



Data Science Seed
Grant Progress Report
Integrative
Hierarchical Clustering

Hongmei Zhang

Division of Epidemiology, Biostatistics, and
Environmental Health
School of Public Health



The objectives of the project

- To effectively **utilize the Data Science seed grant** to produce **preliminary results** on the proposed method to demonstrate the strength and **feasibility of the method** as well as the **capability of our research team.**



Motivation

- An R21 proposal submitted to NIH received a promising percentile of 22%.
 - The focus of the proposed project was to design statistical methods to learn the patterns revealed by omics data, including genetics, epigenetics, and gene expressions.
 - One of the methods proposed was integrative hybrid clustering.
 - Preliminary findings will substantially increase the feasibility of the proposed projects and demonstrate the research group's strong potential of collaboration.



The research team

- Hongmei Zhang (PI)
- Bernie Daigle (Co-I)
- Yu Jiang (Co-I)
- Jiasong Duan (Graduate Assistant)
- Liang Li (Graduate Assistant)



Goal and specific tasks

- Goal: to evaluate existing clustering methods (partitional and hierarchical clustering methods) and their ability to cluster variable vectors.
- Specific tasks (Green colored tasks have been finished):

Task 1. Design a distance metric.

Task 2. Design clustering algorithms.

Task 3. Simulation plans and simulation studies.

Task 4. Apply the method to GEO data sets



Achievement – the distance metric

- Proposed distance metric

$$d_{s1,s2} = \sum_{g=1}^G \left((p_1 \sum_{m=1}^{M_g} w_m D(DNAM_{gm,s1}, DNAM_{gm,s2}) + p_2 D(GE_{g,s1}, GE_{g,s2})) \times (\sum_{j=1}^{p_g} D(SNP_{gj,s1}, SNP_{gj,s2})) \right).$$

- This distance metric evaluates the distance between subjects 1 and 2 based on weighted gene-specific distances among CpG sites in DNA methylation and distance in expression of genes.
- The weight is determined by a weighted agreement (in terms of minor allele frequency) in SNPs between the two subjects. This weight has the potential to assess the joint activities of genetic and epigenetic factors.



Achievement – clustering algorithms

- In the simulation study, two clustering techniques were utilized
 - Partition around medoids (PAM)
 - Hierarchical clustering (divisive and agglomerative)
 - Hybrid clustering → to be designed
- Selection of a distance metric to be compared with the proposed metric
 - Gower distance – has the ability to deal with mixture of continuous and categorical variables as it allows different distance assessment depending on the continuation of a variable.

$$d_{ij} = \frac{\sum_{k=1}^p w_{ij,k} d_{ij}^{(k)}}{\sum_{k=1}^p w_{ij,k}}$$

Gower (1971) A general coefficient of similarity and some of its properties. Biometrics 27 857–874.



Achievement – simulation scenarios

- Assume 100 genes and each gene has one CpG site
- Each cluster has 100 subjects
- Two scenarios are included to guide our final simulation settings
 - Scenario 1: Regularity of DNA methylation (DNAm) on expression of genes (GE), methQTL, as well as interactions between SNPs
 - Three clusters
 - Cluster 1: low DNAm → high GE, SNP1=1 (dominant), SNP2=0 (recessive) (first 60 genes), and high DNAm → low GE with one SNP per gene (0/1) randomly generated for the remaining genes.
 - Cluster 2: low DNAm → high GE, SNP1=0, SNP2=1 (first 60 genes), and high DNAm → low GE with one SNP per gene (0/1) randomly generated for the remaining genes.
 - Cluster 3: high DNAm → low GE, SNP1=0, SNP2=0 (first 60 genes), and low DNAm → high GE, randomly generate one SNP per gene for the remaining genes.

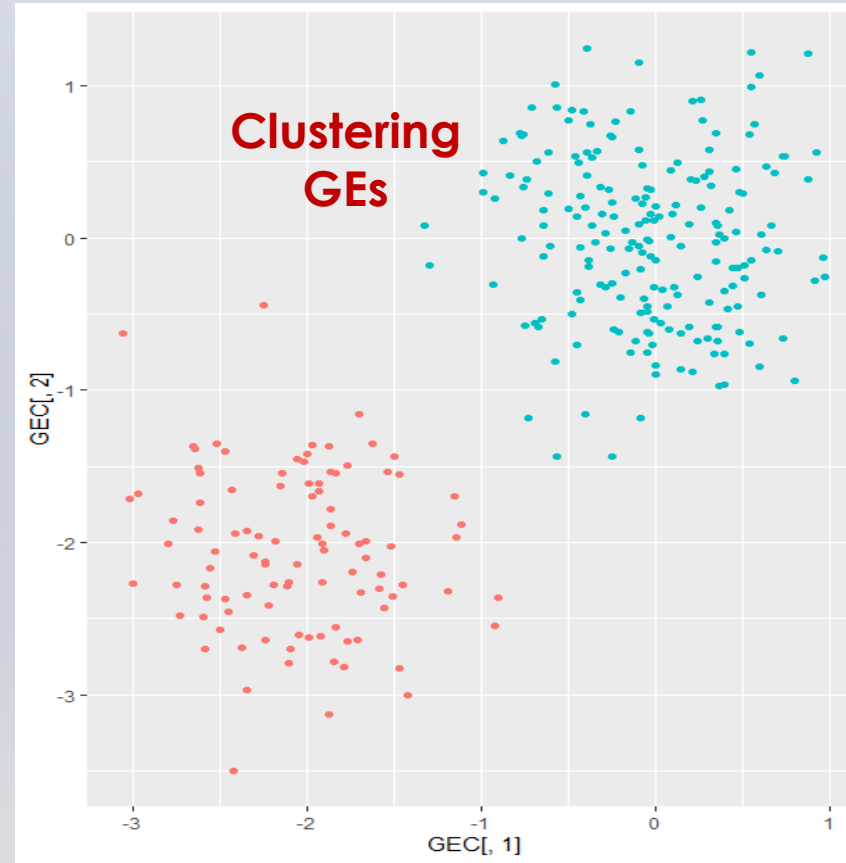
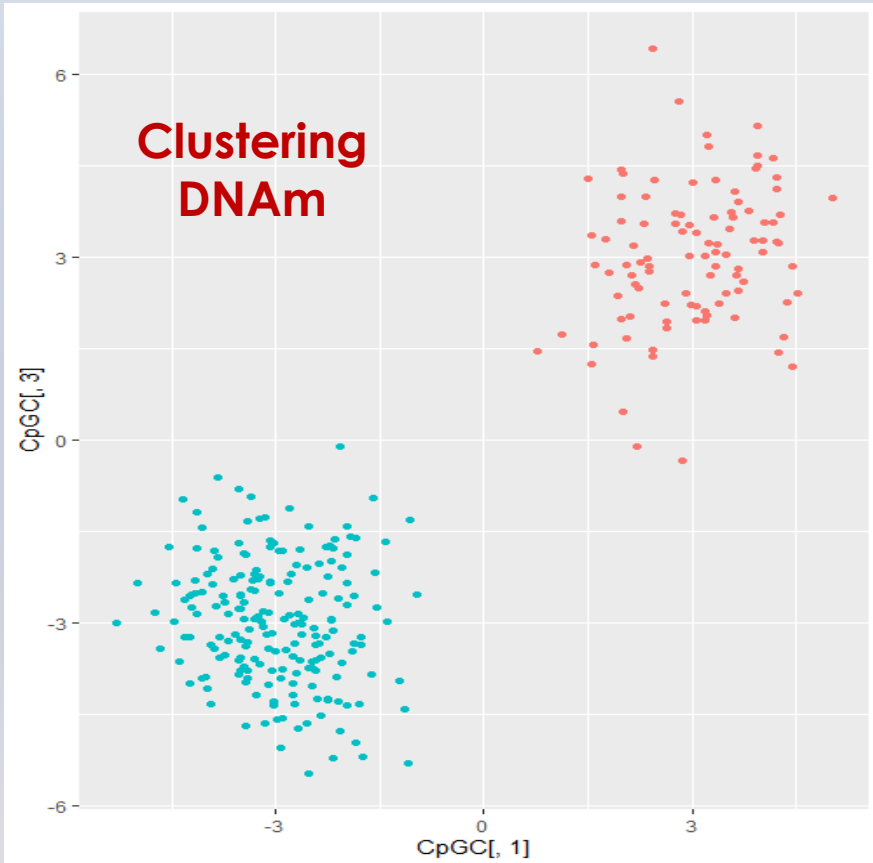


Achievement – simulation scenarios

- Two scenarios are considered (cont'd)
 - Scenario 2: Underlying clustering due to genetic differentiations
 - Two clusters
 - Cluster 1: first 50 with genotype 0, and the other 50 with genotype 1
 - Cluster 2: first 50 with genotype 1, and the remaining with genotype 2
 - In both clusters, GE and DNAm were generated from uniform distributions.

Achievement – simulation results (3 clusters)

- When clustering based on GEs only or DNAm only via K-means → two clusters

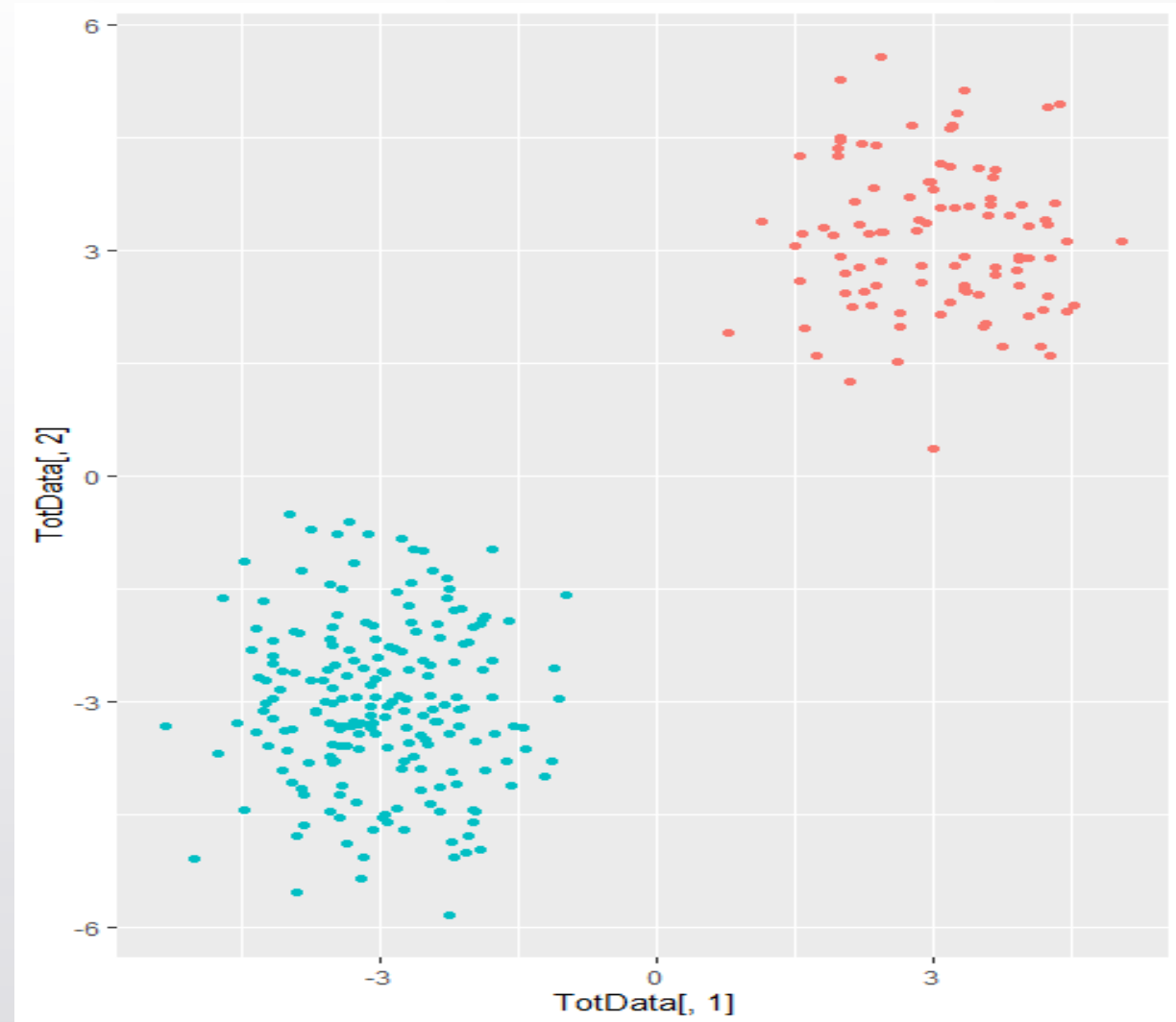


Possible reasons for the wrong results:

Difference between clusters was due to methQTL or SNPxSNP or SNP and DNA methylation interaction → would not be seen if only on expression or DNA methylation data.

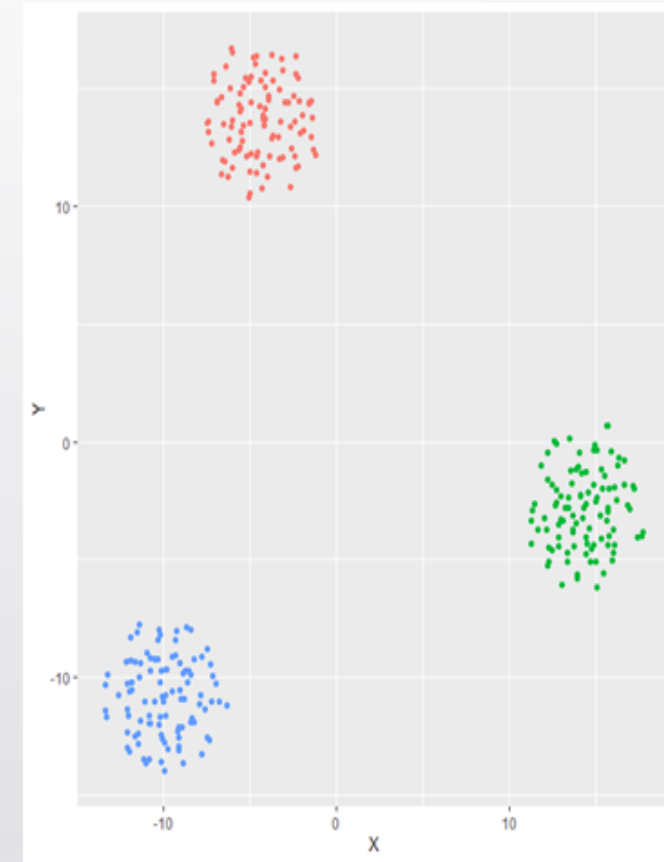
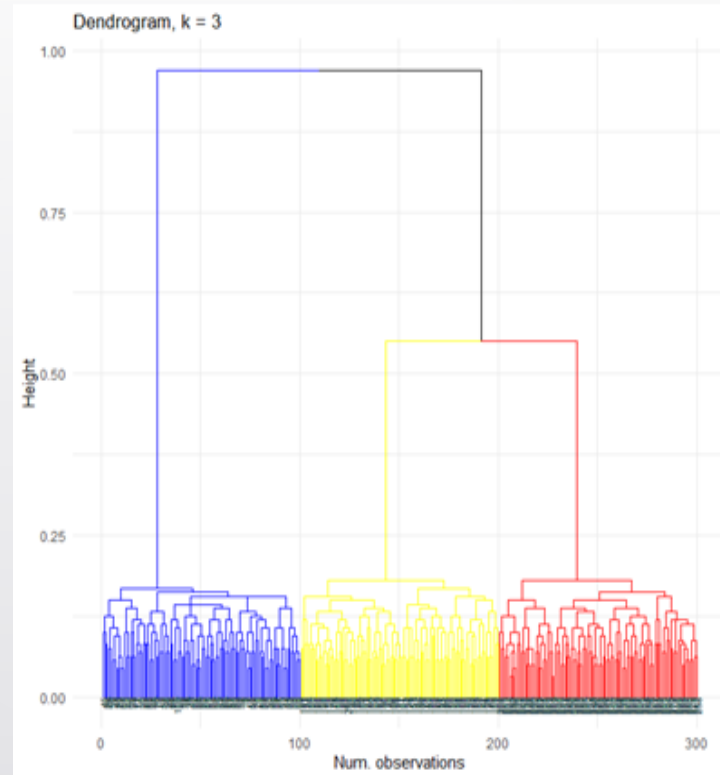
Achievement – simulation results (3 clusters)


- Cluster the three types of data integratively (SNPs, DNAm, and GE) via K-means → two clusters.
- **Reason for the wrong results:** We believe this was due to the distance metric implemented in the clustering (Euclidean).



Achievement – simulation results (3 clusters)

- Implement the Gower distance metric
- Cluster the three types of data integratively (SNPs, DNAm, and GE)
 - PAM
 - Hierarchical clustering
- All the three clusters are correctly identified.



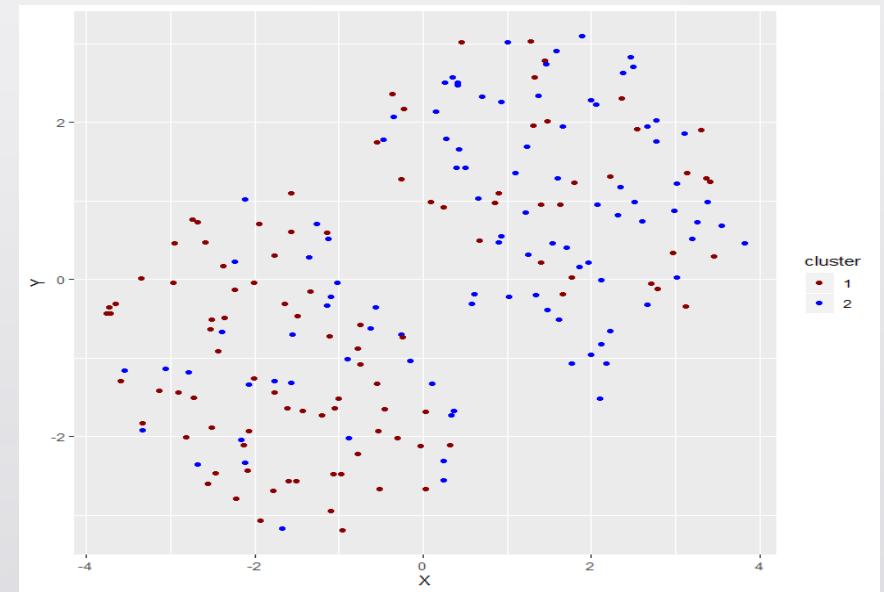
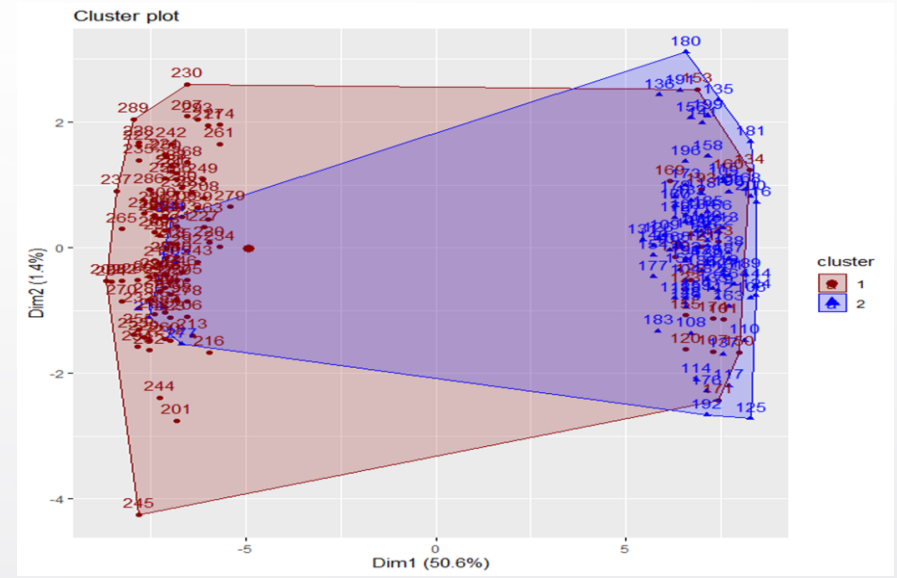


Achievement – simulation results (2 clusters) The need of a better distance metric

- It seemed the Gower distance metric is good enough to deal with these three types of data.
- Do we need any new metric?
- Further assessment using simulated data with 2 clusters (scenario 2).

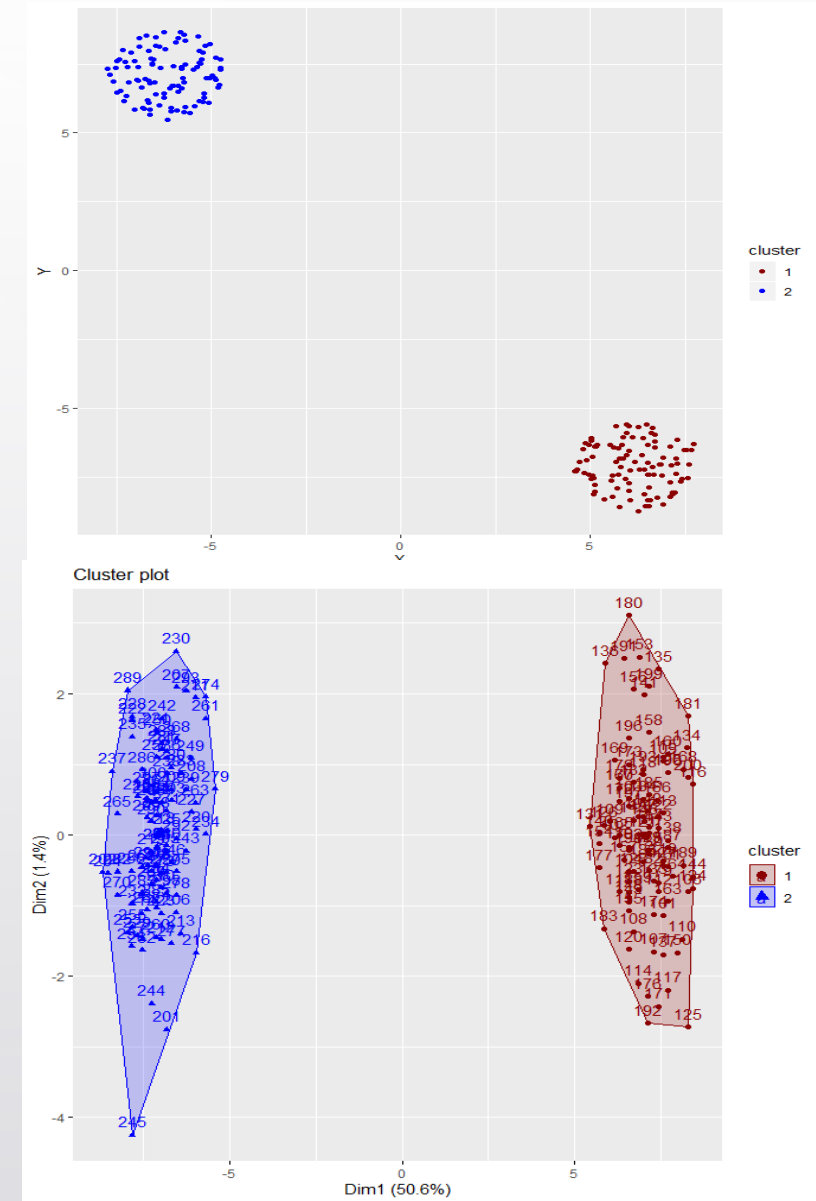
Achievement – simulation results (2 clusters) The need of a better distance metric

- Using Gower distance is not able to differentiate between the two clusters
- Clustering accuracy is not high
 - PAM (101, 99): 67.5%
 - Hierarchical clustering (113, 87): 89.5%
- Reason: the distance metric in Gower (Euclidean for continuous and agreement categorical, and they are additive) does not fit complex situations as in genetic and epigenetic studies.



///
 Achievement – simulation results (2 clusters)
 The need of a better distance metric

- Results from the proposed distance metric
 - All three clusters are correctly identified.
 - Reason: the proposed metric has the potential to capture interactions, methQTL, as well as frequencies of minor alleles.





Next steps

- Incorporate the results to the next submission of the proposal.
- Plan to write a short report and submit to a journal for publication.

**Truly appreciated the support
from this seed grant and
believe the preliminary results
will help future submissions
substantially.**