Published in final edited form as:

Int J Data Min Bioinform. 2018 ; 20(4): 362–379. doi:10.1504/IJDMB.2018.094891.

DiffGRN: differential gene regulatory network analysis

Youngsoon Kim,

Department of Computer Science, Kennesaw State University, Marietta, GA, USA

Jie Hao,

Analytics and Data Science Institute, Kennesaw State University, Kennesaw, GA, USA

Yadu Gautam,

Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA

Tesfaye B. Mersha, and

Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA

Mingon Kang*

Department of Computer Science, Kennesaw State University, Marietta, GA, USA

Abstract

Identification of differential gene regulators with significant changes under disparate conditions is essential to understand complex biological mechanism in a disease. Differential Network Analysis (DiNA) examines different biological processes based on gene regulatory networks that represent regulatory interactions between genes with a graph model. While most studies in DiNA have considered correlation-based inference to construct gene regulatory networks from gene expression data due to its intuitive representation and simple implementation, the approach lacks in the representation of causal effects and multivariate effects between genes. In this paper, we propose an approach named Differential Gene Regulatory Network (DiffGRN) that infers differential gene regulation between two groups. We infer gene regulatory networks of two groups using Random LASSO, and then we identify differential gene regulations by the proposed significance test. The advantages of DiffGRN are to capture multivariate effects of genes that regulate a gene simultaneously, to identify causality of gene regulations, and to discover differential gene regulators between regression-based gene regulatory networks. We assessed DiffGRN by simulation experiments and showed its outstanding performance than the current state-of-the-art correlation-based method, DINGO. DiffGRN is applied to gene expression data in asthma. The DiNA with asthma data showed a number of gene regulations, such as ADAM12 and RELB, reported in biological literature.

Keywords

DiNA; differential network analysis; gene regulatory network

*Corresponding author Mingon Kang, Department of Computer Science, Kennesaw State University, Marietta, GA, USA, mkang9@kennesaw.edu.

1 Introduction

Identification of differential biological processes (e.g., gene regulation) with significant changes under disparate conditions is essential to understand complex biological functions and their roles in a disease. Over the past decade, Differential Expression Analysis (DEA) has been mainly considered to identify differential genes associated with a particular phenotype. DEA tests differences of gene expression levels in a single gene between two sample groups (de la Fuente, 2010; Anders and Huber, 2010). However, DEA has two following shortcomings. First, DEA often fails to capture small changes in expression for a single gene with multiple testing corrections (e.g., pairwise *t*-tests with Bonferroni corrections). Second, DEA assumes that gene expression is independent each other, although genes often cooperate with other genes in a biological system. It often causes a lack in identification of the multivariate effects of genes (Ideker and Krogan, 2012).

Gene Set Enrichment Analysis (GSEA) and network-based differential analysis have been considered to tackle the limitations (Rahmatallah et al., 2014; Rahmatallah et al., 2017). GSEA investigates differential expressions of gene sets (biological pathways or processes) instead of a single gene, which has shown robust analysis results to noise and outliers. Specifically, GSEA with biological pathway databases (e.g., KEGG and REACTOME) incorporates prior biological knowledge into the analysis models and enables one to interpret differential biological processes of pathways (Rahmatallah et al., 2014; Rahmatallah et al., 2017; Hung et al., 2012; Glaab et al., 2012).

Since biological processes are complex and involve a large number of interactions of different biological components, the consideration of their causal and interaction effects in the model is central to differential analysis. Differential Network Analysis (DiNA) examines different biological processes from biological networks (e.g., gene regulatory network), each of which is inferred from gene expression data of a group (Ideker and Krogan, 2012; Cho et al., 2012; Fukushima, 2013; Gambardella et al., 2013; Lichtblau et al., 2017). The differences in network topologies between two groups (i.e., disease versus healthy) show changes in interactions between genes.

Gene Regulatory Networks (GRNs) represent regulatory interactions between genes with a graph model, where a node and an edge indicate a gene and functional relationships between genes respectively (Marbach et al., 2012; Chai et al., 2014). GRNs can be computationally approximated by reverse engineering from gene expression data, which is called GRN *inference* (Hickman and Hodgman, 2011; Araki et al., 2013; Banf and Rhee, 2017). GRN inference can be mainly categorised into three: (1) correlation-based, (2) Bayesian-based, and (3) regression-based inference. Correlation-based GRN inference, which is the simplest approach of reconstructing GRN, builds a co-expression network by measuring interactions between genes with correlation coefficients, and the connections of the network are determined by a threshold. Correlation-based GRNs are undirected graphs, and only pairwise relationships between genes are considered. In contrast to the correlation-based GRNs, Bayesian-based approaches can capture a regulatory direction between genes by approximating conditional probabilities of genes. Moreover, Bayesian-based approaches can cope with a missing value problem and can integrate prior biological knowledge with prior

probabilities. However, Bayesian-based GRNs are infeasible to deal with large-scale genomic data because of its computational complexity, and is also challenging to accurately estimate the distribution with a small number of samples. Regression-based GRN inference considers multivariate effects of genes that regulate a gene based on a regression model, so it constructs a directed graph for GRN. Regression-based inference has shown the outstanding performance with ensemble learning in the assessment of GRN inference (Hill et al., 2016).

Most studies in DiNA have mainly considered correlation-based inference to construct gene regulatory networks from gene expression data, due to its intuitive representation and simple implementation. Differential Coexpression Profile (DCp) and Differential Coexpression Enrichment (DCE) were proposed to identify differential gene pairs of coexpression networks (Yu et al., 2011). DINGO examined genes belonging to biological pathways in glioblastoma and inferred differential correlation-based GRNs by decomposing them to global and group-specific components (Ha et al., 2015). Differential coexpression networks were constructed from gene expression data and protein-protein interaction datasets (Cheng et al., 2018). A Two Dimensional Joint Graphical Lasso (TDJGL) model enhanced the performance of coexpressed networks with gene expression profiles collected across multiple databases (Zhang et al., 2016).

However, differential coexpression networks lack in the representation of causal effects and multivariate effects between genes. Coexpression networks have mainly employed correlation coefficients, mutual information, and conditional mutual information, which construct indirect network graphs. However, multiple genes may regulate a gene simultaneously rather than only a single gene is involved in. Thus, large numbers of multivariate effects may fail to be captured in differential coexpression networks, since coexpression examines the interaction effects between genes by pairwise testing. Consequently, only substantially significant gene pairs would be identified. Whereas, regression-based GRN can capture multivariate effects of multiple genes that regulate a gene in test of significance. Furthermore, regression-based GRNs can construct directed graphs, which help to investigate casual effects between genes.

In this paper, we propose an approach named Differential Gene Regulatory Network (DiffGRN) that identify differential gene regulations between two groups. First, we infer gene regulatory networks of two groups by using Random LASSO. Then, we identify differential gene regulations by the proposed significance test. To the best of our knowledge, DiffGRN is the first study of DiNA based on regression-based gene regulatory network, while most differential network analyses are based on correlation-based gene regulatory network. The advantages of DiffGRN over existing methods include (1) to capture multivariate effects of genes that regulate a gene, (2) to identify causality of gene regulations, and (3) to discover differential gene regulators between regression-based gene regulatory networks.

The manuscript is organised as follows. We elucidate DiffGRN in Section 2. In Section 3, we show the experimental results with simulation data and compare the performance with a current state-of-the-art method. We investigate DiffGRN applying to asthma gene expression

data and induce the differential network between asthma and control groups in Section 4. Finally, we describe the biological findings by DiffGRN in Section 5.

2 Methods

We develop a novel method for DiffGRN analysis that infers differential networks between two groups (e.g., patients versus controls) by using high-dimensional gene expression data. DiffGRN infers group-specific GRNs and provides a statistical solution to perform test of significance between the group-specific linear regression-based GRNs.

2.1 Notations

We describe the notations used throughout this paper. Let $\mathbf{G}$ denote a matrix of gene expression, $\mathbf{G} \in \mathbb{R}^{n \times p}$, where n is a number of samples, and p is a number of genes, i.e., $\mathbf{G} = \{g_1, \dots, g_p\}$ and $g_i \in \mathbb{R}^n (1 \leq i \leq p)$. $\mathbf{G}_{-i}$ is a matrix $\mathbf{G}$ where the column vector of the i -th gene is replaced with ones, which represents gene expression data other than gene i . In this study, we consider datasets of two groups (control and treatment). The superscript of matrix or vector shows the group of the data. For instance, $\mathbf{G}^C$ and $\mathbf{G}^T$ indicate gene expression data of the control and treatment groups, respectively. Regulatory relationships between genes are represented by an adjacency matrix $\mathbf{B} \in \mathbb{R}^{p \times p}$. b_i is the i -th vector in $\mathbf{B}$, and b_{ij} is an element of $\mathbf{B} (1 \leq i, j \leq p)$. We use i and j to indicate an index of row and column in a matrix, respectively. We assume that there is no self-regulation in the gene regulatory network, i.e., $b_{ii}=0$.

2.2 Differential gene regulatory network inference

DiffGRN identifies different molecular regulators that interact with each other from group-specific gene regulatory networks. DiffGRN includes two steps: (1) to infer group-specific GRNs from gene expression data of two groups and (2) to identify statistically significant regulators between them (see Figure 1).

We consider a Linear Regression-based Gene Regulatory Network (LR-GRN) inference approach for constructing a group-specific gene regulatory network, while most differential gene regulatory networks are based on pairwise correlation metrics to measure the strength of associations between genes. A challenge of LR-GRN inference is accurate inference of the network with high-dimension, low-sample size data. To tackle the problem, we take advantages of Random LASSO for inferring a group-specific gene regulatory network.

Given group-specific gene regulatory networks, we perform statistical significance tests for DiNA. For the statistical significance test, we derive a formula to compute a p -value. Note that the test of significance is formulated based on LR-GRN. The details of the method are elucidated in the following sections.

2.3 Gene regulatory network inference

We first construct group-specific gene regulatory networks by a linear regression-based network inference approach. The gene i can be represented by linear effects of other genes. However, since only a few genes should be involved as gene regulators rather than all other

genes, a sparse regularisation (L-1 norm) is applied in the linear model. Then, an adjacency matrix $\mathbf{B}$ of a group-specific gene regulatory network is constructed by identifying of molecular regulators for each gene. For gene i , a sparse linear model is formulated as:

$$g_i = \mathbf{G}_{-i}b_i + \varepsilon_i, \text{ subject to } \|b_i\| \leq C_i, \quad (1)$$

where b_i is the coefficients of gene expressions other than gene i , $\|\cdot\|$ is an L-1 norm, and the residual is denoted as ε_i . The j -th coefficient element in b_i indicate a regulatory relationship from gene j to gene i (with a direction) in the linear model, where zero shows non-relationship between them. In contrast with correlation-based gene regulatory networks, linear regression-based gene regulatory networks can capture main effects of multiple genes. Correlation-based gene regulatory networks may fail to infer gene regulation if the correlation is not significantly high and if multiple genes regulate

simultaneously. The coefficient vector b_i for gene i is used to construct the adjacency matrix $\mathbf{B}$ of the group-specific gene regulatory network, i.e., $\mathbf{B} = \{b_1, \dots, b_p\}^T$. Then, equation (1) can be optimised by:

$$\operatorname{argmin} \|g_i - \mathbf{G}_{-i}b_i\|^2 + \lambda \|b_i\|, \quad (2)$$

where λ is a hyper-parameter for sparsity regularisation, and $\|\cdot\|$ is an L-2 norm of a vector.

The optimisation problem of equation (2) can be solved by a LASSO solution. For reliable gene regulatory network inference, we adapt Random LASSO (Wang et al., 2011). Random LASSO includes the two steps: (1) generating importance of features for genes and (2) training LASSO with the features weighted by importance. Since Random LASSO trains a small set of variables by using bootstrapping technique instead of directly training with the whole variables, the coefficient estimation is more reliable on each training with high-dimensional data. In this study, a significance test of coefficients (t -test with $\alpha_{GRN} = 0.05$) is evaluated on the sub-training of bootstrapping in the second step of Random LASSO to include only statistically significant factors in the network, where the coefficients of insignificant variables are set to zero. The group-specific gene regulatory networks are constructed separately on the two group data. The procedure of constructing the adjacency matrices $\mathbf{B}^T$ and $\mathbf{B}^C$ is described in Algorithm 1 for the group-specific gene regulatory network inference.

Algorithm 1:

```

1: for  $i \in \{1, \dots, p\}$  do
2:    $b_i^T = \text{Random LASSO in equation (2)}$ 
3:    $b_i^C = \text{Random LASSO in equation (2)}$ 
4: end for
5:  $\mathbf{B}^T = \{b_1^T, \dots, b_p^T\}$ 
6:  $\mathbf{B}^C = \{b_1^C, \dots, b_p^C\}$ 

```

2.4 Differential analysis from gene regulatory networks

In this section, we propose a strategy for significance test for differential analysis between the two group-specific gene regulatory networks. Let b_{ij}^C and b_{ij}^T be regulatory effects (coefficients) of gene j to gene i within the control and treatment group respectively, and $\hat{b}_{ij}^C$ and $\hat{b}_{ij}^T$ be estimated regression coefficients of b_{ij}^C and b_{ij}^T . The distribution of estimated differences between the two coefficients, $D_{ij} = \hat{b}_{ij}^T - \hat{b}_{ij}^C$, is as follows:

$$D_{ij} \sim \mathcal{N}\left(b_{ij}^T - b_{ij}^C, \sigma_{D_{ij}}^2\right), \quad (3)$$

where $\hat{b}_{ij}^T \sim \mathcal{N}\left(b_{ij}^T, \sigma_{\hat{b}_{ij}^T}^2\right)$, $\hat{b}_{ij}^C \sim \mathcal{N}\left(b_{ij}^C, \sigma_{\hat{b}_{ij}^C}^2\right)$ and $\sigma_{D_{ij}}^2$ is a variance of D_{ij} . The null and alternative hypotheses of difference are defined as

$$H_0: b_{ij}^T = b_{ij}^C, H_1: b_{ij}^T \neq b_{ij}^C. \quad (4)$$

The hypothesis test follows the procedure of a significance test between two sample means. Under the null hypothesis, the hypothesis is often tested by z-test (Clogg et al., 1995). A z-score (z_{ij}) for the differential analysis between the pair of gene i and j is computed by:

$$z_{ij} = \frac{\hat{b}_{ij}^T - \hat{b}_{ij}^C}{\sqrt{SE(\hat{b}_{ij}^T)^2 + SE(\hat{b}_{ij}^C)^2}}, \quad (5)$$

where $SE(\hat{b}_{ij}^T)$ and $SE(\hat{b}_{ij}^C)$ are the estimates of the standard error of coefficients associated with the treatment and control groups respectively, where the estimates of the variances of the sampling distribution are unbiased (Brame et al., 1998). Then, we can obtain a p -value from z-score of equation (5) using normal distribution. Note that z_{ij} follows the standard normal distribution, $Z \sim \mathcal{N}(0, 1)$. Given a threshold (ξ), significance tests are performed for the

all pairwise genes, i.e., b_{ij} , and an adjacency matrix $\mathbf{D} = \{d_{ij} | 1 \leq i, j \leq p\}$ is constructed for a differential network. The optimal p -value threshold is determined by using False Discovery Rate (FDR) with a significance level α . The procedure is described in Algorithm 2.

Algorithm 2

```

1: for  $i \in \{1, \dots, p\}$  do
2:   Compute  $\mathbf{B}^C$  and  $\mathbf{B}^T$  by Algorithm 1
3:    $b_{ij}^T = \begin{cases} \hat{b}_{ij}^T, & \text{if } \hat{b}_{ij}^T \text{ is nonzero and } p\text{-value}(\hat{b}_{ij}^T) < \alpha_{GRN} \\ 0, & \text{otherwise} \end{cases}$ 
4:    $b_{ij}^C = \begin{cases} \hat{b}_{ij}^C, & \text{if } \hat{b}_{ij}^C \text{ is nonzero and } p\text{-value}(\hat{b}_{ij}^C) < \alpha_{GRN} \\ 0, & \text{otherwise} \end{cases}$ 
5:    $SE(b_{ij}^T) = \begin{cases} SE(\hat{b}_{ij}^T), & \text{if } \hat{b}_{ij}^T \text{ is nonzero and } p\text{-value}(\hat{b}_{ij}^T) < \alpha_{GRN} \\ 0, & \text{otherwise} \end{cases}$ 
6:    $SE(b_{ij}^C) = \begin{cases} SE(\hat{b}_{ij}^C), & \text{if } \hat{b}_{ij}^C \text{ is nonzero and } p\text{-value}(\hat{b}_{ij}^C) < \alpha_{GRN} \\ 0, & \text{otherwise} \end{cases}$ 
7:   Compute  $z$ -score by equation (5)
8:    $p_{ij} = P(Z > |z_{ij}| \times 2)$ 
9:    $d_{ij} = \begin{cases} 1, & \text{if } p_{ij} < \xi \text{ for } \alpha \\ 0, & \text{otherwise} \end{cases}$ 
10:  Construct a differential network from the adjacency matrix  $\mathbf{D}$ 
11: end for

```

3 Simulation studies

We conducted simulation experiments to evaluate our proposed method. Due to few available well-known true models of biological networks, the assessment with real biological data of complex organisms such as human is challenging. Thus, the performance was indirectly evaluated with simulation data that implements biological networks of two different conditions where a true model is known.

In the simulation studies, we aim (1) to verify that our proposed method produces robust performance to identify the true model of DiffGRN from gene expression data of two groups, and (2) to compare the performance with a current state-of-the-art method. We carried out the following three experiments with the simulation data: (1) Receiver Operating Characteristic (ROC) curve, (2) False Discovery Rate (FDR), and (3) Sensitivity analysis.

3.1 Simulation settings

We generated the simulation data under the assumption that we hypothesised in this paper, where main effects of multiple genes were considered to generate the gene expression data. Synthetic gene expressions ($\mathbf{G}^T$ and $\mathbf{G}^C$) were generated with the pre-designed ground truth adjacency matrices ($\mathbf{Z}^T$ and $\mathbf{Z}^C$) for gene regulatory networks of two groups. Binary adjacency matrices, each of which is a sparse acyclic graph without self-loop, were

randomly pre-defined. The probability of edges between two arbitrary vertices was 0.005. Then, the synthetic gene expression data matrix was generated by:

$$\mathbf{G}^T = \mathbf{E}^T(\mathbf{I} - \mathbf{Z}^T)^{-1}, \mathbf{G}^C = \mathbf{E}^C(\mathbf{I} - \mathbf{Z}^C)^{-1}, \quad (6)$$

where $\mathbf{I} \in \mathbb{R}^{p \times p}$ is an identity matrix, and $\mathbf{E}^{T/C} \in \mathbb{R}^{n \times p}$ is a matrix with normally distributed random values for noise, i.e., $\mathbf{E}^{T/C} \sim N(0, 0.01)$.

We compared the performance of our differential gene regulation between groups with correlation-based gene regulatory network called DINGO (Ha et al., 2015).

3.2 Experimental results with simulation data

First, we evaluated the performance by computing the Area Under the Receiver Operating Characteristic Curve (AUROC). We generated 100 samples of 200 genes ($p = 200$) for each group ($n_C = 100$ and $n_T = 100$). Confusion matrices were computed with various thresholds of p -value. The confusion matrix is defined as:

- True Positive (TP): correctly identifies the presence of differential gene regulations,
- False Positive (FP): incorrectly identifies the non-presence of differential gene regulations,
- False Negative (FN): incorrectly identifies the non-presence of differential gene regulations, and
- True Negative (TN): correctly identifies the non-presence of differential gene regulations.

For each experiment, we constructed a true model of differential network and generated gene expression data of two groups. Then, we examined True Positive Rate ($\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$) and False Positive Rate ($\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$) over various thresholds of p -value. We repeated the simulation experiment five times and averaged ROC curve was traced. Figure 2 illustrates the ROC curves of DiffGRN and DINGO, and their AUROCs are shown in Table 1. The AUROCs show that DiffGRN (0.9606) produces substantially higher performance than DINGO (0.5110).

Secondly, we conducted experiments for False Discovery Rate (FDR) to obtain a cut-off p -value with the confidence level of 0.05. FDR was measured by generating gene expression data of empty ground truth network. All significant gene pairs were considered as misclassified. The p -value that controls the FDR at level $\alpha = 0.05$ was $1e-45$ for DiffGRN.

Lastly, we measured the sensitivity of DiffGRN with the p -value threshold ($1e-45$) obtained by the FDR experiments. DiffGRN produced the sensitivity of 0.7609 ± 0.0290 . In contrast, DINGO produced no misclassification in this simulation data (i.e., $\text{FDR} = 0$). It is because correlation-based DINGO shows the statistical significance only if the correlation is significantly high. It shows that correlation-based DiNA is stringent, so it may be a suitable

to maximise true positives, but also increase false negatives. Moreover, the lower AUROC of DINGO indicates that correlation-based DiNA lack to identify multivariate effects of gene regulations from multiple genes.

4 Differential network analysis in asthma

4.1 Data description and preprocessing

We applied DiffGRN to gene expression in asthma for identifying differential networks. The study population consisted of inner-city children aged 6 to 12 years with asthma (cases, $n = 97$) and without asthma (healthy control subjects, $n = 97$) recruited by Inner-City Asthma Consortium (Busse and Mitchell, 2007). Cases of asthma were required to meet the following criteria: (1) a physician diagnosis of asthma, (2) persistent or uncontrolled disease as defined by the National Asthma Education and Prevention Program (National Heart Lung and Blood Institute, 2007), (3) physiologic evidence of asthma ($FEV1 < 85\%$ predicted, or $FEV1/FVC$ ratio $< 85\%$ and bronchodilator responsiveness ($> 12\%$), or $PC20 < 8$ mg/ml of methacholine), and (4) positive prick skin-test to at least one of a panel of indoor aeroallergens (i.e., dust mite, cockroach, mould, cat, dog, rat, or mouse). Controls were required to have: (1) no medical history of asthma, rhinitis, sinusitis, and atopic dermatitis, (2) an $FEV1 > 85\%$ predicted, and (3) no positive prick skin-tests.

Peripheral Blood Mononuclear Cells (PBMCs) were isolated from whole blood by using the Ficoll density gradient separation. RNA was isolated from the PBMCs using the AllPrep RNA kit (Qiagen, Germantown, MD), RNA samples were quantified and purity assessed by using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE). RNA integrity was determined using the Bioanalyser (Agilent, Santa Clara, CA). RNA was reverse transcribed and amplified using the TransPlex Whole Transcriptome Amplification (WTA) Kit (Sigma-Aldrich, St. Louis, MO) followed by Cy3 labelling using Nimblegen one-color labelling kits, hybridisation on Nimblegen Human Gene Expression arrays ($12 \times 13.5k$), washed on the Little Dipper station (SciGene, Sunnyvale, CA), and scanned using Nimblegen's scanner (Roche Nimblegen, Madison, WI). The images were processed using Nimblegen's DEVA software. All samples were assayed once with 194 arrays performed in 9 batches. All mRNA data met the quality control criteria established by the Tumour Analysis Best Practices Working Group. Gene expression was assessed on Nimblegen Human Gene Expression arrays ($12 \times 13.5k$).

To reduce the genes, we performed pairwise t -test for the gene expression data of the two groups, and selected 214 genes with p -value less than 0.05. Then, the asthma gene expression data of 214 genes was normalised on each group. In short, we finally considered gene expression data that includes 194 samples of 214 genes, which consists of 97 samples evenly in asthma and control groups.

To control the false discovery rate for reliable significance test of differential networks, we examined the optimal p -value cut-off by randomising the data set. First, we shuffled the normalised data of both two groups and computed FDRs with various p -values. After repeating the processes ten times, the cut-off threshold was selected as the p -value with the

confidence level of 0.05. Figure 3 shows the average of FDR over various p -values. The p -value that controls FDR of 0.05 was $1e-8$. We used the p -value for the differential analysis.

4.2 Experimental result with asthma data

A total of 1,826 differential gene regulators were identified by DiffGRN. Among them, we depicted a sub-set of the differential network in asthma (p -value $< 1e-150$ and FDR=0.002) in Figure 4. In the figure, six genes of GZMA, WASHC1, CD83, TRBV19, RELB, and AGTR2 show higher in degrees than others. Among them, GZMA (Radom-Aizik et al., 2009), CD83 (Pfeffer et al., 2017), AGTR2 (Wang, 2008), and RELB (Weng et al., 2018; Tan et al., 2016) have been reported as genes that play important roles in development of asthma in biological literature.

There are nine differential comparisons between asthma and control for the effects in differential gene regulators, since the signs of the effects can be {Negative (N), Zero (Z), Positive (P)} on each group. Positive (P) and Negative (N) indicate significant effects of positive and negative signs respectively, while Zero (Z) shows insignificant effect of a gene. We denote the nine cases as NN, NZ, NP, ZN, ZZ, ZP, PN, PZ, and PP, where the first letter indicates the effect type in asthma and the second stands for in the control group. For instance, NZ shows that the effect of gene regulators is negative in asthma (N), while there is no significant effect (Z) in the control group. In our experimental results, we found four major categories of the effects of {NZ, PZ, ZN and ZP} in differential gene regulators. The effects of the gene regulators are visualised in the Figure 5. For instance, RPS28P4 is regulated by five genes: PTPN9, PPP2R5D, PPP3CB (Figure 4) and RPL5, SELPLG (Figure 4). These genes show significant negative or positive effects to regulate gene RPS28P4 simultaneously in asthma, but no significant effect in the control group.

The pathways which are related to the identified genes are listed in Table 2. Immune system pathway is the most enriched pathways based on differential gene regulators, where 10 genes of the pathways are involved in the differential network. Asthma involves adaptive and innate, antigen-independent immune system dysfunction. Immune system pathways have been implicated in the development and progression of asthma.

Our results show multiple genes related to asthma are differentially regulated or involve in differential regulation of other genes. ADAM12 is known to play important roles in tissue growth and remodelling and linked to progression and diagnosis of several disease conditions including asthma. ADAM12 is shown to unregulated the airways epithelial cells during allergic reaction and may contribute to the remodelling of the bronchial airways and increase neutrophil recruitment within airways mucosa, leading to chronic asthma symptoms (Nyren-Erickson et al., 2013; Estrella et al., 2009). In Egyptian children, ADAM12 genetic variants are highly associated to expression level of gene expression of ADAM12 in sputum (Shalaby et al., 2016). DiffGRN identified two differential gene regulators RNF135 and SEMA4F of ADAM12 with RNF135 showed positive and SEMA4F showed negative effect in asthmatic while both genes have no effect on control (see Figure 4 and Figure 4). The role of RNF135 and SEMA4F in asthma is not known, however, RNF135, a member of Ring finger protein, is essential for the regulation of innate immune response against viral infection (Oshiumi et al., 2009; Oshiumi et al., 2010). RELB is essential for the dendritic

cells function and maturation and deficiency of RELB is associated with increased inflammatory cell influx into the airways, levels of chemokines and T-helper cell type 2-associated cytokines (IL-4/5) in lung tissues, serum IgE, and airway remodelling (Nair et al., 2017). DiffGRN identified multiple differential regulators for RELB such as SMAD7 and IL2RG (Figure 4). Studies have shown that overexpression of SMAD7 alleviates allergic inflammation and airways remodelling (Luo et al., 2014), which is in line with our results showing that SMAD7 has positive effects on RELB among asthmatic while no effects on control (Figure 4). IL2RG is an important signalling component of many interleukin receptors, including those of interleukins IL-2, IL-4, and IL -7 which are important biomarkers of asthma (Kim et al., 2016).

5 Conclusions

We propose a DiffGRN for identifying differential biological processes shown in disparate conditions. DiffGRN is the first DiNA model which is based on regression-based gene regulatory networks. Thus, DiffGRN can capture differential multivariate effects of multiple genes that regulate a gene, which is more plausible model than correlation-based models. We demonstrated our novel approach using simulation and real datasets. We showed that our new method has improved ROC than existing methods in the simulation studies. Through asthma gene expression real data, our differential regulated network analysis showed that immune system pathway as most enriched pathways. Immune system pathways have been implicated in the development and progression of asthma.

The integration of multi-omics data has been currently highlighted due to complex multi-modality in biological systems. DiffGRN can be easily extended for multi-omics data for further analysis. The interaction effects with multi-omics data (e.g., DNA methylation and Copy Number Variations (CNV)) can be integrated into the regression model of gene regulatory network inference in equation (1), as being proposed in Zarayeneh et al., 2017.

The consideration of whole gene expression in the model without gene selection may be extremely challenging due to the curse of dimensionality problem. Although the regression-based gene regulatory network inference may cause less computational costs than correlation-based and Bayesian-based approaches, the problem has yet to be studied in depth. The solution would be to utilise parallel computing systems and big data frameworks such as Apache Spark. This study would play an important role in further research such as biomarker discovery and drug development.

Acknowledgement

This research was supported in part by the National Institutes of Health (NIH) grant R01HL132344.

Biographical notes:

Youngsoon Kim is a Postdoctoral Researcher in the Department of Computer Science at Kennesaw State University. She received her Master's and PhD degrees in Information Processing at the Gyeongsang National University in South Korea. She has conducted various researches in data mining and big data analytics as a senior researcher in

GyeongNam Development Institute. Her research interests include bioinformatics, machine learning, and big data analytics.

Jie Hao is a PhD Candidate in Analytics and Data Science Institute at Kennesaw State University. She received her Master's degree in Mathematics at East Tennessee State University and BS degree in Statistics (First Class) from North China University of Technology. Her research areas of interest are machine learning, deep learning, and bioinformatics.

Yadu Gautam is currently a Postdoctoral Fellow at Cincinnati Children's Hospital Medical Centre. He is working on genomics projects aimed at detecting predisposing genes and environmental exposure risk factors in asthma and related phenotypes. He is interested in both application and computational tools development in big biological data to investigate the genetics and environment risk factors for common and complex diseases

Tesfaye Mersha is currently an Associate Professor at the Cincinnati Children's Hospital Medical Centre and University of Cincinnati, where he leads the Population Genetics, Ancestry, and Bioinformatics (pGAB) Laboratory. His research combines quantitative, ancestry and statistical genomics to unravel genetic and non-genetic contributions to complex diseases and racial disparities in human populations, particularly asthma and asthma-related allergic disorders. Much of his research is at the interface of genetic ancestry, statistics, bioinformatics, and functional genomics, and he is interested in cross-line disciplines to unravel the interplay between genome and environment underlying asthma risk.

Mingon Kang is an Assistant Professor in the Department of Computer Science at Kennesaw State University, where he is a Director of DataX Research Lab. He received his Master's and PhD degrees in Computer Science at the University of Texas at Arlington. His research interests include bioinformatics, machine learning, and big data analytics.

References

- Anders S and Huber W (2010) 'Differential expression analysis for sequence count data,' *Genome Biology*, Vol. 11, No. 10.
- Araki R, Seno S, Takenaka Y and Matsuda H (2013) 'An estimation method for a cellular-state-specific gene regulatory network along tree-structured gene expression profiles,' *Gene*, Vol. 518, No. 1, pp.17–25. [PubMed: 23266644]
- Banf M and Rhee SY (2017) 'Computational inference of gene regulatory networks: approaches, limitations and opportunities,' *Biochimica et Biophysica Acta – Gene Regulatory Mechanisms*, Vol. 1860, No. 1, pp.41–52. [PubMed: 27641093]
- Brame R, Paternoster R, Mazerolle P and Piquero A (1998) 'Testing for the equality of maximum-likelihood regression coefficients between two independent equations,' *Journal of Quantitative Criminology*, Vol. 14, No. 3, pp.245–261.
- Busse WW and Mitchell H (2007) 'Addressing issues of asthma in inner-city children,' *Journal of Allergy and Clinical Immunology*, Vol. 119, No. 1, pp.43–49. [PubMed: 17208585]
- Chai LE, Loh SK, Low ST, Mohamad MS, Deris S and Zakaria Z (2014) 'A review on the computational approaches for gene regulatory network construction', *Computers in Biology and Medicine*, Vol. 48, pp.55–65. [PubMed: 24637147]

- Cheng L, Han Y, Zhao X, Xu X and Wang J (2018) 'Identifying pathway modules of tuberculosis in children by analyzing multiple different networks,' *Experimental and Therapeutic Medicine*, Vol. 15, No. 1, pp.755–760. [PubMed: 29399082]
- Cho D-Y, Kim Y-A and Przytycka TM (2012) 'Chapter 5: network biology approach to complex diseases,' *PLoS Computational Biology*, Vol. 8, No. 12, p.e1002820. [PubMed: 23300411]
- Clogg CC, Petkova E and Haritou A (1995) 'Statistical methods for comparing regression coefficients between models,' *American Journal of Sociology*, Vol. 100, No. 5, pp.1261–1293.
- de la Fuente A (2010) From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases', *Trends in Genetics*, Vol. 26, No. 7, pp.326–333. [PubMed: 20570387]
- Estrella C et al. (2009) 'Role of a disintegrin and metalloprotease-12 in neutrophil recruitment induced by airway epithelium,' *American Journal of Respiratory Cell and Molecular Biology*, Vol. 41, No. 4, pp.449–458. [PubMed: 19213876]
- Fukushima A (2013) 'DiffCorr: an R package to analyze and visualize differential correlations in biological networks,' *Gene*, Vol. 518, No. 1, pp.209–214. [PubMed: 23246976]
- Gambardella G, Moretti MN, de Cegli R, Cardone L, Peron A and di Bernardo D (2013) 'Differential network analysis for the identification of condition-specific pathway activity and regulation,' *Bioinformatics*, Vol. 29, No. 14, pp.1776–1785. [PubMed: 23749957]
- Glaab E, Baudot A, Krasnogor N, Schneider R and Valencia A (2012) 'EnrichNet: network-based gene set enrichment analysis,' *Bioinformatics*, Vol. 28, No. 18, pp.i451–i457. [PubMed: 22962466]
- Ha MJ, Baladandayuthapani V and Do KA (2015) 'DINGO: differential network analysis in genomics,' *Bioinformatics*, Vol. 31, No. 21, pp.3413–3420. [PubMed: 26148744]
- Hill SM et al. (2016) 'Inferring causal molecular networks: empirical assessment through a community-based effort,' *Nature Methods*, Vol. 13, No. 4, pp.310–322. [PubMed: 26901648]
- Hickman GJ and Hodgman TC (2011) 'Inference of gene regulatory networks using boolean-network inference methods,' *Journal of Bioinformatics and Computational Biology*, Vol. 7, No. 6, pp.1013–1029.
- Hung JH, Yang TH, Hu Z, Weng Z and DeLisi C (2012) 'Gene set enrichment analysis: performance evaluation and usage guidelines,' *Briefings in Bioinformatics*, Vol. 13, No. 3, pp.281–91. [PubMed: 21900207]
- Ideker T and Krogan NJ (2012) 'Differential network biology', *Molecular Systems Biology*, Vol. 8, p. 565. [PubMed: 22252388]
- Kim HY and Umetsu DT and Dekruyff RH (2016). 'Innate lymphoid cells in asthma: will they take your breath away?', *European Journal of Immunology*, Vol. 46, No. 4, pp.795–806. [PubMed: 26891006]
- Luo X et al. (2014) 'In vivo disruption of TGF – Signaling by Smad7 in airway epithelium alleviates allergic asthma but aggravates lung carcinogenesis in mouse,' *PLoS ONE*, Vol. 5, No. 4.
- Lichtblau Y, Zimmermann K, Haldemann B, Lenze D, Hummel M and Leser U (2017) 'Comparative assessment of differential network analysis methods,' *Briefings in Bioinformatics*, Vol. 18, No. 5, pp.837–850. [PubMed: 27473063]
- Marbach D et al. (2012) 'Wisdom of crowds for robust gene network inference,' *Nature Methods*, Vol. 9, No. 8, pp.796–804. [PubMed: 22796662]
- Nair PM et al. (2017) 'RelB-deficient dendritic cells promote the development of spontaneous allergic airway inflammation,' *American Journal of Respiratory Cell and Molecular Biology*, Vol. 58, No. 3, pp.352–365.
- National Heart Lung and Blood Institute (2007) 'Expert panel report 3 (EPR-3): guidelines for the diagnosis and management of asthma-summary report 2007,' *Journal of Allergy Clinical Immunology*, Vol. 120, No. 5 Suppl, pp.S94–138. [PubMed: 17983880]
- Nyren-Erickson EK, Jones JM, Srivastava DK and Mallik S (2013) 'A disintegrin and metalloproteinase-12 (ADAM12): function, roles in disease progression, and clinical implications,' *Biochimica et Biophysica Acta*, Vol. 1830, No. 10, pp.4445–4455. [PubMed: 23680494]
- Oshiumi H, Matsumoto M, Hatakeyama S and Seya T (2009) 'Riplet/RNF135, a RING finger protein, ubiquitinates RIG-I to promote interferon-beta induction during the early phase of viral infection,' *Journal of Biological Chemistry*, Vol. 284, No. 2, pp.807–817. [PubMed: 19017631]

- Oshiumi H et al. (2010) 'The ubiquitin ligase riplet is essential for RIG-I-dependent innate immune responses to RNA virus infection,' *Cell Host and Microbe*, Vol. 8, No. 6, pp.496–509. [PubMed: 21147464]
- Pfeffer PE, Ho TR, Mann EH, Kelly FJ, Sehlstedt M, Pourazar J, Dove RE, Sandstrom T, Mudway IS and Hawrylowicz CM (2017) 'Urban particulate matter stimulation of human dendritic cells enhances priming of naive CD8 T lymphocytes,' *Immunology*, Vol. 153, No. 4, pp.502–512. [PubMed: 29044495]
- Rahmatallah Y, Emmert-Streib F and Glazko G (2014) 'Gene sets net correlations analysis (GSNCA): amultivariate differential coexpression test for gene sets,' *Bioinformatics*, Vol. 30, No. 3, pp.360–368 [PubMed: 24292935]
- Rahmatallah Y, Zybaïlov B, Emmert-Streib F and Glazko G (2017) 'GSAR: bioconductor package for Gene Set analysis in R,' *BMC Bioinformatics*, Vol. 18, No. 1.
- Radom-Aizik S, Zaldivar F Jr., Leu S-Y and Cooper DM (2009) 'Brief bout of exercise alters gene expression in peripheral blood mononuclear cells of early- and late-pubertal males,' *Pediatric Research*, Vol. 65, No. 4, pp.447–452. [PubMed: 19127215]
- Shalaby SM et al. (2016) 'ADAM33 and ADAM12 genetic polymorphisms and their expression in Egyptian children with asthma,' *Annals of Allergy, Asthma and Immunology*, Vol. 116, No. 1, pp. 31–36.
- Tan H, Pan P, Zhang L, Cao Z, Liu B, Li H and Su X (2016) 'Nerve growth factor promotes expression of costimulatory molecules and release of cytokines in dendritic cells involved in Th2 response through LPS-induced p75NTR,' *Journal of Asthma*, Vol. 53, No. 10, pp.989–998. [PubMed: 27437725]
- Wang S, Nan B, Rosset S and Zhu J (2011) 'Random lasso,' *Annals of Applied Statistics*, Vol. 5, No. 1, pp.468–485. [PubMed: 22997542]
- Wang T, Yin KS, Liu KY, Lu GJ, Li YH and Chen JD (2008) 'Effect of valsartan on the expression of Angiotensin II receptors in the lung of chronic antigen exposure rats,' *Chinese Medical Journal*, Vol. 121, No. 22, pp.2312–2319. [PubMed: 19080339]
- Weng C-M, Lee M-J, He J-R, Chao M-W, Wang C-H and Kuo H-P (2018) 'Diesel exhaust particles up-regulate interleukin-17A expression via ROS/NF- κ B in airway epithelium,' *Biochemical Pharmacology*, Vol. 151, pp. 1–8. [PubMed: 29499168]
- Yu H, Liu B-H, Ye Z-Q, Li C, Li Y-X and Li YY (2011) 'Link-based quantitative methods to identify differentially coexpressed genes and gene pairs,' *BMC Bioinformatics*, Vol. 12.
- Zhang X-F, Ou-Yang L, Zhao XM and Yan H (2016) 'Differential network analysis from cross-platform gene expression data,' *Scientific Reports*, Vol. 6.
- Zarayeneh N, Ko E, Oh JH, Suh S, Liu C, Gao J, Kim D and Kang M (2017) 'Integration of multi-omics data for integrative gene regulatory network inference,' *International Journal of Data Mining and Bioinformatics*, Vol. 18, No. 3, pp.223–239. [PubMed: 29354189]

Figure 1.
DiffGRN workflow

Figure 2.
Overall ROC curves

Figure 3.
FDR curve for p value

Table 1

AUROC_s of DiffGRN and DINGO

	DiffGRN	DINGO
AUROC	0.9606	0.5110

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Enriched pathways from differential gene regulatory network between asthma versus controls

<i>Pathway name</i>	<i>Pathway size</i>	<i>Genes</i> [*]	<i>Gene name</i>
Immune system	407	10	IL2RG, RNF41, MAP3K8, ATF2, IFI6, PPP2R5D, KIR2DL4, USP18, OAS3, CSF2RB
Metabolism of lipids and lipoproteins	368	9	LDLR, PPP1CB, ABCG1, LSS, ACADM, ACAA1, MTMR1, AGPAT4, SPTLC2
Adaptive immune system	417	7	RNF41, MAP3K8, PSME2, PPP2R5D, KIR2DL4, BTRC, UBE3C
Cytokine signaling in immune system	232	7	IL2RG, MAP3K8, IFI6, USP18, BTRC, OAS3, CSF2RB
Cytokine receptor interaction	218	6	CCL23, IL2RG, CCR3, BMPR2, IL18R1, CSF2RB
Class I PI3K signalling events	58	6	IL2RG, PPP3CB, ATF2, RELB, IL18R1, GZMA
Haemostasis	397	6	SERPINB2, ACTN2, HIST1H3H, ITGA5, HIST1H3G, PPP2R5D
Regulation of nuclear SMAD2/3 signalling	75	5	ATF2, PIAS3, SMAD7, SKIL, TGIF1
Cell cycle	330	5	POLA1, AKAP9, PSME2, PPP2R5D, BTRC

* indicates the number of genes which are found in the result of the differential analysis in Asthma.