Single Cell Transcriptome Meta-network Analysis Pipeline to Discover Regulators of Metastasis in Breast Cancer

*Junha Cha[1], Junhan Kim[1], Insuk Lee[1,2]**

*Correspondence: insuklee@yonsei.ac.kr
[1] Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Korea
[2] Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul 03722, Korea

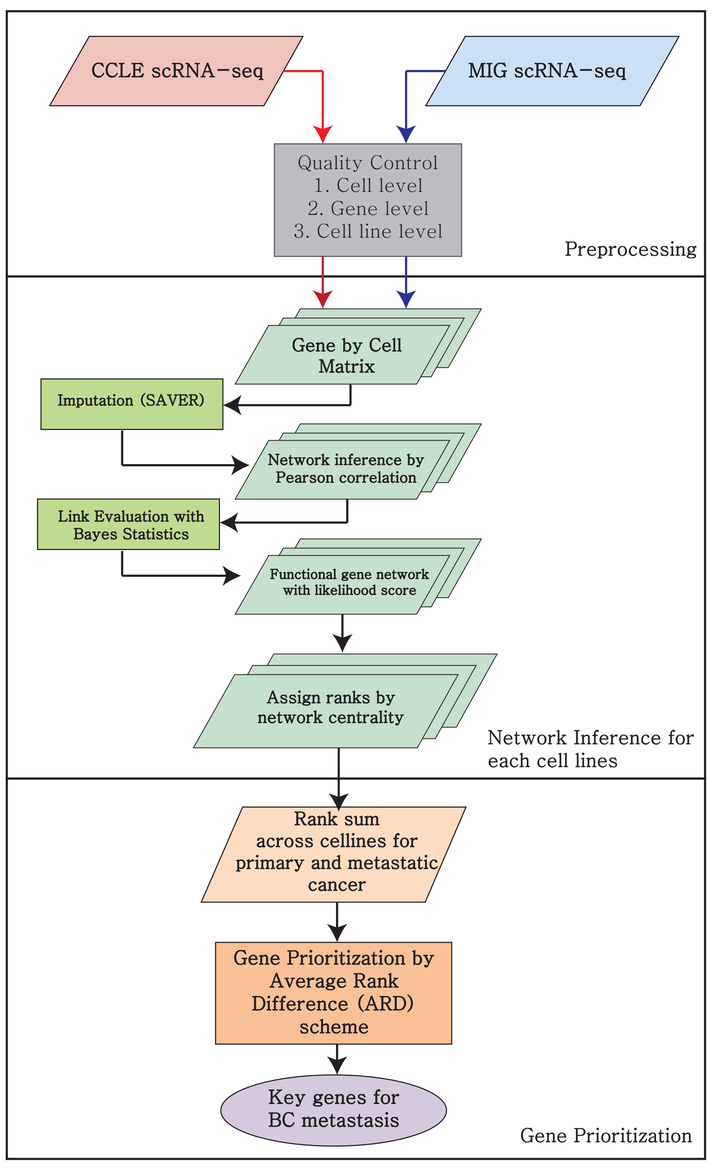
I. Introduction

- Single cell transcriptome data provides unique opportunity to assess and analyze network properties of a system at a cell-type specific level.
- However, single cell dataset with its insufficient capture rate and high dimensionality presents construction of cell-type specific functional gene network(FGN) a challenging task.
- In this research, we constructed functional gene network at a cell line specific level (scNET), using coexperssion pattern of genes by each cell lines. Bayesian Statistics and imputation was used to derive functional links.
- Moreover, we show that gene prioritization through network topology is bioinformatically evaluated and validated using 664 curated gene set assoicated with breast cancer metastasis.

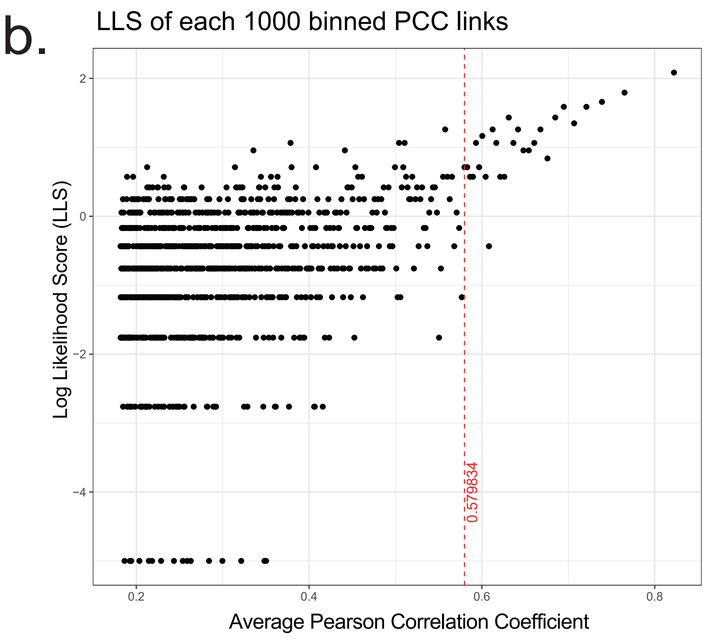
II. Study design & scNET Pipeline

- Two publically available breast cancer single cell dataset was selected (CCLE and MIG) to construct cell-line specific gene functional network. (Figure a)
- For each of 23 cancer cell lines, cells and genes were qualtiy controlled through with mitochondrial content (20%), unique feature counts (2000 < x < 9000), cell fraction percentage of expression (1%), and viable cell line (cell count over 100).
- For each cell-line count matrix, imputation was performed with SAVER and each gene pair pearson correlation was calcuated for log likelihood score with a true positive functional gene link sets derived from high confidence evidence of GOBP database. (Figure b) This resulted in a total of 16 cell-line specific scNET for downstream analysis.
- LLS score for each gene pair was calculated using a regression function from 1000 binned gene pair, and gene pairs with over a threshold (LLS > 1) was retained as the final cell-line FGN. (Figure c)

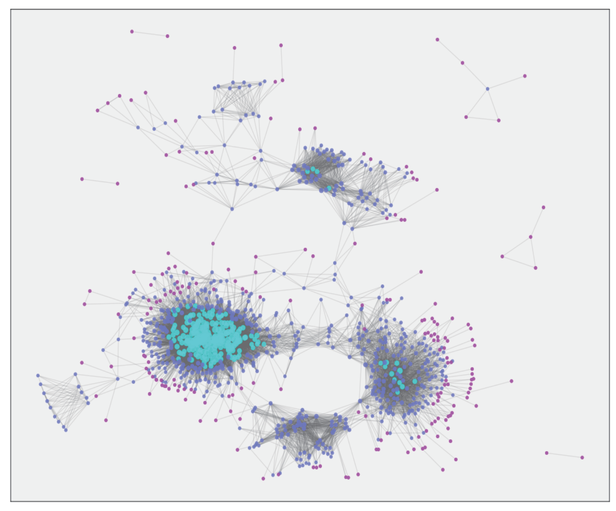
a.



b.



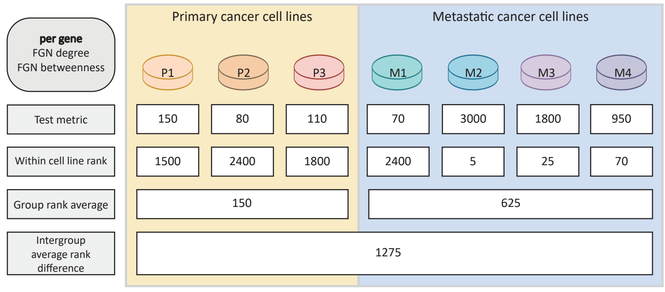
c.



III. Results

- Using network topology of both primary and metastatic cell line networks, each gene was ranked based on either degree or betweenenss.
- Genes that were associated with breast cancer was calcuated as those that showed the most average rank difference (Figure d) between primary cell line networks and metastatic cell line networks. (high ranked in metastatic networks but low ranked in primary networks)
- For each centrality measure (degree and betweeness) and DEG calculated based on wilcoxon test, top top candidates were calcuated for enrichment of 664 curated gene set. Random expected enrichment value is represented as the dotted line (Figure e). In the top candidates we observed that network based gene prioritization(red, green) outperformed candidates derived by DEG analysis.
- We show that candidate genes from DEG and network topology have only minor overlaps (Figure f), suggesting that the two method provide different, mutual information.
- Experimental validation of top candidate genes are currently being performed on cell lines.

d.

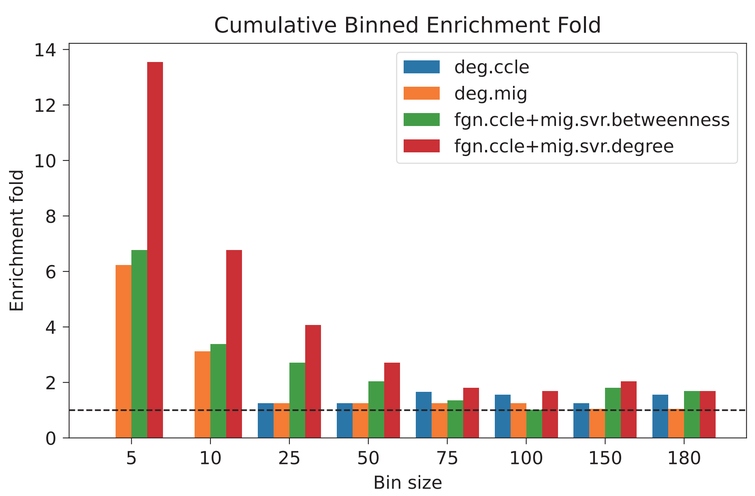


$$Define\ e_{ij}\begin{cases}1\ (i=j)\\0\ (i\neq j)\end{cases}\ and\ r_{x,i}\ as\ rank\ of\ gene\ x\ in\ sample\ i.$$

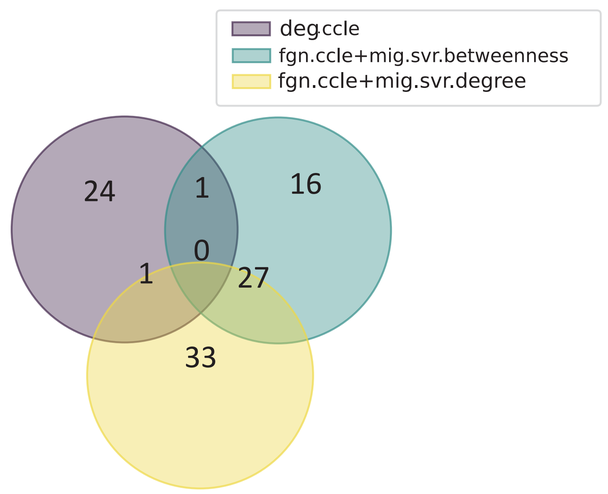
$$Given\ S=\{i_1,i_2,\dots,i_n\}, A\subsetneq S, and\ A\neq\emptyset,$$

$$for\ gene\ x, ARD_{A,x}=\frac{\sum_{i\notin A}\sum_{j=1}^{n}e_{ij}r_{x,j}}{\sum_{i\notin A}\sum_{j=1}^{n}e_{ij}}-\frac{\sum_{i\in A}\sum_{j=1}^{n}e_{ij}r_{x,j}}{\sum_{i\in A}\sum_{j=1}^{n}e_{ij}}$$

e.



f.



III. Conclusions

- By utilizing imputation and Bayes statistics, we show that cell-line specific functional gene network may be constructed with scRNA-seq datasets.
- Gene set prioritization through network topology on multiple Breast cancer cell lines was perfomed with our established ARD pipeline, which was evaluated to have improved performance compared to DEG analysis.