Published in final edited form as:
*Comput Methods Programs Biomed*. 2016 May ; 128: 69–74. doi:10.1016/j.cmpb.2016.02.011.

# *InterSIM*: Simulation tool for multiple integrative 'omic datasets'

**Prabhakar Chalise***, **Rama Raghavan**[1], and **Brooke L. Fridley**[2]
Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, United States

## Abstract

**Background and objective—**Integrative approaches for the study of biological systems have gained popularity in the realm of statistical genomics. For example, The Cancer Genome Atlas (TCGA) has applied integrative clustering methodologies to various cancer types to determine molecular subtypes within a given cancer histology. In order to adequately compare integrative or "systems-biology"-type methods, realistic and related datasets are needed to assess the methods. This involves simulating multiple types of 'omic data' with realistic correlation between features of the same type (e.g., gene expression for genes in a pathway) and across data types (e.g., "gene silencing" involving DNA methylation and gene expression).

**Methods—**We present the software application tool *InterSIM* for simulating multiple interrelated data types with realistic intra- and inter-relationships based on the DNA methylation, mRNA gene expression, and protein expression from the TCGA ovarian cancer study.

**Results—**The resulting simulated datasets can be used to assess and compare the operating characteristics of newly developed integrative bioinformatics methods to existing methods. Application of *InterSIM* is presented with an example of heatmaps of the simulated datasets.

**Conclusions—***InterSIM* allows researchers to evaluate and test new integrative methods with realistically simulated interrelated genomic datasets. The software tool *InterSIM* is implemented in R and is freely available from CRAN.

**Keywords**

Integrative; Simulation; Clustering; NMF

## 1. Introduction

Identification of molecular subtypes of cancer using high throughput molecular data has been frequently accomplished through the use of clustering [1,2]. Clustering involves the grouping of objects across a disjoint set of classes such that objects within the same class are more similar to one another as compared to the objects in different classes. A large number

**Corresponding author*. Tel.: +1 913 945 7987. pchalise@kumc.edu (P. Chalise).. [1] Tel.: +1 913 945 9412. rraghavan@kumc.edu (R. Raghavan). [2] Tel.: +1 913 945 5039. bfridley@kumc.edu (B.L. Fridley).

Conflict of interest
None declared.

Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cmpb.2016.02.011.

of clustering methods are available that use a single data type; such as, hierarchical, *k*-means [3], and non-negative matrix factorization (NMF) [4]. In addition to these methods, a few integrative clustering methods have been proposed that utilizes information from multiple data types collected on the same set of samples including: *iCluster* [5], integrative NMF [6], and mixture model based integrative clustering [7]. A summary of the above mentioned methods can be found in Chalise et al. [8]. However, in order to adequately assess such integrative methods, realistic and interrelated datasets are needed. *InterSIM* bridges this gap by simulating complex interrelated realistic genomic datasets.

Although clustering methods can be used to classify either genes or subjects, the proposed simulation tool focuses on clustering of subjects with the goal of identifying molecular subtypes of disease. In developing the simulation tool, we focused on generating three data types, DNA methylation, mRNA gene expression, and protein expression, on a set of samples with realistic correlation between and within data types. Here are a few examples of the types of relationships we included in the simulation of the data: CpG sites within the same CpG-island would have strong positive correlation, high methylation for a CpG-island upstream of a gene would result in lower mRNA expression or “gene silencing”, and higher mRNA gene expression is likely to result in higher downstream protein expression [9]. Such intra- and inter-feature relationships among the data types were based on real data collected on ovarian cancer tumors from The Cancer Genome Atlas (TCGA).

## 2. Methods

The simulation tool is based on three real datasets from the ovarian cancer study from TCGA – DNA methylation, mRNA gene expression, and protein expression data. In estimation of the relationship in this study, we restricted the tumors to 384 that were common across the three datasets. The level 3 methylation data consists of 27,578 CpG probes from 555 subjects measured using the Illumina 27K, level 3 mRNA gene expression data consists of 17,814 genes from 544 subjects measured with the Agilent G4502A platform, and level 3 RPPA protein expression data contains 187 probes from 412 subjects. Both the methylation and mRNA data were downloaded from https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp, and the protein data were downloaded from http://app1.bioinformatics.mdanderson.org/tcpa/design/basic/download.html. Using the CpG to gene annotation and protein to gene annotation information, 367 CpGs and 160 protein probes were found to map to 131 common genes. Based on these three data types measured on 384 subjects with the common mapped features, we estimated the intra- and inter-relationship between the features for use in the simulation of the realistic datasets, Fig. 1.

In simulating the data, we first consider the case where there are no clusters (i.e., only one cluster, $k = 1$, with effect = 0); and then the case where the number of clusters could vary from $k = 2$ to $K$, where $K$ is the user specified number of clusters. The number of clusters was determined by a handful of features in the various data types. That is, a set of features was selected to have a mean shift in their values so that they would be able to be distinguished among various subgroups or clusters. We start by simulating DNA methylation data, followed by mRNA gene expression, and finally protein expression. Details on the simulation of the data types and the correlation structures are outlined in the following sections.

### 2.1. Methylation data

The methylation *β*-values at a CpG site, *j* is the proportion of methylation ranging from 0.0 to 1.0 which are assumed to follow a beta distribution. The logit transformation of such *β*-values, denoted as *M*-values, then range from $-\infty$ to $\infty$ and can be assumed to follow a Gaussian distribution. The methylation data is then generated from the multivariate normal distribution which uses the covariance structure from the real data,

$$X_i \sim N\left(\mu_{i \times p_d}, \sum\nolimits_{p_d \times p_d}\right),$$

where $p_d$ and $\Sigma_{p_d \times p_d}$ are the number of CpGs and the covariance matrix of the real methylation data, respectively. Here, $\mu_{i \times p_d}$ is the effect size with the cluster mean shift as given by

$$\mu_{i \times p_d} = \mu_d + DMP \times \delta_{DMP},$$

where $\mu_d$ is the vector of mean of the *M*-values at each CpG site, *DMP* is the indicator variable representing differentially methylated CpGs, and $\delta_{DMP}$ is effect size the of the cluster mean shift. The *M*-values are then inverse logit transformed to *β*-values.

### 2.2. Gene expression data

Using the ovarian cancer study, gene-level summaries of the methylation *β*-values were computed for each sample. For each tumor sample, the CpG probes that map to gene were grouped together and the median values of each group of CpGs were computed. Then *M*-values were computed using logit transformation. The Pearson correlation coefficients between the *M*-value and mRNA gene expression for each gene were then computed for using in the simulation.

The gene expression data is generated using the multivariate normal distribution and corresponding covariance structure from the real mRNA gene expression data,

$$Y_i \sim N\left(\mu_{i \times p_g}, \sum\nolimits_{p_g \times p_g}\right),$$

where $p_{\text{gene}}$ and $\Sigma_{p_g \times p_g}$ are the number of genes and the covariance matrix of the real gene expression data, respectively. The effect size $\mu_{i \times p_g}$ is computed as follows:

$\mu_{i \times p_g} = \left(\rho \times \mu_M + \sqrt{1-\rho^2} \times \mu_g\right) + DEG \times \delta_{DEG}$, where $\rho$ is the Pearson correlation coefficient between the *M*-values and the corresponding mRNA gene expression for each gene, $\mu_M$ is a vector of mean *M*-values (gene level), and $\mu_g$ is a vector of mean mRNA gene expression values. Furthermore, in computing the effect size to use in the simulation of the data, *DEG* is an indicator variable representing the differentially expressed genes that are related to the differentially expressed CpGs generated as mentioned above, and $\delta_{DEG}$ is the cluster mean shift. Though, it should be noted that if there is no correlation for a particular gene, then the effect size involves only the mean of the gene expression values.

### 2.3. Protein expression data

Elevated gene expression is likely to result in increased downstream protein expression. Such relationships present in the real datasets are utilized in simulating the protein expression data by computing Pearson correlation coefficients between the expression of each protein and the corresponding gene. The protein expression data are generated using a multivariate normal distribution,

$$Z_i \sim N\left(\mu_{i\times p_p}, \sum\nolimits_{p_p\times p_p}\right),$$

where $p_p$ and $\Sigma_{p_p\times p_p}$ are the number of proteins and the covariance matrix estimated from the real data, respectively. The effect size $\mu_{i\times p_p}$ is computed as follows

$$\mu_{i\times p_p} = \left(\rho\times\mu_g + \sqrt{1-\rho^2}\times\mu_p\right) + DEP\times\delta_{DEP},$$

where $\rho$ is the Pearson correlation coefficient between the protein expression and corresponding gene expression, $\mu_p$ and $\mu_g$ are the mean of protein expression and the corresponding mapped gene expression respectively from the real data. The vector *DEP*, representing differentially expressed proteins that are related to the differentially expressed genes, is a generated indicator variable as mentioned above, and $\delta_{DEP}$ is the cluster mean shift. Similarly to the gene expression data simulation, the effect size only involves the mean of the protein expression when there is no correlation between a particular protein–gene pair.

## 3. Implementation and results

The function has flexibility in specifying the number of samples, effect sizes for each of the three data types, proportion of differentially expressed features that results in samples to form clusters, and proportion of samples in each cluster. The function determines the number of clusters based on how many sample proportions are assigned. The generated data can be visualized using any standard heatmap function. Users can specify whether they want to generate an image plot at the end of the simulation, and what type of image plot – having a clustered pattern or raw data image only (an example with description has been provided with *InterSIM* package documentation). By default, the function utilizes the covariance structures (intra-), correlation (inter-) and the other mapping information from the TCGA data on ovarian cancer. However, users can specify their own covariance matrix for each data type and correlation between the data types. In addition, users can utilize all the mapping information and relevant data from other cancer (or any other disease) types in the function. In order to do that relevant data should be pre-computed as mentioned in the supplementary materials.

The performance of the simulation method was tested using hierarchical and k-means clustering methods designed for a single dataset; as well as, model based integrative clustering method [5] followed by calculating the adjusted rand index (results not shown here). The simulation works well resulting in a higher value of the rand index (measure of agreement between the true clusters and the predicted clusters using any clustering method) when the larger cluster effect size is used and vice versa. The function takes only around 0.5 s to complete a single simulation run on a desktop computer with an Intel i7-2600 CPU 3.40 GHz processor. The heatmaps given in Figs. 2 and 3 were generated using a heatmap function in R package NMF [10]. Fig. 2 shows the comparison between the original and simulated data without having any clusters. Both the density plots and heatmaps by data type show that the distributional properties of the features and the inherent patterns in the data are closely mimicked in the simulated data. Spearman correlation coefficient for each feature (pairwise) and canonical correlation (multivariate correlation) analysis were performed between the original and simulated data in order to assess how closely the simulated data matches with original data. The spearman correlation ranged from 0.8282 to 0.9421 in methylation data, 0.7569 to 0.8914 in mRNA data and 0.7828 to 0.9131 in the protein data. The canonical correlations for the methylation, mRNA and protein data were 1.0, 0.97 and 0.98 respectively. Fig. 3 represents an example of comparison of two sets of simulated data with and without having clusters. Both of the principal components plot and heatmaps by data type show that the function generates data set having desired cluster patterns provided appropriate effect size.

## 4. Conclusion

In order to adequately compare integrative or “systems-biology”-type methods, realistic and related data sets are essential to assess the methods. The goal of this paper is to present, describe and illustrate the software tool we developed to generate multiple types of ‘omics data’ with realistic intraand inter-relationships based on real data. The performance of several available clustering methods on these data suggests that the datasets will be well suited for assessing new clustering methods, and evaluating the relative performances of the existing clustering methods. We hope the availability of such functions will help researchers in developing new integrative methods.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

*Funding*: This research was supported in part by the National Institute of Health (R21 CA182715) and University of Kansas Cancer Center (P30 CA168524). *Data*: TCGA ovarian cancer study.

## REFERENCES

1. Sorlie T, et al. Gene expression patterns of breast cancer carcinomas distinguish tumor subclasses with clinical implications. PNAS. 2001; 98:10869–10874. [PubMed: 11553815]

2. Verhaak RG, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]
3. Hastie, et al. The Elements of Statistical Learning. Springer; New York: 2001.
4. Brunet JP, et al. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:4164–4169. [PubMed: 15016911]
5. Shen R, et al. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast cancer subtype analysis. Bioinformatics. 2009; 25:2906–2912. [PubMed: 19759197]
6. Zhang S, et al. Discovery of multidimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012; 40:9379–9391. [PubMed: 22879375]
7. Kormaksson M, et al. Integrative model based clustering of microarray methylation and expression data. Ann. Appl. Stat. 2012; 6:1327–1347.
8. Chalise P, et al. Integrative clustering methods for high dimensional molecular data. Transl. Cancer Res. 2014; 3:202–216. [PubMed: 25243110]
9. Gry, M., et al. Correlations between RNA and Protein expression profiles in 23 human cell lines. BMC Med. Genomics. 2009. http://dx.doi.org/10.1186/1471-2164-10-365
10. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinform. 2010; 11:367.

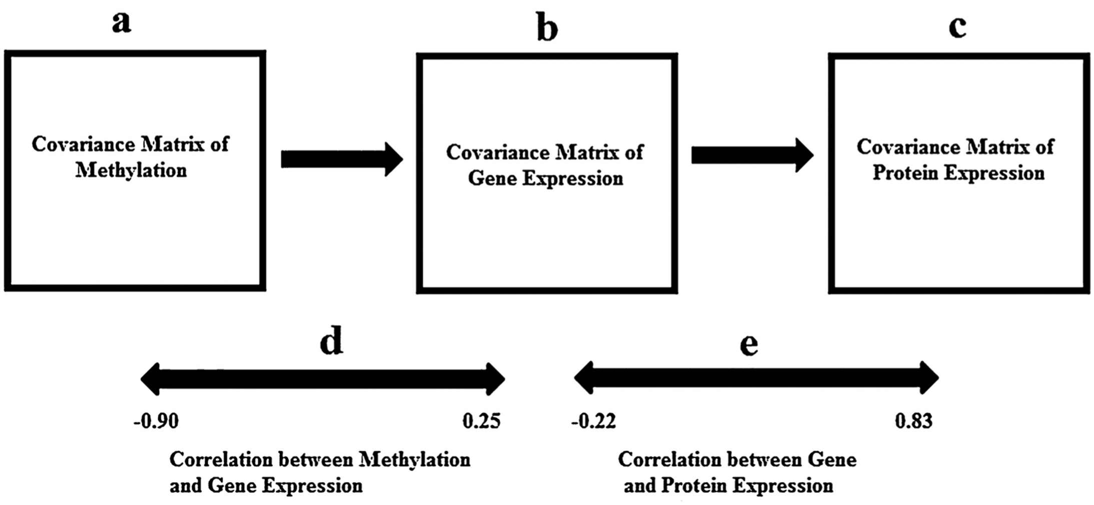
a
b
c
Covariance Matrix of Methylation
Covariance Matrix of Gene Expression
Covariance Matrix of Protein Expression
d
e
-0.90
0.25
-0.22
0.83
Correlation between Methylation and Gene Expression
Correlation between Gene and Protein Expression

**Fig. 1.**

Diagram showing the intra- and inter-correlation structure among the features used in the simulation within and between (a) methylation, (b) gene expression and (c) protein expression data from the TCGA studies on ovarian cancer; (d) represents the correlation between the gene level summary of methylation profile and corresponding gene expression (102 pairs were negatively correlated with minimum value of −0.91 and 29 pairs were positively correlated with maximum value of 0.25); (e) represents correlation between the protein expression and corresponding mapped gene expression (14 pairs were negatively correlated with minimum value of −0.22 and 146 pairs were positively correlated with maximum value of 0.83).

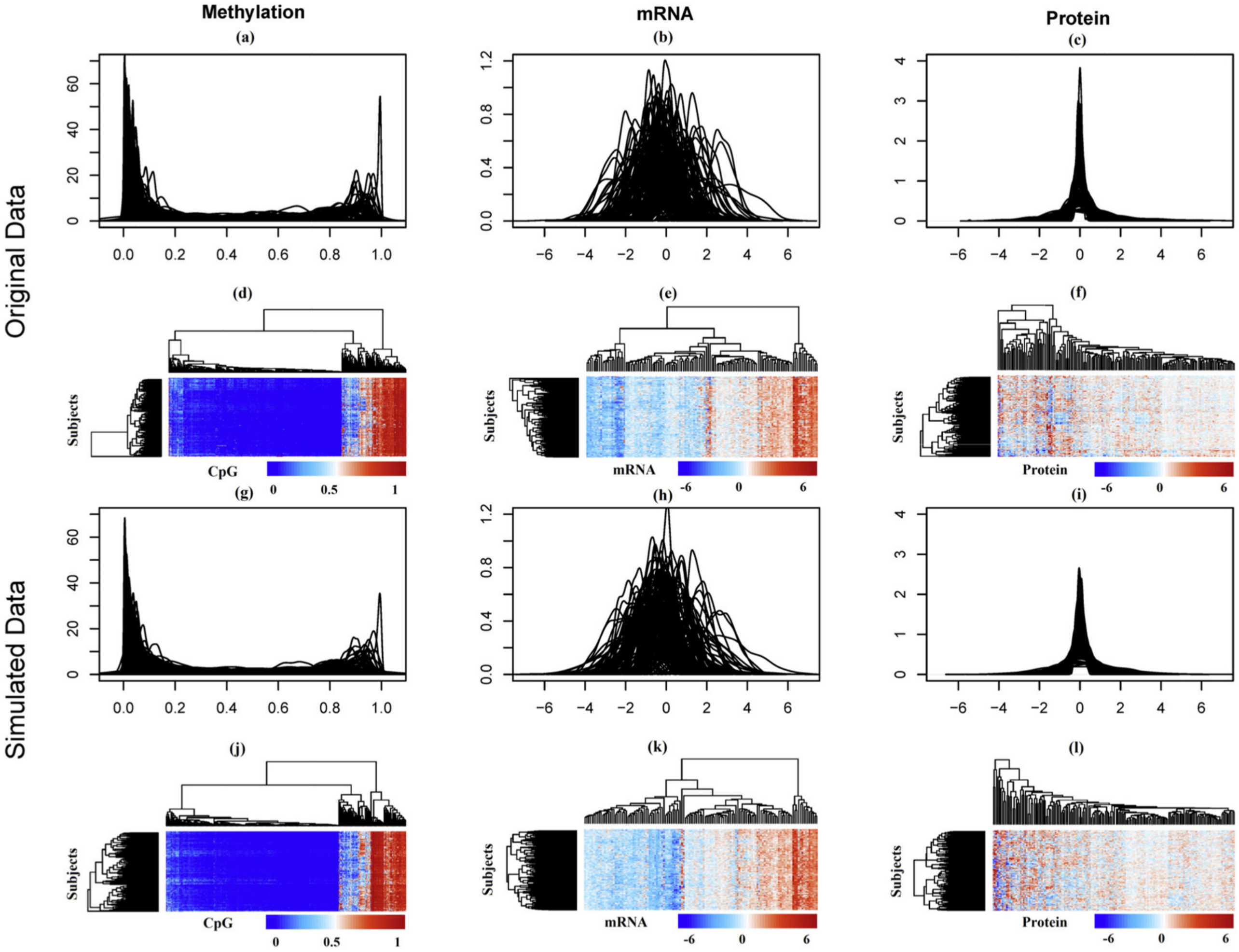

**Fig. 2.**

Comparison of original and simulated data (without cluster-shift effect). (a) and (g) represent the density plots of CpGs in the original data and simulated data respectively; similarly (b)–(h) and (c)–(i) represent the density plots of mRNAs and proteins in the original and simulated data; (d), (e) and (f) represent the heatmaps of the original data and (j), (k) and (l) represent the heatmaps of the simulated data by data type.

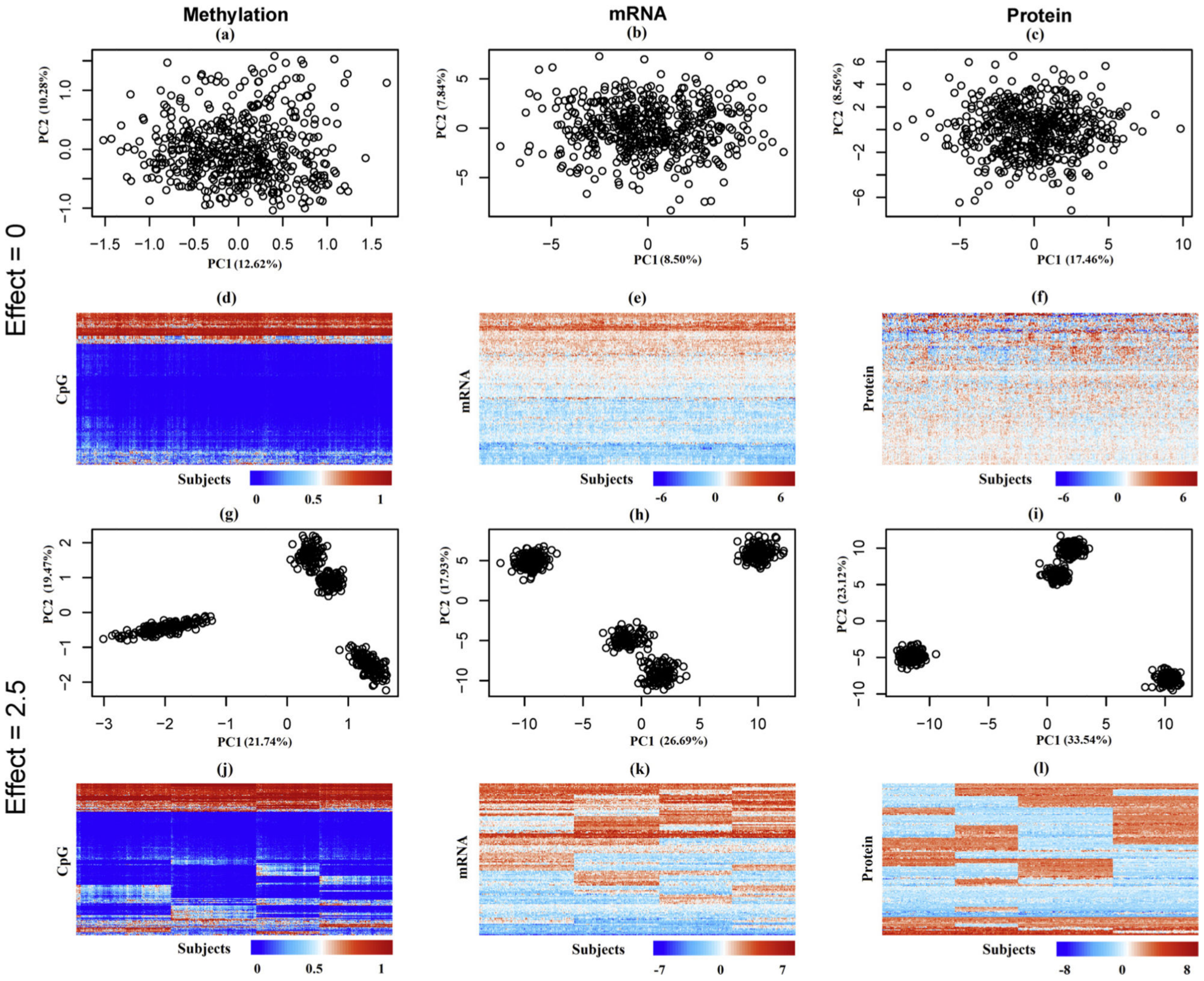

**Fig. 3.**

Example of two sets of simulated data with and without cluster shift effect; (a) and (g) represent the plot between first and second principal components of the methylation data. The numbers in parentheses represent the percentage of variation explained by the first and second principal components; similarly (b)–(h) and (c)–(i) represent the principal components plot of mRNA and protein data respectively; (d), (e) and (f) represent the heatmaps of the first set of simulated data and (j), (k) and (l) represent the heatmaps of the second set of simulated data by data type. The proportion of subjects in the clusters was assigned as 0.20, 0.30, 0.27 and 0.23. The percentages in the parenthesis of the plots (a)–(c) and (g)–(i) represent the percentage of variation explained by the first and second principal components.