Extracting Gene Signatures Using Single-Cell Analysis in Head and Neck Cancer

Janet Li[1] and Alborz Bejnood[2]
[1]Palo Alto High School, [2]Broad Institute

INTRODUCTION

Head and neck squamous cancer (HNSC) constitutes about 4% of all cancers in the US. In general, cancer results from DNA changes that cause uncontrolled cell division. The genes that are “expressed,” or turned on, in a particular cell determine what that cell can do (Mehanna, 2010).

GOAL: To investigate how differences in gene expression affect phenotypes and survival for HNSC patients, identifying genes that cancer cells express but normal cells do not (or vice versa) so that a treatment can be designed that will kill cancer cells but won't kill regular cells.

RESEARCH METHODS

STEP 1: Download single-cell gene expression data (Puram, 2017), and bulk TCGA gene expression and survival data (Grossman, 2016)

STEP 2: Use SingleR to cluster cells and view single-cell data at a deeper level of granularity (Aran, 2019)

STEP 3: Use STREAM to obtain pseudotime plots (Chen, 2018)

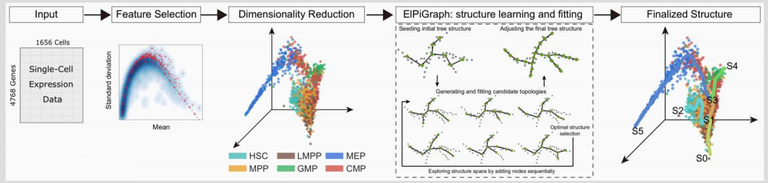

Figure 1. STREAM Pipeline for Trajectory Inference

STEP 4: Use the single-cell data in combination with bulk data for which we have survival information to come up with gene signatures that have statistically significant survival relevance

- Use GSEA to look for survival differences in bulk data between patients with high expression and low expression of the genesets identified in single-cell data (Mootha, 2003)
- Use MSigDB to find overlaps between our genesets and other annotated genesets (Subramanian, 2005)
- Kaplan-Meier plots to see how those genes are significant in terms of survival (Kaplan, 1958)

CONCLUSIONS AND ANALYSIS

The gene expression data is a snapshot of the cells at a single point in time, but in reality, there is a progression of cell states. From the pseudotime plots, we can see that cancer cells change over time as they become metastatic; similarly, fibroblasts also change over time, suggesting a correlation between fibroblasts and cancer cells.

GSEA identified 95 enriched genes for fibroblasts and 70 enriched genes for macrophages. This means that those genes are more highly expressed in patients who are alive than dead in a statistically significant way. These genes comprise a “gene signature,” a set of genes that when taken together, can differentiate between how well a patient survives.

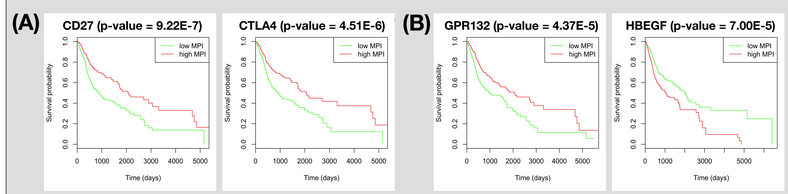

Figure 5. Survival Analysis

(A) Kaplan-Meier survival curves of 2 genes for fibroblasts.
(B) Kaplan-Meier survival curves of 2 genes for macrophages.

DATA AND FINDINGS

Single-Cell Data (Puram, 2017)

- 5,902 cells of 10 different cell types
- 23,686 genes
- 18 patients

Bulk Data (Grossman, 2016)

- 36,899 genes
- 545 patients
- Survival information (status and days)

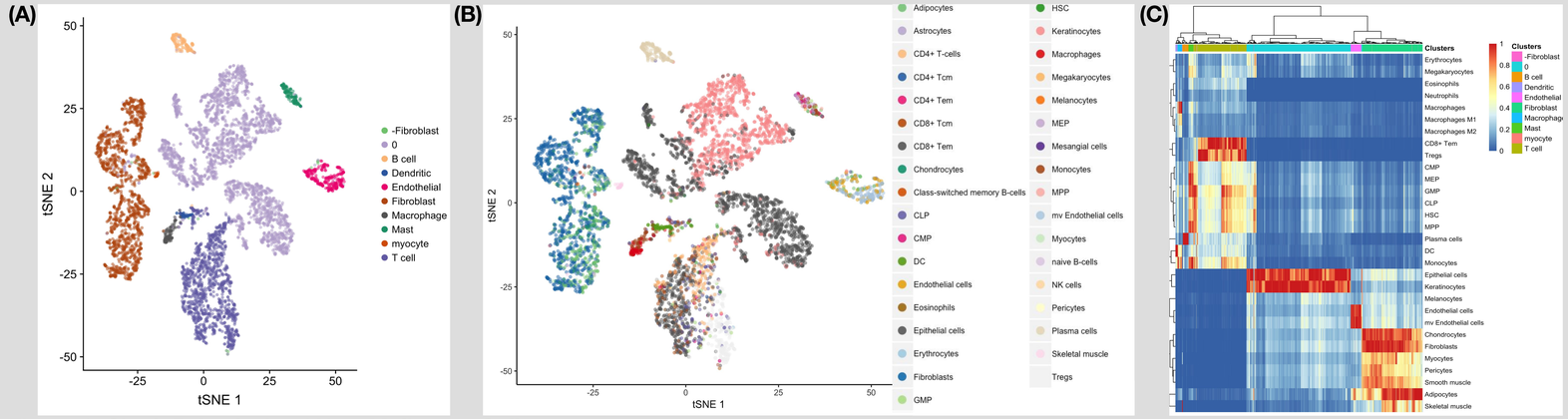

Figure 2. SingleR Analysis of Non-Malignant Cells

(A) Seurat t-SNE (t-distributed stochastic neighbor embedding) plot. Clusters are assigned to indicated cell types by differentially expressed genes.
(B) SingleR t-SNE plot offers increased granularity, showing distinctions between cell types within each of the original clusters.
(C) Heatmap compares the original cell identities to the SingleR annotations.

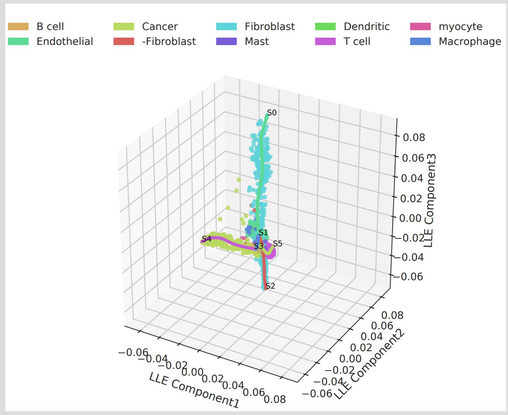

Figure 3. Pseudotime Plot of Single-Cell Data

Fibroblasts and cancer cells have distinct time trajectories, while the other cell types remain relatively static.

(A)

	PROBE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	UBXN11	15	0.211	0.0177	Yes
2	LIMD2	31	0.200	0.0344	Yes
3	SH3BP2	40	0.196	0.0512	Yes
4	APBB3	113	0.179	0.0635	Yes
5	IGFLR1	160	0.173	0.0765	Yes

(B)

	PROBE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	ZAP70	7	0.222	0.0119	Yes
2	MASP1	11	0.212	0.0234	Yes
3	CTLA4	12	0.212	0.0350	Yes
4	KLRK1	16	0.211	0.0465	Yes
5	FGD3	24	0.204	0.0574	Yes

(C)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
MODULE_45 [583]	Whole blood genes.	19		$1.29\ e^{-20}$	$2.3\ e^{-16}$
CHEN_METABOLIC_SYNDROM_NETWORK [1210]	Genes forming the macrophage-enriched metabolic network (MEMN) claimed to have a causal relationship with the metabolic syndrom traits.	23		$9.9\ e^{-20}$	$8.81\ e^{-16}$
SMID_BREAST_CANCER_NORMAL_LIKE_UP [476]	Genes up-regulated in the normal-like subtype of breast cancer.	16		$1.22\ e^{-17}$	$7.27\ e^{-14}$
GNF2_ITGB2 [57]	Neighborhood of ITGB2	8		$2.34\ e^{-14}$	$1.04\ e^{-10}$
TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_ON_HSC_DN [187]	Genes down-regulated in normal hematopoietic progenitors by RUNX1-RUNX1T1 [GeneID=861;862] fusion.	10		$2.31\ e^{-13}$	$8.24\ e^{-10}$

(D)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
SMID_BREAST_CANCER_NORMAL_LIKE_UP [476]	Genes up-regulated in the normal-like subtype of breast cancer.	26		$3.8\ e^{-30}$	$6.76\ e^{-26}$
GO_IMMUNE_SYSTEM_PROCESS [1984]	Any process involved in the development or functioning of the immune system, an organismal system for calibrated responses to potential internal or invasive threats.	36		$2.41\ e^{-25}$	$2.15\ e^{-21}$
GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_UP [200]	Genes up-regulated in comparison of naive CD8 T cells versus day 0 monocytes.	18		$6.36\ e^{-25}$	$3.78\ e^{-21}$
MODULE_45 [583]	Whole blood genes.	23		$1.93\ e^{-23}$	$8.58\ e^{-20}$
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_MYELOID_UP [200]	Genes up-regulated in comparison of systemic lupus erythematosus CD4 (GeneID=920) T cells versus systemic lupus erythematosus myeloid cells.	16		$2.12\ e^{-21}$	$6.3\ e^{-18}$

Figure 4. GSEA and MSigDB Outputs

(A) GSEA output of enriched genes (top 5 genes shown) for macrophages
(B) GSEA output of enriched genes (top 5 genes shown) for fibroblasts
(C) Top 5 MSigDB gene sets that overlap with GSEA enriched genes for macrophages
(D) Top 5 MSigDB gene sets that overlap with GSEA enriched genes for fibroblasts

IMPLICATIONS AND NEXT STEPS

These findings are important because researchers can target those signature genes with drugs or reprogram those genes to help diagnose and treat HNSC. Furthermore, we now have a more complete picture of the different cell types in HNSC. In particular, the previously-identified fibroblast population looks to be a mix of two different types of fibroblasts, one of which may be cancer-associated.

In the future, these methods can be repeated on other cell types within this single-cell data set (we only analyzed fibroblasts and macrophages) or on a different single-cell data set (we only used 1 data set of 18 patients). This research is a potential framework for using single-cell data to find features that impact survival.

ACKNOWLEDGEMENTS / REFERENCES

Special thanks to Alborz Bejnood, Ms. Angell and the AAR program for making this project possible.

Works Cited:

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., ... & Butte, A. J. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 1.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411.

Chen, H., Albergante, L., Hsu, J. Y., Lareau, C. A., Bosco, G. L., Guan, J., ... & Langenau, D. M. (2018). STREAM: Single-cell trajectories reconstruction, exploration and mapping of omics data. *BioRxiv*, 302554.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, *375*(12), 1109-1112.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

Mehanna, H., Paleri, V., West, C. M. L., & Nutting, C. (2010). Head and neck cancer—part 1: epidemiology, presentation, and prevention. *Bmj*, *341*, c4684.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., ... & Houstis, N. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, *34*(3), 267.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., ... & Deschler, D. G. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, *171*(7), 1611-1624.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550.