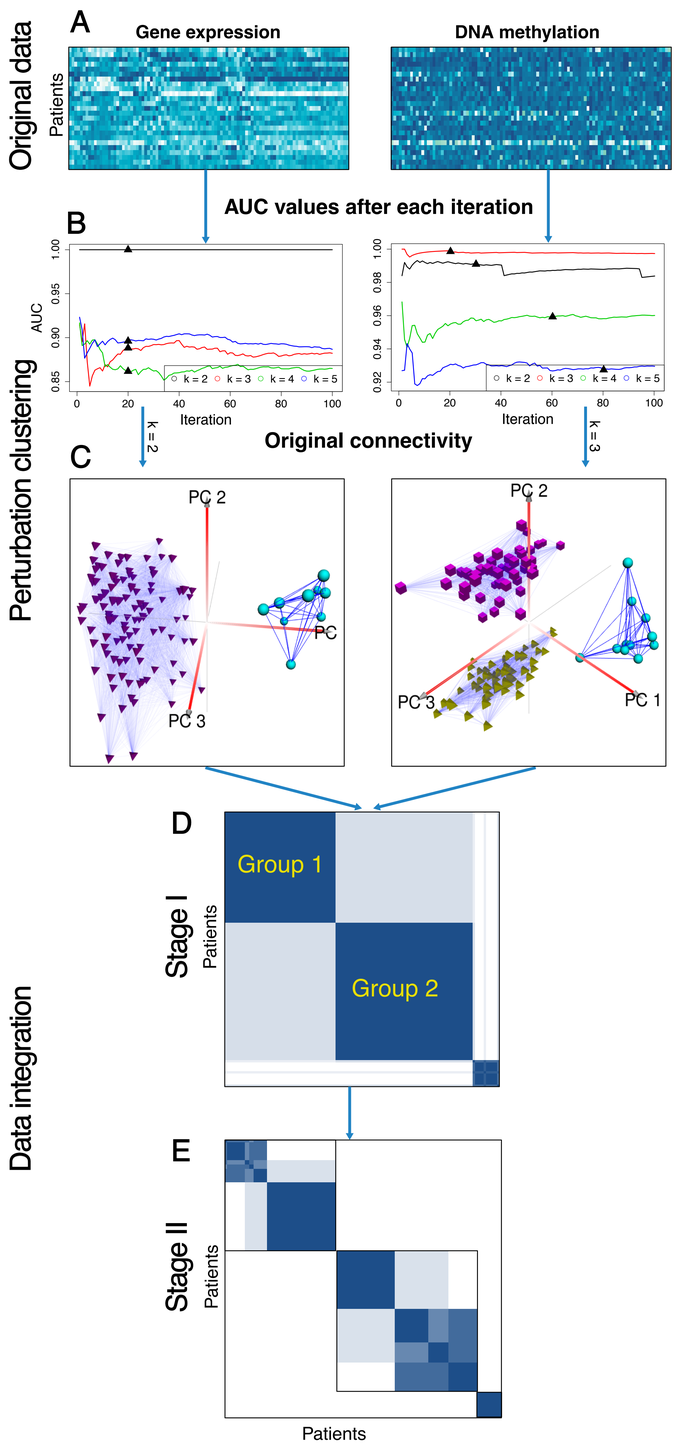
PINSPlus: a fast yet efficient tool for disease subtype discovery using multi-omics data

*Hung Nguyen, Sangam Shrestha and Tin Nguyen**
Computer Science and Engineering Department, University of Nevada, Reno
Contact: tinn@unr.edu

Introduction

Individual analysis of data might not always be sufficient to depict the overall biological phenomenon; Integrative methods are desired to solve the limitation and allow more efficient utilization of clinical data. PINSPlus [1], an improvised version of Perturbation clustering for data Integration and disease Subtyping (PINS) [2], is an unsupervised clustering approach for the meaningful integration of data and discovery of molecular disease subtypes. Unlike many other clustering algorithms, PINSPlus is able to automatically compute the optimal number of clusters based on the stability of membership in each group. Overall, PINSPlus is a powerful tool in disease subtyping and stands out among similar approaches with its distinct features including parallel processing and multi-omics data integration.

Availability: https://cran.r-project.org/package=PINSPlus



Algorithms

Perturbation Clustering: The algorithm begins by generating connectivity matrices of both original data and perturbed data with the range of a predefined number of clusters. With each *k*, Perturbation Clustering performs clustering with the original data and builds its connectivity matrix. Perturbation Clustering then, in an iteration, perturbs input data and performs clustering to generate perturbed connectivity matrices. The iteration continues until stopping criteria are satisfied: i) it reaches the maximum number of iteration or ii) the merged-perturbed connectivity matrix is converging. The returned value of Perturbation Clustering consists of the optimal number of clusters and its membership, and a list of original and perturbed connectivity matrices.

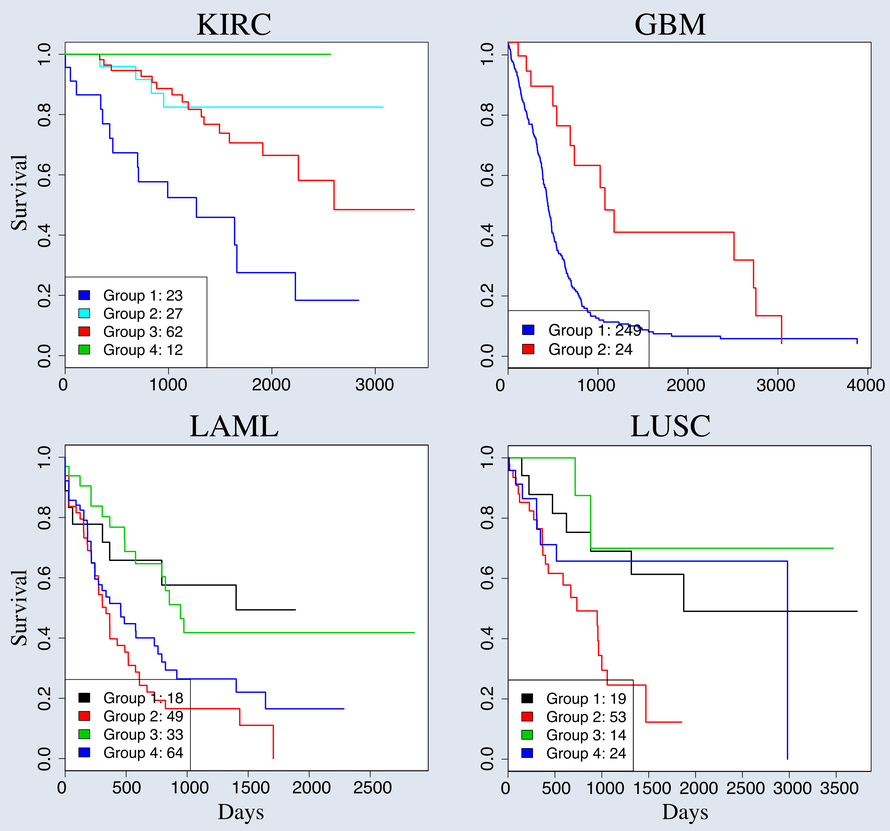
Multi-Omics Data Intergation: In Stage I, PINSPlus combines the connectivity matrices generated from Perturbation clustering from each data type. If all the original connectivity matrices are highly in consent with each other, PINSPlus computes groups for this step by applying hierarchical clustering on the average original connectivity matrix by cutting the height that provides maximum cluster separation. In stage II, PINSPlus aims to further split each discovered group from stage I into subgroups if possible by performing Perturbation Clustering on discovered groups.

Application

Cox p-value of subtypes discovered by PINSPlus, Consensus Clustering (CC) [4], Similarity Network Fusion (SNF) [2], and iClusterPlus [3] for the kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), acute myeloid leukemia (LAML), lung squamous cell carcinoma (LUSC), and the Molecular Taxonomy of Breast Cancer International Consortium (Discovery and Validation).

Datasets	#Samples	PINS+	PINS	CC	SNF	iCluster+
KIRC	124	6e−5	1.3e−4	0.104	0.662	0.011
GBM	273	1.2e−4	8.7e−5	0.039	0.043	0.108
LAML	164	8.7e−4	2.4e−3	0.035	1.5e−3	2.1e−3
LUSC	110	8.4e−3	9.7e−3	0.794	0.071	0.314
Discovery	997	1.8e−9	6.5e−10	2.5e−5	2.3e−5	0.167
Validation	995	3.4e−5	4.3e−5	0.012	0.01	1.9e−3

Kaplan-Meier survival curves for subgroups discovered by PINSPlus from four datasets (KIRC, GBM, LAML, and LUSC)



Running time

PINS, CC and SNF were run using only 1 core whereas PINS+ and iClusterPlus were run using 8 cores. The time is rounded to minute.

Datasets	PINS+	PINS	CC	SNF	iCluster+
KIRC	<1m	14m	< 1m	< 1m	1675m
GBM	2m	80m	< 1m	< 1m	3598m
LAML	<1m	20m	< 1m	< 1m	2011m
LUSC	<1m	10m	< 1m	< 1m	1602m
Discovery	19m	382m	14m	4m	5155m
Validation	11m	344m	14m	2m	5153m

Conclusion

PINSPlus inherits advantages of PINS, including:

- **Finds the optimum number of clusters automatically.**
- **Allows to integrate any data types of samples to refine the discovered clusters.**
- **Produces high accurate results with both single data types and combination of multi-omics data.**
- **Is robust against natural and technical noise.**

PINSPlus enhances PINS performance by:

- **Supporting multicores processing.**
- **Introducing efficient stopping criteria to determine whether perturbation processes should be continued.**

PINSPlus is a flexible version of PINS, which allows users to customize the clustering algorithm and the perturbation method.

PINSPLus is a friendly clustering package, whose required input is a matrix or multiple numeric matrices representing a collection of items and its features.

References

[1] Nguyen, H., Shrestha, S., and Nguyen, T. (2018). PINSPlus: Clustering algorithm for data integration and disease subtyping. Available at: https://CRAN.R-project.org/package=PINSPlus

[2] 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.

[3] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3), 333.

[4] Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences, 110(11), 4245–4250.

[5] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. 52(1-2), 91–118.