressions and the pathways shed some light on the mechanistic properties behind cancer, which can potentially lead to a viable approach for biomarker identification and further research.

Introduction

Dysregulation of genes plays a significant role in cancer. Cancers caused as a result of genetic alterations are often the genesis of aberrant gene expression. Oral cancer, which is a subtype of head and neck cancer, consists of 90% of the cases are oral squamous cell carcinoma. A study by Alexandra Iulia Irimie et al (2017), provides a perspective of ncRNA and its derivatives. Another paper by Gibb et al documents the first evaluation of the lncRNA expression profile for oral mucosa.

Despite the glaring evidence of dysregulated genes playing a critical role in the biological processes of human diseases, very few efforts have been made to identify their association with the disease and assess their prognostic values. In 2014, Steven B Cogill and Liangjiang Wang published an article highlighting gene co-expression relational analysis for the identification and annotation of long noncoding RNAs (lncRNAs) and thus verifying their relation to the cancer disease. To achieve their goal, they used the weighted gene co-expression network analysis (WGCNA) method, which yielded hub lncRNA genes and enriched functional annotation terms within the modules. Recently another paper by Shervin Alaei et al (2019), used similar network construction and module detection with the help of WGCNA to identify novel key regulators in esophageal squamous cell carcinoma. They performed gene ontology and pathway enrichment analysis. This proved to be beneficial for estimating the biological processes or pathways that the lncRNAs co-expressed.

In a review article by Claire Jean Quartier et al, in-silico methods involved in cancer research have been highlighted. The article goes on to explain the importance of the TCGA database, and methods pertaining to computational validation, classification, and prediction using mathematical and statistical analysis.

Cancer recurrence is a major problem affecting patients who have been affected by the disease. In oral cancer, the case is no different. Surgery has been the preferred treatment for both cases. Even with the advancements in chemotherapy and radiotherapy, the treatment remains poor due to the local invasion and metastasis, which leads to recurrence. With a rate of 30% survival rate of patients with recurring oral cancer, the need to identify factors that may be able to identify the factors responsible for recurrence

becomes crucial. This study aims at identifying biomarkers that would help improve the prognosis prediction in specific types of oral cancer (M. Zhou et al, 2018; B. Wang et al, 2013).

In this project, we have focused our attention on the oral cancer subtype of gingivobuccal squamous cell carcinoma (GBSCC), and finding the genes being expressed in the subtype portraying carcinogenic behavior. The reason for choosing this subtype alludes to the fact that GBSCC is often overlooked and very little research has been done on it, even though due to tobacco abuse it is one of the most prevalent subtypes of oral cancer, and is seen to affect humans in a much more severe manner. Therefore to gain a better insight, the need to investigate the genes expressing carcinogenicity becomes a crucial step forward in understanding the mechanism of the disease.

RNA-Seq is used to analyze the continuously changing cellular transcriptome. Specifically, RNA-Seq facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations/SNPs and changes in gene expression over time, or differences in gene expression in different groups or treatments. Here we used RNA sequencing to find the expressed genes in oral cancer data Altschulsets which we will be using to analyze the pathways of those expressed genes.

Materials And Methods

Data Collection

The requisite files, namely the FASTQ files and the GTF files, for the RNA seq data analysis were downloaded from NCBI's Gene Expression Omnibus (GEO).

The sequencing data deposited in the NCBI GEO under the accession number GSE101547, consisting of gingivobuccal cancer tissues of 24 patients (Singh et al., 2017), were downloaded and stored in a separate folder. The respective files were extracted from their zip files and the zip files were kept as backup.

Data Pre-Processing

The FASTQ files were processed using a standard transcriptomic bioinformatics analysis pipeline, to obtain dysregulated genes from the data. The dysregulated genes provided us with the opportunity to further understand the relational nature of the genes with the disease. The raw sequence reads were initially checked for quality using FASTQC.

Alignment

The reads were then aligned with the help of a splice-aware aligner HISAT2 (Kim et al., 2015) with the hg38 reference genome. We were able to achieve a > 90% alignment rate for all the samples. Of the 53 million paired-end reads per sample, approximately, 50 million reads were aligned successfully to the hg38 reference genome. Low and inconsistent reads were subsequently removed.

Count data generation

For the generation of count data, we employed the use of the HTSeq-count tool (Anders et al., 2015) which gave us the count data for the overlapping exons of each gene. To sort out the dysregulated genes we used a GENCODE v37 GTF annotation file, which helped us sort out all the genes present in the reads and obtain a count data. Since the data provided with the data set lacked any information about the strandness of the reads, neither was it considered in the paper by R.Singh et al, we also determined the strandness of the raw reads, with the help of the Salmon quantification tool (Patro et al., 2017), as it is an important aspect when looking for count data.

Data Analysis

We were able to determine that the pair-ended raw reads were unstranded with the help of Salmon. Finally, for differential gene expression data analysis, we used the DESeq2 tool (Love et al., 2014) and we were able to report that the gingivobuccal cancer datasets yielded 32869 differentially expressed genes. Wald Test was employed for statistical analysis which is a way to find out if the explanatory variable in a model is significant. “Significant” means that they add something to the model; variables that add nothing can be deleted without affecting the model in any meaningful way. Excluding the outliers, low counts, and keeping an adjusted p-value threshold of < 0.001, we sorted out the dysregulated genes based on our threshold value of < 0.0001 and obtained 1351 significant dysregulated genes.

Dysregulation is explained as the involvement of the genes in the disruption of normal pathways. The dysregulated genes provide us with an incentive to further explore the causality and their involvement in the disease. A MA-plot for the Tumor vs Normal dysregulated genes were generated using the R-Studio as shown in Fig. 1 below.

Gene annotation

The Ensembl IDs obtained after the analysis were converted to their respective official gene names with the help of ENSEMBL Gene ID to Gene Symbol Converter.

BLAST

For the purpose of comparison and also completing the list of genes devoid of gene names and identity, proper annotations were required which were done and completed with the help of BLAST (Altschul et al., 1990). We also checked for the virulence factors by analyzing the e values of the protein genes. The positive and negative e values were checked and thus computed as a ratio of the positive and negative samples e value.

Panther database data mining and Pathway analysis

The genes which we obtained after the RNASeq data analysis were sorted based on the gene names. The common and exclusive protein genes were separated and made into two datasets and were analyzed in

the Panther database sequentially. Keeping the default search parameters for all the datasets, the genes were analyzed concerning their specific organism. The search results gave us the list of all the genes which were documented in the Panther database. The genes which gave no hits to the pathways were deleted. From the list we obtained, biological conditions and processes of the respective genes were noted and made into another dataset. A figure denoting all the pathways are shown in Fig. 2 below.

PANTHER database pathway analysis (Fig. 2) shows the prevalence of Nicotinic Acetylcholine Receptor Signaling which is responsible for tumor growth and metastasis. Nicotine is a major component found in tobacco, accompanied by other nitrosamines. These in a combined effort act to regulate the nAChRs on nonneuronal cells. As a result of which metastasis, tumor growth, and chemoresistance come into play through the regulation of various pathway functionality (Singh S et al., 2011).

To analyze the respective pathways, we decided on fixing criteria for the genes by which we were going to sort out the virulent genes showing pathogenicity. The criteria we fixed for sorting the datasets were genes that showed or gave a response to stress and stimulus.

miRNA enrichment analysis

From our data, we were able to obtain 60651 genes which we analyzed to be dysregulated. Further filtering the data we found out that 1351 genes were significantly dysregulated based on their p-values. Those sets of genes were selected and passed through the various databases like TargetScan (Grimson A et al., 2007)and miRWalk (Sticht C et al., 2018). We were able to deduce the miRNA targets from the genes. Those targets were then analyzed and associated with head and neck cancer data genes. We were able to find out 5 miRNAs significantly associated with the genes obtained from our dataset.

Results

From a data set of 24 samples of 12 normal and 12 tumor replicates, 60651 genes were analyzed. Excluding the outliers and low counts, 32869 genes were considered for examination, which yielded 1351 dysregulated genes based on the adjusted p values. A histogram of the p-values was generated as shown in Fig 3 below.

Data sorting according to exclusivity

The genes obtained after performing RNA sequencing for the cancer data sets were analyzed using the Wald test to give us the significant genes expressed in the bacteria. The criteria for which were based on the in-vivo and control samples. After performing the Wald test, we received 1369 significant genes. These protein genes were then cross-checked for authenticity in BLAST and PANTHER databases. The genes were separately analyzed using the Panther database and Wikigenes Pathway database for pathway analysis. Upon completing our analysis we obtained 575 genes for head and neck cancer which were documented for having significant pathways. Out of which 45 genes showed response to stress as we decided to follow up on genes showing stress conditions in the gene expression as they were

susceptible to show carcinogenicity among which there were genes like *BCL2L12, TYRP1, PAX9, CRNN, KRT4.*

miRNA enrichment data analysis

Table 1: miRNA-Gene correlation table found in oral squamous cell carcinoma.

hsa-mir-99a-5p	ADCY9, CAPNS1, COL4A1, GNAL, IFIT3, PLSCR1, PPM1A, SALL2, CCDC6, CUL3, CTDSPL, TPPP3, RAVER2, ZBTB4, SUDS3
hsa-mir-100-5p	ALDH9A1,CAPNS1,COL4A1,DNMT1,ACSL3,GRB2,FOXN2,IFIT3,KPNA2,ABLIM1,SMAD7,MMP13,NFIA,PLK1,PSMA5,RAP1B,RRM2,CTDSPL,DEAF1,CBX3,DDAH1,SGTB,DCAF6,MARC1,UBN2,DCBLD1,FAM221A
hsa-mir-7-5p	ADCY9,ALDH3A2,BAX,BCAT1,CCNE1,CDC25B,CKS2,DTYMK,EIF4EBP2,FLNA,GLS,GRB2,GRIN2D,HOXB3,IL12RB2,IRS1,LAMC2,NFYA,ROR1,OAS2,PAK1,PIK3CD,PTK7,SLIT3,TCOF1,TOP2A,XRCC2,LUZP1,GAN,AKAP1,RDH16,PLPP3,TRPA1,XPR1,SECISBP2L,GJC1,MFSD10,RAB32,CHP1,MGLL,MYEOV,VPS4A,UHRF1,EHD3,IL21R,ZDHHC3,EIF3L,UBE2D4,VPS13D,GOLPH3L,PHF10,RNF114,SERTAD4,EIF5A2,POLE4,KCNK13,KIAA1143,FNDC4,SLC25A23,FYCO1,RHBDF2,BCL2L12,STK40,RIOX2,PCED1B,SNX29,ADAMTS17,LDHD,IFNE,SIGLEC14
hsa-mir-138-5p	FOXC1,MMP3,MYBL2,PPARG,RELN,SNAI2,TERT,PPM1D,EED,HIST1H2BJ,PPP1R13B,MTHFD1L,PLEK2,TP53INP2,RMND5A,HIST1H2BK,CYTOR,PPM1L,CASTOR2
hsa-mir-143-3p	ABAT,BRAF,CASP5,CDC25B,COL1A1,COL5A1,COL5A2,SP110,LIMK1,MMP9,MMP13,MMP14,NFIC,OAS3,SCN2B,FSCN1,TERT,TLR2,HIST1H2BG,SKAP2,SECISBP2L,ACOT9,ZBTB44,CMPK1,FGD6,STOX2,AGAP1,MCOLN2,MACC1,CLEC17A

The 5 miRNAs obtained after target identification and analysis led us to a gene-miRNA correlation where we were able to associate the genes involved in the head and neck cancer tissues. The 5 miRNAs and their respective correlations are listed in the table above (Table 1).

Discussion

The dysregulated genes which were obtained after comparing the normal data set with the tumor data set yielded data that can be used to further identify and explore the relational nature of the genes to the disease. The dysregulated genes serve as an indication of the disruption of the normal genomic process. This disruption can be traced back to the source of the disease at hand and thus can be further analyzed and used as a biomarker or subsequent target for drug identification. For identification and analysis of dysregulated genes, comparison between the expression in normal and affected tissues will confirm the involvement of those genes in the disease or the lack thereof. Another point of interest in our study was the identification of Nicotinic Acetylcholine Receptor Signaling which is responsible for tumor growth and metastasis via the help of the PANTHER database. Thus, the dysregulated genes along with their

subsequent cell signaling pathways will help us understand their relevance in tumor growth and maintenance.

Conclusion

Differentially expressed genes that negatively regulate gene expression, have been associated with cell invasiveness and cell dissemination, tumor recurrence, and metastasis. Thus a comparison between the expression of the normal and affected tissues will confirm the involvement of dysregulated genes in the mechanism of disease or the lack thereof. Increasing evidence points towards the need explore the possibilities of genome-scale expression of differentially expressed genes in cancer. It would also be beneficial to gain knowledge about their potential biological functions as information is severely lacking in these sectors.

The recurrence of oral cancers is one of the most important aspects of the disease. Identifying factors that affect the recurrence of these cancers to reduce postoperative recurrence is an emerging issue in the clinic. Since these genes have been linked with the cause of recurrence, a detailed analysis might lead us to the identification of prognostic biomarkers related to the recurrence gains significance of paramount proportions.

Declarations

Ethical Approval and Consent to participate: Not applicable.

Human and Animal Ethics: Not applicable.

Consent for publication: Not applicable.

Availability of supporting data: Will be made available upon considerable request.

Competing interests: The authors declare no competing interests.

Funding: Not applicable.

Authors' contributions: AGNIK HALDAR: Conceptualization, Data Curation, Draft Preparation, Investigation, Methodology, Software, Visualization, Validation, Writing.

AJAY KUMAR SINGH: Supervision, Investigation, Methodology, Software, Visualization, Validation, Writing- Reviewing and Editing

Acknowledgements: We would like to thank the Department of Bioinformatics, Central University of South Bihar for providing us with the provisions to carry out this experiment.

Authors' information:

1. Dr. Ajay Kumar Singh (**Corresponding Author**)

Associate Professor

Department of Bioinformatics

Center for Biological Sciences (Bioinformatics)

Central University of South Bihar,

Panchanpur Road, Fathehpur, Tekari - Gaya-824236

Phone No: 91-9935686230

E-mail: ajaysingh@cusb.ac.in

2. Agnik Haldar

Department of Bioinformatics

Center for Biological Sciences (Bioinformatics)

Central University of South Bihar,

Panchanpur Road, Fathehpur, Tekari - Gaya-824236

Phone No: 91-8582853140

Email: halderagnik@gmail.com, agnikhaldar@cusb.ac.in

References

1. Irimie, A. I., Braicu, C., Sonea, L., Zimta, A. A., Cojocneanu-Petric, R., Tonchev, K., Mehterov, N., Diudea, D., Buduru, S., & Berindan-Neagoe, I. (2017). A Looking-Glass of Non-coding RNAs in oral cancer. *International journal of molecular sciences*, *18*(12), 2620. https://doi.org/10.3390/ijms18122620

2. Gibb, E. A., Enfield, K. S., Stewart, G. L., Lonergan, K. M., Chari, R., Ng, R. T., Zhang, L., MacAulay, C. E., Rosin, M. P., & Lam, W. L. (2011). Long non-coding RNAs are expressed in oral mucosa and altered in oral premalignant lesions. *Oral oncology*, *47*(11), 1055–1061. https://doi.org/10.1016/j.oraloncology.2011.07.008

3. Cogill, S. B., & Wang, L. (2014). Co-expression Network Analysis of Human lncRNAs and Cancer Genes. *Cancer informatics*, *13*(Suppl 5), 49–59. https://doi.org/10.4137/CIN.S14070

4. Alaei, S., Sadeghi, B., Najafi, A., & Masoudi-Nejad, A. (2019). LncRNA and mRNA integration network reconstruction reveals novel key regulators in esophageal squamous-cell carcinoma. *Genomics*, *111*(1), 76–89. https://doi.org/10.1016/j.ygeno.2018.01.003

5. Jean-Quartier, C., Jeanquartier, F., Jurisica, I. *et al. In silico* cancer research towards 3R. *BMC Cancer* 18, 408 (2018). https://doi.org/10.1186/s12885-018-4302-0

6. Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., & Su, J. (2018). Recurrence-Associated Long Non-coding RNA Signature for Determining the Risk of Recurrence in Patients with Colon Cancer. *Molecular therapy. Nucleic acids*, *12*, 518–529. https://doi.org/10.1016/j.omtn.2018.06.007

7. Wang, B., Zhang, S., Yue, K., & Wang, X. D. (2013). The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases. *Chinese journal of cancer*, *32*(11), 614–618. https://doi.org/10.5732/cjc.012.10219

8. Singh, R., De Sarkar, N., Sarkar, S., Roy, R., Chattopadhyay, E., Ray, A., Biswas, N. K., Maitra, A., & Roy, B. (2017). Analysis of the whole transcriptome from gingivo-buccal squamous cell carcinoma reveals deregulated immune landscape and suggests targets for immunotherapy. *PloS one*, *12*(9), e0183606. https://doi.org/10.1371/journal.pone.0183606

9. Sticht C, De La Torre C, Parveen A, Gretz N (2018) miRWalk: An online resource for prediction of microRNA binding sites. PLOS ONE 13(10): e0206239. https://doi.org/10.1371/journal.pone.0206239

10. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Molecular Cell, 27:91-105 (2007)

11. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

12. Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

13. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

14. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

16. Singh, S., Pillai, S., & Chellappan, S. (2011). Nicotinic acetylcholine receptor signaling in tumor growth and metastasis. *Journal of oncology*, *2011*, 456743. https://doi.org/10.1155/2011/456743

17. Mi, H., & Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods in molecular biology (Clifton, N.J.)*, *563*, 123–140. https://doi.org/10.1007/978-1-60761-175-2_7

18. Chen, E.Y., Tan, C.M., Kou, Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128 (2013). https://doi.org/10.1186/1471-2105-14-128

Figures

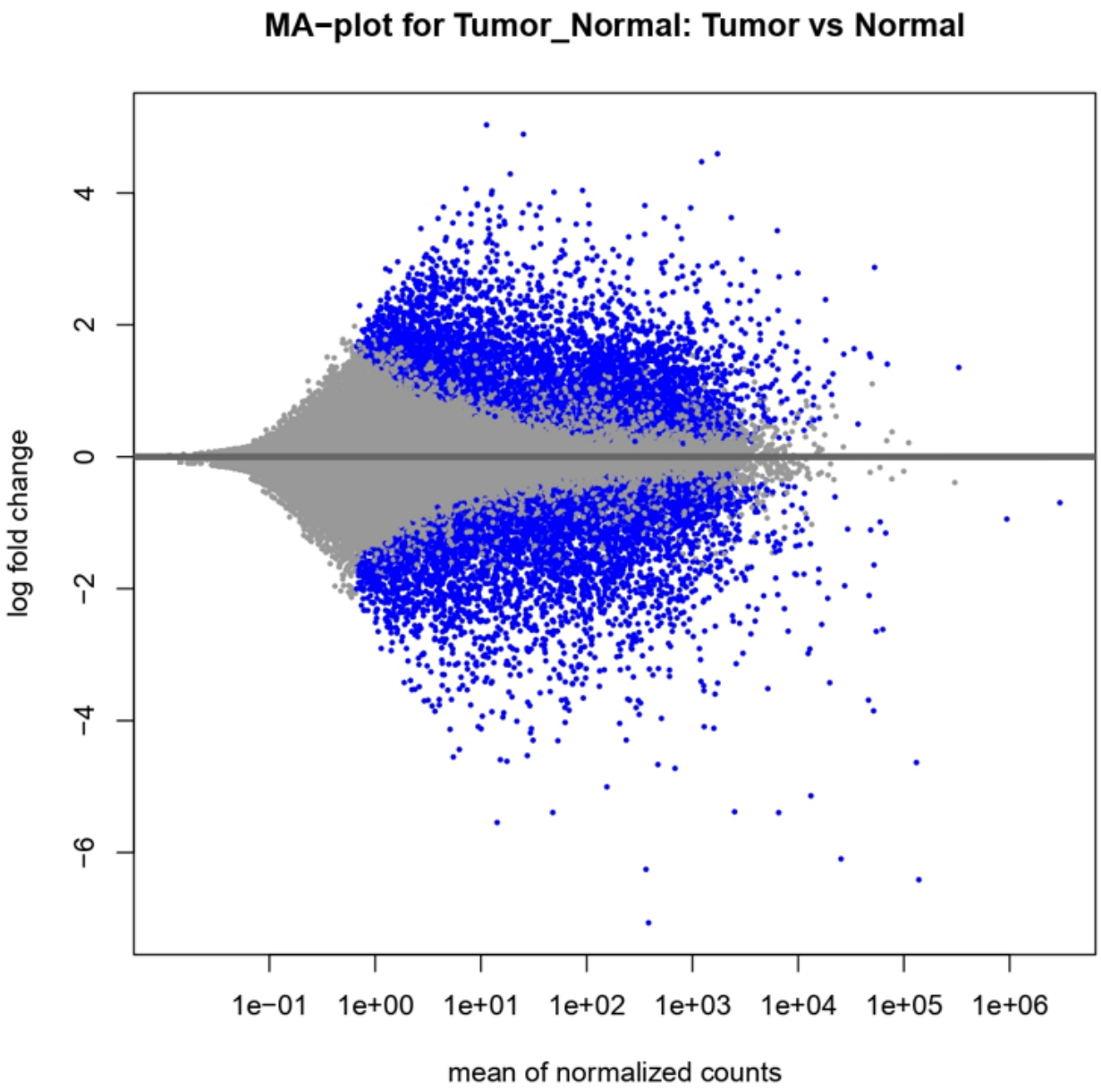

Figure 1

The dysregulation of genes portrayed in the form of a MA plot.

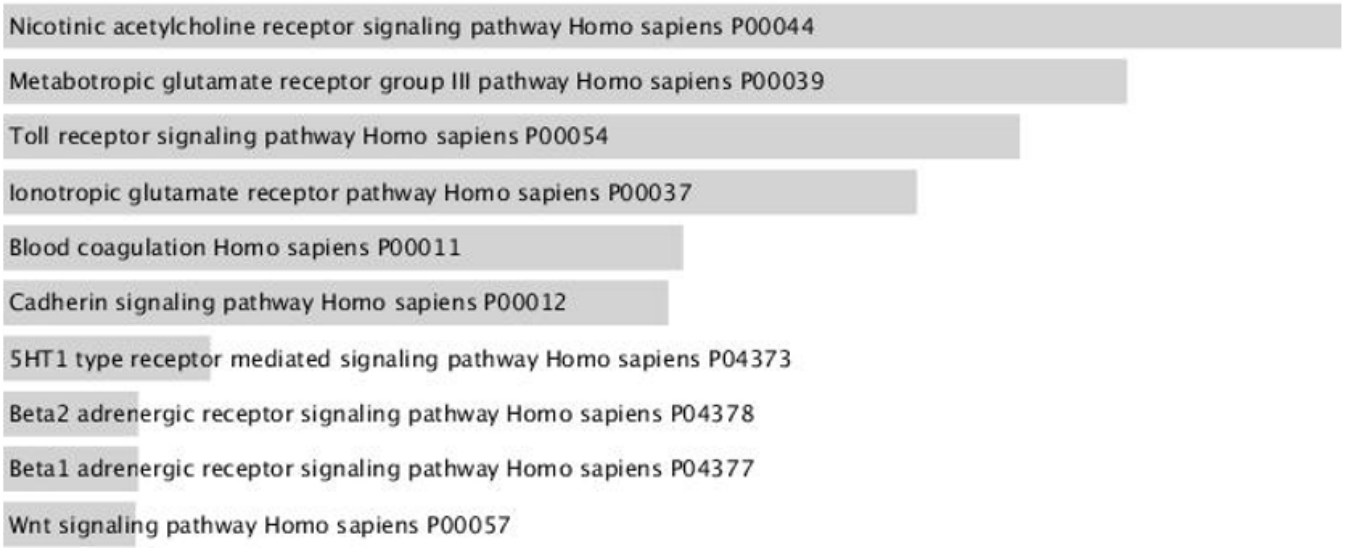

Figure 2

PANTHER database pathway analysis shows the prevalence of Nicotinic Acetylcholine Receptor Signaling which has a history in Tumor Growth and Metastasis

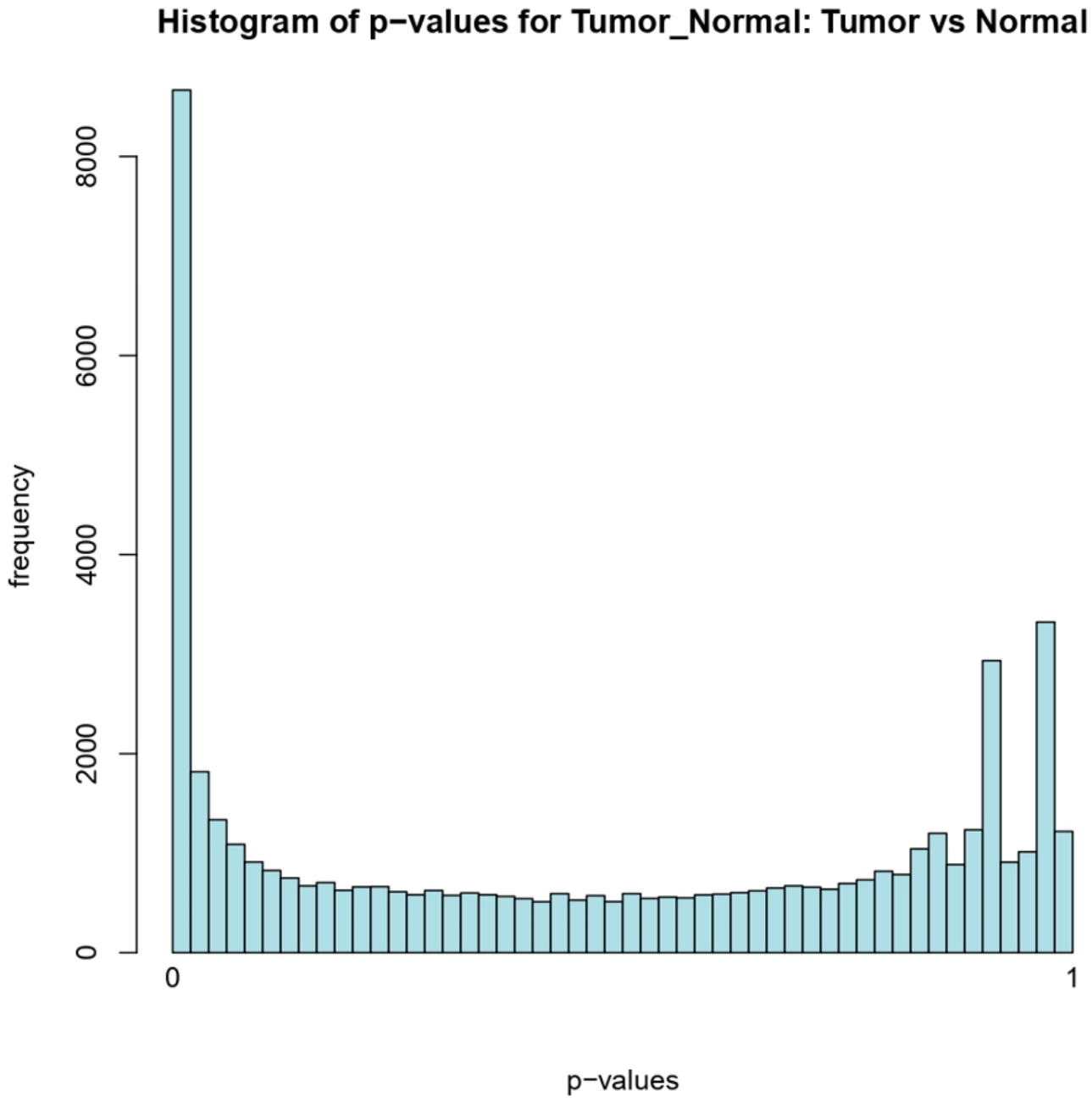

Figure 3

p-values of the Tumor vs Normal genes