PINSPlus: a novel tool for molecular subtyping and multi-omics integration

Hung Nguyen, Sangam Shrestha, Tin Nguyen*
Computer Science and Engineering, University of Nevada, Reno
Contact: tinn@unr.edu, Website: https://www.cse.unr.edu/~tinn/

Background

Over-diagnosis:

- After decades of screening, the chance of person being diagnosed with prostate or breast cancer has doubled
- The number of patients with advanced disease has been reduced only marginally

➔substantial harm of excess detection and over-diagnosis.

Under-diagnosis:

- 30-50% of patients with non-small cell lung cancer develop recurrence and die after curative resection
- Adjuvant therapy (chemo & radiation) is NOT routinely recommended although has shown to significantly improve survival

➔many patients die because they did not receive needed treatment

The problem

- Our current inability to distinguish between patient subgroups (respondent vs. non-respondent) and disease subtypes (aggressive vs. non-aggressive)

The challenge

- Mining high-throughput molecular data to discover subtypes characterized by clinical differences, such as survival and disease recurrence.

Our solution: PINSPlus

The goal is to discover subgroups of patients that share a common clinical behavior.

- Perturbation clustering: ensures robustness against the unstable nature of quantitative assays
- Multi-omics integration: discovers subtypes that can be triggered at different levels (mRNA, miRNA, epigenetics, etc.)
- Availability: CRAN R package (https://cran.r-project.org/package=PINSPlus)

Perturbation clustering

- Data perturbation reveals the true structure of molecular data
- Patient connectivity is persevered, regardless of the clustering algorithms (k-means, hierarchical clustering) or the number of clusters

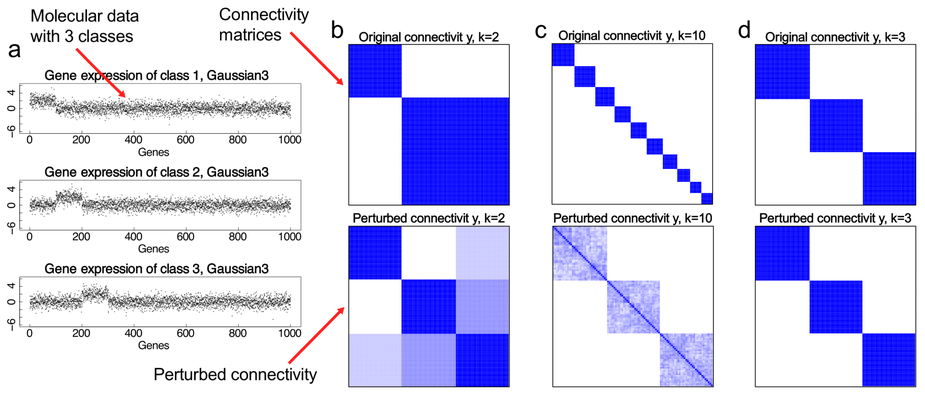

***Fig. 1**: Resilience of pair-wise connectivity*

Multi-omics integration

- Cluster ensemble to find the consistent patterns
- Multiple stages to reveal the hierarchical data structure

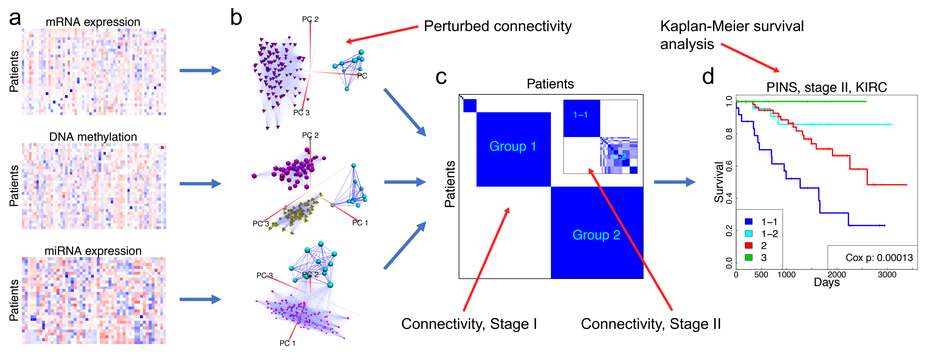

***Fig. 2**: Subtyping of kidney renal clear cell carcinoma (KIRC)*

Validation data

The Cancer Genome Atlas (TCGA):

- KIRC: Kidney renal clear carcinoma, 124 patients
- GBM: glioblastoma multiforme, 273 patients
- LAML: acute myeloid leukemia, 164 patients
- LUSC: lung squamous cell carcinoma, 110 patients

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC):

- Discovery cohort: 997 patients,
- Validation cohort: 995 patients

Validation results

- Metric: We use Cox regression [1] to assess statistical significance of survival differences
- Methods: PINSPlus [2,3], Similarity Network Fusion (SNF) [4], Consensus Clustering (CC) [5], and iClusterPlus [6].
- Results: PINSPlus substantially outperforms other state-of-the-art subtyping approaches in discovering subtypes with significantly different survival profiles.

***Table 1**: Cox p-values of discovered subtypes. Cells highlighted in green have the most significant Cox p-value*

Datasets	#Patient	PINS+	CC	SNF	iCluster+
KIRC	124	6e−5	0.104	0.662	0.011
GBM	273	1.2e−4	0.039	0.043	0.108
LAML	164	8.7e−4	0.035	1.5e−3	2.1e−3
LUSC	110	8.4e−3	0.794	0.071	0.314
Discovery	997	1.8e−9	2.5e−5	2.3e−5	0.167
Validation	995	3.4e−5	0.012	0.01	1.9e−3

***Table 2**: Running time (in minutes)*

Datasets	#Patient	PINS+	CC	SNF	iCluster+
KIRC	124	<1m	< 1m	< 1m	1675m
GBM	273	2m	< 1m	< 1m	3598m
LAML	164	<1m	< 1m	< 1m	2011m
LUSC	110	<1m	< 1m	< 1m	1602m
Discovery	997	19m	14m	4m	5155m
Validation	995	11m	14m	2m	5153m

References

1. Therneau, T. M. and Grambsch, P. M. Modeling Survival Data: Extending the Cox Model (Springer, 2000).
2. Nguyen, H., Shrestha, S., and Nguyen, T. (2018). PINSPlus: Clustering algorithm for data integration and disease subtyping. CRAN R package.
3. Nguyen, T., Tagett, R., Diaz, D., and Draghici, S. (2017). A novel approach for data integration and disease subtyping. Genome research, 27(12), 2025–2039.
4. Wang et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333.
5. Monti et al. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. 52(1-2), 91–118.
6. Mo et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences, 110(11), 4245–4250.