*Bioinformatics*, 2019, 1–4
doi: 10.1093/bioinformatics/bty1049
Advance Access Publication Date: 24 December 2018
Applications Note





Genome analysis

# PINSPlus: a tool for tumor subtype discovery in integrated genomic data

**Hung Nguyen[1], Sangam Shrestha[1], Sorin Draghici[2] and Tin Nguyen [1,*]**

[1]Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557, USA and [2]Department of Computer Science, Wayne State University, Detroit, MI 48202, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on July 18, 2018; revised on December 2, 2018; editorial decision on December 10, 2018; accepted on December 19, 2018

## Abstract

**Summary:** Since cancer is a heterogeneous disease, tumor subtyping is crucial for improved treatment and prognosis. We have developed a subtype discovery tool, called PINSPlus, that is: (i) robust against noise and unstable quantitative assays, (ii) able to integrate multiple types of omics data in a single analysis and (iii) dramatically superior to established approaches in identifying known subtypes and novel subgroups with significant survival differences. Our validation on 12,158 samples from 44 datasets shows that PINSPlus vastly outperforms other approaches. The software is easy-to-use and can partition hundreds of patients in a few minutes on a personal computer.

**Availability and implementation:** The package is available at https://cran.r-project.org/package=PINSPlus. Data and R script used in this manuscript are available at https://bioinformatics.cse.unr.edu/software/PINSPlus/.

**Contact:** tinn@unr.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

After decades of screening, the chance of a person being diagnosed with prostate or breast cancer has doubled (Esserman *et al.*, 2009). However, the number of patients with advanced disease has only been marginally reduced, suggesting that current methods of screening result in either false positives or over-diagnosis. At the same time, 30–55% of patients with non-small cell lung cancer develop recurrence and die after curative resection (Uramoto and Tanaka, 2014), suggesting that a subset of patients would have benefited from more aggressive treatments at early stages. The ability to accurately diagnose patients would allow for better patient prognoses.

The purpose of subtyping multi-omics data is to identify molecular patterns that are similar not only at one level (e.g. mRNA), but from a holistic perspective, that can take into consideration phenomena at various levels (e.g. proteomics, epigenetics). Recent efforts to address the challenge of integration often rely on joint statistical modeling (Mo *et al.*, 2013; Wang *et al.*, 2014), which are limited by strong assumptions of the data distribution and by the gene selection step used to reduce computational complexity. In addition, these approaches are sensitive to even a slight change in molecular measurements or parameter settings.

We have developed a radically different approach, Perturbation clustering for data INtegration and disease Subtyping (PINSPlus) which is built upon the resilience of patient connectivity and cluster ensembles to ensure robustness against noise and bias. Our analysis on 12 158 cancer samples demonstrates that PINSPlus overwhelmingly outperforms existing approaches in identifying known subtypes and in discovering novel patient subgroups with significant survival differences.

## 2 Materials and methods

PINSPlus is an unsupervised approach for subtype discovery without using any a priori knowledge (such as clinical variables or known subtypes). The method is based on the observation that small changes in quantitative assays will be inherently present between

individuals, even in a truly homogeneous population. If distinct molecular subtypes do exist, they must be stable with respect to small changes in quantitative assays (Supplementary Fig. S1). In order to discover reliable subtypes, we estimate how often each pair of patients is grouped together in the following scenarios: (i) when data are perturbed, (ii) when using different data types and (iii) when using different clustering techniques. We then partition patients into subgroups that are strongly connected in all scenarios.

PINSPlus optimizes two algorithms of PINS (Nguyen *et al.*, 2017; Nguyen, 2017): (i) PerturbationClustering() to cluster a single data type (Fig. 1A) and (ii) SubtypingOmicsData() to integrate omics data (Fig. 1B). Given a single data type, the function PerturbationClustering() repeatedly perturbs the data (by adding Gaussian noise) and partitions the patients using different values for cluster number. The number of clusters that gives the most stable connectivity is considered optimal. The corresponding connectivity is considered the optimal connectivity.

For data integration, the input of SubtypingOmicsData() consists of multiple matrices for the same set of patients (rows) where each matrix represents a data type. The function outputs: (i) subtyping results using each data type, (ii) subtyping results using multi-omics data in stage I and (iii) subtyping results in stage II (Fig. 1B).

In order to integrate omics data, we represent patient connectivity from each data type as a graph, with patients as nodes and connectivity as edges. Our goal is to identify subgraphs that are strongly connected across all data types. We merge the connectivities of all data types into a similarity matrix that represents the overall connectivity between patients (Fig. 1B). We use several similarity-based algorithms to cluster the similarity and choose the partitioning that agrees the most with the partitionings of individual data types. This ensemble strategy ensures that the identified subtypes are consistent across all data types, and are robust against the choice of clustering algorithms. This completes stage I.

We also add an additional step to check whether the data has a hierarchical structure, i.e. there are subgroups of patients within discovered subtypes. Since our approach is an unsupervised approach, we do not have prior information to take into account important covariates, such as gender, race or demographics. If these signals are predominant, we are likely to miss the real subtypes (Supplementary Fig. S2). Another motivation is that there are often heterogeneous subgroups of patients that share clinically relevant characteristics even within a subtype. For example, in breast cancer, Luminal A and Luminal B are both estrogen receptor positives and are likely to be grouped together. One-round clustering would likely overlook

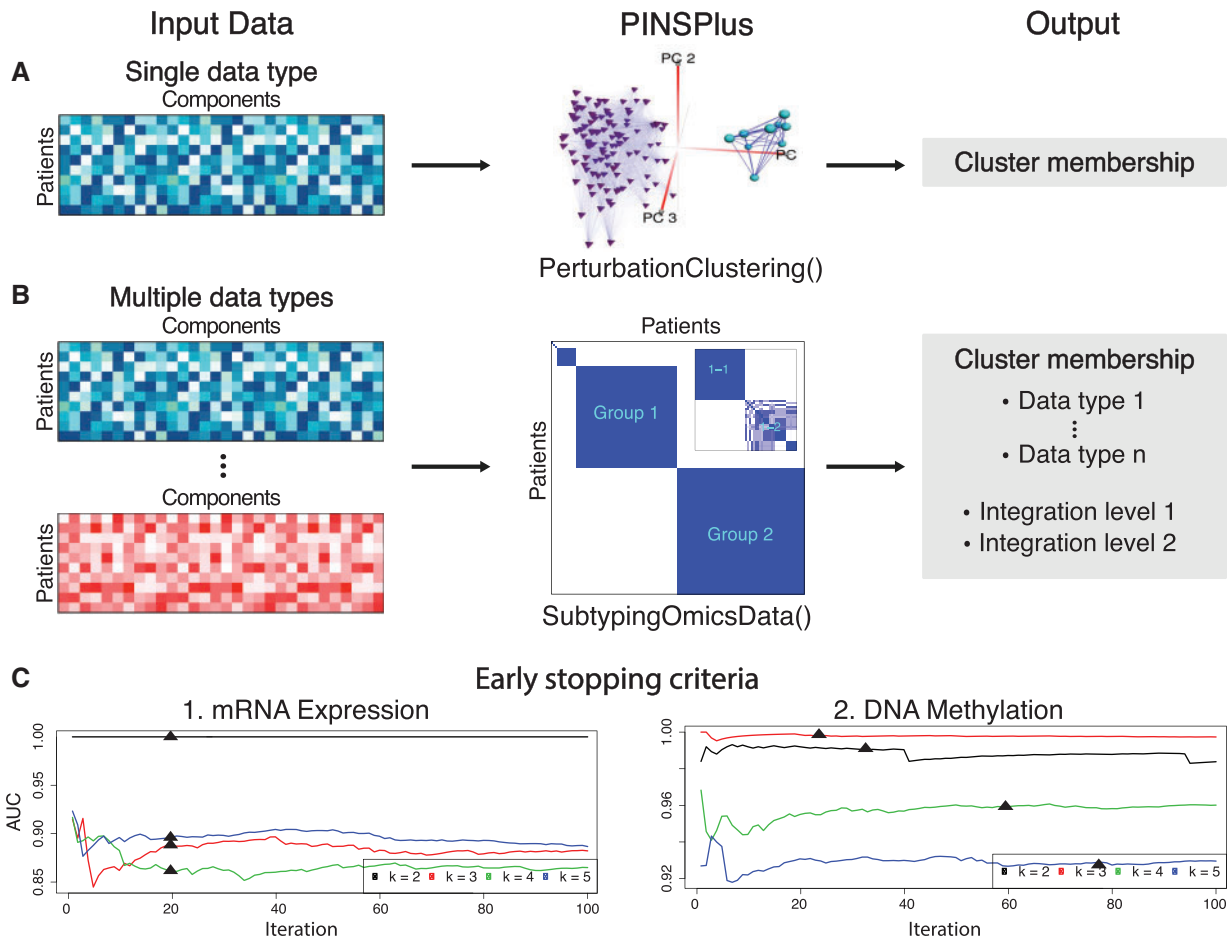


**Fig. 1.** Overall workflow of PINSPlus. (**A**) Subtyping using a single data type. The function PerturbationClustering() reads a single matrix and yields the optimal number of subtypes, as well as cluster membership for each patient. (**B**) Subtyping using multi-omics data. The input consists of multiple matrices (data types) for the same set of patients (rows). The function SubtypingOmicsData() clusters each data type and combines the connectivities to subtype the multi-omics data in stage I. In stage II, the algorithm also attempts to split each discovered group. The output is the cluster membership of each patient, for each data type, and for each of stage I and II. (**C**) Early stoppage criterion. Besides parallel computing, PINSPlus also implemented an early stoppage criterion to speed up the analysis without compromising the results. The triangle symbols indicate the early stop point for each *k* in PINSPlus. We stop the iterations when the AUC values for a given number of cluster (*k*) converge

the subgroups within the groups identified in stage I. In stage II, we attempt to split each discovered group individually, based on reasonable conditions set to avoid over-splitting: (i) stage I clustering has to be extremely imbalanced and (ii) the splitting must be supported by a strong signal across all data types.

PINSPlus also implements an early stopping criterion for the process of generating perturbed connectivity matrices (Fig. 1C). At each iteration, we check whether the AUC values converge. The two panels in Figure 1C show an example using kidney renal clear cell carcinoma (KIRC) data. Each curve represents the AUCs for a value of $k$ (number of clusters). The triangle symbols in each panel indicate the early stopping point for each k in PINSPlus. For mRNA data, the AUC values for $k = 2$ are consistently larger than the rest and thus we terminate all iterations for all values of $k$ after only 20 iterations. For methylation, each curve converges before reaching the maximum number of iterations.

It currently only takes several minutes for the software to cluster hundreds of patients with three or more types of data and tens of thousands of features. The parallel computing allows users to efficiently analyze datasets with tens of thousands of patients. The software uses $k$-means as the default clustering algorithm. We strongly suggest that users run PINSPlus with this setting since it has been extensively tested. However, we also provide hierarchical clustering and partitioning around medoids as built-in alternatives. Users can also incorporate their own algorithm, distance metrics or customized perturbation techniques into PINSPlus (Supplementary Section S1 and Supplementary Fig. S3).

## 3 Results

We tested PINSPlus on the datasets that were analyzed in the original PINS paper (Nguyen *et al.*, 2017): eight mRNA datasets with known subtypes and six multi-omics datasets with known survival (KIRC, GBM, LAML, LUSC and two METABRIC datasets). In addition, we also tested PINSPlus on 30 new omics datasets obtained from TCGA, for a total of 36 multi-omics datasets. We compared PINSPlus with three established subtyping algorithms: Consensus Clustering (CC) (Monti *et al.*, 2003), Similarity Network Fusion (SNF) (Wang *et al.*, 2014) and iClusterPlus (Mo *et al.*, 2013). Supplementary Tables S1–S3 show the details of datasets. Supplementary Table S5 shows the running time of each method.

For the eight mRNA datasets with known subtypes, we use the Rand Index (RI) and Adjusted Rand Index (ARI) to assess the performance of the resulted subtypes. PINSPlus yields the highest RI and ARI values for every single dataset tested (Supplementary Table S4). For the 36 omics datasets with survival information, we use Cox regression to assess the survival difference of the discovered subtypes. Table 1 shows the Cox $P$-values of the subtypes discovered by each of the four approaches. There are nine datasets for which no method is able to identify subtypes with significantly different survival profiles. For the remaining 27 datasets, PINSPlus has significant $P$-values in all of them whereas CC, SNF and iClusterPlus has significant $P$-values in only 8, 14 and 10 datasets, respectively. More importantly, PINSPlus has the most significant $P$-values in 23 datasets.

## 4 Conclusions

As an unsupervised approach, PINSPlus relies solely on molecular data to discover disease subtypes. One caution is that a cluster of samples could be determined not only by molecular measures but also by other variables like environmental or clinical variables. These variables could represent confounders and they should be considered explicitly when available. This problem can be addressed in a number of ways, for instance by integrating the connectivity matrices obtained from clinical variables. We plan to extend PINSPlus in the future to exploit clinical data whenever possible.

Nevertheless, PINSPlus is a fast and powerful software for subtype discovery. PINSPlus overwhelmingly outperforms established approaches in identifying known subtypes and discovering novel subgroups of patients with significant survival differences. The software is flexible enough to be applied in many areas to tackle unsupervised machine learning problems involving either single or multiple types of high-dimensional data.

*Conflict of Interest*: none declared.

**Table 1.** Cox $P$-values of subtypes discovered by PINSPlus, CC, SNF and iClusterPlus for 2 METABRIC breast cancer datasets and 34 TCGA datasets

| Dataset | Size | PINS+ | CC | SNF | iCluster+ |
|---|---|---|---|---|---|
| METABRIC | | | | | |
| 1. Discovery | 997 | 1.8e − 9 | 0.022 | 2.3e − 5 | 0.378 |
| 2. Validation | 983 | 3.4e − 5 | 0.096 | 0.010 | 0.031 |
| TCGA | | | | | |
| 1. KIRC | 124 | 6e − 5 | 0.118 | 0.691 | 0.058 |
| 2. GBM | 273 | 8.7e − 5 | 0.014 | 0.021 | 0.103 |
| 3. LAML | 164 | 8.7e − 4 | 0.292 | 0.002 | 0.083 |
| 4. LUSC | 110 | 0.008 | 0.688 | 0.087 | 0.224 |
| 5. BLCA | 404 | 0.019 | 0.089 | 0.109 | 0.17 |
| 6. HNSC | 228 | 0.046 | 0.428 | 0.366 | 0.364 |
| 7. LIHC | 366 | 0.03 | 0.622 | 0.334 | 0.072 |
| 8. STAD | 362 | 0.002 | 0.428 | 0.041 | 0.434 |
| 9. THYM | 119 | 0.013 | 0.139 | 0.097 | 0.24 |
| 10. GBMLGG | 510 | 7.5e − 17 | 5.2e − 4 | 4.8e − 14 | 5.4e − 14 |
| 11. LGG | 510 | 7.7e − 25 | 2e − 6 | 1.6e − 14 | 2.7e − 14 |
| 12. PAAD | 178 | 2.5e − 4 | 0.013 | 7.4e − 4 | 6.3e − 4 |
| 13. SKCM | 439 | 0.048 | 0.604 | 0.478 | 0.108 |
| 14. COADREAD | 294 | 0.003 | 0.946 | 0.66 | 0.178 |
| 15. UCEC | 234 | 0.001 | 0.105 | 0.018 | 0.619 |
| 16. CESC | 304 | 0.03 | 0.376 | 0.51 | 0.201 |
| 17. COAD | 220 | 0.001 | 0.419 | 0.128 | 0.884 |
| 18. BRCA | 622 | 0.007 | 0.008 | 0.119 | 0.014 |
| 19. STES | 545 | 0.007 | 0.301 | 0.157 | 0.46 |
| 20. KIRP | 271 | 1.1e − 9 | 0.367 | 0.005 | 0.013 |
| 21. KICH | 65 | 0.028 | 0.955 | 0.701 | 0.788 |
| 22. UVM | 80 | 7.5e − 4 | 0.005 | 1.7e − 4 | 0.003 |
| 23. ACC | 79 | 0.007 | 0.014 | 4.3e − 5 | 7.1e − 4 |
| 24. SARC | 257 | 0.03 | 0.148 | 0.044 | 4e − 4 |
| 25. MESO | 86 | 7.3e − 4 | 0.272 | 4.2e − 4 | 2.2e − 4 |
| 26. READ | 74 | 0.649 | 0.737 | 0.762 | 0.249 |
| 27. UCS | 56 | 0.458 | 0.207 | 0.859 | 0.983 |
| 28. OV | 286 | 0.319 | 0.859 | 0.445 | 0.062 |
| 29. ESCA | 183 | 0.33 | 0.791 | 0.392 | 0.16 |
| 30. PCPG | 179 | 0.866 | 0.938 | 0.332 | 0.55 |
| 31. LUAD | 428 | 0.099 | 0.926 | 0.501 | 0.118 |
| 32. PRAD | 493 | 0.349 | 0.638 | 0.475 | 0.879 |
| 33. THCA | 499 | 0.166 | 0.64 | 0.62 | 0.111 |
| 34. TGCT | 134 | 0.531 | 0.758 | 0.838 | 0.58 |

*Notes*: Cells highlighted in yellow have significant Cox $P$-values at the threshold of 5%. Cells highlighted in green have the most significant Cox $P$-value. PINSPlus substantially outperforms the other methods in identifying subtypes with significant survival differences. (Color version of this table is available at *Bioinformatics* online.).

## References

Esserman,L. *et al.* (2009) Rethinking screening for breast cancer and prostate cancer. *J. Am. Med. Assoc.*, **302**, 1685–1692.

Mo,Q. *et al*. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA*, **110**, 4245–4250.

Monti,S. *et al*. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learn.*, **52**, 91–118.

Nguyen,T. (2017) Horizontal and vertical integration of bio-molecular data. PhD Thesis, Wayne State University, Detroit, Michigan, USA.

Nguyen,T. *et al*. (2017) A novel approach for data integration and disease subtyping. *Genome Res.*, **27**, 2025–2039.

Uramoto,H. and Tanaka,F. (2014) Recurrence after surgery in patients with NSCLC. *Trans. Lung Cancer Res.*, **3**, 242–249.

Wang,B. *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.