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 —
Methods —
Results —
Conclusion —
Keywords: Atopic dermatitis, transcriptomics, differential gene expression, unsupervised learning, prediction model.

1. Introduction

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 [10]. 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 [11].

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 such as NOD1, NOD2, TLR2, CD14, and DEFB1, all of which encode the integral factors in cutaneous immunologic response to non-specific antigens may also experience mutations and cause AD [6]. 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], [6].

This study attempts to examine the degree of dissimilarity between the transcriptomic expression between lesional and non-lesional samples in participants with atopic dermatitis. (ELABORATE MORE HERE...)

2. Materials and Methods

2.1. Data Source

The data at hand comes from the Microbes in Allergy and Autoimmunity Related to the Skin (MAARS). The purpose of this multidisciplinary research consortium is to better understand the characteristics surrounding two major chronic inflammatory skin diseases: atopic dermatitis and psoriasis [**?**]. In the MAARS study, these two illnesses that (AFFECT X PEOPLE IN THE WORLD) are investigated from different scientific perspectives and one of the research directions involved exploring the transcriptomic data of lesional and non-lesional samples in patients with atopic dermatitis, psoriasis and neither of the two studied skin diseases.

Data Management: Demographic data for all patients were available alongside their known allergies, concomitant medication and transcriptomic data for sequenced samples. Because it is only of interest to assess RNA expression levels in samples from patients with atopic dermatitis, two sets of data was discarded prior to conducting subsequent analyses: participants with AD who have no sequenced samples and genetic information for samples from healthy individuals and ones with psoriasis. The normalization procedure for transcriptomic data was already performed prior to obtaining the data for this anlysis. No genetic information was missing, so statistical methods for imputation was not needed for this study.

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, an MA diagram was plotted alongside several clustering methods which were used to help depict possible underlying patterns from the MAARS transcriptomic data: principal component analysis (PCA) and hierarchical clustering [9] [5]. 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" [5]. 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 [8] [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 [12]. In a hypothesis framework, the following distributional assumptions for β_{1j} are assumed for all genes j:

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

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 significance level of 0.05 was used in determining the genes highly associated with the outcome of interest, the lesional status of sequenced samples; a volcano plot was used to highlight potential significant genes.

A ensemble learning prediction model was then implemented to assess predictability of sample lesional status; this binary model was chosen due to its strengths in obtaining better predictions as it weights multiple models according to how well each perform on this specific task [5]. 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 (AUROC) was used here to quantify the goodness of the prediction model [7]. Internal validation was performed using leave-one-out cross-validation method whereby the AUROC was evaluated at every iteration [4].

3. Results

In Fig. 1, a simple 2D visualization of the results obtained from PCA is available, whereby lesional samples are displayed in black and non-lesional samples are displayed in red.

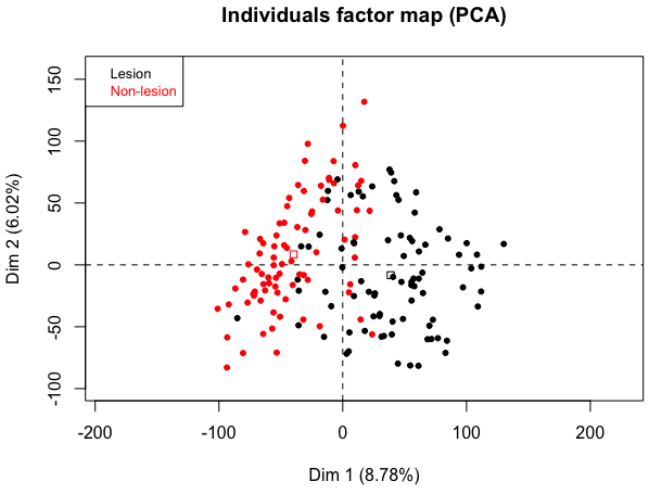


Fig. 1: testing

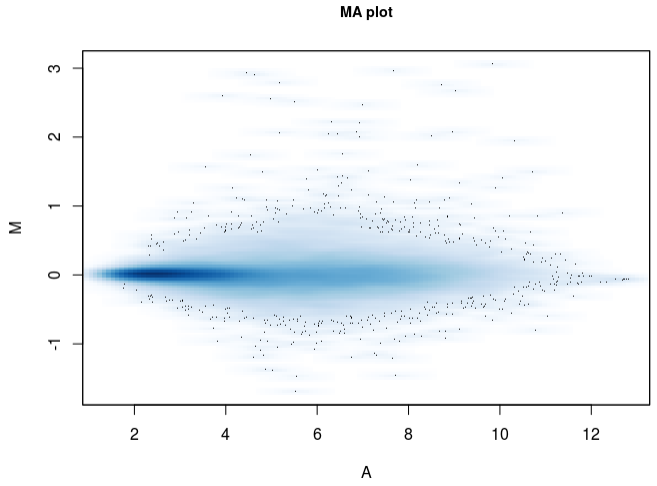
The horizontal axis represents the first component obtained through PCA and the vertical axis the second; they respectively explain 8.78% and 6.02% of the variance in the transcriptomic data.

A Differential Gene Analysis (DGE) was conducted using a linear model for paired data. The results indicated 81 genes that differed significantly between the lesioned and non-lesioned skin samples: 65 that are up-regulated and 16 that down-regulated. A heatmap illustrating the normalized gene expression levels is displayed below.

insert heatmap here

talk about clustering methods

An MA plot was used to visually represent the genes and determine which are up- and down-regulated. In the plot, each dot represents a gene. The x-axis is the average expression of the gene over the mean of normalized counts; the y-axis is the log2 fold change between the lesional statuses. The MA plot below indicates some genes that are up-regulated and fewer that are down-regulated.



Additionally, an sPLS analysis was conducted to identify genes that can "best explain" the variability in the lesional status of the sequenced samples. Information regarding the ten first components were retained whereby a different number of genes was selected for each component; the number of variables selected was determined using a 5-fold cross-validation method. This analysis resulted in the identification of 32 genes, 7 of which were found to be common in the two conducted analyses: DGE and sPLS. Genes KRT16, PRSS27, S100A9, S100A12, PI3, GJB2, ARK1B10 were found common in both analyses. Gene annotations using GSAn (Gene Set Annotation) CITATION, which includes a search on the Gene Ontology Database (GOA) as well, were performed for the seven genes that were identified as significant in both the DGE and the sPLS analyses. Using both the gene identifiers and their synonyms, GSAn retained 19 terms, covering 13 out of 16 genes and 6 of them are synthetic terms.

The figure below 2 provides information about the annotated genes within (GOA), the genes covered by GSAn, and the group-wise similarity between them. The first gauge shows excellent gene coverage of 100% within the GOA file. The second gauge shows that 81% of the genes were covered with the GSAn analysis. The third gauge shows that 44% of the genes share a term in both datasets.

Figure 3 displays the information content and the gene coverage of the synthetic terms. Functional annotation categories from lesional AD samples include serine endopeptidase activity, protein metabolic process, antimicrobial humoral response, cytoskeleton and toll-like receptor binding.

Figure 4 provides more detailed information about the representative terms. Further exploration of the tree will provide additional information on the terms sharing the same informative ancestor or the genes annotated by more than one term. Using this information, information can be obtained on the biological role of these genes in the occurrence of lesional AD.

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] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

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

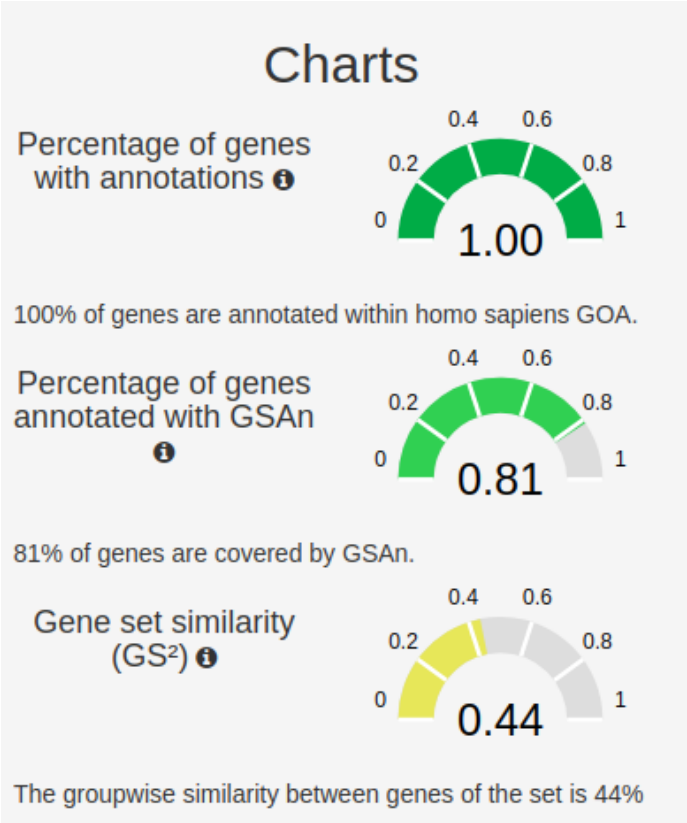


Fig. 2: Descriptive statistics of results obtained with GSAn on the 7 genes of interest.

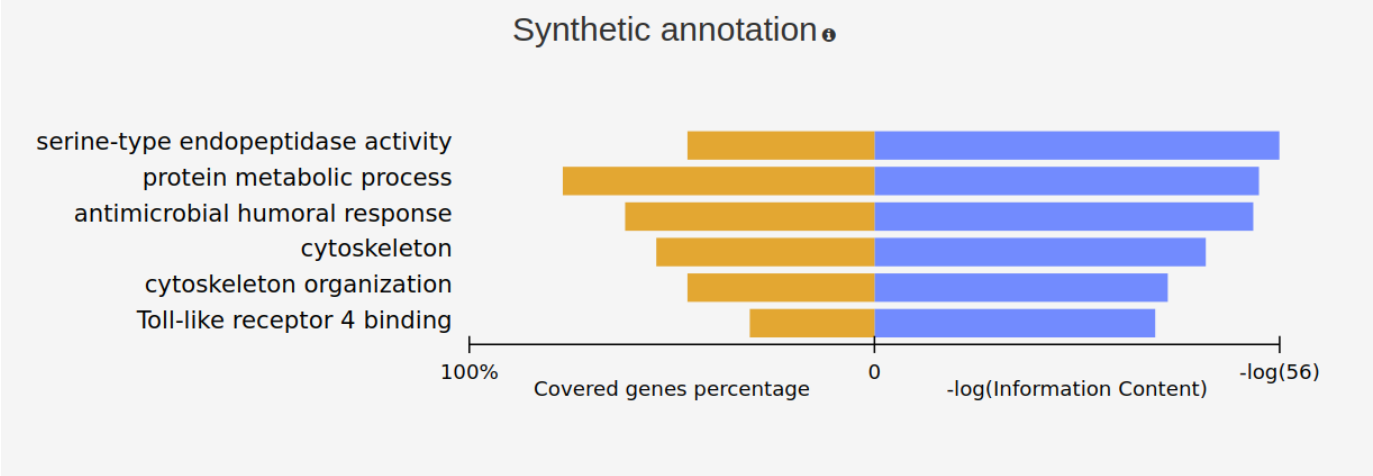


Fig. 3: Summary statistics of information and role of the genes of interest.

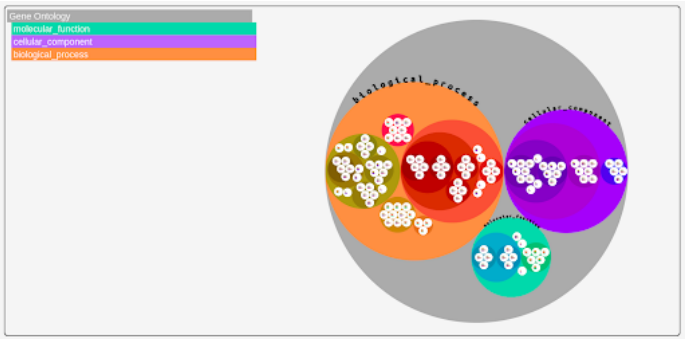


Fig. 4: Diagram showing the groupings of genetic information in the identified genes into three overarching categories: biological processes, cellular components and molecular functions.