# Integrative clustering analysis for omics data with missingness and its application to prostate cancer

Yinqi Zhao,[1] Burcu Darst,[1] Stephen J. Chanock,[2] Sonja I. Berndt,[2] Lynne R. Wilkens,[3] Loic Le Marchand,[3] Demetrius Albanes,[2]
David V. Conti,[1] Christopher A. Haiman[1]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, 90033, USA ;
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, 20892, USA; [3] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, 96822, USA

## Background and Aims

1. Integration of omics data facilitates understanding towards the biological mechanisms between genetic risk factor and complex disease. However, the occurrence of missing data may lead to biased interpretation of statistical results.
2. The Latent Unknown Clustering Integrating omics Data (LUCID)[1] is an integrative model that jointly estimates the latent cluster characterized by omics profiles and genetic risk factor, and associate the latent cluster to risk of disease
3. We extend the LUCID model to handle missing pattern in omics measurement for (1) omics-wise missingness and (2) a general sporadic missing pattern.
4. We demonstrate the performance of our model through simulations and prostate cancer (PCa) data from Multiethnic Cohort (MEC)

## Methods

### 1. Statistical modeling of LUCID model

Based on the DAG in Figure 1, the expected complete likelihood of LUCID model can be written as

$$E_{\boldsymbol{X}} l\left(\boldsymbol{\Theta}\middle|\boldsymbol{D},\boldsymbol{\Theta}^{(t)}\right)$$
$$= \sum_{i=1}^{n}\sum_{j=j}^{k} r_{ij}^{(t)} S\left(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta}\right) + \sum_{i=1}^{n}\sum_{j=j}^{k} r_{ij}^{(t)} \Phi\left(\boldsymbol{M}_i\middle|X_i = j,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)$$
$$+ \sum_{i=1}^{n}\sum_{j=j}^{k} r_{ij}^{(t)} \Phi\left(Y_i\middle|X_i = j,\delta_j,\sigma_j^2\right)$$

## Methods

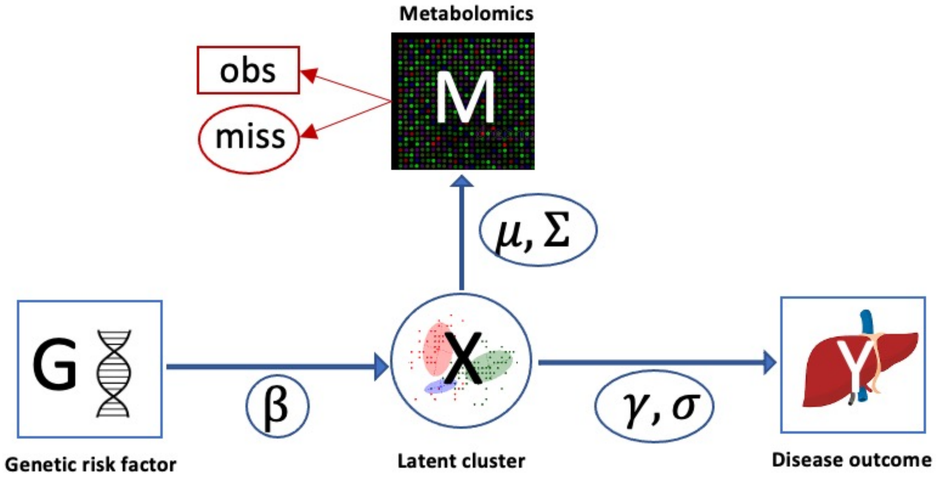


**Figure 1:** DAG illustration of LUCID model.
The square represent observed data; the circle represent unobserved data and model parameters

where $S(\cdot)$ denotes a softmax function, $\Phi(\cdot)$ denotes multi-variate Gaussian distribution and $r_{ij}$ is the responsibility

$$r_{ij} = \frac{s(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta})\Phi\left(\boldsymbol{M}_i\middle|X_i = j,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)\Phi\left(Y_i\middle|X_i = j,\delta_j,\sigma_j^2\right)}{\sum_{j=1}^{k} s(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta})\Phi\left(\boldsymbol{M}_i\middle|X_i = j,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)\Phi\left(Y_i\middle|X_i = j,\delta_j,\sigma_j^2\right)}$$

The Expectation-Maximization (EM) algorithm is applied to estimate model parameters.

**References:** 1. Cheng Peng et al., *Bioinformatics* (2020) 2. Solène Cadiou et al., *Environment International* (2021)

**Support** from NIH U01CA261339, U01CA257328, and P01CA196569

Keck School of Medicine of USC

# Integrative clustering analysis for omics data with missingness and its application to prostate cancer

Yinqi Zhao,[1] Burcu Darst,[1] Stephen J. Chanock,[2] Sonja I. Berndt,[2] Lynne R. Wilkens,[3] Loic Le Marchand,[3] Demetrius Albanes,[2]
David V. Conti,[1] Christopher A. Haiman[1]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, 90033, USA ;
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, 20892, USA; [3] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, 96822, USA

## METHODS

### 2. Extending LUCID to deal with missing omics data

2.1 For omics-wise missing pattern (Figure 2a), we partition the likelihood into $l_{\text{obs}}(\boldsymbol{\Theta}|\boldsymbol{D})$, where omics data $\boldsymbol{M}$ is available and the likelihood is the same as original LUCID, and $l_{\text{miss}}(\boldsymbol{\Theta}|\boldsymbol{D})$, which is

$$E_X l_{\text{miss}}\left(\boldsymbol{\Theta}|\boldsymbol{D},\boldsymbol{\Theta}^{(\text{t})}\right) = \sum_{i=1}^{n_{\text{miss}}}\sum_{j=j}^{k} r_{ij}^{(t)} S(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta}) + \sum_{i=1}^{n_{\text{miss}}}\sum_{j=j}^{k} r_{ij}^{(t)}\,\Phi\left(Y_i|X_i = j,\delta_j,\sigma_j^2\right)$$

The responsibility in $l_{\text{miss}}\left(\boldsymbol{\Theta}|\boldsymbol{D},\boldsymbol{\Theta}^{(\text{t})}\right)$ is only related to exposure and outcome,

$$r_{ij(\text{miss})}^{(t)} = \frac{S(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta})\Phi\left(Y_i|X_i = j,\delta_j,\sigma_j^2\right)}{\sum_{j=1}^{k} S(X_i = j|\boldsymbol{E}_i,\boldsymbol{\beta})\Phi\left(Y_i|X_i = j,\delta_j,\sigma_j^2\right)}$$

2.2 For a general sporadic missing pattern (Figure 2b), we propose a dynamic imputation to impute the missing values through the iteration of EM algorithm. The missing value is updated by

$$M_{il}^{(t+1)} = \sum_{j=1}^{k} r_{ij}^{(t)}\mu_{jl}^{(t)}$$

where $\mu_{jl}^{(t)}$ is the estimate mean of multivariate Gaussian distribution.

## METHODS

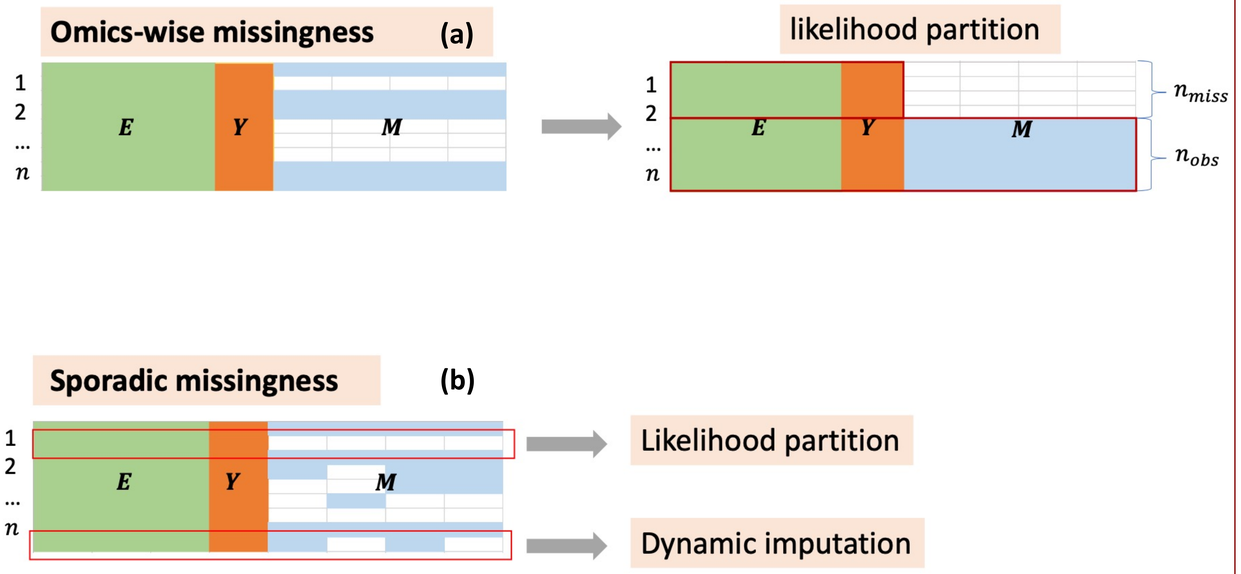


**Figure 2:** Illustration of 2 missing scenarios for extended LUCID model

**REFERENCES:** 1. Cheng Peng et al., *Bioinformatics* (2020) 2. Solène Cadiou et al., *Environment International* (2021)

**SUPPORT** from NIH U01CA261339, U01CA257328, and P01CA196569

Keck School of Medicine of USC

# Integrative clustering analysis for omics data with missingness and its application to prostate cancer

Yinqi Zhao,[1] Burcu Darst,[1] Stephen J. Chanock,[2] Sonja I. Berndt,[2] Lynne R. Wilkens,[3] Loic Le Marchand,[3] Demetrius Albanes,[2]
David V. Conti,[1] Christopher A. Haiman[1]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, 90033, USA ;
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, 20892, USA; [3] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, 96822, USA

## RESULTS

### 1. Simulation study

Data are simulated under LUCID model, with $K = 2$, $\beta_G = \log 2$, $\Delta Z = 0.6$, $\Delta Y = 0.4$. 300 replicates, each replicate generates 2000 observations.

1.1 Omics-wise missing pattern: The data are analyzed by (1) extended LUCID model with likelihood partition; (2) impute missing value + original LUCID; (3) listwise deletion + original LUCID

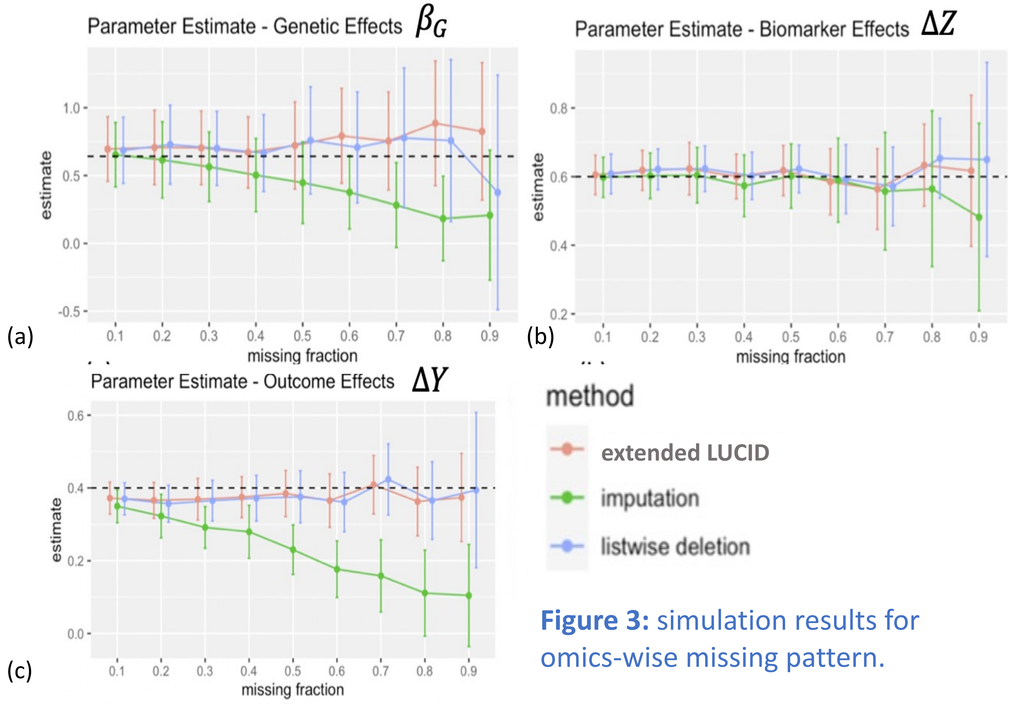


**Figure 3:** simulation results for omics-wise missing pattern.

## RESULTS

1.2 Sporadic missing pattern, analyzed by the extended LUCID model with dynamic imputation.

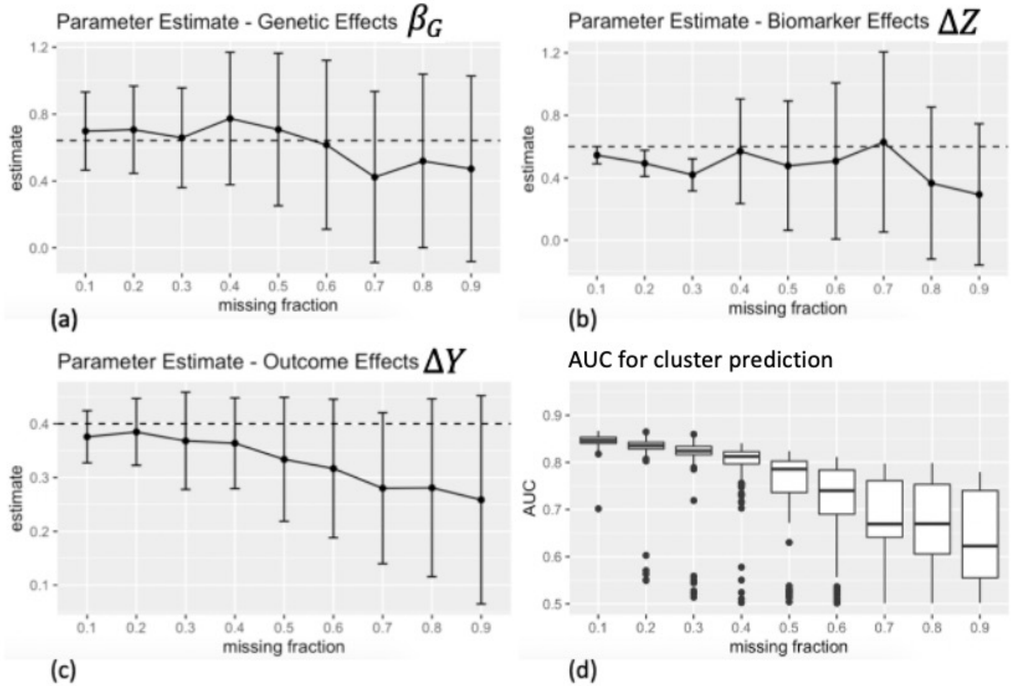


**Figure 4:** simulation for sporadic missing pattern

**REFERENCES:** 1. Cheng Peng et al., *Bioinformatics* (2020) 2. Solène Cadiou et al., *Environment International* (2021)

**SUPPORT** from NIH U01CA261339, U01CA257328, and P01CA196569

Keck School of Medicine of USC

# Integrative clustering analysis for omics data with missingness and its application to prostate cancer

Yinqi Zhao,[1] Burcu Darst,[1] Stephen J. Chanock,[2] Sonja I. Berndt,[2] Lynne R. Wilkens,[3] Loic Le Marchand,[3] Demetrius Albanes,[2]
David V. Conti,[1] Christopher A. Haiman[1]

[1] Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, 90033, USA ;
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, 20892, USA; [3] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, 96822, USA

## Results

### 2. Application to prostate cancer

Data description:

We analyze the genetic risk of prostate cancer in 611 African ancestry men from the Multiethnic Cohort. The pre-diagnostic serum metabolomics contains 1146 features and has sporadic missing pattern

Methods:

We use a "Meet in the middle" approach[2] to screen the metabolomics and select 18 metabolites to fit the LUCID model. (missing ratio ranges from 0 to 0.47). 95% CIs for estimates of LUCID model are derived by Bootstrap.

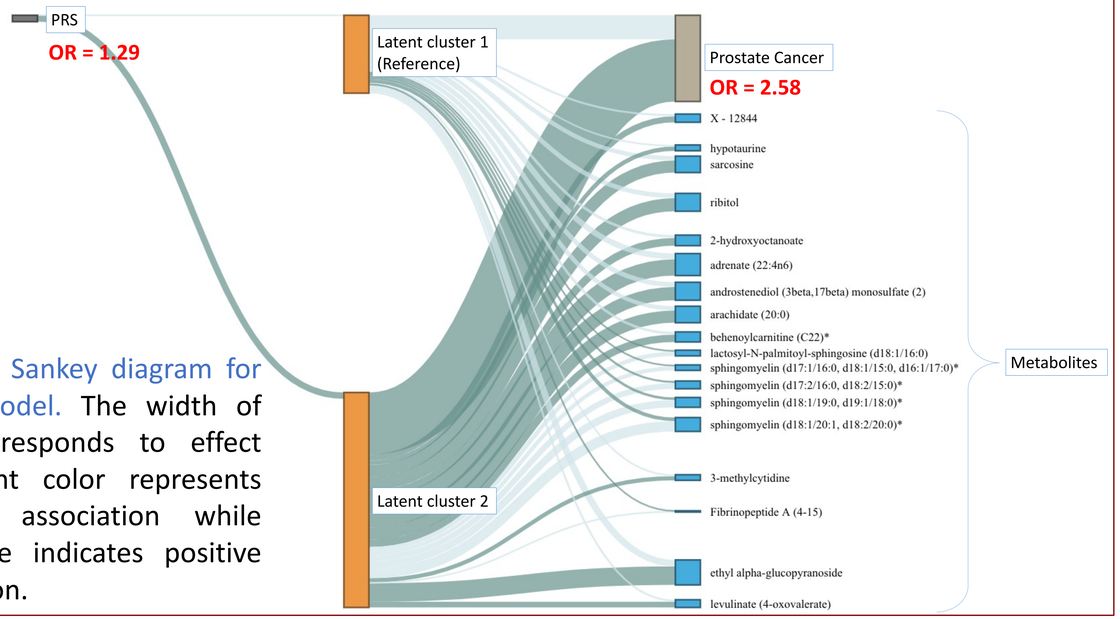


**Figure 5:** Sankey diagram for LUCID model. The width of flow corresponds to effect size. Light color represents negative association while dark blue indicates positive association.

## Results

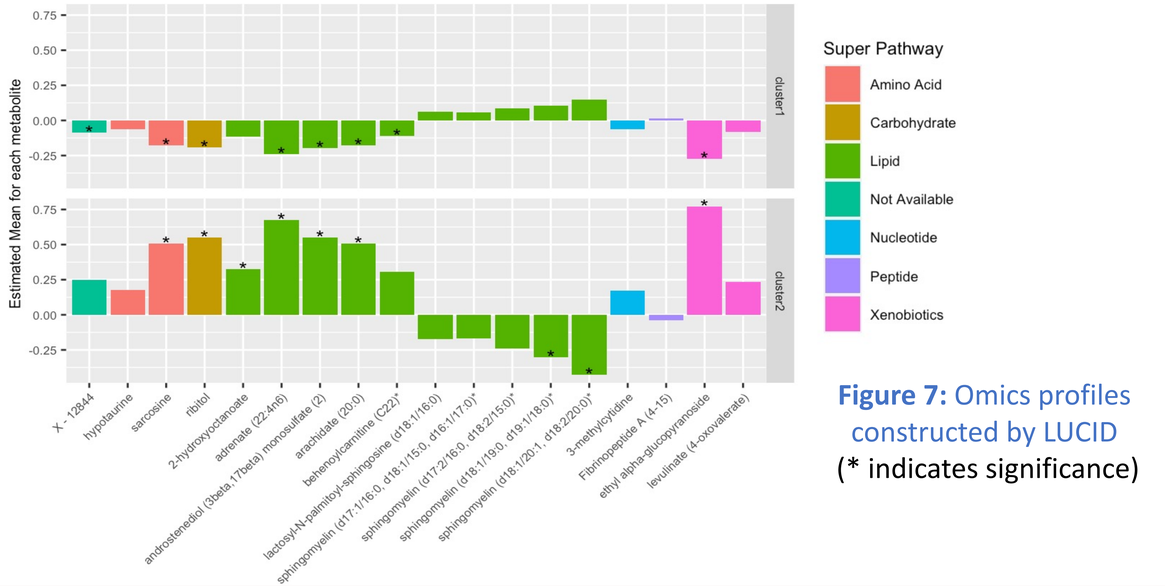


**Figure 7:** Omics profiles constructed by LUCID (* indicates significance)

## Conclusions

1. For omics-wise missing pattern, the extended LUCID model is less biased comparing to imputation at mean and produces parameter estimates with smaller variances comparing to list-wise deletion.
2. For sporadic missing pattern, the extended LUCID model produces estimates for a genetic risk factor $\beta_G$ with small bias, even with a high missing ratio.
3. The extended LUCID model successfully identified high risk group of prostate cancer (OR = 2.58, 95% CI = (1.79, 4.22)) among 611 African American men. The genetic risk factor PRS is associated high risk group (latent cluster 2) with OR = 1.29, 95% CI = (1.01, 1.68).

**References:** 1. Cheng Peng et al., *Bioinformatics* (2020) 2. Solène Cadiou et al., *Environment International* (2021)

**Support** from NIH U01CA261339, U01CA257328, and P01CA196569

Keck School of Medicine of USC