Identifying Relevant Genes Related Atopic Dermatitis using Transcriptomic Data

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 — Atopic dermatitis (AD), a skin disease mainly characterized by the inflammation of the skin, affects $15-20\%$ children and $1-3\%$ adults worldwide. While some literature regarding its causes and its pathogenesis are available, further investigation regarding the transcriptomic expression of AD samples could lead to new findings.

Methods — Transcriptome data for lesional and non-lesional samples was provided by the MAARS study, a cohort study whose goal is to better understand skin diseases due to allergies and autoimmune disorders. Exploratory data analysis using principle component analysis (PCA) was first performed in the attempt to visualize the separability of the AD samples from the controls. Two data-driven methods in investigating the genes prone to be associated with AD were then used in this study: differential gene expression (DGE) analysis in a multiple testing framework and sparse partial least squares (sPLS).

Results — A total of 81 genes were found to be significantly differentially expressed between lesional and non-lesional skin samples using DGE, whereas 42 were obtained from sPLS. In these sets of identified genes, 7 were found to be common to both sets: KRT16, PRSS27, S100A9, S100A12, PI3, GJB2, ARK1B10.

Conclusion — The 7 identified genes were found to be linked to inflammatory processes, lipid metabolism and skin barrier pathways. However, potential confounding in the results due to race, patient age and allergies will further investigation.

Keywords: Atopic dermatitis, transcriptomics, gene annotation, differential gene expression, sparse partial least squares.

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 [1]. 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 [2].

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 [3]. 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 [3], [4].

This study attempts to examine the degree of dissimilarity between the transcriptomic expression between lesional and non-lesional samples in participants with atopic dermatitis. Data-driven methods are employed to identify relevant genes that are related to a flare-up in better understanding the molecular mechanisms.

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 [5]. In the MAARS study, these two illnesses are investigated from different scientific perspectives and one of the research directions involved exploring the transcriptome data of lesional and non-lesional samples in patients with atopic dermatitis, psoriasis (PSO) and neither of the two studied skin diseases.

Table 1. Baseline Characteristics of Patients with AD

Descriptor	Mean (SD) or frequency (%)
Number of patients	86
Sex	
▪ Male	48 (56%)
▪ Female	38 (44%)
Age	31 (8)
Race	
▪ **White**	77 (90%)
▪ **Black**	7 (8%)
▪ **Other**	2 (2%)
Number of sequenced samples	
▪ 1	8 (9%)
▪ 2	78 (91%)
Age (in years)	9.8 (1.5)
Known Allergies	
▪ Pseudo drug	2 (2%)
▪ Dust mite	15 (17%)
▪ Food	37 (43%)
▪ Pollen	47 (55%)
▪ Contact	14 (16%)
▪ Drug	7 (8%)
▪ Animal	33 (38%)
Concomitant medication	
▪ Anti-hypertensive	14 (16%)
▪ Anti-inflammatory	8 (9%)
▪ Other hormones	6 (7%)
▪ Thyroid hormones	4 (5%)
▪ Statins	7 (8%)
▪ Other medication	38 (44%)
Other concurrent diseases	
▪ Hyperlipidemia	5 (6%)
▪ Diabetes (type II)	3 (3%)
▪ Thyroid dysfunction	5 (6%)
▪ Asthma	14 (16%)
▪ Hypertension	16 (19%)
▪ Other	15 (17%)

Fig. 1: Summary of descriptive statistics in patient with AD

Data Management: Demographic data for all patients were available alongside their known allergies, concomitant medication and transcriptome data for sequenced samples. A total of 1317 patients with AD, PSO or neither illness were examined and 618 samples were sequenced. However, because it was only of interest to assess RNA expression levels in samples from patients with atopic dermatitis, two sets of data were discarded prior to conducting subsequent analyses: participants with AD who have no sequenced samples and genetic information for samples from healthy individuals and those with psoriasis. The normalization procedure for transcriptome data was already performed prior to obtaining the data for this anlysis. No genetic information was missing, so statistical methods for imputation were not needed for this study. Descriptive statistics are available in Table 1, displaying information of the 86 patients with AD[1].

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 transcriptome data: principal component analysis (PCA) and hierarchical clustering [6], [7]. 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" [7]. 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 transcriptome information.

2.3. ***Statistical Analysis***

The overarching goal of the statistical analysis was to determine the relevant genes from a transcriptome dataset that are associated to AD. The analysis plan was performed two-fold; the identification of genes highly associated with AD was first performed using differential gene expression analysis (DGE) within a multiple hypothesis testing framework. The transcriptome data was available as a continuous variable since the microarray data has been normalized. Secondly, an analysis using sparse partial least squares (sPLS) 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 and a different number of genes was selected for each constructed component; the number of variables selected was determined using a 10-fold cross-validation method.

In the DGE analysis, linear models were fitted for every single gene. 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. Because multiple samples can come from a same individual, confounding bias caused by intersample

[1]Note that two patients with AD don't have any samples sequenced. These patients have been discarded for further analysis since interest lies only in the transcriptome data

correlation could have occured if not taken into account. Let $\mathcal{P}$ denote the set of all patients and $pat(i) \in \mathcal{P}$ is used to represent the patient to which belongs sample i. The following linear model was fitted:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_i + \beta_{pat(i)} \mathbf{1}_{pat(i)} + \epsilon_{ij}$$

where the coefficient $\beta_{pat(i)}$ accounts for the intersample correlation for samples which are obtained from the same patient. Originally developped for two-color microarray experiments, it is now commonplace to use an empirical Bayes method to analyze differential gene expression with linear models like the one above [8]. In a hypothesis framework, the following distributional assumptions for β_{1j} are assumed for all genes j:

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

where the p-values were obtained using the rank of each gene's t-statistics as described by G. Smyth (2004) and they were 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 [8], [9]. 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; an MA plot was used to visually quantify the amount of potentially significant genes. As a final step, further investigation of the gene pathways was performed using Gene Set Annotation (GSAn 1.0.5).

All exploratory and inferential analyses were performed in R version 3.6.1 using limma version 3.40.6 and mixOmics version 6.8.5 packages respectively for efficiency reasons [**?**], [**?**].

3. Results

In Figure 2, 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.

The horizontal axis represents the first principle component (PC) obtained through PCA and the vertical axis the second; they respectively explain 8.78% and 6.02% of the variance in the transcriptome data. The Scree plot available in Figure 3 is another method of visualizing the amount of variation explained in each constructed PC. While the one of the primary purposes of this plot would help determine the moderate number of PC dimensions needed to explain the a majority of the variation, this plot can also illustrate the decreasing trend of explained variance across constructed PCs.

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.

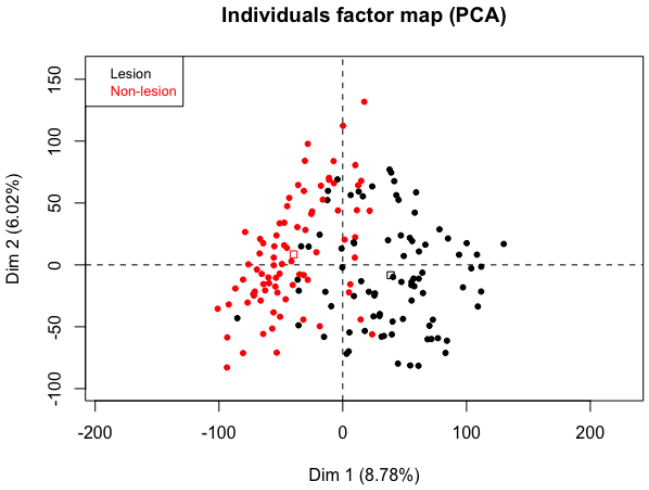

Fig. 2: 2D representation of samples using the two first principle components obtained by PCA.

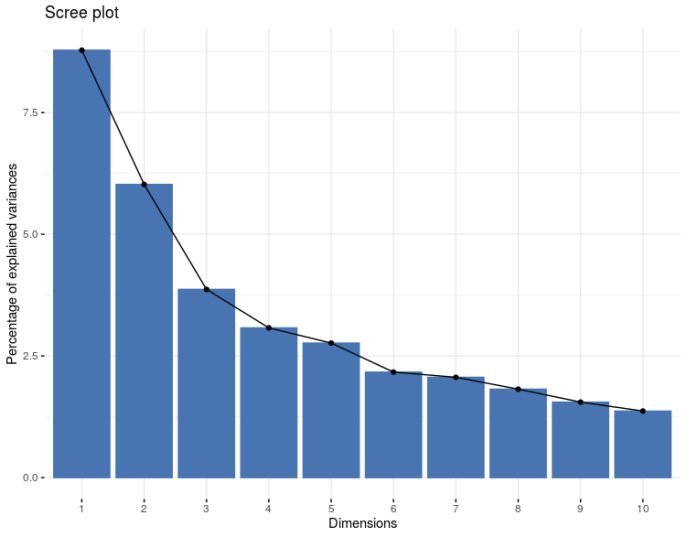

Fig. 3: Percentage of variation explained by the first 10 principle components obtained by PCA.

A MA plot disaplyed in Figure 4 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.

The analysis using sPLS resulted in the identification of 42 genes, 7 of which were found to be common in the two conducted analyses: DGE and sPLS. Genes KRT16,

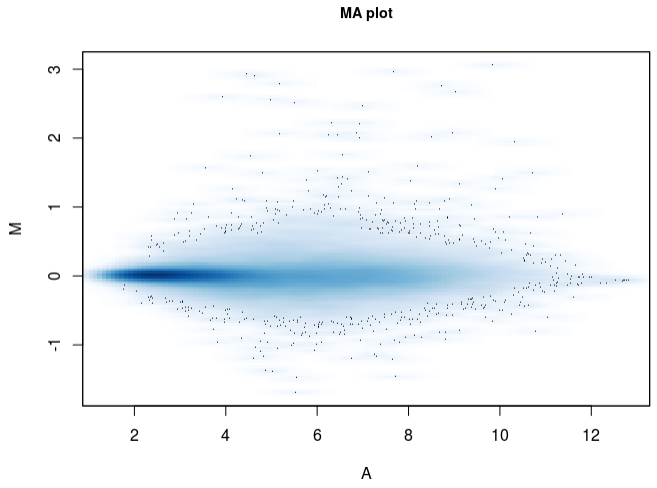

Fig. 4: MA plot for visualization of upregulated and downregulated genes.

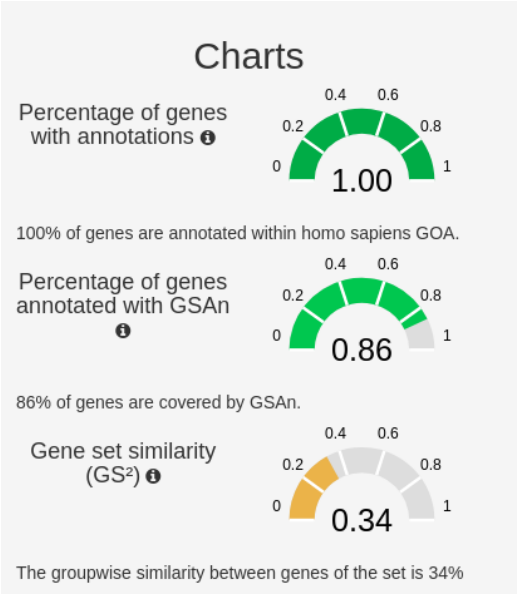

Fig. 5: Descriptive statistics of results obtained with GSAn on the 7 genes of interest.

PRSS27, S100A9, S100A12, PI3, GJB2, ARK1B10 were found common in both analyses. Gene annotations using GSAn (Gene Set Annotation), 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 7 terms, 3 of them are synthetic covering 6 out of 7 genes

Figure 5 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 86% of the genes were covered with the GSAn analysis. The third gauge shows that 34% of the genes share a term in both datasets.

Figure 6 provides more detailed information about the representative terms. Further exploration of the table will provide additional information on the terms annotating the genes. Using this information, information can be obtained on the biological role of these genes in the occurrence of lesional AD. From the annotation results we can say 5 out of 7 genes are annotated by response to external biotic stimuli (GO:0043207). 3 out of 7 genes are annotated by protein metabolic process (GO:0019538) and 2 out of 7 by secretory granule lumen (GO:0034774).

Figure 7 displays the information content and the gene coverage of the synthetic terms. Functional annotation categories from lesional AD samples include protein metabolic process, secretory granule lumen and response to external biotic stimuli.

4. Discussion

The current analysis considered all atopic dermatitis samples and compared lesional and non-lesional subtypes from the same patients. The current analysis was able to detect some degree of variation in transcriptomic expression between lesional and non-lesional AD skin samples: up to 25% of the variation in the transcriptome data can be explained by five principal components. Differential gene expression analysis helped understand the patterns of expressed genes in the diseased samples which can be used as biomarkers to differentiate them from the control samples. However, as displayed in Figure 2, PCA demonstrates that the genetic data at hand warrants further investagation as lesional and non-lesional samples visually appears to form two separate groups. Moreoever, in PCA, each eigenvector defines a principal component and their corresponding eigenvalue is proportional to its explained variance. The eigenvectors with large eigenvalues are the ones that exhibits the most information; the two first components explain 14.8% of the variance in the transcriptome data.

The set of seven genes that were identified in both the DGE and sPLS processes were used for gene annotation. The functional annotation of these seven genes shows their roles in inflammatory process (S100A9/A12), lipid metabolism (AKR1B10, AKR1B11) and skin barrier pathways (KRT16 , PRSS27, GJB2, PI3).

- KRT16 gene regulates innate immunity by producing signals in response to any epidermal barrier breach and mutations in this gene contributes to the initiation and exacerbation of AD [10], [11].
- Peptidase Inhibitor 3 (PI3) is an antimicrobial peptide and prevents elastase-mediated tissue proteolysis. It is increased during the onset of AD [12].

GSAn retained 7 terms, 3 of them being synthetic
6 out of 7 genes are covered

Show 10 entries Search:

GOID	Name	Ontology	IC	Covered genes	Synthetic
GO:0019538	protein metabolic process	BP	45.6	3	true
GO:0034774	secretory granule lumen	CC	27.06	2	true
GO:0043207	response to external biotic stimulus	BP	16.7	5	true
GO:0051092	positive regulation of NF-kappaB transcription factor activity	BP	283	2	false
GO:0043312	neutrophil degranulation	BP	144.93	2	false
GO:0030593	neutrophil chemotaxis	BP	130.04	2	false
GO:0035556	intracellular signal transduction	BP	23.37	2	false

Fig. 6: Table showing the groupings of genetic information in the identified genes into three overarching categories: biological processes, cellular components and molecular functions.

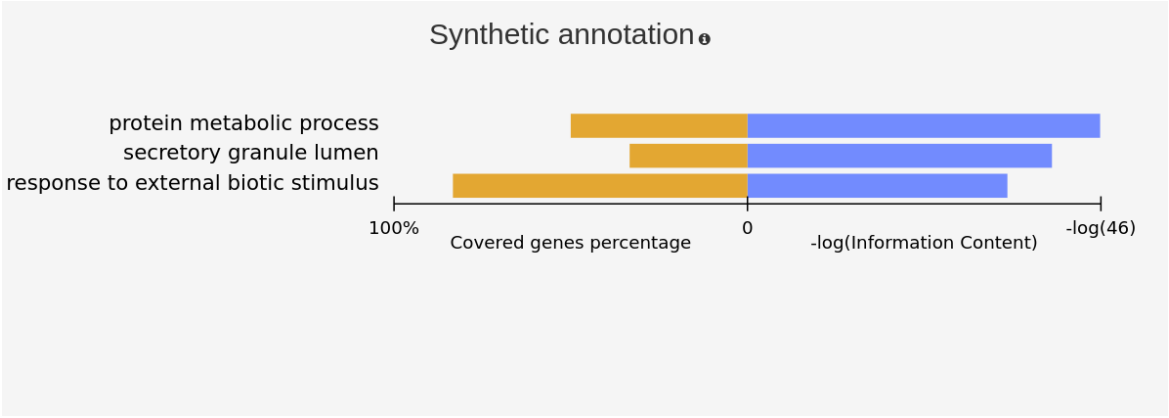

Fig. 7: Summary statistics of information and role of the genes of interest.

- Gap Junction Protein Beta 2(GJB2) expression increases the number of keratinocytes and causes thickness of epidermis which is observed in AD patients [13].
- Alto-keto reductase (AKR1B10 and AKR1B11) regulates keratinocyte differentiation. Alterations in AKR1B10 expression is seen in atopic dermatitis [14], [15].
- Expression of S100A9 causes abnormal differentiation and hyper proliferation of the epidermal cells; the upregulation of this gene can occur due to stress and certain drugs [16].
- S100A12 is known to exert antimicrobial activity and the skin barrier genes involved in AD enhances the expression of S100A12 by prevents it from exerting its activity [17].

*4.1. **Strengths and Limitations***

The current study successfully answered the research question; the data at hand suggests that there are genes expressed at significantly different levels between lesional and non-lesional skin samples obtained from individuals with atopic dermatitis. One strength of the study is that all of the genes identified in the analysis were found in previous literature to play an important role in pathogenesis of AD. Therefore, our results are consistent with previous findings.

A second strength is that the study used two types of analysis to determine the final list of genes. Additionally, the sample groups in this study were different from each other and showed no correlation. A high number of upregulated genes in DGE analysis were identified.

One limitation of the study is that the data provided does not include people of multiple races or age groups which may affect the findings of the study. A review of relevant literature shows that the FLG gene is a common cause of AD yet this gene did not appear in the results. Perhaps this gene is found in both lesional and non-lesional samples of individuals with AD and that is why it did not appear in the current analyses. More research is required to fully investigate.

5. **Conclusion**

The current study found some evidence to support that the genetic makeup of lesional and non-lesional skin samples are different between same-subject samples from individuals affected by AD. The genetic differences that were found may provide some insight into the functionality of healthy and lesioned tissue, mainly surrounding the

inflammatory process, skin barrier pathways and lipid metabolism.

However, some findings in the current study can still lead to research questions that may be of more interest. For example, future research may want to focus on the genetic differences of atopic dermatitis as related to race, age, and allergies.

- Race: The literature suggests that there is evidence that Black individuals may be more prone to eczema than individuals of other races, while at the same time less likely to be treated for the condition. The racial makeup (nearly 90% White) of the sample provided for analysis does not permit genetic comparisons among the races [18], [19].
- AD often changes over time; either improving or worsening as the individual ages. Therefore, it would be of interest to determine how the genetic structure changes over time in individuals and how those changes impact the progression of the condition.
- Allergies: A substantial portion of the sample provided has allergies to various substances, including pollen (48.2%), food (37.8%), and animals (33.6%). This information may be of importance since people with AD are more likely to develop hay fever and are advised to avoid allergens in an effort to prevent a flare-up of symptoms. Although not the topic of the current paper, investigating the differential gene expression of AD patients with and without various allergies may be of interest [20].

Further research using a more diverse sample of patients is needed to understand the genetic transgression from healthy to disease state in AD.

References

[1] S. Nutten, "Atopic dermatitis: global epidemiology and risk factors," *Annals of Nutrition and Metabolism*, vol. 66, no. Suppl. 1, pp. 8–16, 2015.

[2] W. Peng and N. Novak, "Pathogenesis of atopic dermatitis," *Clinical & Experimental Allergy*, vol. 45, no. 3, pp. 566–574, 2015.

[3] E. Guttman-Yassky, A. Waldman, J. Ahluwalia, P. Y. Ong, and L. F. Eichenfield, "Atopic dermatitis: pathogenesis," *Semin Cutan Med Surg*, vol. 36, no. 3, pp. 100–103, 2017.

[4] L. Bin and D. Y. Leung, "Genetic and epigenetic studies of atopic dermatitis," *Allergy, Asthma & Clinical Immunology*, vol. 12, no. 1, p. 52, 2016.

[5] "Microbes in allergy and autoimmunity related to the skin." https://www.maars.eu/. Accessed: 2019-11-03.

[6] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[7] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

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

[9] Y. Benjamini, "Discovering the false discovery rate," *Journal of the Royal Statistical Society: series B (statistical methodology)*, vol. 72, no. 4, pp. 405–416, 2010.

[10] U. Meyer-Hoffert, "Epidermal serine proteases and their inhibitors in atopic dermatitis," in *Atopic Dermatitis-Disease Etiology and Clinical Management*, IntechOpen, 2012.

[11] M. Hello, H. Aubert, C. Bernier, A. Néel, and S. Barbarot, "Atopic dermatitis of the adult," *La Revue de medecine interne*, vol. 37, no. 2, pp. 91–99, 2016.

[12] Y. Mansouri and E. Guttman-Yassky, "Immune pathways in atopic dermatitis, and definition of biomarkers through broad and targeted therapeutics," *Journal of clinical medicine*, vol. 4, no. 5, pp. 858–873, 2015.

[13] G. Suriyaphol, S. Theerawatanasirikul, and P. Chansiripornchai, "Association of gap junction beta 2 and transglutaminase 1 gene expression with canine atopic dermatitis," *The Thai Journal of Veterinary Medicine*, vol. 44, no. 3, pp. 279–285, 2014.

[14] D. Ghosh, L. Ding, U. Sivaprasad, E. Geh, J. B. Myers, J. A. Bernstein, G. K. K. Hershey, and T. B. Mersha, "Multiple transcriptome data analysis reveals biologically relevant atopic dermatitis signature genes and pathways," *PLoS One*, vol. 10, no. 12, p. e0144316, 2015.

[15] Y. Gao, X. Yi, and Y. Ding, "Combined transcriptomic analysis revealed akr1b10 played an important role in psoriasis through the dysregulated lipid pathway and overproliferation of keratinocyte," *BioMed research international*, vol. 2017, 2017.

[16] C. Kerkhoff, A. Voss, T. E. Scholzen, M. M. Averill, K. S. Zänker, and K. E. Bornfeldt, "Novel insights into the role of s 100 a 8/a 9 in skin biology," *Experimental dermatology*, vol. 21, no. 11, pp. 822–826, 2012.

[17] M. Mikus, C. Johansson, N. Acevedo, P. Nilsson, and A. Scheynius, "The antimicrobial protein s100a12 identified as a potential autoantigen in a subgroup of atopic dermatitis patients," *Clinical and translational allergy*, vol. 9, no. 1, p. 6, 2019.

[18] "Eczema in skin of color: What you need to know.." https://nationaleczema.org/eczema-in-skin-of-color/. Accessed: 2019-11-03.

[19] A. H. Fischer, D. B. Shin, D. J. Margolis, and J. Takeshita, "Racial and ethnic differences in health care utilization for childhood eczema: An analysis of the 2001-2013 medical expenditure panel surveys," *Journal of the American Academy of Dermatology*, vol. 77, no. 6, pp. 1060–1067, 2017.

[20] J. McIntosh, "What's to know about eczema?." https://www.medicalnewstoday.com/articles/14417.php. Accessed: 2019-11-03.