Involvement of RP11-334E6.12 lncRNA in gastric cancer pathogenesis

Network based data analysis course report

Stefano Cretti

May 2022

Abstract

We performed an exploratory research on a dataset containing RNA-seq data, both from gastric cancer biopsy and healthy tissue from the same patients. The aim was to try and identify potential RNA based bio-markers for clinical applications, which ideally could be detected from serum. We found that a lncRNA, RP11-334E6.12, might be strongly linked to gastric cancer, due to its high contribution in several predictive models. In literature, this lncRNA shows no prior connection to gastric cancer, and more generally only correlation with breast cancer aggressive phenotype is reported. Studies on serum levels are needed to evaluate potential clinical applications.

Introduction

Gastric cancer (GC) is one of the most common type of tumors among both sexes (albeit with different incidence) and it has a very high mortality rate. Early-onset gastric cancer (EOGC) is a relatively rare subset of gastric cancer, representing about 10% of all GCs. EOGC manifests before the age of 45, therefore it is strongly linked to genomic mutations rather than prolonged exposition to carcinogenic substances; this fact makes EOGCs ideal to study and identify early or major mutations in the carcinogenic process that leads, overtime, to GC [1, 2].

In this paper we analyze a dataset from the study "Proteogenomic Characterization of Human Early-Onset Gastric Cancer"[3]; this dataset, available on recount3 [4, 5] (SRA Study: SRP172499), contains RNA-Seq data of paired healthy stomach and gastric cancer samples for 79 EOGC patients (mostly diffuse-type) plus 1 gastric cancer sample from another EOGC patient. The focus of the original paper is to identifying sub-types of EOGC, adding to the RNA-seq data information from other sources, namely proteomics, protein phosphorylation, and protein N-glycosylation. On the other hand, our analysis will focus on statistical methods to classify a sample into healthy tissue or gastric cancer, functional enrichment analysis and network-based analysis; this could lead to a better understanding of the highly heterogeneous pathogenetic processes that originate GC, plus potentially identifying new genes never before associated to GC that could be used as bio-markers.

Methods

Most of the analysis was performed using RStudio [6](full session details can be found in the supplementary materials, Suppl0). Before each step entailing randomness, `set.seed` function was used in order to obtain reproducible results (with seed equal to 1234 unless specified otherwise).

Firstly, sample metadata, gene annotation (gencode v26) and raw gene counts were directly

downloaded from Sequence Read Archive using `create_rse_manual` (`recount3` library [4, 5]). The raw gene counts were then normalized for both inter-sample and intra-sample variability using GeTMM (Gene length corrected TMM) (`edgeR` library [7, 8, 9]) and then rescaled in counts per million (cpm). After that, the counts were filtered, keeping only those with more than 1 cpm in at least 20% of the samples (`genefilter` library [10]).

Then, different statistical methods, both supervised and unsupervised, were tested on the dataset. Throughout all methods, only the division between healthy and cancer tissue is considered, since the rest of the metadata is mostly homogeneous among the cohort of patients. The following unsupervised methods were performed: principal component analysis, k-means clustering, k-means clustering on $\log_{10}$ transformed counts and hierarchical clustering (all from base `stats` library [11]).

A loose feature selection was then applied using Wilcoxon rank sum test (`WilcoxCV` library [12], created 2 data partitions with 80% of the samples in the training set using `generate.split` function, performed Wilcoxon rank sum test on both of them using `wilcox.selection.split` function with `algo="new"`, adjusted p-values using Holm–Bonferroni correction, selected higher p-value of the two iterations, kept features with p-value below the arbitrary threshold 0.001).

The following supervised methods were performed: random forest (`randomForest` library [13]), linear discriminant analysis (`MASS` library [14], data partition with 75% of the samples in the training set using `caret` library [15]), lasso regression (`glmnet` library [16], both `glmnet` and `cv.glmnet` functions with parameters `standardize = FALSE, family = "binomial"`) and Signature-based Clustering for Diagnostic Purposes (SCUDO) (`rScudo` library [17], data partition with 50% of the samples in the training set using `caret` library and seed equal to 2345 in order to get a plottable graph, train model using `scudoTrain`, perform validation using `scudoTest`, evaluate performance using `scudoClassify`; parameters `nTop=200, nBottom=200, alpha=0.05, N=0.3` where needed).

Classifiers were evaluated in terms of accuracy, sensitivity and specificity, and then compared to determine the best performing method.

After associating the p-values obtained from Wilcoxon rank sum test, the list of genes obtained from the best performing method is used to perform functional enrichment analysis and network based analysis, both on the top 200 genes from the list. Functional enrichment analysis was performed using the `gost` function with parameter `ordered_query = TRUE` (`gprofiler2` library [18]). Network based analysis was instead performed using the `run_pathfindR` function (`pathfindR` library [19]).

Further manual analysis was performed on the non-coding RNA genes found in the list, especially on RP11-334E6.12 lncRNA, which is among the top scoring features for the random forest classifier.

Results

After pre-processing and filtering, 21′893 of the original 63′856 features were kept and each sample displays a symmetric distribution of the cpms of the features (Suppl1). We then performed PCA (Figure 1) and obtain that the first 3 dimensions explain over 90% of the variance (Suppl2); moreover the GC samples cluster tightly together, while the HT ones are more spread. This might seem counter-intuitive at first but it might be due to the fact that the tumorigenic process is guided by similar patterns despite the heterogeneous tumor origin.

When performing K-means clustering we do not get a good classifier (59 out of 79 HT samples classified as GC); this is mostly due to the wide spread of the HT samples, which skews cluster formation (Suppl3). Repeating this operation applying a log-transform we get more reliable results (0.90 accuracy) (Figure 2). Hierarchical clustering is fairly good at predicting GC samples, but not as good as log-transformed K-means; moreover it leads to the creation of many single-samples branches which cannot be reliably identified (Suppl4).

We then performed feature selection, keep-

ing the $3'546$ most influential genes according to Wilcoxon-rank sum test. Fitting a random forest classifiers we get almost perfect classifications (over 0.98 accuracy) and we notice a stark contrast in the expression levels of the top influential genes (Figure 3); moreover we notice that the first 200 genes already explain most of the model (Suppl5). Performing LDA we also obtain a very good separation (Suppl6) (0.92 accuracy on test data). Using lasso regression we get a classifier with decent performance but significantly lower compared to the other methods (0.85 accuracy on test data). We finally performed SCUDO analysis, which yields very good results too (0.94 accuracy) (Suppl7).

Values of accuracy, sensitivity and specificity of all tested methods are reported in the following table (PCA and hierarchical clustering are not included). Some models could be further optimized and cross validation could be performed in order to get more robust estimates; however, since the random forest classifier already performs extremely well, the list of genes sorted by contribution to this model was used for the downstream analyses.

Methods	Acc.	Sens.	Spec.
K-means	0.629	1.000	0.253
K-means ($\log_{10}$)	0.899	0.875	0.924
Random forest	0.981	0.988	0.975
LDA	0.923	0.900	0.947
Lasso regression	0.846	0.889	0.810
SCUDO	0.937	0.949	0.925

When performing functional enrichment analysis on the top 200 genes by importance (Figure 4), we notice a significant enrichment in tissue remodelling functions and extra cellular matrix deposition, which is consistent with what is expected from GC [20].

Performing network enrichment analysis we get similar results (Figure 5), with the top enriched terms belonging to extracellular matrix interaction, regular gastric functions, inflammatory and pathogenic processes.

It is interesting to notice that among the 25 most significant genes, 4 of them are non-coding RNA genes. Looking at different resources (GeneCards, Pubmed and LNCipedia), 2 of them have been already linked with gastric cancer (MIR4435-2HG and RP1-102K2.8), while the other 2 have no prior connection to gastric cancer (RP11-334E6.12 and WISP1-OT1) (Suppl8). Of these, RP11-334E6.12, also known as THY1-AS1, is particularly interesting since it has been linked with aggressive breast cancer phenotype in a single publication[3]. Moreover the supposed target of this antisense lncRNA, THY1 (or CD90) has been shown to be relevant in several types of cancer among which gastric cancer[21].

Discussion

It seems that using RNA-seq on gastric biopsies it is not too challenging to distinguish between healthy tissue and gastric cancer, since several methods, both supervised and unsupervised, allow a confident classification. That being said, gastric biopsy is an invasive procedure and not feasible for screening or repetitive follow up after the pathology. Looking through the genes used by the best performing classifier (random forest after feature selection), we notice the presence of several non-coding RNAs, of which some are not clearly characterised yet. Understanding the role and the exact mechanism of action of these transcripts and studying their presence in either serum or extracellular vesicles could provide a specific non invasive bio-marker which is still needed in clinical settings.

2

References

[1] Julita Machlowska et al. "Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies". In: *International Journal of Molecular Sciences* (2020).

[2] Elizabeth C Smyth et al. "Gastric cancer". In: *The Lancet* (2020).

[3] Dong-Gi Mun et al. “Proteogenomic Characterization of Human Early-Onset Gastric Cancer”. In: *Cancer Cell* (2019).

[4] Leonardo Collado-Torres. *Explore and download data from the recount3 project*. 2022.

[5] Christopher Wilks et al. “recount3: summaries and queries for large-scale RNA-seq expression and splicing”. In: *Genome Biol* (2021).

[6] RStudio Team. *RStudio: Integrated Development Environment for R*. 2022.

[7] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* (2010).

[8] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. In: *Nucleic Acids Research* (2012).

[9] Yunshun Chen, Aaron A T Lun, and Gordon K Smyth. “From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline”. In: *F1000Research* (2016).

[10] Robert Gentleman et al. *genefilter: genefilter: methods for filtering genes from high-throughput experiments*. 2022.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. 2022.

[12] Anne-L. Boulesteix. “WilcoxCV: an R package for fast variable selection in cross-validation”. In: *Bioinformatics* (2007).

[13] Andy Liaw and Matthew Wiener. “Classification and Regression by randomForest”. In: *R News* (2002).

[14] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

[15] Max Kuhn. *caret: Classification and Regression Training*. 2022.

[16] Noah Simon et al. “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent”. In: *Journal of Statistical Software* (2011).

[17] Matteo Ciciani, Thomas Cantore, and Mario Lauria. “rScudo: an R package for classification of molecular profiles using rank-based signatures”. In: *Bioinformatics* (2019).

[18] Liis Kolberg et al. “gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler”. In: *F1000Research* (2020).

[19] Ege Ulgen, Ozan Ozisik, and Osman U Sezerman. “pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks”. In: *Frontiers in Genetics* (2019).

[20] Yukiko Oya, Yoku Hayakawa, and Kazuhiko Koike. “Tumor microenvironment in gastric cancers”. In: *Cancer Science* (2020).

[21] Yun Hu et al. “Multiple roles of THY1 in gastric cancer based on data mining”. In: *Translational Cancer Research* (2020).

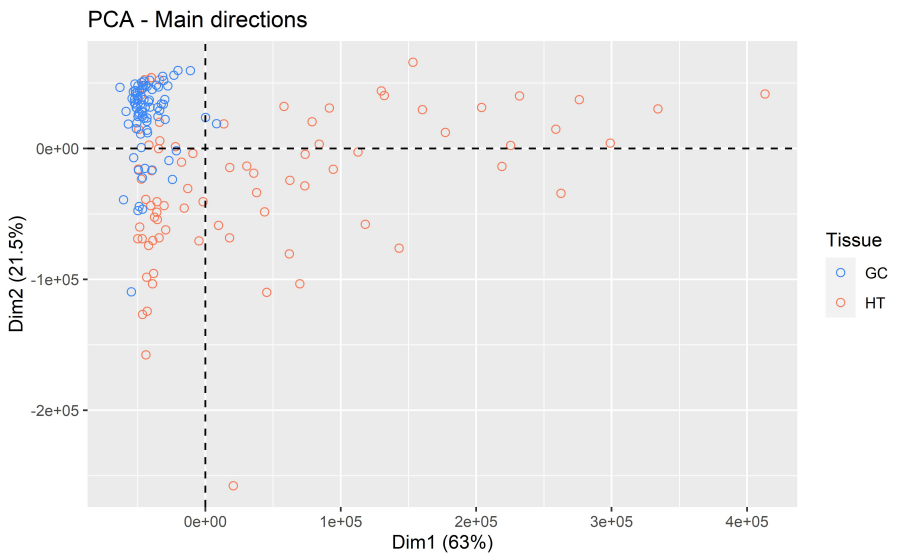


Figure 1: Principal component analysis plot; we notice that the first dimension mostly contributes to separating the healthy tissue samples from each other, while the second defines a tight cluster of gastric cancer samples.

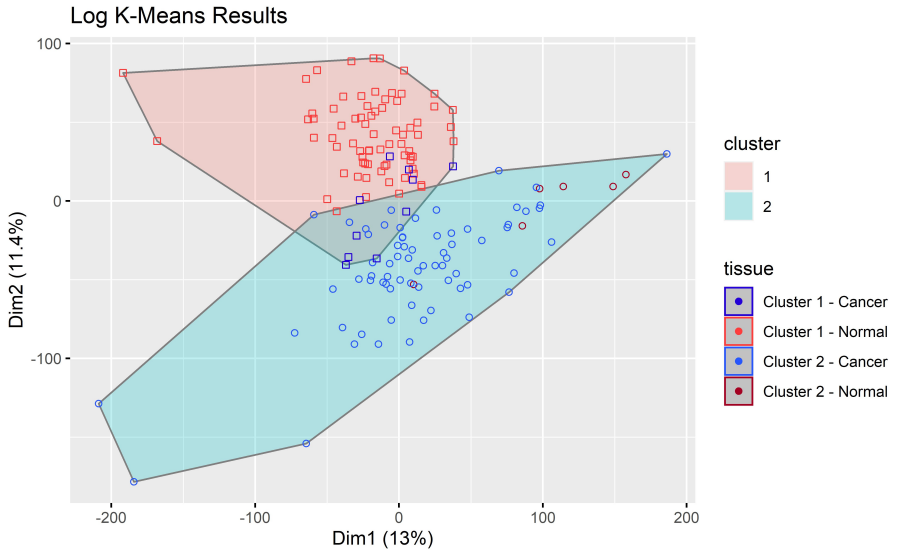


Figure 2: K-means plot of log10 transformed counts; by setting the number of clusters to 2, we see that red cluster mostly contains normal samples while the blue one contains mostly cancer samples.

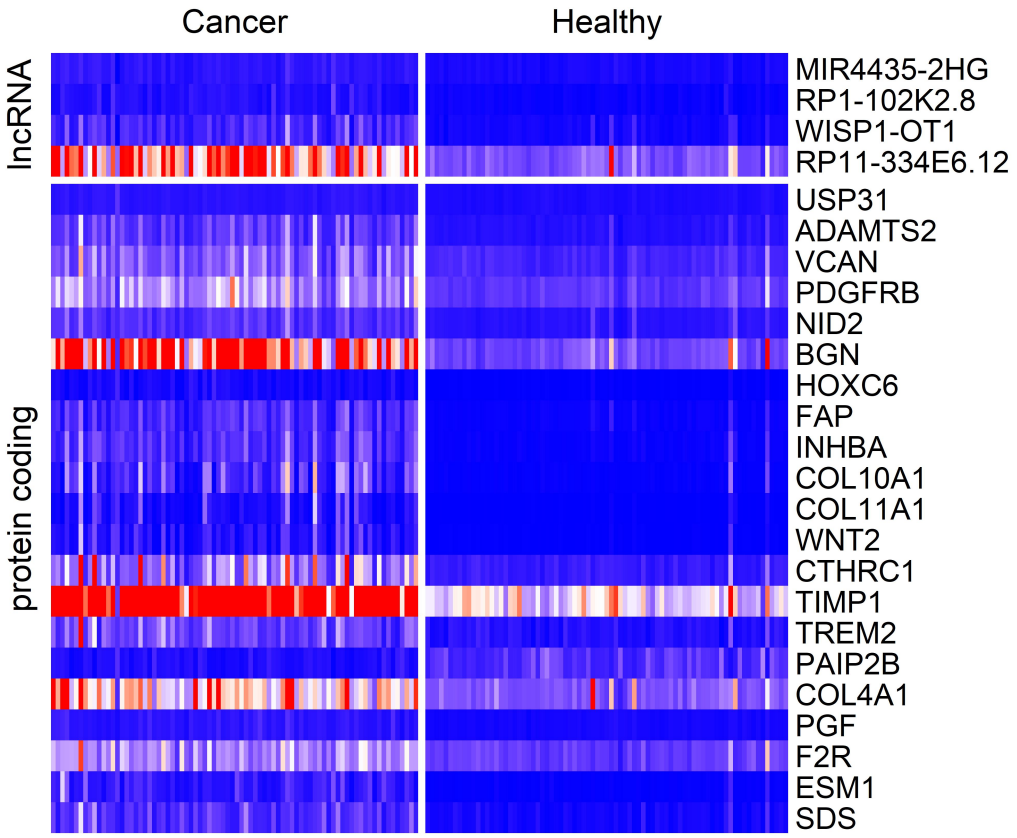
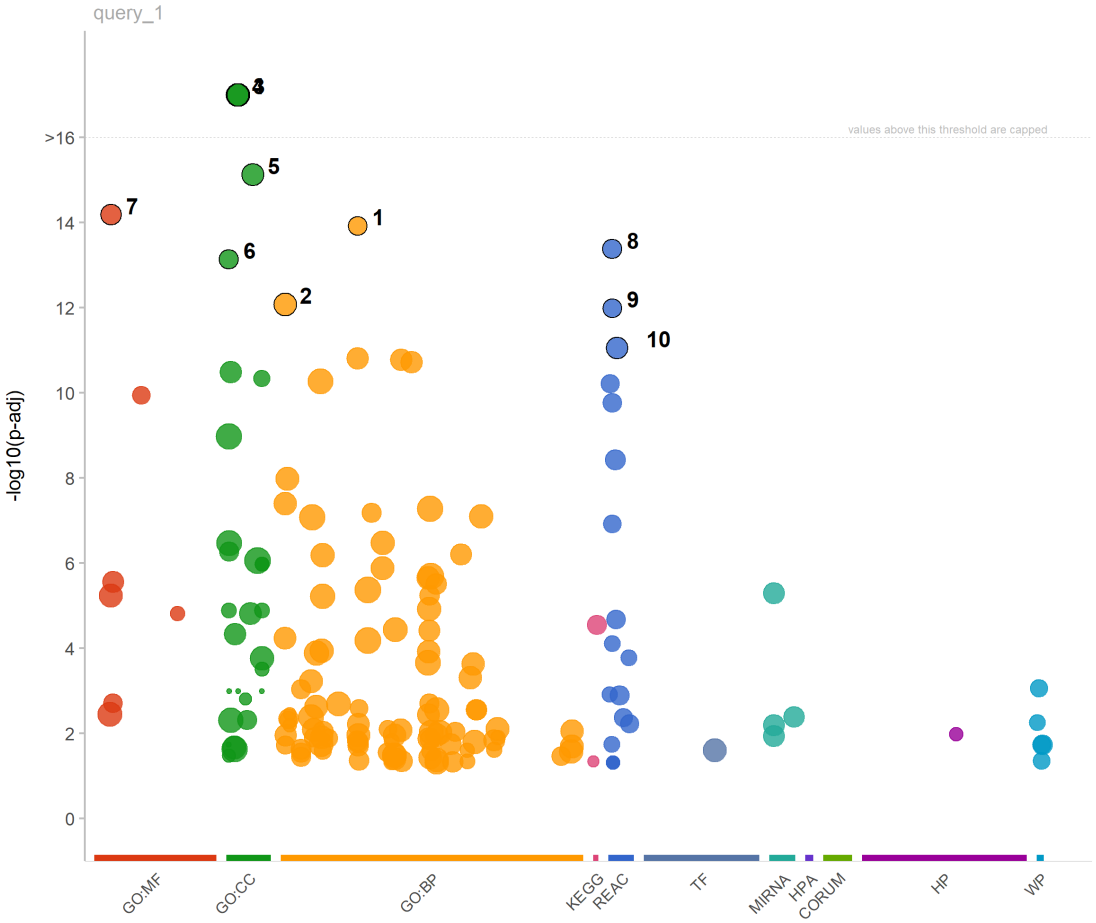


Figure 3: Heat map of the top 25 genes according to random forest classifier; in general, the top 25 genes analyzed are over-expressed in the cancer sample. Values above 200 are capped. 4 genes are RNA coding while the other 21 are protein coding.



id	source	term_id	term_name	term_size		p_value
1	GO:BP	GO:0030199	collagen fibril organization	73		1.2e-14
2	GO:BP	GO:0001501	skeletal system development	535		8.3e-13
3	GO:CC	GO:0031012	extracellular matrix	561		6.0e-18
4	GO:CC	GO:0030312	external encapsulating structure	562		6.3e-18
5	GO:CC	GO:0062023	collagen-containing extracellular matrix	423		7.6e-16
6	GO:CC	GO:0005581	collagen trimer	93		7.4e-14
7	GO:MF	GO:0005201	extracellular matrix structural constituent	173		6.6e-15
8	REAC	REAC:R-HSA-1474290	Collagen formation	89		4.3e-14
9	REAC	REAC:R-HSA-1650814	Collagen biosynthesis and modifying enzymes	67		1.0e-12
10	REAC	REAC:R-HSA-1474244	Extracellular matrix organization	298		8.7e-12

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 4: Enrichment analysis plot; from the plot we notice a great enrichment in terms related to extracellular matrix and tissue remodeling, as expected in a cancerous tissue.

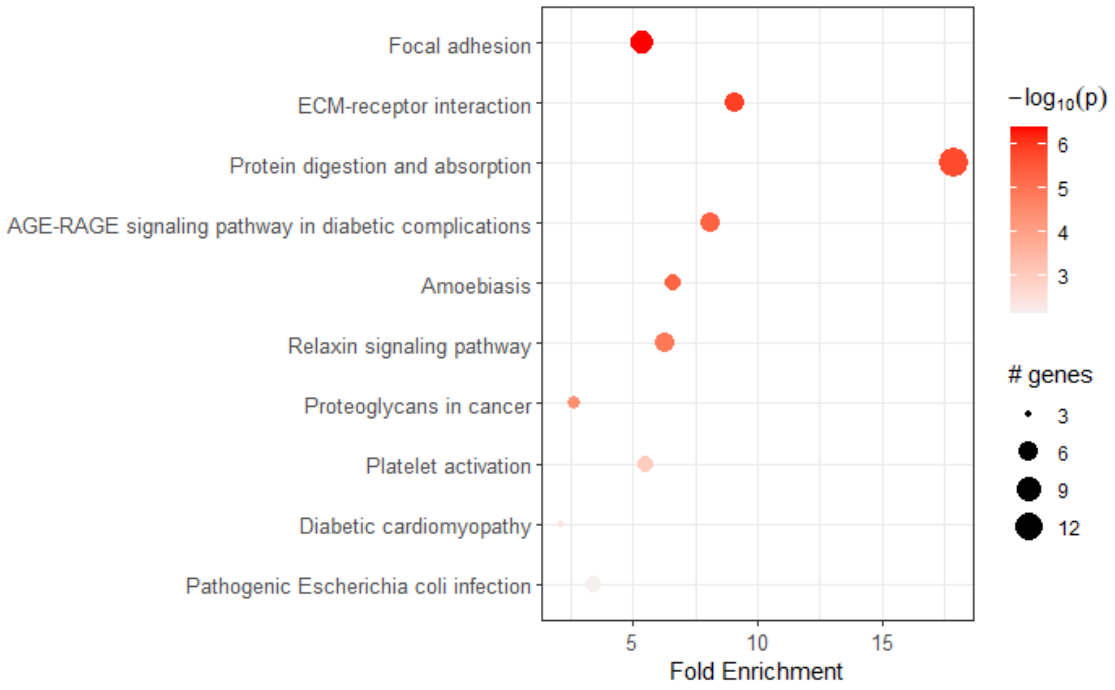


Figure 5: Pathfindr plot; we observe a sizeable over-representation of tissue remodeling and gastric functions.