Analyzing Dermatitis/eczema from the UK Biobank GWAS

Abstract

Eczema is a group of skin diseases affecting roughly 20% of populuations without cure currently. To further the understanding of genetic and environmental causes of eczema, I utilized genome-wide association study summary statistics from the UK BioBank. From gene-set enrichment analysis and variant annotation, I show that variants in the 17 independent loci reaching genome-wide significance are enriched in genes related to immune responses and epidermal functions. In addition, while previous studies suggested environmental factors such as exposure to smoke and vitanim D deficiency can induce eczema, I observe no association between these two factors and the disease status.

Introduction

Eczema (Dermatitis) is a group of inflammatory skin diseases that cause redness, itchiness and rritation. It affects 15%-20% of populations (Paternoster et al. (2015)). Its pathogenesis includes skin barrier dysfunction and cell mediated immune response. Environmental factors such as vitamin D deficiency, humidity and exposure to smoke and detergent have been proposed as (Kantor et al. (2016), Mesquita (2013)). Multiple studies have performed genome wide association stuies (GWAS) on eczema, with variants in the *FLG* gene showing the strongest signals. Despite its prevalence there currently is no cure for eczema. Moreover, interpretation for genome-wide significant variants are usually not obvious. Recent effort on Biobank establishment has enabled large-scaled association studies with electronic health record. In this paper, I utilized recently developed computational methods that require only summary statistics for post-GWAS analyses to further the understanding of eczema. I performed gene-set enrichment analysis for genes whose imputed expression is strongly associated with the disease status. I also investigated environmental risk factors with Mendelian randomization.

Result

Manhattan plot with LocusZoom

Pre-harmonized GWAS summary statistics is publicly available from the MRC-IEU UK Biobank GWAS pipeline. The self-reported eczema GWAS consists of 337119 individuals with 8718 cases. I first visualized associations between self-reported eczema status and each genetic variant with a Manhattan plot (Fig. 1). The most significant variant rs61816761 (1:152313385_G/A, $-10\log_{10} p = 85.6$) is a loss-of-function mutation in the *FLG* gene, which encodes the protein profilaggrin that constitutes the epidermis, and has been frequenctly reported in eczema studies. Another variant rs3093553 (6:31581779_T/G $-10\log_{10} p = 13.09$) is a mutation in the *LBT* gene, which is linked to lymploid development and imflammotory response (NCBI Entrez gene entry).

Population structure

It is concerning whether deviations from the expected p-values at the upper tail are caused by population structure (Fig. 2). LD-score regression, which regresses χ^2 statistics of a variants against its ld-score, differentiates polygenicity from population stratification by testing if the resulting intercept differs from 1. With 1000 Genome Project European individual as a LD reference panel, the inferred intercept is 1.0102 (s.e. = 0.0067), suggesting polygenicity rather than population structure.

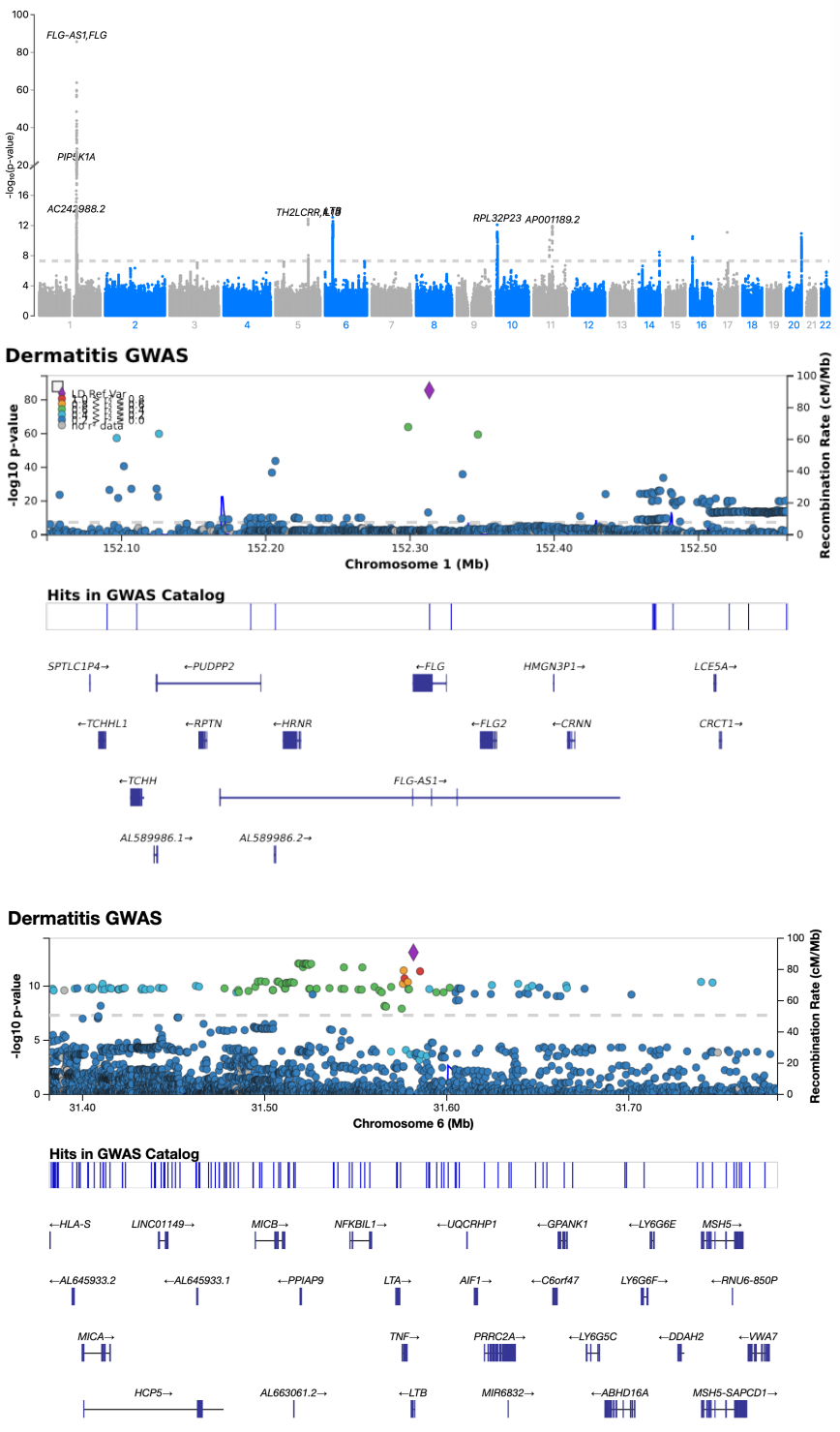

Figure 1: Manhattan plot (top), Locuz zoom plot of rs61816761 (bottom)

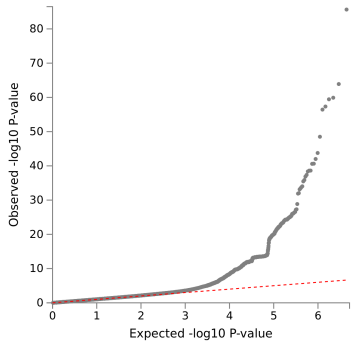

Figure 2: QQplot for Dermatitis

Functional annotation

Variants achieving genome-wide significance from GWAS usually suffer from poor interpretability. I performed functional annotation and visualization with a web application FUMA. There are four lead SNPs reported that had passed the p-value threshold (5×10^{-6}) and independent from each other (r^2 threshold 0.1) (Table 1). The significant variants are enriched in UTR3 regions (Fig. 3). At gene-levels, there are only two genes reaching genome-wide significance: MAP4K4 and FSTL4, and MAP4KA has been linked to inflammation.

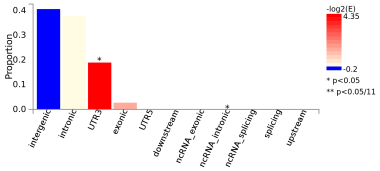

Figure 3: functional consequences of candidate SNPs

Table 1: lead SNPS from FUMA

unique ID	rsID	chromosome	position	p-value
2:102378254:A:G	rs6750020	2	102378254	5.0e-07
5:132713335:C:T	rs2897442	5	132713335	2.0e-07
6:408079:C:T	rs1050976	6	408079	1.3e-06
9:7783468:C:T	rs79418311	9	7783468	4.2e-06

S-Predixcan and gene set enrichment analysis

In addition, I used S-predixcan, which tests for association between imputed expression-level of genes and trait values requiring only summary statistics. Skin tissues and lymphocytes are especially of interest here since studies have suggested correlation between eczema and other immune responses. Inferred z-scores from the three tissue types are highly correlated (Fig. 4). The *TNXB* gene is statistically significant across these tissues ($Z > 6$), and it regulates the production of collagen. Multiple genes in the *LCE* complex are the most significant as well, and they regulate epidermal differentiation. In addition, I ran gene set enrichment analysis (GSEA) through the web application `WebGestalt`. Interleukin-17 production, response to epidermal growth factor, fluid transport are enriched among GO terms under biological processes with the skin tissue results (Fig. 5), while the lymphocyte results did not produce any significant terms.

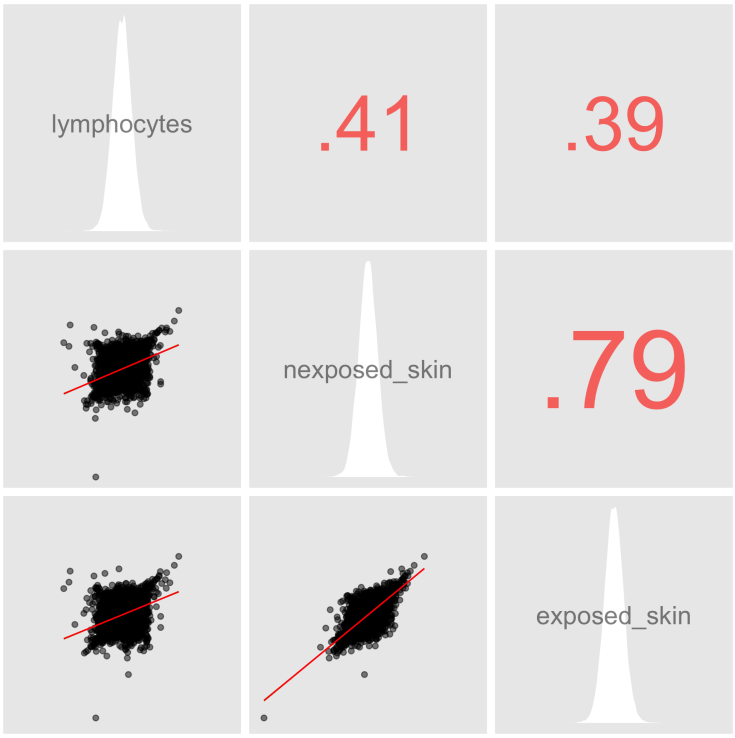

Figure 4: correlation of Z-scores from genes computed by S-prediXcan between each tissue type

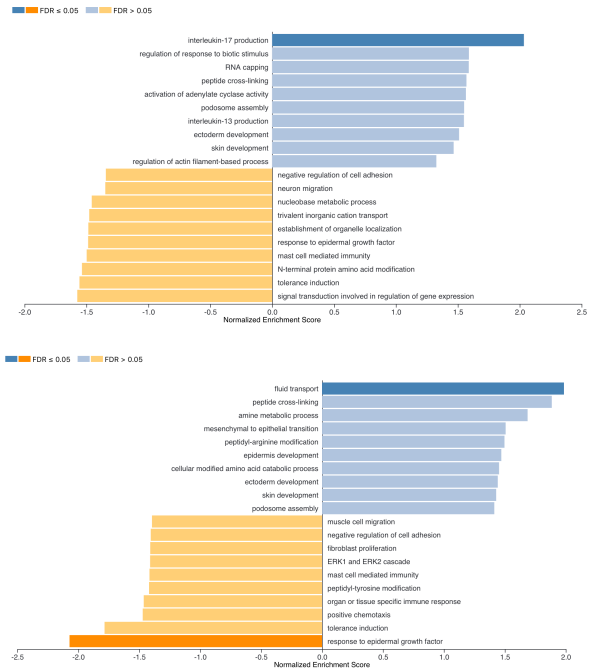

Figure 5: GSEA: skin sun exposed, GSEA: skin sun unexposed

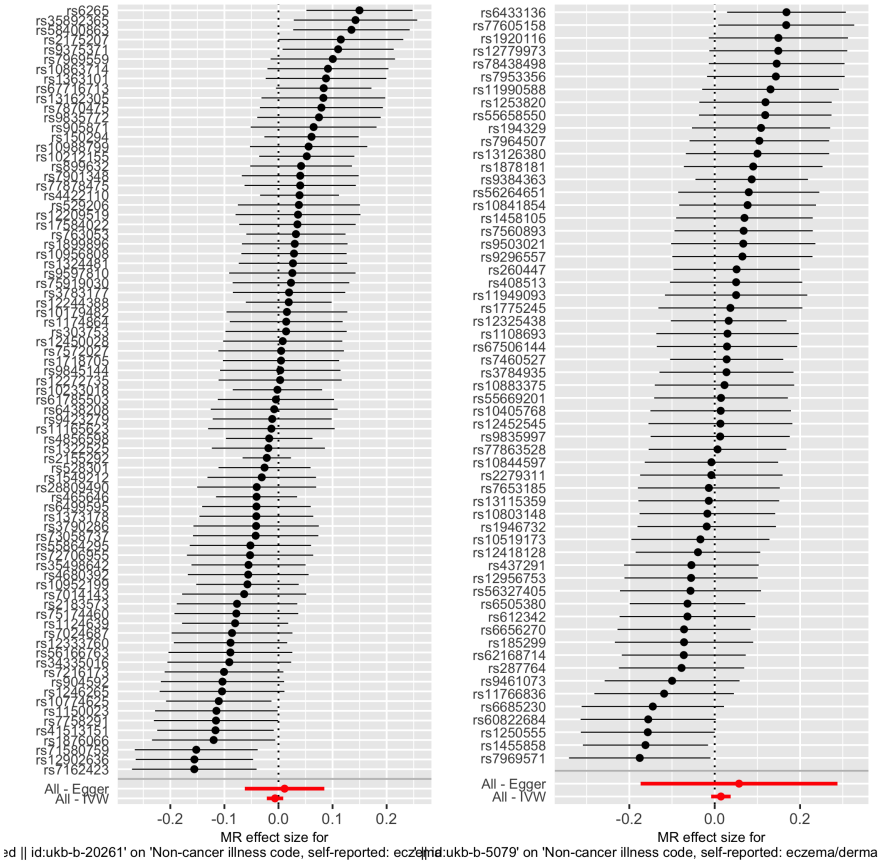

Figure 6: forrest plots with significant SNPs in ever smoked (left) Vitamin D (right) as instruments for mendelian randomization

Mendelian randomization

Eczema is a complex disease caused by both genetic and environmental components, and I performed mendelian randomization to test hypotheses on contribution from environmental variables. It has been proposed that vitamin D plays an important role in epidermal barrier dysfunction, so deficiency in Vitamin D maybe associated Eczema. In addition, studies have also suggested that exposure to smoke negatively affects immune systems and leads to higher risk for Eczema. Across all summarization schemes, the null hypotheses that smoking or Vitamin D levels not being associated with Eczema could not be rejected from the UK biobank summary statistics (Fig. 6).

Discussion

By intergrating computational tools for post-GWAS analyses, I showed that genes linked to immune responses or epidermal structures are associated with eczema, while environmental variables such as smoking and deficienty in vitamin D is not significantly correlated with the disease status. There are several limitations in the study above. First, self-reported dermatitis is a vague term that can lead to pooling atopic dermatitis, contact dermatitis, stasis dermatitis. Second, the UK Biobank suffers from bias due to the use of European cohort. There are observations that, hay fever, eczema, and asthma frequently exist in the same individulas. Hence, further steps such as colocalization can further elucidate genetic causes for eczema.

Data availability

The scripts used for the analysis, and raw outputs from predixcan are availble here:

Reference

Kantor, Robert, Ashley Kim, Jacob P. Thyssen, and Jonathan I. Silverberg. 2016. "Association of Atopic Dermatitis with Smoking: A Systematic Review and Meta-Analysis." *Journal of the American Academy of Dermatology* 75 (6). Elsevier: 1119–1125.e1. https://doi.org/10.1016/j.jaad.2016.07.017.

Mesquita, Ana Carolina de Souza Machado AND Costa, Kleyton de Carvalho AND Igreja. 2013. "Atopic dermatitis and vitamin D: facts and controversies." *Anais Brasileiros de Dermatologia* 88 (December). scielo: 945–53.

Paternoster, Lavinia, Marie Standl, Johannes Waage, Hansjörg Baurecht, Melanie Hotze, David P. Strachan, John A. Curtin, et al. 2015. "Multi-Ancestry Genome-Wide Association Study of 21,000 Cases and 95,000 Controls Identifies New Risk Loci for Atopic Dermatitis." *Nature Genetics* 47 (12): 1449–56. https://doi.org/10.1038/ng.3424.