Bayesian linear mixed model with multiple random effects for prediction analysis on high dimensional multi-omics data

Bioinformatics, October 26, 2023

Yang Hai, Jixiang Ma, Kaixin Yang, Yalu Wen

Woobeen Jeong
March 14, 2024

Interdisciplinary Program in Bioinformatics, Seoul National University

Contents

- **Introduction**
- **Materials and Methods**
  - Two step Bayesian linear mixed model
  - Step1: Multi-omics data integration for each region
  - Step2: Risk prediction with multi-omics data
- **Result**
  - Simulation study:
    1. Single omics data vs. Multi omics data
    2. Real data: Alzheimer's disease PET neuro-imaging data
- **Discussion**

Introduction

- Biological phenomena can be clarified in the context of multiple level omics system.

- Two main goal of omics integration is:

  1. multi level pathway inference
  2. detect underlying molecular patterns

- Through analysis and interpretation of such multi-omics data, accurate disease risk prediction is possible, contributing to precision medicine.

- Previous common approaches in ML/DL:

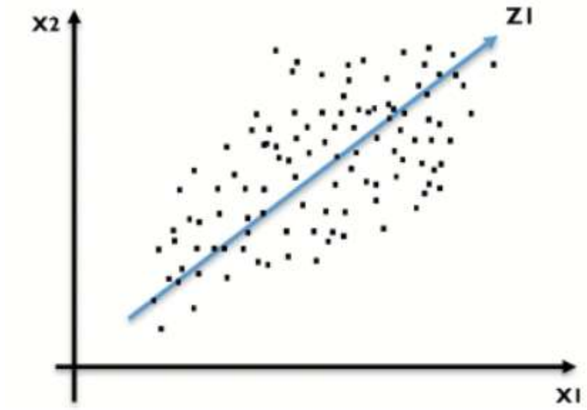
  - Summarize each omics layer from the dataset as a latent variable and then integrates them.
  - Calculate similarities between two classes of samples in networks based analysis.
  - Dimensionality reduction plays a crucial role.

Limitation of Dimensional reduction

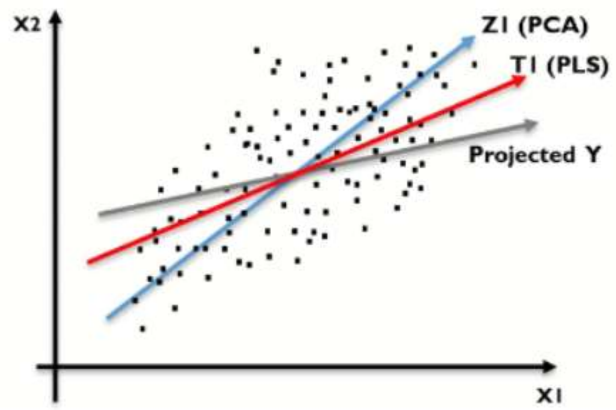
- PCA (Principal Component Analysis)
  - Selects principal components (eigenvectors) in a way that maximizes the variance of the data, by eigen value decomposition on the covariance matrix

- PLS (Partial Least Squares)
  - maximize the variance of the linear combination of X(=t), while simultaneously maximizing the covariance between the linear combination of X, observable as in PCA, and Y.

PCA



PLS



$$t = Xw, w = \text{weight},$$

$$\text{maximize } \; corr(t, Y)Var(t)$$

Limitation of Dimensional reduction

- LRA (Low rank approximation)
  - Approximates a given matrix with a lower-rank matrix, capturing the most important underlying structure by eigenvalue decomposition on the covariance matrix.

- CCA (Canonical Correlation Analysis)
  - Identifies linear combinations of variables in two datasets that maximize their correlation.

- Can simplify the structure of the data and facilitate interpretation.

- However, these kinds of reduction are not designed to maximize the prediction accuracy.

1. Normal distribution for every omics layer may not represent the true effect size distribution.
2. Vast amount of noise can result in an similarity estimation process.

- To address this issues, they proposed two way linear mixed model(TBLMM).

Background: Kernel function (kernel trick)

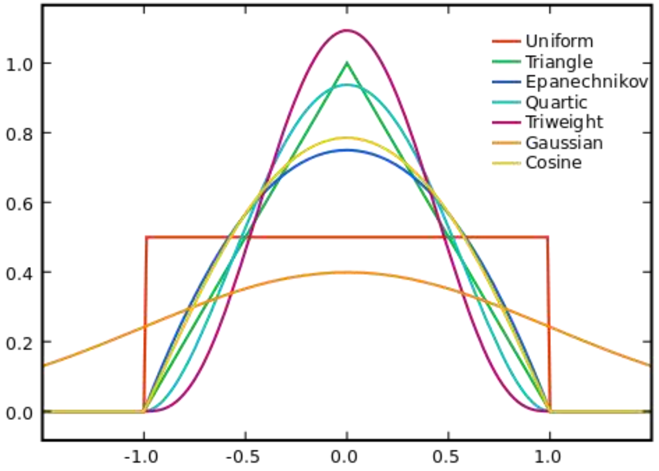
- To operate in a high-dimensional data, implicit feature space by computing inner product. (as a similarity, also called as similarity function K)

$k(x, x'), \quad \forall x, x' \in X, \quad k\colon X \times X \to \mathbb{R} \quad \textit{that satisfies Mercer's condition}$

1) symmetry: $k(x, x') = k(x', x)$

2) positive definiteness: $\sum_{i=1}^{n} \sum_{J=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \qquad \forall c_i, c_j \in \mathbb{R}$

3) Infinite dimensional feature mapping: $\int\int \mathrm{g(x)k(x, x')g(x')}dxdx' < \infty$



as characteristics:

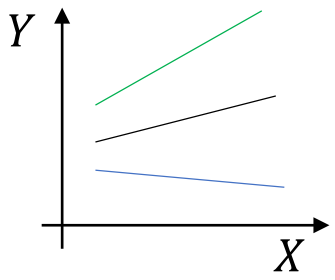
1) positive semi – definiteness: $\int\int g(x)k(x, x')g(x')dxdx' \geq 0, \qquad \alpha^T K\alpha \geq 0$

2) non – negative: $k(x, x') = c, \qquad c\int g(x)dx \int g(x')dx' = c\left(\int g(x)dx\right)^2$

Background: LMM

- Linear mixed model (LMM) consists of fixed effect (β) and random effect(u) also known as blocking:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X_\mathrm{i} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma)$$



$$Y_{i2} = (\beta_0 + u_{02}) + (\beta_1 + u_{12})X_\mathrm{i} + \epsilon_i$$

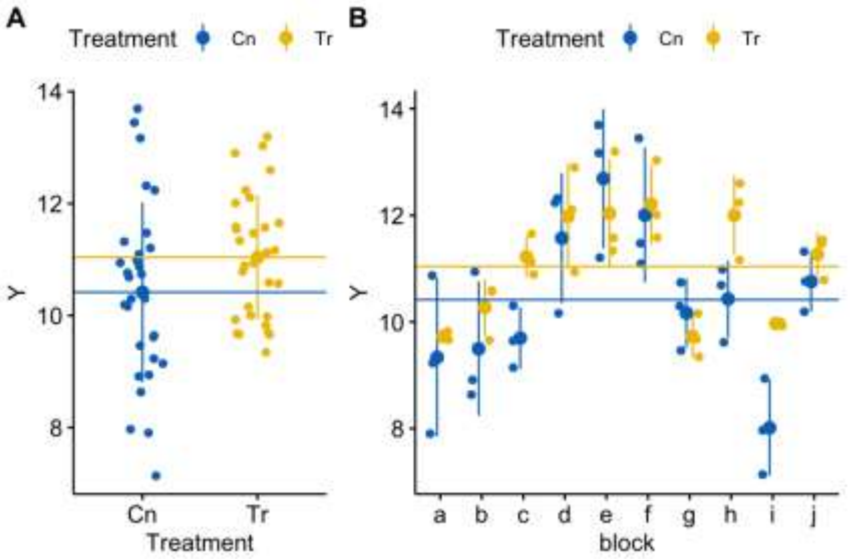
$$Y_i = \beta_0 + \beta_1 X_\mathrm{i} + \epsilon_i$$

$$Y_{i1} = (\beta_0 + u_{01}) + (\beta_1 + u_{11})X_\mathrm{i} + \epsilon_i$$

- For more complex mixed model,

$$Y = X\beta + Zu + \epsilon$$

Fixed effect vector (β) = population mean do not vary.
Random effect vector (u) = parameter can vary depend on blocks(subject, time point, …).



- We can blocking the model by LMM
- Limitations:
  - It simply assumes all genetic variants have the same effect-size distribution that can be sensitive to the underlying disease model.
  - For genomic data, single nucleotide polymorphisms (SNPs) that come from different genetic regions are unlikely to have the same type of effect sizes.

(Applied Statistics for Experimental Biology 2018 Jeffrey A. Walker / Lecture: Linear mixed model - Yue Jiang, Duke University)
(A Bayesian linear mixed model for prediction of complex traits, Bioinformatics Yang et al.)

Background: BLMM

- Bayesian Linear mixed model (BLMM) can accommodate various model assumption, by specifying different prior distributions on cumulative effect(g_m).

$$Y = X\Gamma\beta + \sum_{m=1}^{M} g_m + \epsilon_i,\ \epsilon_i \sim N(0, I\sigma_\epsilon^2), \qquad \beta \sim N(0, \sigma_\beta^2), \qquad \Gamma = \mathrm{diag}(\gamma), \gamma = \left(\gamma_1, ..\gamma_\mathrm{p}\right)^\mathrm{T}, \gamma_i \sim Bernoulli(\theta_0)$$

- $\theta_0 \in [0,1]$ as the tuning parameter for controlling sparsity indicating whether genetic variant exist.
- Using γ as a variable selection avoids underestimating the posterior variance.

- While set a multivariate normal prior for each cumulative effect(g_m), σ_m^2 reflects the effect sizes for predictors that allows for difference across regions(m) as inversed gamma.

$$g_m | K_m \sim N(0, K_m \sigma_m^2), m = 1, \ldots, M, \qquad \sigma_m^2 \sim IG(a, b), \qquad K_m = G_m W_m G_m^T / p_m,$$

$$W_m = diag\left(w_1, \ldots, w_{p_m}\right), w_i = \frac{1}{MAF_i(1 - MAF_i)}$$

- K_m is the genetic similarity for region, where G_m is genotype matrix and number of genetic marker(p_m).
- W_m is the weights of the rare variants, where MAF_i is minor allele frequency for the i th variant (1,.., p_m).
- Because σ_m^2 are expected to be small, the hyperparameters(a, b) are set to be 0.1 for all regions.

(A Bayesian linear mixed model for prediction of complex traits, Bioinformatics Yang et al.)

TBLMM: Two step Bayesian LMM

- To model the outcome as a sum of region-wise predictive effect from region $m \in \{1, \dots, M\}$:

$$Y = \sum_{m=1}^{M} F_m + \epsilon_n,\ \epsilon_n \sim N(0, I\sigma_\epsilon^2)$$

- F_m as the joint predictive effect from all omics data,
  Decomposed into large effect from a few predictors $X_m\beta_m$,
  small effect from a large number of predictors O_m as joint predictive effects.
  S_m is the set of all effects (marginal effect of genome, interaction between genome and methylome)

$$where\ X \in \mathbb{R}^{n \times p_o^m}, \qquad X_m = [X_{expression}^m, X_{methylation}^m, \dots, X_{genomics}^m]$$

$$Y = \sum_{m=1}^{M} (X_m\beta_m + O_m) + \epsilon_n = \sum_{m=1}^{M} (X_m\beta_m + \sum_{j \in S_m} o_j^m) + \epsilon_n, \qquad \epsilon_n \sim N(0, I\sigma_\epsilon^2), \qquad o_j^m \sim N(0, K_j^m \sigma_{mj}^2)$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{m=1}^{M} \begin{bmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{bmatrix} [\beta_1 \cdots \beta_m] + \sum_{m=1}^{M} \sum_{j \in S_m} o_j^m + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

TBLMM: C+T prediction method for ΣXmβm

- Previously in BLMM, due to the high computational burden with high-dimensional multi-omics input, Using a SCT (Stacked Clumping and Thresholding) method to get C+T score. (polygenic score)

$$X_i^{(k)}(INFO_T, r_c^2, w_c^2, p_T) = \sum_{\substack{j\in S\_clumping(k,INFO_T,r_c^2,w_c^2)\\ p_j < p_T}} \widehat{\beta_j}\cdot G_{i,j} \qquad in \quad Y_{BLMM} = X_{\mathrm{m}}\beta_m + \sum_{m=1}^{M} g_m + \epsilon_i$$

- $INFO$ score ($INFO_T$) is the thresholds on genotype imputation $\in [0,1]$,
- r_c^2 is squared correlation for clumping threshold, w_c^2 is clumping window size divided by r_c^2.
- $\widehat{\beta_j} = p_j$ are the p-value as the effect size from GWAS with $INFO$ score ≤ $INFO_T$ (below the threshold).

- Similarly, Using univariate analysis to estimate effect size of each predictor, Select a fraction of predictors with the largest effects in each region m.

- Get estimate effect size $\widehat{\beta_{pm}}$ from $n \times p_o^m$ dimensional X_m (by linear regression)

$$in \quad Y_{TBLMM} = \sum_{m=1}^{M}(X_m^r\beta_m^r + O_m) + \epsilon_i$$

- 5% Largest effect within region m selected : $n \times p_o^m \rightarrow n \times p_r^m$ $(m_r = 0.05 m_o)$ as X_m^{r}

Step 1: Integration for each region

- To model the outcome as a sum of region-wise predictive effect from region $m = 1, \dots, M$, Consider different omics data(T_1^m), within layer interaction(T_2^m), and between layer interaction(T_3^m).

$$from \quad Y = X_m\beta_m + \sum_{j \in S_m} o_j^m + \epsilon_n, \qquad \mathrm{Y} = X_m\beta_m + \sum_{t=1}^{T_1^m} o_t^m + \sum_{t'=1}^{T_2^m} W_{t'}^m + \sum_{t''=1}^{T_3^m} B_{t''}^m + \epsilon_n$$

- Consider joint effect(o_t^m) of omics data by using random effect term, similar to BLMM,

$$o_t^m \sim N\left(0, K_{o,t}^m \sigma_{o,mt}^2\right), \qquad W_{t'}^m \sim N\left(0, K_{w_1,t'}^m \sigma_{w_2,mt'}^2 + K_{w_1,t'}^m \sigma_{w_2,mt'}^2\right), \qquad B_{t''}^m \sim N\left(0, K_{b,t''}^m \sigma_{b,mt''}^2\right)$$

- Calculate marginal predictive effect from each of them by using linear kernel function.

Step 1: Multi omics data integration

- To model the outcome as a sum of region-wise predictive effect from region $m = 1, \ldots, M$, Consider different omics data(T_1^m), within layer interaction(T_2^m), and between layer interaction(T_3^m).

$$from \quad Y = X_m\beta_m + \sum_{j \in S_m} o_j^m + \epsilon_n, \qquad \mathrm{Y} = X_m\beta_m + \sum_{t=1}^{T_1^m} o_t^m + \sum_{t'=1}^{T_2^m} W_{t'}^m + \sum_{t''=1}^{T_3^m} B_{t''}^m + \epsilon_n$$

- Consider joint effect(o_t^m) of omics data by using random effect term, similar to BLMM,

$$o_t^m \sim N\big(0, K_{o,t}^m \sigma_{o,mt}^2\big), \qquad W_{t'}^m \sim N\left(0, K_{w_1,t'}^m \sigma_{w_2,mt'}^2 + K_{w_1,t'}^m \sigma_{w_2,mt'}^2\right), \qquad B_{t''}^m \sim N\big(0, K_{b,t''}^m \sigma_{b,mt''}^2\big)$$

- Calculate marginal predictive effect from each of them by using linear kernel function.

Step 1: Multi omics data integration

$$o_t^m \sim N\left(0, \underline{K_{o,t}^m}\sigma_{o,mt}^2\right), \qquad W_{t'}^m \sim N\left(0, \underline{K_{w_1,t'}^m}\sigma_{w_2,mt'}^2 + \underline{K_{w_2,t'}^m}\sigma_{w_2,mt'}^2\right), \qquad B_{t''}^m \sim N\left(0, \underline{K_{b,t''}^m}\sigma_{b,mt''}^2\right)$$

- Each Kernel defined as:
- While $\mathrm{Z}_{\mathrm{mt}}^k, \mathrm{Z}_{\mathrm{mt}}^l$ are the vectors of t th omics data for individual k, l
- p_{mt} is the number of variants for t th omics layer.
- $\theta_{mt'}^{w2w}$ is the parameter for t th omics layer indicates the rate of decay of the covariance.

$$\underline{K_{o,t}^m}(\mathrm{Z}_{\mathrm{mt}}) = \left(\frac{1}{\sqrt{p_{mt}}}\mathrm{Z}_{\mathrm{mt}}^k\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{p_{mt}}}\mathrm{Z}_{\mathrm{mt}}^l\right)$$

$$\underline{K_{b,t''}^m} = K_{t_1''}^m \circ K_{t_2''}^m$$

$\circ$ = Hadamard product

$$\underline{K_{w_1,t'}^m}(\mathrm{Z}_{\mathrm{mt'}}) = \left(\left(\frac{1}{\sqrt{p_{mt'}}}\mathrm{Z}_{\mathrm{mt'}}^k\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{p_{mt'}}}\mathrm{Z}_{\mathrm{mt'}}^l\right)\right)^2$$

$$\underline{K_{w_2,t'}^m}(\mathrm{Z}_{\mathrm{mt'}}; \theta_{mt'}^{w2w}) = \frac{1}{2\pi}\sin^{-1}\left(\frac{\frac{1}{p_{mt'}}\left(\mathrm{Z}_{\mathrm{mt'}}^k\right)^T \mathrm{Z}_{\mathrm{mt'}}^l}{\sqrt{\left(\theta_{mt'}^{w2w} + \left|\mathrm{Z}_{\mathrm{mt'}}^k\right|^2/p_{mt'}\right)\left(\theta_{mt'}^{w2w} + \left|\mathrm{Z}_{\mathrm{mt'}}^l\right|^2/p_{mt'}\right)}}\right)$$

Step 1: Multi omics data integration

$$o_t^m \sim N\left(0, \underline{K_{o,t}^m}\sigma_{o,mt}^2\right), \qquad W_{t'}^m \sim N\left(0, \underline{K_{w_1,t'}^m}\sigma_{w_2,mt'}^2 + \underline{K_{w_2,t'}^m}\sigma_{w_2,mt'}^2\right), \qquad B_{t''}^m \sim N\left(0, \underline{K_{b,t''}^m}\sigma_{b,mt''}^2\right)$$

- Variance-covariance matrix can be sum of kernels:

$$\Sigma^m = \sum_{t=1}^{T^m} \underline{K_{o,t}^m}\sigma_{o,mt}^2 + \sum_{t=1}^{T_1^m} \underline{K_{w_1,t'}^m}\sigma_{w_2,mt'}^2 + \sum_{t=1}^{T_2^m} \underline{K_{w_2,t'}^m}\sigma_{w_2,mt'}^2 + \sum_{t=1}^{T_3^m} \underline{K_{b,t''}^m}\sigma_{b,mt''}^2 + I_n\sigma_o^2$$

- All coefficients σ_{ml}^2 are non-negative, where

$$\sigma_{ml}^2 \in \left[\sigma_{o,m1}^2, \dots, \sigma_{w1,m1}^2, \dots, \sigma_{w2,m1}^2, \dots, \sigma_{b,m1}^2\right], \qquad l \in \{1, \dots L = T_1^m + T_2^m + T_3^m\}, \qquad K^m = \sum_{l=1}^{L} \frac{\sigma_{ml}^2}{\sum_{l=1}^{L}\sigma_{ml}^2} K_l^m$$

- So that,

$$\mathrm{Y} = X_m\beta_m + \sum_{t=1}^{T_1^m} o_t^m + \sum_{t'=1}^{T_2^m} W_{t'}^m + \sum_{t''=1}^{T_3^m} B_{t''}^m + \epsilon_n \iff \mathrm{Y} = \sum_{m=1}^{M}\left(X_m\beta_m + O_m\right) + \epsilon_n, O_m \sim N\left(0, K^m\sigma_m^2\right)$$

Step 2: Risk prediction

- Under the Bernoulli-Gaussian prior for each β_m^r, re-parameterized by binary variable γ_m

$$Y = \sum_{m=1}^{M} (X_m^r \beta_m^r + O_m) + \epsilon_n, O_m \sim N(0, K^m \sigma_m^2\ )$$

$$\Leftrightarrow\ Y = \sum_{m=1}^{M} (X_m^r \Gamma_m \beta_m^r + O_m) + \epsilon_n, \qquad \Gamma_\mathrm{m} = \mathrm{diag}(\gamma_m), \gamma_m = \left(\gamma_1, ..\gamma_\mathrm{p}\right)^\mathrm{T}, \gamma_m \sim Bernoulli(\theta_0)$$

- Through TBLMM, we can select predictive regions(O_m) when multiple regions are considered.

$$O_m | K^m,\quad \sigma_m^2 \sim D(r_m) N(0,\ K^m \sigma_m^2\ ) + \left(1 - D(r_m)\right)\delta_1,\ r_m \sim Ber(\delta_0), D(r_m),\ \delta_1 = 0$$

- $D(r_m)$ is the probability of success for a Bernoulli random trial of r_m

Simulation study

- To evaluate the performance of **TBLMM**, compared with **TBLMM-LIN** for only linear kernel and **OmicKrig** that widely used method.

- For Multi-omics dataset,

1. Gene expression data from ANDI (Alzheimer's disease Neuroimaging Initiative, n=712) contains an average number of approximately 1,300 mutations per genes.
   - To mimic the real human genome, get the gene expression levels for each gene region by region.

2. Genomic data from ANDI, same.

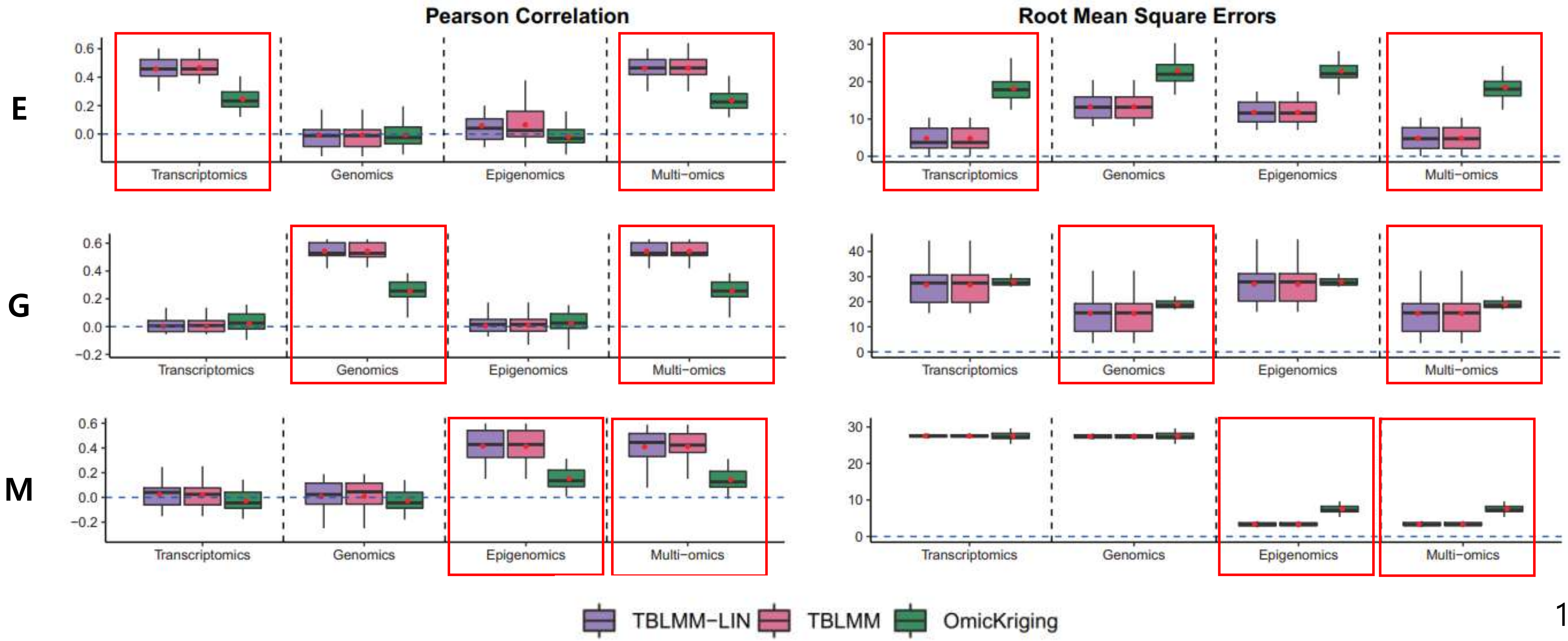
3. Methylation data generated by methylKit

- 80% of train set were randomly selected, remain 20% for calculation of Pearson correlation and RMSE (Root mean square error).
  - higher correlation indicates better prediction on model
  - lower RMSE, Root mean square of error between predicted and actual values indicates the better one.

- Repeat simulation 100 times each.

Simulation study: single omics data

- The performance of TBLMM is better even when the outcome is affected by a single omics only,
- such as gene expression (E) with linear effect, genomic (G), or methylation(epigenomic) (M).
- Of all, even single omics data was given, TBLMM demonstrates better performance.



Real data

- For practical application, real data from ANDI (Alzheimer's disease Neuroimaging Initiative, n=712)

- To predict the illness based on the positron emission tomography (PET) image outcome, Two different baseline datasets, AV45 (n=639) and FDG (n=501), were used.

1. Whole genome sequencing (WGS, n=818) performed on blood sample with illumine Hiseq2000.

2. Gene expression profiling accompanied with WGS (n=811) using U219 Array.

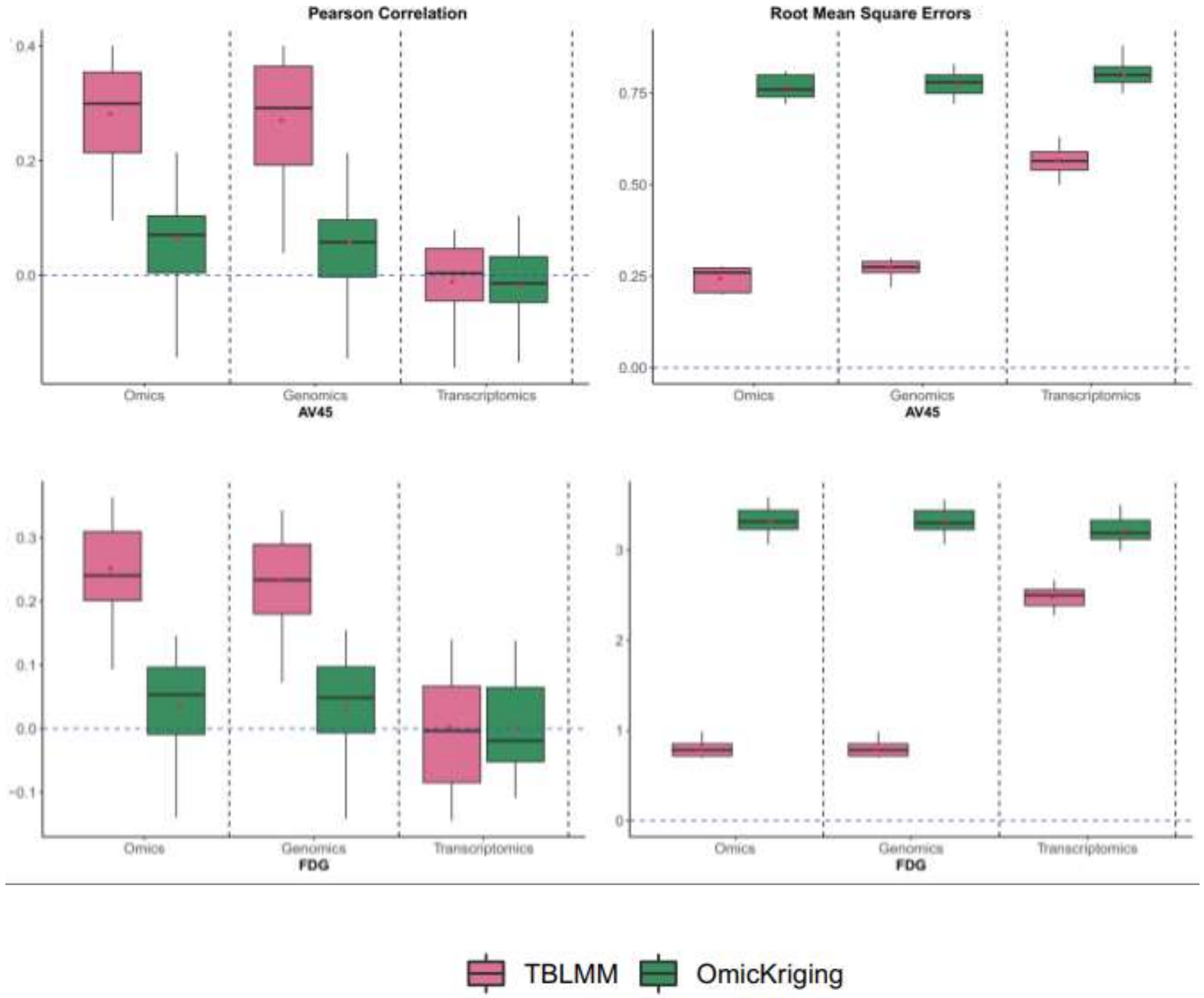
- Total n=712 after quality control were used.

- The datasets were split into train and test sets with an 80% to 20% ratio, respectively, and this process was repeated 100 times.

Result

- For both AV45 and FDG, TBLMM showed higher prediction performance than OmicKrig.

- Comparing model with omics data and single layer omics only, variance of outcome can be mainly explained by genetic effects.



Discussion

- Proposed TBLMM is a flexible model because of a two step procedure,
  first step focusing on dimensional reduction via kernel fusion
  while second step for detect predictors through Bayesian linear mixed model

- It can accommodate various disease model.

- By selecting appropriate kernel, TBLMM can capture not only key predictors but also discriminate between additive and non-additive omics effects.

- Based on real data from Alzheimer's patients, TBLMM performs better than OmicKrig with higher prediction accuracy.

Thank You