Identifying Relevant Genes Related Atopic Dermatitis to using Transcriptomic

Creel, Kristin
ISPED
University of Bordeaux
kristin.creel@etu.u-bordeaux.fr

Dong, Larry
Department of Epidemiology, Biostatistics
and Occupational Health
McGill University
larry.dong@mail.mcgill.ca

Nama Ravi, Sneha Keerthi
ISPED
University of Bordeaux
sneha-keerthi.nama-ravi@etu.u-bordeaux.fr

Abstract

Background — test1
Methods — test2
Results — test3
Conclusion — test4
Keywords: Atopic dermatitis, transcriptomics, differential gene expression, unsupervised learning, prediction model.

1. Introduction

test

Atopic dermatitis (AD) or atopic eczema is an itchy, inflammatory skin condition characterised by poorly defined erythema with edema vesicles, and weeping in the acute stage and lichenification in the chronic stage. The global prevalence is 15-20 % in children and 1-3% in adults, posing a significant burden on health-care resources and patients' quality of life [9].

The etiological factors associated with the initiation and progress of the disease are known to be genetic, environmental and immunological that affects the epithelial barrier-immunity interplay [10].

Clinical investigations and discoveries in molecular medicine have positively identified 46 genes linked to AD. Mutations in filaggrin (FLG) genes (influencing intermediate filament protein filaggrin expression) are most common in the AD diseased population, it affects 10-50% of AD patients worldwide.

Few additional barrier genes encoded by the epidermal differentiation complex (EDC) locus chromosome 1q21, including claudins, loricrin (LOR), involucrin (IVL), SPINK5, AND tmem79/matt, are also associated with AD.

The genes of innate immune system like NOD1, NOD2, TLR2, CD14, and DEFB1, that encode the integral factors in cutaneous immunologic response to non-specific antigens may also experience mutations and cause AD [5].

Studies have identified FLG gene mutation to be the most significant risk factor for AD, followed by the genes in the type 2 T helper lymphocyte (Th2) signalling pathways. Additionally, gene profiling assays demonstrated AD is associated with decreased gene expression of epidermal differentiation complex genes and elevated Th2 and Th17 genes. Hypomethylation of TSLP and FCER1G in AD were also reported; and miR-155, which targets the immune suppressor CTLA-4, was found to be significantly over-expressed in infiltrating T cells in AD skin lesions [2], [5].

2. Materials and Methods

2.1. Data Source

Data came from MAARS

Data Cleaning:

2.2. Exploratory Data Analysis

Exploratory data analysis was conducted to better understand the data at hand from a data-driven perspective. To this end, several clustering methods were used to help depict possible underlying patterns from the MAARS transcriptomic data: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and hierarchical clustering [8] [4]. In statistical machine learning, clustering techniques falls under unsupervised learning, which is a paradigm that attempts to perform inference on data without response variables or "labels" [4]. In the recent years, this area of machine learning has gathered much attention from researchers; data separability without the use of labels is known as disentangling in the autoencoder literature [7] [3]. Here, the sample labels are the lesional status of the skin sample and the primary purpose of performing such unsupervised analyses was to visually assess the separability of the samples using their transcriptomic information.

2.3. Statistical Analysis

2.3.1) Differential Gene Expression Analysis: The overarching goal of the statistical analysis was the determine the relevant genes from a transcriptomic datatset that are predictive of AD. The analysis plan was performed two-fold; the identification of genes highly associated with AD was first performed using differential gene expression analysis within a multiple hypothesis testing framework. The transcriptomic data was available as a continuous variable since the microarray data has been normalized.

Let $Y_{ij} \in \mathbb{R}$ denote the gene expression level of gene j in sample i and let $X_i \in \{0, 1\}$ represent its lesional status. The following linear model was fitted:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} X_i + \sum_{k:\, k \in \mathcal{P}} \beta_k \mathbf{1}_{pat(i)=k} + \epsilon_{ij} \\ &= \beta_{0j} + \beta_{1j} X_i + \beta_{pat(i)} \mathbf{1}_{pat(i)} + \epsilon_{ij} \end{aligned}$$

where the $\mathcal{P}$ is the set of all patients and $pat(i) \in \mathcal{P}$ is used to represent the patient to which belongs sample i. The coefficient β_k accounts for the intersample correlation for samples which are obtained from the same patient. By construction, $\mathbf{1}_{pat(i)=k} = 1$ for exactly one $k \in \mathcal{P}$ and $\left|\mathcal{P}\right| < n$.

Originally developped for two-color microarray experiments, it is now commonplace to employ an empirical Bayes method to analyze differential gene expression with linear models like the one above [11]. In a hypothesis framework, the following distributional assumptions for β_{1j} are assumed for all genes j:

$$\begin{aligned} H_0 : &\qquad \beta_{1j} \,\big|\, \sigma_j^2 \sim \mathcal{N}\left(0, \sigma_j^2\right) \\ H_1 : &\qquad \beta_{1j} \,\big|\, \sigma_j^2 \not\sim \mathcal{N}\left(0, \sigma_j^2\right) \end{aligned}$$

whereby p-values were ultimately obtained and adjusted using the Benjamini-Hochberg procedure to reduce the error in false discovery rate prone to be inflated due to performing multiple parallel statistical tests [1].

A prediction model was then implemented to assess predictability of sample lesional status. MENTION VARIABLE SELECTION. The receiving operating characteristic (ROC) curve is often used to evaluate data-driven binary prediction models and the area under this curve was used here to quantify the goodness of the prediction model [6].

3. Results

In

4. Discussion

5. Conclusion

References

[1] Yoav Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 72(4):405–416, 2010.

[2] Lianghua Bin and Donald YM Leung. Genetic and epigenetic studies of atopic dermatitis. *Allergy, Asthma & Clinical Immunology*, 12(1):52, 2016.

[3] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[5] Emma Guttman-Yassky, Andrea Waldman, Jusleen Ahluwalia, Peck Y Ong, and Lawrence F Eichenfield. Atopic dermatitis: pathogenesis. *Semin Cutan Med Surg*, 36(3):100–103, 2017.

[6] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

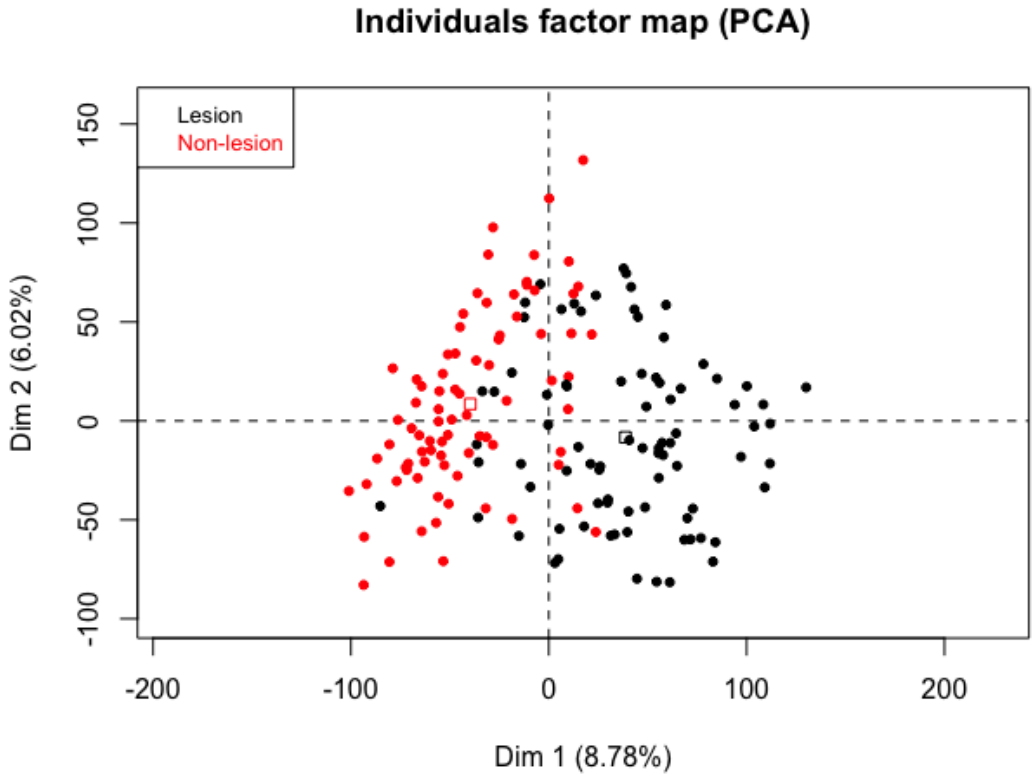
[7] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[9] Sophie Nutten. Atopic dermatitis: global epidemiology and risk factors. *Annals of Nutrition and Metabolism*, 66(Suppl. 1):8–16, 2015.

[10] W Peng and N Novak. Pathogenesis of atopic dermatitis. *Clinical & Experimental Allergy*, 45(3):566–574, 2015.

[11] Gordon Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 1.


Individuals factor map (PCA)
Lesion
Non-lesion
Dim 2 (6.02%)
150
100
50
0
-50
-100
-200
-100
0
100
200
Dim 1 (8.78%)