Using a Machine Learning Approach to Identify Low-Frequency and Rare *FLG* Alleles Associated with Remission of Atopic Dermatitis

Ronald Berna[1], Nandita Mitra[2], Ole Hoffstad[1], Bradley Wubbenhorst[3], Katherine L. Nathanson[3] and David J. Margolis[1,2]

Atopic dermatitis (AD) is a common relapsing inflammatory skin disease. *FLG* is the gene most consistently associated with AD. Loss-of-function variants in *FLG* have been previously associated with AD. Low-frequency and rare alleles (minor allele frequency < 5%) in this gene have been given less attention than loss-of-function variants. We fine sequenced the *FLG* gene in a cohort of individuals with AD. We developed a machine learning–based algorithm to associate low-frequency and rare alleles with the disease. We then applied this algorithm to the *FLG* data, searching for associations between groups of low-frequency and rare *FLG* alleles and AD remission. A group of 46 rare and low-frequency *FLG* alleles was associated with increased AD remission ($P = 2.76e\text{-}11$). Overall, 16 of these 46 *FLG* variants were identified in an independent cohort and were associated with decreased AD incidence ($P = 0.0007$). This study presents an application of statistical methods in AD genetics and suggests that low-frequency and rare alleles may play a larger role in AD pathogenesis than previously appreciated.

JID Innovations (2021);**1:**100046 doi:10.1016/j.xjidi.2021.100046

INTRODUCTION

Atopic dermatitis (AD) is a common chronic inflammatory skin disease that typically presents with red itchy patches on the flexural parts of the limbs (Abramovits, 2005; Akdis et al., 2006; Bieber, 2008; Leung and Bieber, 2003). It is common, affecting up to 20% of children and 3.2−10.2% of adults in the industrialized world (Chiesa Fuxench et al., 2019; Pellerin et al., 2013).

Genetic studies of AD have suggested that both barrier dysfunction and immunodysregulation are key in the disease pathogenesis (Leung and Bieber, 2003). Genetic variation in the cytokines involved in the T helper type 2 response, such as *TSLP*, and in epidermal surface barrier proteins, primarily FLG, have been associated with variation in AD onset and persistence (Irvine et al., 2011; Kim et al., 2019; Palmer et al., 2006). Of these, the most consistently associated and widely studied is FLG. Loss-of-function (LoF) mutations in *FLG* lead to decreased production of FLG protein and are thought to result in a barrier defect that predisposes an individual to AD (Irvine et al., 2011; Margolis et al., 2012; Palmer et al., 2006; Quiroz et al., 2020).

The FLG protein belongs to a family of S100 fused-type proteins, many of which are part of the development of the skin's cornified envelope (Pellerin et al., 2013; Wu et al., 2011b). The genes are found in a section of chromosome 1 called the epidermal differentiation complex (Mischke et al., 1996; Pellerin et al., 2013; Wu et al., 2011b). Other skin barrier proteins in this family include FLG-2, TCHHL1, and hornerin (Margolis et al., 2014a; Pellerin et al., 2013; Pendaries et al., 2015; Wu et al., 2011b). In a previous study, we conducted next-generation sequencing of *TSLP* and *IL7R* and identified specific genetic variants likely to have a function in AD (Berna, 2021). However, this study was limited by the inability to effectively examine low-frequency (1% < minor allele frequency [MAF] < 5%) and rare (MAF < 1%) variants (Berna, 2021). Indeed, a key limitation of examining uncommon variants with next-generation sequencing studies is that one requires very large sample sizes to effectively examine such variants, an expensive and laborious endeavor.

This paper aims to use machine learning methods to identify low-frequency and rare alleles in *FLG*, in a diseased population, associated with AD remission. We developed a genetic algorithm–based approach to low-frequency and rare allele association, assessed general characteristics of this model using a simulated dataset, applied this model to low-frequency and rare alleles in *FLG*, and sought to identify the low-frequency and rare alleles associated with AD remission in a large longitudinal African American cohort.

[1]*Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA;* [2]*Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA; and* [3]*Division of Translational Medicine and Human Genetics, Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

Correspondence: Ronald Berna, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Boulevard, Philadelphia, Pennsylvania 19104, USA. E-mail: ronald.berna@pennmedicine.upenn.edu

Abbreviations: AD, atopic dermatitis; GAD, genetics of atopic dermatitis; GEE, generalized estimating equation; LoF, loss of function; MAF, minor allele frequency; PEER, Pediatric Eczema Elective Registry; PoC, probability of clearance

Received 1 February 2021; revised 28 June 2021; accepted 29 June 2021; accepted manuscript published online XXX; corrected proof published online XXX

Cite this article as: JID Innovations 2021;1:100046

RESULTS

Simulation results

Before applying our algorithm to Pediatric Eczema Elective Registry (PEER) genetic data, we conducted a number of simulation studies to determine the effectiveness of our algorithm in identifying low-frequency and rare alleles associated with longitudinal measures of disease. First, we constructed three longitudinal datasets: one in which the first 10 alleles were associated with more severe disease, one in which the first 10 alleles were associated with less severe disease, and one in which no alleles were differentially associated with disease. A total of 100 simulations of our algorithm were run on each dataset individually, with results showing clear identification of disease-associated alleles and good discrimination for alleles unassociated with disease (Figure 1a–c). The difference between these groups was statistically significant, with $P < 0.001$ for alleles associated with disease.

We next assessed the algorithm's sensitivity to initial conditions, evaluating its ability to identify the associated

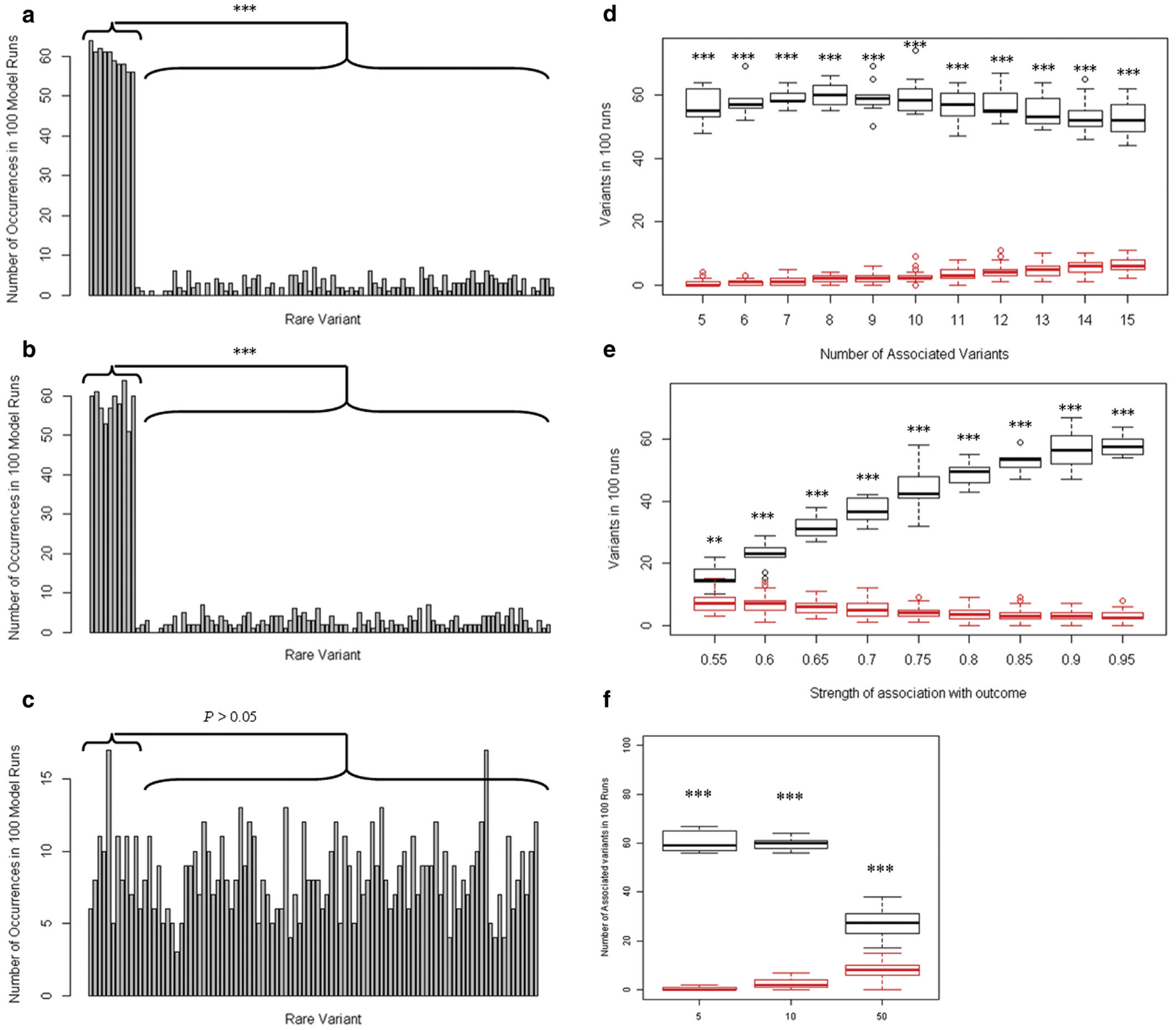


Figure 1. Simulation studies of genetic algorithm model (**a–c**). Three simulation studies showing the ability of the model to identify associated alleles. The first 10 alleles were associated with **(a)** moderate disease control and **(b)** more severe disease control. Associated alleles are identified ~60% of the time. By contrast, when no alleles are associated with disease control **(c)**, the algorithm picks up no clearly associated alleles. **(d)** Identification of alleles as the number of associated alleles increases. The black box plots represent the number of times the associated alleles were identified, the red box plots represent the number of times the unassociated alleles were identified. Each column represents the results of 100 independent runs of the model. **(e)** The proportion of identified alleles associated with disease, as the alleles' association with disease varied. Each column represents the results of 100 independent model runs. The black boxes correspond to the associated alleles, and the red boxes correspond to the unassociated alleles. **(f)** The proportion of alleles associated with disease, as population size increases. A total of 100 simulations were run for sample sizes of 50, 100, and 500, which corresponded to 5, 10, and 50 associated alleles, respectively. The black boxes represent the truly associated alleles, and the red boxes represent the unassociated alleles. ** $P < 0.01$, *** $P < 0.001$.

variants when the number of associated alleles increased or decreased (Figure 1d), when the associated alleles were more or less strongly associated with the outcome measure (Figure 1e), and when the size of the cohort increased or decreased (Figure 1f). These analyses showed the identification of approximately 60% of associated alleles on an average run (Figure 1d), with improving identification for more strongly associated alleles (Figure 1e). The algorithm also performs strongly when the population size is quite large and when the number of associated alleles increases (Figure 1f). All group differences in Figure 1d–f were statistically significant with $P < 0.01$ or $P < 0.001$.

PEER results

Within our PEER genetic cohort, 42% were male, the mean age of AD onset was 2.14 years, and the mean duration of observation was 97 months. Demographic information is provided in Table 1. A total of 60% of all surveys reported disease clearance; the proportion of surveys reporting clearance at any given time point is presented in Figure 2.

Massively parallel sequencing revealed 583 unique alleles in *FLG*, of which 337 had a MAF < 5%. After filtering *FLG* LoF alleles and alleles strongly correlated with *FLG* LoF, 322 *FLG* low-frequency and rare alleles remained (Figure 3). We performed 100 runs of the genetic algorithm on the *FLG* low-frequency and rare alleles. Evaluation of the alleles revealed a group of 46 low-frequency and rare alleles in *FLG* significantly associated with increased AD remission (OR = 5.19; 95% confidence interval = 3.52–7.66, adjusted *P* = 2.76e-11). The *P*-value was adjusted for 250,000 independent tests. Annotation of these variants is presented in Table 2. The frequencies in a control population, the African American subset of the Allele Frequency Aggregator, are provided in Table 3 (Phan et al., 2020). Genotype frequencies are provided in Table 4. A total of 40 of the identified alleles were within exon 3, *FLG*'s largest exon (Figure 4). This group of low-frequency and rare alleles is also associated with greater reports of disease clearance at almost every follow-up survey (Figure 5). These 46 low-frequency and rare *FLG* alleles are present in 55 different individuals, representing 16.9% of the PEER African American population. These 46 alleles are largely not in linkage disequilibrium with each other (Figure 6).

Table 1. Participant Demographics: Basic Demographic Data Regarding African Americans within PEER

Demographics	African Americans
Number	326
Age of AD onset, y, mean (SD)	2.14 (2.85)
Sex, male, n (%)	136 (41.98)
Asthma, n (%)	182 (56.17)
Seasonal allergies, n (%)	228 (70.37)
Observation time in months, mean (95% CI)	97 (101.2–92.9)
Disease control[1], mean (SD)	2.57 (0.74)

Abbreviations: AD, atopic dermatitis; CI, confidence interval; PEER, Pediatric Eczema Elective Registry.

[1] Disease control measured on a 4-point scale, with 1 representing complete disease control, and 4 representing uncontrolled disease.

Genetics of AD results

As a secondary study, we evaluated the contribution of these alleles to AD risk in a different population, the genetics of AD (GAD) group. We include the GAD data because we are seeking to show the significance of our low-frequency and rare alleles in an alternate cohort and thereby show that these alleles have significance beyond the PEER cohort.

Within the GAD genetic group, 316 individuals were African American. Massively parallel sequencing identified 1,094 unique alleles in *FLG*. Of the 46 low-frequency and rare alleles associated with less persistent AD in PEER, 16 were identified in the GAD African American group. Table 5 presents the MAFs of these alleles in GAD cases and controls. The OR of the association between these alleles and the presence of AD was 0.376, with a *P*-value of 0.0007.

DISCUSSION

In this study, we used a genetic algorithm–based approach to identify low-frequency and rare alleles associated with longitudinal measures of AD. We identified a group of low-frequency and rare alleles within *FLG* associated with AD. This *FLG* low-frequency and rare allele group represents a larger proportion of the PEER population than *FLG* LoF variants. This group is associated with lower odds of having AD in an independent dataset (the GAD group).

A key contribution of this report was the use of an algorithm to assess the association of low-frequency and rare alleles (MAF < 5%) with longitudinal measures of disease. Traditionally, it has been very difficult to identify low-frequency and rare alleles associated with common diseases because (i) these alleles occur infrequently and there is typically insufficient statistical power to generate accurate ORs, (ii) existing methods (burden tests and variant component tests) are primarily useful for examining whether a whole gene, and not particular variants, are implicated in a disease, and (iii) there are few methods for associating low-frequency and rare alleles with a disease when outcome measures are longitudinal (Lee et al., 2012; Wu et al., 2011a). Our genetic algorithm–based approach has several strengths. By grouping alleles, as is done in burden tests and variance component tests, we can increase the statistical power of any given generalized estimating equation (GEE) calculation. By iteratively refining the alleles in our group, we can better localize low-frequency and rare alleles that are likely to be disease-associated. By focusing on optimization of the OR of the association, we can associate low-frequency and rare alleles with a disease even when the outcome measure is nonbinary.

Numerous studies have examined the role of FLG in skin barrier formation. One recent report suggested that it is essential for the assembly of keratohyalin granules in the terminally differentiating epidermal layers (Quiroz et al., 2020). The association between *FLG* LoF variation and AD has been established in multiple cohorts and holds across ancestral groups (Barker et al., 2007; Marenholz et al., 2006; Margolis et al., 2014b; Margolis et al., 2018; Palmer et al., 2006; Pigors et al., 2018; Weidinger et al., 2007). However, few studies have associated low-frequency and rare *FLG* alleles with disease. None, to our knowledge, have identified non-LoF alleles associated with AD. Our group of 46

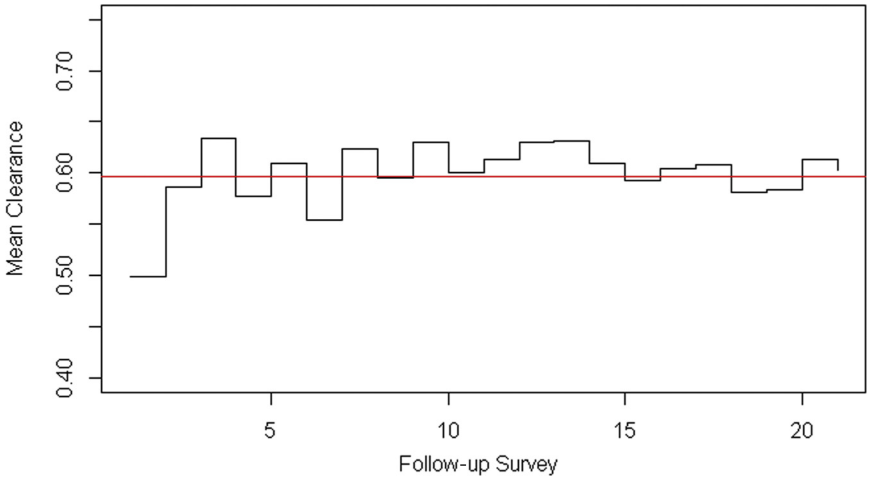


Figure 2. Proportion of subjects showing the outcome of disease clearance at each month of follow-up. The red line shows the proportion of surveys reporting disease clearance across the entire study period, 0.6.

low-frequency and rare alleles, associated with five-fold less severe AD, is present in 16.9% of the PEER population. By contrast, *FLG* LoF variants are present in only 11.3% of this study population. The association between our rare variant group and AD is visually apparent, with greater reports of disease clearance at almost every follow-up survey (Figure 5).

Six of the 46 variants identified represent synonymous amino acid changes. This is both interesting and unexpected. It is possible that these regions represent regulatory regions within coding exons (Dong et al., 2010). Further study is needed to identify the function of these alleles.

These *FLG* alleles are also protective against AD in an independent population, the GAD group. The fact that these variants are protective in the GAD group whereas they decrease persistence in the PEER cohort suggests that these variants mitigate AD presence and severity. The GAD results suggest that these variants have significance beyond the PEER population.

Although our analyses do not show causality between these *FLG* alleles and AD remission, these findings are valuable for two reasons. First, these low-frequency and rare alleles could be incorporated into genetic risk models predicting AD severity. Evaluation of these alleles in alternate populations could further verify these associations. Second, these alleles could be investigated in future studies of the molecular mechanisms of AD.

This study has several limitations. We only examined African American individuals, so our results may not apply to AD in different races/ethnicities. Because African Americans represent a distinct subset of those with African ancestry, our results may not generalize to all those of African ancestry. As with all observational studies, our analyses show that low-frequency and rare alleles are associated with AD severity. This association is not yet causal. Third, although associated with AD persistence in PEER, our allele groups may not generalize to the broader AD population. Studies in different populations will be needed to confirm these findings.

In this paper, we uncovered associations between low-frequency and rare alleles and AD remission using an approach employing a genetic algorithm to group these alleles. We identified a group of 46 low-frequency and rare alleles in *FLG* associated with decreased AD remission. These variants represent a contribution to AD genetics independent of *FLG* LoF insofar as all LoF variants and all alleles strongly correlated with LoF variants were removed before generating these allele groups. These alleles were present at a clinically significant frequency in this study population and accounted for a larger proportion of the African American PEER population than *FLG* LoF alleles alone. A subset of these alleles is associated with AD risk in an independent population, the GAD group. This study presents an application of statistical methods in AD genetics and uncovers genetic associations that may be valuable for future study of the mechanisms and epidemiology of AD.

MATERIALS AND METHODS

Rare allele model

Existing methods for associating low-frequency and rare alleles with a disease, such as burden tests and variance component tests, are intended to identify whether a gene is significant rather than whether any given low-frequency and rare variants are significant. However, machine learning methods enable the identification of clusters of the covariates associated with a specific outcome and have been shown to be useful in studying AD in the past (Berna et al., 2020; Paternoster et al., 2018; Thijs et al., 2017). We approached the identification of low-frequency and rare alleles associated with disease as a clustering problem. Low-frequency and rare alleles (defined as SNPs with a MAF $\leq$ 5%) were considered predictor variables, and a measure of disease outcome was considered the outcome variable. By iteratively drawing the subsets of our predictor variables and evaluating their association with disease outcome, we sought to identify groups or clusters of low-frequency and rare alleles associated with different disease outcomes.

Our outcome measure (the predictor variable) is described below. We used a genetic algorithm, implemented by the rbga.bin function from the genalg 0.2.0 package in R, to identify clusters of low-frequency and rare alleles (Chatterjee et al., 1996). The rgba.bin function was used to select alleles, which were then evaluated with GEE for association with AD remission/persistence. To iteratively improve the group of alleles selected, the rgba.bin function requires an outcome metric. This outcome metric was the GEE model's OR for the association. Each run of the genetic algorithm drew 2,500 subsets of low-frequency and rare alleles (divided into 50 iterations, each with a population size of 50), with replacement, from the set of

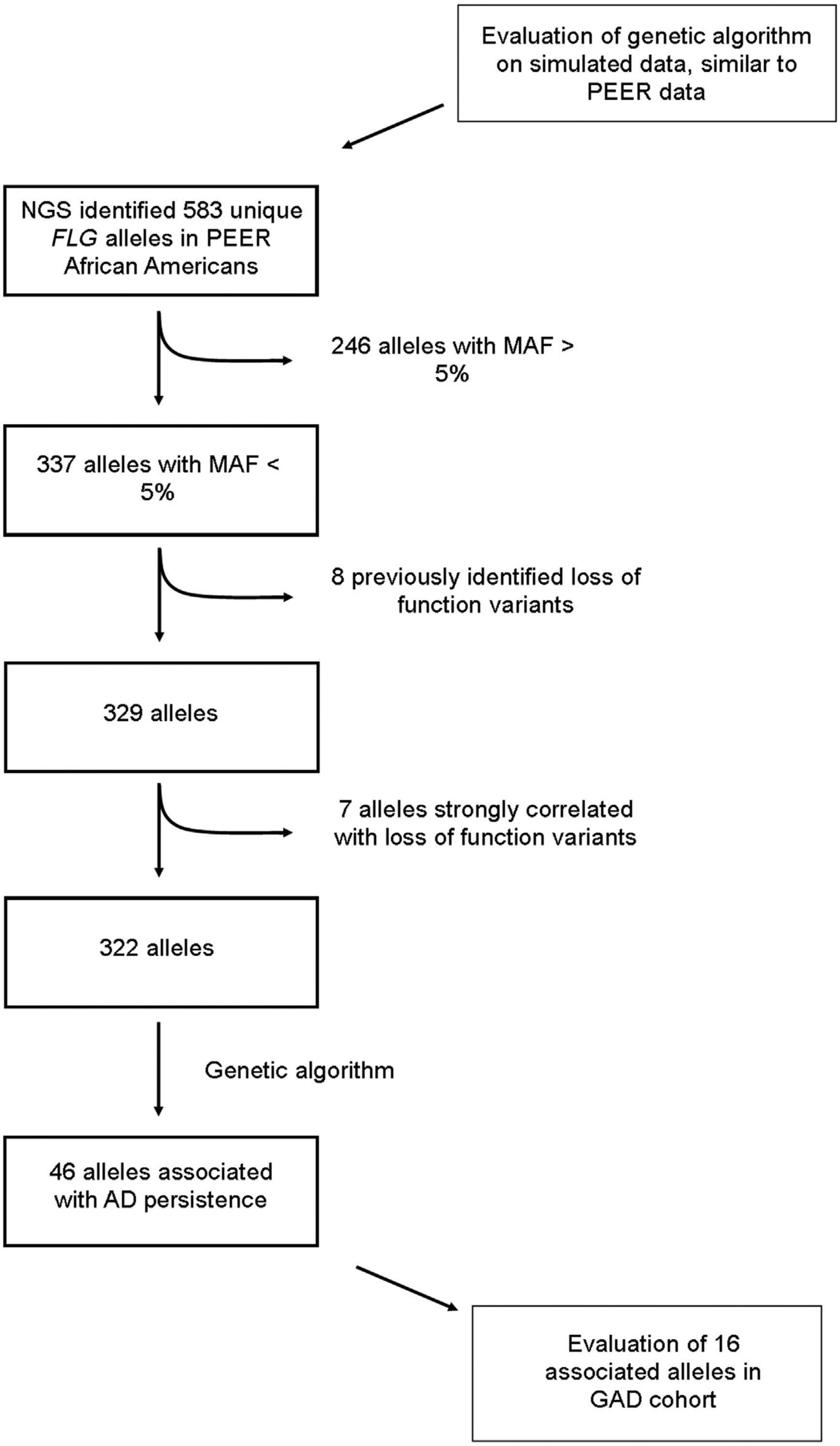

Evaluation of genetic algorithm on simulated data, similar to PEER data
NGS identified 583 unique *FLG* alleles in PEER African Americans
246 alleles with MAF > 5%
337 alleles with MAF < 5%
8 previously identified loss of function variants
329 alleles
7 alleles strongly correlated with loss of function variants
322 alleles
Genetic algorithm
46 alleles associated with AD persistence
Evaluation of 16 associated alleles in GAD cohort

Figure 3. Analysis plan for *FLG* alleles. AD, atopic dermatitis; GAD, genetics of atopic dermatitis; MAF, minor allele frequency; NGS, next-generation sequencing; PEER, Pediatric Eczema Elective Registry.

all low-frequency and rare alleles. Multiple runs of the algorithm were computed, and the results across runs were aggregated; alleles that occurred in ≥20 runs were included in the final rare allele groups.

Because the OR of the association was used to inform allele selection, this method represents a supervised clustering approach. *P*-values were computed from the GEE models. Because numerous independent estimates were computed to obtain the final rare allele groups, *P*-values were corrected for multiple comparisons according to a Bonferroni correction, accounting for the total number of estimates computed.

Simulation studies

To assess the performance of our genetic algorithm–based approach and to verify the model's ability to accurately identify the alleles associated with disease, we conducted sensitivity analyses. We evaluated the model's ability to identify (i) low-frequency and rare alleles associated with more severe disease, (ii) low-frequency and

Table 2. Annotation of the *FLG* Alleles Associated with Decreased AD Persistence in African Americans

Location	Nucleotide Change	RSID	Amino Acid Change	MAF	Functional Region
152275810	G>A	rs150496930	p.A3851V	0.002	Repeat domain 12
152276248	C>A	rs371128626	p.G3705V	0.005	Repeat domain 11
152276858	A>G	rs150047484	p.S3502P	0.002	Repeat domain 11
152276976	G>A	rs140294281	p.S3462S	0.002	Repeat domain 11
152276979	C>T	rs145466389	p.G3461G	0.003	Repeat domain 11
152277184	T>C	rs146234375	p.H3393R	0.002	Repeat domain 10
152277637	G>A	[1]	p.S3242F	0.002	Repeat domain 10
152277738	G>C	rs774463249	p.S3208R	0.002	Repeat domain 10
152277769	A>C	rs143183339	p.V3198G	0.002	Repeat domain 10
152277794	C>T	rs146288788	p.D3190N	0.002	Repeat domain 10
152277905	G>A	rs148315024	p.R3153C	0.002	Repeat domain 10
152278525	C>T	[1]	p.S2946N	0.002	Repeat domain 9
152278706	G>A	rs141172870	p.R2886C	0.002	Repeat domain 9
152278787	G>A	[1]	p.H2859Y	0.002	Repeat domain 9
152279561	C>T	rs146849256	p.D2601N	0.002	Repeat domain 8
152280131	A>G	rs966449727	p.S2411P	0.002	Repeat domain 7
152280134	G>A	rs141651911	p.R2410C	0.003	Repeat domain 7
152280330	T>A	rs368784083	p.S2344S	0.003	Repeat domain 7
152281133	G>A	rs151189270	p.R2077C	0.002	Repeat domain 6
152281389	C>T	rs138652718	p.A1991A	0.005	Repeat domain 6
152282238	G>A	rs151199504	p.D1708D	0.002	Repeat domain 5
152282384	T>A	rs200033409	p.T1660S	0.002	Repeat domain 5
152282684	G>A	rs151103850	p.R1560C	0.012	Repeat domain 5
152283239	G>A	rs772994159	p.H1375Y	0.002	Repeat domain 4
152283742	G>T	rs142660239	p.S1207Y	0.002	Repeat domain 4
152283962	T>C	rs140646945	p.T1134A	0.005	Repeat domain 4
152284289	C>G	rs145939718	p.A1025P	0.003	Repeat domain 3
152284376	C>T	rs149106390	p.G996R	0.002	Repeat domain 3
152284382	C>T	rs149595328	p.G994S	0.012	Repeat domain 3
152284540	C>T	rs547196696	p.R941H	0.003	Repeat domain 3
152285021	C>G	rs148739675	p.D781H	0.003	Repeat domain 2
152285119	G>C	rs201522026	p.T748S	0.003	Repeat domain 2
152285647	C>T	rs749798893	p.R572Q	0.002	Repeat domain 2
152285807	G>T	rs12036682	p.H519N	0.002	Repeat domain 2
152285981	G>A	rs184361545	p.R461W	0.002	Repeat domain 2
152286006	C>A	[1]	p.E452D	0.002	Repeat domain 1
152286029	G>A	rs141885805	p.L445L	0.002	Repeat domain 1
152286043	TG>T	[1]	p.H440fs	0.002	Repeat domain 1
152286061	G>T	rs144808372	p.T434K	0.002	Repeat domain 1
152286118	C>A	[1]	p.R415L	0.002	Repeat domain 1
152287645	G>A	rs991098106	—	0.002	Intron variant
152289544	C>T	rs1010015022	—	0.002	Intron variant
152291412	G>C	rs775513539	—	0.002	Intron variant
152293991	A>T	[1]	—	0.002	Intron variant
152296514	C>T	rs554188171	—	0.003	Intron variant
152296874	C>G	rs550028010	—	0.002	Intron variant

Abbreviations: AD, atopic dermatitis; CI, confidence interval; MAF, minor allele frequency; RSID, Reference SNP cluster ID.
The OR for the association of these alleles with AD persistence is 5.19 (95% CI = 3.52–7.66, adjusted *P* = 2.76e-11).
[1] RSID unavailable.

rare alleles associated with less severe disease, and (iii) when no low-frequency and rare alleles are disease-associated. A simulated dataset of 100 alleles, each with a MAF = 0.01, was constructed, for which binary outcome measures of disease clearance were recorded at 6-month intervals for 12 total years (the simulated data were intentionally reflective of the PEER data, in which binary outcome measures were recorded at 6-month intervals for ~12 years). Outcomes for each allele were binomially distributed, with a mean probability of a positive outcome varying between 0 (outcome never occurs) and 1 (outcome always occurs). For initial simulations, alleles were assigned outcomes according to a binary distribution with means of 0.95, 0.05, and 0.5 to represent a strong positive association with AD clearance, strong negative association with AD clearance, and no association with AD clearance,

Table 3. Frequencies of Low-Frequency and Rare *FLG* Alleles in PEER Versus Frequencies in African American Subset of ALFA

Location	Nucleotide Change	MAF in PEER	ALFA Allele Frequency, African Americans
152275810	G>A	0.002	0
152276248	C>A	0.005	0.0014
152276858	A>G	0.002	0.001
152276976	G>A	0.002	0.0026
152276979	C>T	0.003	0.0027
152277184	T>C	0.002	0
152277637	G>A	0.002	NA
152277738	G>C	0.002	0
152277769	A>C	0.002	0
152277794	C>T	0.002	0.0036
152277905	G>A	0.002	0.0009
152278525	C>T	0.002	NA
152278706	G>A	0.002	0
152278787	G>A	0.002	NA
152279561	C>T	0.002	0.0006
152280131	A>G	0.002	0
152280134	G>A	0.003	0.0047
152280330	T>A	0.003	0.0003
152281133	G>A	0.002	0.0003
152281389	C>T	0.005	0.0012
152282238	G>A	0.002	0
152282384	T>A	0.002	0
152282684	G>A	0.012	0.0119
152283239	G>A	0.002	0
152283742	G>T	0.002	0.0006
152283962	T>C	0.005	0.0047
152284289	C>G	0.003	0
152284376	C>T	0.002	0
152284382	C>T	0.012	0.0012
152284540	C>T	0.003	0
152285021	C>G	0.003	0.0053
152285119	G>C	0.003	0
152285647	C>T	0.002	0
152285807	G>T	0.002	0.0002
152285981	G>A	0.002	0
152286006	C>A	0.002	NA
152286029	G>A	0.002	0
152286043	TG>T	0.002	NA
152286061	G>T	0.002	0
152286118	C>A	0.002	NA
152287645	G>A	0.002	0
152289544	C>T	0.002	0
152291412	G>C	0.002	0
152293991	A>T	0.002	NA
152296514	C>T	0.003	0.0025
152296874	C>G	0.002	0.0004

Abbreviations: ALFA, Allele Frequency Aggregator; MAF, minor allele frequency; NA, not applicable; PEER, Pediatric Eczema Elective Registry.

Table 4. Genotype Frequencies of *FLG* Low-Frequency and Rare Alleles

		Genotype Frequencies		
Location	Nucleotide Change	AA	Aa	aa
152275810	G>A	0.997	0.003	0
152276248	C>A	0.991	0.009	0
152276858	A>G	0.997	0.003	0
152276976	G>A	0.988	0.012	0
152276979	C>T	0.994	0.006	0
152277184	T>C	0.997	0.003	0
152277637	G>A	0.997	0.003	0
152277738	G>C	0.997	0.003	0
152277769	A>C	0.994	0.006	0
152277794	C>T	0.997	0.003	0
152277905	G>A	0.997	0.003	0
152278525	C>T	0.997	0.003	0
152278706	G>A	0.991	0.009	0
152278787	G>A	0.997	0.003	0
152279561	C>T	0.991	0.009	0
152280131	A>G	0.997	0.003	0
152280134	G>A	0.994	0.006	0
152280330	T>A	0.994	0.006	0
152281133	G>A	0.994	0.006	0
152281389	C>T	0.991	0.009	0
152282238	G>A	0.997	0.003	0
152282384	T>A	0.997	0.003	0
152282684	G>A	0.914	0.086	0
152283239	G>A	0.997	0.003	0
152283742	G>T	0.972	0.028	0
152283962	T>C	0.991	0.009	0
152284289	C>G	0.979	0.021	0
152284376	C>T	0.988	0.009	0.003
152284382	C>T	0.975	0.025	0
152284540	C>T	0.994	0.006	0
152285021	C>G	0.994	0.006	0
152285119	G>C	0.994	0.006	0
152285647	C>T	0.997	0.003	0
152285807	G>T	0.988	0.009	0.003
152285981	G>A	0.997	0.003	0
152286006	C>A	0.997	0.003	0
152286029	G>A	0.997	0.003	0
152286043	TG>T	0.997	0.003	0
152286061	G>T	0.997	0.003	0
152286118	C>A	0.994	0.006	0
152287645	G>A	0.997	0.003	0
152289544	C>T	0.997	0.003	0
152291412	G>C	0.997	0.003	0
152293991	A>T	0.997	0.003	0
152296514	C>T	0.994	0.006	0
152296874	C>G	0.997	0.003	0

respectively. First, the model was run 100 times on a dataset where 10 alleles had a probability of clearance (PoC) of 0.95 (the remaining alleles were assigned a PoC of 0.5). Second, the model was run 100 times on a dataset where 10 alleles had a PoC of 0.05 (the remaining alleles were assigned a PoC of 0.5). Third, the model was run 100 times on a dataset in which all 100 alleles had a PoC = 0.5.

We next evaluated the algorithm's ability to identify a variable number of associated alleles. We applied the algorithm to a simulated dataset of 100 alleles, as mentioned earlier. In this simulation, X alleles were assigned a PoC of 0.95 (the remaining alleles were assigned a PoC = 0.5). We ran 100 simulations each for every value of X between 5 and 15.

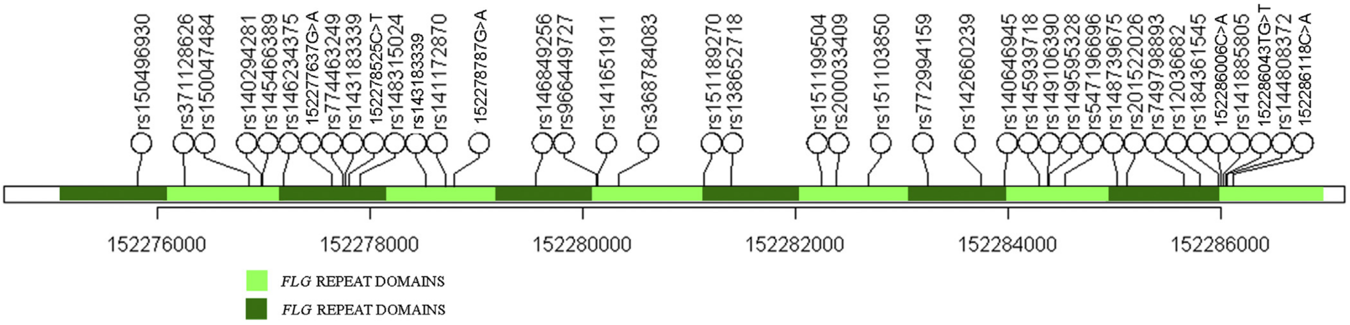


Figure 4. Plot of low-frequency and rare alleles along the *FLG* gene (for only those alleles lying in exon 3). Alternating light and dark green segments correspond to the 12 tandem repeats in *FLG*. Allele locations based on Margolis et al. (2019).

We also evaluated the model's ability to identify alleles as the strength of association varied. In this case, we assigned 10 of 100 alleles a PoC of Z (the remainder had a PoC = 0.5). We ran 10 simulations for each value of Z between 0.55 and 0.95, incrementing by 0.05 each time.

We finally evaluated the model's ability to identify low-frequency and rare alleles as the pool of low-frequency and rare alleles increased in size. Three simulations of 100 runs each were computed for population sizes of 50, 100, and 500. Truly associated alleles represented 10% of the total alleles.

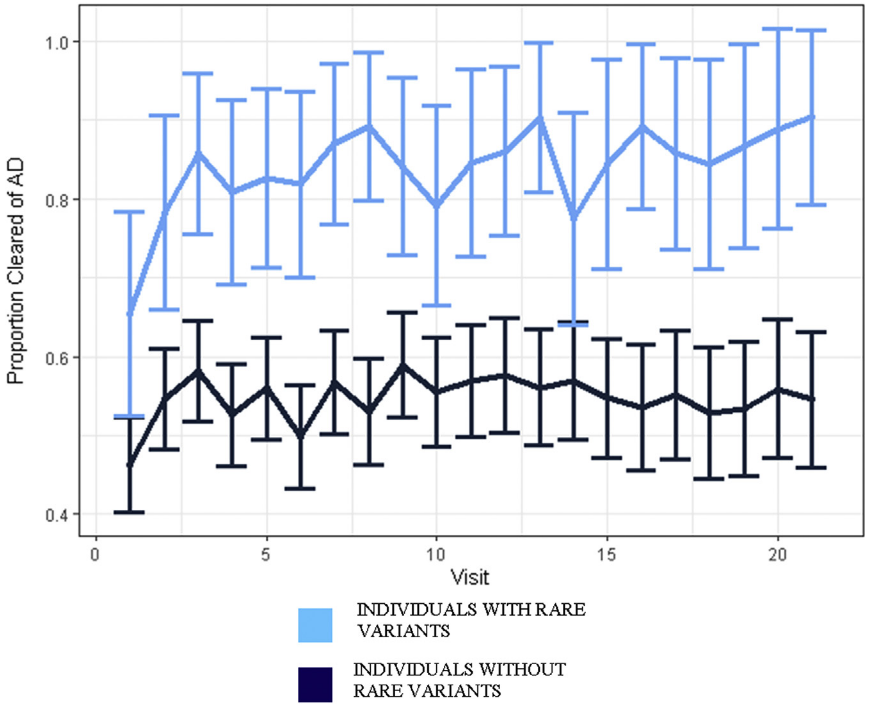
t-Tests were used to compare group means. *P*-values for these *t*-tests are presented in the respective figures.

Participants and genetic data

Genetic data were obtained from a subset of the PEER study for which DNA samples were available. The PEER study was approved by the Institutional Review Board of the University of Pennsylvania (Philadelphia, PA), and written informed consent was obtained from all participants or from their caregivers. Both the overall PEER cohort and the subset with genetic data have been previously described (Berna et al., 2021; Margolis et al., 2014c). This study examines only individuals self-described as African American. Self-described ancestry previously has been determined to strongly correlate with genetic markers of race within this cohort (Lou et al., 2019). We chose to strictly examine African American individuals because of the increased genetic variability observed in this population and the relative paucity of AD genetic studies of African Americans.

DNA was collected with Oragene DNA collection kits (DNA Genotek, Ottawa, Canada). Massively parallel sequencing genotyping of *FLG* was conducted on the 326 African American PEER individuals with sufficient DNA. Genotyping was by targeted capture using the Agilent SureSelect Platform (Agilent, Santa Clara, CA). Sequencing was performed on an Illumina Hiseq 4000 (Illumina, San Diego, CA). Raw sequencing data were aligned and mapped to the reference genome GRCh37 using the Burrows–Wheeler Aligner, version 0.7.17-r1188 (Li and Durbin, 2010). Single nucleotide variant and insertion and deletion calling were accomplished using the Genome Analysis Toolkit

Figure 5. Application of the model to *FLG* low-frequency and rare alleles. Average clearance over time for individuals with (blue) and without (black) low-frequency and rare alleles listed earlier. Error bars represent 95% confidence intervals. AD, atopic dermatitis.



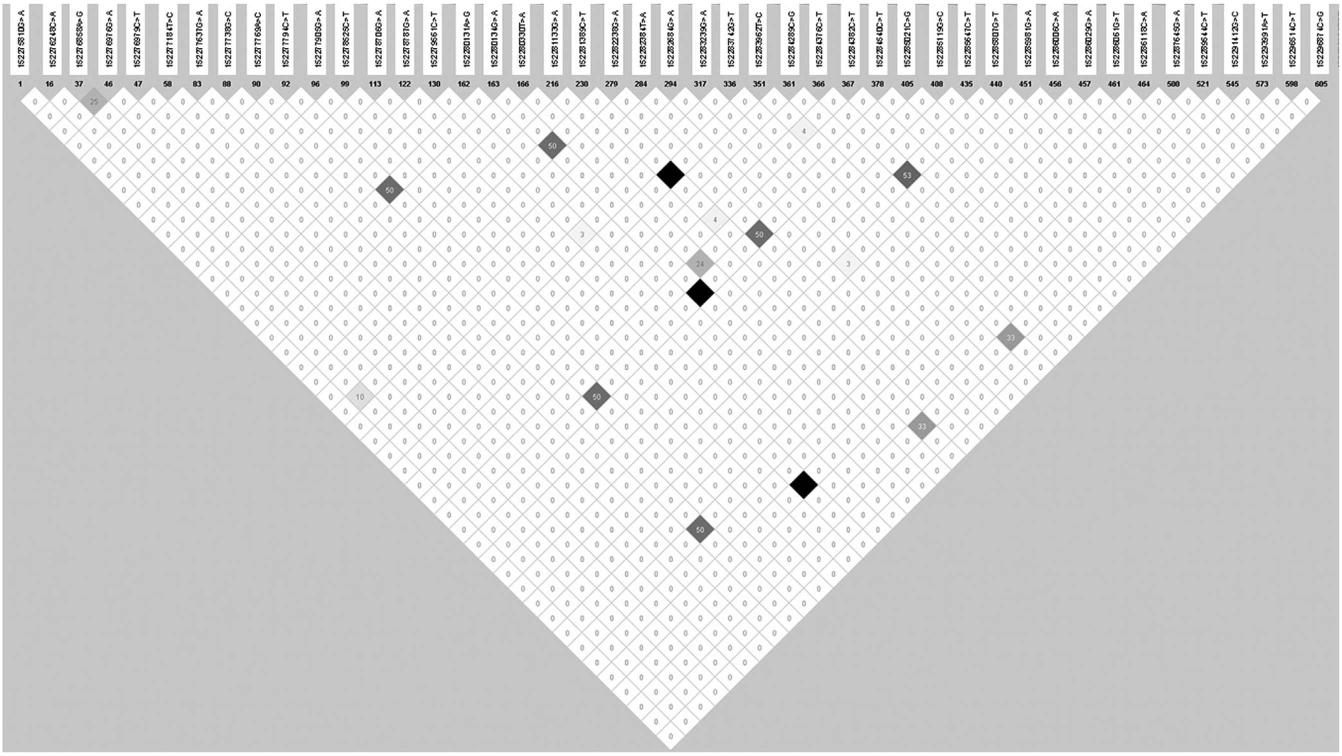


Figure 6. LD plot of the 46 low-frequency and rare *FLG* alleles associated with AD remission. 152286043TG>T not shown. Numbers overlying the shaded boxes represent the R^2 values. AD, atopic dermatitis; LD, linkage disequilibrium.

HaplotypeCaller, version 3.7 (Broad Institute, Boston, MA), after following Genome Analysis Toolkit best practices realignment and recalibration (DePristo et al., 2011; Poplin et al., 2018[1]; Van der Auwera et al., 2013).

Uncommon allele model to PEER low-frequency and rare alleles

Before applying the rare allele model to low-frequency and rare *FLG* alleles found in PEER, alleles were filtered as shown in Figure 3. First, we a priori chose to examine only alleles with a MAF < 5%. Then, we removed the previously identified LoF variants (as identified in Margolis et al. [2018]). Then, alleles strongly correlated ($R^2 > 0.1$) with LoF alleles were removed. The uncommon variant model was applied to the remaining alleles.

All alleles were read to a mean depth ≥30, with the majority read to a depth ≥100. After extensive discussion with our genetics partners, we concluded that this sequencing is more than adequate to accurately call even the rarest of these alleles. Sequencing depth for each of the 46 uncommon alleles of interest is presented in Table 6.

Outcome measure

Disease clearance was defined using a self-reported outcome of whether or not a child's skin was symptom free during the previous 6 months. Because children in PEER could be followed for >10 years, participants could have multiple reports of this outcome over time. The association between these outcomes and individual low-frequency and rare alleles was evaluated with GEEs for binary outcomes, assuming an exchangeable working correlation structure with empirical standard errors. This GEE model provided an estimate of the likelihood of AD improvement over time, which we interpreted as a measure of AD remission. An OR > 1 indicates that a risk factor increases the odds of remission; an OR < 1 indicates that a risk factor decreases the odds of disease remission. *P*-values reported are from these GEE estimates. GEE models were implemented through the geeglm function from the R package geepack 1.2-1.

Table 5. MAFs in GAD Cases and Controls

RSID	Location	MAF, GAD Controls	MAF, GAD Cases
rs371128626	152276248	0	0.005
rs140294281	152276976	0.025	0.010
rs145466389	152276979	0	0.005
rs146288788	152277794	0.017	0.005
rs146849256	152279561	0.008	0
rs141651911	152280134	0.025	0.005
rs151189270	152281133	0	0.005
rs138652718	152281389	0.058	0.015
rs200033409	152282384	0.050	0.050
rs151103850	152282684	0.008	0.020
rs142660239	152283742	0.008	0
rs145939718	152284289	0.008	0
rs149595328	152284382	0.008	0.020
rs148739675	152285021	0.033	0.005
rs201522026	152285119	0.108	0.010
rs12036682	152285807	0	0.005

Abbreviations: GAD, genetics of atopic dermatitis; MAF, minor allele frequency; RSID, Reference SNP cluster ID.

[1] Poplin R, Ruano-Rubio V, De Pristo MA, Fennell TJ, Carniero MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2018.

Table 6. Sequencing Depth for Each Allele in Our 46 Allele Composite

Location	Nucleotide Change	RSID	Mean Total Depth	Mean Alternate Allele Depth	Alternate Allele Fraction	Number of Individuals with Each Allele
152275810	G>A	rs150496930	463	218	0.471	1
152276248	C>A	rs371128626	339	158	0.466	3
152276858	A>G	rs150047484	285	137	0.481	1
152276976	G>A	rs140294281	395	190	0.481	4
152276979	C>T	rs145466389	530	276	0.521	2
152277184	T>C	rs146234375	179	89	0.497	1
152277637	G>A	[1]	197	84	0.426	1
152277738	G>C	rs774463249	95	53	0.558	1
152277769	A>C	rs143183339	398	213	0.535	2
152277794	C>T	rs146288788	754	371	0.492	1
152277905	G>A	rs148315024	113	56	0.496	1
152278525	C>T	[1]	60	33	0.55	1
152278706	G>A	rs141172870	590	175	0.297	3
152278787	G>A	[1]	126	112	0.889	1
152279561	C>T	rs146849256	357	183	0.513	3
152280131	A>G	rs966449727	141	64	0.454	1
152280134	G>A	rs141651911	353	184	0.521	2
152280330	T>A	rs368784083	243	122	0.502	2
152281133	G>A	rs151189270	548	156	0.285	1
152281389	C>T	rs138652718	679	182	0.268	3
152282238	G>A	rs151199504	137	75	0.547	1
152282384	T>A	rs200033409	224	86	0.384	1
152282684	G>A	rs151103850	241	140	0.581	28
152283239	G>A	rs772994159	384	189	0.492	1
152283742	G>T	rs142660239	440	214	0.486	9
152283962	T>C	rs140646945	343	166	0.484	3
152284289	C>G	rs145939718	230	108	0.47	1
152284376	C>T	rs149106390	469	288	0.614	4
152284382	C>T	rs149595328	422	183	0.434	7
152284540	C>T	rs547196696	286	121	0.423	2
152285021	C>G	rs148739675	450	215	0.478	2
152285119	G>C	rs201522026	214	103	0.481	2
152285647	C>T	rs749798893	473	225	0.476	1
152285807	G>T	rs12036682	388	220	0.567	5
152285981	G>A	rs184361545	503	263	0.523	1
152286006	C>A	[1]	157	74	0.471	1
152286029	G>A	rs141885805	404	209	0.517	1
152286043	TG>T	[1]	161	80	0.497	1
152286061	G>T	rs144808372	356	158	0.444	1
152286118	C>A	[1]	268	133	0.496	1
152287645	G>A	rs991098106	66	32	0.485	1
152289544	C>T	rs1010015022	260	116	0.446	1
152291412	G>C	rs775513539	194	98	0.505	1
152293991	A>T	[1]	125	53	0.424	1
152296514	C>T	rs554188171	184	99	0.538	2
152296874	C>G	rs550028010	348	186	0.534	1

Abbreviation: RSID, Reference SNP cluster ID.

[1] RSID unavailable.

Statistics and visualizations

Demographic characteristics are presented with means and SDs, as appropriate. Plots of clearance over time (with 95% confidence intervals) were constructed to show temporal differences in outcome measures for different groups. It is important to note that although these graphs provide an intuitive visualization of longitudinal differences in the disease course, the GEE models are not computing ORs on the basis of these curves.

Plots of low-frequency and rare alleles along a gene were created using the lolliplot function in R, utilizing the locations in the Single Nucleotide Polymorphism Database (National Center for Biotechnology Information, Bethesda, MD), using reference genome GRCh37 (https://www.ncbi.nlm.nih.gov/snp/).

The R^2 between the associated alleles identified in the analysis discussed earlier was calculated to show the correlation between alleles within this study population and to show that these alleles are largely not collinear. R^2 plots were generated in Haploview (Broad Institute, Boston, MA) (https://www.broadinstitute.org/haploview/haploview). Because these R^2 calculations represent the true R^2 values within our population of interest and are not intended to represent the R^2 within any broader population, they are not powered to any condition.

GAD cohort

As a secondary study, we evaluated the contribution of these alleles to AD risk in a different population. Specifically, we evaluated our low-frequency and rare alleles with AD in an independent cohort, the GAD group. Individuals in GAD were examined by dermatologists with expertise in the diagnosis of AD (from the University of Pennsylvania Perelman School of Medicine [Philadelphia, PA], Children's Hospital of Philadelphia [Philadelphia, PA], Pennsylvania State University/Hershey Medical Center [Philadelphia, PA], and Washington University School of Medicine in St Louis [St Louis, MO]). All subjects had a history and an examination consistent with AD (cases) or had no history of AD by history and examination (controls). For this study, we analyzed only individuals who were African American, by self-report. All subjects or legal guardians provided written informed consent or, if appropriate, assent approved by their appropriate Institutional Review Board.

Genotyping in the GAD cohort was carried out as previously described (Margolis et al., 2020). We then assessed for the association between our rare allele composite and the presence of AD using a chi-square test. *P*-values presented for all GAD cohort analyses represent *P*-values from chi-square tests. All analyses were implemented in R, version 3.6.1.

Data availability statement

The R code for the entire project was uploaded to FigShare and is publicly available.

The DOI for this data is https://doi.org/10.6084/m9.figshare.14569467.v1.

The Pediatric Eczema Elective Registry data (source) are not currently publicly available. The Pediatric Eczema Elective Registry study is an ongoing study sponsored by Valeant in response to a postmarketing commitment with the Food and Drug Administration. The GAD data (source) are not currently publicly available because this study is still enrolling.

A Browser Extensible Data file of the probes used for targeted capture is available on request from the corresponding author.

ORCIDs

Ronald Berna: http://orcid.org/0000-0003-0520-1218
Nandita Mitra: http://orcid.org/0000-0002-7714-3910
Ole Hoffstad: http://orcid.org/0000-0002-0261-903X
Bradley Wubbenhorst: http://orcid.org/0000-0001-8489-3659
Katherine L. Nathanson: http://orcid.org/0000-0002-6740-0901
David J. Margolis: http://orcid.org/0000-0002-0506-8085

AUTHOR CONTRIBUTIONS

Conceptualization: RB, DJM; Data Curation: RB, OH, BW; Formal Analysis: RB, DJM; Funding Acquisition: DJM; Investigation: RB, DJM; Methodology: RB, DJM, NM, KLN; Project Administration: RB, DJM, OH; Resources: DJM, KLN; Software: RB, OH, DJM, BW; Supervision: DJM, NM, KLN; Validation: RB, DJM, OH; Visualization: RB, DJM; Writing - Original Draft Preparation: RB, DJM; Writing - Review and Editing: RB, DJM, NM, KLN

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Institutes for Health (National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, MD) R01-AR060962 (principal investigator: DJM). The Pediatric Eczema Elective Registry study is funded as the Atopic Dermatitis Registry by Valeant Pharmaceuticals International (principal investigator: DJM).

CONFLICT OF INTEREST

DJM is a consultant for Pfizer, LEO, and Sanofi with respect to studies of atopic dermatitis and serves on an advisory board for the National Eczema Association. The remaining authors state no conflict of interest.

REFERENCES

Abramovits W. Atopic dermatitis. J Am Acad Dermatol 2005;53(Suppl. 1):S86–93.

Akdis CA, Akdis M, Bieber T, Bindslev-Jensen C, Boguniewicz M, Eigenmann P, et al. Diagnosis and treatment of atopic dermatitis in children and adults: European Academy of Allergology and Clinical Immunology/American Academy of Allergy, Asthma and Immunology/PRACTALL Consensus Report. Allergy 2006;61:969–87.

Barker JN, Palmer CN, Zhao Y, Liao H, Hull PR, Lee SP, et al. Null mutations in the filaggrin gene (FLG) determine major susceptibility to early-onset atopic dermatitis that persists into adulthood. J Invest Dermatol 2007;127:564–7.

Berna R, Mitra N, Hoffstad O, Wan J, Margolis DJ. Identifying phenotypes of atopic dermatitis in a longitudinal United States cohort using unbiased statistical clustering. J Invest Dermatol 2020;140:477–9.

Berna R, Mitra N, Lou C, Wan J, Hoffstad O, Wubbenhorst B, et al. TSLP and IL-7R variants are associated with persistent atopic dermatitis. J Invest Dermatol 2021;141:446–50.e2.

Bieber T. Atopic dermatitis. N Engl J Med 2008;358:1483–94.

Chatterjee S, Laudato M, Lynch LA. Genetic algorithms and their statistical applications: an introduction. Comp Stat Data Anal 1996;22:633–51.

Chiesa Fuxench ZC, Block JK, Boguniewicz M, Boyle J, Fonacier L, Gelfand JM, et al. Atopic dermatitis in America study: a cross-sectional study examining the prevalence and disease burden of atopic dermatitis in the US adult population. J Invest Dermatol 2019;139:583–90.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.

Dong X, Navratilova P, Fredman D, Drivenes Ø, Becker TS, Lenhard B. Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. Nucleic Acids Res 2010;38:1071–85.

Irvine AD, McLean WH, Leung DY. Filaggrin mutations associated with skin and allergic diseases. N Engl J Med 2011;365:1315–27.

Kim J, Kim BE, Leung DYM. Pathophysiology of atopic dermatitis: clinical implications. Allergy Asthma Proc 2019;40:84–92.

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762–75.

Leung DY, Bieber T. Atopic dermatitis. Lancet 2003;361:151–60.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95.

Lou C, Mitra N, Wubbenhorst B, D'Andrea K, Hoffstad O, Kim BS, et al. Association between fine mapping thymic stromal lymphopoietin and atopic dermatitis onset and persistence. Ann Allergy Asthma Immunol 2019;123:595–601.e1.

Marenholz I, Nickel R, Rüschendorf F, Schulz F, Esparza-Gordillo J, Kerscher T, et al. Filaggrin loss-of-function mutations predispose to phenotypes involved in the atopic march. J Allergy Clin Immunol 2006;118:866–71.

Margolis DJ, Apter AJ, Gupta J, Hoffstad O, Papadopoulos M, Campbell LE, et al. The persistence of atopic dermatitis and filaggrin (FLG) mutations in a US longitudinal cohort. J Allergy Clin Immunol 2012;130:912–7.

Margolis DJ, Gupta J, Apter AJ, Ganguly T, Hoffstad O, Papadopoulos M, et al. Filaggrin-2 variation is associated with more persistent atopic dermatitis in African American subjects. J Allergy Clin Immunol 2014a;133:784–9.

Margolis DJ, Gupta J, Apter AJ, Hoffstad O, Papadopoulos M, Rebbeck TR, et al. Exome sequencing of filaggrin and related genes in African-American children with atopic dermatitis. J Invest Dermatol 2014b;134:2272–4.

Margolis DJ, Mitra N, Berna R, Hoffstad O, Kim BS, Yan A, et al. Associating filaggrin copy number variation and atopic dermatitis in African-Americans: challenges and opportunities. J Dermatol Sci 2020;98:58–60.

Margolis DJ, Mitra N, Gochnauer H, Wubbenhorst B, D'Andrea K, Kraya A, et al. Uncommon filaggrin variants are associated with persistent atopic dermatitis in African Americans [published correction appears in J Invest Dermatol 2018;138:2084–5]. J Invest Dermatol 2018;138:1501–6.

Margolis DJ, Mitra N, Wubbenhorst B, D'Andrea K, Kraya AA, Hoffstad O, et al. Association of filaggrin loss-of-function variants with race in children with atopic dermatitis. JAMA Dermatol 2019;155:1269–76.

Margolis JS, Abuabara K, Bilker W, Hoffstad O, Margolis DJ. Persistence of mild to moderate atopic dermatitis. JAMA Dermatol 2014c;150:593–600.

Mischke D, Korge BP, Marenholz I, Volz A, Ziegler A. Genes encoding structural proteins of epidermal cornification and S100 calcium-binding proteins form a gene complex ("Epidermal Differentiation Complex") on human chromosome 1q21. J Invest Dermatol 1996;106:989–92.

Palmer CN, Irvine AD, Terron-Kwiatkowski A, Zhao Y, Liao H, Lee SP, et al. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. Nat Genet 2006;38:441–6.

Paternoster L, Savenije OEM, Heron J, Evans DM, Vonk JM, Brunekreef B, et al. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. J Allergy Clin Immunol 2018;141:964–71.

Pellerin L, Henry J, Hsu CY, Balica S, Jean-Decoster C, Méchin MC, et al. Defects of filaggrin-like proteins in both lesional and nonlesional atopic skin. J Allergy Clin Immunol 2013;131:1094–102.

Pendaries V, Le Lamer M, Cau L, Hansmann B, Malaisse J, Kezic S, et al. In a three-dimensional reconstructed human epidermis filaggrin-2 is essential for proper cornification. Cell Death Dis 2015;6:e1656.

Phan L, Jin Y, Zhang H, Qiang W, Shekhtman E, Shao D, et al., ALFA: allele frequency aggregator. National Center for Biotechnology Information, US National Library of Medicine. www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/. 2020 (accessed November 8, 2020).

Pigors M, Common JEA, Wong XFCC, Malik S, Scott CA, Tabarra N, et al. Exome sequencing and rare variant analysis reveals multiple filaggrin mutations in Bangladeshi families with atopic eczema and additional risk genes. J Invest Dermatol 2018;138:2674–7.

Quiroz FG, Fiore VF, Levorse J, Polak L, Wong E, Pasolli HA, et al. Liquid-liquid phase separation drives skin barrier formation. Science 2020;367:eaax9554.

Thijs JL, Strickland I, Bruijnzeel-Koomen CAFM, Nierkens S, Giovannone B, Csomor E, et al. Moving toward endotypes in atopic dermatitis: identification of patient clusters based on serum biomarker analysis [published correction appears in J Allergy Clin Immunol 2018;142:714. J Allergy Clin Immunol 2017;140:730–7.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11.10.1–33.

Weidinger S, Rodríguez E, Stahl C, Wagenpfeil S, Klopp N, Illig T, et al. Filaggrin mutations strongly predispose to early-onset and extrinsic atopic dermatitis. J Invest Dermatol 2007;127:724–6.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011a;89:82–93.

Wu Z, Latendorf T, Meyer-Hoffert U, Schröder JM. Identification of trichohyalin-like 1, an S100 fused-type protein selectively expressed in hair follicles. J Invest Dermatol 2011b;131:1761–3.