Research Article ISSN 2639-9458

Microbiology & Infectious Diseases

Machine-Learning Analysis of Single-Cell RNA-Seq Biomarkers in Gastric Cancer Caused By *Helicobacter Pylori*

Eric Li[1], Valentina L. Kouznetsova[2,4], Santosh Kesari[6] and Igor F Tsigelny[2–5,*]

[1]*REHS program, San Diego Supercomputer Center, University of California, San Diego, Calif., USA.*

[2]*San Diego Supercomputer Center, University of California, San Diego, Calif., USA.*

[3]*Department of Neurosciences, University of California, San Diego, Calif., USA.*

[4]*Biana, La Jolla, Calif, USA.*

[5]*CureScience, San Diego, Calif, USA.*

[6]*Pacific Neuroscience Institute, Santa Monica, Calif, USA.*

***Correspondence:**

Igor F Tsigelny, San Diego Supercomputer Center, University of California, San Diego, Calif., USA, (Orcid ID: 0000-0002-7155-8947).

Received: 17 Mar 2022; **Accepted:** 22 Mar 2022; **Published:** 07 Apr 2022

Citation: Eric Li, Kouznetsova VL, Kesari S, et al. Machine-Learning Analysis of Single-Cell RNA-Seq Biomarkers in Gastric Cancer Caused By *Helicobacter Pylori* . Microbiol Infect Dis. 2022; 6(2): 1-19.

ABSTRACT

Gastric cancer is one of the most prevalent and deadly cancers in the world. One of the biggest factors for this disease is Helicobacter pylori (H. pylori)*, infecting roughly half of the world's population. However, there is a limited understanding of the* H. pylori *infection at single-cell level. In this study, single-cell RNA-Seq datasets from intestinal metaplasia samples were analyzed.*

Using bioinformatics methods, the cells were clustered and cell types were identified with cell type specific marker genes. For each cell type, H. pylori *infected cells were compared with control cells using statistical analysis in order to find significant genes and pathways. Then, machine-learning (ML) approaches were used to build models to distinguish* H. pylori *positive and negative cells, and the severity of infection.*

It is found that H. pylori *infection is linked to an increase in enterocytes and a decrease in pit mucous cells (PMCs). These changes may promote disease progression from gastritis to gastric cancer. Significantly differentially expressed genes and several pathways such as the MHC class II antigen presentation pathway and the PD-1 pathway were identified. The random forest-based models achieved an accuracy of higher than 97% for detecting positivity and severity.*

We identified the specific type of the host cells along with signaling pathways related to H. pylori *infection and signaling pathways leading to gastric cancer. We demonstrated that ML methods are useful in detection of the affected by* H. pylori *PMC cells.*

Keywords

Helicobacter pylori, Machine learning, Cancer, PMC cells.

Acronyms and Abbreviations

FDR: False Discovery Rate; GMC: Gland Mucous Cell; *H. pylori*: *Helicobacter pylori*; IMS: Intestinal Metaplasia, Severe; IMW: Intestinal Metaplasia, Wild; MHC: Major Histocompatibility Complex; ML: Machine Learning; NGS: Next-Generation Sequencing; NMD: Nonsense-Mediated Decay; PCA: Principal Component Analysis; PD-1: Programmed Death-1; PMC: Pit

Mucous Cell; RF: Random Forest; RNA-Seq: RNA Sequencing; scRNA-Seq: Single Cell RNA Sequencing; tSNE: t-Distributed Stochastic Neighbor Embedding.

Introduction

Gastric cancer is one of the deadliest and most widespread cancers in the world, killing an estimated 738 000 people in 2018 [1]. One of the biggest factors for this disease is the bacteria *Helicobacter pylori* (*H. pylori*) [2], which was discovered in 1982 by Marshall and Warren [3] and was found to be linked to gastric cancer in 1994. These bacteria shockingly infect around half of the world's population, but most infections are asymptomatic [4]. Besides gastric cancer, *H. pylori* also causes many other issues, such as gastritis, ulcers, allergies and much more. Despite knowing what possible symptoms may occur from this infection, it is still near impossible to predict what exactly will happen when a certain individual is infected due to factors such as bacterial strain and human/environmental determinants. There are many known mechanisms in a *H. pylori* infection, with the most well-known being production and injection of CagA protein to the target cell, the first identified bacterial protein correlated in cancer [5]. Most strains of *H. pylori* contain the CagA gene, with around 60–70 percent of western strains and almost all eastern strains containing it. Additionally, not all CagA are the same, with the protein structure in eastern and western strains being slightly different. They differ in their EPIYA and CM motifs, which are segments of the protein repeated throughout the sequence.

CagA is delivered into the cell by a structure called a T4SS syringe, which is formed from many adhesins such as BabA, BabB and SabA [5]. After entry, CagA attaches to the plasma membrane and has two different mechanisms in which this is achieved, depending on the polarity of the cell. If the cell is polar, the central region is responsible for binding. If the cell is nonpolar, the C-terminal region is responsible for binding. Then, the EPIYA-C/EPIYA-D regions interact with the SH2 domain of SHP2, which is an enzyme, which then triggers signaling also triggered by growth factors. Because of this, the cell elongates into what is called a hummingbird phenotype, leading to increased cell motility, which contributes to tumor metastasis. When comparing western and eastern CagA, eastern CagA is a greater risk for gastric cancer than a western strain with one EPIYA-C motif, but the strains with multiple EPIYA-C pose a greater risk [6].

RNA-Seq is a method to study the gene expression in biological samples using next-generation sequencing (NGS) technology [7]. RNA-Seq can be used to identify the differentially expressed genes in disease samples so that disease biomarkers can be identified, and used to find pathways of interest that can be used to develop drugs that treat certain diseases and conditions. In RNA-Seq projects, the sequences are first mapped to the human genome or the genome of the species from which the samples were taken. Then the expression of the genes are calculated and genes from different groups of samples, such as control and disease groups, are compared using statistical methods so that significantly up- and downregulated genes can be found. Because traditional RNA-Seq studies the bulk of tissue samples, the gene expression measured by RNA-Seq is a mixture of the gene expression of different cell types in the sample. It is difficult to study heterogeneous systems made from different cell types. In recent years, a new technology called single cell RNA-Seq (scRNA-Seq) has been developed [8]. scRNA-Seq is the most advanced method to study gene expressions, because it can detect the RNA expression of each individual cell, with a single experiment able to study hundreds of thousands of cells. In a scRNA-Seq study, each cell is captured in a small droplet and a molecular barcode is attached to all RNAs in the cell. Afterwards, the gene expression for each cell can be identified using the barcode [9].

The analysis of scRNA-Seq data is similar to the bulk sample RNA-Seq data. However, there are several unique analyses, which are specific only to scRNA-Seq. First, the cells are clustered using gene expression data. Second, the cell type is identified according to cell type specific gene markers. Currently, there are many available bioinformatics methods for the analysis of scRNA-Seq data. The Seurat package is one of the most popular tools [10].

scRNA-Seq is more powerful than the traditional RNA-Seq, because it can find biomarkers or pathways for specific cell types. But since this is a very novel technology, there are not many existing research projects that study gastric cancer and *H. pylori* using this method. Zhang and coauthors, used scRNA-Seq to investigate gastric cancer samples [11]. This study found and compared cell types between tissues from different stages of cancer and found interesting biomarkers between different stages of cancer. We analyzed the published scRNA-Seq datasets to identify gene and pathway markers for *H. pylori* infection for different cell types. Machine-learning models were constructed for the prediction of *H. pylori* infection and classification of severity of infection.

Methods

We analyzed the scRNA-Seq datasets [11]. The gene-expression datasets GSE134520 of this study were downloaded from the NCBI GEO database. These datasets have samples from many tissues, including non-atrophic gastritis, chronic atrophic gastritis, early gastric cancer, and intestinal metaplasia. Only the intestinal metaplasia samples (Table S1), which have both *H. pylori* positive and negative samples, were selected for our analysis. Intestinal metaplasia is an early transformation that can lead to gastric cancer. Therefore, it can be useful to study the *H. pylori* related transformation to cancer. We used wild intestinal metaplasia (IMW) and severe intestinal metaplasia (IMS) datasets for analysis.

The scRNA-Seq data were analyzed in R using the Seurat package [10]. The raw data were first trimmed to remove data points without a cell, with multiple cells or with dead cells. Data points with less than 200 genes were considered without a cell. Data points with more than 20 000 total expression counts were considered multiple cells. Data points with more than 20% mitochondrial genes were considered dead cells. The expression data were normalized and

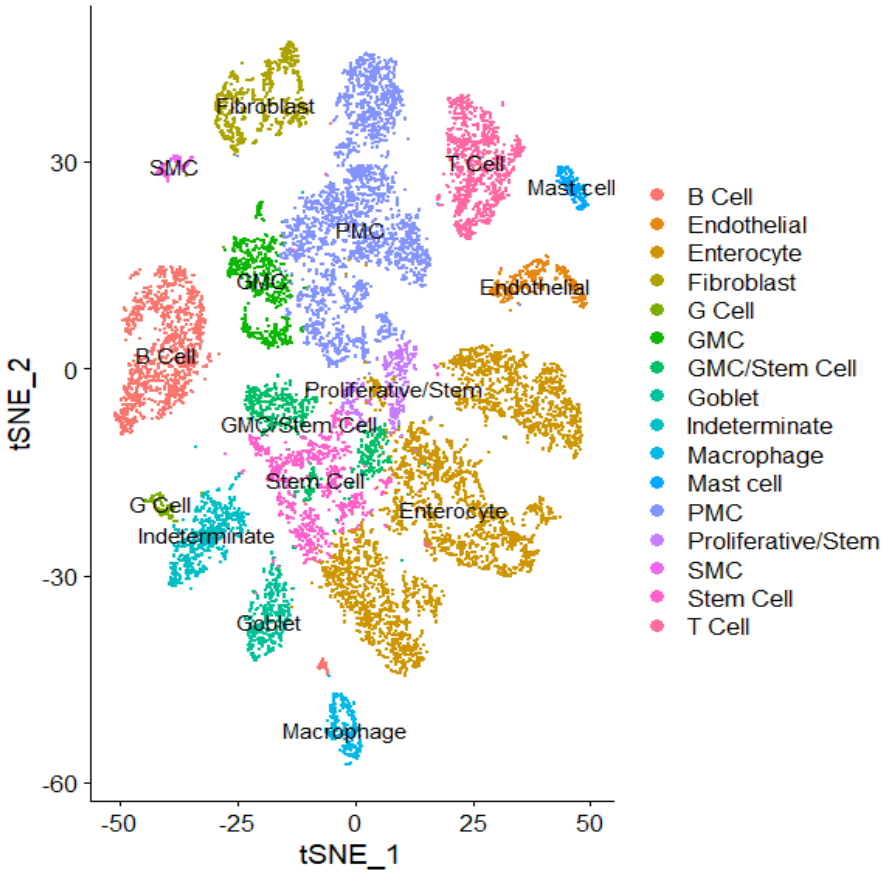

Figure 1: tSNE plot of cells by Cell type.

then the 2000 most variable genes were identified using the Find Variable Features from the Seurat package. The expression data were further scaled using the Scale Data command. For cell clustering, a principal component analysis (PCA) was first performed on the variable genes. Then cells were clustered by the Find Neighbors and the Find Clusters command from the Seurat package. After cell clustering, the cell types were identified using cell type specific marker genes described in previous study, listed in Table S2 [11]. For each cell type, significant differentially expressed genes between the *H. pylori* positive and negative cells were identified using the Wilcoxon Rank Sum test [12]. The *p*-values were adjusted using the Bonferroni method in R [13]. Significantly upregulated genes (adjusted *p*-value < 0.05) in *H. pylori* cells were uploaded to reactome.org to find the significant pathways that are enriched with these genes [14,15].

Machine-learning models were then built to predict *H. pylori* positive and negative cells. For each cell type, significant genes with adjusted *p*-value < 0.05 were used to train random forest (RF) models [16]. The Caret package [17] was used to optimize the parameters and calculate the model accuracy using a 5-fold cross validation repeated 5 times, from which the average accuracy was taken from. This was done using the train Control function from the Caret package. The default value of 500 trees were used in the RF models. A grid search was performed to optimize the *mtry* parameter, which refers to the number of variables used in each node in a tree. A value of *mtry* = 8 had the highest accuracy of 97.48 and a kappa value of 94.12. Also, using only the *H. pylori* positive cells, RF models were trained using the Caret package to predict the severity of *H. pylori* infection, distinguishing severe vs wild intestinal metaplasia cells. The same procedures of 5-fold cross validation and the grid search for *mtry* parameter were used, and with a *mtry* value of 14 the accuracy was 98.21 and the kappa was 95.89.

Results and Discussion

After cell clustering, 26 clusters were created (Figure S1). A cell type was then assigned to each cluster (Figures S2–S20, Table S3) by the expression of cell type specific marker genes (Table S2). The cell types are shown in Figure 1 and number of cells are listed in Table 1. Enterocyte and PMC are the most abundant cell types in these samples. There are more enterocytes in the *H. pylori* positive samples than the negative samples, but there are less PMCs in the positive samples. PMCs were selected to build machine-learning models. The intention of using PMCs is for the detection of *H. pylori* infection in the early stage of gastric cancer. Enterocytes are intestinal absorptive cells, which are found in patients with intestinal metaplasia. Prior to the intestinal metaplasia stage, enterocytes are rare in gastric tissues. However, PMCs are common in gastric tissues for all patients. Therefore, PMCs-based machine-learning models can be applied to a wider range of patients. There

are also large differences shown in Table 1 with Gland Mucous Cells (GMCs) and Stem cells. Despite this, however, we chose to not use these two types of cells in finding results. The difference in stem cells could be explained by the clusters assigned to GMCs/Stem Cells, as if we assume that the majority of cells in that cluster are stem cells, then the amount of stem cells would be relatively equal. GMCs were not used because they are not well studied and are very similar to PMCs.

Table 1: Number of cells for each cell type in *H. pylori* positive and negative samples.

Cell type	Number and percentage of cells (*H. pylori* +)	Number and percentage of cells (*H. pylori* −)
Enterocyte	2275 (32.57%)	1453 (19.78%)
Pit mucous cell (PMC)	869 (12.44)	1858 (25.30%)
B Cell	665 (9.52%)	645 (8.78%)
Stem Cell	186 (2.66%)	764 (10.40%)
T Cell	357 (5.11%)	593 (8.07%)
Fibroblast	340 (4.87%)	411 (5.60%)
Gland mucous cell (GMC)	559 (8.00%)	123 (1.67%)
GMC/Stem Cell	566 (8.10%)	88 (1.20%)
Indeterminate	91 (1.30%)	520 (7.08%)
Endothelial	223 (3.19%)	213 (2.90%)
Goblet	252 (3.61%)	157 (2.14%)
Proliferative/Stem	267 (3.82%)	128 (1.74%)
Macrophage	178 (2.55%)	111 (1.51%)
Mast cell	83 (1.19%)	108 (1.47%)
Smooth muscle cell (SMC)	51 (0.73%)	75 (1.02%)
G Cell	22 (0.32%)	97 (1.32%)

Both enterocytes and PMCs are epithelial cells. All our samples are from intestinal metaplasia. Intestinal metaplasia is a transformational stage where some of the cells that make up the lining of the stomach are replaced by the cells found at the lining of the intestine. Enterocytes are intestinal absorptive cells, kind of epithelial cells, which lines the inner surface of the small and large intestines. Therefore, *H. pylori* possibly plays a role in the development of intestinal metaplasia or could transform the intestinal metaplasia in a more severe stage tissue.

The significant genes for the Enterocytes and PMCs are shown in the heat maps in Figure 2. In these two heat map figures, the top 50 significantly differentially expressed genes in *H. pylori* positive (y) and negative (n) cells are shown, with the more significant genes on top. From these heat maps, we can also see that there are more PMCs in negative than positive samples, and there are more enterocytes in positive than in negative samples. In PMCs, several upregulated genes in positive samples, including LCN2, CD74, REG1A, BPIFB1, PIGR, HLA-DRB1, REG3A, MT-CO2, HLA-DRA, HLA-DPB1, and HLA-DPA1 were found. Some of these genes were also upregulated in the enterocytes. Significant genes were also identified for some additional cell types, which are shown in the heat maps (Figures S21, S22).

For each cell type, the upregulated genes in *H. pylori* positive cells were uploaded into the Reactome website (https://reactome.org/) and the enriched pathways for these genes were identified (Figures S23, S24). The immune system has the most significant pathways,

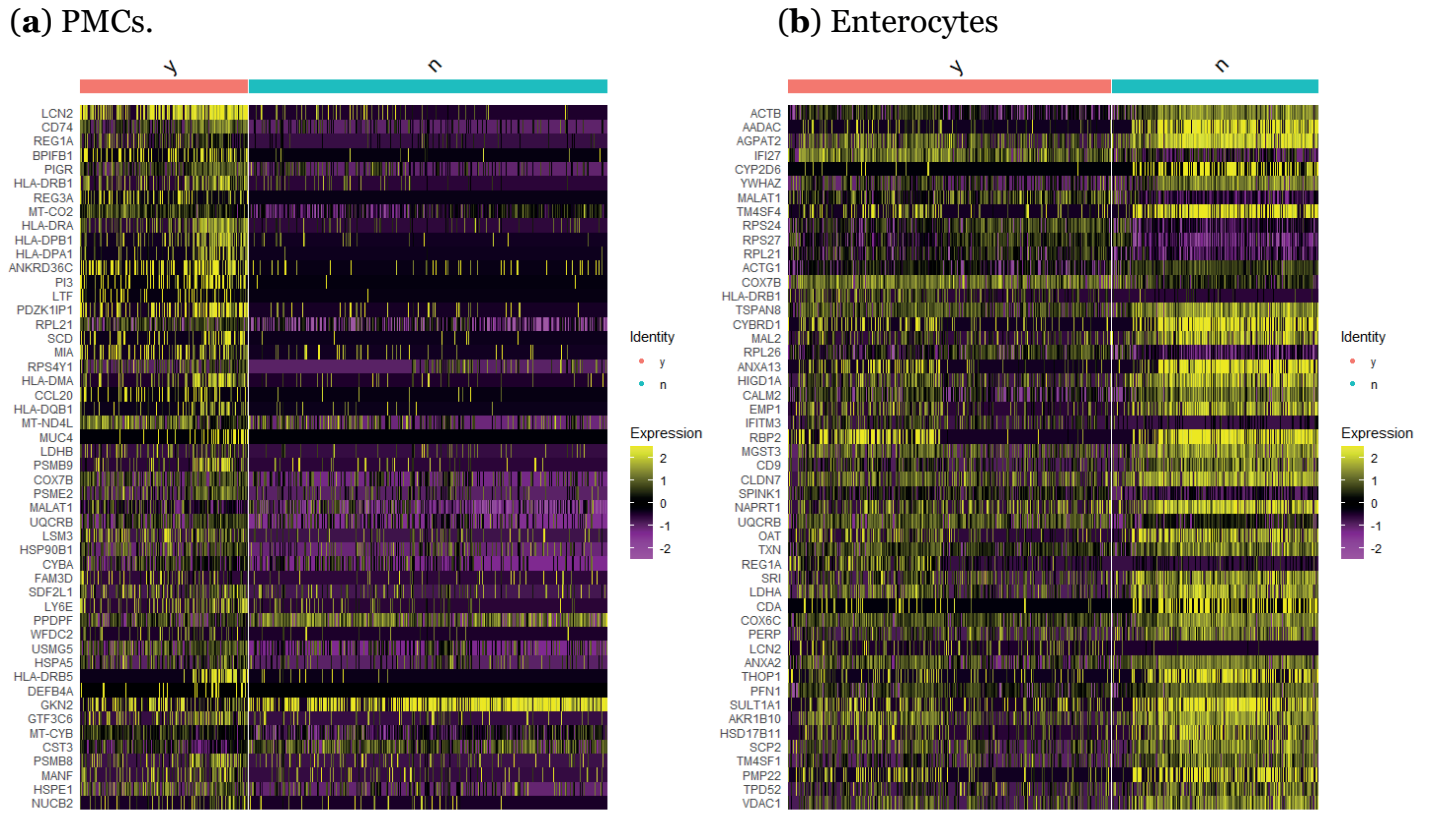

Figure 2: Two heat maps showing the top 50 most significantly differentiated genes in PMCs and enterocytes. In each heat map, the cells at the left side under the red bar denoted by letter "y" and the cells at the right side under the green bar denoted by letter "n" are *H. pylori* positive and negative cells respectively.

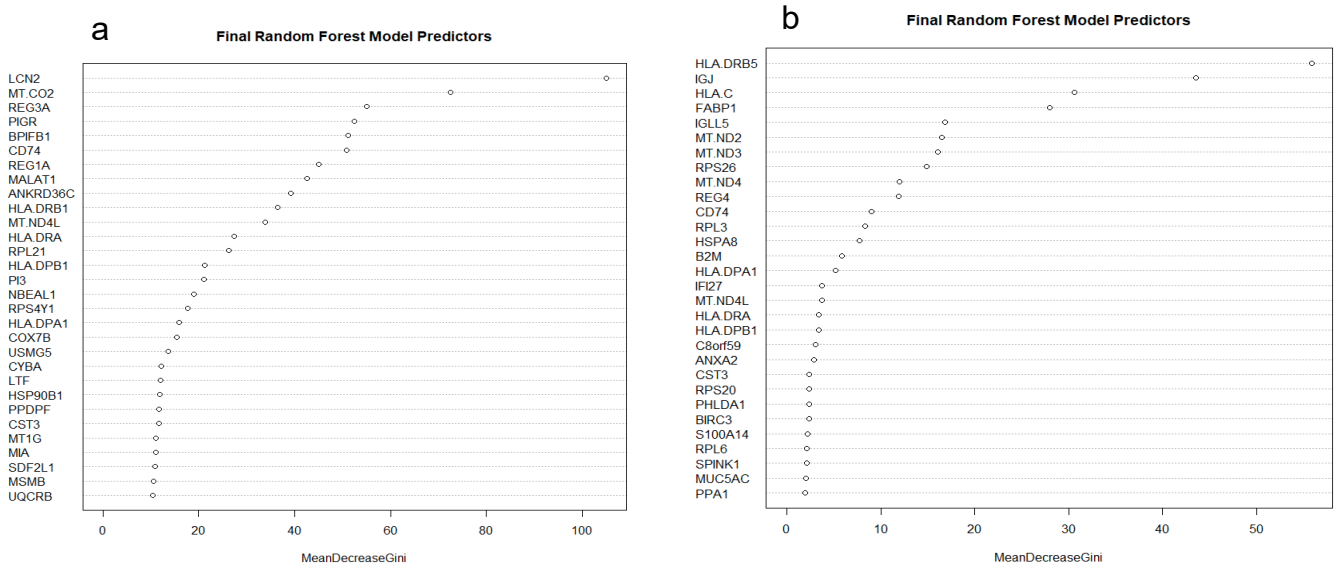

Figure 3: Important predictor for detecting *H. pylori* positivity (**a**) and severity (**b**) of infected PMCs.

followed by signal transduction. The pathways that are enriched in both the *H. pylori* infected PMC and Enterocyte cells include:

- MHC class II antigen presentation (p-value = 1.81×10^{-12}, false discovery rate (FDR) = 8.23×10^{-11})
  - Plays a role in immune protection, and reacts with urease produced by *H. pylori*;
- PD-1 signaling (p-value = 2.22×10^{-16}, FDR = 2.51×10^{-14})
  - Responsible in stopping antitumor responses in the immune system;
- Nonsense-Mediated Decay (p-value = 1.11×10^{-16}, FDR = 1.44×10^{-15})
  - Starts the destruction of mRNA with premature termination codons.

Machine-learning methods were applied to predict *H. pylori* infection using the scRNA-Seq data. Here, the gene expression data from a single cell type were used to build machine-learning models. Among the identified cell types (Table 1), enterocytes and PMCs are the most abundant in these samples, making approximately 45% of all the cells. Therefore, models based on enterocytes or PMCs can be trained from a larger number of cells to achieve better prediction accuracy. In this study, the PMCs were selected to build machine-learning models. The reason for using PMCs is for the detection of *H. pylori* infection in the early stage of gastric cancer. Enterocytes are the intestinal absorptive cells, which are found in patients with intestinal metaplasia. Prior to the intestinal metaplasia stage, enterocytes are much less abundant in gastric tissues. However, PMCs are common in gastric tissues for all patients. Therefore, PMCs-based machine-learning models could be applied to a wider range of patients.

The scaled expression data for significantly differentially expressed genes between *H. pylori* positive and negative PMCs with adjusted p-value < 0.05 were used to train and validate machine-learning models. Random forest (RF) technique is a powerful machine-learning method based on an ensemble of decision trees built from a randomly selected subset of predictors. RF models were trained to classify *H. pylori* positive and negative PMCs. The procedures for building the RF models and the 5-fold cross validation are mentioned in *Materials and Methods* section. The final RF model has an accuracy of 97% in classifying *H. pylori* positive and negative PMCs. The important genes in the model are shown in Figure 3a. The y-axis in the graphs represents genes and the x-axis represents the mean decrease Gini, which is a measure of how important a variable is in the model. The topmost important predictors, in the order of decreasing Gini, are LCN2, MT.CO2, REG3A, PIGR, BPIFB1 and CD74, which are found among the highly expressed genes in *H. pylori* positive cells. Next, using the similar machine-learning method, predictive models were built to classify the severity of *H. pylori* infection. Here only the data of *H. pylori* positive PMCs were used. Cells from IMS and IMW were considered severely and non-severely infected. The scaled expression data for significantly differentially expressed genes between the IMS and IMW with adjusted p-value < 0.05 were used to train and validate RF models. The final model also reached an accuracy of 97%. Among the important genes of this model (Figure 3b), the top ones are HLA-DRB5, IGJ, HLA-C and FABP1, which is a different set of genes from the previous model for predicting *H. pylori* infected PMCs.

Conclusion

In this project, we analyzed the scRNA-Seq datasets from patients infected by *H. pylori* and controls with Intestinal metaplasia, an early-stage transformation that may lead to gastric cancer. We used R Studio, Seurat package, Reactome and bioinformatics tools to analyze these datasets. The cell types were identified, and the number of cells were compared between the *H. pylori* positive cells and controls. Significant genes were identified between the *H. pylori* positive and negative cells. We also found several significant pathways, which could also be used to elucidate the mechanisms of *H. pylori* impact on cancer development.

H. pylori may affect the development of intestinal metaplasia or

perhaps make it more severe. Additionally, the difference in PMCs concentrations may suggest that cells in *H. pylori* positive subjects may be vulnerable to stomach acid since there are less mucous-producing cells. Another explanation could be that the reduction in PMCs may be a way to combat the *H. pylori* that hide in the epithelial layer in order to escape gastric acid.

H. pylori produces urease to protect itself from stomach acid and the urease binds to MHC class II antigens on epithelial cells when first colonizing a host. When binding, a signal will be sent out which causes an increased rate of apoptosis to epithelial cells lining the stomach. This not only will help *H. pylori* infection but may also lead to the stomach lining being more vulnerable to other factors, including cancers and diseases [18].

PD-1 is another pathway that is highly enriched in *H. pylori* positive cells. This pathway is responsible for inhibiting immune responses, and because of this plays a large part in both cancer treatment and infection. There have been multiple studies showing that inhibition of PD-1 is effective in improving immune response towards cancer, so since the *H. pylori* samples have this pathway enriched, it would put the samples more at risk of getting cancer [19].

The NMD (Nonsense-Mediated Decay) pathway is involved in making sure mRNA is of good quality. However, tumor cells can use this pathway to their advantage by destroying tumor-suppressing mRNA, and adjusting the activity of NMD to increase their rate of growth. This may cause *H. pylori* positive samples to have higher chances of getting cancer [20].

scRNA-Seq technology and machine-learning methods together could be used for early detection of gastric cancer and diagnostics of cancer progression with high accuracy.

Authors' Contribution

EL analyzed the datasets and created the machine learning models. IFT, SK and VLK proposed the topic of the project and focused the analysis and model development. EL and VLK wrote the manuscript. All authors participated in the final corrections of the manuscript.

Acknowledgements

We would like to thank Mrs. Lani Keller for giving many useful suggestions for this project. We would also like to thank Dr. Inhan Lee for advice in next generation sequencing data analysis, RNA-Seq, and R studio.

References

1. Rawla P, Barsouk A. Epidemiology of gastric cancer. Global trends, risk factors and prevention. Prz Gastroenterol. 2019; 14: 26.
2. Polk DB, Peek RM Jr. Helicobacter pylori: Gastric cancer and beyond. Nat Rev Cancer. 2010; 10: 403-414.
3. Watts G. Nobel prize is awarded to doctors who discovered H. pylori. BMJ. 2005; 331: 795.
4. Hooi JKY, Lai WY, Ng WK, et al. Global prevalence of Helicobacter pylori infection: Systematic review and meta-analysis. Gastroenterology. 2017; 153: 420-429.
5. Backert S, Clyne M, Tegtmeyer N. Molecular mechanisms of gastric epithelial cell adhesion and injection of CagA by Helicobacter pylori. Cell Commun Signal. 2011; 9: 1-11.
6. Hatakeyama M. Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer. Proc Jpn Acad Ser B Phys Biol Sci. 2017; 93: 196-219.
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57-63.
8. Zheng GXY, Terry GM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8: 1-12.
9. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161: 1202-1214.
10. Butler A, Hoffman, P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; 36: 411-420.
11. Zhang P, Yang M, Zhang Y, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep. 2019; 27: 1934-1947.
12. Dalgaard P. Introductory Statistics with R. 2nd ed. Springer-Verlag New York Inc. New York, 2008; 99-100.
13. Bretz F, Hothorn T, Westfall P. Multiple Comparisons using R. 1st ed. Chapman and Hall/CRC: New York. 2010; 32.
14. Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018; 46: D649-D655.
15. Griss J, Viteri G, Sidiropoulos K, et al. ReactomeGSA - Efficient multi-omics comparative pathway analysis. Mol Cell Proteomics. 2020; 19: 2115-2125.
16. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002; 23:18-22.
17. Kuhn M. Building predictive models in R using the Caret package. J Stat Softw. 2008; 28: 1-26.
18. Fan X, Gunasena H, Cheng Z, et al. Helicobacter pylori urease binds to class II MHC on gastric epithelial cells and induces their apoptosis. J Immunol. 2000; 165: 1918-1924.
19. Han Y, Liu D, Li L. PD-1/PD-L1 pathway: Current researches in cancer. Am J Cancer Res. 2020; 10: 727-742.
20. Popp MW, Maquat LE. Nonsense-mediated mRNA decay and cancer. Curr Opin Genet Dev. 2018; 48: 44-50.

Supplementary Materials

Acronyms and Abbreviation: GMC, gland mucous cell; PMC, pit mucous cell

Table S1

Sample metadata.

Sample	Subject	*H. pylori* infection
IMW1	P5	y
IMW2	P6	n
IMS1	P7	y
IMS2	P7	y
IMS3	P8	n
IMS4	P8	n

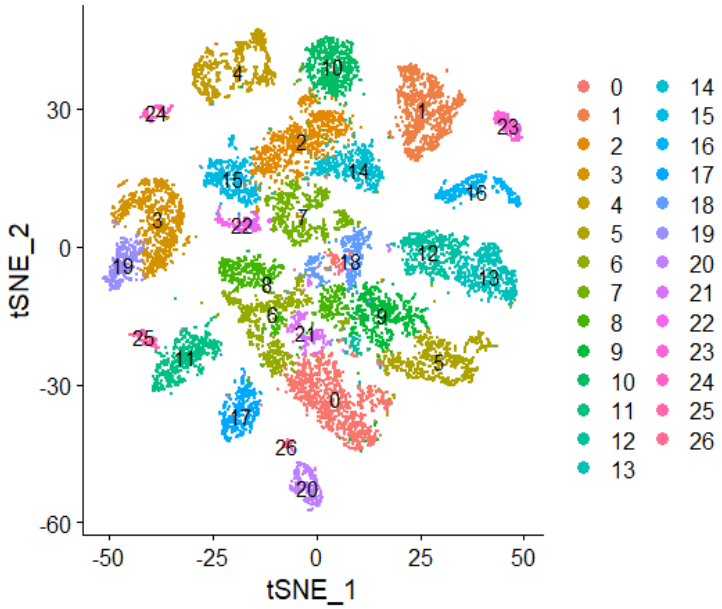

Figure S1. Cell clusters. The plot was made using the TSNEPlot function in the Seurat package (Ref S1, #10 in the text).

Table S2

Marker genes for different cell types according to ref S2 (#11in the text).

Cell group	Cell Type	Marker Gene(s)
Mucous & secretory	Pit mucous cell (PMC)	MUC5AC
Mucous & secretory	Gland mucous cell (GMC)	MUC6
Mucous & secretory	Parietal cell	ATP4A, ATP4B, GIF
Mucous & secretory	Chief Cell	PGA4, PGA3, LIPF
Endocrine	G cell	GAST
Endocrine	X cell	GHRL
Endocrine	D cell	SST
Immune cells	T Cell	CD2, CD3D,CD3E,CD3G
Immune cells	B Cell	CD79A,CD19
Immune cells	Mast Cell	TPSAB1,TPSB2
Immune cells	Macrophage	CD14, CD163, CD68, CSF1R
Stromal cells	Fibroblasts	FAP, PDPN,COL1A2,DCN, COL3A1, COL6A1
Stromal cells	Endothelial cells	PECAM1,VWF,ENG,MCAM
Stem Cells	Stem cell	OLFM4,SOX2,LGR5,CCKBR
Myocytes	Smooth muscle cell	ACTA2,ACTN2,MYL2,MYH2
Proliferative cell	Proliferative Cell	MKI67,BIRC5,CDK1
Intestinal cells	Goblet cell	TFF3,SPINK4,MUC2
Intestinal cells	Enteroendocrine cell	CHGA,CHGB,TAC1,TPH1,NEUROG3
Intestinal cells	Enterocytes	FABP1,CA1,VIL1

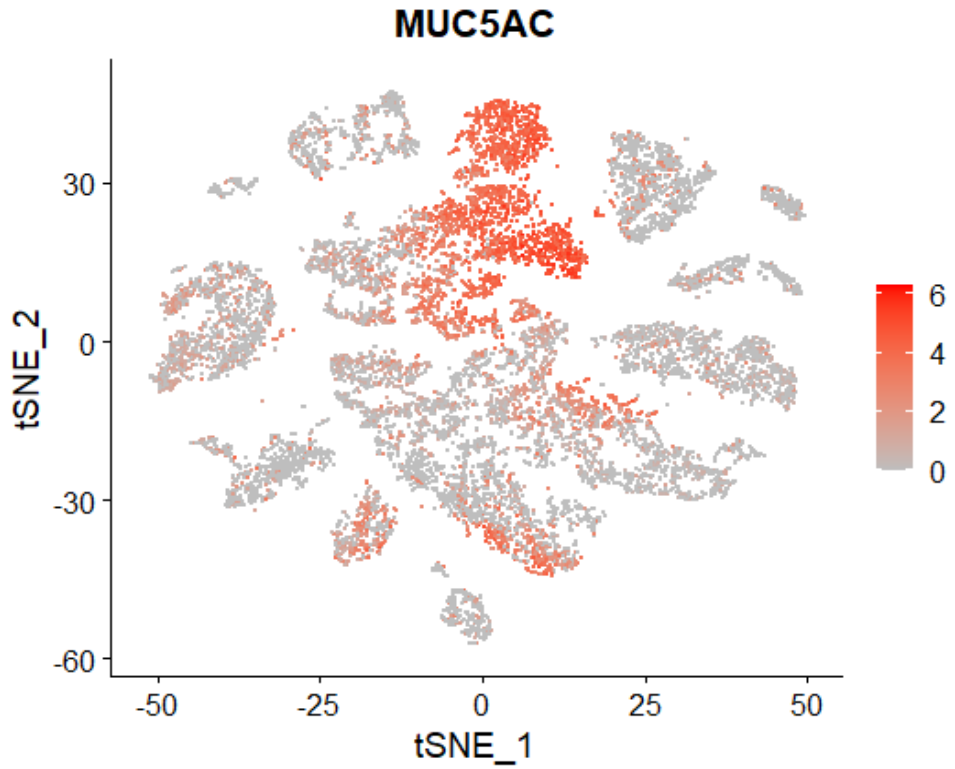

Figure S2. Cell clusters colored by gene expression of PMC specific marker MUC5AC.

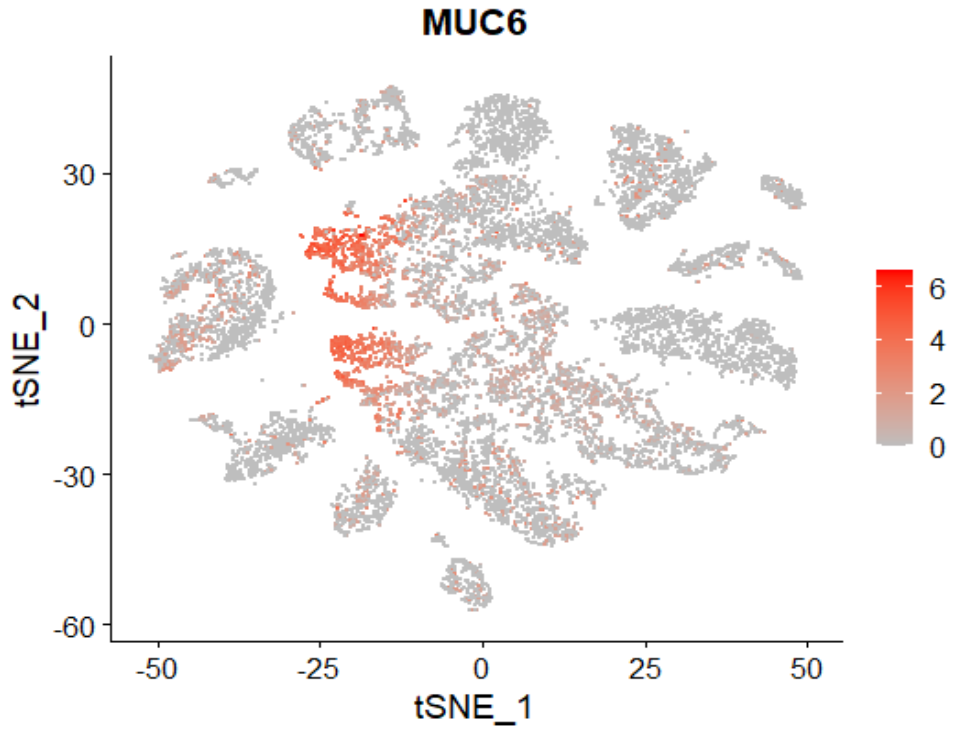

Figure S3. Cell clusters colored by gene expression of GMC specific marker MUC6.

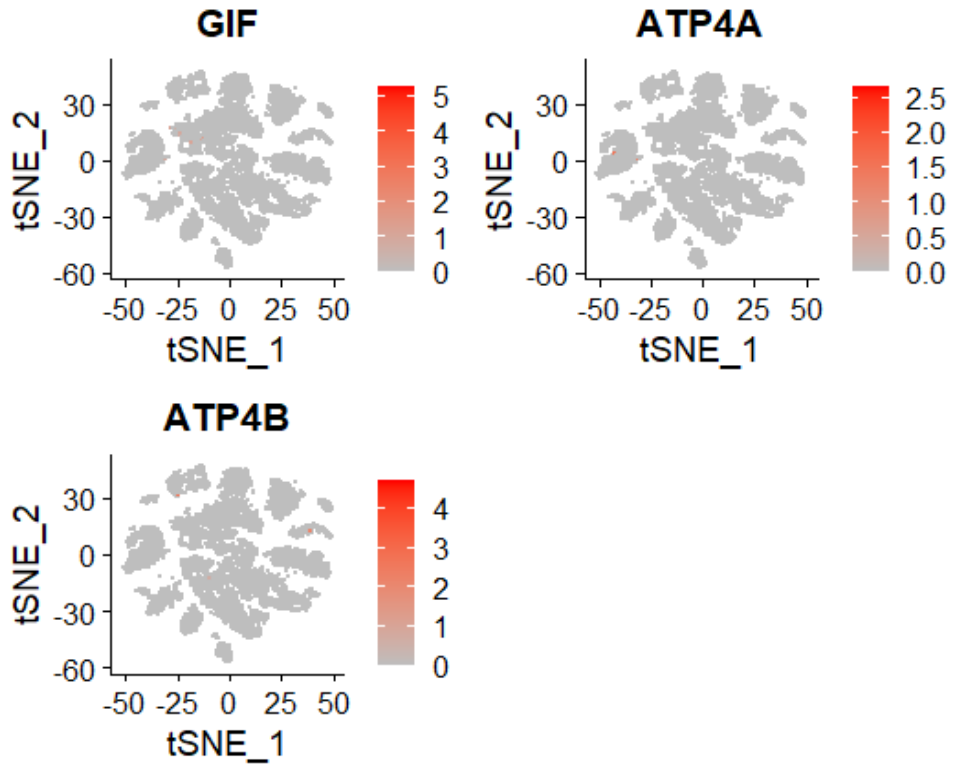

Figure S4. Cell clusters colored by gene expression of Parietal cell specific markers GIF, ATP4A and ATP4B. Parietal cells were not detected.

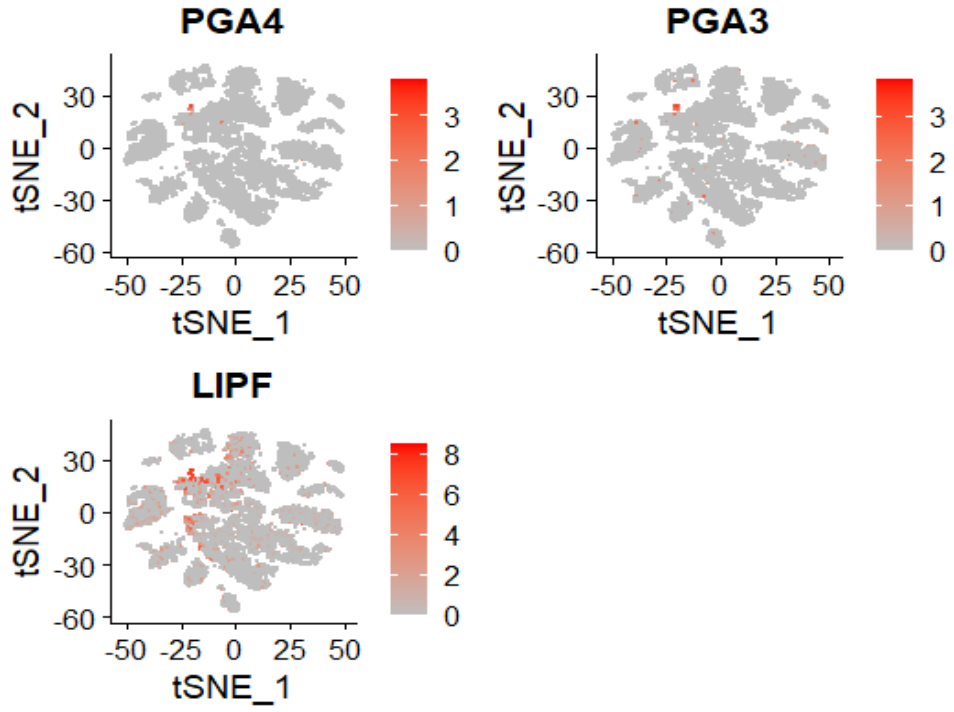

Figure S5. Cell clusters colored by gene expression of chief cell specific markers PGA4, PGA3 and LIPF.

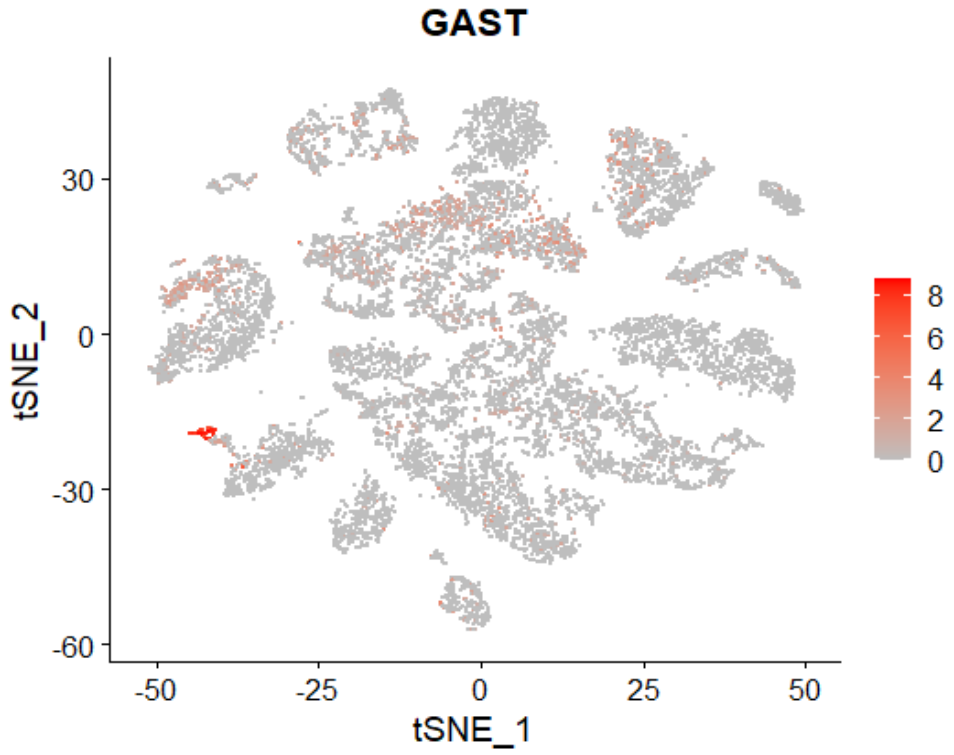

Figure S6. Cell clusters colored by gene expression of G cell specific marker GAST.

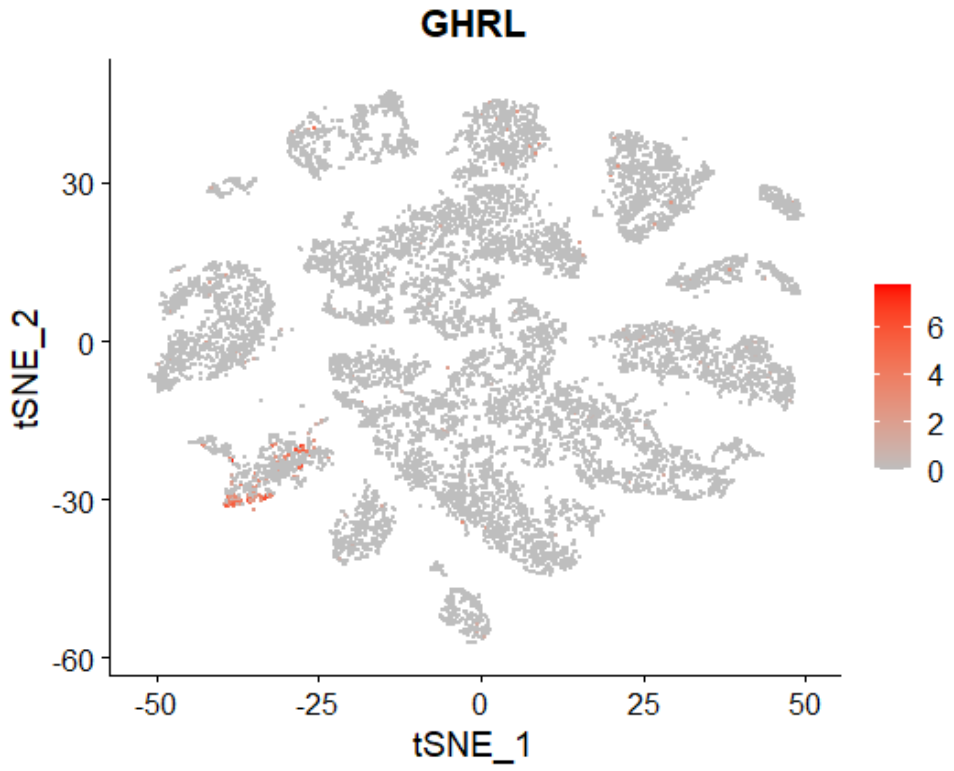

Figure S7. Cell clusters colored by gene expression of X cell specific marker GHRL.

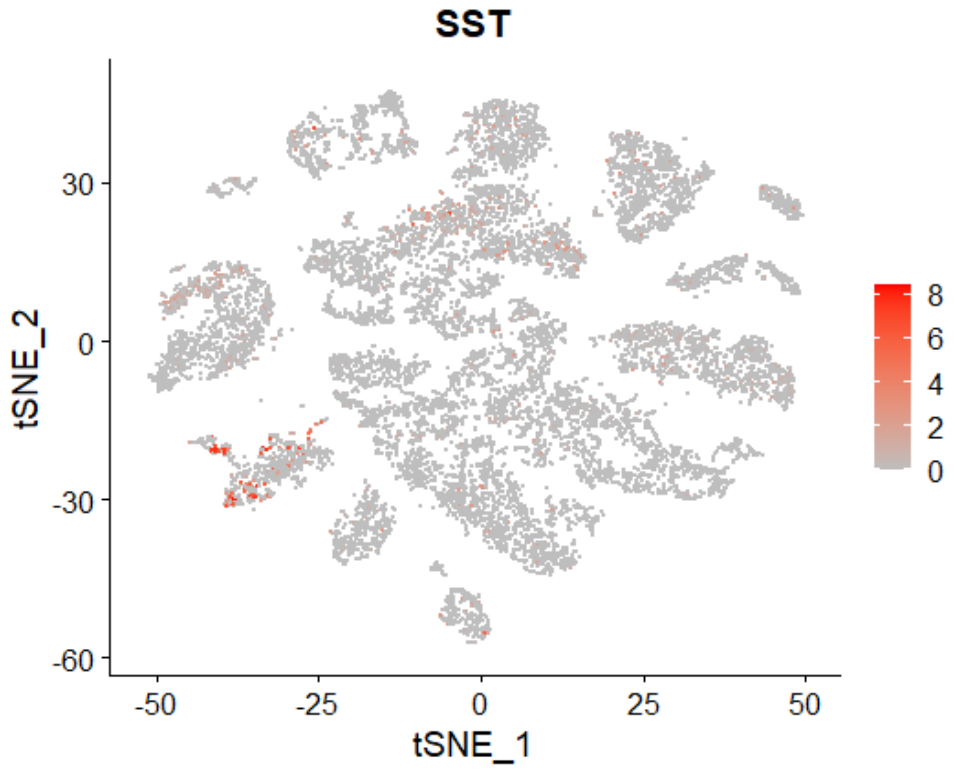

Figure S8. Cell clusters colored by gene expression of D cell specific marker SST.

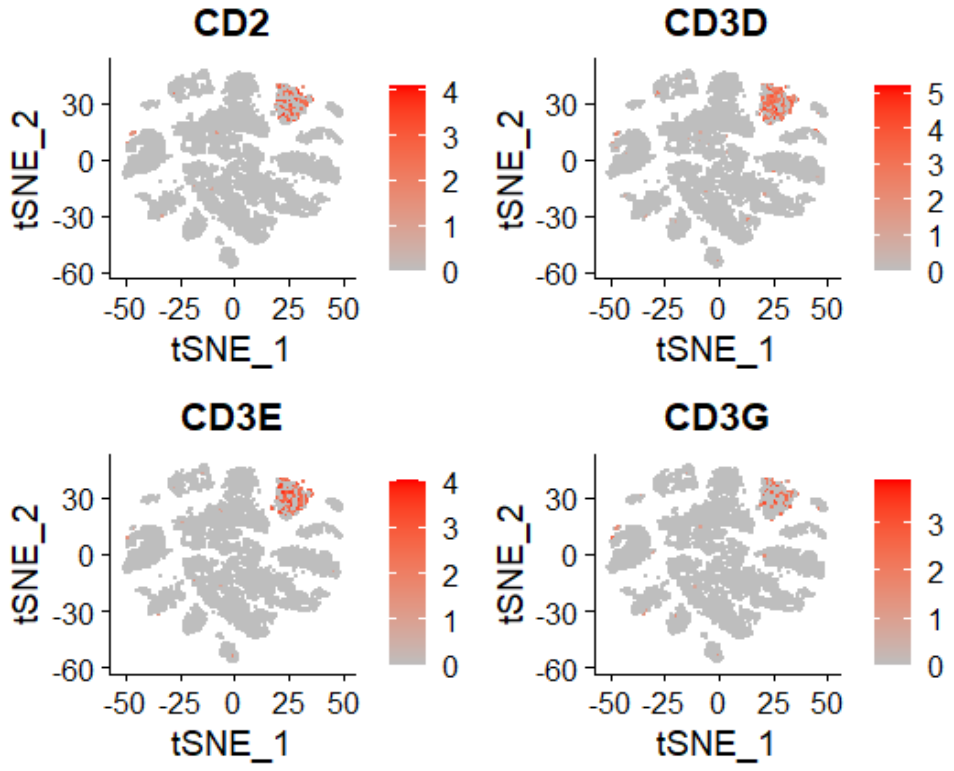

Figure S9. Cell clusters colored by gene expression of T cell specific markers CD2, CD3D, CD3E and CD3G.

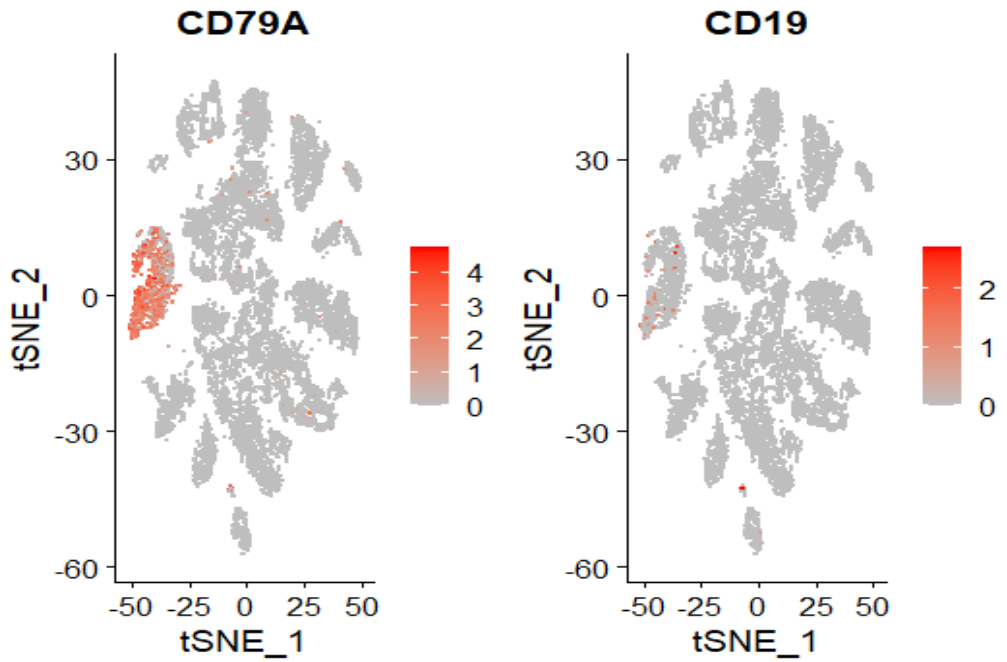

Figure S10. Cell clusters colored by gene expression of B cell specific markers CD79A and CD19.

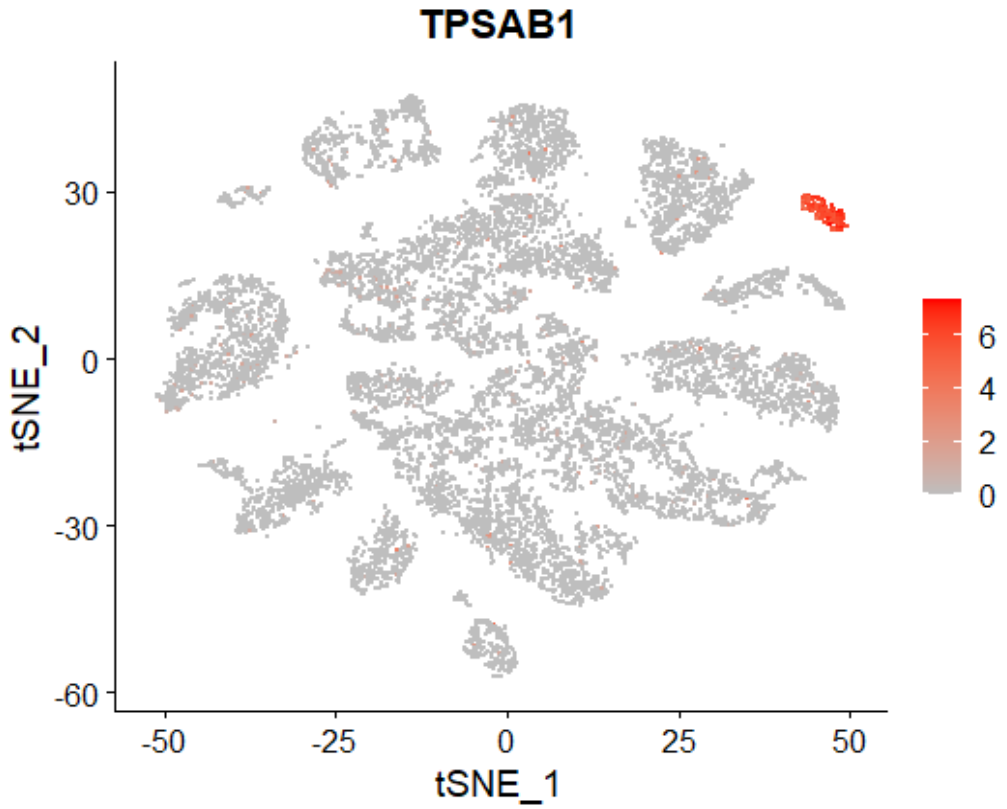

Figure S11. Cell clusters colored by gene expression of Mast cell specific marker TPSAB1.

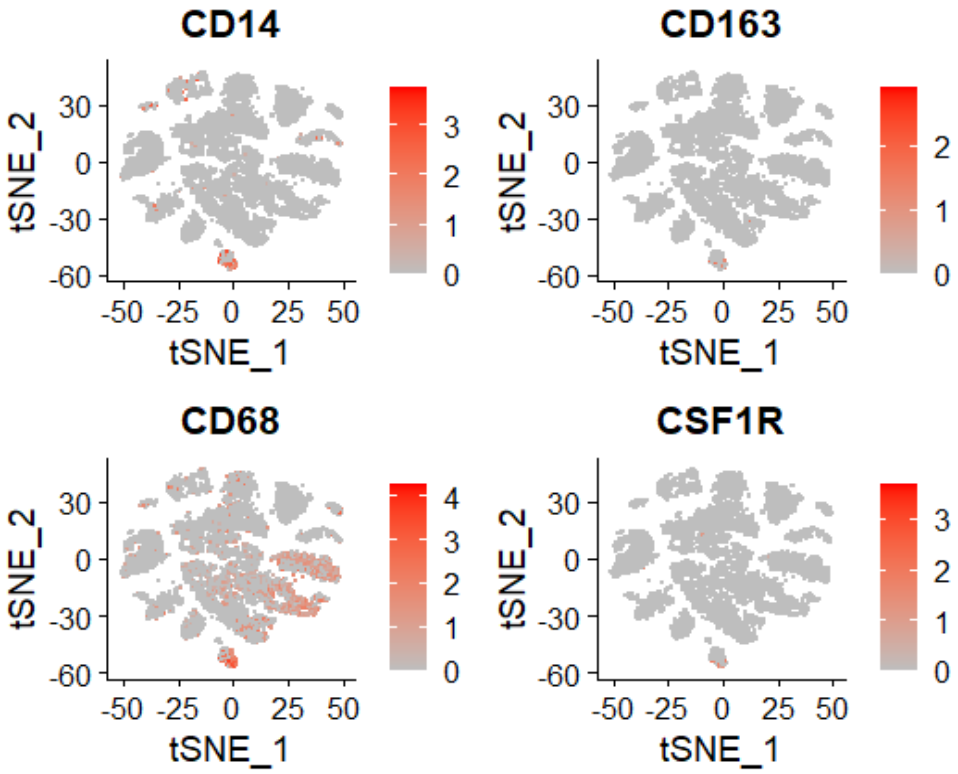

Figure S12. Cell clusters colored by gene expression of Macrophage cell specific markers CD14, CD163, CD68 and CSF1R.

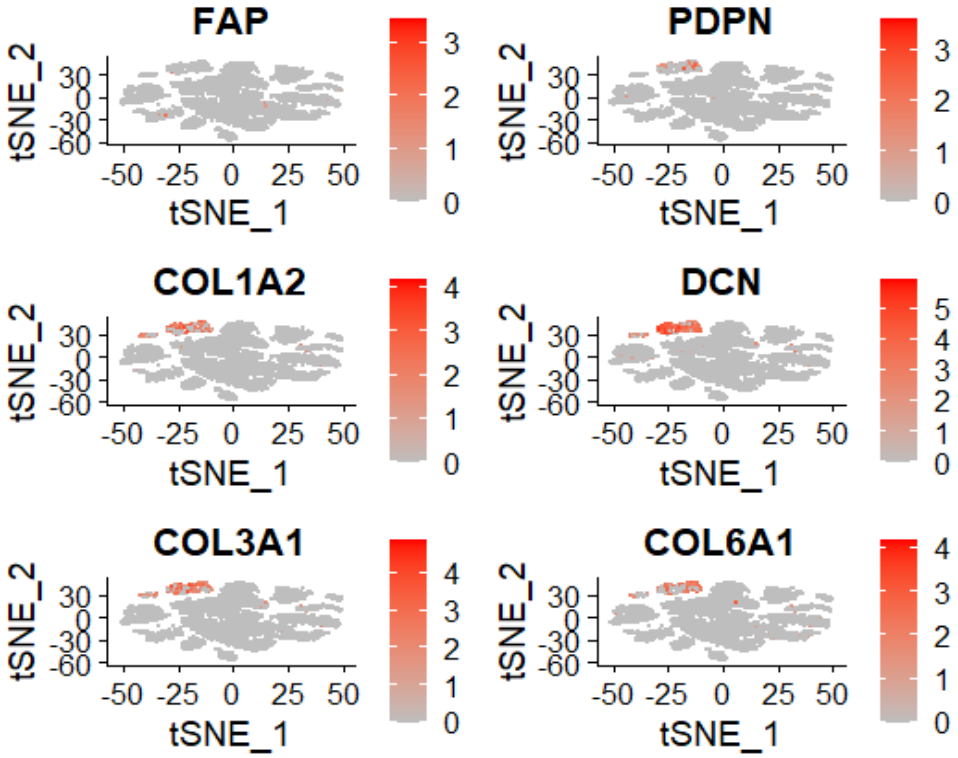

Figure S13. Cell clusters colored by gene expression of Fibroblasts cell specific markers FAP, PDPN, COL1A2, DCN, COL3A1 and COL6A1.

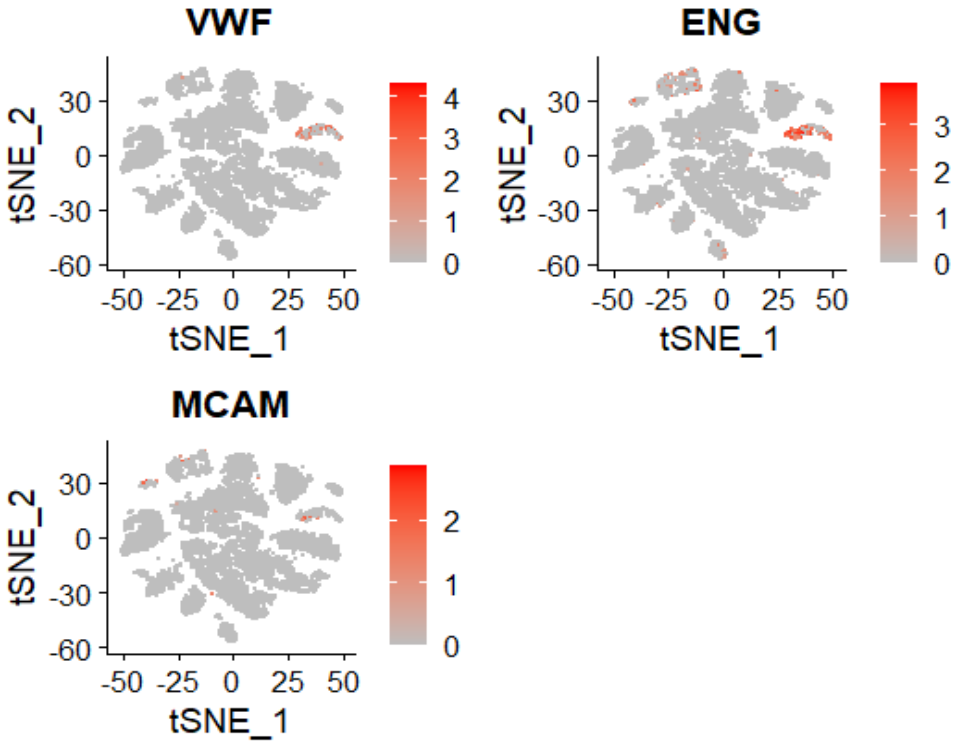

Figure S14. Cell clusters colored by gene expression of Endothelial cell specific markers VWF, ENG and MCAM.

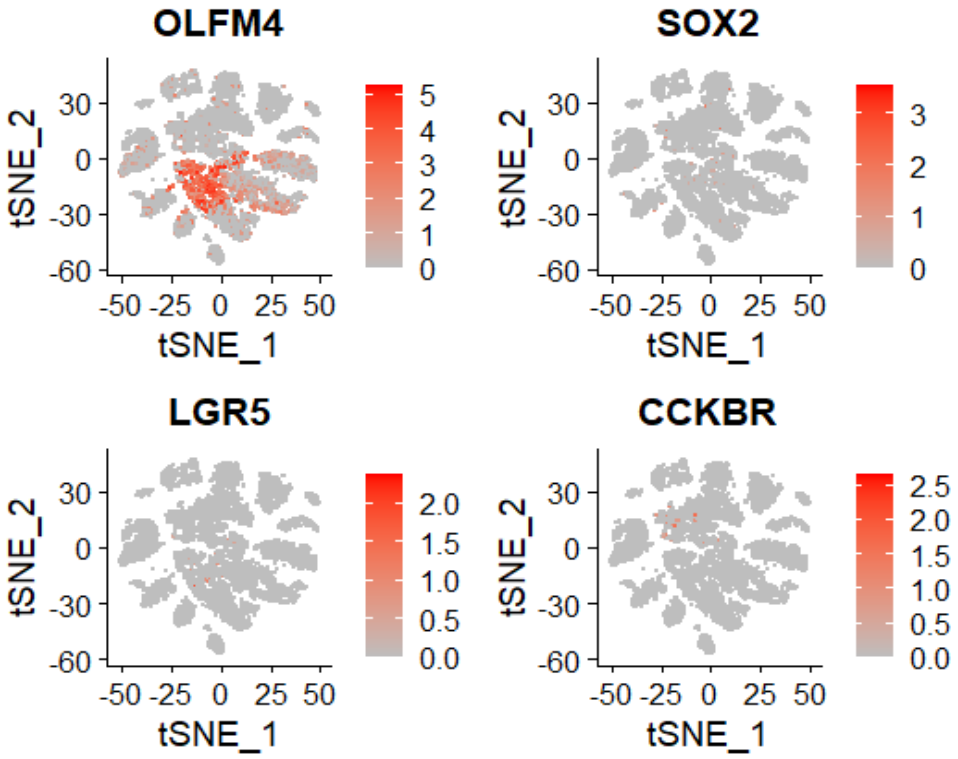

Figure S15. Cell clusters colored by gene expression of Stem cell specific markers OLFM4, SOX2, LGR5 and CCKBR.

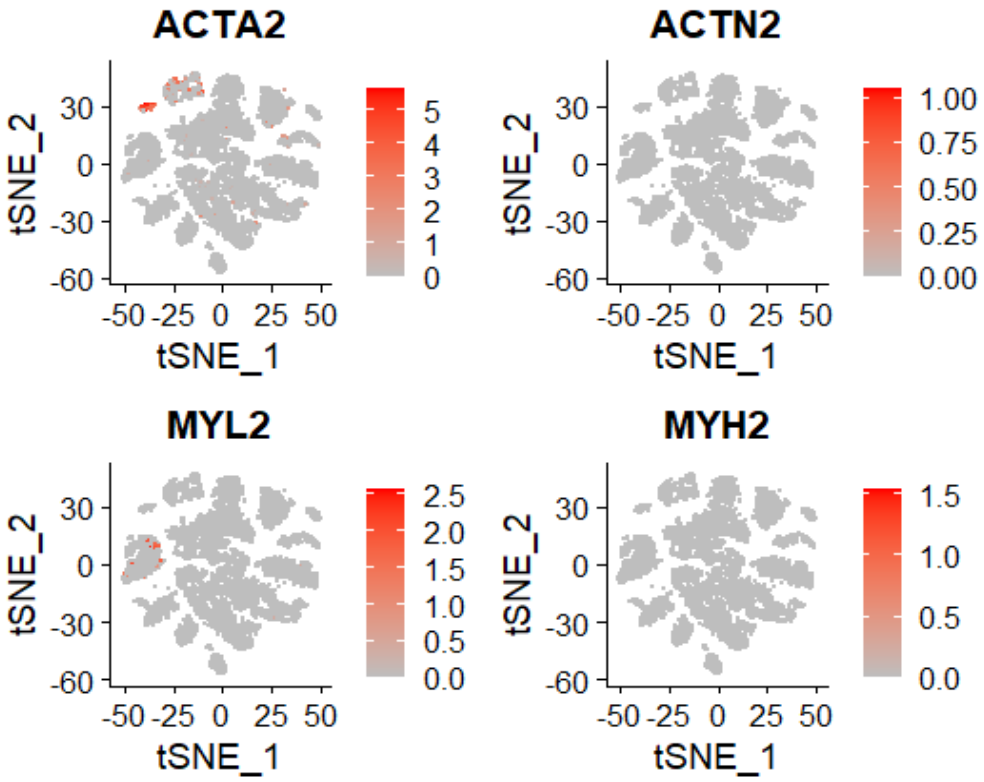

Figure S16. Cell clusters colored by gene expression of SMC cell specific markers ACTA2, ACTN2, MYL2 and MYH2.

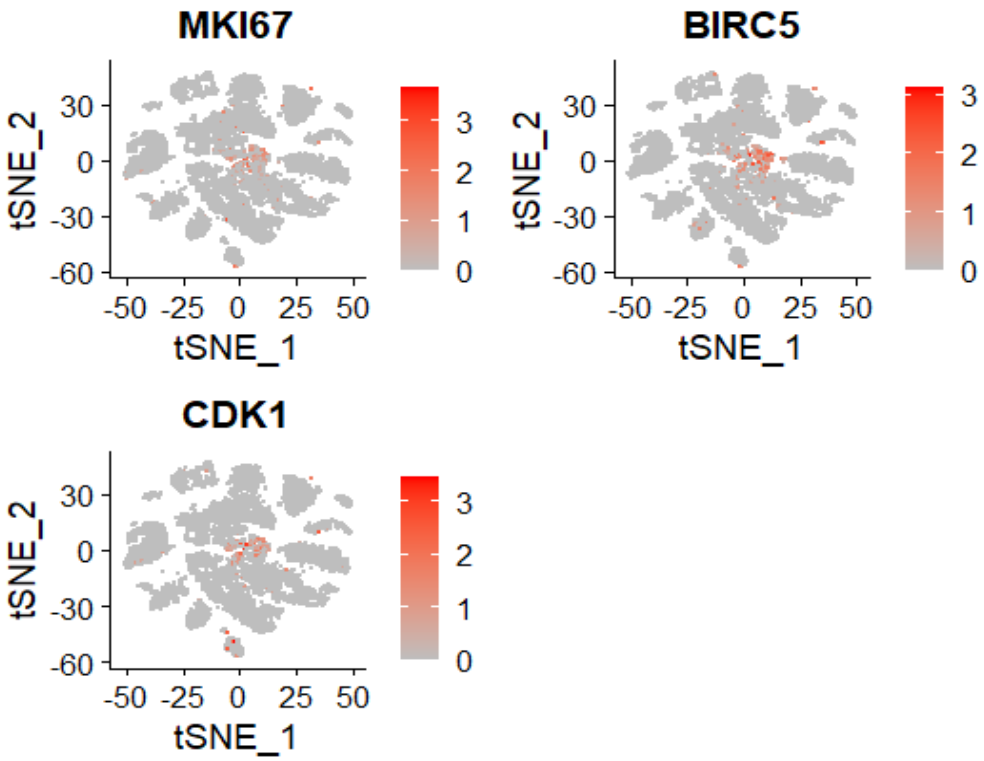

Figure S17. Cell clusters colored by gene expression of Proliferative cell specific markers MKI67, BIRC5 and CDK1.

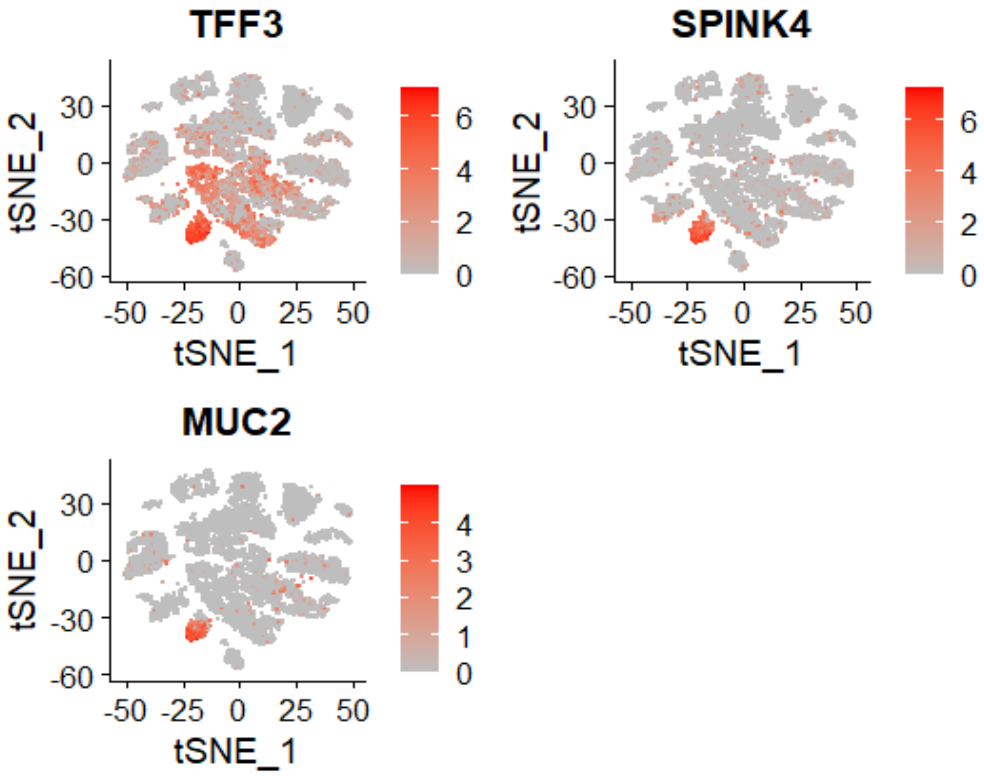

Figure 18. Cell clusters colored by gene expression of Goblet cell specific markers TFF3, SPINK4 and MUC2.

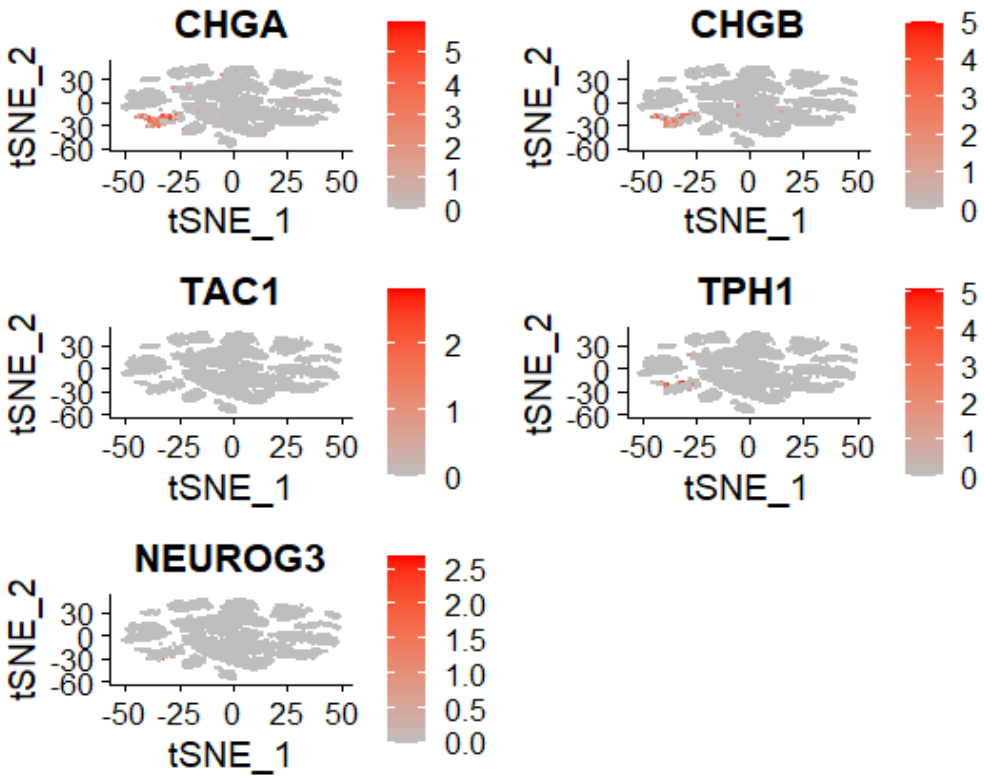

Figure S19. Cell clusters colored by gene expression of Enteroendocrine cell specific markers CHGA, CHGB, TAC1, TPH1 and NEUROG3.

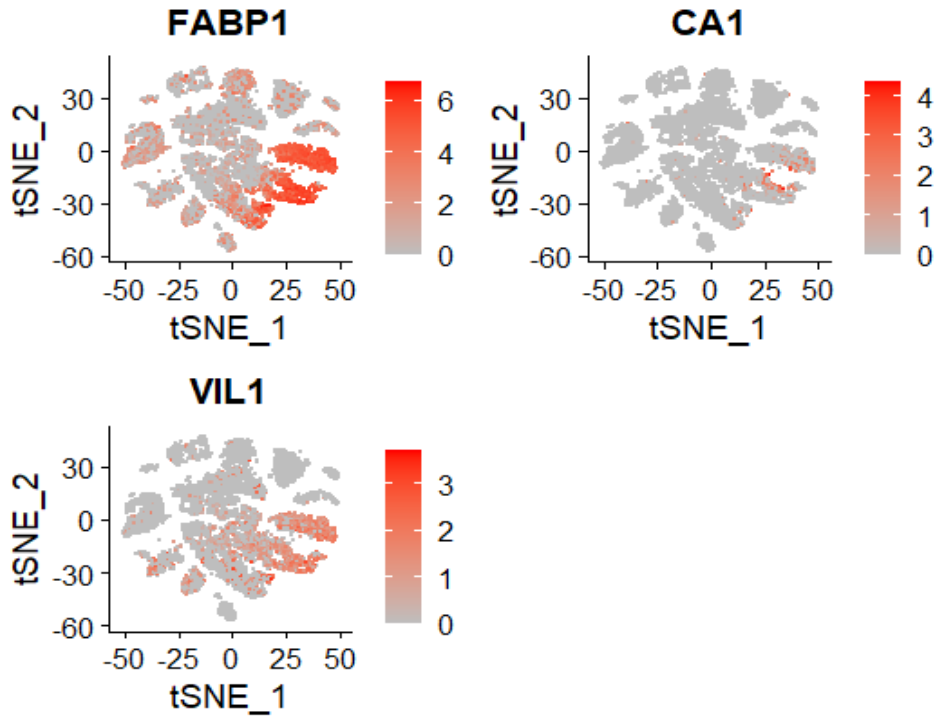

Figure S20. Cell clusters colored by gene expression of Enterocytes cell specific markers FABP1, CA1 and VIL1.

Table S3
Cell clusters and corresponding them cell types.

Cluster	Cell type
0	Enterocyte
1	T Cell
2	PMC
3	B Cell
4	Fibroblast
5	Enterocyte
6	Stem Cell
7	PMC
8	GMC/Stem Cell
9	Enterocyte
10	PMC
11	Indeterminate
12	Enterocyte
13	Enterocyte
14	PMC
15	GMC
16	Endothelial
17	Goblet
18	Proliferative/Stem
19	B Cell
20	Macrophage
21	Stem Cell
22	GMC
23	Mast cell
24	SMC
25	G cell
26	B Cell

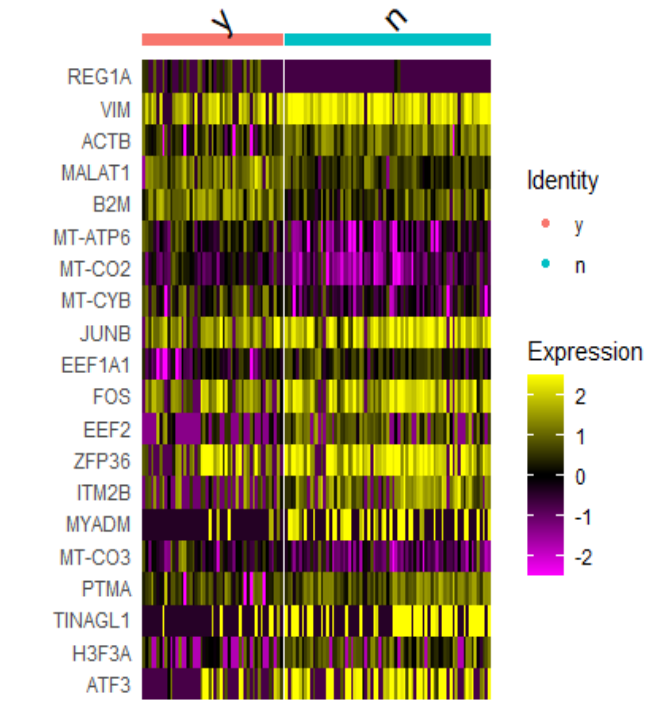

Figure S21. Gene expression heatmap of top 20 most differentially expressed genes in SMCs. the cells at the left side under the red bar denoted by letter "y" and the cells at the right side under the green bar denoted by letter "n" are *H. pylori* positive and negative cells respectively.

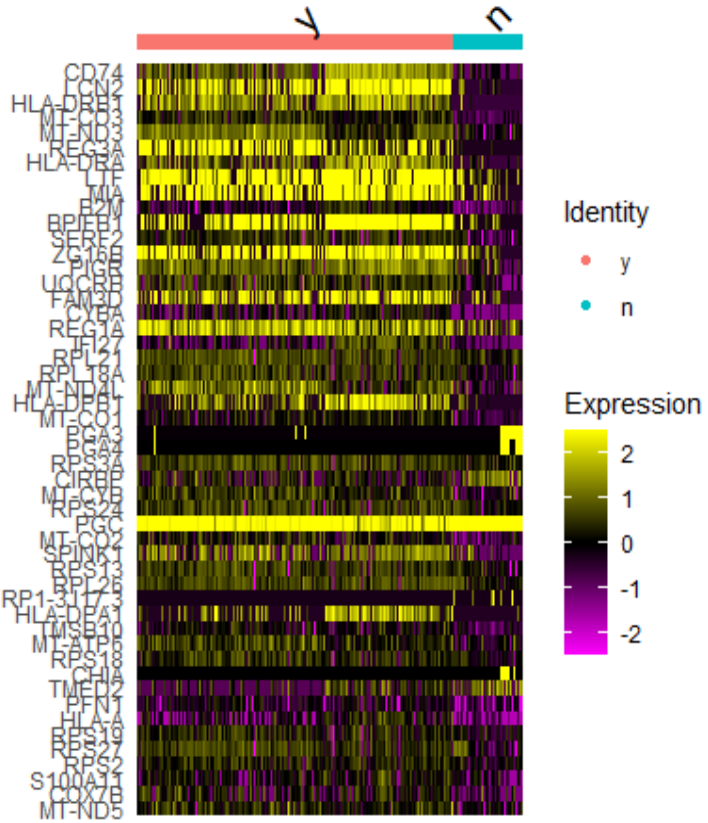

Figure S22. Gene expression heatmap of top 50 most differentially expressed genes in GMCs. The cells at the left side under the red bar denoted by letter "y" and the cells at the right side under the green bar denoted by letter "n" are *H. pylori* positive and negative cells respectively.

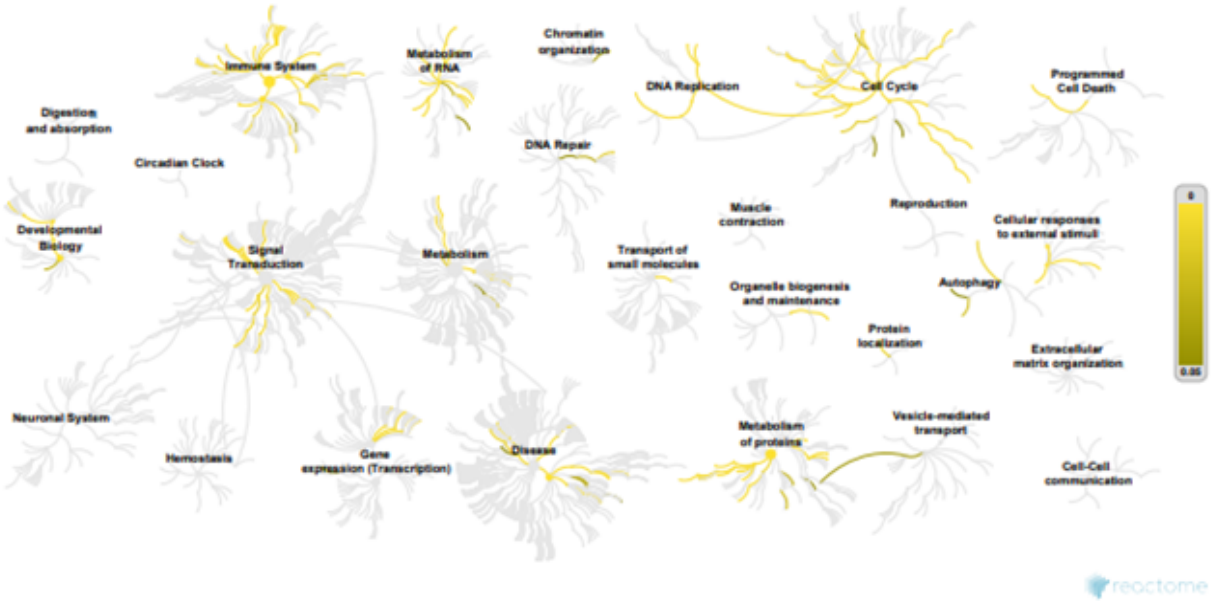

Figure S23. Enriched pathways in *H. pylori* infected PMCs.

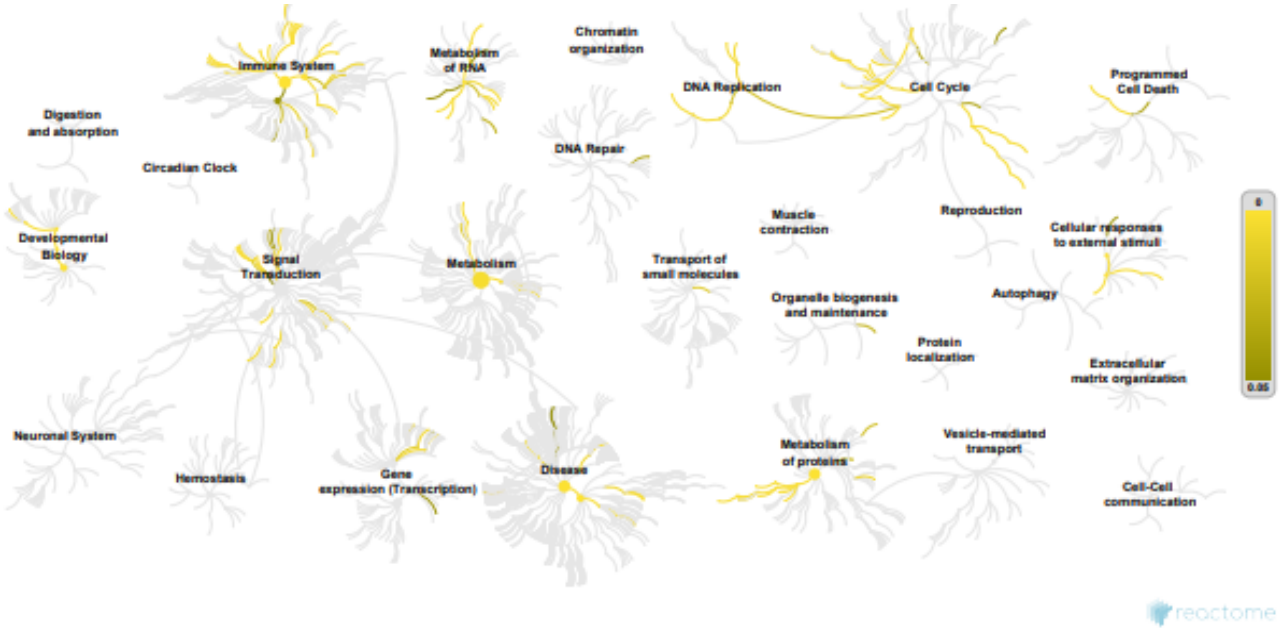

Figure S24. Enriched pathways in *H. pylori* infected enterocytes.

R code

R Code for analysis could be found at this web site: https://pastebin.com/Pf8b7USt

Supplementary References

1. Hoffman P. Dimensional reduction plot. Satija Lab and Collaborators. https://satijalab.org/seurat/reference/dimplot. (Ref. [10] in the main text.)
2. Zhang P, Yang M, Zhang Y, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep. 2019; 27(6):1934–1947. https://doi.org/10.1016/j.celrep.2019.04.052 (Ref. [11] in the main text.)