

## Research Article

# GEE-TGDR: A Longitudinal Feature Selection Algorithm and Its Application to lncRNA Expression Profiles for Psoriasis Patients Treated with Immune Therapies

Suyan Tian <sup>1</sup>, Chi Wang <sup>2,3</sup> and Mayte Suarez-Farinas <sup>4,5</sup>

<sup>1</sup>Division of Clinical Division, First Hospital of Jilin University, Changchun, Jilin, China 130021

<sup>2</sup>Department of Internal Medicine, College of Medicine, University of Kentucky, 800 Rose St., Lexington, KY 40536, USA

<sup>3</sup>Markey Cancer Center, University of Kentucky, 800 Rose St., Lexington, KY 40536, USA

<sup>4</sup>Department of Population Health Science & Policy, The Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA

<sup>5</sup>Department of Genetics and Genomics, The Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA

Correspondence should be addressed to Suyan Tian; [windytian@hotmail.com](mailto:windytian@hotmail.com)  
and Mayte Suarez-Farinas; [mayte.suarezfarinas@mssm.edu](mailto:mayte.suarezfarinas@mssm.edu)

Received 23 September 2020; Revised 6 March 2021; Accepted 29 March 2021; Published 10 April 2021

Academic Editor: Dominique Monlezun

Copyright © 2021 Suyan Tian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the fast evolution of high-throughput technology, longitudinal gene expression experiments have become affordable and increasingly common in biomedical fields. Generalized estimating equation (GEE) approach is a widely used statistical method for the analysis of longitudinal data. Feature selection is imperative in longitudinal omics data analysis. Among a variety of existing feature selection methods, an embedded method—threshold gradient descent regularization (TGDR)—stands out due to its excellent characteristics. An alignment of GEE with TGDR is a promising area for the purpose of identifying relevant markers that can explain the dynamic changes of outcomes across time. We proposed a new novel feature selection algorithm for longitudinal outcomes—GEE-TGDR. In the GEE-TGDR method, the corresponding quasilielihood function of a GEE model is the objective function to be optimized, and the optimization and feature selection are accomplished by the TGDR method. Long noncoding RNAs (lncRNAs) are posttranscriptional and epigenetic regulators and have lower expression levels and are more tissue-specific compared with protein-coding genes. So far, the implication of lncRNAs in psoriasis remains largely unexplored and poorly understood even though some evidence in the literature supports that lncRNAs and psoriasis are highly associated. In this study, we applied the GEE-TGDR method to a lncRNA expression dataset that examined the response of psoriasis patients to immune treatments. As a result, a list including 10 relevant lncRNAs was identified with a predictive accuracy of 70% that is superior to the accuracies achieved by two competitive methods and meaningful biological interpretation. A widespread application of the GEE-TGDR method in omics longitudinal data analysis is anticipated.

## 1. Introduction

With fast evolution of high-throughput technology, longitudinal omics experiments have become affordable and increasingly common in many biomedical fields for exploring dynamically or temporally changed biological systems or processes. Usually, the analysis strategies focus on analyzing individual time points separately. As many investigators have pointed out [1–4], a failure to incorporate information contained in the dependent structure of time course data

results in inefficient estimation of the standard errors, leading to an inadequate statistical power. Especially in big omics studies, this problem stands out since the sample size of such data is usually small. Furthermore, an oversimplified consideration by combining the results from marginal analysis at individual time points tends to fail to detect a meaningful pattern of changes over time.

The generalized estimating equation (GEE) approach [5] is a well-established and widely used statistical method to analyze longitudinal data. GEE considers the first two

marginal moments (i.e., mean and variance) of data and a working correlation matrix to model correlated responses, artfully avoiding the specification of full joint likelihood function. The appeal of GEE lies in that it yields consistent estimators for the parameters of interest, even if the working correlation structure is incorrectly specified. Naturally, GEE has been modified or extended to identify differentially expressed genes over time for high-throughput data. Such modifications and/or extensions are not simple due to the high dimensionality of omics data, although some efforts have been made [1, 2, 6].

Like its cross-sectional counterpart, feature selection is imperative in the learning process for longitudinal omics data. Feature selection is aimed at eliminating irrelevant genes, avoiding overfitting, speeding up the learning process, and achieving a final model that is parsimonious (i.e., the number of selected genes is as least as possible). Consequently, a modification to GEE to analyze high-dimensional data necessitates the involvement of feature selection. In the literature, there are several such algorithms. For example, Wang et al. [2] used a smoothly clipped absolute deviation (SCAD) penalty term [7] which is a novel extension to the  $L_1$  penalty to equip the GEE models with feature selection capacity. The  $L_1$  penalty, also known as LASSO [8], forces genes with small estimated coefficients out of the final model, rendering a sparse solution by the means of which feature selection occurs. However, two subsequent works on this topic [1, 6] showed that this algorithm usually fails to converge when the number of covariates is much larger than the number of samples. This drawback is more apparent and fatal in longitudinal omics data, where the sample size is typically smaller than that of a cross-sectional study.

Among a variety of existent feature selection algorithms, we have devoted dramatic efforts on the threshold gradient descent regularization (TGDR) [9] method (see the Methods section for its description). Previously, we had extended TGDR for classification task of multiple groups ( $>2$ ) and for identification of subgroup-specific prognostic genes with a survival outcome [10–14]. By applying these TGDR extensions to different types of omics data including microarray, RNA sequencing, and mass spectrometry (MS) data, we have shown that TGDR and its respective extensions have many merits including easy-to-moderate programming intensity, good predictive performance, and biologically meaningful implications of the resulting signatures. In a recent work [4], we show that the TGDR algorithm can be regarded as an optimization strategy and that the final models given by TGDR have superior predictive performance and more meaningful biological interpretation than the LASSO models optimized by the coordinate descent method [15]. Therefore, an integration of GEE with TGDR may overcome the drawbacks of existing approaches for the purpose of longitudinal feature selection.

Long noncoding RNAs (lncRNAs) are posttranscriptional and epigenetic regulators and have the characteristics of lower expression levels and more tissue-specific compared with protein-coding genes [16]. Once being regarded as evolutionary junks, lncRNAs have been demonstrated to play essential roles in many complex diseases, especially in cancer

[16]. As pointed out by our previous study [17], psoriasis is an ideal model for examining the effects of targeted immune treatments given that it is well characterized by molecular profiles, displays low placebo effects, and possesses easily accessible diseased tissues. So far, the implication of lncRNAs in psoriasis remains largely unexplored and poorly understood. Among the limited research carried out to explore the roles of lncRNAs play in psoriasis; however, some encouraging results have turned up. For example, a very recent study [18] has shown that LOC285194 can serve as a sponge for miR-616 that regulates the expression of GATA3 though binding to its 3'-untranslated region using Western blotting, quantitative real-time PCR, and dual-luciferase reporter assays. Specifically, the expression level of LOC285194 was lower in the affected skin of patient with psoriasis compared to the unaffected skin. Furthermore, Rakhshan et al. [19] showed that one SNP (i.e., rs12826786) of the HOX Transcript Antisense RNA (HOTAIR) is associated with a higher risk of developing psoriasis (TC+TT versus CC: OR = 1.59,  $p = 0.02$ ). Therefore, we believe that the roles of lncRNAs play in psoriasis deserve to be explored deeply and widely.

In this article, we proposed a new feature selection algorithm, referred to as GEE-TGDR, specifically for longitudinal data mining and feature selection. In the GEE-TGDR method, the corresponding quasilielihood function of a GEE model is the objective function to be optimized, while the optimization and feature selection are accomplished by the TGDR method. We applied this method to a longitudinal microarray gene expression data that is aimed at assessing the treatment efficacy of two immune therapies for psoriasis patients and identified the relevant lncRNAs that can predict the temporal changes of psoriasis area and severity index (PASI) scores that is utilized to determine if a patient with psoriasis responds to the treatments, with the objectives of revealing the underlying mechanisms of these two treatments from the perspective of lncRNAs.

Following the structures of a review by [20], the article is organized as follows. In Section 2, the details about the proposed GEE-TGDR method are given. In Section 3, the application of the GEE-TGDR method to psoriasis longitudinal lncRNA expression data and the analysis results are presented. Then, the biological relevance of identified lncRNA signature to psoriasis is discussed in detail. In Section 4, the limitations of the present study in addition to contributions and future work are discussed. Lastly, conclusions are given.

## 2. Materials and Methods

**2.1. Experimental Data.** The microarray dataset [17] used to characterize the proposed GEE-TGDR algorithm was in the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>) under the accession number of GSE85034. There were 179 arrays in this experiment, including the gene expression profiles of 30 patients with moderate to severe psoriasis at the baseline nonlesion skins and baseline lesion skins and at weeks 1, 2, 4, and 16. Of the 30 patients, half were administrated with adalimumab (ADA), and the other half were treated with methotrexate (MTX). One patient on the ADA arm had no expression measurements of week 16

since his/her psoriasis area and severity index (PASI) score already had experienced a 75% decrement at the week 4. In original paper, a treatment response was based on a reduction of 75% in PASI score after week 12 or later. Longitudinal profiles of PASI scores (baseline lesion skins, at weeks 1, 2, and 4) were the outcomes of interest, and the lncRNA expression values of the baseline lesional skins serve as potential predictors to investigate if they are relevant to the PASI scores of psoriasis patients over time.

In this study, the preprocessed data were directly downloaded from the GEO database. No alternative preprocessing had been carried out. By matching the gene symbols of lncRNAs in the GENCODE (<https://www.genecodegenes.org/>) database (version 32) to those of genes annotated by the Illumina HumanHT-12 V 4.0 bead chips, 662 unique lncRNAs were identified and included in the downstream analysis.

**2.2. Statistical Methods.** In this paper, we conceive a new novel feature selection algorithm called GEE-TGDR specifically for selecting relevant features associated with the temporal changes of longitudinal outcomes, in which GEE is equipped with TGDR just as its name implies. We briefly described both GEE and TGDR methods before proceeding to the proposed integration. Here, to keep it the most relevant, we focused on the case of continuous outcomes.

**2.2.1. Threshold Gradient Descent Regularization.** For continuous outcomes, the TGDR algorithm is based on a linear model, where a response variable  $Y_i$  ( $i = 1, \dots, n$ ,  $n$  is the sample size) is modelled by a  $P$ -dimensional vector of observed covariates  $X_{ip}$  (here,  $p = 1, \dots, P$ ) as  $E(Y | X) = X^T \beta$ . Here,  $\beta$ 's represent the coefficients of covariates for the magnitudes of association between covariates and the outcome. For continuous outcomes, a normal distribution is usually assumed, and then, the corresponding likelihood function is used as a response function/an objective function in the TGDR algorithm. With some algebraic simplification, the response function can be written as

$$\text{Res}(\beta) = n^{-1} \sum_{i=1}^n (Y_i - X_i^T \beta)^2. \quad (1)$$

The TDGR algorithm started from that the  $\beta$ 's were initially set at zero's (corresponding to the null model). Using  $\Delta v$  to denote a small positive increment (e.g., 0.01) in the gradient descent search, and for iteration  $k$ ,

- (1) Upon current estimate  $\beta^{(k)}$ , a negative gradient matrix  $g$  with its  $p^{\text{th}}$  component as  $g_p^{(k)}$  are calculated as

$$g_p^{(k)} = n^{-1} \sum_{i=1}^n X_{ip} (Y_i - X_i^T \beta^{(k)}). \quad (2)$$

- (2) Let  $f(k)$  represent the threshold vector of size  $P$  at iteration  $k$  and  $I(x)$  is an indicator (if the condition

$x$  is true, this indicator returns 1; otherwise, its value is 0), then its  $p^{\text{th}}$  component (for the  $p^{\text{th}}$  gene) is

$$f_p^{(k)} = I(|g_p^{(k)}| \geq \tau \times \max_{l \in \{1, 2, \dots, P\}} (|g_l^{(k)}|)). \quad (3)$$

- (3) Update  $\beta_p^{(k+1)} = \beta_p^{(k)} + \Delta v \times g_p^{(k)} \times f_p^{(k)}$  and  $k = k + 1$
- (4) Repeat steps 1-3 for  $K$  times.  $K$  can be regarded as a tuning parameter, with a large value corresponding to a dense model (more nonzero coefficients) and a small value to a sparse model (less nonzero coefficients). The optimal value of  $K$  is determined by crossvalidations (CVs).

In the TGDR method, no explicit penalty term is added to the objective function (i.e., response function). The regularization on coefficients (thus the selection of features) is made possible by introducing the threshold function  $f^{(k)}$  in step 2, which determines if the gradient of a coefficient is large enough to descent or more precisely speaking to be updated. For more detailed description of the TGDR method, the works [9, 21] are referred.

**2.2.2. Generalized Estimating Equation.** In the longitudinal notation, the  $j^{\text{th}}$  time point/measurement of the  $i^{\text{th}}$  subject, a  $t$ -dimensional vector of response variables  $Y_{ij}$  (here,  $i = 1, \dots, n$  and  $j = 1, \dots, t$ ) and covariates  $X_{ijp}$  (here,  $p = 1, \dots, P$  represents  $p^{\text{th}}$  covariate) are observed. Thus  $Y_i = (Y_{i1}, \dots, Y_{it})^T$  denotes the vector of responses at  $t$  different time points for subject  $i$ , and  $X_{ij} = (X_{ij1}, \dots, X_{ijP})^T$  is  $P$  covariates for subject  $i$  at time point  $j$ .

In the GEE model, the first two marginal moments of  $Y_{ij}$  are denoted by  $\mu_{ij}(\beta) = E(Y_{ij} | X_{ij})$  (the expectation of  $Y_{ij}$  given  $X_{ij}$ ) and  $\sigma^2(\beta) = V(Y_{ij})$  (the variance of  $Y$ ). Here,  $\beta$ 's are the coefficients representing the magnitude of association between covariates and outcomes, with  $\beta_{jp}$  representing how attribute  $p$  is associated with the value of outcome  $Y_j$  (meaning the outcome at time point  $j$ ). Those  $\beta$ 's are parameters of interest. Furthermore, the distribution of  $Y_{ij}$  is assumed to belong to an exponential family with a canonical link function. Let  $\mu_i(\beta) = (\mu_{i1}(\beta_1), \dots, \mu_{it}(\beta_t))^T$  and  $A_i(\beta) = \text{diag}(\sigma_{i1}^2(\beta_1), \dots, \sigma_{it}^2(\beta_t))$ , then under a canonical link function  $V_i(\beta) = A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$ . Here,  $R_i(\alpha)$  is an  $t \times t$  working correlation matrix with  $\alpha$  as the finite dimensional parameter vector for correlations, which would be usually estimated by the residual-based moment method. In a GEE model, the quasilielihood function can be written as

$$QL(\beta) = n^{-1} \sum_{i=1}^n (Y_i - \mu_i(\beta))^T V_i^{-1}(\beta) (Y_i - \mu_i(\beta)). \quad (4)$$

Four structures are commonly used for the working correlation matrix  $R_i(\alpha)$ —first-order autoregressive (AR1), exchangeable, unstructured, and independent structure.

**2.2.3. GEE-TGDR.** The conventional TGDR method only deals with univariate outcomes. As far as longitudinal outcomes that are multivariate are concerned, the method needs to be extended.

In this study, we proposed to replace the likelihood function with the corresponding quasilielihood function and to extend TGDR as GEE-TGDR. With  $\Delta v$  denoting a small positive increment (e.g., 0.01) in gradient descent search, then at  $k$  iteration,

- (1) Upon current estimate  $\beta^{(k)}$ , a negative gradient matrix  $g$  with its  $(j, p)^{\text{th}}$  component as  $g_{jp}^{(k)}$  are calculated

$$g_{jp}^{(k)} = n^{-1} \sum_{i=1}^n X_{ijp}^T A_i^{1/2}(\beta^{(k)}) R^{-1}(\alpha) A_i^{-1/2}(\beta^{(k)}) (Y_i - \mu_i(\beta^{(k)})). \quad (5)$$

- (2) Let  $f_j^{(k)}$  represent the threshold vector of size  $P$  for the  $j^{\text{th}}$  time point ( $j = 1, \dots, t$ ) at iteration  $k$ , then its  $p^{\text{th}}$  component (for the  $p^{\text{th}}$  gene) is

$$f_{jp}^{(k)} = I(|g_{jp}^{(k)}| \geq \tau \times \max(|g_{jl}^{(k)}|)), \quad \forall l \in (1, 2, \dots, P). \quad (6)$$

- (3) Update  $\beta_{jp}^{(k+1)} = \beta_{jp}^{(k)} + \Delta v \times g_{jp}^{(k)} \times f_{jp}^{(k)}$  and  $k = k + 1$
- (4) Calculate the residuals, viz.  $Y_i - \mu_i(\beta^{(k)})$ , and based on them, to estimate the nuisance parameters involved in  $R(\alpha)$  (for different correlation structures, the parameters are different) and  $A_i(\beta^{(k)})$ . Of note, since at different time points, we have different threshold function, the selected genes at different time points are expected to differ. In this way, the selection of critical time points is possible
- (5) Repeat steps 1-4 for  $K$  times.  $K$  is a tuning parameter, the same as in the conventional TGDR method. The optimal value of  $K$  is also determined by CVs

In this study, we only developed the GEE-TGDR algorithm for continuous outcomes given in the motivated database; PASI scores which are continuous were the outcomes of interest, then the corresponding expectations of  $Y_i$ 's given  $X_i$ 's are  $[X_{i1}^T \beta_1, \dots, X_{it}^T \beta_t]$ . Here, let  $j = 1, 2, \dots, t$  represent the time points measured; then  $X_{ij}$  are for the gene expression profiles at time point  $j$  for subject  $i$ , and  $\beta_j$  are for the corresponding coefficients of those gene expression values

at time point  $j$ . Figure 1 gives the graphical illustration of the GEE-TGDR algorithm for continuous longitudinal outcomes.

Since the outcomes were continuous, the mean squared error (MSE) statistic was calculated to evaluate the performance of resulting gene signatures. It is worth pointing out that for the outcomes of other types, an extension suitable for the underlying data type of GEE-TGDR algorithm is straightforward, with the corresponding quasilielihood function serving as the objective function/response function.

**2.3. Statistical Language.** Statistical analysis was carried out in the R language version 3.6.1 (<http://www.r-project.org>).

### 3. Results

**3.1. Identified LncRNA Signatures.** In this study, we propose to extend the feature selection algorithm TDGR to account for correlation structure of longitudinal data. This is accomplished by defining the objective function of TDGR as the corresponding quasilielihood function, which as in GEE is specified based on the first two moments and a working correlation matrix. TDGR-GEE is described in the Materials and Methods section. In this section, we illustrate the application of the proposed method while looking for biomarkers that predict clinical resolution of psoriasis after being treated with two immune therapies.

Gene expression profiles of baseline lesional skin biopsies were obtained for 30 subjects followed up to 16 weeks after treatment with adalimumab and methotrexate. Clinical resolution at weeks 1, 2, and 4 was measured by PASI. In this example, we would like to identify a signature of genes whose baseline expression values correlate with changes in PASI, our continuous longitudinal outcome. We used 662 lncRNA as covariates in the proposed GEE-TGDR model, under 4 different working correlation structures. The performance statistics (i.e., MSEs) and identified lncRNA genes are presented in Table 1.

In this application, the results obtained under working correlation structures exchangeable, unstructured, and independent barely differ, with similar sets of biomarkers leading to similar performance. This reflects a well-known robust characteristic of GEE, where when predictors are correctly given, the GEE estimates remain consistent even if the correlation structures are misspecified. Under the AR1 structure, GEE-TGDR identified only one lncRNA as being related to PASI scores, leading to an underfitting and inferior to the performance when compared to the other three correlation structures.

Due to the patient burden and budgetary restrictions, longitudinal omics data are usually very short and unevenly spaced. In this case, AR1 is not well suited and the unstructured correlation may be the most suitable structure, even though that this structure corresponds to a model with more nuisance parameters involved in the corresponding working correlation structure.

Crossvalidation (CV) results gave us an idea for the variability in the model performance in this regard; CV results indicated that all correlation structures but AR1 structure

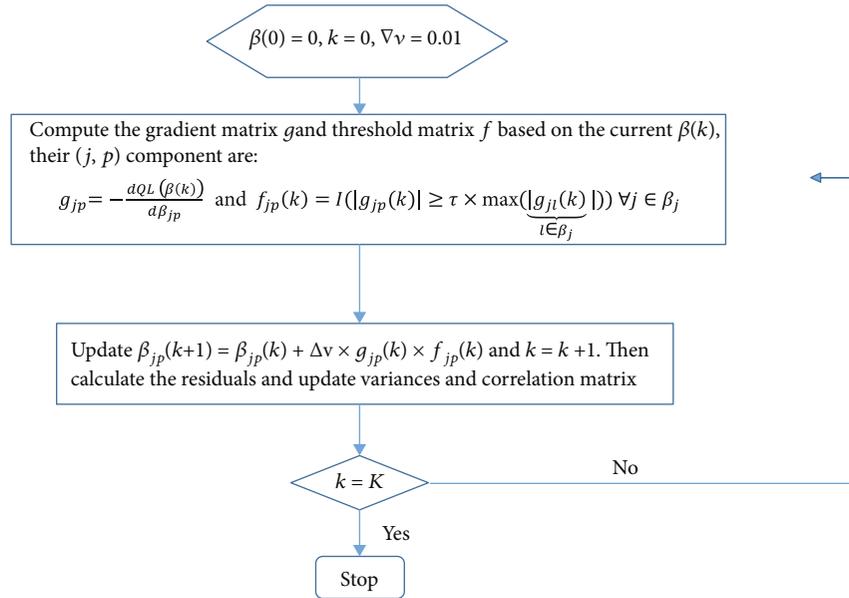


FIGURE 1: Flowchart of the proposed GEE-TGDR algorithm.

TABLE 1: Results of psoriasis lncRNA longitudinal data.

	Ave. of MSE (5-fold CVs)	SD of MSE (CVs)	MSE (all data)	Identified lncRNAs (using all data)			
				Baseline	Week 1	Week 2	Week 4
AR1	14.456	3.258	2.101	RAMP2-AS1	RAMP2-AS1	RAMP2-AS1	RAMP2-AS1
Unstructured	3.725	0.498	0.793	XIST RAMP2-AS1 MIR205	LRRC75A-AS1 PAXIP1-AS1 LINC00667 RAMP2-AS1 MIR205	LRRC75A-AS1 TMEM99 LINC01018 PAXIP1-AS1 LINC01139 RAMP2-AS1	TMEM99 LINC01018 PAXIP1-AS1 LINC01139 RAMP2-AS1
Exchangeable	2.758	1.649	0.767	XIST RAMP2-AS1 MIR205	LRRC75A-AS1 XIST LINC01139 SDHAP2 RAMP2-AS1	TMEM99 LINC01139 RAMP2-AS1	XIST LINC01018 PAXIP1-AS1 LINC01139 RAMP2-AS1
Independent	2.675	1.694	0.760	SNHG5 LINC01139 RAMP2-AS1 MIR205	SNHG5 RAMP2-AS1 MIR205	SNHG5 TMEM99 RAMP2-AS1 MIR205	SNHG5 XIST LINC01018 LINC01139 RAMP2-AS1 MIR205

Only baseline expression values were used. AR1: autoregressive order 1; MSE: mean squared error; SD: standard deviation; CV: crossvalidation.

provided similar results, with both the exchangeable and independent structures having the least MSEs but a bigger variability and the unstructured structure having a larger MSE but the smaller variations.

Even though that at individual time points, the identified features varied substantially for the unstructured, exchangeable, and independent working correlation structures (Figure 2), the unions of lncRNA lists across time are essentially the same, including 9 lncRNAs identified by all these three structures and one lncRNA selected by the independent structure alone (Figure 3).

3.2. Comparison with Competing Methods. In order to further characterize the GEE-TGDR method, a comparison with two competing methods was made. One competing method under consideration is the GEE-screening method [1] in which a GEE model was fit with the PASI scores over time as the outcome and the expression values of a certain lncRNA as the covariate. The GEE-screening method filtered genes one by one. Of note, in the GEE-TGDR method and the GEE-screening model, we only considered the unstructured working correlation structure. In the other competing method, namely, linear mixed model-based screening

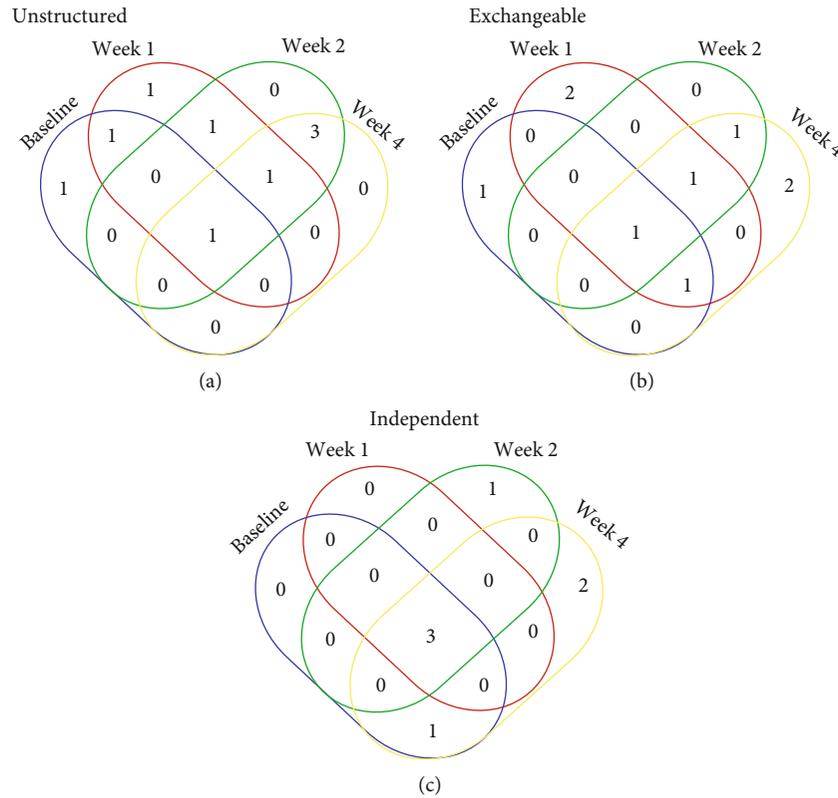


FIGURE 2: Venn diagram of identified lncRNAs for baseline, at weeks 1, 2, and 4, respectively, by different working correlation structures. (a) Under the unstructured working correlation structure. (b) The exchangeable working structure. (c) The independent working structure.

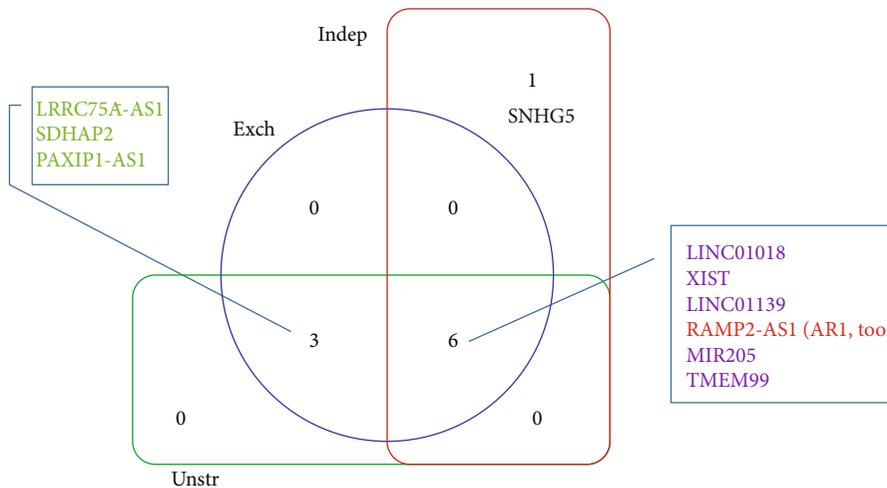


FIGURE 3: Venn diagram of integrated lncRNAs by three working correlation structures.

method, a GEE model was replaced by a linear mixed model (the outcome and the covariate are the same as those in the GEE-screening model), and the intercept term was regarded as a random effect. The lncRNAs with corresponding  $p$  values of the coefficients  $<0.05$  were selected as being relevant in both competing methods. Then, a support vector machine model was fit using the response status as the outcome and the identified lncRNAs by a specific method as predictors. According to the 10-fold crossvalidation results (to estimate the predictive performance of each method), the GEE-

TGDR method achieved the best predictive accuracy (Table 2). Of note, even GEE-TGDR has the best performance compared to the other two competitive methods, its predictive accuracy is estimated as 70%, which is far from 100%, leaving a large space to be improved.

**3.3. Biological Relevance.** In order to gain biological insight-identified biomarkers, we evaluated the relevance to psoriasis of the 10 identified lncRNA using disease confidence scores, where a high score represents a solid support by the literature

TABLE 2: Comparison between the GEE-TGDR method and two competing algorithms.

Method	Size	Predictive error
GEE-TGDR	9	30%
GEE-based screening	50	40%
Linear mixed model-based screening	27	33.33%

\*The predictive errors were calculated on the basis of 10-fold crossvalidations. Here, the response status, i.e., if the PASI score experienced a reduction of 75% from the baseline affected skin after week 12 or later. Size: the number of identified lncRNAs by a specific method; here, the sizes trained on the whole dataset were given; in crossvalidations, these numbers were subject to changes since the training sets were a subset of the whole dataset. For GEE-TGDR- and GEE-based screening, only unstructured working correlation matrix was considered.

according to the GeneCards database. None of the 10 lncRNAs were directly related to psoriasis while 5 lncRNAs, listed in a descending order for the confidence scores and thus descending support by the literature according to the GeneCards database, *MIR205*, *XIST*, *SNHG5*, *LINC01139*, and *SDHAP2* were associated with immunity.

Little meaningful information was extracted from currently annotated lncRNA databases, no surprisingly since that psoriasis remains largely unexplored from the perspective of lncRNAs. We thus focused on studying the mRNAs correlated or targeted by these lncRNAs. Specifically, we identified the genes whose baseline lesional expression was strongly correlated with at least one of the 10 lncRNA ( $|\text{Spearman correlation coefficient}| > 0.6$ , 5) and identified 225 mRNAs genes. According to the GeneCards database [22], approximately 30% of these mRNAs (64) were directly related to psoriasis, most notably *IL10*, *FABP5*, *KRT16*, *CCR6*, *IL18*, *STAT3*, *GATA3*, and *SERPINB3*, providing some validation of the lncRNA biomarkers identified by the GEE-TGDR method. In contrast, among the 29 target mRNAs identified by the lncRNA Disease 2.0 database [23] as targeted by the 10 lncRNA panel (all of which were identified by the correlation approach), GeneCards claimed that *CCR10*, *AOC3*, *UBB*, and *WNK4* were directly related to psoriasis, but only *CCR10* had a large confidence score for its relevancy to psoriasis. Of note, among the 10 lncRNAs, only *RAMP2-AS1*, *PAX1P1-AS1*, *TMEM99*, and *LIN01018* have many correlated mRNAs, but the other five have few or no correlated mRNAs at all.

**3.4. Enriched Pathways by Target mRNAs.** A gene-set over-representation analysis was carried out on the 225 mRNAs identified as targeted by the 10 lncRNA biomarker panel using the STRING software [24] on KEGG and GO collections. About 346 enriched biological process (BP) terms, 23 molecular function (MF) terms, and 21 cellular component (CC) terms were identified in the GO collection reflecting the immune pathophysiology of the disease. The top 3 enriched KEGG pathways [25] reflected the inflammatory processes not only identifying inflammatory bowel diseases (FDR < 0.001) and cytokine-cytokine receptor interaction (FDR = 0.005) but also zeroing on the hallmark pathway in psoriasis: Th17 differentiation (FDR = 0.031).

Lastly, among the 225 mRNA, we selected the top 10 in terms of psoriasis-relevance (confidence score for relevancy > 15) and constructed a lncRNA-mRNA interaction network, visualized by Cytoscape software [26] (Figure 4). We observed that the target mRNAs are highly connected, with *IL10* serving as a hub gene. It is well-known that *IL10* is an immunosuppressive cytokine and enables to maintain immunological homeostasis [27]. Based on this, we anticipate that identified lncRNAs may regulate the expression of important cytokines such as *IL10* and warrant further investigation.

## 4. Discussion

**4.1. Limitations and Future Work.** At current stage, the GEE-TGDR method has several limitations. First, no grouping structure is taken into account, and thus, the GEE-TGDR method belongs to the conventional embedded feature selection category. So far, accumulated studies [28–31] have shown that a pathway-based method that considers grouping information is superior to its gene-based counterpart in which grouping information is ignored. Thus, how to extend the proposed GEE-TGDR method to account for correlations among genes is a research avenue we will pursue in the near future.

Second, the TGDR method is much slower than the coordinate descent (CD) [15] method as shown by our previous study [4]. Given that the GEE-TGDR extension has the TGDR method as an optimization strategy, its speed of convergence is expected to be very slow. A method that combines the merits of these two algorithms together is definitely in demand. Alternatively, a sine cosine algorithm [20] may be integrated into the gradient descent step for a faster updating and a better tuning of hyperparameters (tuning parameters). Furthermore, the step increment  $\Delta v$  is fixed at a constant value in the current version. In the future, this parameter will be modified to update along the iterations, as in the Adam algorithm, which may boost the computing efficiency and avoid being stuck in a local minimum value as well.

Third, the GEE-TGDR method only takes time-invariant covariates in its current version. For longitudinal gene expression profiles, a summary score would be utilized to summarize each gene's expression values over time as one overall value. Consequently, covariates became time-invariant again. For example, the mean values of lncRNA expression profiles at baseline and week 1 can be used to represent the corresponding lncRNAs and then as the covariates to investigate they are associated with PASI scores at week 1, week 2, and week 4 or the change of PASI scores at those time points from the baseline levels. On the other hand, the GEE-TGDR method can be certainly extended to handle time-varying covariates, which can examine the impact of dynamic changes in gene expression values on the outcomes of interest and thus facilitate a timely adjustment on treatment strategies accordingly. Lastly, right now, the only type of outcomes is continuous; yet certainly, it can be extended to handle outcomes of other types, with the corresponding quasilielihood function acting as the objective function.

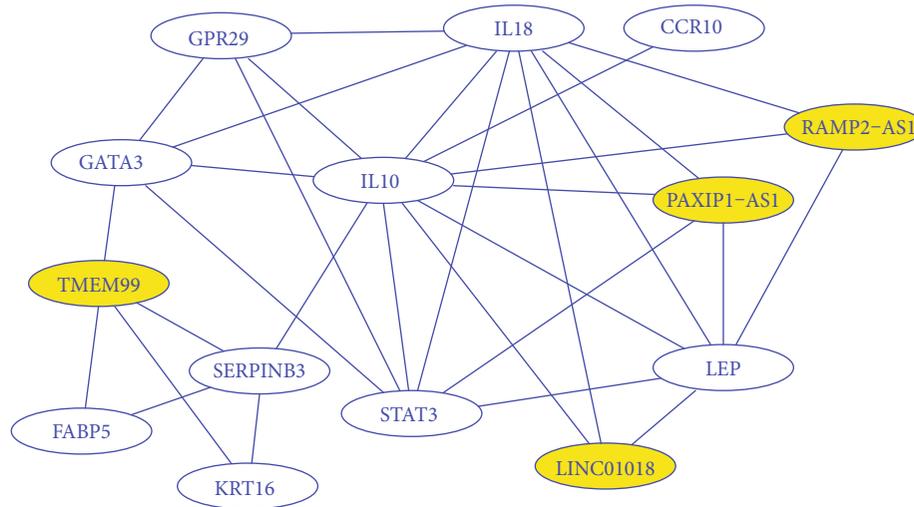


FIGURE 4: Resulting interaction network of identified lncRNAs and their correlated mRNAs. Here, only mRNAs with high enough confidence scores for the relevancy to psoriasis were considered. From the network, it is observed that IL10 is a hub gene directly connecting several other mRNAs and three identified lncRNAs. Four lncRNAs were highlighted in yellow, and the other six lncRNAs without correlated mRNAs were omitted from the graph.

**4.2. Contributions.** In this study, we propose a new feature selection algorithm that is capable of analyzing longitudinal outcomes and investigating the associations between gene expression profiles and the temporal changes of outcomes. In the psoriasis application, overfitting might be possible on the basis of the large discrepancy in MSE statistics between the whole training set and the crossvalidations. Even worse but more realistic, overfitting and underfitting may accompany each other to exist in a feature selection process. Since for real-world applications, the true relevant genes are unknown so the biological relevance is usually resorted to abstract some insight about the appropriation of identified gene lists. Nevertheless, for psoriasis and the underlying mechanism of immune treatments to combat this disease, little has been investigated from the perspective of lncRNAs to mine such relevant information. To the best of our knowledge, our work here is one of first efforts to unveil the mechanisms of psoriasis and its immune treatments using lncRNA expression profiles and a feature selection method specific for longitudinal data.

After the limitations of the GEE-TGDR method are addressed in the near future, we believe that a lncRNA signature will be harvested to tell precisely which patients would respond to a specific treatment from those who would not and thus facilitating personalized regimens or at least complementing other molecular markers for precise treatment strategies.

## 5. Conclusions

In this study, we proposed a novel feature selection algorithm—GEE-TGDR—capable of handling longitudinal outcomes and identifying relevant genes associated with the temporal changes of such outcomes.

Our future work will focus on eliminating the limitations of the GEE-TGDR method. In addition, extensions of the

current procedure to analyze other types of outcomes rather than continuous ones and a more efficient and faster implementation of updating coefficients are at the top of this list.

It is worth mentioning that besides dealing with longitudinal clinical outcomes, the GEE-TGDR can be adopted to inference the associations between lncRNAs and mRNAs and thus construct lncRNA-mRNA interaction networks. For example, using well-known cancer-related mRNAs as outcomes, the lncRNAs that may potentially regulate/target those mRNAs could be found with the aid of the GEE-TGDR method, which is also one of our future works. Therefore, we anticipate a widespread application of the GEE-TGDR method in omics data analysis.

## Data Availability

Preprocessed gene expression data (accession no.: GSE85034) along with patient's clinical information were downloaded from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>).

## Conflicts of Interest

No competing interests have been declared.

## Authors' Contributions

ST conceived and designed the study. ST and CW analyzed the data. CW and ST interpreted data analysis and results. ST and CW wrote the paper. All authors reviewed and approved the final manuscript.

## Acknowledgments

We thank Dr. Danna Gilbreath for the English editing. This study was supported by a fund (No. 31401123) from the

National Natural Science Foundation of China. Dr. Suarez-Farinas was also supported by the Irma T. Hirschl/Monique Weill-Coulier Research Award.

## References

- [1] P. Xu, L. Zhu, and Y. Li, "Ultrahigh dimensional time course feature selection," *Biometrics*, vol. 70, no. 2, pp. 356–365, 2014.
- [2] L. Wang, J. Zhou, and A. Qu, "Penalized generalized estimating equations for high-dimensional longitudinal data analysis," *Biometrics*, vol. 68, no. 2, pp. 353–360, 2012.
- [3] S. Tian, C. Wang, and H. H. Chang, "A longitudinal feature selection method identifies relevant genes to distinguish complicated injury and uncomplicated injury over time," *BMC Medical Informatics and Decision Making*, vol. 18, no. S5, p. 115, 2018.
- [4] S. Tian and C. Wang, "Feature selection for longitudinal data by using sign averages to summarize gene expression values over time," *BioMed Research International*, vol. 2019, Article ID 1724898, 12 pages, 2019.
- [5] S. L. Zeger and K. Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, vol. 42, no. 1, pp. 121–130, 1986.
- [6] Y. Zheng, Z. Fei, W. Zhang et al., "PGS: a tool for association study of high-dimensional microRNA expression data with repeated measures," *Bioinformatics*, vol. 30, no. 19, pp. 2802–2807, 2014.
- [7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [8] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.
- [9] J. H. Friedman and B. E. Popescu, *Gradient directed regularization for linear regression and classification*, Technical Report, Statistics Department, Stanford University, 2003.
- [10] S. Tian, H. H. Chang, C. Wang, J. Jiang, X. Wang, and J. Niu, "Multi-TGDR, a multi-class regularization method, identifies the metabolic profiles of hepatocellular carcinoma and cirrhosis infected with hepatitis B or hepatitis C virus," *BMC Bioinformatics*, vol. 15, no. 1, p. 97, 2014.
- [11] S. Tian and M. Suárez-Fariñas, "Multi-TGDR: a regularization method for multi-class classification in microarray experiments," *PLoS One*, vol. 8, no. 11, article e78302, 2013.
- [12] S. Tian and M. Suárez-fariñas, "Hierarchical-TGDR," *Systems Biomedicine*, vol. 1, no. 4, pp. 278–287, 2013.
- [13] S. Tian, C. Wang, and M.-W. An, "Test on existence of histology subtype-specific prognostic signatures among early stage lung adenocarcinoma and squamous cell carcinoma patients using a Cox-model based filter," *Biology Direct*, vol. 10, no. 1, pp. 1–17, 2015.
- [14] S. Tian, "Identification of subtype-specific prognostic genes for early-stage lung adenocarcinoma and squamous cell carcinoma patients using an embedded feature selection algorithm," *PLoS One*, vol. 10, no. 7, article e0134630, 2015.
- [15] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
- [16] X. Chen, C. C. Yan, X. Zhang, and Z. You, "Long non-coding RNAs and complex diseases : from experimental results to computational models," *Briefings in Bioinformatics*, vol. 18, no. 4, pp. 558–576, 2017.
- [17] J. Correa, J. Kim, S. Tian, L. E. Tomalin, and J. G. Krueger, "Shrinking the psoriasis assessment gap: early gene-expression profiling accurately predicts response to long-term treatment," *Journal of Investigative Dermatology*, vol. 137, no. 2, pp. 305–312, 2017.
- [18] J. Lin, Y. Fang, M. Tao et al., "LOC285194 inhibits proliferation of human keratinocytes through regulating miR-616/GATA3 pathway," *Molecular and Cellular Probes*, vol. 53, p. 101598, 2020.
- [19] A. Rakhshan, N. Zarrinpour, A. Moradi et al., "A single nucleotide polymorphism within HOX Transcript Antisense RNA (HOTAIR) is associated with risk of psoriasis," *International Journal of Immunogenetics*, vol. 47, no. 5, pp. 430–434, 2020.
- [20] L. Abualigah and A. Diabat, "Advances in sine cosine algorithm: a comprehensive survey," *Artificial Intelligence Review*, 2021.
- [21] S. Ma and J. Huang, "Regularized gene selection in cancer microarray meta-analysis," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–12, 2009.
- [22] M. Safran, I. Dalah, J. Alexander et al., "GeneCards version 3: the human gene integrator," *Database : the journal of biological databases and curation*, vol. 2010, article baq020, pp. 1–16, 2010.
- [23] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, "LncRNADisease 2.0 : an updated database of long non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 47, pp. D1034–D1037, 2019.
- [24] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, pp. D808–D815, 2013.
- [25] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG databases at GenomeNet," *Nucleic Acids Research*, vol. 30, no. 1, pp. 42–46, 2002.
- [26] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics (Oxford, England)*, vol. 27, no. 3, pp. 431–432, 2011.
- [27] L. Isac and J. Song, "Interleukin 10 promotor gene polymorphism in the pathogenesis of psoriasis," *Acta Dermatovenerologica Alpina Pannonica et Adriatica*, vol. 28, no. 3, pp. 119–123, 2019.
- [28] T. Zeng, W. W. Zhang, X. T. Yu et al., "Edge biomarkers for classification and prediction of phenotypes," *Science China Life Sciences*, vol. 57, no. 11, pp. 1103–1114, 2014.
- [29] H. Sun, W. Lin, R. Feng, and H. Li, "Network-regularized high-dimensional Cox regression for analysis of genomic data," *Statistica Sinica*, vol. 24, no. 3, pp. 1433–1459, 2014.
- [30] S. Tian, C. Wang, and B. Wang, "Incorporating pathway information into feature selection towards better performed gene signatures," *BioMed Research International*, vol. 2019, 12 pages, 2019.
- [31] J. Liu, J. Huang, and S. Ma, "Incorporating group correlations in genome-wide association studies using smoothed group Lasso," *Biostatistics*, vol. 14, no. 2, pp. 205–219, 2013.