Analysis of gastric adenocarcinoma microenvironment with single-cell RNA-seq

Adam Cicherski

Abstract

The main goal of that project was to perform single-cell RNA-seq analysis of gastric cancer samples from a primary tumor, non-tumoral tissue, and different organs metastasis with the use of common sc-RNA bioinformatics tools: Seurat for cell clustering, Monocle3 for pseudo-time trajectories inference, inferCNV for large scale copy number variation analysis and CellChat for exploration of signaling pathways.

1 Introduction

Gastric cancer is the fifth-leading type of cancer and the third-leading cause of death from cancer, making up 7% of cases and 9% of death of diagnosis. It is also known to have a generally poor prognosis due to resistance to chemotherapy. Still, surgical treatment is the most common form of management, but novel kinds of treatment such as immunotherapy and gene therapy are intensively studied. New generation sequencing methods and especially single-cell sequencing could lead us to a better understanding of the complex biological processes that underlie the development and growth of gastric cancer and also help us in finding potential targets for new drugs.

2 Materials and Methods

2.1 Data set

scRNA-seq analysis was performed on 10 samples from 6 patients downloaded from Gene Expression Omnibus (GEO accession: GSE163558). Samples in-

clude one non-tumoural tissue sample (NT1), three primary tumor samples (PT1-3), and six metastasic samples (two lymph node metastasis(LN1-2), two liver metastasis(Li1-2), one ovary metastasis(O1) and one peritoneal metastasis(P1)). Patients had not been administrated therapy.

2.2 Seurat Clustering

All samples were fetched to R by the Read10XData function and then Seurat's [1] objects for each sample were created with default parameters. Next, the percent of mitochondrial genes in each sample was calculated. Cells which does not satisfy the criteria: more than 200 genes expressed, less than 5000 genes expressed, and less than 20 percent of mitochondrial genes were excluded from further analysis. In the next step, 1295 cells (number of cells in the smallest sample) were randomly chosen from each sample. The data normalization, cell cycle scoring, and FindVariableFeature method were applied to all samples. Next, the integration features were selected, and based on that integration anchors were found. After integration with the IntegrateData method, data were scaled and principle component analysis was performed. After inspection of the ElbowPlot, 20 PC were selected for TSNE dimensionality reduction and neighbors graph construction. Finally, cells were clustered with the Louvain algorithm with a resolution of 0.8. To identify clusters method FindAllMarkers was used with parameter only.pos=TRUE.

2.3 Malignancy score

Epithelial cells were subsetted from the integrated Seurat object. FindVariableFeature method from Seurat was again applied to the RNA assay of that subset, and then ScaleData and RunPCA methods were used. The dimension of the data set was reduced with UMAP based on 20 PC. The malignancy score of epithelial cells was calculated with the Seurat AddModuleScore method, based on the top fifty differentially expressed genes in tumor cells from bulk RNA-seq of gastric cancer. In the same way, the non-malignant score was calculated based on normal tissue DEGs. Next, the difference between malignant and non-malignant scores was calculated, and based on the growth curve, -0.02 was chosen as a point of discrimination between malignant and non-malignant cells. The proportion of malignant epithelial cells in normal tissue, primary tumors, and metastasis was calculated.

2.4 Pseudo time trajectories inferences

Information from the Seurat object of epithelial cells was fetched into Monocle3 [2] cell_data_set object. Monocle method learn_graph was used, and then cells were ordered in pseudo time with a root node chosen based on the proportion of non-malignant cells in the node's neighborhood. Finally, the graph_test method was applied to get genes that are expressed as a function of pseudo-time progression.

2.5 Copy Number Variation Inferency

To obtain information about large structural rearrangements of cancer cells, R package InferCNV was used. Data was fetched from Seurat's object subsetted to epithelial cells (used as observation) and B cells (used as control) and annotated due to the position of the genes on chromosomes with the AnnoProbe package. InferCNV::run method was used, with the application of noise filter, predictions based on hidden Markov model, and validation with Bayesian Network Latent Mixture Model.

2.6 Cell-Cell Comunication

To explore patterns of communication between cells in the gastric tumor environment I used the CellChat R package [3]. First, I have created a CellChat object based on previous Seurat data. Then Cellchat database was limited to "human" and "secreted signaling". Over-expressed genes and over-expressed interactions were identified and then Communication probability was calculated and filtered with a criterium of a minimum of ten cells. Then communication probability pathway and aggregated net were calculated. The signaling role of each cell line in each pathway and the contribution of each ligand-receptor pair were calculated. Five incoming and five outgoing general communication patterns were recognized. Finally, pathways were clustered two times, based on their functional similarity, and based on structural similarity

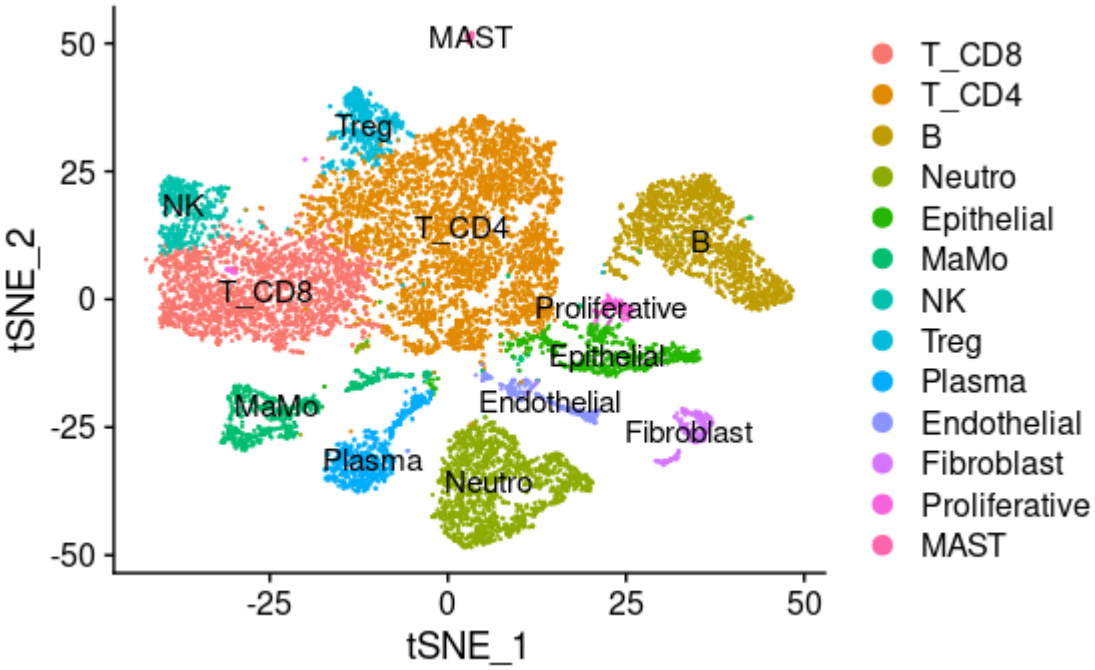

Figure 1: Unsupervised clustering of integrated samples

3 Results

3.1 Cell lineages

After unsupervised cell clustering main cell lineages were identified based on canonical markers obtained from Jiang et al. article and PnaglaoDB [4]. These lineages include: 6737 T-cells (markers: CD2, CD3D, CD3D ILR7), 1600 B-cells (MS4A1, CD79A), 613 Plasma Cells(JCHAGIN, IGHG1, SSR4), 677 Epithelial cells (EPCAM, KRT18, KRT8), 251 Endothelial cells (PECAM1, VWF), 236 Fibroblast cells (CLO1A2,COL6A2), 153 Proliferative cells(MKI67, STMN1), 1449 Neutrophils (CSF3R), 707 Myeloid mononuclear phagocytose system cells(CD68,CTSB, labeled as MaMo on all plots), 479 NK cells (KLRD1, KLRF1, GNLY) and 48 Mast cells(KIT, TPSB2, TPSAB1).

T cells were divided into three main clusters: 2154 T CD8+ cells (CD8A, CD8B), 4112 T CD4+ cells(CD4, CCR7, LEF1), 471 T regulatory cells(FOXP3, IL2RA) The proportion of each cluster differs among samples, showing the tu-

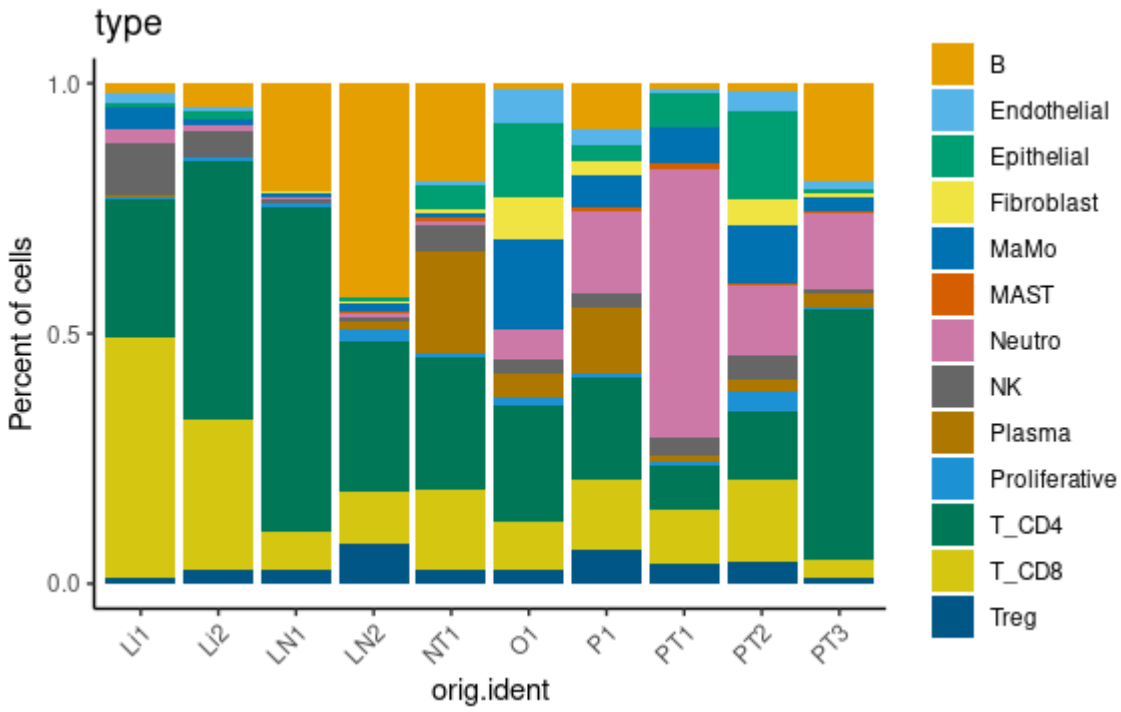

Figure 2: Proportion of different cells lineages among samples

mors heterogeneity of gastric cancer.

The cell cycle phase affection on clustering was investigated, and no signs of impact were found, so like in the original paper I decided to not regress out cell cycle scores from data. The proportion of malignant epithelial cells also differs among samples. When PT1 and PT2 contain almost only malignant cells, half of the cells from PT3 were classified as non-tumoral. Similarly, the liver metastasis samples also contain a higher proportion of non-tumor cells than other metastasis. It is worth noting, that the NT1 sample contains about 19 percent of tumor cells, what's consistent with Jiang et al. results.

3.2 Trajectory

Pseudo-time trajectory inferred with the monocle3 is shown in Fig.5. 10 genes with the most significant q-vale, which are expressed as a function of pseudo time are: JUN, ILR7, NEAT1, PLCG2, MKI67, S100A6, S100A10, TOP2A, FTH1

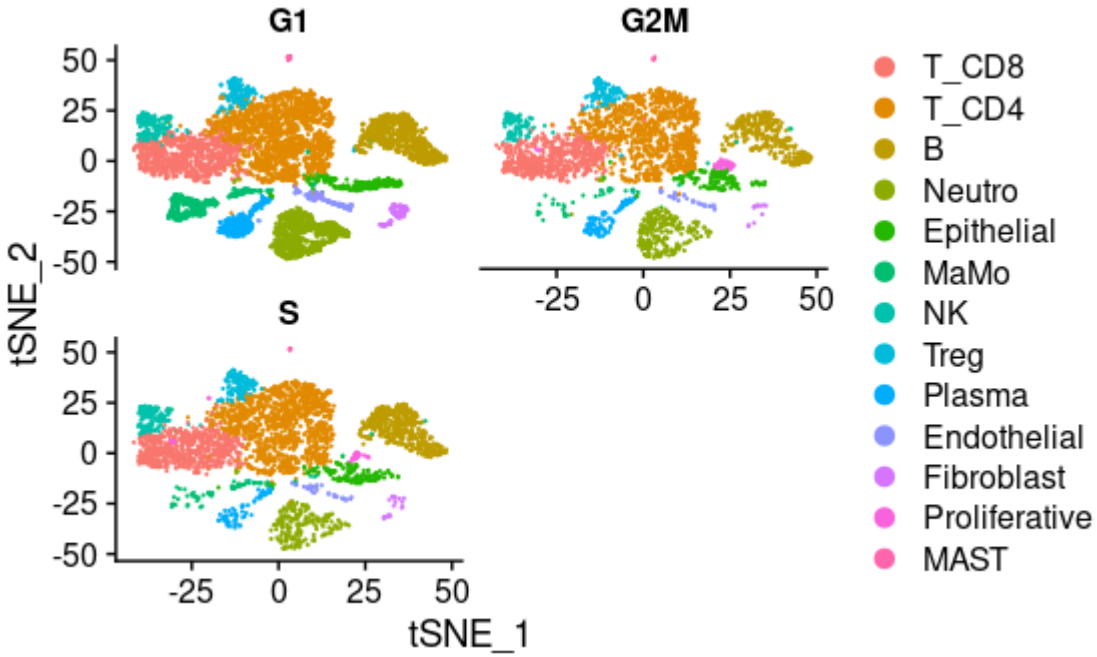

Figure 3: data split by cell cycle phases

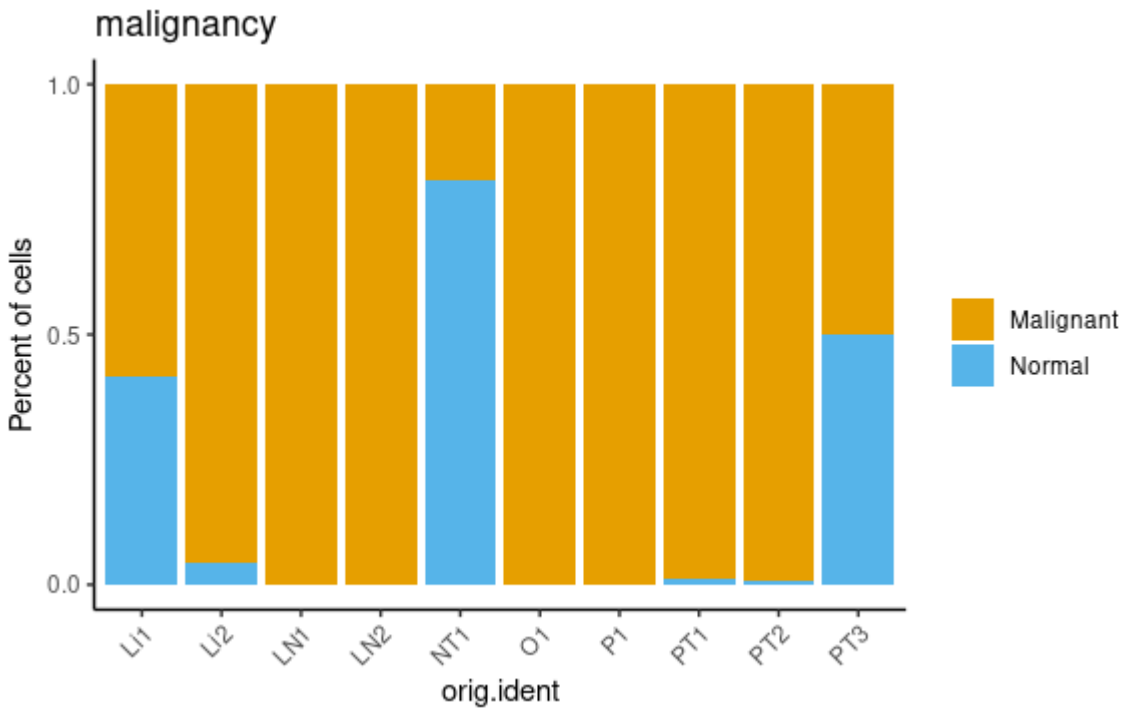

Figure 4: proportion of malignant epithelial cells in each sample

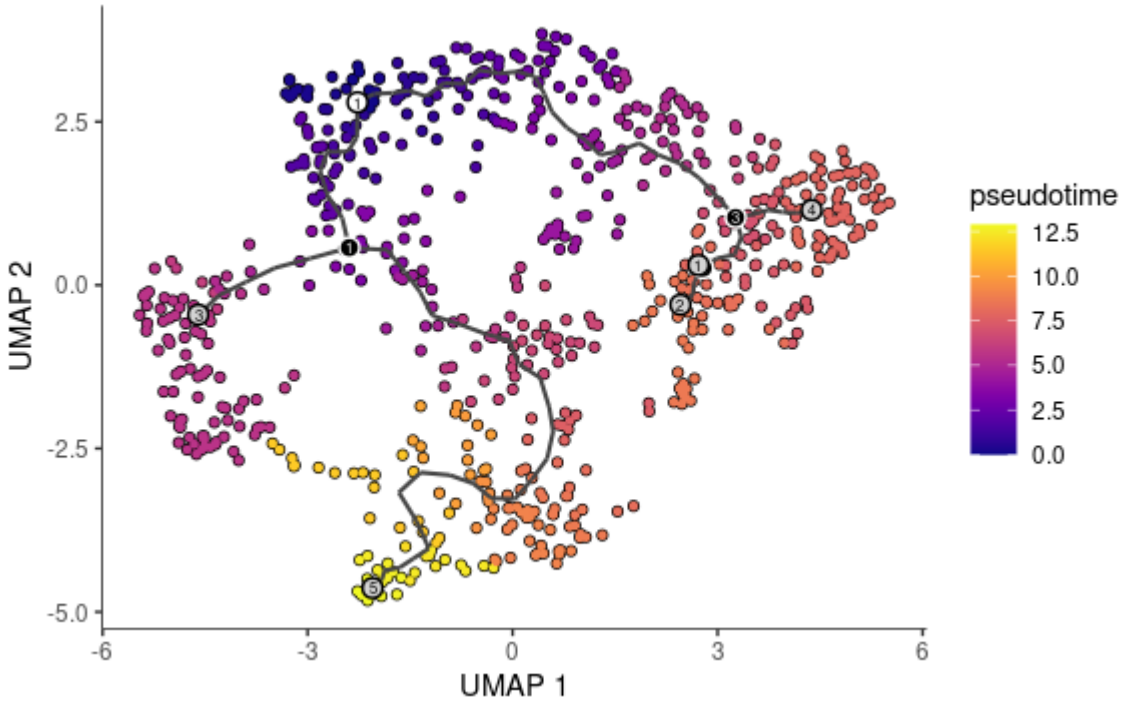

Figure 5: pseudo-time trajectory of epithelial cells

and S100A11

3.3 Chromosomal rearangment

As shown in Fig6, the most common structural variation between epithelial cells and control include loss of signal on chromosomes 6 and 13 and gain of signal on chromosomes 20, 21, and 16. However, there is also a loss and a gain of signal among control cells on chromosome 6 which potentially might be the symptom of the MHC genes cluster affection on results in that region.

3.4 Signaling pathways

39 significantly expressed signaling pathways were recognized. Chords diagrams, ligand-receptor pairs contribution, and signaling role of cells plots for all that pathways are available in a supplement. Here, I will describe only a few, potentially the most important results.

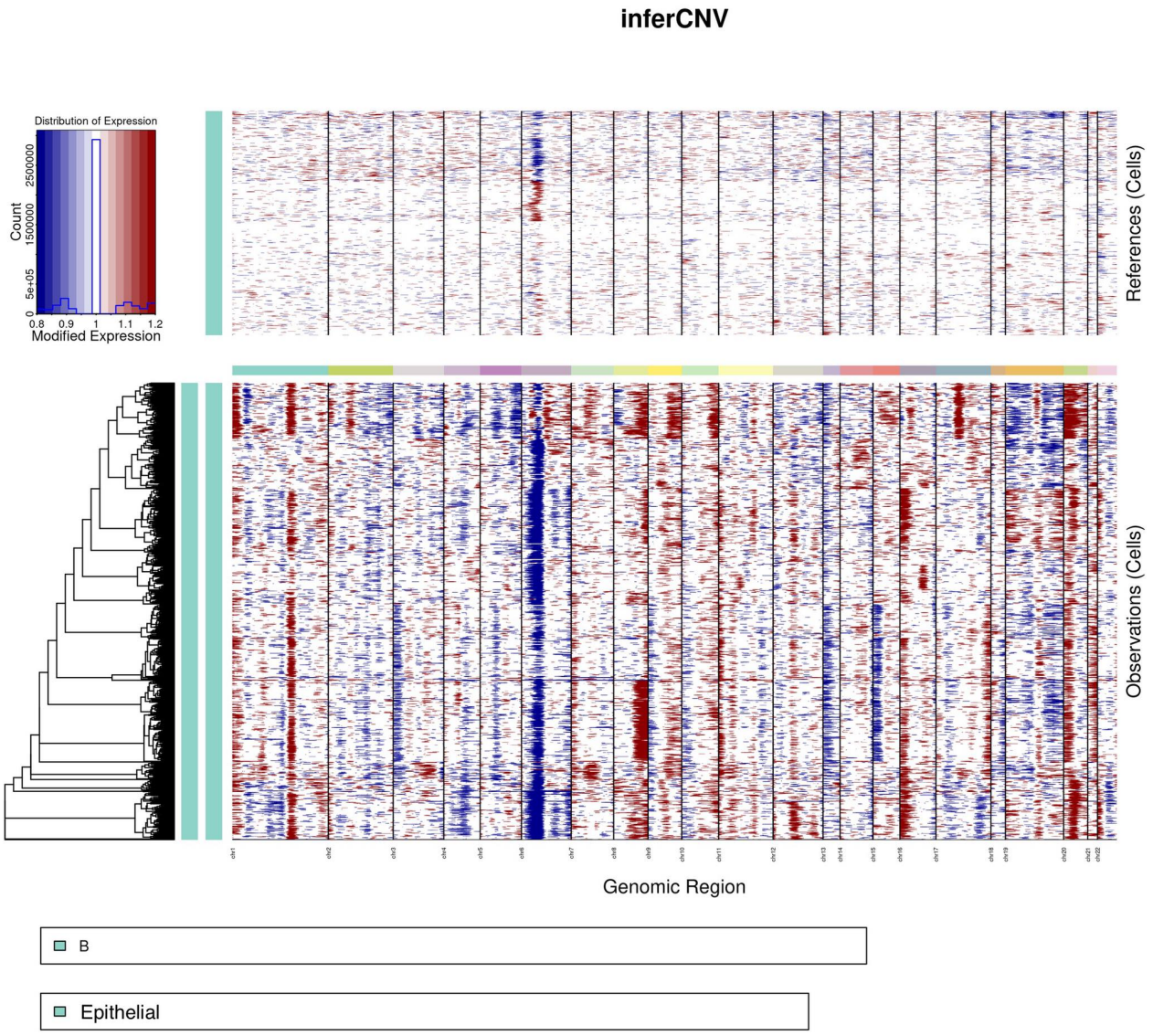

Figure 6: Large scale copy number variations

3.4.1 IGF pathway

The activation of IGF signaling pathways can promote growth, metastasis, and drug resistance in many types of human tumors, including mesenchymal, epithelial, and hematopoietic cancer [5]. Two cell lineages that communicate by this pathway are epithelial cells (playing the role of receiver and influencer) and fibroblast cells (sender and influencer). The main ligand-receptor pair that contribute to this pathway is IGF1(ITGA6+ITGB4).

3.4.2 VEGF pathway

VEGF signaling pathway is responsible for stimulation of the formation of blood vessels and due to that also play important role in the metastasis process. Ligand-receptor pairs that contribute to that pathway are VEGFAVEGFR1, VEGFAVEGFR1R2, VEGFAVEGFR2, VEGFBVEGFR1. Endothelial cells play the role of the receiver while Mast and Myeloid mononuclear phagocytose system cells (and in less portion fibroblast and Epithelial cells) play as senders.

3.4.3 HGF pathway

Cancer-associated fibroblasts have been shown to promote the growth, survival, and migration of cancer cells in an HGF-dependent manner [6]. HGF signaling in cancer cells is associated with increased tumor aggressiveness and predicts poor outcomes. The only ligand-receptor pair that contributes to that pathway is HGFMET. The fibroblast is a sender and the epithelium is a receiver in this pathway.

3.4.4 Pathways clustering and general communication patterns

Based on functional similarity four clusters were recognized:

- FGF, ANGPT, ANGPTL, PDGF, PERIOSTIN,ncWNT, VEGF
- PTN, COMPLEMENT, CSF, CXCL, MK, MIF
- GALECTIN, CCL, INF-II, TNF, ANNEXIN, IL16, LIGHT
- VISFATIN, CD137, PARs, LIFR, TGFb, GNR, BAFF, SPP1, CALCR, GRN, OSM, EGF

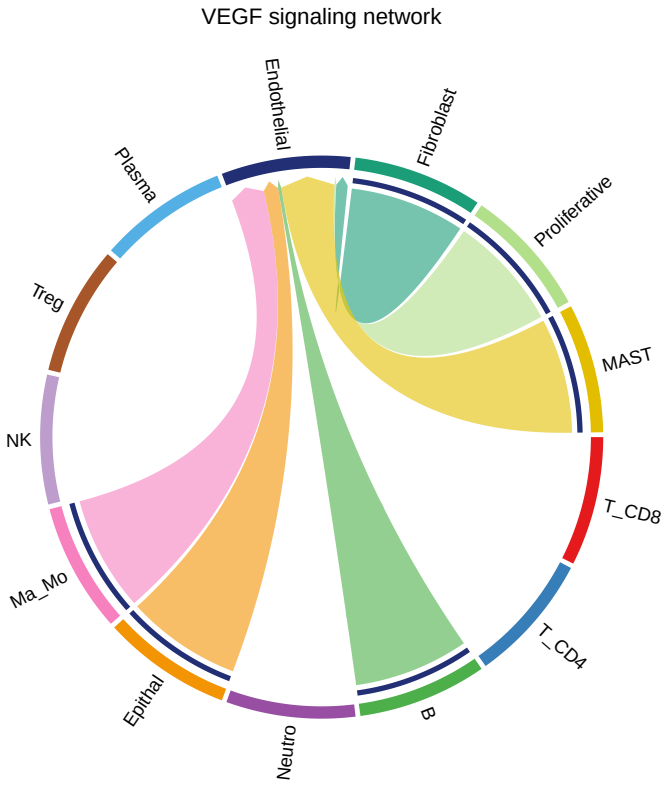

Figure 7: VEGF pathway chord diagram

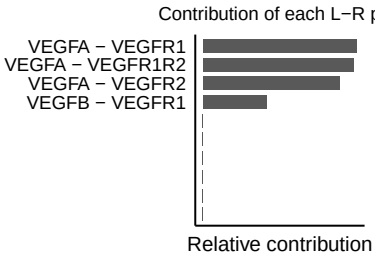

Figure 8: VEGF pathway Ligand-Receptor pairs contribution

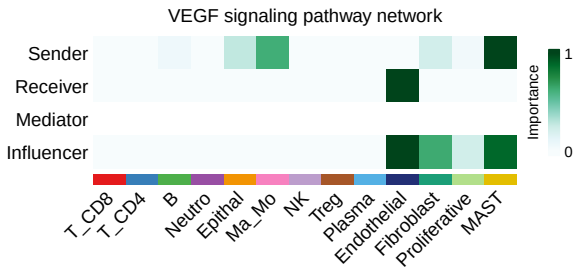

Figure 9: VEGF pathway cells signaling roles

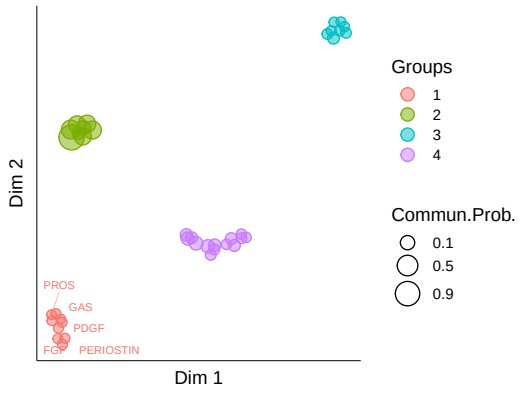

Figure 10: Pathways clustering based on structural similarity

Clustering based on structural similarity also recognizes four clusters:

- PROS, GAS, IGF, HGF, FASLG, TWEAK, FGF, PDGF, PERIOSTIN
- VISFATIN, CXCL, SPP1 GALECTIN, PTN, ANNEXIN, MK, MIF
- ANGPT, CD137, COMPLEMENT, KIT, LIGHT, ncWNT, GRN, IL1
- TGFb, IL16, TNF, CCL, PARs VEGF, ANGPTL, BAFF, LIFR, CSF, CALCR, OSM, EGF, IFN-II

Five incoming general patterns include

- pattern1: T CD8+, T CD4+, B, NK, T regulatory, Plasma, Proliferative cells, Mast
- pattern2: Fibroblast, Epithelial
- pattern3: Neutrophiles
- pattern4: Endothelial
- pattern5: Myeloid mononuclear phagocytose system cells

Five outgoing general patterns include:

- pattern1: T CD4+, B
- pattern2: Fibroblast

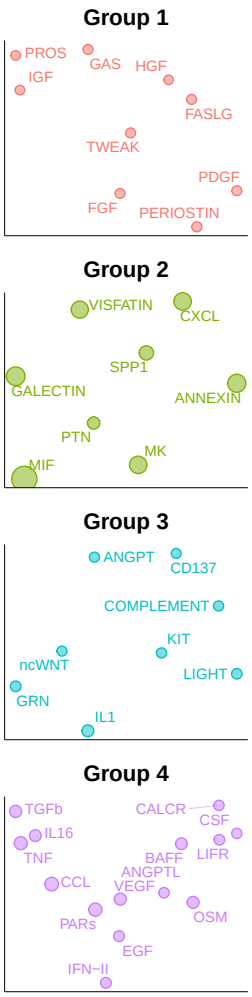

Figure 11: Pathways clustering based on structural similarity - zoom

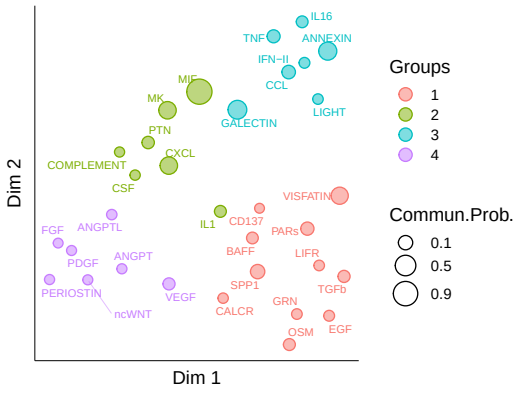

Figure 12: Pathways clustering based on functional similarity

- pattern3: Mast
- pattern4: Neutrophiles
- pattern5: TCD8+, NK

3.5 Discussion

Gastric cancer shows a high degree of heterogeneity during the progression. The proportion of different cell lineages differs among metastasis samples, and also among primary tumors. The heterogeneity is shown also in the pseudo-time trajectory of epithelial cells, which has a few branch points corresponding to the cell fate decisions . Cell-cell communication pathways suggest, that malignant epithelial cells are strongly affected by interactions with cancer-related fibroblast and endothelial cells, and probably are evidence of epithelial-mesenchymal transition. The results obtained in this project may become the basis for further, more complex analysis, which probably would include a comparison of activation of different pathways among samples and also clustering of immune cells into more specific lineages, based on a full data set.

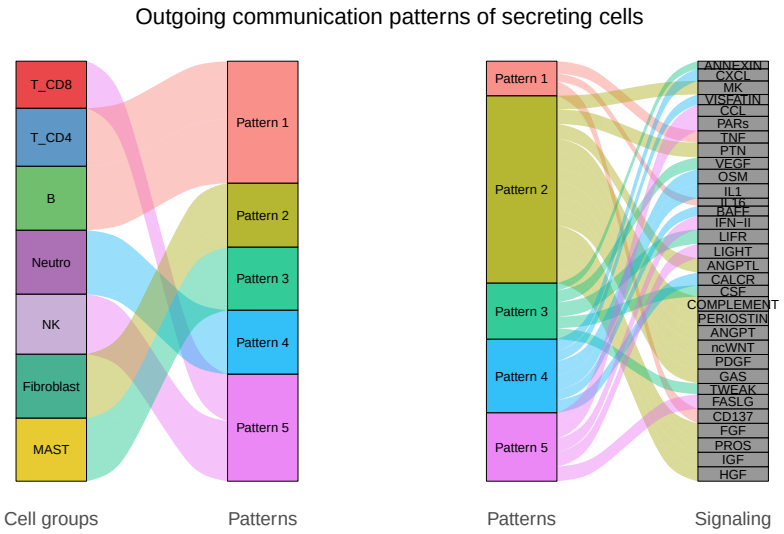

Figure 13: Outgoing patterns

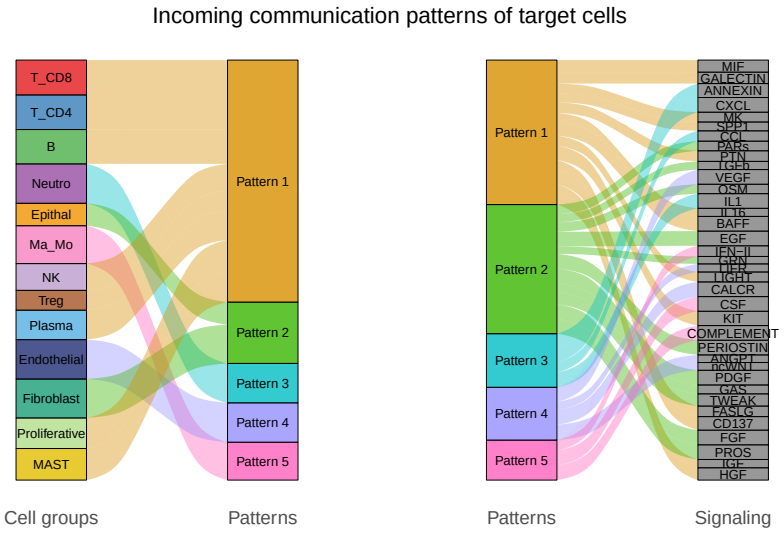

Figure 14: Incoming patterns

4

References

[1] Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021). URL `https://doi.org/10.1016/j.cell.2021.04.048`.

[2] Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979–982 (2017).

[3] Jin, S. *et al.* Inference and analysis of cell-cell communication using cellchat. *Nature communications* **12**, 1–20 (2021).

[4] Franzén, O., Gan, L.-M. & Björkegren, J. L. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database* **2019** (2019).

[5] Hua, H., Kong, Q., Yin, J., Zhang, J. & Jiang, Y. Insulin-like growth factor receptor signaling in tumorigenesis and drug resistance: a challenge for cancer therapy. *Journal of hematology & oncology* **13**, 1–17 (2020).

[6] Owusu, B. Y., Galemmo, R., Janetka, J. & Klampfer, L. Hepatocyte growth factor, a key tumor-promoting factor in the tumor microenvironment. *Cancers* **9**, 35 (2017).