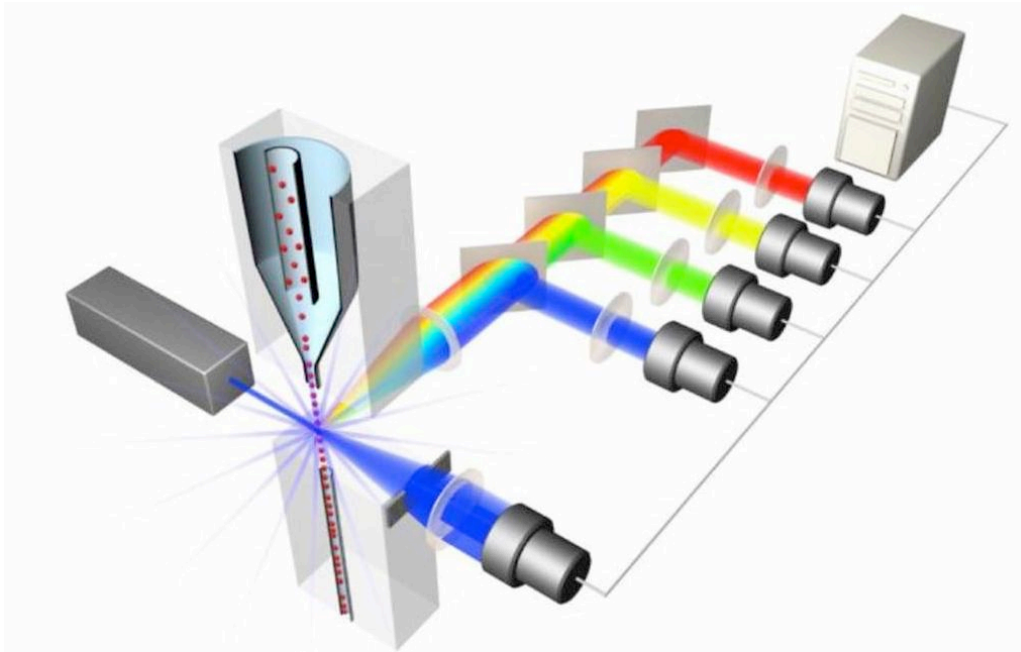


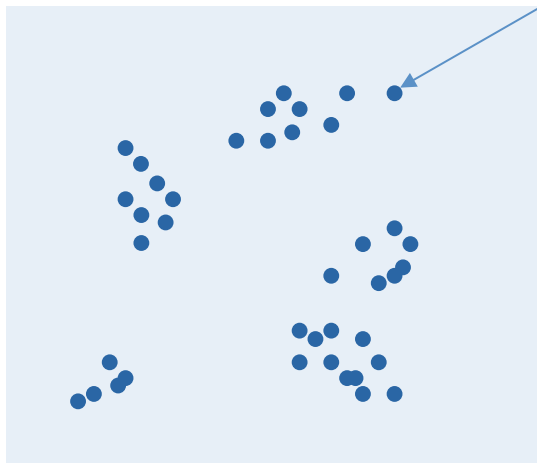
# Identifying Leukemic Cell Populations in Comparative Flow Cytometry



Ariful Azad  
Johannes Langguth  
Youhan Fang  
Alan Qi  
Alex Pothen

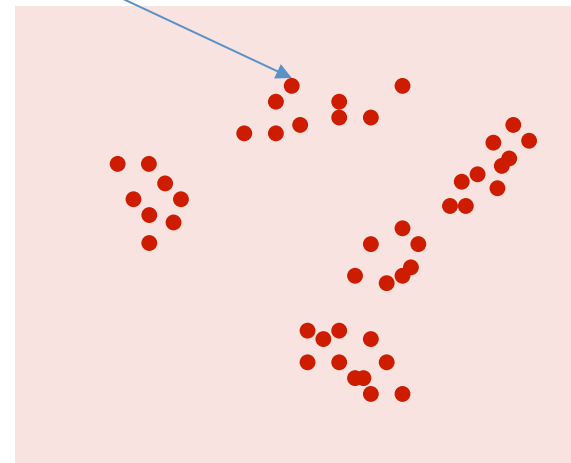
# Problem description

Each point represents 6/7 dimensional data of a single cell, obtained from a flow cytometry experiment.



Leukemic sample

Oncogene PML-RAR $\alpha$   
expressed and Leukemia  
developed in mice

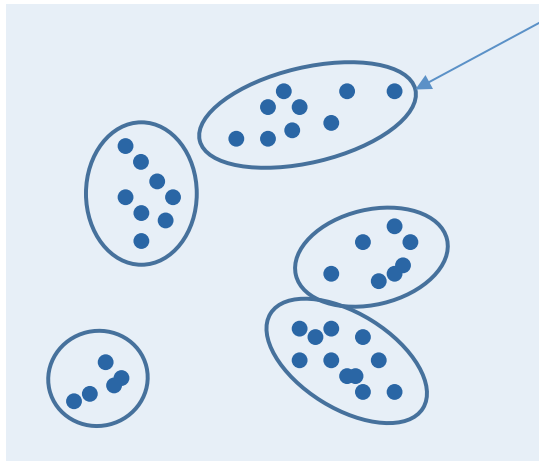


Wild Type (non-leukemic) sample

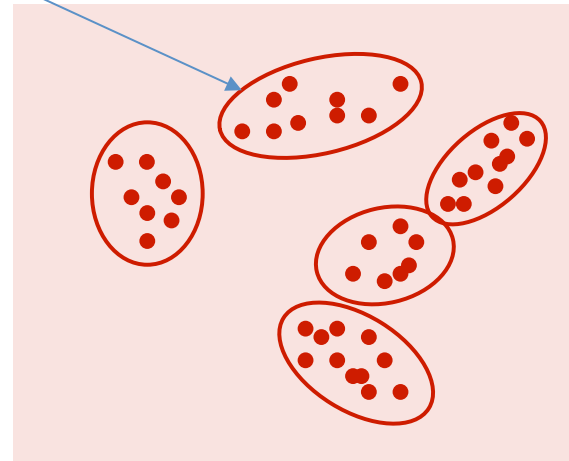
Oncogene PML-RAR $\alpha$  not  
expressed

# Problem description

To facilitate comparison across samples, cells are first clustered in each sample. Each cluster is represented by its mean and covariance matrix.



Leukemic sample

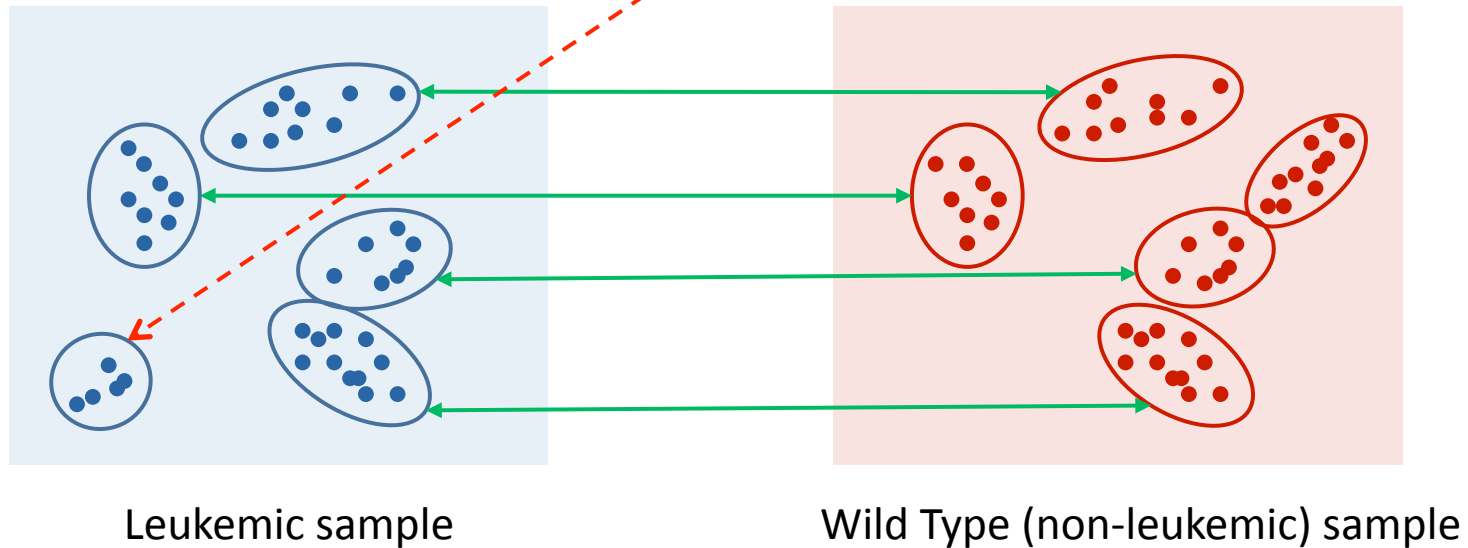


Wild Type (non-leukemic) sample

# Objective - I

## Identifying Leukemic Clusters by pairwise comparison

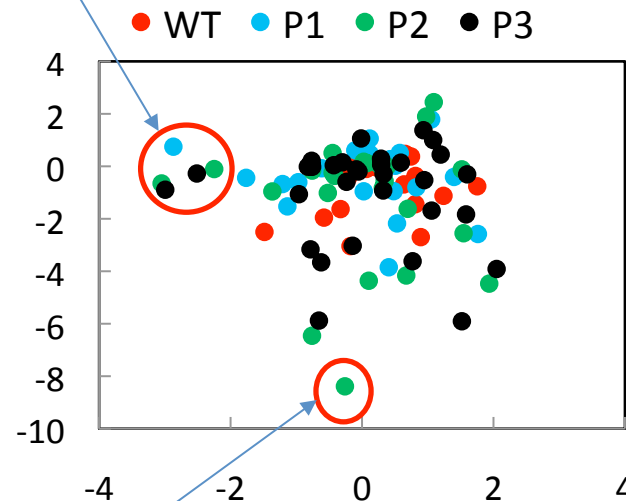
Corresponding clusters are matched based on a similarity measure (KL divergence). Unmatched clusters are characteristic of **Leukemia**.



## Objective - II

### Identifying Leukemic Clusters by group comparison

From multiple Leukemic samples identify clusters similar to each other but dissimilar to any WT cluster => **common Leukemic clusters**.



If a cluster in a Leukemic sample is dissimilar to clusters in any other Leukemic sample as well as to any WT cluster => **distinctive Leukemic cluster**.

# Previous work and our contribution

- Most works on gating/clustering not downstream analysis.
- The closest work is FLAME by Pyne et. al. (2009)
  - Clustering with skew-t distribution (parametric)
  - Formation of meta-clusters using Partition Around Medoids (PAM)
  - Matching cluster in each sample to meta-clusters using weighted b-matching with an integer programming formulation.
- Our contributions
  - Use of non-parametric clustering – automatic and fast.
  - Developed a **Generalized Edge Cover** formulation to identify local diseased clusters by comparing a Leukemic sample with WT .
  - Defined three metrics and used Branch and Bound algorithm to identify Leukemic clusters from groups.
  - Also developed statistical metrics to assess the affinity of a cluster to a group which allows **soft** membership.

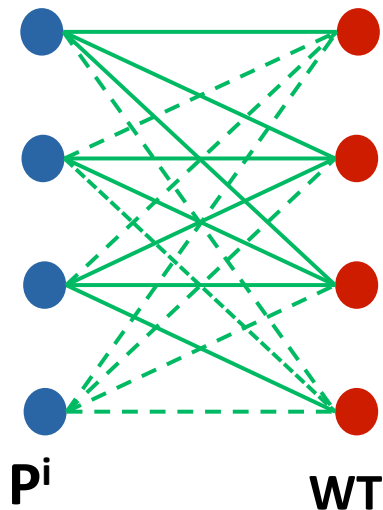
# Datasets

- Mouse bone marrow cell data obtained from Wojiski et al., Harvard Medical School.
- Each sample was clustered using Nonparametric **Dirichlet Process Mixture model** (DPM) .
  - Dissimilarity between a pair of clusters is calculated using KL-divergence (mutual information)

Dataset 1				Dataset 2			
Sample	D	Cells	Clusters	Sample	D	Cells	Clusters
WT	6	115k	18	WT	7	49k	21
Pre-leukemic	6	132k	23	Pre-leukemic	7	69k	22
Leukemic-1	6	107k	22	Leukemic-1	7	78k	21
Leukemic-2	6	132k	28	Leukemic-2	7	6k	12
Leukemic-3	6	236k	31	Leukemic-3	7	49k	21

# Pairwise comparison: Generalized Edge Cover in a Bipartite Graph

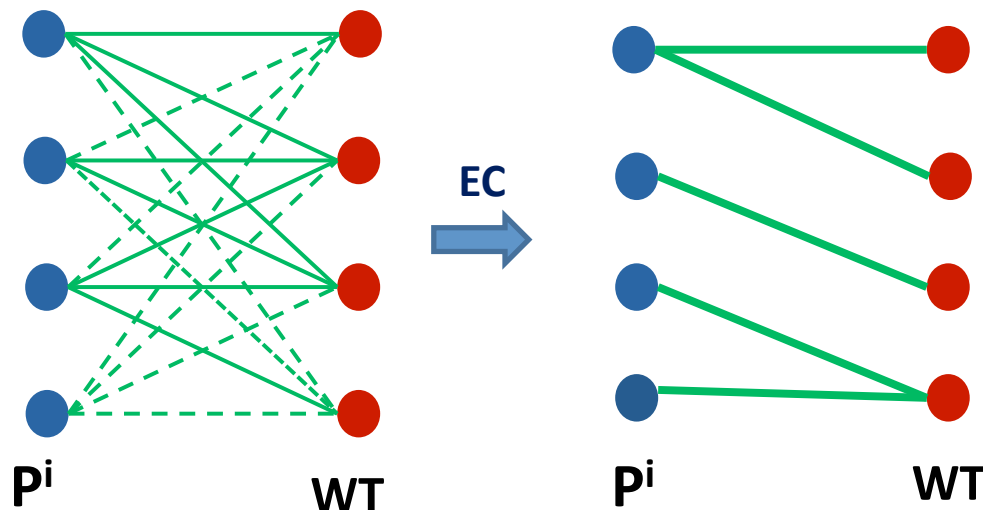
- Construct a complete bipartite graph with clusters from a Leukemic sample on one side and from WT on the other side.
- Edge weights are KL-divergence between corresponding clusters.
- Edges with large weights are shown in dashed line.



# Pairwise comparison: Generalized Edge Cover in Bipartite Graph

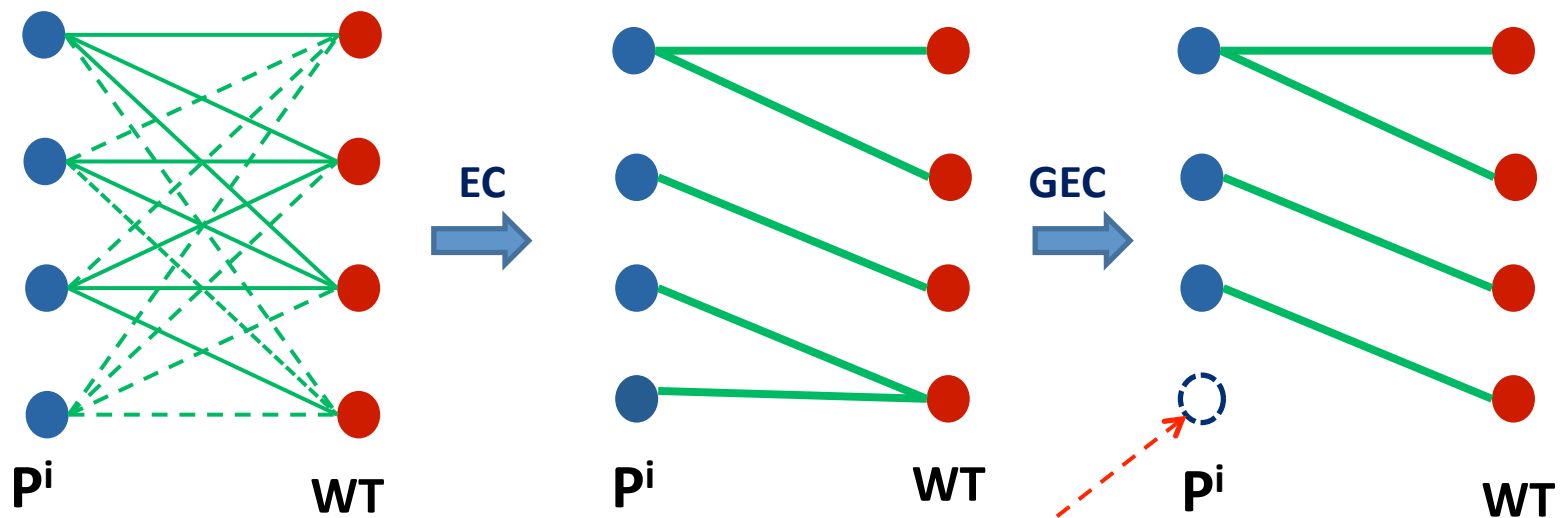
- A **minimum-weight edge cover (EC)** is a subset of edges such that every vertex is incident to at least one edge of the set and summation of weight of edges in the set is minimum.

- Objective function: 
$$\min \left( \sum_{(v_i, v_j) \in EC} c_{ij} \right)$$



# Pairwise comparison: Generalized Edge Cover in Bipartite Graph

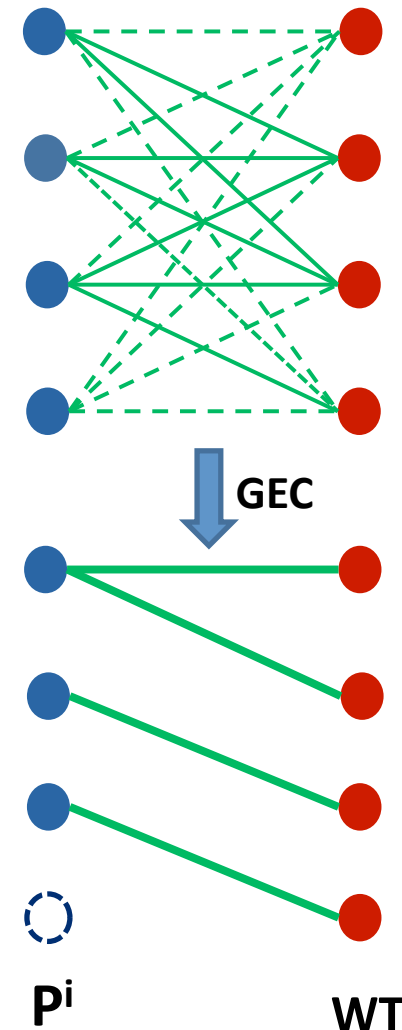
- A **generalized edge cover (GEC)** is an edge cover which allows few uncovered vertices at the cost of a penalty ( $\lambda$ ).
- Objective function: 
$$\min \left( \sum_{(v_i, v_j) \in EC} c_{ij} + \lambda * |V_{uc}| \right)$$
- Obtained from minimum-weight perfect matching in an intermediate graph (detail shown in the paper).



Leukemic cluster

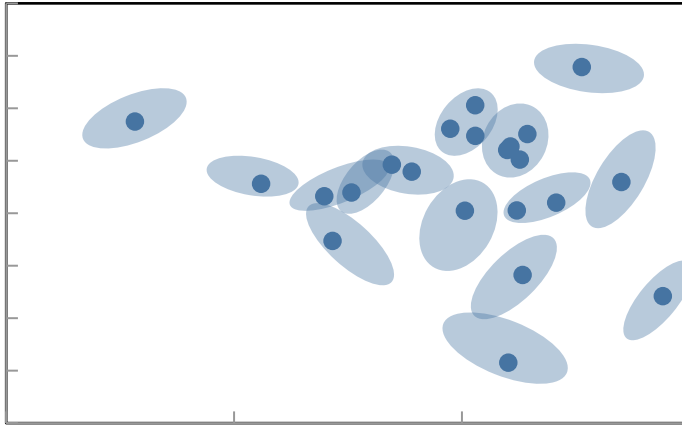
# Pairwise comparison: Generalized Edge Cover in Bipartite Graph

- **Significance:** Uncovered vertices (clusters) in a Leukemic sample do not have any close correspondence in WT and are **markers for leukemia**.

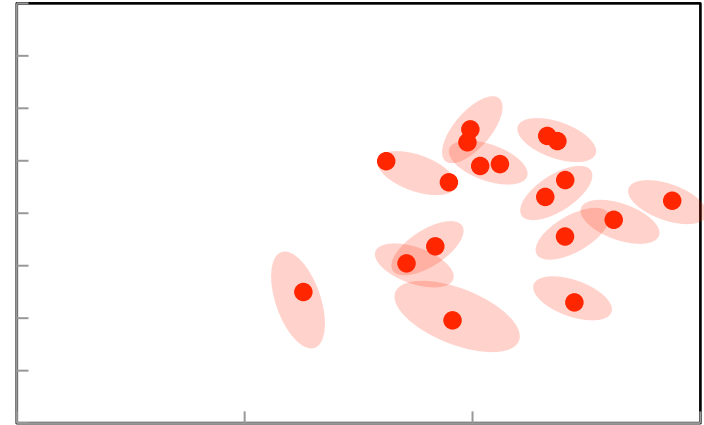


# Generalized edge cover Results

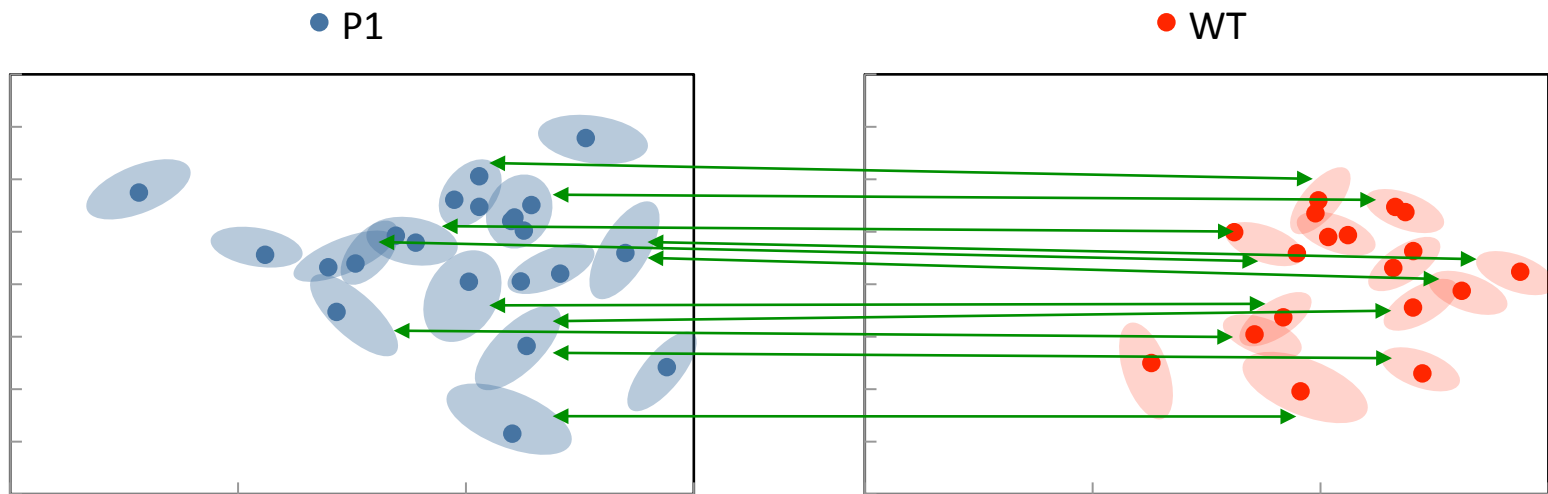
• P1



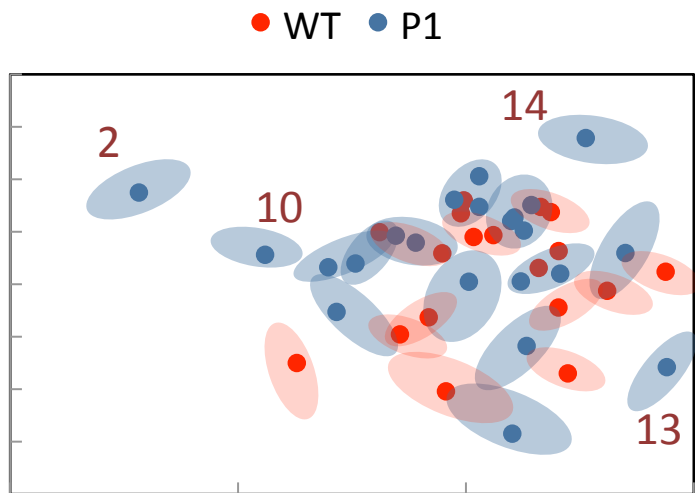
• WT



# Generalized edge cover Results



# Generalized edge cover Results and Limitations



Patient	Leukemic clusters
P <sup>1</sup>	2,10,13,14
P <sup>2</sup>	4,12,17,18,20,25,28,24
P <sup>3</sup>	11,12,13,16,17,18,19,29

## ■ Limitations

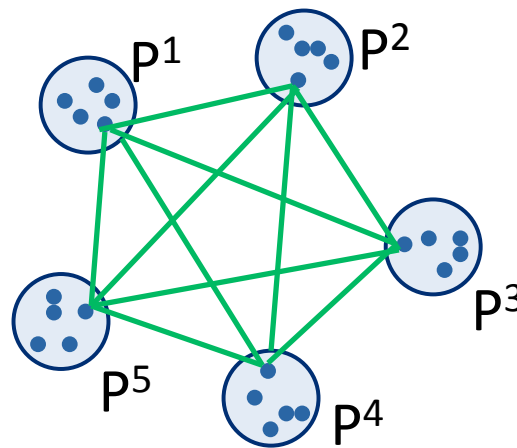
- **Bipartite Model:** Bipartite model can only identify Leukemic clusters in one sample.

# Identifying Global Leukemic Clusters

## Basic Definitions

### ■ Within Class Distance (WCD)

- For a leukemic cluster we select clusters from other Leukemic samples to form a group with minimum average distance.
- Average of the weights of the **green** edges.
- A group found with **small WCD** means, Leukemic clusters in the group are similar to each other.

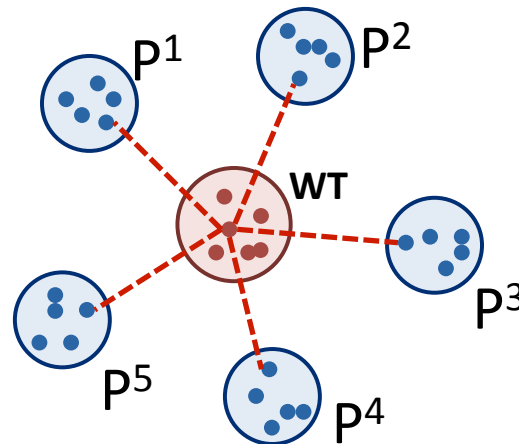


# Identifying Global Leukemic Clusters

## Basic Definitions

- **Between Class Distance (BCD)**

- For a group find the nearest WT with minimum average distance.
- Average of the weights of the broken red edges.
- Minimum BCD of a group is **large** means Leukemic clusters in the group are dissimilar from any WT cluster.



# Identifying Global Leukemic Clusters

## Distinctive and Common Leukemic Clusters

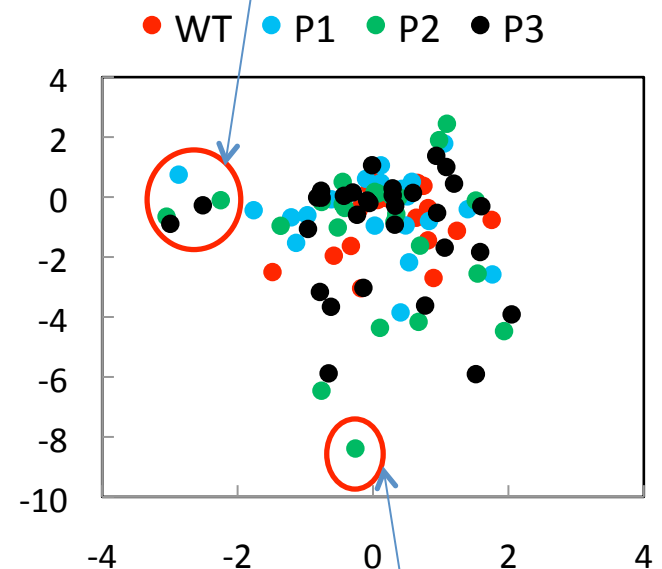
### ■ Distinctive Leukemic Cluster:

- A cluster in a single leukemic sample Dissimilar to any WT cluster as well as to clusters in other leukemic samples.
- Cluster fails to form a group with low value of WCD

### ■ Common Leukemic Clusters:

- A group of clusters from Leukemic samples similar to each other (**low WCD**) but dissimilar to any WT clusters (**high BCD**).
- Can be identified by high values of a comprehensive metric, Coherence Confidence (CC) .

Common Leukemic Cluster



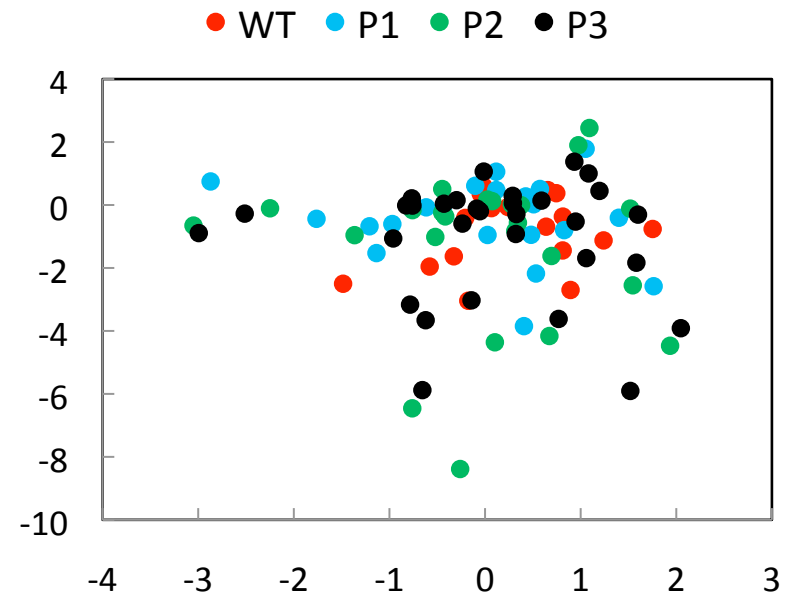
Distinctive Leukemic Cluster

$$CC(S) = \frac{BCD(S) - WCD(S)}{BCD(S) + WCD(S)} \left[ 1 - a^{-(BCD(S)) + WCD(S)} \right]$$

# Identifying Global Leukemic Clusters

## Results

P1	P2	P3	WCD	CC
<b>4</b>	<b>7</b>	<b>8</b>	1.40	.40
<b>11</b>	<b>13</b>	<b>14</b>	1.27	.39
18	<b>26</b>	<b>31</b>	3.05	.41
15	11	<b>21</b>	49.91	.08
2	<b>4</b>	18	296.7	.19
<b>17</b>	<b>21</b>	<b>17</b>	1.68	.64

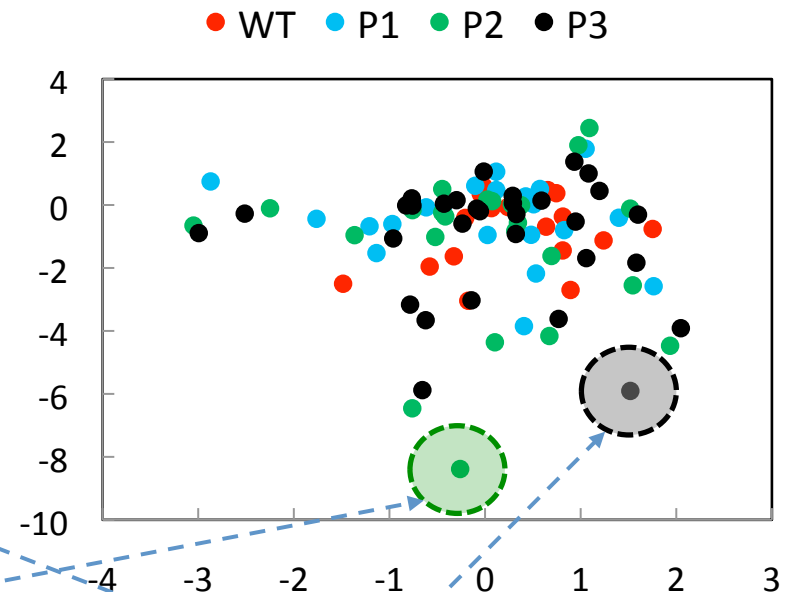


This is the compact group (minimum WCD) for the clusters shown in **blue**

# Identifying Global Leukemic Clusters

## Results

P1	P2	P3	WCD	CC
4	7	8	1.40	.40
11	13	14	1.27	.39
18	26	31	3.05	.41
15	11	21	49.91	.08
2	4	18	296.7	.19
17	21	17	1.68	.64

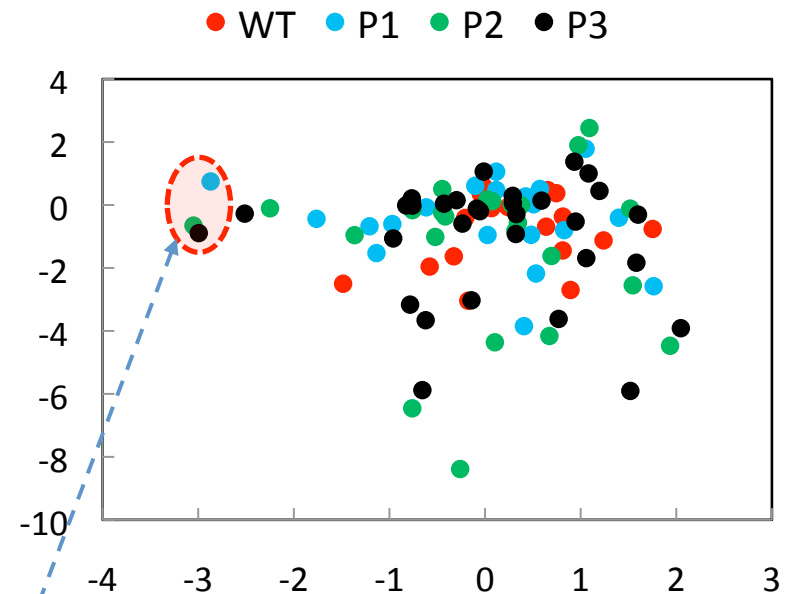


Distinctive Leukemic Clusters (high **WCD**)

# Identifying Global Leukemic Clusters

## Results

P1	P2	P3	WCD	CC
4	7	8	1.40	.40
11	13	14	1.27	.39
18	26	31	3.05	.41
15	11	21	49.91	.08
2	4	18	296.7	.19
17	21	17	1.68	.64

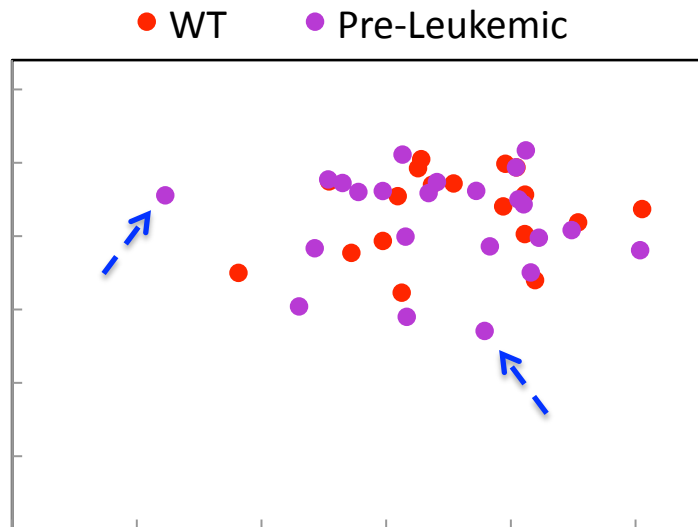


Common Leukemic Clusters (low WCD, high BCD => **high CC**)

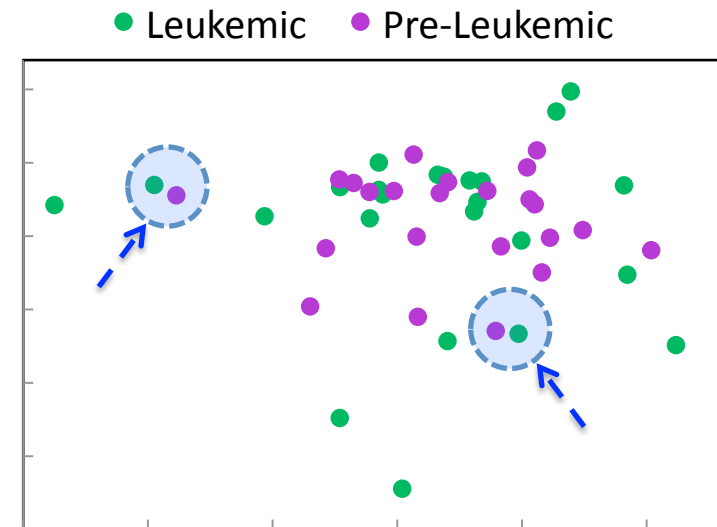
# Effect of Acute Promyelocytic Leukemia (APL) on bone marrow cells

- Wojiski et.al. (2009) reported datasets from flow cytometry experiments on bone marrow cells of mice.
- 3 classes of samples
  - **Wild Type (WT)**: Oncogene PML-RAR $\alpha$  not expressed.
  - **Pre-Leukemic (H)**: Oncogene PML-RAR $\alpha$  expressed but has not developed APL yet.
  - **Leukemic (P<sup>i</sup>)**: Oncogene PML-RAR $\alpha$  expressed and developed APL.
- WT and Pre-Leukemic samples had similar cell populations of certain cell types (certain types of stem and progenitor cells).
- In leukemic mice (P), most of the stem and progenitor cell populations are reduced. However one specific progenitor cell type (GMPs) is increased, relative to the WT and Pre-Leukemic mice.

# Effect of Acute Promyelocytic Leukemia (APL) on bone marrow cells



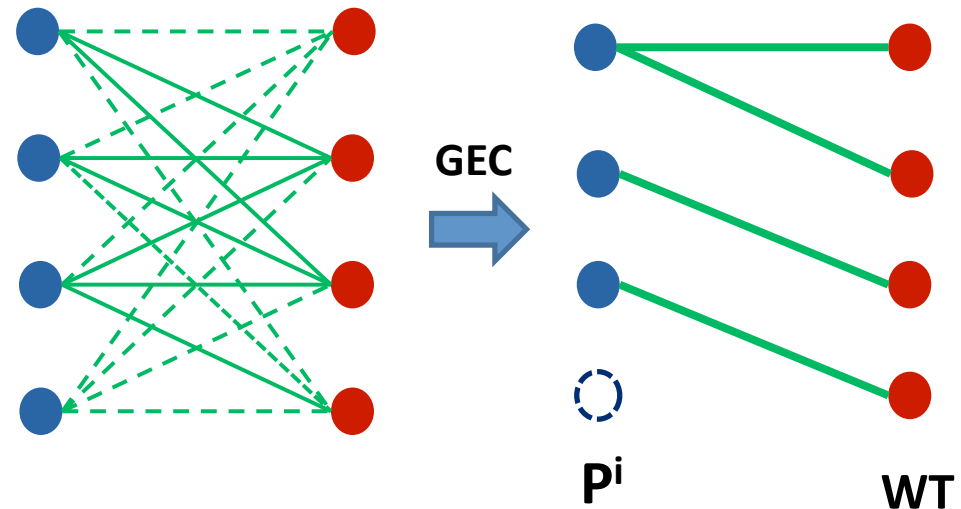
1. Pre-Leukemic clusters matched with WT clusters well, agreeing with Wojiski et.al. (2009).
2. But we find few Pre-Leukemic clusters are quite different.



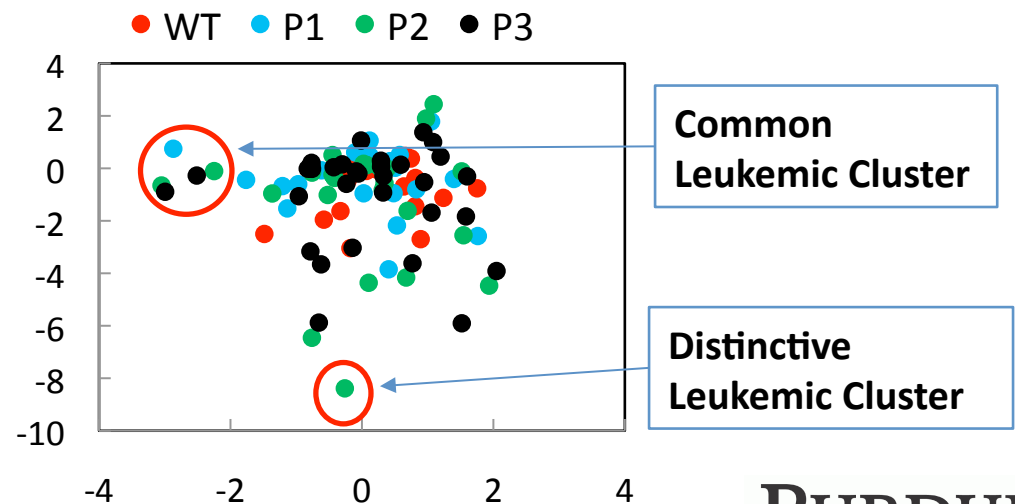
1. Generally pre-Leukemic clusters are closer to WT than to Leukemic clusters.
2. But a few pre-Leukemic clusters are closer to Leukemic clusters.
3. Significant for early detection of Leukemia

## Recap

- Identified Leukemic clusters using generalized edge cover.



- Identified Distinctive and Common Leukemic clusters using group formation with branch-and-bound.



Questions ?

# Supporting Slides

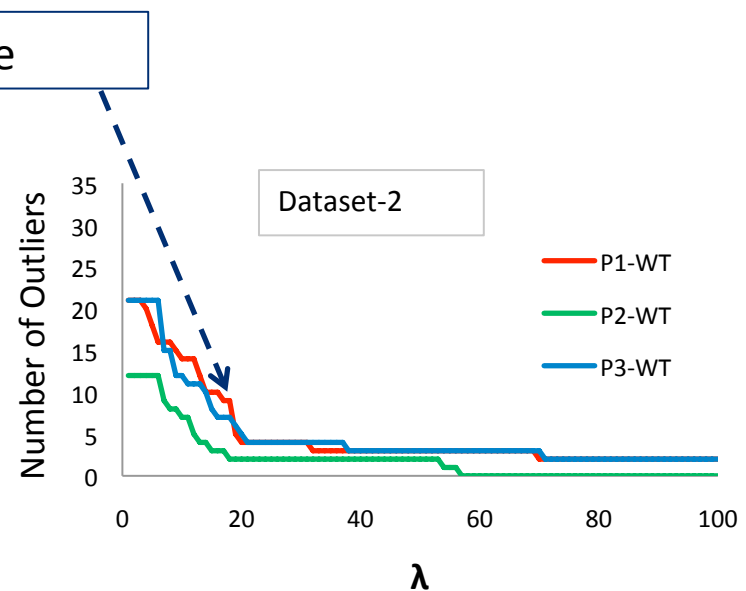
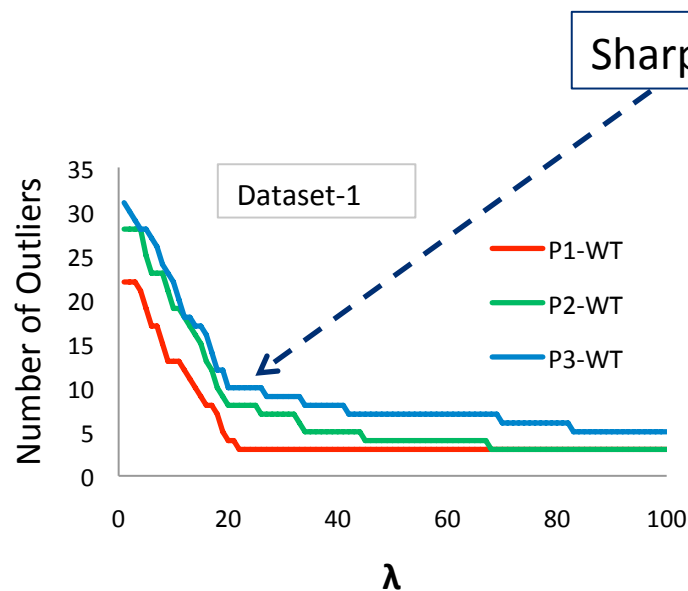
# Methods

- Clustering each sample individually
  - Dirichlet Process Mixture Model (DPM)
- Compare each Leukemic sample with the wild type sample to identify Local Leukemic clusters.
  - Generalized Edge Cover
- Compare all Leukemic samples in a dataset to WT to identify Global Leukemic clusters.
  - Branch and Bound heuristic Tree search
- Statistically justify the significance of Leukemic clusters
  - Permutation test
- Probabilistic model for group assignment

# Generalized edge cover

## Selection of cut-off value, $\lambda$

- Too large or too small cut-off value may underestimate or overestimate the number of Leukemic clusters.
- A break point in the curve is ideal choice.
- $\lambda=20$  is reasonable for current datasets.



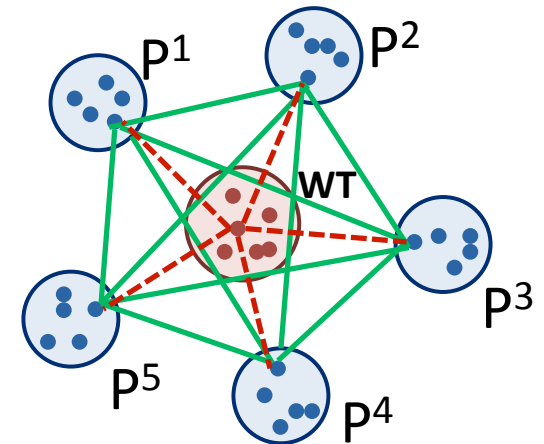
# Identify Global Leukemic Clusters

## Basic Definitions

- Given a set of clusters from Leukemic Samples,  $S = \{u_1, \dots, u_N\}$  where  $u_i$  is a cluster from  $P^i$  and cluster  $w$  belongs to WT

$$\text{Within Class Distance, } WCD(S) = \frac{2}{N(N-1)} \sum_{\substack{u_i, u_j \in S \\ i \neq j}} d(u_i, u_j)$$

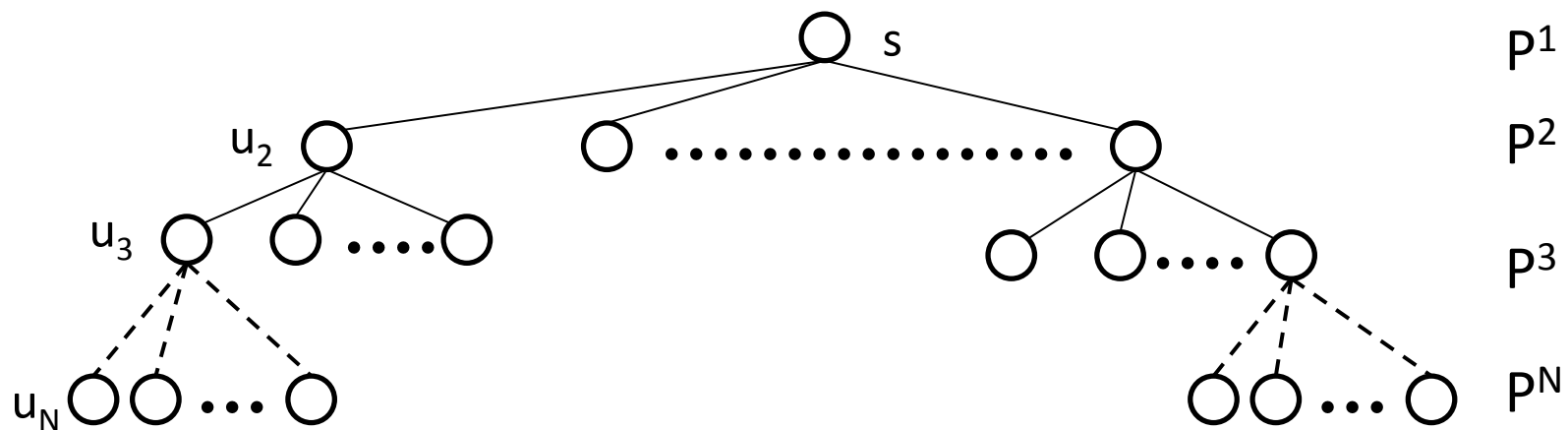
$$\text{Between Class Distance, } BCD(S) = \frac{1}{N} \min_{w \in WT} \left\{ \sum_{u_i \in S} d(w, u_i) \right\}$$



- WCD – average of the weight of the green edges.**
  - Small** : clusters in  $S$  are similar to each other
- BCD – average of the weight of the broken red edges.**
  - Large**: clusters in  $S$  are dissimilar from any WT cluster

# Group Construction

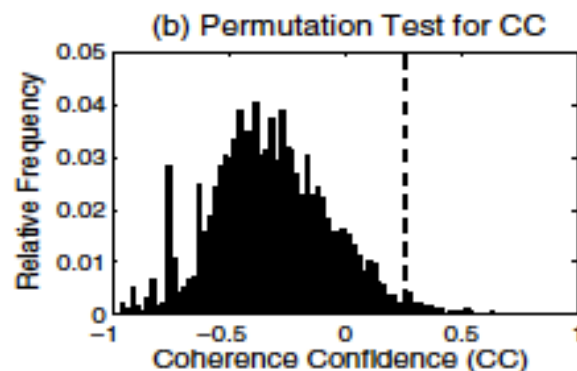
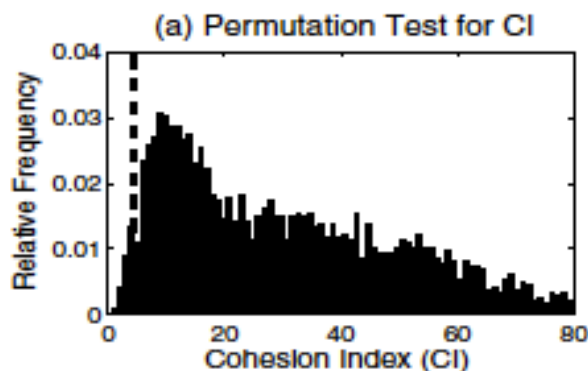
- Construct the most compact group for each cluster with lowest WCD score. (NP-hard problem)
- A greedy method need to visit all nodes in the tree. With  $N$  sample each with  $m$  clusters the time is  $O(m^{N-1})$ .
- Branch and Bound approach prunes lots of branches and find the optimal group reasonably fast.



# Justification of statistical significance

## Permutation test

- **Permutation Test:** Construct a group by picking one cluster from each patient randomly and calculate WCD and CC. repeating this process 1,00,000 times we can assess the chance of getting a specific value of WCD or CC.
- **Significance Level:** The values to the left of broken vertical line (WCD) or right to the vertical line (CC) have probability of less than .05 which can occur rarely by chance.

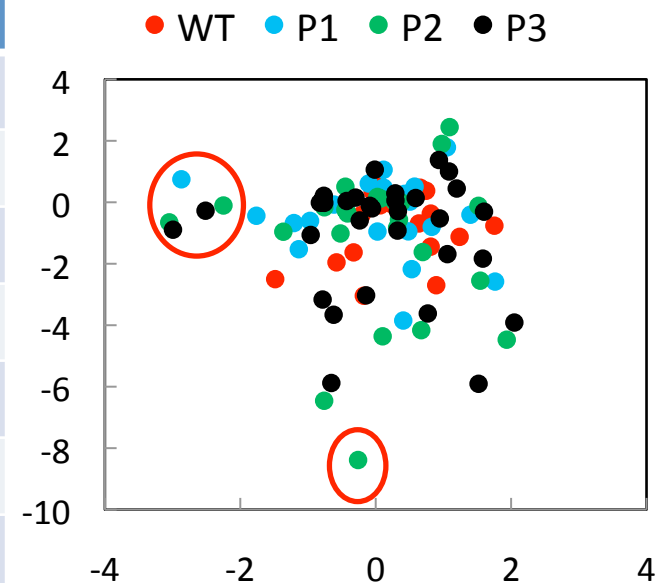


# Identifying Global Leukemic Clusters

## Results

Color	Meaning
Green	Starting cluster for a group (Seed)
Red	Distinctive Leukemic Clusters (high WCD)
Light Red	Common Leukemic Clusters (low WCD, high BCD => high CC)

P1	P2	P3	WCD	CC
4	7	8	1.40	.40
11	13	14	1.27	.39
18	26	31	3.05	.41
15	11	21	49.91	.08
10	26	18	154.6	.35
17	21	17	1.68	.64
9	5	3	1.76	.58



## Probabilistic model for group assignment

- Each cluster from every patient form the best group for itself.
- For N patients each with K clusters – NK groups.
- A cluster can appear multiple times across multiple groups.
- We can calculate the affinity of a cluster to a group by the following equation

$$P(u_i | S) = \frac{\textit{f frequency of being member in group, S}}{\textit{f frequency of being member in any group}} = \frac{\sum_{\substack{u_j \in S \\ i \neq j}} f_{ij}}{\sum_{i \neq j} f_{ij}}$$

- Hence the uniqueness of a group can be calculated:

$$\textit{Group uniqueness} = P(S) = \prod_{\substack{u_i \in S \\ 1 \leq i \leq N}} P(u_i | S)$$

# Results

## Group refinement

Color	Meaning					
	Member with strong support					
	Only strong member in the group – Distinctive Leukemic Clusters					
	Member with weak support (can be deleted)					

$u_1$	$P(u_1)$	$u_2$	$P(u_2)$	$u_3$	$P(u_3)$	$p(S)$
11	1	13	1	14	1	1
17	.29	16	1	6	1	.29
19	.83	22	.83	28	.33	.23
2	.33	4	1	18	.33	.11
21	1	6	.20	27	.25	.05
18	.17	19	.23	27	.25	.01

## Combined result with Pre-Leukemic(H) included

