From the Institute of Epidemiology at Helmholtz Munich and the Institute for Medical Information Processing, Biometry, and Epidemiology

Dissertation
submitted for the degree of Doctor of Philosophy (Ph.D.) in
Medical Research - Epidemiology and Public Health
at the Medical Faculty
of the Ludwig-Maximilian University of Munich

Longitudinal assessment of the development of allergic diseases, aeroallergen sensitization and the role of DNA methylation

Submitted by:

Anna Kilanowski

From:

Lahnstein, Germany

2023

With approval of the Medical Faculty
of the Ludwig-Maximilian University of Munich

First evaluator (1.TAC member):	Prof. Dr. Annette Peters
Second evaluator (2. TAC member):	Prof. h.c. Dr. med. habil. Dr. h.c. Berthold Koletzko
Third evaluator:	Prof. Dr. med. Bianca Schaub
Fourth evaluator:	Dr. Burkhard Summer
Direct supervisor:	Dr. Marie Standl
Dean:	Prof. Dr. med. Thomas Gudermann
Date of oral defense:	19-Sep-2023

Contents

Abbreviations

BAMSE	Barn (=Children) Allergy Milieu Stockholm Epidemiology
BIC	Bayesian information criterion
CpG	Cytosine and Guanine only separated by their phosphate backbone
DACT	Divide-Aggregate Composite-Null Test
DNAm	DNA methylation
DOHaD	Developmental origins of health and disease
EWAS	Epigenome-wide association study
gHMA	Gene-based High-dimensional Mediation Analysis
GINIplus	German Infant Study on the Influence of Nutrition Intervention plus Air pollution and Genetics on Allergy Development
GWAS	Genome-wide association study
HIMA	High-dimensional mediation analysis
HRC	Haplotype Reference Consortium
IF	Impact factor
IgE	Immunoglobulin E
ISAAC	International Study of Asthma and Allergies in Childhood
LCA	Latent class analysis
LD	Linkage disequilibrium
LISA	Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany study
MAF	Minor allele frequency
MRS	Methylation Risk Score
PDGF-Rα	Platelet-derived growth factor receptor α
PRS	Polygenic Risk Score
RAST	Radio-Allergo-Sorbent-Test
SNP	Single Nucleotide Polymorphism
Th2	Type 2 T helper cells

Part I

Introduction

1 Background

Allergic diseases have seen an enormous increase in prevalence in the second half of the 20th century[1;2]. This increase usually coincides with societies' industrialization and modernization and has been declared an "epidemic" since[2;3].
It is estimated that up to 40% of the population is affected by allergic sensitization worldwide[4], 400 million suffer from allergic rhinitis alone[5], leading to a huge and non-negligible burden of disease in both social and healthcare costs and reduced quality of life[5;6]. Asthma further is the top chronic, non-communicable disease in childhood[7], causing over 2.8 million disability-adjusted life years in 2019[5].
The prevalence of the three common manifestations of allergic diseases, asthma, atopic dermatitis and allergic rhinitis, reach up to 23.3%, 13.7% and 20.6%, respectively, for allergic symptoms in 13- to 14-year-olds, as assessed in the third phase of the International Study of Asthma and Allergies in Childhood[8]. In a cross-sectional German study, physician-diagnosed allergic diseases affect 5.1%, 5.5% and 12.4% of 11- to 13-year-olds, respectively for asthma, atopic dermatitis and allergic rhinitis[9].
Recently, new research indicates that the increase in allergic disease prevalence might have plateaued on a high level[10], especially in westernized countries. However, further development has to be observed[11].
Potential solutions to ease the burden of allergies might tackle development of allergic diseases, primary prevention, but also progression of symptoms as secondary prevention. For once apparent allergic diseases there is currently no cure available and the only treatment options for patients include the medication of symptoms, but not the diseases itself, or the avoidance of allergic triggers.
Possible primary preventive measures, to reduce the risk for allergic disease occurrence, are summarized in the respective guidelines of the German society for allergology and clinical immunology and include for example full breastfeeding for 4 to 6 months or the importance of biodiversity in the current version from 2022[12]. However, with the present prevalence, these measures and their causal mechanisms have to be further investigated to facilitate the best evidence-based suggestions to tackle the allergy epidemic. In order to achieve this goal, it is of utmost importance to further characterize allergic diseases and comprehend their underlying biological mechanisms.

1.1 Allergic disease definitions

Allergic diseases are caused by an overreactions of the immune system in response to usually harmless substances. The most common biological mechanisms for allergic diseases are Type I reactions, whose standard mechanism is related to Immunoglobulin E (IgE) antibodies. Type I allergic diseases start with an immune response to certain antigens, in this case usually harmless allergens, which activate type 2 helper T cells (Th2) in some individuals. These Th2 cells initiate the production of IgE antibodies, which can, upon re-contact with the specific allergen, activate mast-cells and a release of histamines, leading to allergic symptoms, such as itching, sneezing or wheezing[6;13]. This IgE-mediated responses are further described under the term atopy.

Atopy is defined as "a personal or familial tendency to produce IgE antibodies in response to low doses of allergens, usually proteins, and to develop typical symptoms such as asthma, rhino-conjunctivitis, or eczema/dermatitis"[14].
According to the nomenclature for allergies, allergic diseases fall under the category of allergic hypersensitivity, which assumes an immunological involvement in disease development. These might be IgE-mediated reactions or not. The definition of IgE-mediated diseases is further divided into non-atopic and atopic reactions, which include allergic asthma, allergic rhinitis and atopic dermatitis[14].

Asthma is a chronic respiratory disease, which is characterized by chronic airway inflammation[15] and symptoms include coughing, wheezing, shortness of breath and airway obstruction[16]. Asthma might be IgE-mediated or not, and is characterized by episodes of wheezing often triggered by allergic reactants such as pollen but also potentially by air pollution or exercise. In children the mean prevalence is 11.6% for 6- to 7-year-old and slightly higher with 13.7% for 13- to 14-year-olds[17].
Atopic dermatitis has a life-time prevalence of up to 20% and is characterized by an itchy rash, which persists over at least 6 months and is typically located in the face of infants or around knees, elbows, hands and feet of older children and adolescents[18]. It usually starts in infancy but is now also recognized as lifelong disposition and attributed to a huge non-fatal health burden[19]. The original diagnostic criteria were published in 1980 and must contain at least itching, specified locations over the body and a chronic and relapsing disease history for a diagnosis[20].
Allergic rhinitis displays symptoms of a runny nose, including sneezing, mucous discharge and nasal obstruction[21] and current treatment guidelines are specified by the Allergic Rhinitis and its Impact on Asthma initiative[22]. Around 20% of all children are affected by allergic rhinitis[8] and about 80% of rhinitis patients develop the disease before the age of 20[23].
Differing patterns in onset, persistence and remission of asthma, dermatitis and rhinitis further impede treatment and prognosis as these allergic diseases do not develop at the same time points during childhood[24].

A joint mechanism of all atopic diseases is the activation of the immune system upon contact with allergens and thus the production of before mentioned IgE antibodies. Measuring these IgE antibodies in blood can yield an objectively measured indicator of allergic sensitization (=presence of specific antibodies), which is more common in children suffering from any of the described allergic diseases[25]. In the other direction, up to 53.7% of sensitized 8-year-olds had at least one other allergic disease[26].
This allergic multimorbidity has been previously reported as allergic diseases develop jointly "more often than expected by chance alone"[26] and within the investigated consortium 3.7% of 4 year olds and 4.4% of 8 year old had more than one allergic disease at the respective time. It has further been described that atopic dermatitis in infancy increases both the risk for allergic rhinitis[27] and asthma[28]. Even though the often postulated "atopic march", describing the progression from atopic dermatitis in infancy to subsequent allergic rhinitis and asthma, is being discussed[29;30], these associations may indicate common early-life determinants and shared genetic origins[31].

1.2 Early-life determinants

As allergic diseases often develop in infancy and childhood[24], it is portraying an especially important window of vulnerability. According to the developmental origins of health and

disease (DOHaD) hypothesis, during this time a child's health is greatly influenced by its environment[32]. Furthermore, the DOHaD hypothesis has been applied to the area of allergic diseases and several factors are discussed[33]. Previous research described perinatal and early-life determinants influencing allergic disease development, including, but not limited to, maternal smoking during pregnancy[34], pet ownership[35], socio-economic status[36] or breastfeeding[37].
More holistic approaches to explain the development of and the increase in allergic diseases include the hygiene and the biodiversity hypothesis.
The hygiene hypothesis builds on observations introduced by Strachan in 1989[38] and suggests that infections early in life due to contact with older siblings or other external parties might be protective of allergic diseases and that therefore allergy prevalence is rising as households tend to get smaller and more hygienic. This has been revisited and although household size does not explain all of the variability in allergic prevalence, a cleaner and more sterile environment in early life might still play an important role in the development of allergic diseases[39].
The biodiversity hypothesis, proposed by Haahtela in 2019, suggests that not necessarily early infections but in general a healthy exposure to the natural environment and microbial diversity promotes balance of the immune system and protects from allergic diseases[40]. Both hypotheses have in common that especially early windows of immune modulation are important for predicting and preventing allergic disease development and underline the importance to study risk and protective factors in early childhood.

1.3 Genetics of allergic diseases

With family history being the most important risk factor for allergic diseases[41], disentangling the genetic effect on allergic diseases from a shared environment has been in the interest of research over the past decades. Previous research approaches identified specific genes associated with allergic diseases, such as the Filagrin (*FLG*) gene for atopic dermatitis[42], using candidate-gene studies or estimated the general heritability of allergic diseases using twin studies[43].
In order to identify more genetic risk variants using hypothesis-free approaches and taking advantage of available arrays with hugely extended coverage, current endeavors have now turned to Genome-wide Association Studies (GWAS) in the general population. Since high-throughput technologies became available and affordable, sample sizes for conducted GWAS increased substantially over the past years. These GWAS aim to identify single-nucleotide polymorphisms (SNPs) associated with allergic diseases, which also share genetic signals mostly in the immune response process[31]. So far, several SNPs associated with asthma[44], atopic dermatitis[45;46;47] and rhinitis[48] but also any form of allergic disease[31;49] have been identified. Importantly, these genetic variations might help to further distinguish disease types, as, for example, several SNPs at the 17q12-21 locus have been linked with childhood-onset but not adult-onset asthma[50;44].
As theses SNPs usually have only small effect sizes and only describe parts of the genetic predisposition, their sole assessment might not capture the whole genetic disease contribution, both regarding statistical power and complexity of the diseases. Approaches aggregating the available evidence to Polygenic Risk Scores (PRS) have therefore been proposed for multiple complex diseases[51]. PRS are weighted scores, summing up the respective individual allele dosages weighted with the identified GWAS effect estimate. They therefore aggregate the available evidence over all identified genetic risk variants for single study participants and enable a risk estimation in comparison to the general population. For allergic diseases, these PRS might enhance predictive performance and enable the identification of at-risk populations.

1.4 Epigenetics of allergic diseases

Not only genetics alone are influencing disease development but also the activity of respective genes. Epigenetics describe multiple molecular mechanisms, such as histone modifications, which relate to changes in gene activity. The most widely studied mechanism is DNA methylation (DNAm), describing the addition of a methyl- (CH3) group to a cytosine locus in the genome, forming a so called CpG locus, together with the neighboring guanine and their collective phosphate back[52]. While these additions do not change the genome, they are influencing gene expression, for example by inhibition of transcription factor binding to the DNA, potentially silencing the following gene[53]. Thus, DNAm leads to a variety of different cell types as not all genes are expressed simultaneously[53] but may further be a potential mechanism through which the environment affects gene expression and allergic disease development, which is investigated in the present dissertation.
It has previously been described that environmental factors may change the regulation of gene expression[52] through DNAm, for example prenatal smoking[54;34] or air pollution[55;56]. Especially exposures in early time windows might lead to life-long diseases and DNAm might represent a potential and reversible mechanism for these developmental origins[57;58].
Epigenome-wide association studies (EWAS), though they do not yet reach sample sizes of recent GWAS, have already identified CpGs associated with allergic phenotypes, such as aeroallergen sensitization[59], childhood asthma[60] and any allergy[61] in collaborative efforts.
Similar to limitations of GWAS, analyses of DNAm are often limited by small effect sizes and the complexity of diseases, being affected by more than one gene. Methylation Risk Scores (MRS), which follow a similar approach as described PRS and provide aggregated scores on available evidence for single participants, might address this issue[62].
As DNAm might be modified by environmental factors, it can further be used to investigate causal or predictive patterns. It is currently unknown, whether changes in DNAm patterns lead to disease manifestation or if in reverse DNAm is changed by the disease itself. Both effect direction are supported by previous literature, as Davidson et al. reported that environmental factors can induce changes in DNAm and thus alter gene expression and phenotypes[63;64] but further Min et al. observed changes in DNAm following disease onset[65]. Addressing the question of causality and effect modification based on genetic and environmental exposures might provide a better understanding of molecular mechanisms in the development of allergic diseases.

2 Aims and outline

Based on the need to better understand the development of allergic diseases, their early-life determinants and molecular mechanisms, the present dissertation aims to investigate the longitudinal development of asthma, atopic dermatitis and rhinitis, as well as the contribution and effect directions between allergic sensitization and DNAm.
The cumulative work includes three publications tackling

1. The joint longitudinal development of allergic diseases and their early-life and genetic determinants, employing clustering approaches and PRS,
2. The establishment of MRS as valid biomarkers of aeroallergen sensitization, underlining previous evidence from EWAS and
3. The question of causality between aeroallergen sensitization and DNAm to further elucidate molecular mechanisms and to identify potential biomarkers and predictors of allergic diseases.

Publication 1: Kilanowski, A, Thiering, E, Wang, G, et al. Allergic disease trajectories up to adolescence: Characteristics, early-life, and genetic determinants. Allergy. 2022; 00: 1-15. doi: 10.1111/all.15511 [Impact factor (IF) = 14.710]

In the first publication, we investigate the joint development of allergic diseases from birth to adolescence in the German GINIplus and LISA birth cohorts. Unsupervised k-means clustering was applied to derive allergic disease trajectories based on parent-reported doctor's diagnoses. Associations between the derived classification and known early-life determinants, such as birth order or family history, were investigated. The trajectories aligned with clinical observations as examined by their association with lung function and allergic sensitization measures, all used in the diagnosis of different allergic diseases. We further studied the genetic association of these trajectories with PRS of any allergic disease or the specific disease PRS for asthma, atopic dermatitis and rhinitis. Our results were successfully replicated in the independent Swedish BAMSE cohort.

Publication 2: Kilanowski, A, Chen J, Everson T, et al. Methylation Risk Scores for Childhood Aeroallergen Sensitization: Results from the LISA Birth Cohort. Allergy. https://doi.org/10.1111/all.15315 [IF = 14.710]

For the second publication, we focus on epigenetic DNAm data and apply a methodology similar to the PRS on DNAm to calculate MRS for present aeroallergen sensitization as main outcome. DNAm data was available for two time points in the LISA study, 6 and 10 years of age, and we assessed the association of MRS with aeroallergen sensitization in both cross-sectional and prospective settings, as well as in a dose-response context. Furthermore, the effects of MRS were compared to those from other known determinants of allergic diseases, such as early-life determinants and PRS.

Publication 3: Kilanowski, A, Merid SK, Abrishamcar S, et al. DNA methylation and aeroallergen sensitization: The chicken or the egg?. Clin Epigenet 14, 114 (2022). https://doi.org/10.1186/s13148-022-01332-5 [IF = 7.291]

In the third publication, we assess the underlying causal mechanisms between genetic and environmental risk factors with aeroallergen sensitization and whether these are predicted by DNAm or if rather present aeroallergen sensitization influences changes in DNAm. To determine causal tendencies, we employed high-dimensional and causal mediation analysis and tested both effect directions, from exposure over DNAm to sensitization and over allergic sensitization to DNAm changes. To incorporate the high-dimensional nature of DNAm data, MRS as already aggregated DNAm variables, were employed but also three new approaches for high-dimensional mediation analyses. These were used to identify single CpGs which could act as potential mediators or predictors of allergic sensitization. This analysis was possible due to the longitudinal character of the LISA data and follows up both, identified single CpGs but also the previously established MRS, as mediators and outcomes of the two inspected effect direction, in a causal mediation analysis.

3 Methods

Detailed descriptions of the employed methods and data sources used for the present dissertation are included in the respective publications. Nonetheless, the available data, outcome and allergic disease definitions as well as genetic and DNAm data and main statistical approaches are summarized briefly in this section.

3.1 Study design and population

The present thesis is based on data from the two prospective German birth cohorts, GINIplus and LISA.
The German Infant Study on the Influence of Nutrition Intervention plus Air pollution and Genetics on Allergy Development (GINIplus) study recruited 5591 full-term healthy newborns, born between 1995 and 1998, in the study centers Munich and Wesel. It further included a randomized clinical trial on the effect of different hydrolyzed cow-milk formulas on the development of allergic diseases and latest results are published in Gappa et al., 2021[66]. The Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany (LISA) study included 3094 full-term healthy newborns born in Munich, Wesel, Bad Honnef and Leipzig between 1997 and 1999. Ethical approval for both studies was granted by the respective local ethics committees and written consent from parents and participants was obtained. Participants were followed up at 1, 2, 4, 6, 10 and 15 years of age and additionally at 3 years in GINIplus and 0.5 and 1.5 years in LISA. Both studies are still ongoing with currently the 25-year follow-up taking place. Further details of both studies can be found elsewhere[67].
The prospective nature of both studies and the mostly harmonized methodology, enabled the observation of allergic disease development covering the vulnerable time windows in both infancy and adolescence and thus allow answering the research questions of this dissertation.

3.2 Variable definitions and datasets

Main outcomes and data sources used in this thesis are described in the following paragraphs. This includes allergic disease definitions, early-life determinants and genetic as well as DNAm datasets.

3.2.1 Allergic diseases

Allergic diseases were the main outcomes investigated in this thesis. Here, we can classify them in three groups:

1. **Main allergic disease definitions -** Asthma, atopic dermatitis and rhinitis were defined according to parent-reported doctor's diagnoses, which were asked during each study follow-up covering all years since the last follow-up. Diagnoses were aggregated to harmonized follow-up time points at 1, 2, 4, 6, 10 and 15 years of age.

2. **Symptoms of allergic diseases -** For sensitivity analyses, we further used parent-reported allergic symptoms: Wheezing for asthma, itchy rash at pre-defined body regions for atopic dermatitis and runny nose without having a common cold for rhinitis. These allergic symptoms are further being used by the international ISAAC study[68].
3. **Aeroallergen sensitization -** Aeroallergen sensitization was measured in blood at 6, 10 and 15 years of age with the Radio-Allergo-Sorbent-Test (RAST) test using the CAP-RAST FEIA system by Pharmacia Diagnostics (Freiburg, Germany) according to the manufacturer's instructions. A screening test for specific IgE antibodies against the most common aeroallergens (SX1) was employed. Positive sensitization to the screening test was defined by a specific IgE threshold of > 0.35 kU/L, representing RAST class 1 or higher.

3.2.2 Early-life determinants

Early-life determinants associated with allergic diseases have been further assessed in the context of this dissertation: In the first publication to characterize allergic trajectories, as comparative predictors to MRS in the second publication and as exposures (family history of allergic diseases and maternal smoking) in the mediation analyses of the third publication. All early-life determinants were selected based on existing literature and relevant studies investigating their influence on allergic diseases and their definitions in the GINIplus/LISA studies are included in the respective supplementary material of each publication.

3.2.3 Genetic data

Genetic data was available from sub-populations of both birth cohorts (GINIplus and LISA) in the two study centers Munich and Wesel. For Munich, samples were genotyped using the Affymetrix Chips 5.0 and 6.0 (Thermo Fisher, USA) and for Wesel using the Infinium Global Screening Array GSA v2 MD. After quality control, the final samples consist of 1511 participants with available genetic data for Munich and 792 for Wesel. For imputation, the Haplotype Reference Consortium (HRC) version 1.1[69] was used.
Polygenic risk scores (PRS) were calculated based on available GWAS summary statistics for any allergic disease[31;49], asthma[44], atopic dermatitis[45] and rhinitis[48]. In brief, we performed an initial quality control of reported SNPs and excluded multi-allelic ones, rare variants (Minor allele frequency (MAF) < 0.1), correlated (Linkage disequilibrium (LD) R2 > 0.7) and badly imputed ones (Rsq < 0.4). Afterwards, weighted sums of individual allele dosages with the published GWAS effect estimates were calculated and standardized to z-scores.

3.2.4 Epigenetic data

Epigenetic data on DNAm was available for participants from the LISA study based on paired measurements at 6 and 10 years. Blood-clot samples from both time points were analyzed using the Illumina EPIC Chip for 256 participants. Paired samples for each participant were always placed on the same chip to avoid batch effects and pre-processing and quality control were performed in R[70]. The final dataset contains 461 samples from 240 participants of the LISA study, 234 measurements at 6 and 227 measurements at 10 years, with 774,330 CpG sites each.
MRS were calculated similar to the PRS based on published EWAS summary statistics as weighted sum scores, this time the beta-values of the respective CpG loci were weighted with the EWAS-identified effect estimates. In addition, we used clumping and thresholding for score

refinement: *Clumping* describes the identification of correlated CpGs with the CoMeBack method[71] to retain only the most significant one and to exclude multiple correlated effects within close proximity. *Thresholding* of the included CpGs, uses varying p-value cut-offs to include only CpG loci up to a certain significance level, CpGs with a lower significance are not included in the final scores.

3.3 Statistical considerations

Statistical methods presented within this dissertation can be grouped into three main topics: Unsupervised clustering of allergic trajectories from birth to adolescence, polygenic and methylation risk scores as a method to aggregate high-dimensional genetic and epigenetic data into easily usable scores and high-dimensional mediation analyses, intended to investigate the causal mechanisms between prenatal exposures and DNAm leading to allergic phenotypes. All statistical analyses were performed in R[70].

3.3.1 Clustering of allergic diseases

Clustering techniques used within this dissertation are k-means clustering for the main analysis of the first publication and latent class analysis (LCA) for the respective sensitivity analysis.
K-means clustering was implemented using the *kml3d* package[72] in R, which can implement both the time axis and different diseases as input. In short, k-means assigns k random observations as cluster centers and allocates all surrounding observations to their nearest center. Afterwards, new centers are assigned based on the most central observation and the process is repeated until the allocations do not change anymore. We repeated these steps for differing number of clusters (k) from 2 to 9 and chose the best number of allergic disease clusters according to the quality criterion of Davies-Bouldin[73].
LCA was performed as a sensitivity analysis for deriving the allergic trajectories but cannot implement time progression and different diseases in the algorithm but rather uses the variables in the pre-defined order. LCA identifies latent patterns in the observations to reduce dimensionality of the data and was here performed using the *depmixS4* package[74] and determining the best partition based on Bayesian information criterion (BIC) and log-likelihood.
Summarizing, we applied two different unsupervised clustering methods to derive valid clusters of allergic disease trajectories and validated our main results, derived using k-means clustering, with LCA.

3.3.2 Polygenic and methylation risk scores

In the area of high-dimensional omics data, reduction methods to accumulate the previous genetic or epigenetic evidence into single variables have gained interest in the scientific community. Within this dissertation, we used polygenic (PRS) as well as methylation risk scores (MRS) to aggregate results from published GWAS and EWAS studies.
In brief, both are weighted sum scores based on previously published summary statistics, where the effect sizes of the identifying studies are multiplied with the dosage of the respective SNP (for PRS) or the beta-value of the respective CpG loci (for MRS). Both, PRS and MRS, are transformed to z-scores for analysis to make them comparable, despite differing numbers of included variants identified in respective GWAS or EWAS.

3.3.3 High-dimensional mediation analysis

Within the present dissertation, we applied three recent high-dimensional mediation analysis methods, namely the Divide-Aggregate Composte Null Test (DACT)[75], high-dimensional mediation analysis (HIMA)[76] and gene-based high-dimensional mediation analysis (gHMA)[77], to identify single CpGs or genes potentially mediating the pathway from genetic and environmental exposures to aeroallergen sensitization. Traditional causal mediation analysis was used to validate them.
There are several challenges in high-dimensional mediation analysis, which we tackled with the different methods, as each of them brings their unique advantages to the field:

1. *Multiple testing burden* - As the number of tested potential mediators is much higher than the sample size of the analysis, we took advantage of previous knowledge on significant EWAS hits to filter the number of mediators for DACT. For HIMA and gHMA, as hypothesis free approaches, the first step of the algorithm selects the top CpGs/genes with the highest effect size via sure-independence-screening and only uses these in further analyses.

2. *Composite structure of the null hypothesis* - There are three scenarios all leading to non-significant mediation (a) the association between exposure and mediator is not significant; (b) the association between mediator and outcome is not significant or (c) both associations are not significant. DACT leverages the high-dimensionality of the data by using the provided p-values from both association paths (exposure-mediator and mediator-outcome) to estimate the proportions of each scenario to increase the power for their newly calculated p-values.

3. *Small effect sizes of single CpGs* - As single CpGs only explain a small fraction of epigenetic variability and often only employ very small effect sizes, HIMA and gHMA deal with this problem by identifying only the top CpGs/genes with the highest effects and in the case of gHMA additionally aggregating them on gene-level. The idea behind the latter is that not single CpGs but rather whole genes are mediating disease courses and the evidence can be aggregated into this unit.

4 Results

The general aim of this dissertation was to improve the understanding of the longitudinal development of allergic diseases as well as the role of their genetic and epigenetic variations. Here, key findings are summarized for each publication.

4.1 Key findings

In the first publication, we aimed to characterize the joint development of the allergic diseases, asthma, atopic dermatitis and rhinitis, in the GINIplus and LISA studies. Additionally, factors determining these trajectories, as early-life determinants and PRS, were described and assessed. We replicated theses in the Swedish BAMSE cohort and found robust replication of the seven allergic ("Intermittently allergic", "Rhinitis", "Early-resolving dermatitis", "Mid-persisting dermatitis", "Multimorbid", "Persistent dermatitis + rhinitis" and "Early-transient asthma") and a pre-specified non-allergic cluster. The trajectories were further tested for clinical characteristics, such as spirometry and allergic sensitization. Here, we saw that clusters comprising a high percentage of asthma cases present lower lung function values and aeroallergen sensitization was highest in rhinitis-containing ones. A differentiation of allergic development is presented based on the PRS for asthma, differentiating between the rhinitis and multimorbid cluster and based on food allergen sensitization, differentiating early-resolving from persisting dermatitis cases.
Looking at the clinical perspective, 6.2% of our sample in GINIplus and LISA were sorted into allergic multimorbidity cluster, whereas 35.9% belonged to specific allergic disease clusters. These numbers increased to 14.8% and 54.5% in BAMSE, respectively, based on the higher prevalence of symptoms compared to doctor's diagnoses.

Key finding 1: There are seven robust allergic disease trajectories, which follow clinical observations and characteristics and can further be differentiated based on disease-specific PRS, hence underlining the genetic component in allergic disease development. The information on shared genetic and early-life determinants might help to understand phenotypic differences in clinical practice and start to facilitate future predictions of disease courses.

For the second publication, we built on pre-existing knowledge for PRS and published EWAS on allergic phenotypes to calculate MRS in a subset of the LISA cohort. As MRS are not as established as PRS, we first evaluated their association with our outcome of aeroallergen sensitization and found an around 80% increased risk of aeroallergen sensitization with a one standard deviation increase in MRS in a cross-sectional model at 6 years. We further present dose-response relationships between MRS and RAST classes of aeroallergen sensitization. Results were stronger in cross-sectional analyses compared to prospective ones, thus we concluded that MRS are applicable as cross-sectional biomarkers of the disease but not as predictors of future aeroallergen sensitization. Cell type specific effects were noticed in regard to the replication of previous EWAS hits and should be investigated further.

Key finding 2: MRS can be applied as cross-sectional biomarkers of aeroallergen sensitization, which can be used in future research but also support the association between DNAm and allergic diseases and underline previous EWAS findings. Moreover, the cross-sectional nature of MRS performing as biomarkers, rather than predictors, supports the hypothesis that allergic diseases are influencing DNAm profiles.

In the last publication of this dissertation, we conducted high-dimensional and causal mediation analyses to determine whether changes in DNAm are caused by aeroallergen sensitization, as indicated by the findings of the second publication, or if there are further DNAm patterns which are predicting allergic sensitization. We identified five CpGs which are causally mediating the relationship between environmental and genetic risk factors (maternal smoking, family history of allergies and PRS for any allergic disease) and aeroallergen sensitization at consecutive time points. On the other hand, we could confirm that our previously calculated MRS are mediated via prior sensitization. Thus, we provided evidence for both causal directions, DNAm influences allergic phenotypes and allergic phenotypes change DNAm, for future research.

Key finding 3: There are two sets of DNAm markers, one as a consequence of prior allergic sensitization. The other set mediates the association between genetic effects and allergic sensitization and thus works as a predictor of aeroallergen sensitization.

4.2 Strengths and limitations

Please note that publication-specific strengths and limitations are included in the respective publications and only general strengths and limitations are discussed here.
The present dissertation is based on data from two well-established German birth cohorts and can therefore rely on a vast amount of prospectively collected data over the years. This strength is amplified by the addition of available genetic data in GINIplus and LISA and DNAm data in LISA at two time points, which enables participation in up-to-date biomedical research combined with long-time follow-up data on the same participants.
Several state of the art methods, such as PRS and MRS but also high-dimensional mediation approaches, were applied in the context of this dissertation, demonstrating their applicability in the context of allergic diseases and promoting understanding of longitudinal development as well as its background in genetic and epigenetic DNAm variations.

We also want to note a few limitations: As both studies employ a long follow-up time up to adolescence, we naturally had to deal with missing data due to loss to follow up. For the first publication, we included only participants with less than 3 missing diagnoses per allergic disease and applied last observation carried forward imputation for the remaining participants. This might lead to an overestimation of allergic disease prevalences, especially for atopic dermatitis, which displays also resolving courses. Furthermore, prevalence estimates in general might not reflect the general population but are still valid for our studies. Publications 2 and 3 only include complete observations in their respective time windows.
Further, sample size is still an issue and while the full GINIplus and LISA studies provide a sufficient sample size for regular statistical analyses, this may pose a limitation in high-dimensional analyses relying on our DNAm data. It is still noteworthy, that calculated MRS relied on external evidence and demonstrated applicability in this new context.

5 Conclusion and outlook

Allergic diseases pose an important challenge for public health, as they are affecting large portions of the population in especially industrialized countries and there is currently no cure for them. The present dissertation aimed to investigate the longitudinal development of allergic diseases, allergic sensitization and the contributing role of DNAm.
First, we identified 7 distinct and robust allergic disease trajectories, which are associated with genetic factors but also with common early-life determinants. These associations enabled us to describe and characterize the allergic trajectories and describe their differences. For example does the multimorbid cluster, presenting all three allergic diseases, show significant associations with the PRS for any allergy, asthma and rhinitis, which differentiates it from the rhinitis cluster. Hence, we hypothesized that an increased genetic predisposition might be one factor distinguishing these two clusters. As the trajectories follow known clinical observations and are further supported by their associations with lung function and allergic sensitization, this classification provides further information for both pediatricians and the scientific community. Further, early-life determinants were associated with distinct trajectories and it should be determined, which of these are the most powerful drivers usable for prediction. PRS approaches might further be applied to stratify the population in at-risk individuals, who could benefit from certain preventive measures.
In the second publication, we established MRS as biomarkers of aeroallergen sensitization, which (1) demonstrated the applicability of this rather new concept to allergic phenotypes and (2) supports previous EWAS results and their validity. Associations and dose-response trends between MRS and aeroallergen sensitization were independent of ancestries of original EWAS, number of included CpGs or allergic phenotypes. Summarizing, the results indicate that MRS are applicable as cross-sectional biomarkers of aeroallergen sensitization.
Assessing the causality between DNAm and aeroallergen sensitization, we found evidence for both effect directions and DNAm could be both a cause and a consequence of allergic diseases. More specifically, the results suggest that there are two distinct sets of CpGs, one altered by previous aeroallergen sensitization and one predicting it. CpGs and their respective genes, which were identified as mediators preceding allergic sensitization, could be studied as results of environmental exposures or as future targets of drugs. Further, establishing a biomarker based on predicting CpGs formed into a MRS in combination with the presence of allergic sensitization might facilitate earlier prediction of allergies.
As both effect directions, DNAm influencing aeroallergen sensitization and sensitization influencing DNAm, seem to be occurring in allergic diseases, it is an interesting field of future research. This will help to combine genetic and environmental determinants inducing changes in DNAm, which might further explain mechanisms in allergic development.

6 Author contributions

Publications included in this thesis are published either in Allergy (Publications 1 & 2; Impact factor = 14.710) or Clinical Epigenetics (Publication 3; Impact factor = 7.291), which rank 1st of 27 and 19th of 175 in the categories Allergy and Genetics & Heredity according to the Journal Citation Report^{TM} 2021, respectively.
I am the first author in all of them and was significantly involved in the research question development. I performed all main analyses, wrote the initial drafts, incorporated co-author and reviewer comments and lead the submission process. For the first and third publication, I provided analysis scripts to the replicating BAMSE cohort and collected and prepared their results for publication.
Research ideas were developed under the supervision of my direct supervisors at Helmholtz Munich, Dr. Marie Standl and Dr. Elisabeth Thiering (first publication) and under the supervision of Dr. Marie Standl and Dr. Anke Hüls (Emory University, Atlanta, USA) for the second and third publication. Ideas were always supported by my main supervisor Prof. Dr. Annette Peters and two further members of my thesis advisory committee, Prof. h.c. Dr. med. Habil. Dr. h.c. Berthold Koletzko and Prof. Dr. Irina Lehmann.
Further, I performed the quality control and data processing of our LISA DNA methylation data, which I used in the second and third publication. I adapted preprocessing scripts from Rory Wilson and Dr. Nicole Gladish (Stanford University, California, USA), ran the processing and wrote the respective documentation. We received further help from Dr. Melanie Waldenberger and Nadine Lindemann.

6.1 Further projects

Next to the three publications included in this thesis, I was involved in several other projects. In cooperation with Dr. Lavinia Paternoster (University of Bristol, UK), we worked on a GWAS meta-analysis on atopic dermatitis, which is currently submitted for publication in Nature Communications. In brief, we collected GWAS summary data from over 40 cohorts with 1,086,394 participants for the discovery and 3,604,027 for the replication and performed both European-ancestry and multi-ancestry meta-analyses. We found 35 novel loci and could confirm further 61 loci (Appendix A1). Together with Dr. Ashley Budy-Aggrey (University of Bristol, UK) I performed the main quality control of incoming cohort data. Further, I ran the European-ancestry meta-analysis and contributed the supplementary material on Manhattan, QQ and Forest plots as well as on regional association plots.
Building on the work of a previous master student, I was further involved in the analysis of the platelet-derived growth factor receptor alpha (PDGF-Rα) pathway in our GINIplus/LISA data, in regard to lung function and lung diseases. This pathway is known to be associated with lung disease in preterm newborns and reduced lung function later in life. Here, we extended this known paradigm to adolescence. For the later part of the analysis, I calculated PRS, reran the analyses and prepared the results for statistical interpretation. The resulting publication is currently in preparation.
Furthermore, I was involved in the supervision of another Master student, who worked on developing a prediction model for allergic disease trajectories, which were derived for my

first publication. He used different feature-selection and prediction models to predict allergic trajectories with risk factors until the age of four, building onto my first publication. This project directly builds upon my first publication and I was involved by providing guidance on the outcome, analyses and thesis submission.

Next to my PhD, I was further engaged in the doctoral initiative (DINI) at Helmholtz Munich since June 2020, where I was involved in the development of new PhD guidelines, planning of social events and communication with the graduate school HELENA. For the year 2021, I was the elected doctoral representative at Helmholtz Munich.
Further, I was the active representative of Helmholtz Munich in the Helmholtz Juniors (August 2020 to November 2022), the German-wide representation of doctoral researchers within the Helmholtz association. Here, I was speaker of the survey working group and developed, ran, analyzed and reported the harmonized N2-Survey 2021 together with representatives from the Max-Planck society and Leibniz association. This will result in the publication and dissemination of three publications on working conditions, power abuse and mental health of doctoral researchers within Germany's non-university research organizations, a full report of results collected in the Helmholtz association and center-specific reports for 17 of the 18 Helmholtz centers communicated to the respective general administrations.

7 References

[1] Devereux G. The increase in the prevalence of asthma and allergy: food for thought. Nature Reviews Immunology. 2006 Nov;6(11):869-74. Number: 11 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nri1958.

[2] Platts-Mills TAE. The allergy epidemics: 1870-2010. Journal of Allergy and Clinical Immunology. 2015 Jul;136(1):3-13. Available from: https://www.sciencedirect.com/science/article/pii/S0091674915005849.

[3] Holgate ST. The epidemic of allergy and asthma. Nature. 1999 Nov;402(6760 Suppl):B2-4.

[4] World Health Organization. White Book on Allergy 2011-2012 Executive Summary. By Prof. Ruby Pawankar, MD, PhD, Prof. Giorgio Walkter Canonica, MD, Prof. Stephen T. Holgate, BSc, MD, DSc, FMed Sci and Prof. Richard F. Lockey, MD.; 2012.

[5] Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet. 2020 Oct;396(10258):1204-22. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30925-9/abstract.

[6] Dierick BJH, van der Molen T, Flokstra-de Blok BMJ, Muraro A, Postma MJ, Kocks JWH, et al. Burden and socioeconomics of asthma, allergic rhinitis, atopic dermatitis and food allergy. Expert Review of Pharmacoeconomics & Outcomes Research. 2020 Oct;20(5):437-53.

[7] Asher MI, Ellwood P. The global asthma report 2014. 2014.

[8] Mallol J, Crane J, von Mutius E, Odhiambo J, Keil U, Stewart A. The International Study of Asthma and Allergies in Childhood (ISAAC) Phase Three: A global synthesis. Allergologia et Immunopathologia. 2013 Mar;41(2):73-85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0301054612001097.

[9] Thamm R, Poethko-Müller C, Hüther A, Thamm M. Allergische Erkrankungen bei Kindern und Jugendlichen in Deutschland – Querschnittergebnisse aus KiGGS Welle 2 und Trends. Journal of Health Monitoring. 2018 Mar;3(3).

[10] Strachan DP, Rutter CE, Asher MI, Bissell K, Chiang CY, El Sony A, et al. Worldwide time trends in prevalence of symptoms of rhinoconjunctivitis in children: Global Asthma Network Phase I. Pediatric Allergy and Immunology. 2022;33(1):e13656. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pai.13656. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/pai.13656.

[11] Grisanti K, Grayson MH. The Allergy Epidemic. In: Stukus DR, editor. Allergies and Adolescents: Transitioning Towards Independent Living. Cham: Springer International Publishing; 2018. p. 1-14.

[12] Kopp MV, Muche-Borowski C, Abou-Dakn M, Ahrens B, Beyer K, Blümchen K, et al. S3 guideline Allergy Prevention. Allergologie Select. 2022;6:61-97.

[13] Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clinical and Experimental Immunology. 2010 Apr;160(1):1-9.

[14] Johansson SGO, Hourihane Jo, Bousquet J, Bruijnzeel-Koomen C, Dreborg S, Haahtela T, et al. A revised nomenclature for allergy: An EAACI position statement from the EAACI nomenclature task force. Allergy. 2001;56(9):813-24. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1398-9995.2001.00002.x-i1. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1398-9995.2001.00002.x-i1.

[15] for Asthma GI. Global Strategy for Asthma Management and Prevention; 2022. Available from: Availablefrom:www.ginasthma.org.

[16] Papadopoulos NG, Miligkos M, Xepapadaki P. A Current Perspective of Allergic Asthma: From Mechanisms to Management. Handbook of Experimental Pharmacology. 2022;268:69-93.

[17] Pearce N, Aït-Khaled N, Beasley R, Mallol J, Keil U, Mitchell E, et al. Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax. 2007;62(9):758-66.

[18] Weidinger S, Beck LA, Bieber T, Kabashima K, Irvine AD. Atopic dermatitis. Nature Reviews Disease Primers. 2018 Jun;4(1):1.

[19] Weidinger S, Novak N. Atopic dermatitis. The Lancet. 2016 Mar;387(10023):1109-22. Available from: http://www.sciencedirect.com/science/article/pii/S014067361500149X.

[20] Hanifin J, Rajka G. Diagnostic Features of Atopic Dermatitis. Acta Dermato-Venereologica. 1980;60(92):44-7. Available from: https://www.medicaljournals.se/acta/content/abstract/10.2340/00015555924447.

[21] Hansel FK. Clinical and histopathologic studies of the nose and sinuses in allergy. Journal of Allergy. 1929 Nov;1(1):43-70. Available from: https://www.sciencedirect.com/science/article/pii/S0021870729900836.

[22] Klimek L, Bachert C, Pfaar O, Becker S, Bieber T, Brehler R, et al. ARIA guideline 2019: treatment of allergic rhinitis in the German health system. Allergologie Select. 2019 Dec;3(1):22-50. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7066682/.

[23] Nur Husna SM, Tan HTT, Md Shukri N, Mohd Ashari NS, Wong KK. Allergic Rhinitis: A Clinical and Pathophysiological Overview. Frontiers in Medicine. 2022;9. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2022.874114.

[24] Nissen SP, Kjær HF, Høst A, Nielsen J, Halken S. The natural course of sensitization and allergic diseases from childhood to adulthood. Pediatric Allergy and Immunology. 2013;24(6):549-55. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/pai.12108.

[25] Rönmark E, Perzanowski M, Platts-Mills T, Lundbäck B. Different sensitization profile for asthma, rhinitis, and eczema among 7-8-year-old children: Report from the Obstructive Lung Disease in Northern Sweden studies. Pediatric Allergy and Immunology. 2003;14(2):91-9. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1034/j.1399-3038.2003.00042.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1399-3038.2003.00042.x.

[26] Pinart M, Benet M, Annesi-Maesano I, von Berg A, Berdel D, Carlsen KCL, et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitised and non-IgE-sensitised children in MeDALL: a population-based cohort study. The Lancet Respiratory Medicine. 2014 Feb;2(2):131-40.

[27] Hopper JL, Bui QM, Erbas B, Matheson MC, Gurrin LC, Burgess JA, et al. Does eczema in infancy cause hay fever, asthma, or both in childhood? Insights from a novel regression model of sibling data. Journal of Allergy and Clinical Immunology. 2012 Nov;130(5):1117-22.e1. Available from: https://www.sciencedirect.com/science/article/pii/S0091674912012857.

[28] Saunes M, Øien T, Dotterud CK, Romundstad PR, Storrø O, Holmen TL, et al. Early eczema and the risk of childhood asthma: a prospective, population-based study. BMC Pediatrics. 2012 Oct;12(1):168. Available from: https://doi.org/10.1186/1471-2431-12-168.

[29] Dharmage SC, Lowe AJ, Tang MLK. Revisiting the Atopic March Current Evidence. American Journal of Respiratory and Critical Care Medicine. 2022 Oct;206(8):925-6. Publisher: American Thoracic Society - AJRCCM. Available from: https://www.atsjournals.org/doi/10.1164/rccm.202206-1219ED.

[30] Khan SJ, Dharmage SC, Matheson MC, Gurrin LC. Is the atopic march related to confounding by genetics and early-life environment? A systematic review of sibship and twin data. Allergy. 2018;73(1):17-28. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/all.13228. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.13228.

[31] Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nature genetics. 2017 Dec;49(12):1752-7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5989923/.

[32] Gluckman PD, Buklijas T, Hanson MA. Chapter 1 - The Developmental Origins of Health and Disease (DOHaD) Concept: Past, Present, and Future. In: Rosenfeld CS, editor. The Epigenome and Developmental Origins of Health and Disease. Boston: Academic Press; 2016. p. 1-15. Available from: https://www.sciencedirect.com/science/article/pii/B9780128013830000013.

[33] Wegienka G, Zoratti E, Johnson CC. The Role of the Early-Life Environment in the Development of Allergic Disease. Immunology and allergy clinics of North America. 2015 Feb;35(1):1-17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4427897/.

[34] Thacher JD, Gehring U, Gruzieva O, Standl M, Pershagen G, Bauer CP, et al. Maternal Smoking during Pregnancy and Early Childhood and Development of Asthma and Rhinoconjunctivitis - a MeDALL Project. Environmental Health Perspectives. 2018 Apr;126(4):047005.

[35] Hu C, Duijts L, Erler NS, Elbert NJ, Piketty C, Bourdès V, et al. Most associations of early-life environmental exposures and genetic risk factors poorly differentiate between eczema phenotypes: the Generation R Study. The British Journal of Dermatology. 2019 Dec;181(6):1190-7.

[36] Uphoff E, Cabieses B, Pinart M, Valdés M, Antó JM, Wright J. A systematic review of socioeconomic position in relation to asthma and allergic diseases. European Respiratory Journal. 2015 Aug;46(2):364-74. Publisher: European Respiratory Society Section: Original Articles. Available from: https://erj.ersjournals.com/content/46/2/364.

[37] Lodge C, Tan D, Lau M, Dai X, Tham R, Lowe A, et al. Breastfeeding and asthma and allergies: a systematic review and meta-analysis. Acta Paediatrica. 2015;104(S467):38-53. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apa.13132. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/apa.13132.

[38] Strachan DP. Hay fever, hygiene, and household size. British Medical Journal. 1989 Nov;299(6710):1259-60. Publisher: British Medical Journal Publishing Group Section: Research Article. Available from: https://www.bmj.com/content/299/6710/1259.

[39] Strachan DP. Family size, infection and atopy: the first decade of the 'hygiene hypothesis'. Thorax. 2000 Aug;55(Suppl 1):S2-S10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1765943/.

[40] Haahtela T. A biodiversity hypothesis. Allergy. 2019;74(8):1445-56. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/all.13763. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.13763.

[41] Ballardini N, Kull I, Lind T, Hallner E, Almqvist C, Ostblom E, et al. Development and comorbidity of eczema, asthma and rhinitis to age 12: data from the BAMSE birth cohort. Allergy. 2012 Apr;67(4):537-44.

[42] 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. The Journal of Allergy and Clinical Immunology. 2012 Oct;130(4):912-7.

[43] Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunological Reviews. 2011 Jul;242(1):10-30.

[44] El-Husseini ZW, Gosens R, Dekker F, Koppelman GH. The genetics of asthma and the promise of genomics-guided drug target discovery. The Lancet Respiratory Medicine. 2020 Oct;8(10):1045-56.

[45] Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature Genetics. 2015 Dec;47(12):1449-56.

[46] Grosche S, Marenholz I, Esparza-Gordillo J, Arnau-Soler A, Pairo-Castineira E, Rüschendorf F, et al. Rare variant analysis in eczema identifies exonic variants in DUSP1, NOTCH4 and SLC9A4. Nature Communications. 2021 Nov;12(1):6618.

[47] Sliz E, Huilaja L, Pasanen A, Laisk T, Reimann E, Mägi R, et al. Uniting biobank resources reveals novel genetic pathways modulating susceptibility for atopic dermatitis. The Journal of Allergy and Clinical Immunology. 2021 Aug:S0091-6749(21)01303-8.

[48] Waage J, Standl M, Curtin JA, Jessen LE, Thorsen J, Tian C, et al. Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. Nature Genetics. 2018;50(8):1072-80.

[49] Ferreira MAR, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Eleven loci with new reproducible genetic associations with allergic disease risk. Journal of Allergy and Clinical Immunology. 2019 Feb;143(2):691-9. Available from: http://www.sciencedirect.com/science/article/pii/S009167491830558X.

[50] Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A Large-Scale, Consortium-Based Genomewide Association Study of Asthma. The New England journal of medicine. 2010 Sep;363(13):1211-21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4260321/.

[51] Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics. 2018 Sep;50(9):1219-24. Available from: https://www.nature.com/articles/s41588-018-0183-z.

[52] Melén E, Koppelman GH, Vicedo-Cabrera AM, Andersen ZJ, Bunyavanich S. Allergies to food and airborne allergens in children and adolescents: role of epigenetics in a changing environment. The Lancet Child & Adolescent Health. 2022 Aug. Available from: https://www.sciencedirect.com/science/article/pii/S2352464222002152.

[53] Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology. 2013 Jan;38(1):23-38. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/npp2012112.

[54] Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. American Journal of Human Genetics. 2016 Apr;98(4):680-96.

[55] Aguilera J, Han X, Cao S, Balmes J, Lurmann F, Tyner T, et al. Increases in ambient air pollutants during pregnancy are linked to increases in methylation of IL4, IL10, and IFNγ. Clinical Epigenetics. 2022 Mar;14(1):40.

[56] Park J, Kim WJ, Kim J, Jeong CY, Park H, Hong YC, et al. Prenatal Exposure to Traffic-Related Air Pollution and the DNA Methylation in Cord Blood Cells: MOCEH Study. International Journal of Environmental Research and Public Health. 2022 Mar;19(6):3292.

[57] Barouki R, Melén E, Herceg Z, Beckers J, Chen J, Karagas M, et al. Epigenetics as a mechanism linking developmental exposures to long-term toxicity. Environment International. 2018 May;114:77-86.

[58] Wu H, Zhang Y. Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. Cell. 2014 Jan;156(0):45-68. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3938284/.

[59] Zhang H, Kaushal A, Merid SK, Melén E, Pershagen G, Rezwan FI, et al. DNA methylation and allergic sensitizations: A genome-scale longitudinal study during adolescence. Allergy. 2019;74(6):1166-75.

[60] Reese SE, Xu CJ, Dekker HTd, Lee MK, Sikdar S, Ruiz-Arenas C, et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. Journal of Allergy and Clinical Immunology. 2019 Jun;143(6):2062-74. Publisher: Elsevier. Available from: https://www.jacionline.org/article/S0091-6749(18)32788-X/abstract.

[61] Xu CJ, Gruzieva O, Qi C, Esplugues A, Gehring U, Bergström A, et al. Shared DNA methylation signatures in childhood allergy: The MeDALL study. Journal of Allergy and Clinical Immunology. 2021 Mar;147(3):1031-40. Publisher: Elsevier. Available from: https://www.jacionline.org/article/S0091-6749(20)31766-8/abstract.

[62] Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics. 2019 Jul;0(0):1-11. Available from: https://doi.org/10.1080/15592294.2019.1644879.

[63] Davidson EJ, Yang IV. Role of epigenetics in the development of childhood asthma. Current Opinion in Allergy and Clinical Immunology. 2018 Apr;18(2):132-8.

[64] Hernandez-Pacheco N, Kere M, Melén E. Gene-environment interactions in childhood asthma revisited; expanding the interaction concept. Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology. 2022 May;33(5):e13780.

[65] Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nature Genetics. 2021 Sep;53(9):1311-21. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Epigenetics;Genetics research Subject_term_id: epigenetics;genetics-research. Available from: https://www.nature.com/articles/s41588-021-00923-x.

[66] Gappa M, Filipiak-Pittroff B, Libuda L, von Berg A, Koletzko S, Bauer CP, et al. Long-term effects of hydrolyzed formulae on atopic diseases in the GINI study. Allergy. 2021 Jun;76(6):1903-7.

[67] Heinrich J, Brüske I, Cramer C, Hoffmann U, Schnappinger M, Schaaf B, et al. GINIplus and LISAplus – Design and selected results of two German birth cohorts about natural course of atopic diseases and their determinants. Allergologie Select. 2017 Aug;1(1):85-95. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6040001/.

[68] Asher MI, Montefort S, Björkstén B, Lai CK, Strachan DP, Weiland SK, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. The Lancet. 2006 Aug;368(9537):733-43. Available from: http://www.sciencedirect.com/science/article/pii/S0140673606692830.

[69] McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016 Oct;48(10):1279-83. Number: 10 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng.3643.

[70] Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/.

[71] Gatev E, Gladish N, Mostafavi S, Kobor MS. CoMeBack: DNA methylation array data analysis for co-methylated regions. Bioinformatics (Oxford, England). 2020 May;36(9):2675-83.

[72] Genolini C, Alacoque X, Sentenac M, Arnaud C. kml and kml3d: R Packages to Cluster Longitudinal Data. Journal of Statistical Software. 2015 Jun;65(1):1-34. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v065i04.

[73] Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979 Apr;PAMI-1(2):224-7.

[74] Visser I, Speekenbrink M. depmixS4: An R Package for Hidden Markov Models. Journal of Statistical Software. 2010;36(7). Available from: http://www.jstatsoft.org/v36/i07/.

[75] Liu Z, Shen J, Barfield R, Schwartz J, Baccarelli AA, Lin X. Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. Journal of the American Statistical Association. 2022 Jan;117(537):67-81. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1914634. Available from: https://doi.org/10.1080/01621459.2021.1914634.

[76] Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics (Oxford, England). 2016 Oct;32(20):3150-4.

[77] Fang R, Yang H, Gao Y, Cao H, Goode EL, Cui Y. Gene-based mediation analysis in epigenetic studies. Briefings in Bioinformatics. 2021 May;22(3). Available from: https://doi.org/10.1093/bib/bbaa113.

[78] Love SB, Kilanowski A, Yorke-Edwards V, Old O, Barr H, Stokes C, et al. Use of routinely collected health data in randomised clinical trials: comparison of trial-specific death data in the BOSS trial with NHS Digital data. Trials. 2021 Sep;22(1):654. Available from: https://doi.org/10.1186/s13063-021-05613-x.

[79] Kilanowski A, Chen J, Everson T, Thiering E, Wilson R, Gladish N, et al. Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort. Allergy. 2022;77(9):2803-17. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15315.

[80] Kress S, Kilanowski A, Wigmann C, Zhao Q, Zhao T, Abramson MJ, et al. Airway inflammation in adolescents and elderly women: Chronic air pollution exposure and polygenic susceptibility. Science of The Total Environment. 2022 Oct;841:156655. Available from: https://www.sciencedirect.com/science/article/pii/S0048969722037524.

[81] Kilanowski A, Thiering E, Wang G, Kumar A, Kress S, Flexeder C, et al. Allergic disease trajectories up to adolescence: Characteristics, early-life, and genetic determinants. Allergy. 2022;n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15511.

[82] Kilanowski A, Merid SK, Abrishamcar S, Feil D, Thiering E, Waldenberger M, et al. DNA methylation and aeroallergen sensitization: The chicken or the egg? Clinical Epigenetics. 2022 Sep;14(1):114. Available from: https://doi.org/10.1186/s13148-022-01332-5.

[83] Abrishamcar S, Chen J, Feil D, Kilanowski A, Koen N, Vanker A, et al. DNA methylation as a potential mediator of the association between prenatal tobacco and alcohol exposure and child neurodevelopment in a South African birth cohort. Translational Psychiatry. 2022 Sep;12(1):1-12. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41398-022-02195-3.

[84] Feil D, Abrishamcar S, Christensen GM, Vanker A, Koen N, Kilanowski A, et al.. DNA methylation as a potential mediator of the association between indoor air pollution and neurodevelopmental delay in a South African birth cohort; 2022. Available from: https://www.researchsquare.com.

[85] Budu-Aggrey A, Kilanowski A, Sobczyk MK, Shringarpure SS, Mitchell R, Reis K, et al.. European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation; 2022. Publisher: Cold Spring Harbor Laboratory Press _eprint: https://www.medrxiv.org/content/early/2022/10/10/2022.10.05.22279072.full.pdf. Available from: https://www.medrxiv.org/content/early/2022/10/10/2022.10.05.22279072.

Part II

Publications

1 Publication: Allergic disease trajectories up to adolescence: Characteristics, early-life and genetic determinants

Title:	Allergic disease trajectories up to adolescence: Characteristics, early-life and genetic determinants
Authors:	Anna Kilanowski, Elisabeth Thiering, Gang Wang, Ashish Kuma, Sara Kress, Claudia Flexeder, Carl-Peter Bauer, Dietrich Berdel, Andrea von Berg, Anna Bergström, Monika Gappa, Joachim Heinrich, Gunda Herberth, Sibylle Koletzko, Inger Kull, Erik Melén, Tamara Schikowski, Annette Peters, Marie Standl
Journal:	*Allergy*
Year:	2022
Volume:	Online ahead of print
DOI:	10.1111/all.15511
Supplements:	https://onlinelibrary.wiley.com/doi/full/10.1111/all.15511 (available online)
Impact factor:	14.710 (Journal Citation Reports^{TM} 2021)
Rank:	1/27 in category ALLERGY (Journal Citation Reports^{TM} 2021)

Received: 18 February 2022 | Revised: 27 July 2022 | Accepted: 13 August 2022

DOI: 10.1111/all.15511

Allergy EUROPEAN JOURNAL OF ALLERGY AND CLINICAL IMMUNOLOGY EAACI WILEY

ORIGINAL ARTICLE

Allergic disease trajectories up to adolescence: Characteristics, early-life, and genetic determinants

Anna Kilanowski[1,2,3] | Elisabeth Thiering[1,3] | Gang Wang[4,5] | Ashish Kumar[5] | Sara Kress[6,7] | Claudia Flexeder[1,8,9] | Carl-Peter Bauer[10] | Dietrich Berdel[11] | Andrea von Berg[11] | Anna Bergström[12,13] | Monika Gappa[14] | Joachim Heinrich[9,15] | Gunda Herberth[16] | Sibylle Koletzko[17,18] | Inger Kull[5,19,20] | Erik Melén[5] | Tamara Schikowski[6] | Annette Peters[1,21] | Marie Standl[1,8]

[1]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[2]Institute for Medical Information Processing, Biometry, and Epidemiology, Pettenkofer School of Public Health, LMU Munich, Munich, Germany

[3]Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, Munich, Germany

[4]Department of Integrated Traditional Chinese and Western Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[5]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden

[6]Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany

[7]Medical Research School Düsseldorf, Heinrich Heine University, Düsseldorf, Germany

[8]German Center for Lung Research (DZL), Munich, Germany

[9]Institute and Clinic for Occupational, Social and Environmental Medicine, University Hospital, LMU Munich, Munich, Germany

[10]Department of Pediatrics, Technical University of Munich, Munich, Germany

[11]Research Institute, Department of Pediatrics, Marien-Hospital Wesel, Wesel, Germany

[12]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[13]Centre for Occupational and Environmental Medicine, Stockholm, Sweden

[14]Evangelisches Krankenhaus Düsseldorf, Children's Hospital, Düsseldorf, Germany

[15]Allergy and Lung Health Unit, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

[16]Department of Environmental Immunology, Helmholtz Centre for Environmental Research-UFZ, Leipzig, Germany

[17]Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany

[18]Department of Pediatrics, Gastroenterology and Nutrition, School of Medicine Collegium Medicum, University of Warmia and Mazury, Olsztyn, Poland

[19]Södersjukhuset, Karolinska Institutet, Stockholm, Sweden

[20]Sachs Children's and Youth Hospital, Stockholm, Sweden

[21]Chair of Epidemiology, Ludwig-Maximilians University, Munich, Germany

Abbreviations: BAMSE, Barn (=Children) Allergy Milieu Stockholm Epidemiology; FEV_1, Forced Expiratory Volume in one second; FLG, Filaggrin; FVC, Forced Vital Capacity; GINIplus, German Infant Study on the Influence of Nutrition Intervention plus Air pollution and Genetics on Allergy Development; GLI, Global Lung Initiative; GWAS, Genome-wide association studies; HRC, Haplotype Reference Consortium; IgE, Immunoglobulin E; ISAAC, International Study of Asthma and Allergies in Childhood; LISA, Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany study; LOCF, Last observation carried forward; PRS, Polygenic Risk Score; RRR, Relative Risk Ratio; SNP, Single Nucleotide Polymorphism.

Correspondence
Marie Standl, Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Institute of Epidemiology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany.
Email: marie.standl@helmholtz-muenchen.de

Funding information
Bundesministerium für Bildung und Forschung; Commission of the European Communities, the 7th Framework Program; Federal Ministry for Environment (IUF Düsseldorf), Grant/Award Number: FKZ 20462296; Fondation Nestlé; H2020 European Research Council; Helmholtz Zentrum München; Helmholtz-Zentrum für Umweltforschung; Hjärt-Lungfonden; Karolinska Institutet SFO Epidemiology and Region Stockholm; Leibniz Research-Institute for Environmental Medicine at the University of Düsseldorf; Ludwig-Maximilians-Universität München; Marien-Hospital Wesel; Mead Johnson Nutrition; Pediatric Practice, Bad Honnef; Swedish Research Council; Technische Universität München

Abstract

Background: Allergic diseases often develop jointly during early childhood but differ in timing of onset, remission, and progression. Their disease course over time is often difficult to predict and determinants are not well understood.

Objectives: We aimed to identify trajectories of allergic diseases up to adolescence and to investigate their association with early-life and genetic determinants and clinical characteristics.

Methods: Longitudinal k-means clustering was used to derive trajectories of allergic diseases (asthma, atopic dermatitis, and rhinitis) in two German birth cohorts (GINIplus/LISA). Associations with early-life determinants, polygenic risk scores, food and aeroallergen sensitization, and lung function were estimated by multinomial models. The results were replicated in the independent Swedish BAMSE cohort.

Results: Seven allergic disease trajectories were identified: "Intermittently allergic," "rhinitis," "early-resolving dermatitis," "mid-persisting dermatitis," "multimorbid," "persisting dermatitis plus rhinitis," and "early-transient asthma." Family history of allergies was more prevalent in all allergic disease trajectories compared the non-allergic controls with stronger effect sizes for clusters comprising more than one allergic disease (e.g., RRR = 5.0, 95% CI = [3.1–8.0] in the multimorbid versus 1.8 [1.4–2.4] in the mild intermittently allergic cluster). Specific polygenic risk scores for single allergic diseases were significantly associated with their relevant trajectories. The derived trajectories and their association with genetic effects and clinical characteristics showed similar results in BAMSE.

Conclusion: Seven robust allergic clusters were identified and showed associations with early life and genetic factors as well as clinical characteristics.

KEYWORDS
allergic diseases, epidemiology, longitudinal clustering, polygenic risk score, trajectories

1 | INTRODUCTION

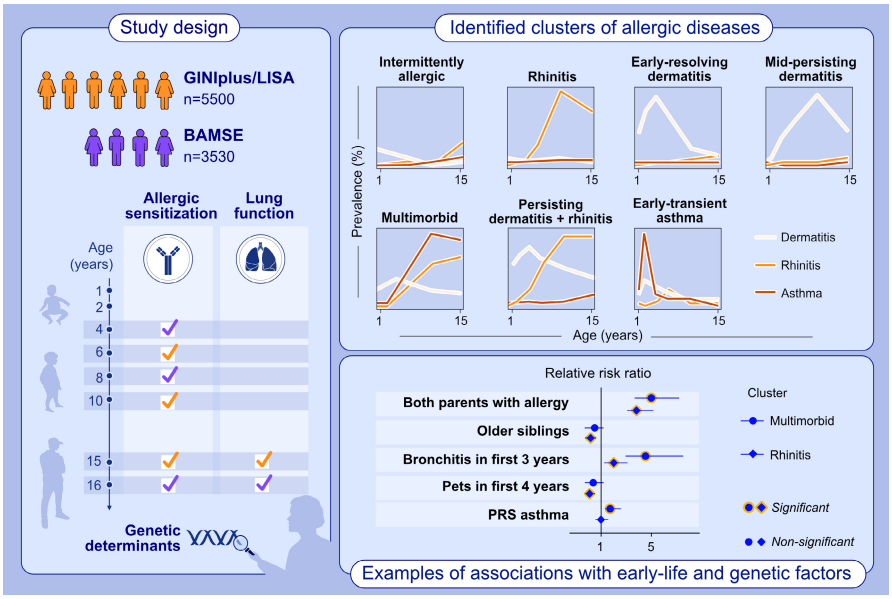
Allergic diseases, such as asthma, atopic dermatitis, and rhinitis, impose high impact on quality of life.[1] While atopic dermatitis often develops in early infancy with high remission rate up to adolescence, asthma, and rhinitis usually occur later in childhood.[2] In general, allergic diseases often develop jointly, in temporal succession and differing severities, highlighting the role of heritability, joint mechanisms, and genetic susceptibility.[2–5]

Prediction of onset, progression or remission of allergic diseases are often difficult to obtain. It is therefore of high importance, to characterize the patterns of joint disease development and their relating factors.

Based on this need, a pioneer investigation of clinical phenotypes taking age and time into account, reported classes of childhood wheezing in 1995.[6] With increasing sample sizes and long-lasting birth cohorts, researchers have now turned to data-driven approaches to identify more detailed developmental patterns of allergic diseases,[7] for example, building on the asthma trajectories in the Millennium Cohort Study up to seven years[8] and in BAMSE up to young adulthood.[9] The combination of the allergic disease trajectories, asthma, rhinitis, and dermatitis, together, was previously described until school-age[10–13] and from school age to adulthood.[14]

Genetic susceptibility to allergic diseases was investigated in several genome-wide association studies (GWAS).[5,15–18] Risk variants, single nucleotide polymorphisms (SNPs), identified in these analyses can be used to calculate individual predisposition. These aggregated polygenic risk scores (PRS) were applied in the context of various diseases,[19] including allergic ones[20] and their derived latent classes.[21] In addition, specific gene expression signatures have been identified for multimorbid asthma, rhinitis, and dermatitis, as compared with single-disease allergic phenotypes,[22] highlighting the current focus on genetic origins of allergic diseases.

However, given the rise of allergic disease prevalence in the recent decades,[23] the increase cannot be explained by genetic factors alone. Several environmental factors have been investigated in the



GRAPHICAL ABSTRACT

We identified seven allergic disease trajectories up to adolescence, which are corresponding to clinical observation in the German GINIplus and LISA studies and replicated the results in the Swedish BAMSE cohort. The clusters can be characterized using polygenic risk scores and early-life determinants, which support the hygiene hypothesis. The clusters also pose clinical implications for allergic sensitization, increasing with number of present allergic diseases, and lung function.

context of the hygiene hypothesis[24,25] and are commonly analyzed as determinants of allergic diseases, for example, in Hu et al.,[26] stating poor distinction of dermatitis phenotypes for example by breastfeeding or pet ownership alone.

Summarizing the need for a more specific distinction of phenotypes and their temporal patterns is apparent. Therefore, the aim of this study is to identify joint trajectories of allergic diseases using independent birth cohorts. Following objectives are to assess their association with (1) early-life determinants, (2) GWAS-derived PRS, and (3) clinical characteristics, such as allergic sensitization and spirometry.

2 | METHODS

2.1 | Study population

Data were obtained from two prospective, population-based German birth cohorts with a focus on the development of allergic diseases, the German Infant Study on the Influence of Nutrition Intervention plus Air pollution and Genetics on Allergy Development (GINIplus) and the Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany (LISA) study.For GINIplus, 5991 full-term, healthy newborns were recruited in Munich and Wesel between 1995 and 1998. The LISA study included 3094 healthy, full-term newborns from the study centers Munich, Wesel, Bad Honnef and Leipzig, born between 1997 and 1999. Ethics approval by the respective ethics committees, and written consent from all participating families was obtained. Both studies had comparable follow-up time points, which were harmonized to time points 1, 2, 4, 6, 10, and 15 years. More details on both studies can be found elsewhere.[27]

2.2 | Allergic diseases

Allergic diseases were defined as parent-reported doctor diagnosis of asthma, atopic dermatitis, hay fever or allergic rhinitis for each year of life, which were aggregated to cover the time periods between follow-ups (Table S1).

2.3 | Early-life determinants

We investigated early-life determinants, previously reported to be associated with allergic diseases. These include sex, parental education level, family history of allergic diseases, Caesarean section, exclusive breastfeeding, presence of older siblings, maternal smoking during pregnancy, second-hand tobacco smoke exposure, pet exposure, urbanicity at birth residency, and early bronchitis infections (Table S2).

2.4 | Aeroallergen and food sensitization

Specific immunoglobulin E (IgE) was measured at 6, 10, and 15 years using the CAP-RAST FEIA system (Pharmacia Diagnostics, Freiburg, Germany) according to the manufacturer's instructions. Sensitization

against common aero- (SX1 mix) and food allergens (FX5 mix) was defined with a cut-off for allergen specific IgE of >0.35kU/l in the screening test (Table S1).

2.5 | Spirometry

Lung function was measured using spirometry at the 15-year follow-up investigation according to guidelines from the American Thoracic Society and the European Respiratory Society[28] and has been described previously.[29] The present study investigates forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), and the Tiffeneau-Index (FEV_1/FVC), all standardized according to the Global Lung Initiatives (GLI) formula to control for non-linear effects of age, height, and sex.[30]

2.6 | Polygenic risk scores

Combining the joint effect of genetic variation, we calculated four different PRS based on current GWAS summary statistics for any allergic disease,[5,15] asthma,[17] dermatitis,[16] and rhinitis.[18] SNPs reported as significant in the GWAS meta-analyses were extracted from existing genome-wide data (Methods S1). After quality control, four PRS were calculated for each participant, weighting the allele dosage with the effect size reported in the GWAS. Finally included SNPs and annotated genes[31] can be found in Table S3a–d and a more detailed description of the calculation in Methods S2. Additionally, the two most common Filaggrin (FLG) loss-of-function mutations (R501X, 2282del4) were genotyped.[32] FLG mutation carrier status is defined as at least one mutation compared with no mutation.

2.7 | Replication study

The prospective, population-based BAMSE (Swedish abbreviation for Children, Allergy, Milieu, Stockholm, Epidemiology) project recruited 4093 newborns in the area of Stockholm between 1994 and 1996. Participants were followed-up at 1, 2, 4, 8, 12, and 16 years, and ethics approval was given by the Regional Ethics Board (EPN). Allergic disease diagnoses were defined based on parent-reported symptoms[33] according to the questionnaire used in the International Study of Asthma and Allergies in Childhood (ISAAC).[34] Further details on the study, including genetic data can be found elsewhere[35] and in the supplementary information (Methods S1 and S3 and Table S1).

The BAMSE cohort was used for replication of clusters, pre-defined early-life determinants, genetic variation, and clinical characteristics.

2.8 | Statistical analysis

All participants with available data on at least three out of six time-points per allergic disease were included in the analysis. Remaining missing observations were filled in carrying the last observation forward (LOCF).

To identify joint trajectories of allergic disease development, longitudinal k-means clustering using the kml3d package[36] in R[37] was employed, which enables to cluster longitudinal data with multiple trajectories. To ensure the focus on allergic development, a case-only approach including only those who reported any allergic disease for at least one time point was used. The non-allergic cluster, comprising all remaining participants, was added afterward. Differing numbers of clusters from two to nine clusters were tested and the optimal number of clusters was chosen based on the quality criterion of Davies-Bouldin[38] in GINIplus/LISA. The cluster number in BAMSE was fixed according to this for comparability and independent clustering results are provided in the supplement. We specified the algorithm to run 40 times with random starting conditions and used standard Euclidean distance. Sensitivity analyses for missingness were conducted using only complete observations as well as performing analyses based on bagging tree imputation of diagnoses in GINIplus/LISA to test the assumptions of LOCF (further described in methods S4) and latent class modeling.[9] Here, the optimal partitioning was based on the Bayesian information criterion and log-likelihood. As it is not possible to consider both dimensions, time and multiple allergic diseases, in this approach, the longitudinal clustering was the preferred approach. We also added symptom-based clustering for GINIplus/LISA to check for differences between the disease definitions within the used studies. K-means clusters are always presented for their optimal solution and pre-set for seven allergic disease clusters to ease comparison.

The total population and the allergic and non-allergic sub-populations were descriptively assessed regarding the distribution of early-life determinants and differences between the sub-populations were analyzed with a Chi-squared test.

Early-life determinants were assessed using post hoc Tukey-tests to identify differences between clusters and compact letter design was used to indicate similarities between the trajectories as well as a multiple-testing adjusted p-value[39] from the respective analysis of variance (ANOVA) (Table S4). Significant factors (FDR-adjusted ANOVA p-value <0.05) were selected for a joint multinomial regression model. Influences of genetic factors on clusters were assessed using a joint multinomial model followed by a fixed-effect meta-analysis between GINIplus/LISA and BAMSE using the meta package.[40] The association with allergic sensitization and spirometric z-scores was also assessed using multinomial regression models for each time point and measurement. All models were additionally adjusted for sex and further for study and study center in GINIplus/LISA. Relative Risk Ratios (RRR) and corresponding 95% confidence intervals are presented. Further sensitivity analyses were performed testing the robustness of genetic association for additional adjustment for (a) family history of allergic diseases, (b) early-life determinants, and (c) interaction effects with sex in GINIplus/LISA.

All analyses were conducted using R[37] version 4.0.3 (GINIplus/LISA) and version 4.0.4 (BAMSE) and code is available on request.

3 | RESULTS

Starting with descriptive assessments, the main analysis comprised 5550 participants of the GINIplus/LISA studies (Figure S1). Population characteristics, cumulative allergic disease prevalences, and investigated determinants are presented in Table 1 for GINIplus/LISA and Table S5 for pre-defined BAMSE determinants. Allergic disease development in GINIplus/LISA shows early dermatitis as most common allergic disease in infancy with decreasing prevalence after four years and increasing rhinitis and asthma prevalence during childhood and adolescence (Figure S2).

3.1 | Joint trajectories of allergic diseases

Within GINIplus/LISA, seven allergic disease cluster plus one non-allergic cluster were identified (Figure 1A) with a comparable cluster allocation observed in BAMSE (Figure 1B): Intermittently allergic (GINIplus/LISA:17.5%, BAMSE:15.5%), rhinitis (7.5%, 5.3%), early-resolving dermatitis (6.3%, 11.5%), mid-persisting dermatitis (4.1%, 12.0%), multimorbid (4.0%, 8.2%), persisting dermatitis plus rhinitis (2.2%, 6.6%), early-transient asthma (0.5%, 10.2%), and non-allergic (57.8%, 30.6%). Names for the clusters were chosen descriptively and do not impose strict clinical definitions. The independent clustering in BAMSE yielded six allergic disease clusters (Figure S3) and only using complete observations in GINIplus/LISA led to nine optimal allergic disease clusters (Figure S4a; see Figure S4b for pre-defined 7 + 1 cluster solution in complete observations only).

The sensitivity analysis using latent class modeling, yielded eight optimal classes, including a non-allergic class in GINIplus/LISA and seven in BAMSE (Figure S5). The identified classes are highly similar and differ only in a split of the multimorbid cluster. Here, we observe, instead of only one cluster featuring all three allergic diseases now one with low level dermatitis and one with high level dermatitis, supporting the original differentiation. GINIplus/LISA looses the early-transient wheeze cluster and BAMSE the intermittent and early-resolving dermatitis ones.

3.2 | Early-life determinants

To assess which factors potentially influence disease development, Table 2 presents results of the multinomial model containing early-life determinants, which were identified using Tukey-Tests (Table S4) in GINIplus/LISA. Male sex is associated with the clusters comprising a high prevalence of rhinitis and inversely with the persisting dermatitis cluster. Parental history of allergies increases the risk for all identified clusters compared with the non-allergic cluster, with intermittently allergic showing the weakest association, significantly differing from the clusters comprising rhinitis and early-transient asthmatics. Pet exposure and presence of older siblings are inversely associated with the rhinitis cluster in GINIplus/LISA. In BAMSE associations with family history show similar patterns, although not all effects reach significance and are mostly visible for only one parent with allergies. Nonetheless, an even stronger inverse relationship can be seen for early bronchitis infections.

3.3 | Genetic factors

Further, we investigated genetic influences on allergic disease trajectories to identify the effects of genetic predisposition on disease development. For this, we used a fixed-effect meta-analysis of the calculated PRS and FLG mutation carrier status on the derived clusters (Table 3; Table S6 for individual studies). In GINIplus/LISA all specific PRS are significantly associated with their respective trajectories and therefore more specific than the PRS for any allergic disease, which is only associated with the rhinitis cluster. Further, the PRS for atopic dermatitis shows a significant association with early-onset or transient but not mid-onset or persisting atopic dermatitis.

Within BAMSE, the PRS for any allergic disease shows significant associations with the intermittently allergic, early dermatitis, multimorbid, and dermatitis plus rhinitis cluster, whereas the specific PRS are less clear associated.

Comparing the independent samples, we observe similar emerging patterns, even though clusters differ in their specifics. Despite differences in significance, effect directions are mostly similar between GINIplus/LISA and BAMSE and the meta-analysis shows significant associations between all allergic clusters and the PRS for any allergic disease, except the mid-persisting dermatitis and early-transient asthma cluster. The multimorbid cluster is specified by additional significant associations with the PRS for asthma and rhinitis, differentiating it from the rhinitis cluster. The FLG-mutation is significantly associated with the multimorbid and persisting dermatitis plus rhinitis cluster.

Our sensitivity analyses showed that genetic associations seen between PRS and clusters are rather robust to adjustment for family history of allergic diseases and early-life determinants (Tables S7 and S8). Investigating genetic effects and their modification by sex yielded no significant interaction (Table S9).

3.4 | Aeroallergen and food sensitization

For the clinical characterization of trajectories, we investigated associations of aero- and food allergen sensitization measured at several time points with the allergic disease trajectories, as shown in Table 4. All clusters, except the early-transient asthma and the mid-persisting dermatitis clusters, show significant associations compared to the non-allergic control in all cohorts. However, the magnitude of effects is distinctly higher in rhinitis containing trajectories. Especially the clusters with more than one allergy are robustly associated and are significantly higher compared with all

TABLE 1 Population characteristics and determinants in GINIplus/LISA

	Total ($N = 5550$)	Allergic disease cases ($N = 2342$)	No allergic diseases ($N = 3208$)	*p*-value (Chi-squared)
Sex				
Male	50.8% (2818)[0]	52.7% (1235)[0]	49.3% (1583)[0]	0.014
Study				
GINI intervention	24.9% (1381)[0]	31.4% (735)[0]	20.1% (646)[0]	<0.001
GINI control	35.5% (1971)[0]	32.4% (759)[0]	37.8% (1212)[0]	<0.001
LISA	39.6% (2198)[0]	36.2% (848)[0]	42.1% (1350)[0]	<0.001
Study center				
Munich	51.5% (2856)[0]	52.9% (1238)[0]	50.4% (1618)[0]	0.079
Leipzig	10.2% (564)[0]	10.5% (247)[0]	9.9% (317)[0]	0.444
Bad Honnef	4.4% (244)[0]	3.8% (90)[0]	4.8% (154)[0]	0.099
Wesel	34.0% (1886)[0]	32.7% (767)[0]	34.9% (1119)[0]	0.104
Cumulative prevalence of allergic disease up to 15 years				
Any allergic disease	42.2% (2342)[0]	100.0% (2342)[0]	0.0% (0)[0]	
Asthma	9.2% (513)[0]	21.9% (513)[0]	0.0% (0)[0]	
Dermatitis	27.9% (1549)[0]	66.1% (1549)[0]	0.0% (0)[0]	
Rhinitis	21.3% (1183)[0]	50.5% (1183)[0]	0.0% (0)[0]	
Family history of allergic diseases				
No parent	43.0% (2272)[263]	33.7% (758)[94]	49.8% (1514)[169]	<0.001
One parent	40.2% (2126)[263]	43.1% (970)[94]	38.0% (1156)[169]	<0.001
Both parents	16.8% (889)[263]	23.1% (520)[94]	12.1% (369)[169]	<0.001
Parental education				
Low	6.6% (366)[23]	6.6% (154)[10]	6.6% (212)[13]	1.000
Medium	27.8% (1536)[23]	27.3% (637)[10]	28.1% (899)[13]	0.520
High	65.6% (3625)[23]	66.1% (1541)[10]	65.2% (2084)[13]	0.528
Caesarean section				
Yes	19.8% (1037)[325]	20.8% (466)[99]	19.1% (571)[226]	0.154
Breastfeeding during first 4 months				
Yes	62.6% (3417)[92]	61.5% (1417) [39]	63.4% (2000)[53]	0.169
Presence of older siblings				
Yes	46.6% (2582)[11]	43.5% (1015)[7]	48.9% (1567)[4]	<0.001
Maternal smoking during 2nd or 3rd trimester of pregnancy				
Yes	9.6% (521)[141]	8.8% (202)[46]	10.2% (319)[95]	0.082
Environmental tobacco smoke exposure in first 4 years				
Yes	37.4% (2012)[172]	36.3% (825)[72]	38.2% (1187)[100]	0.176
Bronchitis infection in first 3 years				
Yes	39.3% (2111)[180]	44.5% (1007)[77]	35.6% (1104)[103]	<0.001
Cat or dog in first 4 years				
Yes	27.1% (1253)[924]	24.3% (480)[368]	29.1% (773)[556]	<0.001
Urbanicity at birth				
City	47.0% (2472)[285]	47.2% (1054)[107]	46.8% (1418)[178]	0.817
Town or suburb	40.6% (2139)[285]	41.4% (925)[107]	40.1% (1214)[178]	0.349
Rural area	12.4% (654)[285]	11.5% (256)[107]	13.1% (398)[178]	0.074

Note: Presented are prevalence, total case numbers and number of missing values for non-allergic participants, allergic participants, and the total sample.

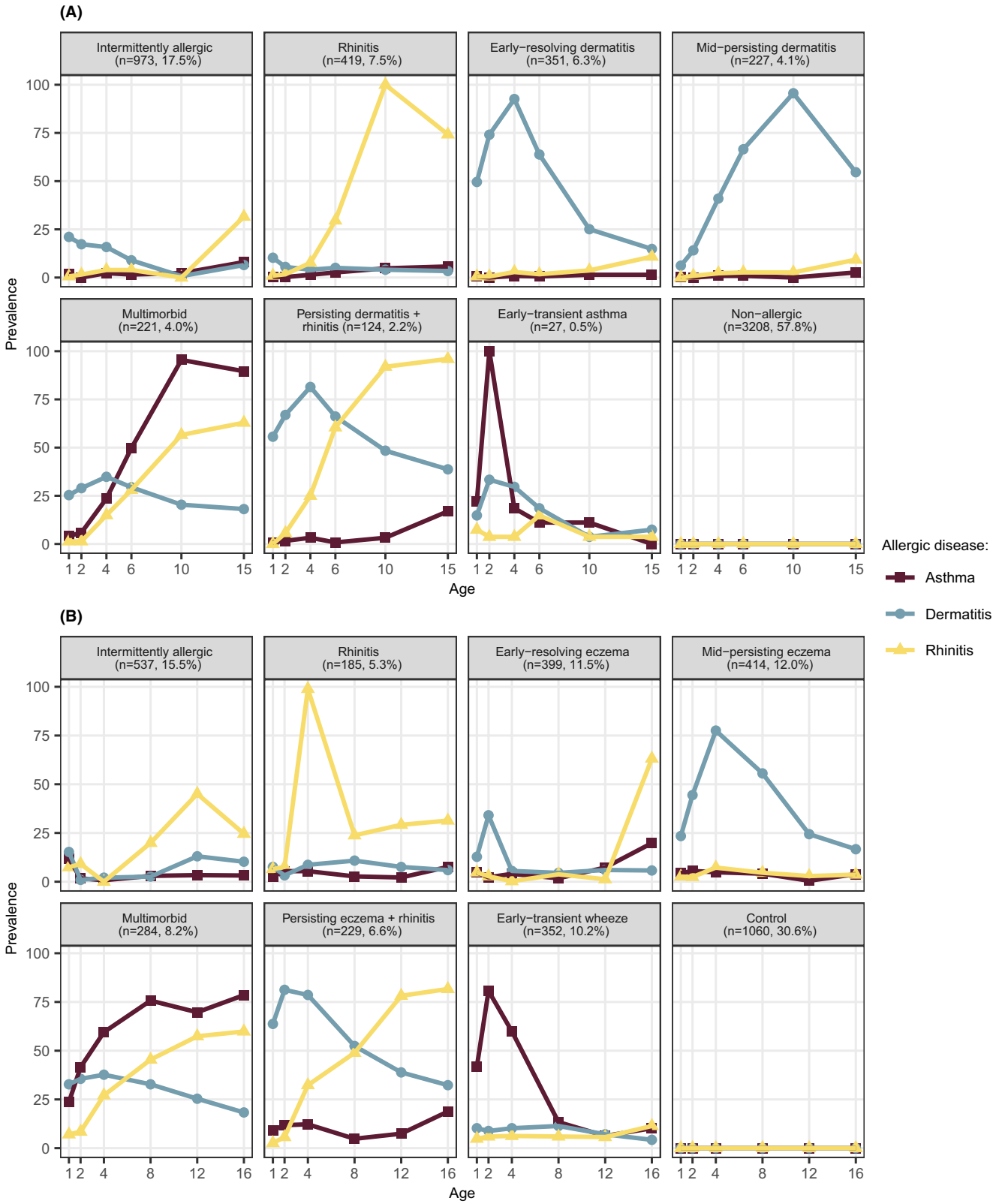


FIGURE 1 Clusters of allergic diseases in (A) GINIplus/LISA and (B) BAMSE. The allergic disease clusters were derived using longitudinal k-means for allergic patients, the non-allergic cluster was added afterward. The best number of clusters was defined in GINIplus/LISA following the quality criterion of Davies-Bouldin.

TABLE 2 Results from multinomial regression of clusters on early-life determinants including all early-life factors, which showed an association in the univariate analysis, at once

	Intermittently allergic	Rhinitis	Early-resolving dermatitis	Mid-persisting dermatitis	Multimorbid	Persisting dermatitis + rhinitis	Early-transient asthma
GINIplus & LISA							
Male sex	1.1 [0.9–1.3]	**1.3* [1–1.6]**	1.0 [0.8–1.3]	**0.6* [0.4–0.8]**	**1.8* [1.3–2.5]**	**1.8* [1.2–2.8]**	2.0 [0.8–4.7]
One parent with allergies	**1.3* [1–1.6]**	**2.3* [1.7–3.1]**	**1.9* [1.3–2.6]**	1.1 [0.7–1.7]	**1.7* [1.1–2.6]**	**3.5* [1.9–6.7]**	**3.7* [1.1–12.1]**
Both parents with allergies	**1.8* [1.4–2.4]**	**3.6* [2.5–5.1]**	**3.4* [2.3–5.1]**	**2.8* [1.7–4.5]**	**5.0* [3.1–8]**	**7.9* [3.9–15.9]**	**15.9* [4.5–56.3]**
Older siblings	**0.8* [0.7–0.9]**	**0.7* [0.5–0.9]**	1.0 [0.8–1.3]	1.1 [0.8–1.5]	0.8 [0.6–1.1]	0.8 [0.5–1.3]	2.7 [1–7.6]
Bronchitis infection in first 3 years	**1.6* [1.3–1.9]**	**1.6* [1.2–2]**	1.2 [0.9–1.6]	1.1 [0.8–1.6]	**4.8* [3.4–6.8]**	**2.0* [1.3–3.1]**	**6.0* [2.4–15.0]**
Cat or dog in first 4 years	1.0 [0.8–1.2]	**0.6* [0.5–0.8]**	0.8 [0.6–1.0]	1.2 [0.8–1.7]	0.7 [0.5–1.1]	0.6 [0.4–1.1]	1.8 [0.7–4.2]
BAMSE							
Male sex	1.2 [1.0–1.5]	1.2 [0.9–1.7]	1.0 [0.8–1.2]	**0.8* [0.6–0.9]**	**1.4* [1.1–1.9]**	1.3 [1.0–1.7]	**1.6* [1.2–2.0]**
One parent with allergies	**1.6* [1.2–2.2]**	1.2 [0.7–1.9]	**1.5* [1.1–2.0]**	**1.7* [1.2–2.3]**	**4.6* [3.4–6.2]**	**2.7* [1.9–3.8]**	**2.7* [2.0–3.6]**
Both parents with allergies	1.2 [0.3–4.5]	2.3 [0.5–11.1]	2.0 [0.6–7.0]	1.0 [0.2–4.8]	**9.8* [3.7–26.5]**	2.1 [0.4–10.0]	**5.3* [1.9–14.9]**
Older siblings	0.9 [0.7–1.1]	0.7 [0.5–1.0]	0.8 [0.7–1.1]	0.9 [0.7–1.1]	1.1 [0.8–1.4]	1.0 [0.7–1.3]	1.3 [1.0–1.6]
Bronchitis infection in first 3 years	**1.5* [1.0–2.3]**	1.7 [0.9–3.1]	1.4 [0.9–2.2]	1.5 [0.9–2.3]	**2.5* [1.6–3.9]**	1.3 [0.7–2.3]	**3.9* [2.7–5.7]**
Cat or dog in first 4 years	**0.7* [0.5–0.8]**	**0.6* [0.4–0.9]**	0.7* [0.6–0.9]	0.8 [0.7–1.1]	**0.5* [0.4–0.8]**	**0.6* [0.4–0.9]**	0.9 [0.7–1.2]

Note: Models in GINIplus/LISA are additionally adjusted for study and study center. Values show RRR and corresponding 95% confidence intervals in comparison to the control cluster. Significance is indicated by bold font and asterisks.

TABLE 3 Results from fixed-effect meta-analysis on multinomial regression of clusters on PRS and FLG mutation status, controlled for sex, study and study center in GINIplus/LISA and sex only in BAMSE

Meta-analysis (fixed effect)	Intermittently allergic	Rhinitis	Early-resolving dermatitis	Mid-persisting dermatitis	Multimorbid	Persisting dermatitis + rhinitis	Early-transient asthma
PRS any allergic disease	**1.1* [1.0–1.3]**	**1.3* [1.1–1.6]**	**1.2* [1.1–1.4]**	1.1 [0.9–1.3]	**1.3* [1.1–1.6]**	**1.6* [1.3–1.9]**	1.1 [0.9–1.3]
PRS asthma	1.1 [0.9–1.2]	1.0 [0.9–1.2]	**1.2* [1.0–1.3]**	1.0 [0.9–1.2]	**1.2* [1.0–1.4]**	1.0 [0.8–1.2]	1.1 [0.9–1.3]
PRS dermatitis	1.1 [1.0–1.2]	1.1 [0.9–1.2]	1.1 [0.9–1.2]	1.1 [1.0–1.3]	1.0 [0.9–1.2]	1.1 [1.0–1.4]	1.0 [0.9–1.2]
PRS rhinitis	1.1 [0.9–1.2]	1.1 [1.0–1.3]	1.0 [0.9–1.2]	1.0 [0.9–1.2]	**1.4* [1.2–1.6]**	1.1 [0.9–1.4]	**1.2* [1.0–1.5]**
Any FLG-mutation	0.9 [0.6–1.3]	0.8 [0.5–1.6]	0.9 [0.5–1.5]	1.6 [1.0–2.5]	**1.7* [1.0–2.8]**	**2.0* [1.2–3.4]**	1.4 [0.8–2.7]

Note: Table shows RRR and corresponding 95% confidence intervals. Significance is indicated by bold script and asterisks. A RRR higher than 1 indicates a positive association between increasing PRS and likelihood to be allocated to this trajectory. Tables for single cohorts are presented in Table S6.

other clusters in BAMSE. Additionally, we see significant association of sensitization already before the peak of rhinitis prevalence. In food sensitization, this might enable differentiation between the early-resolving and mid-persisting dermatitis cases.

3.5 | Spirometry

To assess the impact of trajectories on lung function in adolescence and thus long-term impairment of spirometric function, Table 5 displays the association of spirometry measures z-scores at age 15 years with trajectory allocation. Significant results are observed for FEV_1 and the Tiffeneau-Index in the multimorbid (all cohorts) and early-transient asthma clusters (only BAMSE). Thus, reduced spirometric z-scores can be seen in the clusters comprising a high percentage of asthmatic participants. Borderline significant associations are seen for the intermittently allergic and early-resolving dermatitis cluster.

4 | DISCUSSION

In the present study, we identified seven allergic disease trajectories, using an unsupervised, longitudinal clustering approach in population-based birth cohorts, and one non-allergic control group. The role of early-life determinants was investigated and significant differences in relation to allergic sensitization and spirometric indices between the trajectories were observed. Furthermore, the specific GWAS-derived PRS showed significant associations with their disease-associated trajectories. Successful replication of trajectories, their association with genetic factors, lung function, and sensitization was achieved in the independent Swedish BAMSE cohort, and a comparable classification was obtained when using latent class modeling, further strengthening the robustness of our findings.

Comparing our results with the available evidence, our clusters could confirm results from two of the papers previously characterizing allergic disease trajectories identifying five clusters of allergic diseases up to four years[10] and six up to nine years in the PARIS birth cohort[13] as well as eight classes up to eleven years in UK cohorts.[12] We extend the covered time period up to adolescence, showing the continued developments indicated by the classification found in UK cohorts[12] and being able to distinguish two different dermatitis trajectories, not reported in the previous clustering approaches. Furthermore, the PRS based on one GWAS on any allergic disease[5] was significantly associated with all identified classes in the UK cohorts.[21] Our application of PRS for all single disease entities allows further distinction of sub-phenotypes.

Another publications tracking allergic disease development from nine to 34 years of age identified six classes of allergic diseases ("No symptoms," "Rhinoconjunctivits only," "Late-onset wheeze," "Rhinoconjunctivitis+Wheeze," and "Eczema+Rhinoconjunctivitis+Wheeze") using latent classes.[14] Identified clusters are comparable. The two identified multimorbid clusters resemble our multimorbid and persisting dermatitis plus

TABLE 4 Results from single multinomial regression of clusters on aeroallergen sensitization (SX1 mix) and food allergen sensitization (FX5 mix) at different time points, controlled for sex, study, and study center in GINIplus/LISA and BAMSE

	Intermittently allergic	Rhinitis	Early-resolving dermatitis	Mid-persisting dermatitis	Multimorbid	Persisting dermatitis + rhinitis	Early-transient asthma
Aeroallergen sensitization							
GINIplus & LISA							
Aeroallergen sensitization at 6 years	**2.2* [1.8–2.8]**	**23.0* [16.6–31.8]**	**4.4* [3.2–6.2]**	1.6 [1–2.5]	**18.9* [12.7–28.1]**	**55.3* [27.2–112.4]**	0.7 [0.1–5.1]
Aeroallergen sensitization at 10 years	**3.0* [2.4–3.7]**	**23.0* [15.7–33.9]**	**2.4* [1.7–3.4]**	**2.1* [1.4–3.1]**	**11.9* [7.8–18.1]**	**40.5* [17.4–94.4]**	2.1 [0.5–9.2]
Aeroallergen sensitization at 15 years	**3.1* [2.5–3.8]**	**13.2* [9–19.2]**	**2.8* [2–3.8]**	1.4 [0.9–2.1]	**12.2* [7.5–19.7]**	**22.7* [9.7–52.9]**	1.5 [0.5–4.8]
BAMSE							
Aeroallergen sensitization at 4 years	**4.1* [2.5–6.7]**	**4.3* [2.3–8.0]**	**3.5* [2.0–5.9]**	**2.6* [1.5–4.6]**	**28.7* [17.9–46.1]**	**27.1* [16.5–44.4]**	**2.2* [1.2–4.0]**
Aeroallergen sensitization at 8 years	**5.8* [4.0–8.3]**	**4.2* [2.5–6.9]**	**4.3* [2.9–6.4]**	**2.1* [1.3–3.3]**	**27.0* [18.0–40.6]**	**32.1* [20.8–49.7]**	**1.9* [1.2–3.1]**
Aeroallergen sensitization at 16 years	**3.1* [2.4–4.1]**	**2.2* [1.5–3.2]**	**3.6* [2.7–4.7]**	**1.8* [1.3–2.4]**	**13.5* [9.2–19.9]**	**14.1* [9.3–21.4]**	**1.7* [1.2–2.3]**
Food allergen sensitization							
GINIplus & LISA							
Food allergen sensitization at 6 years	**2.1* [1.5–2.8]**	**4.0* [2.8–5.6]**	**2.5* [1.6–3.9]**	1.0 [0.5–2.0]	**4.9* [3.2–7.4]**	**6.2* [3.8–10.2]**	1.6 [0.2–12.8]
Food allergen sensitization at 10 years	**1.8* [1.4–2.3]**	**5.0* [3.7–6.7]**	**1.8* [1.2–2.7]**	1.3 [0.8–2.2]	**4.2* [2.9–6.2]**	**5.9* [3.7–9.5]**	2.5 [0.5–12.8]
Food allergen sensitization at 15 years	**2.0* [1.5–2.8]**	**4.0* [2.7–5.8]**	**2.4* [1.5–3.8]**	0.9 [0.4–2.0]	**5.1* [3.2–7.9]**	**6.2* [3.6–10.8]**	2.8 [0.6–13.2]
BAMSE							
Food allergen sensitization at 4 years	1.2 [0.8–1.8]	1.2 [0.7–2.1]	1.3 [0.9–2.0]	1.2 [0.8–1.8]	**5.6* [3.9–8.0]**	**3.9* [2.6–5.8]**	0.9 [0.5–1.4]
Food allergen sensitization at 8 years	**1.8* [1.3–2.6]**	**1.9* [1.2–3.1]**	**1.6* [1.1–2.4]**	1.1 [0.8–1.7]	**6.4* [4.5–9.2]**	**6.0* [4.1–8.8]**	1.3 [0.9–2.0]
Food allergen sensitization at 16 years	**1.6* [1.0–2.4]**	1.3 [0.7–2.5]	**1.7* [1.1–2.6]**	1.0 [0.6–1.7]	**7.3* [4.9–10.8]**	**7.0* [4.6–10.6]**	**1.7* [1.0–2.7]**

Note: Table shows RRR and corresponding 95% confidence intervals. Significance is indicated by bold script and asterisks. A RRR lower than 1 indicates a lower sensitization value in this trajectory.

TABLE 5 Results from single multinomial regression of clusters on spirometry measurements, controlled for study and study center in GINIplus/LISA and BAMSE

	Intermittently allergic	Rhinitis	Early-resolving dermatitis	Mid-persisting dermatitis	Multimorbid	Persisting dermatitis + rhinitis	Early-transient asthma
GINIplus & LISA							
FEV1 (GLI)	**0.9* [0.8–1.0]**	0.9 [0.8–1.1]	1.1 [1–1.4]	1.0 [0.8–1.3]	**0.7* [0.6–0.8]**	0.9 [0.7–1.2]	0.5 [0.3–1.0]
FVC (GLI)	0.9 [0.8–1.0]	0.9 [0.8–1.1]	1.1 [0.9–1.3]	0.9 [0.7–1.1]	1.0 [0.8–1.2]	0.9 [0.7–1.2]	0.6 [0.3–1.1]
Tiffeneau-Index (GLI)	0.9 [0.9–1.0]	1.0 [0.8–1.1]	1.1 [0.9–1.3]	1.2 [1.0–1.4]	**0.6* [0.5–0.7]**	1.0 [0.8–1.3]	0.9 [0.5–1.5]
BAMSE							
FEV1 (GLI)	0.9 [0.8–1.1]	1.0 [0.8–1.3]	0.8* [0.7–1.0]	0.9 [0.8–1.1]	**0.6* [0.5–0.7]**	0.9 [0.8–1.1]	**0.7* [0.6–0.8]**
FVC (GLI)	1.0 [0.8–1.1]	1.0 [0.8–1.2]	0.9 [0.7–1.0]	1.0 [0.8–1.1]	0.9 [0.7–1.0]	0.9 [0.7–1.1]	0.9 [0.8–1.1]
Tiffeneau-Index (GLI)	0.9 [0.8–1.1]	1.0 [0.8–1.3]	0.9 [0.8–1.1]	1.0 [0.8–1.1]	**0.6* [0.5–0.7]**	1.1 [0.9–1.3]	**0.6* [0.5–0.8]**

Note: Table shows RRR and 95% corresponding confidence intervals. Significance is indicated by bold script and asterisks. A RRR lower than 1 indicates a lower lung function value in this trajectory.

rhinitis clusters, with lower prevalence of dermatitis and higher prevalence of wheeze, respectively. They further found that all allergic trajectories were already established by adolescence, not changing much after the age of 18 supporting our approach to include participants from infancy to adolescence in our longitudinal clustering, still ensuring comparability to previously reported approaches covering early life windows.

Within the first intermittently allergic cluster it is likely that the allocated participants have mild or transient forms of allergic diseases, as they all reported an allergic disease for at least one time point but prevalences were generally low. Nonetheless, they have a significant risk associated with family history and parents might be more attentive for symptoms of dermatitis and rhinitis.

The rhinitis cluster was inversely associated with environmental factors such as pet exposure and presence of older siblings, which have been discussed in the context of the hygiene hypothesis.[25] This cluster was further significantly associated with the PRS for any allergic diseases and rhinitis, underlining the genetic contribution to this allergic phenotype. It further indicates that within the SNPs identified to be associated with any allergic disease,[5,15] rhinitis is the most common allergic disease and GWAS are commonly not adjusted for comorbid allergic diseases.

As previously reported, early-onset dermatitis is, compared to late-onset dermatitis, associated with sensitization against aeroallergens and higher genetic susceptibility.[41] This is also seen in GINIplus/LISA, with the early dermatitis cluster being significantly associated with the respective PRS and higher aeroallergen sensitization rates at six years. In BAMSE, instead the mid-persisting trajectory is significantly associated with the PRS for dermatitis, but the trajectories prevalence also peaks earlier, potentially blurring the differentiation between the two trajectories, with the early-transient cluster being detected rather in the independent BAMSE clustering solution with six allergic disease trajectories only (Figure S3). Another aspect to consider is the structurally higher percentage of missing diagnoses in the mid-persisting trajectory (Figure S7a), which might lead to falsely carried-forward dermatitis cases in GINIplus/LISA, while non-missing cases might be still following a persisting disease course but on a lower level than assumed here. Further differentiation between the dermatitis clusters is possible in regard to food sensitization, which was observed in the early-resolving but not the mid-persisting cluster, as supported by the literature,[42] which might indicate earlier remission in those children.

The multimorbid cluster is significantly characterized by a higher percentage of male participants, following the previous results reporting an increased risk for allergic multimorbidity in males with family history of allergies.[43] There are also significant associations of the multimorbid cluster with the PRS for asthma and rhinitis in GINIplus/LISA, as well as any allergy and rhinitis in BAMSE.

Both, the multimorbid and rhinitis clusters, show similar rhinitis prevalences and a high sensitization rate. However, one group of participants develops additional asthma, leading to a higher burden of disease and worse lung function. One possible explanation might be higher genetic susceptibility as indicated by the effect of the

asthma PRS; another one could be the higher percentage of bronchitis infections in infancy, negatively affecting spirometry measures[29] as also seen in our results. This development from early bronchitis to asthma and reduced spirometry differentiates the multimorbid from the persisting dermatitis plus rhinitis cluster, where asthma prevalence remains low.

Other early-life factors, such as maternal smoking during pregnancy or second-hand tobacco smoke exposure during early childhood, both risk factors for reduced lung function, did not differ between these clusters. We also did not see significant differences for further common determinants, such as Caesarean section or urbanicity, potentially due to low variance in our sample or low case numbers.

The persisting dermatitis plus rhinitis cluster was previously described in UK studies as "persistent eczema with later-onset rhinitis."[12] It seems to be very robust, as it was found within five different birth cohorts (ALSPAC & MAAS[12]; GINIplus, LISA, and BAMSE). It has the strongest associations with aeroallergen sensitization, nearly double compared with the multimorbid cluster in GINIplus/LISA, but none of the PRS is significantly associated with the trajectory in GINIplus/LISA and only the PRS for any allergic disease in BAMSE.

Nonetheless, FLG mutation, known to be the strongest genetic risk factor for dermatitis,[44] was significantly associated with the persisting dermatitis plus rhinitis cluster in BAMSE. Higher effect sizes, although not significant, can also be seen for the other clusters comprising relevant proportions of dermatitis patients (early-resolving (only GINIplus/LISA), mid-onset and multimorbid). However, the association with FLG seems higher in persisting forms of dermatitis, in comparison with the early-onset resolving cluster.

The last cluster is weakly powered in GINIplus/LISA with only 27 participants, all of which reported an asthma diagnosis in infancy but with a prevalence below 20% later in life. This early-transient asthma sub-phenotype is often identified in clustering approaches as "early-transient wheeze"[9,10] and much more common in English-speaking countries where a higher prevalence of asthma has been reported.[43] This was also observed during our replication in the BAMSE cohort using a symptom-based phenotype definition. Mostly missing sensitization indicates that this cluster does not depict an allergic disease trajectory but rather an early respiratory infection trajectory with potentially ongoing impairment[6] as underlined by the significant association with early bronchitis infections. Previous studies report on the connection between wheezing inducing airway infections and future asthma development[45,46] and although these participants here do not develop further asthma until 15 years of age, they present reduced lung function outcomes in adolescence. The presence of early-asthma or wheezing symptoms is much higher in our symptom-based clustering for GINIplus/LISA (Figure S8) but these cases may just be early respiratory infections instead of asthma, which is difficult to diagnose in infancy.

Similar results in lung function reduction as seen for our multimorbid and early-transient asthma clusters were also reported for cohorts covering nine to 34 and seven to 53 years of age, respectively.[14,47] Forster et al. reported lower Tiffeneau-indices in young adulthood for classes involving wheeze[14] and Bui et al. showing that early-onset persisting asthma trajectories had stronger associations with lower lung function at 53 years than later appearing asthma cases.[47]

One limitation is the difference in allergic disease definitions based on clinical practices in Germany and Sweden. Although all cohorts employ similar time points of follow-up, GINIplus/LISA analyzed parent-reported doctor diagnoses, while in BAMSE, the analysis was based on parent-reported symptoms and prescribed asthma medication use. Remarkably, we were still able to derive the same cluster solution, showing the robustness of the results, further supported by our sensitivity analysis clustering symptom-diagnoses in GINIplus/LISA (Figure S8). Nevertheless, all included studies are observational cohort studies and cannot investigate causality and are impacted by missing validated food allergy assessments, non-measured sensitization in infancy and missing indicators of severity or quality of life. Also, the issue with missing diagnoses due to loss to follow-up needs to be considered in the context of prospective birth cohorts, which collect valuable data but loose participants with each further time point, leading to higher number of missing values at the later time points of this study. Possible solutions for addressing this are different imputation methods but here no clear standard has been established yet. Allergic diseases are chronic but still every imputation might lead to false assumptions or might omit especially transient or late-onset trajectories. Of note, percentage of missing diagnoses further differs not only related to the time points but also between the clusters, especially the mid-persisting dermatitis. Potential reasons include a higher loss-to-follow-up in non-allergic parents (compared to those with positive family history), which are also overrepresented in the mid-persisting dermatitis cluster (Table S4). Furthermore, the early-transient wheeze cluster displays a higher proportion of missing diagnoses at earlier time-points. However, the generally low sample size in this cluster (n = 27) and thus missingness might simply occur by chance in this cluster.

Further, our PRS were calculated purely on genome-wide significant SNP sets and not as proposed by newer publications on complete GWAS results, not restricted by a significance threshold.[48] Nevertheless, in this paper, we aimed to show associations of disease-specific scores, which we assume to be represented by the respective large-scale GWAS results. This approach also enables direct comparison to other publications using a similar PRS for any allergic disease.[21]

The strengths of our study include the long follow-up period, covering both childhood and adolescence and the ability to cover many discussed early-life determinants. The successful replication in an independent cohort further underlines robustness of results. Furthermore, this study demonstrates the association of disease-specific PRS with allergic disease trajectories.

Using the results from this paper, we want to further drive clinical prediction of allergic trajectories, enabling pediatricians to forecast future allergic developments and initiate prevention strategies. Early identification of patients at risk, might help them to mitigate

further risk factors or concretely prevent asthma exacerbations with life-threatening potential.

In conclusion, we aimed to classify allergic disease development to add knowledge about the characteristics and determinants of the derived seven allergic disease trajectories from birth to adolescence. The derived trajectories allow a clearer classification of common allergic disease courses in times of increasing prevalence and burden of disease, which might be further facilitated to improve prediction in the future.

AUTHOR CONTRIBUTIONS

Anna Kilanowski involved in formal analysis, methodology, visualization, writing—original draft preparation, and writing—review and editing. Elisabeth Thiering contributed to conceptualization, methodology, supervision, and writing—review and editing. Gang Wang involved in formal analysis, validation, writing—review and editing. Ashish Kumar and Sara Kress contributed to validation and writing—review and editing. Claudia Flexeder performed writing—review and editing. Carl-Peter Bauer and Dietrich Berdel involved in investigation, resources, and writing—review and editing. Andrea von Berg contributed to funding acquisition, investigation, resources, and writing—review and editing. Anna Bergström performed investigation, resources, and writing—review and editing. Monika Gappa contributed to investigation and writing—review and editing. Joachim Heinrich performed data curation, funding acquisition, investigation, resources, and writing—review and editing. Gunda Herberth contributed to investigation, resources, writing—review and editing. Sibylle Koletzko involved in funding acquisition, investigation, resources, and writing—review and editing. Inger Kull involved in investigation, resources, and writing—review and editing. Erik Melén involved in data curation, funding acquisition, investigation, resources, and writing—review and editing. Tamara Schikowski performed data curation, investigation, resources, and writing—review and editing. Annette Peters involved in supervision and writing—review and editing. Marie Standl involved in conceptualization, data curation, methodology, project administration, resources, supervision, and writing—review and editing.

ACKNOWLEDGMENTS

GINIplus: The authors thank all the families for their participation in the GINIplus study. Furthermore, we thank all the members of the GINIplus Study Group for their excellent work. The GINIplus Study group consists of the following: Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg (Heinrich J, Brüske I, Schulz H, Flexeder C, Zeller C, Standl M, Schnappinger M, Ferland M, Thiering E, Tiesler C); Department of Pediatrics, Marien-Hospital, Wesel (Berdel D, von Berg A); Ludwig-Maximilians-University of Munich, Dr von Hauner Children's Hospital (Koletzko S); Child and Adolescent Medicine, University Hospital rechts der Isar of the Technical University Munich (Bauer CP, Hoffmann U); IUF-Environmental Health Research Institute, Düsseldorf (Schikowski T, Link E, Klümper C, Krämer U, Sugiri D). LISA: The authors thank all the families for their participation in the LISA study. Furthermore, we thank all members of the LISA Study Group for their excellent work. The LISA Study group consists of the following: Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology, Munich (Heinrich J, Schnappinger M, Brüske I, Ferland M, Schulz H, Zeller C, Standl M, Thiering E, Tiesler C, Flexeder C); Department of Pediatrics, Municipal Hospital "St. Georg," Leipzig (Borte M, Diez U, Dorn C, Braun E); Marien Hospital Wesel, Department of Pediatrics, Wesel (von Berg A, Berdel D, Stiers G, Maas B); Pediatric Practice, Bad Honnef (Schaaf B); Helmholtz Centre of Environmental Research—UFZ, Department of Environmental Immunology/Core Facility Studies, Leipzig (Lehmann I, Bauer M, Röder S, Schilde M, Nowak M, Herberth G, Müller J); Technical University Munich, Department of Pediatrics, Munich (Hoffmann U, Paschke M, Marra S); Clinical Research Group Molecular Dermatology, Department of Dermatology and Allergy, Technische Universität München (TUM), Munich (Ollert M, J. Grosch). BAMSE: We thank all the participants and parents participating in the BAMSE cohort and the staff involved in the study through the years. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

GINIplus: The GINIplus study was mainly supported for the first 3 years of the Federal Ministry for Education, Science, Research, and Technology (interventional arm) and Helmholtz Zentrum Munich (former GSF) (observational arm). The 4 year, 6 year, 10 year, and 15 year follow-up examinations of the GINIplus study were covered from the respective budgets of the 5 study centres (Helmholtz Zentrum Munich (former GSF), Research Institute at Marien-Hospital Wesel, LMU Munich, TU Munich and from 6 years onwards also from IUF—Leibniz Research-Institute for Environmental Medicine at the University of Düsseldorf) and a grant from the Federal Ministry for Environment (IUF Düsseldorf, FKZ 20462296). Further, the 15 year follow-up examination of the GINIplus study was supported by the Commission of the European Communities, the 7th Framework Program: MeDALL project, and as well by the companies Mead Johnson and Nestlé. LISA: The LISA study was mainly supported by grants from the Federal Ministry for Education, Science, Research, and Technology and in addition from Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef for the first 2 years. The 4 year, 6 year, 10 year, and 15 year follow-up examinations of the LISA study were covered from the respective budgets of the involved partners (Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef, IUF—Leibniz-Research Institute for Environmental Medicine at the University of Düsseldorf) and in addition by a grant from the Federal Ministry for Environment (IUF Düsseldorf, FKZ 20462296). Further, the 15-year follow-up examination of the LISA study was supported by the Commission of the European Communities, the 7th Framework Program: MeDALL project. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research

and innovation programme (grant agreement No. 949906). BAMSE: The BAMSE study is supported by grants from the Swedish Research Council, the Swedish Heart-Lung Foundation, Karolinska Institutet SFO Epidemiology and Region Stockholm (ALF project, and for cohort and database maintenance).

CONFLICT OF INTEREST

MG reports grants from Nestlé Vevey, Switzerland, during the conduct of the study and personal fees from Aimmune, ALK, AstraZeneca, Boehringer, GSK, Nestle, Novartis, and Sanofi outside the submitted work. SKoletzko reports grants from Mead Johnson Company during the conduct of the study; personal fees from Nestle, Danone, Shire, AbbVie, ThermoFisher, Janssen, Pfizer, Takeda, Mead Johnson, grants from BioGaia outside the submitted work. EM has received lecture and/or advisory board fees from ALK, AstraZeneca, Chiesi, Novartis and Sanofi outside the submitted work. All other authors have no interests to declare.

ORCID

Anna Kilanowski https://orcid.org/0000-0001-5328-6274
Elisabeth Thiering https://orcid.org/0000-0002-5429-9584
Joachim Heinrich https://orcid.org/0000-0002-9620-1629
Gunda Herberth https://orcid.org/0000-0003-0212-3509
Sibylle Koletzko https://orcid.org/0000-0003-2374-8778
Inger Kull https://orcid.org/0000-0001-6096-3771
Annette Peters https://orcid.org/0000-0001-6645-0985
Marie Standl https://orcid.org/0000-0002-5345-2049

REFERENCES

1. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global burden of disease study 2019. *Lancet*. 2020;396(10258):1204-1222. doi:10.1016/S0140-6736(20)30925-9
2. Nissen SP, Kjær HF, Høst A, Nielsen J, Halken S. The natural course of sensitization and allergic diseases from childhood to adulthood. *Pediatr Allergy Immunol*. 2013;24(6):549-555. doi:10.1111/pai.12108
3. Pinart M, Benet M, Annesi-Maesano I, et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitised and non-IgE-sensitised children in MeDALL: a population-based cohort study. *Lancet Respir Med*. 2014;2(2):131-140. doi:10.1016/S2213-2600(13)70277-7
4. Fuertes E, Standl M, Von Berg A, et al. Parental allergic disease before and after child birth poses similar risk for childhood allergies. *Allergy*. 2015;70(7):873-876. doi:10.1111/all.12609
5. Ferreira MA, Vonk JM, Baurecht H, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet*. 2017;49(12):1752-1757. doi:10.1038/ng.3985
6. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. The group health medical associates. *N Engl J Med*. 1995;332(3):133-138. doi:10.1056/NEJM199501193320301
7. Lee E, Hong SJ. Phenotypes of allergic diseases in children and their application in clinical situations. *Korean J Pediatr*. 2019;62(9):325-333. doi:10.3345/kjp.2018.07395
8. Panico L, Stuart B, Bartley M, Kelly Y. Asthma trajectories in early childhood: identifying modifiable factors. *PLOS One*. 2014;9(11):e111922. doi:10.1371/journal.pone.0111922
9. Ödling M, Wang G, Andersson N, et al. Characterization of asthma trajectories from infancy to young adulthood. *J Allergy Clin Immunol Pract*. 2021;16:2368-2376.e3. doi:10.1016/j.jaip.2021.02.007
10. Rancière F, Nikasinovic L, Bousquet J, Momas I. Onset and persistence of respiratory/allergic symptoms in preschoolers: new insights from the PARIS birth cohort. *Allergy*. 2013;68(9):1158-1167. doi:10.1111/all.12208
11. Garcia-Aymerich J, Benet M, Saeys Y, et al. Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. *Allergy*. 2015;70(8):973-984. doi:10.1111/all.12640
12. Belgrave DCM, Granell R, Simpson A, et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. *PLoS Med*. 2014;11(10):e1001748. doi:10.1371/journal.pmed.1001748
13. Bougas N, Just J, Beydon N, et al. Unsupervised trajectories of respiratory/allergic symptoms throughout childhood in the PARIS cohort. *Pediatr Allergy Immunol*. 2019;30(3):315-324. doi:10.1111/pai.13027
14. Forster F, Ege MJ, Gerlich J, et al. Trajectories of asthma and allergy symptoms from childhood to adulthood. *Allergy*. 2021;2:1192-1203. doi:10.1111/all.15075
15. Ferreira MAR, Vonk JM, Baurecht H, et al. Eleven loci with new reproducible genetic associations with allergic disease risk. *J Allergy Clin Immunol*. 2019;143(2):691-699. doi:10.1016/j.jaci.2018.03.012
16. Paternoster L, Standl M, Waage J, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet*. 2015;47(12):1449-1456. doi:10.1038/ng.3424
17. El-Husseini ZW, Gosens R, Dekker F, Koppelman GH. The genetics of asthma and the promise of genomics-guided drug target discovery. *Lancet Respir Med*. 2020;8(10):1045-1056. doi:10.1016/S2213-2600(20)30363-5
18. Waage J, Standl M, Curtin JA, et al. Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet*. 2018;50(8):1072-1080. doi:10.1038/s41588-018-0157-1
19. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z
20. Arabkhazaeli A, Ahmadizar F, Leusink M, et al. The association between a genetic risk score for allergy and the risk of developing allergies in childhood—results of the WHISTLER cohort. *Pediatr Allergy Immunol*. 2018;29(1):72-77. doi:10.1111/pai.12824
21. Clark H, Granell R, Curtin JA, et al. Differential associations of allergic disease genetic variants with developmental profiles of eczema, wheeze and rhinitis. *Clin Exp Allergy*. 2019;49(11):1475-1486. doi:10.1111/cea.13485
22. Lemonnier N, Melén E, Jiang Y, et al. A novel whole blood gene expression signature for asthma, dermatitis, and rhinitis multimorbidity in children and adolescents. *Allergy*. 2020;75(12):3248-3260. doi:10.1111/all.14314
23. Pawankar R, Holgate ST, Canonica GW, Lockey RF, Blaiss MS. WAO white book on allergy 2013 update. 2013.
24. Strachan DP. Hay fever, hygiene, and household size. *Br Med J*. 1989;299(6710):1259-1260. doi:10.1136/bmj.299.6710.1259
25. Roduit C, Frei R, von Mutius E, Lauener R. The hygiene hypothesis. In: Esser C, ed. *Environmental Influences on the Immune System*. Springer; 2016:77-96. doi:10.1007/978-3-7091-1890-0_4
26. Hu C, Duijts L, Erler NS, et al. Most associations of early-life environmental exposures and genetic risk factors poorly differentiate between eczema phenotypes: the generation R study. *Br J Dermatol*. 2019;181(6):1190-1197. doi:10.1111/bjd.17879

27. Heinrich J, Brüske I, Cramer C, et al. GINIplus and LISAplus. Design and selected results of two German birth cohorts about natural course of atopic diseases and its determinants. *Allergol Sel*. 2017;1(1):85-95. doi:10.5414/ALX01455E
28. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J*. 2005;26(2):319-338. doi:10.1183/09031936.05.00034805
29. Luzak A, Fuertes E, Flexeder C, et al. Which early life events or current environmental and lifestyle factors influence lung function in adolescents? – results from the GINIplus & LISAplus studies. *Respir Res*. 2017;18, 138. doi:10.1186/s12931-017-0619-5
30. Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic REFERENCE values for spirometry for the 3–95 year age range: the GLOBAL lung function 2012 equations. *Eur Respir J*. 2012;40(6):1324-1343. doi:10.1183/09031936.00080312
31. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4
32. Weidinger S, Illig T, Baurecht H, et al. Loss-of-function variations within the filaggrin gene predispose for atopic dermatitis with allergic sensitizations. *J Allergy Clin Immunol*. 2006;118(1):214-219. doi:10.1016/j.jaci.2006.05.004
33. Ballardini N, Bergström A, Wahlgren CF, et al. IgE antibodies in relation to prevalence and multimorbidity of eczema, asthma, and rhinitis from birth to adolescence. *Allergy*. 2016;71(3):342-349. doi:10.1111/all.12798
34. Asher MI, Keil U, Anderson HR, et al. International study of asthma and allergies in childhood (ISAAC): rationale and methods. *Eur Respir J*. 1995;8(3):483-491. doi:10.1183/09031936.95.08030483
35. Gref A, Merid SK, Gruzieva O, et al. Genome-wide interaction analysis of air pollution exposure and childhood asthma with functional follow-up. *Am J Respir Crit Care Med*. 2017;195(10):1373-1383. doi:10.1164/rccm.201605-1026OC
36. Genolini C, Alacoque X, Sentenac M, Arnaud C. Kml and kml3d: R packages to cluster longitudinal data. *J Stat Software*. 2015;65(1):1-34. doi:10.18637/jss.v065.i04
37. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing; 2019. https://www.R-project.org/
38. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(2):224-227. doi:10.1109/TPAMI.1979.4766909
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.
40. Schwarzer G, Carpenter JR, Rücker G. Fixed effect and random effects meta-analysis. In: Schwarzer G, Carpenter JR, Rücker G, eds. *Meta-Analysis with R*. Use R! Springer International Publishing; 2015:21-53. doi:10.1007/978-3-319-21416-0_2
41. Gustafsson D, Sjöberg O, Foucard T. Sensitization to food and airborne allergens in children with atopic dermatitis followed up to 7 years of age. *Pediatr Allergy Immunol*. 2003;14(6):448-452. doi:10.1046/j.0905-6157.2003.00093.x
42. Martin PE, Eckert JK, Koplin JJ, et al. Which infants with eczema are at risk of food allergy? Results from a population-based cohort. *Clin Exp Allergy*. 2015;45(1):255-264. doi:10.1111/cea.12406
43. Gough H, Grabenhenrich L, Reich A, et al. Allergic multimorbidity of asthma, rhinitis and eczema over 20 years in the German birth cohort MAS. *Pediatr Allergy Immunol*. 2015;26(5):431-437. doi:10.1111/pai.12410
44. Margolis DJ, Apter AJ, Gupta J, et al. The persistence of atopic dermatitis and filaggrin (FLG) mutations in a US longitudinal cohort. *J Allergy Clin Immunol*. 2012;130(4):912-917. doi:10.1016/j.jaci.2012.07.008
45. Mikhail I, Grayson MH. Asthma and viral infections: an intricate relationship. *Ann Allergy Asthma Immunol*. 2019;123(4):352-358. doi:10.1016/j.anai.2019.06.020
46. Rossi GA, Colin AA. Infantile respiratory syncytial virus and human rhinovirus infections: respective role in inception and persistence of wheezing. *Eur Respir J*. 2015;45(3):774-789. doi:10.1183/09031936.00062714
47. Bui DS, Lodge CJ, Perret JL, et al. Trajectories of asthma and allergies from 7 years to 53 years and associations with lung function and extrapulmonary comorbidity profiles: a prospective cohort study. *Lancet Respir Med*. 2021;9(4):387-396. doi:10.1016/S2213-2600(20)30413-6
48. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759-2772. doi:10.1038/s41596-020-0353-1

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kilanowski A, Thiering E, Wang G, et al. Allergic disease trajectories up to adolescence: Characteristics, early-life, and genetic determinants. *Allergy*. 2022;00:1-15. doi: 10.1111/all.15511

2 Publication: Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort

Title:	Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort
Authors:	Anna Kilanowski, Junyu Chen, Todd Everson, Elisabeth Thiering, Rory Wilson, Nicole Gladish, Melanie Waldenberger, Hongmei Zhang, Juan C. Celedón, Esteban G. Burchard, Annette Peters, Marie Standl, Anke Hüls
Journal:	*Allergy*
Year:	2022
Volume:	77:2803-2817
DOI:	10.1111/all.15315
Supplements:	https://onlinelibrary.wiley.com/doi/10.1111/all.15315 (available online)
Impact factor:	14.710 (Journal Citation Reports^{TM} 2021)
Rank:	1/27 in category ALLERGY (Journal Citation Reports^{TM} 2021)

Received: 2 November 2021 | Revised: 17 March 2022 | Accepted: 22 March 2022

DOI: 10.1111/all.15315

Allergy EUROPEAN JOURNAL OF ALLERGY AND CLINICAL IMMUNOLOGY EAACI WILEY

ORIGINAL ARTICLE

Autoimmunity and Clinical Immunology

Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort

Anna Kilanowski[1,2,3,4] | Junyu Chen[1] | Todd Everson[1,5] | Elisabeth Thiering[3,4] | Rory Wilson[6] | Nicole Gladish[7,8] | Melanie Waldenberger[3,6] | Hongmei Zhang[9] | Juan C. Celedón[10] | Esteban G. Burchard[11] | Annette Peters[3,12] | Marie Standl[3,13] | Anke Hüls[1,5]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

[2]Institute for Medical Information Processing, Biometry, and Epidemiology; Pettenkofer School of Public Health, LMU Munich, Munich, Germany

[3]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[4]Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, Munich, Germany

[5]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

[6]Research Unit Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[7]Centre for Molecular Medicine and Therapeutics, British Columbia Children's Hospital, Vancouver, BC, Canada

[8]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

[9]Division of Epidemiology, Biostatistics, and Environmental Health Sciences, School of Public Health, University of Memphis, Memphis, Tennessee, USA

[10]Division of Pediatric Pulmonary Medicine, UPMC Children's Hospital

Abstract

Background: Epigenomic (e.g., DNA methylation [DNAm]) changes have been hypothesized as intermediate step linking environmental exposures with allergic disease. Associations between individual DNAm at CpGs and allergic diseases have been reported, but their joint predictive capability is unknown.

Methods: Data were obtained from 240 children of the German LISA cohort. DNAm was measured in blood clots at 6 ($N = 234$) and 10 years ($N = 227$) using the Illumina EPIC chip. Presence of aeroallergen sensitization was measured in blood at 6, 10, and 15 years. We calculated six methylation risk scores (MRS) for allergy-related phenotypes, like total and specific IgE, asthma, or any allergies, based on available publications and assessed their performances both cross-sectionally (biomarker) and prospectively (predictor of the disease). Dose–response associations between aeroallergen sensitization and MRS were evaluated.

Results: All six allergy-related MRS were highly correlated ($r > .86$), and seven CpGs were included in more than one MRS. Cross-sectionally, we observed an 81% increased risk for aeroallergen sensitization at 6 years with an increased MRS by one standard deviation (best-performing MRS, 95% confidence interval = [43%; 227%]). Significant associations were also seen cross-sectionally at 10 years and prospectively, though the effect of the latter was attenuated when restricted to participants not sensitized at baseline. A clear dose–response relationship with levels of aeroallergen sensitization could be established cross-sectionally, but not prospectively.

Conclusion: We found good classification and prediction capabilities of calculated allergy-related MRS cross-sectionally, underlining the relevance of altered

Abbreviations: 95% CI, 95% confidence interval; CpG, cytosine and Guanine only separated by their phosphate backbone; DNAm, DNA methylation; ETS, environmental tobacco smoke exposure; EWAS, epigenome-wide association studies; gDNA, genomic DNA; GWAS, genome-wide association studies; IgE, immunoglobulin E; LISA, influence of life-style factors on development of the immune system and allergies in East and West Germany study; MRS, methylation risk score; PRS, polygenic risk score; RAST, Radio-Allergo-Sorbent-Test; ROC, receiver operating characteristic; RR, risk ratio.

Marie Standl and Anke Hüls shared last and corresponding authors.

of Pittsburgh, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[11]UCSF Asthma Collaboratory, University of California, San Francisco, California, USA

[12]Chair of Epidemiology, Ludwig-Maximilians University, Munich, Germany

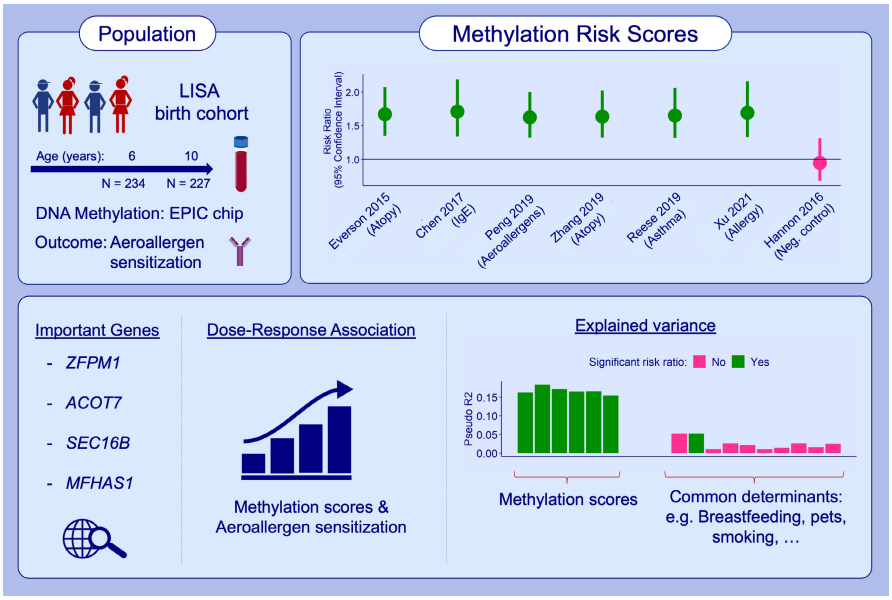
[13]German Center for Lung Research (DZL), Gießen, Germany

Correspondence
Anke Hüls, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd NE, 30322 Atlanta, GA, USA.
Email: anke.huels@emory.edu

Marie Standl, Institute of Epidemiology, Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany.
Email: marie.standl@helmholtz-muenchen.de

gene-regulation in allergic diseases and providing insights into potential DNAm biomarkers of aeroallergen sensitization.

KEYWORDS
allergic diseases, DNA methylation, epidemiology, methylation risk scores, polygenic risk scores



GRAPHICAL ABSTRACT
We calculated six MRS for allergy-related phenotypes and present their association with childhood aeroallergen sensitization. All six allergy-related MRS were highly correlated and seven CpGs were overlapping between the MRS, all located in genes associated with allergic diseases. A clear dose–response relationship with levels of aeroallergen sensitization could be established cross-sectionally, but not prospectively.
Abbreviations: ACOT7, acyl-CoA thioesterase 7; CpG, cytosine–guanosine dinucleotide; LISA, influence of lifestyle factors on development of immune system and allergies in east and west Germany study; MFHAS1, multifunctional ROCO family signaling regulator; MRS, methylation risk score; SEC16B, SEC16 homolog B, endoplasmic reticulum export factor; ZFPM1, zinc finger protein, FOG, family member 1

1 | INTRODUCTION

The link between genetic variation and allergic diseases is already well established by several genome-wide association studies (GWAS).[1] However, non-genetic and environmental determinants, like birth order[2,3] or pet ownership,[4] have also been discussed and might explain further variance in allergic diseases (e.g., asthma and allergic rhino-conjunctivitis) through epigenetic mechanisms such as DNA methylation (DNAm).[5]

Over the past years, epigenome-wide association studies (EWAS) have identified differential DNAm at several CpG (addition of a methyl group to a cytosine in the context of CpG dinucleotides) sites to be associated with allergic phenotypes including atopy, defined as allergic reaction in skin prick test, high total (≥200 kU/L) or

specific immunoglobulin E (IgE) (≥0.35 kU/L),[6–9] childhood asthma[10] or any allergic disease plus sensitization.[11]

Compared to large-scale GWAS, current EWAS often have limited sample size[12] with the maximum sample size in allergic phenotypes being 3493.[10] Age-, tissue-, and cell type-specific differences in DNAm patterns further limit the generalizability of results.[5] Additionally, identified CpGs usually have small effect sizes, similar to single genetic variants for common diseases.[13] Given the unknown generalizability and replicability of recent EWAS of allergy-related phenotypes, mainly due to small sample sizes, a proof of the applicability of their results in a predictive context is of great interest for methylation studies.

Following the same methodology as previously employed for polygenic risk scores (PRS),[14] methylation risk scores (MRS) could be used to evaluate the reproducibility of published atopy-related EWAS and their prediction accuracy cross-sectionally (as biomarkers of aeroallergen sensitization) and prospectively (as predictors of future aeroallergen sensitization). MRS have been reported as biomarkers for exposures like smoking,[15] as predictors of poor survival outcomes in hepatocellular carcinoma[16] and disease indicators for prostate cancer, even outperforming other known risk factors.[17]

Methylation risk scores are calculated by using external evidence from published EWAS and weighting the CpGs in the target cohort with the respective effect sizes from external EWAS on the same phenotype to calculate a weighted average. Thereby, small effects of single CpG sites are accumulated, which increases the statistical power and prediction accuracy.[14]

The objective of this study is to calculate MRS that are derived from published EWAS, in order to classify cross-sectional, and predict prospective, childhood aeroallergen sensitization in the prospective German LISA birth cohort. For this, we evaluated (1) the predictive accuracy of six different MRS in both cross-sectional and prospective models, (2) their overlap and correlations and (3) compared their associations and prediction accuracy to other known determinants of allergic sensitization and individual CpG sites.

2 | METHODS

2.1 | Study population

We used data from the prospective German birth cohort on the Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany (LISA), which recruited 3097 full-term healthy newborns of European ancestry between 1997 and 1999 in four study centers (Munich, Wesel, Leipzig and Bad Honnef). The study was approved by the local ethics committees (Bavarian General Medical Council, Medical Council for North-Rhine-Westphalia and the University of Leipzig), and informed parental consent was given. More information can be found elsewhere.[18]

Allergen-specific serum immunoglobulin E (IgE) concentrations were assayed by the CAP-RAST FEIA system (Pharmacia Diagnostics) according to the manufacturer's instructions and in line with global recommendations.[19] An overall screening test was used to test allergic sensitization against aeroallergens at 6, 10, and 15 years. Our outcome was defined by a specific IgE threshold of >0.35 kU/L (Radio-Allergo-Sorbent-Test [RAST] class 1) to the screening test of common aeroallergens (Dermatophagoides pteronyssinus, cat, dog, rye, timothy grass, Cladosporium herbarum, birch, and mugwort). Further RAST classes were defined according to common cutoffs,[20] where RAST 0 implies no allergic sensitization and RAST 5 or 6 (combined into one category) is the highest possible value. Questionnaire-based information on symptoms of rhino-conjunctivitis (concurrent running nose and itching eyes) and wheezing in the previous 12 months was collected at the same time points.

We assessed potential determinants of allergic diseases, which have been shown to be associated with different allergic diseases or lung function, such as parental education,[21] breastfeeding,[21,22] birth order,[2] pet holding,[21] maternal smoking during pregnancy,[23] environmental tobacco smoke exposure[23] (ETS) or bronchitis infections[23] in early childhood, as well as PRS calculated as weighted scores from genome-wide significant GWAS hits for any allergic disease,[1,24] asthma,[25] dermatitis,[26] allergic rhinitis,[27] and total IgE.[28] Additional information on the study design and on the definition of determinants of allergic diseases can be found in the Appendix S1 (Table S1, Methods 1).

2.2 | DNA methylation (DNAm) data

Samples using genomic DNA (gDNA) from blood clots at 6 and 10 years were analyzed using the MethylationEPIC BeadChip (Illumina, Inc.). Paired samples were placed on the same chip to avoid batch effects among pairs. CpGs on sex chromosomes and those having missing values low intensities were excluded. We used functional normalization[29] to normalize the data and ComBat[30] to adjust for technical variation. After outlier removal, the final dataset includes information on 774,330 CpG probes for 461 DNAm samples, 234 at six and 227 at 10 years, with an overlap of 221 participants with DNAm data at both time points. Cell type proportions were estimated both with the Houseman method[31] using a new reference panel[32] and with the EpiDISH[33] package, which additionally includes eosinophil estimates. Further information on processing and quality control can be found in the Appendix S1 (Methods 2 and Figure S1).

2.3 | Calculation of MRS

We calculated MRS based on the effect estimates or other summary statistics for CpG sites that have previously been associated with allergic diseases[6,8,10,11] or additionally provided summary statistics[7,9] for associations with up to a raw *p*-value of .1 for each EWAS. A weighted sum of DNAm beta values, defined as

estimated methylation level, was then transformed to z-scores, and MRS were produced for each respective EWAS and differing *p*-value thresholds. A literature review identified EWAS of phenotypes related to atopy or high IgE. Further publications for any kind of allergic disease were included, if they were conducted in a larger consortium framework (asthma[10] and any allergic disease[11]). Seven MRS were calculated, one for high IgE,[7] one for aeroallergen sensitization,[8] two for atopy, defined as high total IgE or positive skin-prick test and sensitization, respectively,[6,9] one for asthma[10] and one for any allergic disease[11] as well as one MRS for schizophrenia[34] as negative control. In all seven EWAS, DNAm was measured in whole blood. Varying *p*-value thresholds from 1×10^{-1} to the lowest reported *p*-value per EWAS were considered, resulting in several scores per EWAS with a decreasing number of CpGs for smaller *p*-values, similar to what is known as "thresholding" for PRS.[14] To correct for correlations between included CpG sites, co-methylated regions were calculated using the CoMeBack method,[35] which identifies co-methylated regions based on correlation and proximity of CpGs. In accordance with the original publication, we did this based on residuals corrected for Houseman cell type proportions of the LISA study. Only one CpG per co-methylated region was included in the final MRS, a procedure similar to "clumping" in PRS approaches.[14] All MRS were calculated as z-scores following a standard normal distribution. A more detailed description is further provided in the supplementary information (Methods 3).

2.4 | Statistical analysis

Associations between each MRS and aeroallergen sensitization were estimated using logistic and Poisson regression with robust standard errors. Poisson regression was used to assess risk ratios (RR), as aeroallergen sensitization was not a rare outcome in our sample and thus odds ratios would not resemble RR. All models were adjusted for sex, age, whether the blood was taken in the allergy season (March to August), as current pollen exposure might influence DNAm[36] as well as circulating IgE levels, and estimated cell type proportions using EpiDISH. We applied the following criteria to evaluate and compare the performance of different MRS: 1) RR and corresponding 95% confidence intervals (95% CI) were used to evaluate the strength and accuracy of the association with aeroallergen sensitization; 2) C-statistic, the area-under-the-curve and 3) explained variance (Pseudo R^2) were used to evaluate the prediction accuracy for aeroallergen sensitization. The different MRS were compared and evaluated under four different scenarios: Two cross-sectional models assessing the association at 6 and 10 years and two prospective models, assessing the association between the MRS and subsequent aeroallergen sensitization (MRS at 6 and 10 years as predictor of aeroallergen sensitization at 10 and 15 years, respectively). As a sensitivity analysis, the prospective models were calculated in the non-sensitized population only, excluding all participants with sensitization at the time of DNA methylation measurement, thereby analyzing only those who could develop new sensitization between the two time points. We furthermore calculated the receiver operating characteristic (ROC) for the cross-sectional analyses to assess the diagnostic ability of our MRS.

The best MRS per EWAS were selected based on the highest c-statistic in the cross-sectional model at 6 years. Correlations between the seven "best MRS" (one per EWAS) and the corresponding CpGs were evaluated. All CpGs reported in the available EWAS were tested for replication in the LISA study, both with the Houseman (as done in the original EWAS[6–10]) and EpiDISH cell type proportions, with successful replication being defined as a *p*-value below .05 after adjusting for the total number of tested CpGs from all EWAS using the Benjamini–Hochberg correction.[37]

Associations between the MRS and the six RAST classes were investigated using boxplots to evaluate a potential dose–response relationship with increasing levels of aeroallergen sensitization and ordinal logistic regression analyses.

To compare the strength of association and prediction accuracy of the MRS to those of other common determinants of allergic diseases (including allergy-related PRS, Table S1) and the most common single CpGs, we calculated the explained variance and strength of association (RR and 95% CI) with aeroallergen sensitization and compared it to the performance of the MRS.

We further assessed the association of all MRS with allergic disease symptoms, namely rhino-conjunctivitis and wheezing, using the same approach as described above. In addition, we calculated correlations between the MRS and the different estimated cell type proportions to assess whether a specific cell type was overrepresented in the MRS. In a sensitivity analysis, we tested the impact of co-methylated regions on the robustness of MRS: Namely, we calculated MRS with and without application of the CoMeBack method and used a reference population instead of the LISA study to determine the co-methylated regions (see Gatev et al., 2020[35] for details).

All statistical analyses were run in R[38] V4.1.2.

3 | RESULTS

3.1 | Description of study participants

We included 461 samples, collected from 240 participants of the LISA birth cohort, in our analysis, both from six (N = 234) and ten (N = 227) years of age (Table 1), of which 221 were paired with DNAm data available at both time points (Figure S1). The sample included slightly more males than females (58% vs. 42%) and the prevalence of rhino-conjunctivitis symptoms increased, while that of wheezing symptoms decreased, between 6 and 10 years. Relevant outcome measures used in the 6-year sample are aeroallergen sensitization at 6 years (prevalence: 32.6%, 74 cases) and at 10 years (44.9%, 105 cases). In the 10-year sample, aeroallergen sensitization at 10 years (44.5%, 101 cases) and at 15 years (37%, 84 cases) were analyzed in the main analysis. Differences seen between the two time points are

TABLE 1 Sample information and variable distribution in the final analysis sets at 6 and 10 years of age in the LISA birth cohort

	Analysis sample (*N* = 240)	
	DNAm measured at 6 years (*N* = 234)	DNAm measured at 10 years (*N* = 227)
Male sex, % (*n*) [*n* missing]	57.69% (135) [0]	57.71% (131) [0]
Age at DNA methylation measurement [years], mean (sd)	6.072 (0.15)	10.155 (0.14)
High parental education[a], % (*n*) [*n* missing]	79.49% (186) [2]	79.74% (181) [2]
Aeroallergen sensitization[b], % (*n*) [*n* missing]		
At 6 years	31.62% (74) [0]	29.96% (68) [0]
At 10 years	44.87% (105) [0]	44.49% (101) [0]
At 15 years	37.18% (87) [63]	37.00% (84) [60]
Allergic symptoms in last 12 months, % (*n*) [*n* missing]		
Rhino-conjunctivitis	8.55% (20) [1]	14.98% (34) [4]
Wheezing	10.68% (25) [2]	7.93% (18) [2]
Blood taken in allergy season (Mar - Aug), % (*n*) [*n* missing]		
At 6 years	67.52% (158) [0]	66.96% (152) [0]
At 10 years	52.99% (124) [0]	51.10% (116) [0]
At 15 years	41.45% (97) [55]	39.65% (90) [52]
Polygenic risk scores[c], mean (sd)		
Any allergic disease	0.187 (0.95)	0.187 (0.93)
Asthma	0.147 (1.04)	0.155 (1.05)
Dermatitis	0.002 (0.99)	−0.014 (0.99)
Rhinitis	0.023 (0.97)	0.052 (0.95)
Total IgE	0.083 (0.97)	0.084 (0.97)
Family history of allergic diseases, % (*n*) [*n* missing]		
No parent	29.06% (68) [24]	28.19% (64) [24]
One parent	42.74% (100) [24]	43.17% (98) [24]
Both parents	17.95% (42) [24]	18.06% (41) [24]
Other known risk factors for allergy, % (*n*) [*n* missing]		
Smoking during pregnancy	7.26% (17) [7]	7.05% (16) [6]
Breastfeeding in first 4 months	83.33% (195) [1]	83.70% (190) [1]
Older siblings	49.57% (116) [0]	48.90% (111) [0]
ETS[d] in first 4 years	25.64% (60) [3]	25.11% (57) [3]
Bronchitis infection in first 3 years	64.96% (152) [2]	66.96% (152) [2]
Cat or dog in first 4 years	19.66% (46) [8]	19.82% (45) [9]

Note: If time point of measurement is not mentioned, data were obtained from questionnaire data filled out by the LISA parents at birth, 1, 2, or 4 years of age.

[a]Defined as more than 10 years of education.

[b]Defined by a specific IgE threshold of >0.35 kU/L (Radio-Allergo-Sorbent-Test (RAST) class 1).

[c]z-scores.

[d]Environmental tobacco smoke exposure.

due to sample removal, as originally all samples were paired and are presenting the same baseline characteristics. Baseline characteristics from our analysis sample (*N* = 240) are similar to the total study population of the LISA Munich cohort (*N* = 1464, Table S2).

3.2 | Methylation risk scores

Table 2 shows information on the seven EWAS, phenotype, age group, and sample size from which MRS were calculated. The EWAS reported between 13 and 395 significant signals and varied by age, from 4 to 18 years, and ethnicity, covering not only European but also Hispanic and multi-ethnic populations. The best MRS per EWAS were selected based on the highest c-statistic in the cross-sectional model at 6 years across all *p*-value thresholds that were tested (Figure 1 and Table S3). The best-performing MRS included two (Everson2015,[6] atopy) to 24 (Zhang2019,[9] atopy) CpGs for *p*-value thresholds ranging from 1×10^{-4} (Zhang2019,[9] atopy) to 1×10^{-13} (Peng2019,[8] aeroallergens). CpG sites and the corresponding weights for the best MRS are listed in Table S4.

TABLE 2 Overview of included EWAS, their phenotypes (allergy-related outcome) and age groups

EWAS	Phenotype	No. of significant CpGs with FDR-corrected *p*-values <.05 [with raw *p*-values <.01[a]]	Sample size (discovery +replication)	Age	Ethnicity	*p*-value threshold for best MRS[b]	No. of included CpGs (best MRS)	Cohorts	Chip
Everson2015[6]	Atopy status (IgE ≥200 kU/L)	13	367 + 464	18	European	1×10^{-8}	2	IoW +BAMSE	450K
Chen2017[7]	Log10(IgE)	200 [25 089]	879 (meta)	6–22	Hispanic	1×10^{-8}	3	PR-GOAL, GALA II	450K
Peng2019[8]	Environmental allergen sensitization (≥0.35 IU/ml to common aeroallergens)	395	739 (total)	Mean: 7.7 & 9.8	Multi-ethnic	1×10^{-13}	5	Project Viva, Generation R	450K
Zhang2019[9]	Atopic status (≥3 mm greater than negative control in SPT or IgE ≥ 0.35 kU/L for mix inhalant of food allergens)	35 [775]	376 + 267	10 & 18	European	1×10^{-4}	24	IoW +BAMSE	450K/EPIC
Reese2019[10]	Childhood asthma	179	3493 (meta)	Mean range: 7.1 to 17.01	Multi-Ethnic	1×10^{-8}	9	BAMSE, CHOP, GALA II, ICAC, NFBC 1986, PIAMA, RAINE, STOPPA	450K
Xu2021[11]	Any allergic disease (asthma, eczema, rhinitis) PLUS sensitization against common aeroallergens (≥0.35 IU/ml)	21	1457 + 1436	4 & 8	European	1×10^{-9}	18	BAMSE, INMA, PIAMA +EDEN, ECA, INMA, PIAMA, Karelia	450K
Hannon2016[34]	Schizophrenia (negative control)	2519	675 + 847	Adults	European	1×10^{-5}	867	UCL case-control +Aberdeen case-control, MZ twins cohort	450K

[a]Some studies provided additional summary statistics for all CpG sites with *p*-values <.01, n/a indicates that these additional summary statistics were not provided.

[b]Best MRS per EWAS were selected based on the highest c-statistic in the cross-sectional model at 6 years across all *p*-value thresholds that were tested (see also Table S2, where all MRS are evaluated).

The six allergy-related MRS were highly correlated with each other but not with the negative control (age 6 years: Figure S3, age 10 years: Table S5). A total of seven CpGs are included in more than one allergy-related MRS, with cg11699125 being the most common one included in all but one allergy-related MRS. All of these seven CpGs could be successfully replicated in LISA (Table 3) and were annotated to the genes *ZFPM1*, *ACOT7*, *MFHAS1*, and *SEC16B* using UCSC reference genes from the Illumina annotation file. Replication of reported EWAS signals (1501 in total) in LISA at six (N = 234) and ten (N = 227) years yielded 554 and 288 replicated hits correcting for Houseman and 111 and one replicated hits correcting for EpiDISH cell type estimates, respectively. Thus, we observe highly reduced replication rates when including eosinophils as a cell type confounder. Of note, of the published EWAS,[6–11] only one (including the cohorts in PR-GOAL and GALA-II)[7] controlled for eosinophils in their analyses. Complete results can be found in Table S6.

3.3 | MRS as cross-sectional biomarkers

Figure 1 and Table S3 present results from evaluating MRS that were calculated based on different p-value thresholds and EWAS

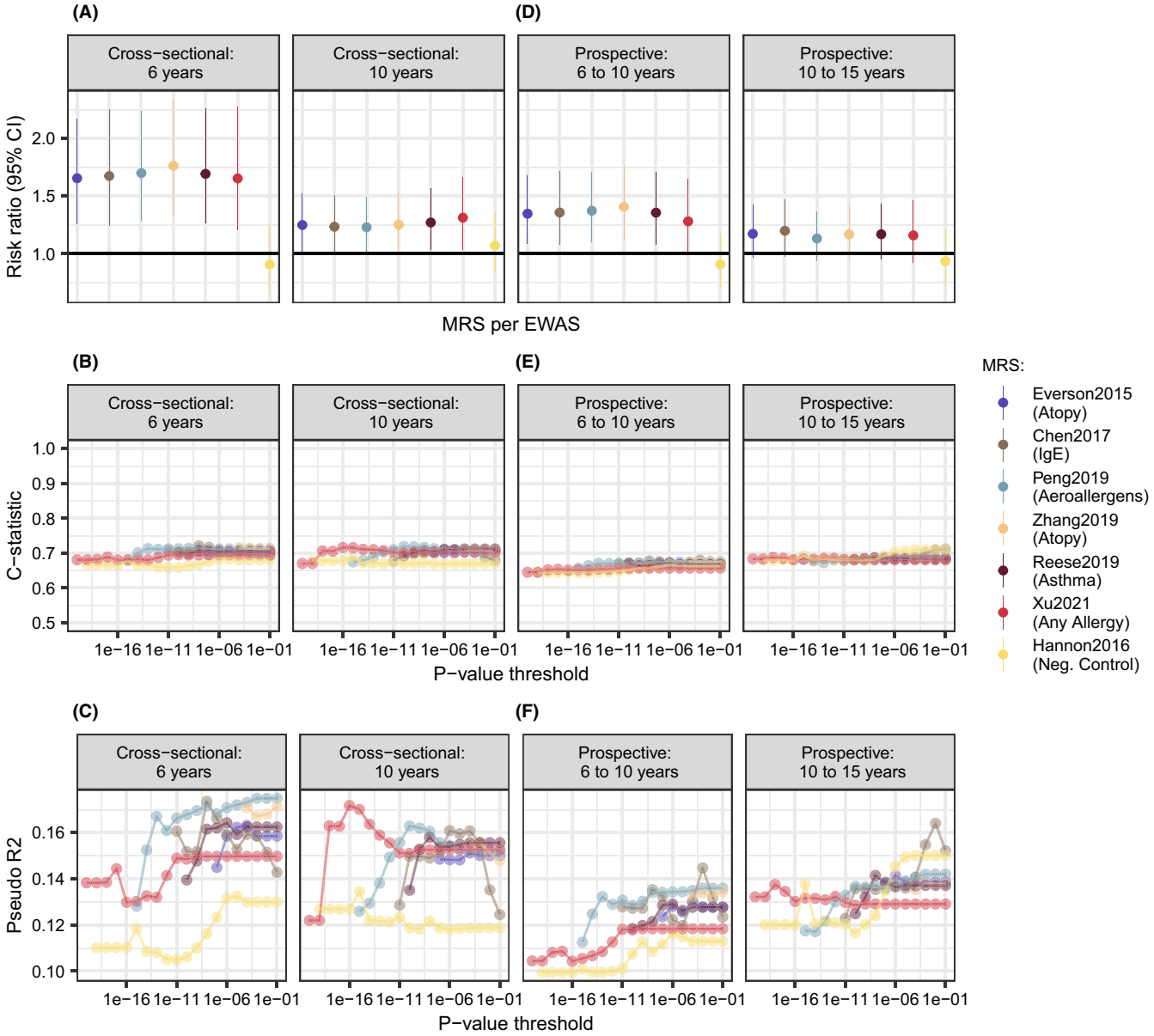


FIGURE 1 Predictive capabilities of MRS on aeroallergen sensitization. Four different models and criterions are displayed, assessing the cross-sectional and prospective impact of MRS as well as their (A) mean effect size per publication over all p-value thresholds, (B) performance and (C) explained variance for the different p-values thresholds (determining how many CpG sites were included in the MRS). All models are adjusted for sex, age, whether the blood was taken within the allergy season and cell type proportions. RR (A) were derived from Poisson regressions, whereas the other criterions (B&C) were calculated using logistic regression. Sample sizes for the four models were n = 234, n = 227, n = 234 and n = 167, respectively

TABLE 3 Overview of CpGs, which are present in more than one of the best-performing MRSs and their association with allergy-related outcomes in the original EWAS (A) and in the LISA cohort (replication, B)

CpG (gene)	Allergy-related outcome	cg04983687 (ZFPM1)	cg09249800 (ACOT7)	cg11699125 (ACOT7)	cg11988722 (intergenic)	cg12077460 (MFHAS1)	cg14011077 (intergenic)	cg17971251 (SEC16B)
A. Effect direction (*p*-values[a]) from original EWAS								
Everson2015[6]	Atopy	−(6.46E−09)	−(8.52E−09)					
Chen2017[7]	IgE	−(1.50E−12)	−(2.60E−11)	−(6.70E−07)				
Peng2019[8]	Aeroallergens	−(7.11E−15)		−(7.28E−12)	−(6.51E−14)			−(4.03E−11)
Zhang2019[9]	Atopy		−(1.17E−04)	−(4.34E−05)	−(9.04E−05)	−(2.98E−05)	−(2.83E−05)	−(1.46E−05)
Reese2019[10]	Asthma	−(1.33E−10)	−(1.19E−08)	−(7.54E−10)			−(7.02E−09)	−(9.52E−09)
Xu 2021[11]	Any allergy			−(5.84E−19)		−(3.73E−13)		
B. Beta coefficients (*p*-values[b]) in LISA								
LISA: 6 years	Aeroallergen sensitization	−11.61[c] (1.62E−03)	−14.71[c] (1.53E−02)	−11.94[c] (3.14E−02)	−21.59[c] (6.31E−04)	−19.65[c] (3.17E−03)	−40.89[c] (1.02E−02)	−26.84[c] (5.19E−04)
LISA: 10 years	Aeroallergen sensitization	−3.93 (1.62E−01)	−9.77[c] (2.79E−03)	−8.41[c] (2.47E−03)	−10.04[c] (2.35E−02)	−6.61 (4.27E−01)	−14.67 (1.22E−01)	−12.90[c] (8.01E−03)

Note: In A, effect direction (+/-) is reported instead of effect estimates because of the different allergy-related outcomes that were used in the original EWAS, which do not allow a direct comparison of effect estimates.

[a]Raw *p*-values corrected for confounding variables as per EWAS (check[6–11] for details).

[b]Raw *p*-values given corrected for Houseman cell type estimates.

[c]These CpGs were all successfully replicated in the LISA cohort (FDR threshold of 0.05).

for the cross-sectional (age 6 and 10 years) as well as prospective analyses (6–10 years and 10–15 years). To improve clarity, Figure 1A presents the mean MRS over all *p*-value thresholds per EWAS. All allergy-related MRS were significantly associated with aeroallergen sensitization in LISA (Figure 1A). Effect sizes were very similar between different MRS ranging from RR = 1.47 [95% CI: 1.19; 1.84] to RR = 1.81 [1.44; 2.27] in the cross-sectional model at 6 years and from RR = 1.12 [0.87; 1.44] to RR = 1.40 [1.19; 1.64] at 10 years (Table S3). Classification accuracy (Figure 1B, c-statistic) was about 0.7 for all allergy-related MRS and the best scores explain more than 15% of variance in aeroallergen sensitization, quantified with pseudo R^2, at 6 years and more than 12% at 10 years (Figure 1C). The negative control (MRS for schizophrenia) was not associated with aeroallergen sensitization in LISA. The ROC curves display similar patterns for allergy-related MRS and the negative control performs worst (Figure S2).

3.4 | MRS as prospective predictors

In the prospective models, all allergy-related MRS are significantly associated with aeroallergen sensitization, even though the effect estimates are smaller than in the cross-sectional models (Figure 1D,A). The prediction accuracy and the explained variance of the prospective models was smaller than in the cross-sectional models. For example, the pseudo R^2 decreased from explaining roughly between 12% and 15% of the variance in the cross-sectional models to only 8%–12% in the prospective ones (Figure 1F). The c-statistic was also slightly lower with ~0.65 in the prospective models instead of ~0.7 in the cross-sectional models. In a sensitivity analysis, we analyzed whether prospective associations are observed because of participants that were already sensitized at the time of DNAm measurement. For this, we ran the prospective analyses restricted to the non-sensitized population only. Looking only at the non-sensitized population (N = 160 from 6 to 10, N = 99 from 10 to 15 years), the effect of MRS on prospective aeroallergen sensitization was further attenuated and no significant association was observed (Table S7). This might imply that DNAm is a consequence or biomarker of aeroallergen sensitization rather than a predictor of sensitization development.

3.5 | Dose–response relationship

Figure 2 shows a clear and significant positive trend between higher MRS and higher RAST classes in the two cross-sectional analyses, except for the negative control (Figure 2A,B). This trend can be seen for all allergy-related MRS, independent of age group, ethnicity, or specific phenotype in the original EWAS. The trend was weaker in the prospective models (Figure 2C,D). The prospective trend from 6 to 10 years was significant for all allergy-related MRS, but with lower odds ratios than in the cross-sectional models. The prospective trend from 10 to 15 years was only significant for two of the six allergy-related MRS (Figure 2D and Table S8).

3.6 | Prediction accuracy of MRS in comparison with known determinants

As seen in Figures 1 and 3, the explained variance of allergy-related MRS is about 15% in the cross-sectional model at 6 years. Explained variance by other common determinants was lower, with family history of allergic diseases explaining around 5% and all others <3%, including PRS (Figure 3). Significant associations with aeroallergen sensitization were only present for the MRS and having two parents with a history of allergic diseases. Of the seven CpGs, present in more than one MRS, all were significantly associated with a reduced risk for sensitization and the pseudo R^2 was similar to the MRS, especially for cg17971251 and cg11988722.

3.7 | Prediction accuracy of allergy-related MRS for other allergic symptoms

Associations between calculated MRS and allergic symptoms, such as rhino-conjunctivitis and wheezing, were weaker than associations with aeroallergen sensitization and those models explained less variance (pseudo R^2 <0.09 for rhino-conjunctivitis and <0.14 for wheeze) (Figure S4A,C). The prediction accuracy for rhino-conjunctivitis is similar to the accuracy for aeroallergen sensitization (c-statistic ~0.7). In contrast to this, the association between MRS and wheezing was stronger in terms of effect estimates and prediction accuracy. However, the higher RRs for wheezing and their wide CIs can also be attributed to the lower case numbers for allergic symptoms (Rhino-conjunctivitis: n = 20 and wheezing: n = 25 at 6 years), and these results should be interpreted cautiously (Figure S4 and Table S9). Results of best-performing MRS at 10 years and their association with symptoms of wheezing and allergic rhinitis can be found in Table S10.

3.8 | Correlations with cell type proportions

During bulk DNAm analysis, several different blood cell types with differing methylation profiles are analyzed. To assess whether a specific cell type is overrepresented in the MRS, we calculated correlations between the MRS and the different estimated cell type proportions. There was little correlation ($r \leq .3$) between the allergy-related MRS and estimated cell types apart from eosinophils (r = [.53; .59]), indicating that the MRS represent differential DNA methylation-related to aeroallergen sensitization independent of most immune cell types, apart from the known association with eosinophils[39] (Figure S5).

3.9 | Robustness of MRS to determination of co-methylated regions

In our main analysis, co-methylated regions were determined using the LISA cohort. Using a reference population instead of our own

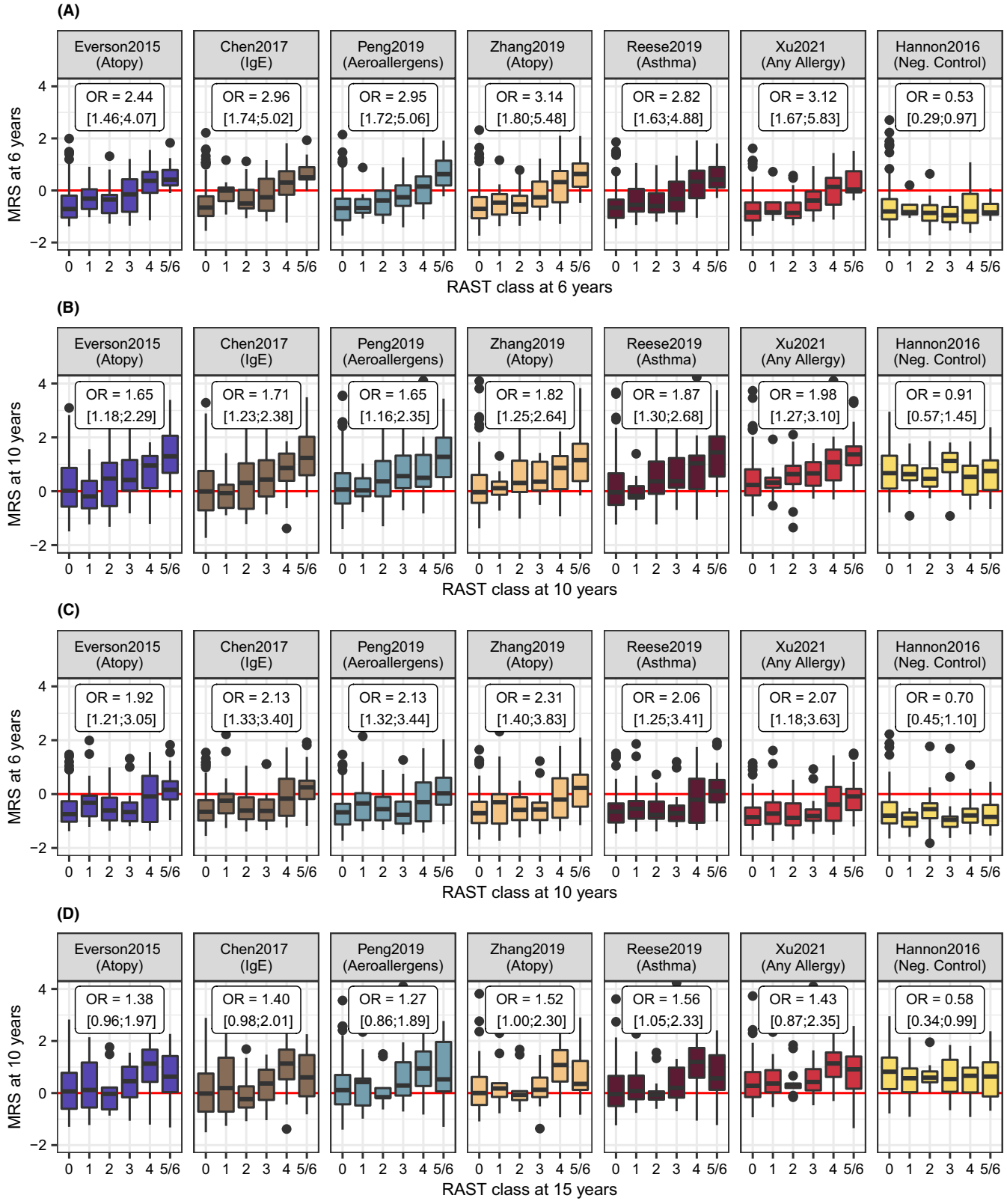


FIGURE 2 Dose–response relationship of MRS z-scores and RAST classes of aeroallergen sensitization cross-sectionally at (A) 6 and (B) 10 years and prospectively from (C) 6 to 10 and (D) 10–15 years. The fifth and sixth RAST classes are combined due to the low sample size in the highest class. Odds ratios (OR) and 95% confidence intervals from ordinal logistic regression analysis of the association between RAST classes and MRS are displayed in their respective panels

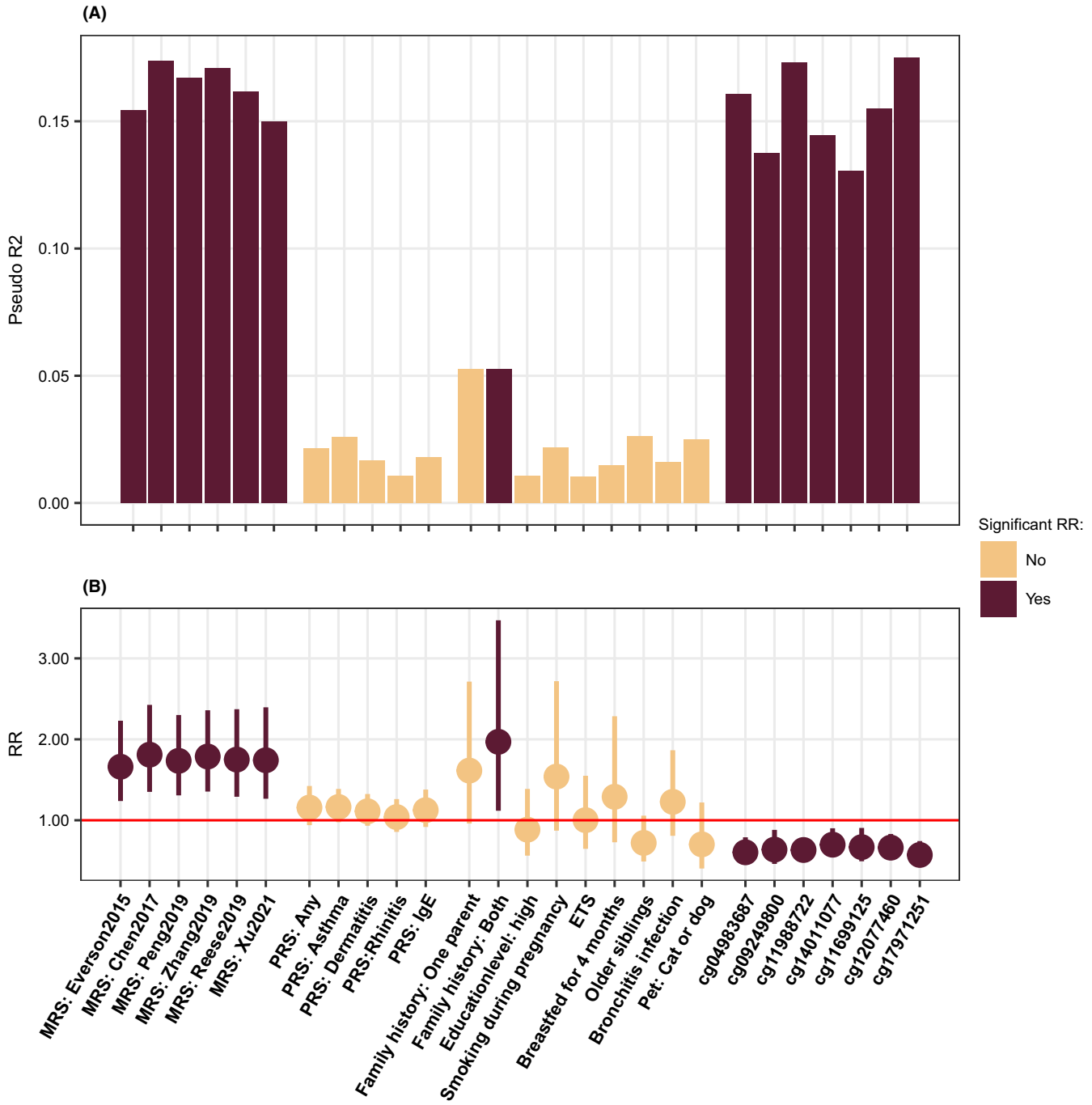


FIGURE 3 Prediction accuracy of MRS in comparison with other known risk factors. (A) Explained variance was assessed using logistic regression and (B) RR and 95% CI using Poisson regression with robust standard errors. For continuous variables (MRS, PRS, and CpGs z-scores were used; hence, the RR estimate can be understood per one standard error increase). All models were adjusted for sex, age and those with MRS and single CpGs additionally for cell type proportions and whether the blood was taken within the allergy season. Significance was determined in the Poisson model with a threshold of 0.05

LISA cohort for the determination of co-methylation region (as described in[35]) or no filtering based on CoMeBack at all did not have an impact on our main results or the number of included CpGs (Tables S11 and S12). Individual CpGs included in the final MRS were correlated, despite application of the CoMeBack method, which only removes correlated CpG sites that are in close proximity to each other (Figure S6).

4 | DISCUSSION

In the present study, we calculated different MRS from available EWAS of atopy, high IgE, asthma, or any allergic disease and assessed their prediction accuracy for childhood aeroallergen sensitization using cross-sectional and prospective data on DNAm and sensitization from the German LISA study. We showed the superior

performance of allergy-related MRS compared to well-known determinants of allergic diseases, like birth order, as well as their high correlation with each other. Seven CpGs were overlapping between the MRS, all located in previously reported genes associated with allergic diseases, and were successfully replicated in the LISA study. The best-performing MRS show a clear dose–response relationship with RAST classes of aeroallergen sensitization and explain more variance in aeroallergen sensitization than common determinants or PRS. However, we noticed differences between cross-sectional and prospective analyses, with the latter showing smaller effect sizes, lower prediction accuracy and less explained variance.

Our results fit with the accumulating evidence of improved disease definition using DNAm patterns and more specifically MRS as biomarkers for exposures[15] or predictor of diseases[16]

Looking at other determinants, MRS outperform them in explained variance, with about 15% of explained variance vs. <7% for the next best determinant. Similar values are achieved by the seven most commonly represented CpGs in the MRS. This highlights the role of DNAm as important allergy-specific factor. Even though other determinants of allergic disease have been widely established and are also included in clinical recommendations,[40] we could only observe significantly increased risk for the epigenetic factors and if both parents had a history of allergic diseases. Lack of associations with the other determinants could be explained by the relatively small sample size in this sub-sample of the LISA cohort. Furthermore, the low predictive capabilities of a PRS for asthma in childhood were published previously[41] and might underline our results of larger epigenetic associations as these lie on a level of omics closer to the phenotype.[42]

We found that the seven most important CpGs included in more than one MRS mapped to the genes *ZFPM1*, *ACOT7*, *MFHAS1*, and *SEC16B*, all of which have been reported in relation to allergic diseases.[43–46] The first three genes affect inflammatory responses through mast cell differentiation and development of cysteinyl leukotrienes. *ACOT7* has also been discussed as an important "cross-tissue allergy-associated methylation site" by one of the discovering EWAS.[11] The functional pathways of *SEC16B* have yet to be elucidated.

The MRS calculated for the 6-year data showed a stronger effect with aeroallergen sensitization at 6 years (cross-sectional analysis) than with aeroallergen sensitization at 10 years (prospective analysis; RR~1.7 vs. ~1.4). The RR were further attenuated and not significant anymore when analyzing only the non-sensitized population in both prospective models. This might indicate that the MRS are in fact not predictive of sensitization at a later time point but coincide with or follow aeroallergen sensitization and the prospective models only capture the effect of already sensitized participants at baseline. However, the prospective analysis of the non-sensitized population is limited by a small sample size and thus limited statistical power. The positive trend between MRS and RAST classes seen in the cross-sectional models could not be seen in the prospective models, hence underlining the cross-sectional but not predictive nature of the association. A prospective prediction capability could have helped with early detection of allergic disease development, and future studies might evaluate the prospective capabilities of combining IgE and DNAm measurements to improve prediction of allergic disease development. Development of an enhanced predictive tool is of great interest in the context of personalized medicine and might include genetic and epigenetic aspects, as well as IgE as already available biomarker. Nevertheless, the observed cross-sectional classification capability of MRS underline the relevance of altered gene-regulation in allergic diseases, aligning with previous publications noting that DNAm changes are more often seen as a consequence rather than the cause of a disease and that especially SNP-CpG associations are not necessarily causal.[47]

We observed reduced replication rates of reported CpGs when adjusting for EpiDISH cell type estimates compared to the often used Houseman estimates (7.4% vs. 36.9% at 6 years). This might indicate that a high portion of previously seen associations may be attributable to eosinophils, which are not estimated in Houseman proportions. Notably, our MRS results remain significant even when adjusting for eosinophils, whereas EWAS replication is highly diminished.

In our study, we did not observe differences with ethnicity for our MRS in an population of European ancestry, as the MRS calculated from an EWAS of Hispanics with multiple racial backgrounds[7] performed just as well and sometimes even better than European-ancestry derived MRS. This aligns with the meta-analysis results from the EWAS on asthma conducted in the pregnancy and childhood epigenetics consortium,[48] which did not see any influence of ethnicity on detected CpG hits. We could, however, not evaluate the applicability of our MRS for non-European populations.

Taking into account, the relatively small sample sizes used in the applied EWAS ranging from only 367 samples in discovery[6] to 3493 used for meta-analysis,[10] the portion of replicated signals in the LISA study (38.8% at 6 years) indicates a good replicability of allergy-related EWAS results. Further, our MRS performed well over all included EWAS, independent of variation in ethnicity or age, ranging from childhood to young adulthood. EWAS results of allergy seem to be rather similar across childhood, indicated by replication of signals at 6 years, although only one EWAS obtained results in participants younger than 6 years,[11] while the others were mostly older (Table 2) and similar patterns in both cross-sectional models.

Robustness of our findings was also confirmed across the different phenotypes used in the published allergy-related EWAS. Although main phenotypes were similar, as our outcome is a direct categorization of aeroallergen sensitization measurements, even broader ones like total IgE[7] or any allergic disease[11] result in the same patterns. Especially the similar findings across EWAS of different phenotypes, for example, sensitization vs. asthma, might hint in the direction of a general allergy phenotype,[49] also in terms of DNAm patterns. MRS were also associated with symptoms of allergic diseases in the LISA cohort, though associations with rhinoconjunctivitis and wheezing were weaker than for aeroallergen sensitization, likely due to the lower prevalence of these symptoms.

We recognize additional study limitations. We could not test the accuracy of MRS across different tissues (e.g., nasal epithelium), as

there are, to the best of our knowledge, no respective large-scale EWAS available for calculating further MRS. However, previous publications could replicate their signals from whole blood in other allergy-relevant tissues.[10,50] Additionally, we extracted gDNA from blood clots, whereas other studies used whole blood, so predictive accuracy of proposed MRS might be even higher using identical sample processing methodology. Our MRS approach uses CoMeBack to remove correlated CpGs located in co-methylated regions from the MRS. Future studies should evaluate if the prediction accuracy of MRS can be further improved by considering all correlations between CpGs instead of only those located in close proximity to each other and does not account for trans-chromosomal correlations. Absence of significant associations between aeroallergen sensitization and established predictors of allergic diseases in our sub-cohort indicate a limited statistical power due to our relatively small sample size for this analysis. Therefore, future studies with larger sample sizes are needed to replicate our findings. However, the strong associations and prediction accuracy that we found for the MRS despite our relatively small study sample also demonstrates the applicability of this approach for small study populations and the robustness of previously reported EWAS results.

Strengths of this study include the objective assessment of aeroallergen sensitization in blood. This makes all of our main associations robust as neither aeroallergen sensitization diagnosis nor DNAm, estimated cell type proportions or sex are subject to recall bias. Moreover, the prospective design of the LISA study enabled us to compare repeated measures at two time points of DNAm with three time points of measured aeroallergen sensitization.

In summary, we established well-working MRS for aeroallergen sensitization, which outperform commonly known determinants in identifying the disease. The presented results confirm the association of DNAm at some CpGs with allergic diseases and underline the relevance of altered gene-regulation in allergic diseases. The results support replication and applicability of available EWAS results and pave the way for future analyses investigating the specific functions between methylation patterns as biomarkers of disease manifestation.

FUNDING INFORMATION

The LISA study was mainly supported by grants from the Federal Ministry for Education, Science, Research and Technology and in addition from Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef for the first 2 years. The 4-year, 6-year, 10-year, and 15-year follow-up examinations of the LISA study were covered from the respective budgets of the involved partners (Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef, IUF—Leibniz-Research Institute for Environmental Medicine at the University of Düsseldorf) and in addition by a grant from the Federal Ministry for Environment (IUF Düsseldorf, FKZ 20462296). Further, the 15-year follow-up examination of the LISA study was supported by the Commission of the European Communities, the 7th Framework Program: MeDALL project.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 949906).

AK was supported by a research fellowship (No. 57504619) from the DAAD (German academic exchange service) to conduct this project with AH at Emory University. AH is supported by the HERCULES Center (NIEHS P30ES019776). HZ is funded by NIAID R01AI121226.

ACKNOWLEDGEMENTS

The authors thank all the families for their participation in the LISA study. Furthermore, we thank all members of the LISA Study Group for their excellent work. The LISA Study group consists of the following: Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology, Munich (Heinrich J, Schnappinger M, Brüske I, Ferland M, Schulz H, Zeller C, Standl M, Thiering E, Tiesler C, Flexeder C); Department of Pediatrics, Municipal Hospital "St. Georg", Leipzig (Borte M, Diez U, Dorn C, Braun E); Marien Hospital Wesel, Department of Pediatrics, Wesel (von Berg A, Berdel D, Stiers G, Maas B); Pediatric Practice, Bad Honnef (Schaaf B); Helmholtz Centre of Environmental Research—UFZ, Department of Environmental Immunology/Core Facility Studies, Leipzig (Lehmann I, Bauer M, Röder S, Schilde M, Nowak M, Herberth G, Müller J); Technical University Munich, Department of Pediatrics, Munich (Hoffmann U, Paschke M, Marra S); Clinical Research Group Molecular Dermatology, Department of Dermatology and Allergy, Technische Universität München (TUM), Munich (Ollert M, J. Grosch). We further want to thank Nadine Lindemann for her work analyzing the DNA methylation samples. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

Dr. Celedón has received research materials from Pharmavite, GSK, and Merck in order to provide medications free of cost to participants in NIH-funded studies, unrelated to the current work. All other authors have no further conflicts of interest to declare.

AUTHOR CONTRIBUTION

A.H., M.S., and A.K. conceived the idea for the presented study. A.K. conducted the analyses and wrote the initial paper draft. A.H. conceptualized the methodology and supervised the analyses. R.W. and N.G. helped with data processing of DNAm data. E.T and T.E. provided input on the statistical analysis. M.W. answered questions regarding laboratory sample processing. J.C. wrote the initial code for calculation of MRS and H.Z. J.C.C and E.B. provided further summary statistics on the used EWAS results. A.H., M.S., and A.P. jointly supervised the present project. All authors revised and commented on the final manuscript version.

DATA AVAILABILITY STATEMENT

Due to data protection reasons, the datasets generated and/or analyzed during the current study cannot be made publicly available. The datasets are available to interested researchers from the corresponding author on reasonable request, provided the release is consistent with the consent given by the LISA study participants. Ethical approval might be obtained for the release and a data transfer agreement from the legal department of the Helmholtz Zentrum München must be accepted.

ORCID

Anna Kilanowski https://orcid.org/0000-0001-5328-6274
Junyu Chen https://orcid.org/0000-0001-9659-0225
Todd Everson https://orcid.org/0000-0003-2732-4550
Elisabeth Thiering https://orcid.org/0000-0002-5429-9584
Nicole Gladish https://orcid.org/0000-0001-6039-7001
Hongmei Zhang https://orcid.org/0000-0003-3557-0364
Juan C. Celedón https://orcid.org/0000-0002-6139-5320
Annette Peters https://orcid.org/0000-0001-6645-0985
Marie Standl https://orcid.org/0000-0002-5345-2049
Anke Hüls https://orcid.org/0000-0002-6005-417X

REFERENCES

1. Ferreira MAR, Vonk JM, Baurecht H, et al. Eleven loci with new reproducible genetic associations with allergic disease risk. *J Allergy Clin Immunol*. 2019;143(2):691-699. doi:10.1016/j.jaci.2018.03.012
2. Nicolaou NC, Simpson A, Lowe LA, Murray CS, Woodcock A, Custovic A. Day-care attendance, position in sibship, and early childhood wheezing: a population-based birth cohort study. *J Allergy Clin Immunol*. 2008;122(3):500-506.e5. doi:10.1016/j.jaci.2008.06.033
3. Strachan DP. Hay fever, hygiene, and household size. *Br Med J*. 1989;299(6710):1259-1260. doi:10.1136/bmj.299.6710.1259
4. Chen CM, Rzehak P, Zutavern A, et al. Longitudinal study on cat allergen exposure and the development of allergy in young children. *J Allergy Clin Immunol*. 2007;119(5):1148-1155. doi:10.1016/j.jaci.2007.02.017
5. Long A, Bunning B, Sampath V, DeKruyff RH, Nadeau KC. Epigenetics and the environment in airway disease: asthma and allergic rhinitis. *Adv Exp Med Biol*. 2020;1253:153-181. doi:10.1007/978-981-15-3449-2_6
6. Everson TM, Lyons G, Zhang H, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med*. 2015;7(1):89. doi:10.1186/s13073-015-0213-8
7. Chen W, Wang T, Pino-Yanes M, et al. An epigenome-wide association study of total serum immunoglobulin E in Hispanic children. *J Allergy Clin Immunol*. 2017;140(2):571-577. doi:10.1016/j.jaci.2016.11.030
8. Peng C, Van Meel ER, Cardenas A, et al. Epigenome-wide association study reveals methylation pathways associated with childhood allergic sensitization. *Epigenetics*. 2019;14(5):445-466. doi:10.1080/15592294.2019.1590085
9. Zhang H, Kaushal A, Merid SK, et al. DNA methylation and allergic sensitizations: a genome-scale longitudinal study during adolescence. *Allergy*. 2019;74(6):1166-1175. doi:10.1111/all.13746
10. Reese SE, Xu CJ, den Dekker HT, et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. *J Allergy Clin Immunol*. 2019;143(6):2062-2074. doi:10.1016/j.jaci.2018.11.043
11. Xu CJ, Gruzieva O, Qi C, et al. Shared DNA methylation signatures in childhood allergy: the MeDALL study. *J Allergy Clin Immunol*. 2021;147(3):1031-1040. doi:10.1016/j.jaci.2020.11.044
12. Bin L, Leung DYM. Genetic and epigenetic studies of atopic dermatitis. *Allergy Asthma Clin Immunol*. 2016;12:52. doi:10.1186/s13223-016-0158-5
13. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z
14. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. *Epigenetics*. 2019;15:1-11. doi:10.1080/15592294.2019.1644879
15. Elliott HR, Tillin T, McArdle WL, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics*. 2014;6(1):4. doi:10.1186/1868-7083-6-4
16. Villanueva A, Portela A, Sayols S, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. *Hepatology*. 2015;61(6):1945-1956. doi:10.1002/hep.27732
17. Neste LV, Partin AW, Stewart GD, Epstein JI, Harrison DJ, Criekinge WV. Risk score predicts high-grade prostate cancer in DNA-methylation positive, histopathologically negative biopsies. *Prostate*. 2016;76(12):1078-1087. doi:10.1002/pros.23191
18. Heinrich J, Brüske I, Cramer C, et al. GINIplus and LISAplus. Design and selected results of two German birth cohorts about natural course of atopic diseases and its determinants. *Allergol Sel*. 2017;1(1):85-95. doi:10.5414/ALX01455E
19. Ansotegui IJ, Melioli G, Canonica GW, et al. IgE allergy diagnostics and other relevant tests in allergy, a World Allergy Organization position paper. *World Allergy Organ J*. 2020;13(2):100080. doi:10.1016/j.waojou.2019.100080
20. Hoffman DR. Comparison of methods of performing the radioallergosorbent test: Phadebas, Fadal-Nalebuff and Hoffman protocols. *Ann Allergy*. 1980;45(6):343-346.
21. Hu C, Duijts L, Erler NS, et al. Most associations of early-life environmental exposures and genetic risk factors poorly differentiate between eczema phenotypes: the Generation R Study. *Br J Dermatol*. 2019;181(6):1190-1197. doi:10.1111/bjd.17879
22. Paternoster L, Savenije OEM, Heron J, et al. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. *J Allergy Clin Immunol*. 2018;141(3):964-971. doi:10.1016/j.jaci.2017.09.044
23. Luzak A, Fuertes E, Flexeder C, et al. Which early life events or current environmental and lifestyle factors influence lung function in adolescents? – Results from the GINIplus & LISAplus studies. *Respir Res*. 2017;18:138. doi:10.1186/s12931-017-0619-5
24. Ferreira MA, Vonk JM, Baurecht H, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet*. 2017;49(12):1752-1757. doi:10.1038/ng.3985
25. El-Husseini ZW, Gosens R, Dekker F, Koppelman GH. The genetics of asthma and the promise of genomics-guided drug target discovery. *Lancet Respir Med*. 2020;8(10):1045-1056. doi:10.1016/S2213-2600(20)30363-5
26. Paternoster L, Standl M, Waage J, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet*. 2015;47(12):1449-1456. doi:10.1038/ng.3424
27. Waage J, Standl M, Curtin JA, et al. Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet*. 2018;50(8):1072-1080. doi:10.1038/s41588-018-0157-1
28. Granada M, Wilk JB, Tuzova M, et al. A genome-wide association study of plasma total IgE concentrations in the Framingham heart study. *J Allergy Clin Immunol*. 2012;129(3):840-845.e21. doi:10.1016/j.jaci.2011.09.029

29. Fortin JP, Labbe A, Lemire M, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*. 2014;15(12):503. doi:10.1186/s13059-014-0503-2
30. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat Oxf Engl*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037
31. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86. doi:10.1186/1471-2105-13-86
32. Salas LA, Koestler DC, Butler RA, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol*. 2018;19:64. doi:10.1186/s13059-018-1448-7
33. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*. 2017;18:105. doi:10.1186/s12859-017-1511-5
34. Hannon E, Dempster E, Viana J, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol*. 2016;17(1):176. doi:10.1186/s13059-016-1041-x
35. Gatev E, Gladish N, Mostafavi S, Kobor MS. CoMeBack: DNA methylation array data analysis for co-methylated regions. *Bioinforma Oxf Engl*. 2020;36(9):2675-2683. doi:10.1093/bioinformatics/btaa049
36. North ML, Jones MJ, MacIsaac JL, et al. Blood and nasal epigenetics correlate with allergic rhinitis symptom development in the environmental exposure unit. *Allergy*. 2018;73(1):196-205. doi:10.1111/all.13263
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.
38. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2019. https://www.R-project.org/.
39. Martin LB, Kita H, Leiferman KM, Gleich GJ. Eosinophils in allergy: role in disease, degranulation, and cytokines. *Int Arch Allergy Immunol*. 1996;109(3):207-215. doi:10.1159/000237239
40. Schäfer T, Bauer CP, Beyer K, et al. S3-Guideline on allergy prevention: 2014 update. *Allergo J Int*. 2014;23(6):186-199. doi:10.1007/s40629-014-0022-4
41. Dijk FN, Folkersma C, Gruzieva O, et al. Genetic risk scores do not improve asthma prediction in childhood. *J Allergy Clin Immunol*. 2019;144(3):857-860.e7. doi:10.1016/j.jaci.2019.05.017
42. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868
43. Kitamura N, Mori A, Tatsumi H, Nemoto S, Hiroi T, Kaminuma O. Zinc finger protein, multitype 1, suppresses human Th2 development via downregulation of IL-4. *Int Arch Allergy Immunol*. 2011;155(Suppl 1):53-56. doi:10.1159/000327292
44. Ferreira GB, Overbergh L, van Etten E, et al. Protein-induced changes during the maturation process of human dendritic cells: a 2-D DIGE approach. *PROTEOMICS – Clin Appl*. 2008;2(9):1349-1360. doi:10.1002/prca.200800110
45. Zhong J, Wang H, Chen W, et al. Ubiquitylation of MFHAS1 by the ubiquitin ligase praja2 promotes M1 macrophage polarization by activating JNK and p38 pathways. *Cell Death Dis*. 2017;8(5):e2763. doi:10.1038/cddis.2017.102
46. Laffleur B, Duchez S, Tarte K, et al. Self-Restrained B cells arise following membrane IgE expression. *Cell Rep*. 2015;10(6):900-909. doi:10.1016/j.celrep.2015.01.023
47. Min JL, Hemani G, Hannon E, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. 2021;53(9):1311-1321. doi:10.1038/s41588-021-00923-x
48. Felix JF, Joubert BR, Baccarelli AA, et al. Cohort profile: pregnancy and childhood epigenetics (PACE) consortium. *Int J Epidemiol*. 2018;47(1):22-23u. doi:10.1093/ije/dyx190
49. Anto JM, Bousquet J, Akdis M, et al. Mechanisms of the development of allergy (MeDALL): introducing novel concepts in allergy phenotypes. *J Allergy Clin Immunol*. 2017;139(2):388-399. doi:10.1016/j.jaci.2016.12.940
50. Hoang TT, Sikdar S, Xu CJ, et al. Epigenome-wide association study of DNA methylation and adult asthma in the agricultural lung health study. *Eur Respir J*. 2020;56(3):2000217. doi:10.1183/13993003.00217-2020

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kilanowski A, Chen J, Everson T, et al. Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort. *Allergy*. 2022;77:2803–2817. doi:10.1111/all.15315

3 Publication: DNA methylation and aeroallergen sensitization - the chicken or the egg?

Title:	DNA methylation and aeroallergen sensitization - the chicken or the egg?
Authors:	Anna Kilanowski, Simon K. Merid, Sarina Abrishamcar, Dakotah Feil, Elisabeth Thiering, Melanie Waldenberger, Erike Melén, Annette Peters, Marie Standl, Anke Hüls
Journal:	*Clinical Epigenetics*
Year:	2022
Volume:	14
DOI:	10.1186/s13148-022-01332-5
Supplements:	https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-022-01332-5 (available online)
Impact factor:	7.291 (Journal Citation Reports™ 2021)
Rank:	19/175 in category GENETICS & HEREDITY (Journal Citation Reports™ 2021)

Kilanowski *et al. Clinical Epigenetics* (2022) 14:114
https://doi.org/10.1186/s13148-022-01332-5

Clinical Epigenetics

RESEARCH Open Access

DNA methylation and aeroallergen sensitization: The chicken or the egg?

Anna Kilanowski[1,2,3,4], Simon Kebede Merid[5], Sarina Abrishamcar[3], Dakotah Feil[3], Elisabeth Thiering[1,4], Melanie Waldenberger[1,6], Erik Melén[5], Annette Peters[1,7], Marie Standl[1,8*†] and Anke Hüls[3,9*†]

Abstract

Background: DNA methylation (DNAm) is considered a plausible pathway through which genetic and environmental factors may influence the development of allergies. However, causality has yet to be determined as it is unknown whether DNAm is rather a cause or consequence of allergic sensitization. Here, we investigated the direction of the observed associations between well-known environmental and genetic determinants of allergy, DNAm, and aeroallergen sensitization using a combination of high-dimensional and causal mediation analyses.

Methods: Using prospectively collected data from the German LISA birth cohort from two time windows (6–10 years: $N = 234$; 10–15 years: $N = 167$), we tested whether DNAm is a cause or a consequence of aeroallergen sensitization (specific immunoglobulin E > 0.35kU/l) by conducting mediation analyses for both effect directions using maternal smoking during pregnancy, family history of allergies, and a polygenic risk score (PRS) for any allergic disease as exposure variables. We evaluated individual CpG sites (EPIC BeadChip) and allergy-related methylation risk scores (MRS) as potential mediators in the mediation analyses. We applied three high-dimensional mediation approaches (HIMA, DACT, gHMA) and validated results using causal mediation analyses. A replication of results was attempted in the Swedish BAMSE cohort.

Results: Using high-dimensional methods, we identified five CpGs as mediators of prenatal exposures to sensitization with significant (adjusted $p < 0.05$) indirect effects in the causal mediation analysis (maternal smoking: two CpGs, family history: one, PRS: two). None of these CpGs could be replicated in BAMSE. The effect of family history on allergy-related MRS was significantly mediated by aeroallergen sensitization (proportions mediated: 33.7–49.6%), suggesting changes in DNAm occurred post-sensitization.

Conclusion: The results indicate that DNAm may be a cause or consequence of aeroallergen sensitization depending on genomic location. Allergy-related MRS, identified as a potential cause of sensitization, can be considered as a cross-sectional biomarker of disease. Differential DNAm in individual CpGs, identified as mediators of the development of sensitization, could be used as clinical predictors of disease development.

†Marie Standl and Anke Hüls contributed equally to this work and shared last authorship

*Correspondence: marie.standl@helmholtz-muenchen.de; anke.huels@emory.edu

[1] Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[3] Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd NE, Atlanta, GA 30322, USA
Full list of author information is available at the end of the article

Keywords: High-dimensional mediation analysis, DNA methylation, Allergic diseases, Epidemiology, Methylation risk scores, Polygenic risk scores, Maternal smoking

Background

With the rise of available DNA methylation (DNAm) data in multiple cohort studies, the number of epigenome-wide association studies (EWAS) demonstrating a connection between DNAm and allergic diseases has increased. Over the last decade, EWAS reported associations of single CpGs (addition of a methyl-group to a cytosine in the context of CpG dinucleotides) with several allergic outcomes: High total immunoglobulin E (IgE) [1, 2], an antibody involved in Type I immune response and highly associated with allergic diseases, specific IgE [3] against certain aeroallergens and specific IgE plus skin-prick test [4] and meta-analyses on asthma [5] and any allergic disease [6]. Many of these CpGs have been successfully replicated in independent cohorts, and we could verify the robustness of these findings via replication of significant hits in the German LISA study [7].

However, it is unknown whether DNAm changes occur in response to allergic disease or if differential DNAm can serve as predictor of future development of allergies. Looking at aeroallergen sensitization, an objectively measured indicator of allergic diseases, we previously reported that methylation risk scores (MRS), which are defined as a weighted sum of methylation beta estimates, can be considered as cross-sectional biomarkers of current sensitization [7]. However, the predictive capabilities in prospective associations with aeroallergen sensitization were limited, indicating that DNAm might be a result rather than a predictor of allergic sensitization. On the other hand, studies investigating DNAm in cord-blood found associations with higher IgE levels later in life [8, 9], indicating a certain predictive potential.

One way to investigate this "chicken or egg—what came first?" question is a causal mediation analysis with data on exposure, mediator and outcome from three subsequent time points. Known determinants of allergic disease that can be used as exposures in such mediation analyses include genetic and environmental factors. Allergic diseases are highly heritable, with heritability estimates for allergic diseases being described as high as 91.7% for asthma [10], 90% for atopic dermatitis [11], 91% for allergic rhinitis and 68% for specific serum IgE (reviewed in Ober and Yao [12]). Additionally, numerous genetic variants associated with allergic diseases have been identified in multiple genome-wide association studies (GWAS), e.g., for atopic dermatitis [13], rhinitis [14] or any allergic disease [15, 16]. Polygenic risk scores (PRS) have been proposed to summarize genetic susceptibility to allergic diseases in one score for allergic trajectories [17] or asthma prediction [18, 19], presenting a significant association and a predictive area-under-the-curve of up to 0.59 for early transient asthma phenotypes and 0.58 for intermediate-onset wheeze [18]. However, as genetic variation in complex diseases represents a risk increase but not a certainty of disease onset as in monogenic diseases, family history of allergic diseases can be additionally considered as a proxy for the combination of allergic inheritance and environmental risk.

Further looking at environmental risk factors, maternal smoking during pregnancy represents a well-established environmental risk factor, which has been shown to influence allergic outcomes, especially asthma [20], and has also been biologically validated in preclinical mouse models [21].

A methodological challenge of investigating the "chicken or egg" question in causal mediation analyses is the high-dimensionality of DNAm data with up to 850K CpG sites being measured with the most recent Illumina DNAm arrays (Illumina MethylationEPIC BeadChip microarray). Several approaches have been proposed to address high-dimensionality in mediation analysis including (1) dimension-reduction methods, e.g., by using MRS, (2) integration of prior knowledge by only focusing on CpG sites with a known association with the exposure or outcome (or both) and (3) hypothesis-generating high-dimensional mediation analyses (HMA).

The objective of this study is to determine the causality of the observed associations between changes in DNAm and the development of allergen sensitization using HMA and MRS. We conduct different HMA at two subsequent time points using well-established determinants of allergic disease (maternal smoking during pregnancy, family history of allergies and a PRS for any allergic disease) as exposures and prospective measurements of DNAm and aeroallergen sensitization as mediators and outcomes.

Methods

Study population

For this study, we used data from a population-based German birth cohort on the Influence of **L**ife-style factors on Development of the **I**mmune **S**ystem and **A**llergies in East and West Germany (LISA). From 1997 to 1999, a total of 3,097 full-term healthy newborns were recruited at four study centers (Munich, Wesel, Leipzig and Bad Honnef). The study was approved by local ethics committees (Bavarian Board of Physicians, Board of

Physicians of North-Rhine-Westphalia and Medical Faculty of the University of Leipzig) and written, informed consent was obtained from the parents or legal guardians. In the present study, only data from participants enrolled in the Munich study center with parental consent for genetic analyses at both six and ten years is included ($N_{max} = 240$).

Aeroallergen sensitization

Positive aeroallergen sensitization was defined as a specific IgE threshold of > 0.35 kU/L (at least Radio-Allergo-Sorbent-Test (RAST) class one), measured for a mix of common aeroallergens (SX1 mix: Dermatophagoides pteronyssinus, cat, dog, rye, timothy grass, Cladosporium herbarum, birch and mugwort). Serum at six, ten and 15 years was analyzed using the CAP-RAST FEIA system (Pharmacia Diagnostics, Freiburg, Germany) according to the manufacturer's instructions.

Risk factors for aeroallergen sensitization

Genome-wide data in the LISA study were measured using the Affymetrix Chip 5.0 and 6.0 (Thermo Fisher Scientific, USA). More information on genetic data can be found in the supplementary material of Grosche et al. [22]. We calculated a PRS for any allergic disease based on the genome-wide significant hits reported in Ferreira et al. [15, 16]. Single nucleotide polymorphisms (SNPs) were extracted for each participant and weighted with the reported effect size. Multiallelic SNPs, highly correlated variants (Linkage disequilibrium R2 > 0.7), those with a low imputation quality (< 0.4) or a minor allele frequency of less than 1% were excluded. Further information on quality control and PRS calculation can be found elsewhere [7, 23].

Information on family history of allergic diseases was collected at birth and defined as a binary factor indicating no family history or at least one biological parent reporting ever experiencing asthma, atopic dermatitis or hay fever.

Maternal smoking during pregnancy was defined as smoking in the second and/or third trimester of pregnancy, with controls defined as either stopped smoking before the second trimester or never smoking. Potential confounders after literature research are sex, age, season at blood withdrawal, cell-type proportions, Body-Mass-Index (BMI), socio-economic status (SES), and air pollution, defined as nitrogen dioxide (NO_2) at birth address (Additional file 2: Table S1).

DNAm data

DNAm was measured for 256 participants from blood clots taken at six and ten years using the Methylation EPIC BeadChip (Illumina, Inc., San Diego, CA). We applied functional normalization [24] and ComBat [25] to normalize the data and remove technical variation. Probes were removed if they were located on the sex chromosomes, had missing values, or failed the detection p value of 0.01 in more than 1% of samples. Samples were removed if they were outliers, sex mismatches, or did not fulfill the bad-sample threshold of methylated and unmethylated intensities. Cell-type proportions were estimated using the *EpiDISH* package [26]. Further information on quality control and data processing can be found elsewhere [7].

Methylation risk scores

MRS were calculated for six allergy-related EWAS, namely high IgE [2], aeroallergen sensitization [3], asthma [5], any allergic disease [6] and two on atopy, defined as high total IgE [1] or positive specific IgE as well as a positive skin-prick test [4]. Details on the calculation and evaluation of these allergy-related MRS have been published previously [7]. In short, we calculated each MRS by weighting the CpG beta-values with the respective effect size identified by the EWAS and transformed to z-scores. The selection of CpG sites was conducted using a pruning and thresholding approach [27]. As described previously [7], the MRS that reached the highest prediction accuracy for allergic sensitization at six years of age across all p-value thresholds was used in the downstream analyses.

Statistical analysis

To evaluate whether changes in DNAm are predictors or consequences of allergic diseases, we tested the following two hypotheses: (H1, DNAm as predictor) The association between exposure (maternal smoking during pregnancy; family history of allergic disease; PRS for any allergies) and allergic sensitization is mediated by prior changes in DNAm (measured by MRS or methylation in individual CpG sites); (H2, DNAm as consequence) The association between exposure (maternal smoking during pregnancy; family history of allergic disease; PRS for any allergies) and changes in DNAm (measured by MRS or methylation in individual CpG sites) is mediated by prior allergic sensitization. In our main analyses, mediators were measured at six years and outcomes at ten years, both for hypothesis (H1) and (H2). In addition, we conducted a secondary analysis for hypothesis (H1), in which mediators (DNAm) were measured at ten years and outcome (aeroallergen sensitization) at 15 years (Fig. 1 and Additional file 1: Figure S1).

Mediation analyses rely on the following three assumptions [28]: (1) no exposure-mediator confounding, (2) no mediator-outcome confounding and (3) no exposure-outcome confounding. To fulfill these assumptions to

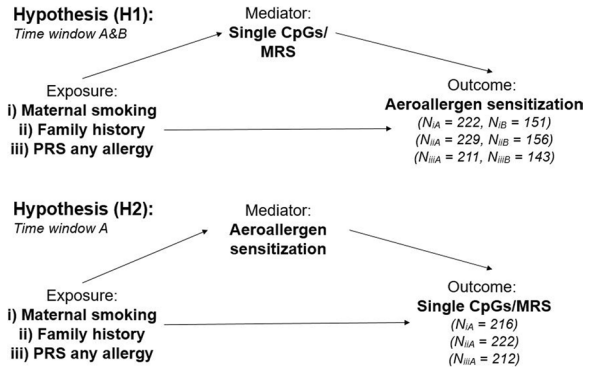


Fig. 1 Display of models used for the identification and validation of potential mediators. Hypothesis (H1) describe the mediation of aeroallergen sensitization through DNAm and hypothesis (H2) the reversed direction that sensitization is mediating DNAm changes. Time window A covers the development from six to ten years and time window B from ten to 15 years. See also Additional file 1: Figure S1

the best of our knowledge, we constructed directed acyclic graphs (DAGs) to visualize each of these paths using *dagitty* [29] (Additional file 1: Figures S2–S9). A minimal sufficient adjustment set was identified for each pathway via the tracing of association directions and elimination of any potential confounders already associated with a precursory confounder. Exposure-mediator models were adjusted for SES (Exposure: maternal smoking during pregnancy), SES and NO_2 exposure at birth (family history) and sex (PRS) for both hypotheses. Mediator-Outcome models were adjusted for all potential confounders according to the DAGs (Additional file 1: Figures S2–S9). A detailed description of the definition and assessment of these covariates is provided in Additional file 1: Table S1.

Associations with continuous outcomes (MRS or DNAm in individual CpG sites) were analyzed using linear regression and associations with binary outcomes (allergic sensitization) were analyzed using logistic regression.

Causal mediation analysis of MRS

Causal mediation analysis, using the R package *mediation* [30], was applied to test the two hypotheses (H1) and (H2) for allergy-related MRS. Results were adjusted for multiple testing using the Benjamini–Hochberg procedure [31] for false-discovery rate (FDR) together within each H1 and H2.

High-dimensional mediation analysis of individual CpGs

High-dimensional mediation analyses (HMA) were used to test the two hypotheses for individual CpGs. H1 was tested using the Divide-Aggregate Composite-Null test (DACT), HIMA, and gene-based HMA (gHMA). H2 was tested using only DACT, because HIMA and gHMA are only applicable for high-dimensional mediators but not for high-dimensional outcomes.

1. Previous knowledge + Divide-Aggregate Composite-Null test (DACT)

   Based on previously published EWAS of total IgE [1, 2], aeroallergen sensitization [3, 4], childhood asthma [5] and any allergic disease [6] we used existing knowledge on allergy-relevant CpGs to reduce the multiple testing burden. Of the 1673 previously reported CpGs, 1501 were available in the LISA cohort and 583 CpGs were significantly associated with aeroallergen sensitization in the LISA cohort at six years [7] (False discovery rate ≤ 0.05; adjustment for Houseman cell -type estimates to resemble the initial discovery analyses), which were further taken as testing-set of potential mediators. Of note, none of these CpGs were significantly associated with any of the exposures after multiple testing correction and adjustment for sex, detailed age and *EpiDISH* cell-type estimates (Additional file 2: Table S2).

   We used DACT for the composite null hypothesis of no mediation effect as suggested by Liu et al. [32] to improve the multiple testing burden. In short, DACT takes the p values from the exposure-mediator and the mediator-outcome model to compute a new joint list of p-values, which will be used to determine significance (p-value < 0.05). This is done by aggregating the weighted p -values of the three possible null-hypotheses leading to no mediation effect and calibrating this using Efron's empirical null framework [33].
2. HIMA
   Whereas the previous approach relied on existing knowledge as a baseline selection of mediators, HIMA as proposed by Zhang et al. [34] uses a three step procedure to identify significant CpGs throughout the whole epigenome. First, the top CpGs with the largest effect sizes (beta of standardized inputs) for the response variable are identified using sure independence screening (SIS) [35]. The total number of top hits (N) varies per model and is calculated by $N = 2^{*}n/\log(n)$, with n being the input sample size. To capture relevant CpGs with our smaller sample size, we applied a looser threshold than the original publication. In a second step, HIMA estimates the mediation effect using minimax concave penalty and performs joint significance testing as a third and final step.

3. Gene-based HMA (gHMA)
   We further applied gene-based high-dimensional mediation analysis (gHMA) as proposed by Fang et al. [36]. The idea behind this approach is that not single CpGs but genes act as biological units and should therefore be analyzed together. The functions further provide different modeling options for linear or nonlinear relationships and an omnibus-test to combine both, which outperformed the single models in their simulation study. First, we annotated every CpG to their nearest gene within 20,000 base pairs as done previously [37], resulting in 40,916 different genes. We then applied gHMA to each of these 40,916 genes, each covering between one and 1758 CpGs, performing the linear, nonlinear and omnibus-test for significance. We used differing kernel-thresholds of 0.7, 0.8 and 0.9 as values for explained variance by the kernel principal components. Results of the omnibus-test were corrected using the Benjamini–Hochberg procedure [31].

Validation of CpG sites using causal mediation analysis

All significant CpG sites identified with the three described methods above are followed up using a causal mediation analysis to determine the direct, indirect, and total effects as well as the proportion mediated. Multiple testing correction followed the one applied for the MRS evaluation by calculating the FDR for all H1 CpGs together, the same correction was applied for H2 CpGs. Models and adjustment are the same as for MRS analyses and single CpGs were afterwards annotated using mQTL databases provided by Gaunt and Hawe et al. [38, 39].

Sensitivity analyses

We conducted a set of sensitivity analyses to evaluate the robustness of associations for any CpG sites that were successfully validated in the causal mediation analysis described above.

First, to further evaluate the impact of differences in cell-type proportions on our findings, we conducted a sensitivity analysis in which we additionally adjusted all exposure-mediator associations for estimated cell types, which are otherwise only included in the mediator-outcome associations.

Second, to focus exclusively on newly developed aeroallergen sensitization in our mediation analyses with aeroallergen sensitization as outcome, we conducted a sensitivity analysis in which we excluded individuals already sensitized at baseline DNAm measurement.

Third, we conducted sex-stratified analyses, as puberty may play a role in allergen sensitization [40].

Replication of potential mediators

Single CpGs moving forward to validation in causal mediation analysis was further replicated in the independent Swedish BAMSE (Swedish abbreviation for Children, Allergy, Milieu, Stockholm, Epidemiology) cohort, which recruited 4093 newborns between 1994 and 1996. Ethical approval was given by the Regional Ethics Board (EPN) and further information is available elsewhere [41]. Here, we used exposure data from birth (maternal smoking in second and/or third trimester of pregnancy, any family history of allergic diseases and the same calculated PRS for any allergic disease [7, 23]), DNAm data measured at eight years of age with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA) [6] and outcome data (positive aeroallergen sensitization to the SX1 mix) from 16 years. Further information on genetic and DNAm data can be found in Additional file 1: Methods S1.

All analyses were performed in R [42] V.4.1.2 in LISA and V.4.1.3 in BAMSE.

Results

The total sample size for the six different models and time windows, from six to ten years (A) and from ten to 15 years (B), varied from 143 to 229, only including participants, who had all necessary data available (respective exposure, DNAm and covariates) (Fig. 1 and Additional file 1: Figure S1). Participants in the overall sample for all models were majority male (57.7%) and their blood samples were collected primarily during the allergy season from March to August. Prevalence of aeroallergen sensitization increased from baseline to follow-up in each time window and missing values for exposures ranged from six (maternal smoking) to twelve missing values in the PRS (Table 1).

Causal mediation analysis for MRS

Allergy-related MRS were not found to be a mediator of the association between family history of allergies and subsequent allergic sensitization (H1, Fig. 2A). However, we found significant indirect effects for the association between family history of allergies and all six allergy-related MRS with prior allergic sensitization as mediators (H2) (e.g., Indirect effect (Chen2017) = 0.081 [0.020; 0.160]). Proportion mediated by allergic sensitization ranged from 33.7% (Everson2015) to 49.6% (Zhang2019) (Table 2 and Fig. 2B). Results were robust to additional adjustment for cell-type estimates as exposure-mediator confounders in our sensitivity analysis (Additional file 2:

Table 1 Description of total sample of LISA participants included in this study

	Time window A (6 and 10 years)	Time window B (10 and 15 years)
Total sample size—*N*	234	227
Confounder		
Male sex—*N* (%) [Nmiss]	135 (57.7%) [0]	131 (57.7%) [0]
Exact age **at baseline**—Mean (sd) [Nmiss]	6.1 (0.2) [0]	10.2 (0.1) [0]
Exact age **at follow-up**—Mean (sd) [Nmiss]	10.2 (0.1) [0]	15.2 (0.2) [52]
Blood taken in allergy season **at baseline**—*N* (%) [Nmiss]	158 (67.5%) [0]	116 (51.1%) [0]
Blood taken in allergy season **at follow-up**—*N* (%) [Nmiss]	124 (53%) [0]	90 (51.4%) [52]
BMI measured **at baseline**—Mean (sd) [Nmiss]	15.3 (1.3) [1]	17.0 (2.5) [1]
BMI measured **at follow-up**—Mean (sd) [Nmiss]	16.9 (2.5) [1]	20.3 (2.8) [59]
High parental education—*N* (%) [Nmiss]	186 (80.2%) [2]	181 (80.4%) [2]
NO_2 at birth address—Mean (sd) [Nmiss]	21.4 (6.1) [1]	21.2 (5.5) [1]
Exposures		
(i) Maternal smoking during pregnancy—*N* (%) [Nmiss]	17 (7.5%) [7]	16 (7.2%) [6]
(ii) At least one parent allergic—*N* (%) [Nmiss]	151 (64.5%) [0]	148 (65.2%) [0]
(iii) PRS—Mean (sd) [Nmiss]	0.2 (1.0) [12]	0.2 (0.9) [11]
Outcome		
Sensitized **at baseline**—*N* (%) [Nmiss]	74 (31.6%) [0]	101 (44.5%) [0]
Sensitized **at follow-up**—*N* (%) [Nmiss]	105 (44.9%) [0]	84 (50.3%) [60]

Baseline is defined as the first time point of the model (six or ten) and follow-up as the second (ten or 15, respectively). The sample sizes for the mediation models with different exposures were as followed: Hypothesis 1A—$N_{MaternalSmoking} = 215$, $N_{FamilyHistory} = 198$, $N_{PRS} = 211$; Hypothesis 1B—$N_{MaternalSmoking} = 163$, $N_{FamilyHistory} = 154$, $N_{PRS} = 158$; Reversed models for Hypothesis 2A (Exposure–Sensitization–DNAm)—$N_{MaternalSmoking} = 216$, $N_{FamilyHistory} = 222$, $N_{PRS} = 212$. See also Additional file 1: Figure S1

Table S3 and S4), while keeping the mediator-outcome confounders, including cell-type estimates, consistent.

We did not find any significant mediation effects for maternal smoking during pregnancy or the PRS for either of the two hypotheses. Full results for all MRS models can be found in Additional file 2: Tables S5 (H1) and S6 (H2) for the time window from six to ten only, as DNAm as an outcome was not measured at 15 years of age.

DACT

We identified 90 unique CpGs as potential mediators (H1) with the DACT approach: For the first time window (A) from six to ten years, we found 18 CpGs for maternal smoking, 51 for family history and six for the PRS. For the second time window from ten to 15 the numbers were 20, 19 and ten, respectively. Of all of these, only one CpG (cg26851984) was validated in causal mediation analyses (significant indirect effect after multiple-testing correction), for time window A and maternal smoking as exposure (Table 3). Differential DNAm at cg26851984 mediates 81% of the association between maternal smoking and aeroallergen sensitization and is robust to additional adjustment for cell-type estimates of the exposure-mediator association. Of note, cg26851984 is also an mQTL with 58 surrounding SNPs as reported in a recent publication by Hawe et al. [39]. A mediation plot for cg26851984 is presented in Fig. 3 (first panel) showing the validated associations with the single CpG as mediators.

In the reversed models investigating sensitization as a potential mediator of subsequent changes in DNAm (H2), we did not identify any mediation effects for individual CpGs in either main model (Additional file 2: Table S7).

HIMA

Dependent on the sample size of the different exposures and time windows, between 58 and 85 CpGs ($N = 2^*n/\log(n)$; Fig. 1) were screened for highest effect sizes during the first step of HIMA and had their estimates calculated and tested for joint significance in HIMA in different models. We identified three CpGs as potential mediators in the time window from six to ten years (time window A), one CpG of the association between each exposure and aeroallergen sensitization. In addition, we identified four CpGs as mediators in the later time window (B) from ten to 15 years, three for PRS as exposure and one for family history (Additional file 2: Tables S7 for full results and S8 for annotated hits). Four of the seven identified CpGs were significantly validated in the causal mediation analysis and none are located in mQTLs (Table 3; Fig. 3 (panels 2–5)).

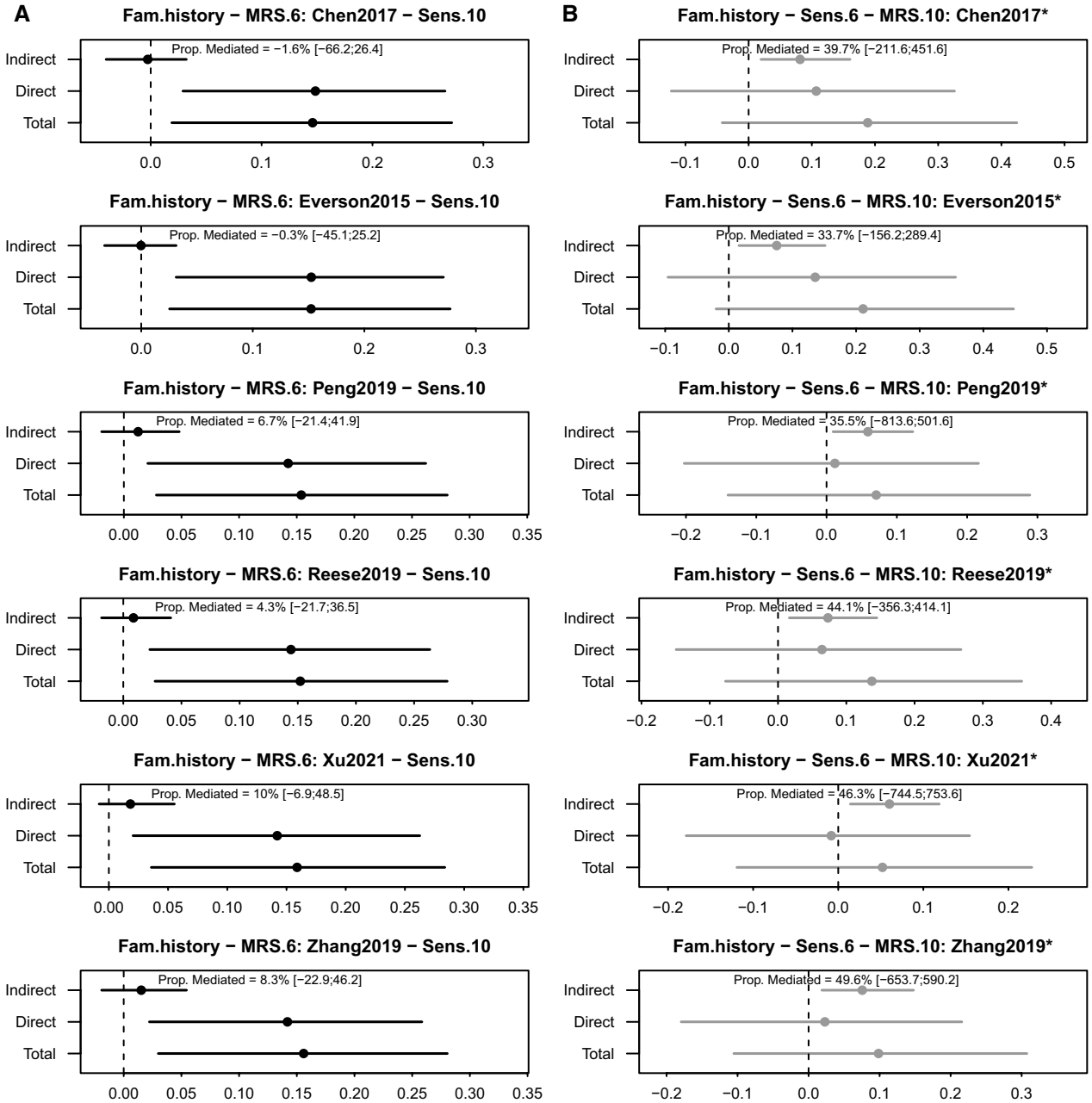


Fig. 2 MRS as predictor or consequence of allergic disease. Significant indirect effects are indicated with an asterisk. The title follows the pattern exposure–mediator–outcome. Evaluation **A** whether the association between family history of allergic disease and allergic sensitization at ten years is mediated by prior changes in DNAm at six years (measured by MRS) or **B** whether the association between family history of allergic disease and changes in DNAm at ten years (measured by MRS) is mediated by prior allergic sensitization at six years. The six MRS can be allocated to the following phenotypes: Chen2017—total IgE, Everson2015—atopy, Peng2019—aeroallergen sensitization, Reese2019—childhood asthma, Xu2021—any allergy and Zhang2019—atopy, respectively

Sensitivity analyses

All CpGs presented in Table 3 showed nominal significant associations after additional adjustment for cell-type proportions between exposure and mediator (Additional file 2: Table S9) and when restricting the analysis sample to those who were not sensitized

Table 2 Significant mediation (FDR < 0.05) between family history as exposure and MRS, mediated by aeroallergen sensitization measured (H2)

Outcome	Indirect effect [95% CI]	Direct effect [95% CI]	Total effect [95% CI]	Prop.Med [95% CI]
(A) Sensitization at 6 and DNAm measured at 10 years				
(ii) Family history of allergic disease (both parents allergic)				
Chen2017 *(IgE)*	0.081 [0.020; 0.160]	0.107 [− 0.122; 0.326]	0.188 [− 0.042; 0.425]	0.397 [− 2.116; 4.516]
Everson2015 *(Atopy)*	0.075 [0.016; 0.151]	0.136 [− 0.096; 0.356]	0.211 [− 0.020; 0.447]	0.337 [− 1.562; 2.894]
Peng2019 *(Aeroallergen)*	0.059 [0.009; 0.123]	0.012 [− 0.202; 0.216]	0.071 [− 0.140; 0.289]	0.355 [− 8.136; 5.016]
Reese2019 *(Asthma)*	0.073 [0.017; 0.145]	0.064 [− 0.149; 0.268]	0.137 [− 0.077; 0.357]	0.441 [− 3.563; 4.141]
Xu2021 *(Allergy)*	0.060 [0.014; 0.119]	− 0.008 [− 0.179; 0.154]	0.052 [− 0.119; 0.227]	0.463 [− 7.445; 7.536]
Zhang2019 *(Atopy)*	0.076 [0.019; 0.148]	0.023 [− 0.179; 0.216]	0.099 [− 0.105; 0.307]	0.496 [− 6.537; 5.902]

No significant associations were found for (i) maternal smoking or (iii) PRS for any allergies

Table 3 DNAm in individual CpG sites as predictors of aeroallergen sensitization (H1). Displayed CpGs were significantly validated in the causal mediation analysis (FDR < 0.05)

HMA Method	Mediator [CpG (UCSC/ nearest gene—UCSC Group)]	Indirect effect [95% CI]	Direct effect [95% CI]	Total effect [95% CI]	Prop.Med [95% CI]	mQTL (Hawe et al.[a])
A. DNAm at six years and sensitization measured at ten years						
(i) Maternal smoking during pregnancy						
DACT	cg26851984 *(/RP11-772E11.1)*	0.139 [0.050; 0.242]	− 0.025 [− 0.226; 0.182]	0.114 [− 0.124; 0.328]	0.811 [− 7.782; 8.513]	58
HIMA	cg17992705 *(ATXN2L - ExonBnd)*	− 0.108 [− 0.193; − 0.041]	0.217 [0.017; 0.397]	0.109 [− 0.102; 0.315]	− 0.601 [− 11.384; 14.469]	
(ii) Family history of allergic disease (both parents allergic)						
HIMA	cg11329030 *(/ATP6V1E1P1)*	0.095 [0.040; 0.158]	0.079 [− 0.025; 0.193]	0.174 [0.060; 0.280]	0.547 [0.218; 1.349]	
(iii) PRS for any allergies						
HIMA	cg04684486 *(SLC31A2 - TSS200)*	0.020 [0.004; 0.042]	0.019 [− 0.043; 0.082]	0.040 [− 0.026; 0.103]	0.390 [− 2.851; 6.302]	
B. DNAm at ten years and sensitization measured at 15 years						
(iii) PRS for any allergies						
HIMA	cg19310430 *(C11orf45 - 5'UTR)*	0.063 [0.026; 0.105]	0.036 [− 0.036; 0.110]	0.099 [0.019; 0.179]	0.632 [0.208; 1.910]	

CpG sites that were identified as mediators in at least one high-dimension mediation analysis (HMA) method (HIMA or DACT) and validated in causal mediation analysis (significant indirect effect) are presented.

[a] No mQTLs from Gaunt et al. [38] were matched to the respective CpGs

at the time of DNAm measurement (Additional file 2: Table S10). However, those associations were not significant after adjustment for multiple testing. We did not find sex-specific differences in mediation effects in terms of effect estimates and direction of effects, but indirect effects were only significant for three of the five CpGs in males (Additional file 2: Table S11) and for none of the CpGs in females (Additional file 2: Table S12), most likely due to the reduced sample size.

gHMA

We did not identify any significant genes for either time window or exposure with the gHMA method.

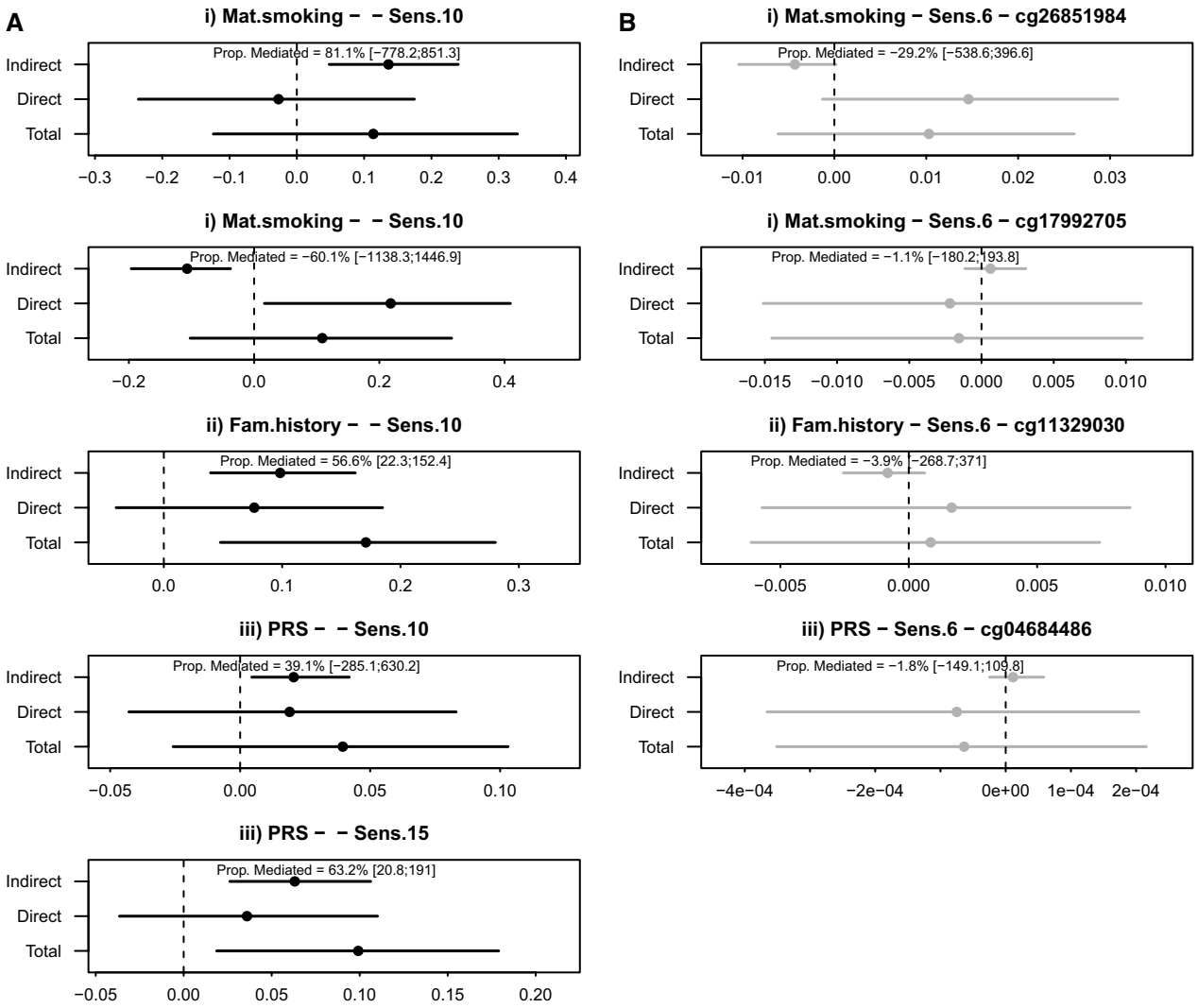


Fig. 3 DNAm in individual CpG sites as predictor or consequence of allergic disease. CpG sites that were identified as mediators in at least one HMA (HIMA or DACT) and validated in causal mediation analysis are presented. Significant indirect effects are indicated with an asterisk and the title follows the pattern exposure–mediator–outcome. Evaluation **A** whether the association between (i) maternal smoking during pregnancy/(ii) family history of allergic disease/(iii) PRS for any allergies and allergic sensitization at ten/15 years is mediated by prior changes in DNAm at six/ten years or **B** whether the association between (i) maternal smoking during pregnancy/(ii) family history of allergic disease/(iii) PRS for any allergies and changes in DNAm at ten years is mediated by prior allergic sensitization at six years. For cg19310430 there is no corresponding model for hypothesis (H2) as DNAm was not measured at 15 years

Replication in BAMSE

Data was available for 445 participants with DNAm measured at eight and aeroallergen sensitization measured at 16 years of age (Additional file 2: Table S13). Table 4 presents the results from BAMSE for our previously validated CpGs (Table 3). Due to the different arrays used in LISA and BAMSE, only two of the five CpGs were available for replication. None of these two CpGs could be replicated in BAMSE, but for cg26851984 the directions of the indirect and direct effects are the same compared to LISA. Full results are included in Additional file 2: Table S14.

Discussion

The present study investigated whether DNAm is a potential cause/predictor or a consequence/outcome of sensitization by conducting causal mediation analyses for well-known risk factors of aeroallergen sensitization as exposures (maternal smoking during pregnancy, family history of allergies, and PRS for any allergy) and data on

Table 4 DNAm in individual CpG sites as predictors of aeroallergen sensitization (H1). Replication of validated CpGs (Table 3) in BAMSE

HMA Method	Mediator [CpG (UCSC/nearest gene - UCSC Group)]	Indirect effect [95% CI]	Direct effect [95% CI]	Total effect [95% CI]	Prop.Med [95% CI]	mQTL (Hawe[a])
A. DNAm at eight and sensitization measured at 16 years						
(i) Maternal smoking during pregnancy						
DACT	cg26851984 *(/RP11-772E11.1)*	0.004 [− 0.012; 0.025]	− 0.035 [− 0.193; 0.134]	− 0.031 [− 0.189; 0.138]	− 0.001 [− 1.667; 1.304]	58
(iii) PRS for any allergies						
HIMA	cg19310430 *(C11orf45 - 5'UTR)*	0.000 [− 0.004; 0.004]	0.101 [0.048; 0.154]	0.101 [0.050; 0.154]	0.000 [− 0.041; 0.050]	

CpGs cg17992705, cg11329030 and cg04684486 are only available on the EPIC array and thus not available in BAMSE, which used the 450K array.

[a] No mQTLs from Gaunt et al. were matched to the respective CpGs

DNAm and aeroallergen sensitization from two consecutive time points as outcomes. We found evidence that DNAm in most previously identified CpG sites (summarized in MRS) was a consequence rather than a cause of aeroallergen sensitization. In addition, we identified five single CpGs that mediated the association between maternal smoking during pregnancy, family history of allergic diseases and a PRS and subsequent aeroallergen sensitization, thus serving as predictors of sensitization. Aggregating both hypotheses, we suggest that DNAm can be a cause as well as a consequence of aeroallergen sensitization, depending on the genomic location.

This study further attempted replication of identified CpGs in the independent Swedish BAMSE cohort but could not significantly replicate any of the five reported CpGs. This might, however, not necessarily negate our findings, as three of the five CpGs were not measured in BAMSE (450K chip vs. EPIC chip in LISA). Furthermore, the time difference is larger between the two assessment points in BAMSE (eight to 16 vs. ten to 15 in LISA). To the best of our knowledge, there are no previous studies investigating causal epigenetic mediation between prenatal exposures and aeroallergen sensitization in childhood and adolescence. Previous studies have reported mediation effects of DNAm for the associations between body-mass-index (BMI) and trajectories with asthma [43], BMI and cardio-metabolic risk [44], and age at puberty onset and lung function [45]. Of note, none of these studies investigated both directions, DNAm as both a predictor (H1) and as a consequence (H2).

Publications investigating mQTLs found that DNAm changes are often seen as a consequence of diseases rather than their cause [46] and this is supported by our findings on the allergy-related MRS. However, in the present study we also identified CpGs which serve as mediators for the association between known determinants of allergies and aeroallergen sensitization. Of note, none of the identified single CpGs are part of the evaluated MRS after clumping and thresholding, even though one has been previously reported by the same EWAS as an associated CpG site (Peng [3]). This might indicate that DNAm acts in both effect directions, represented by differing sets of CpG loci.

On the one hand, our finding that MRS are rather a consequence than a cause of sensitization falls in line with our previous results [7], which might also rely on the fact that the pre-identified CpGs were reported in mostly cross-sectional EWAS. On the other hand, the single CpGs mediating prenatal exposures on aeroallergen sensitization later in life, might be facilitated as early predictors for disease development. These should be followed up in future studies to further determine their clinical relevance.

For cg26851984, which was identified as a mediator of the association between maternal smoking during pregnancy and sensitization with DACT, we identified the closest gene to be *PRPF3*. This gene is associated with eczema [13], eosinophil counts [47] and any allergy [15], supporting the importance of this CpG as a mediator of allergen sensitization. Of note, this CpG was previously reported in an EWAS on aeroallergen sensitization [3], as only previously known CpGs were tested as potential mediators with the DACT method. However, it is not part of the allergy-related MRS previously calculated based on these EWAS [7] after clumping and thresholding. Further, it is a mQTL and its associations have to be interpreted with caution as effects here could be attributable to surrounding SNPs, which may explain the higher mediation effect size (0.139 for maternal smoking as exposure and cg26851984 as mediator) compared to all others ($\leq$ 0.108), but also the higher albeit non-significant proportion mediated of 81.1%.

Other CpGs identified with the hypothesis-generating HIMA approach were also located in proximity to allergy-relevant genes. *ATXN2L*, located in the exon boundary and corresponding to cg17992705, is

associated with forced vital capacity [48], a lung function parameter that is reduced in asthma patients. Further, *DIP2C* (cg12724894) and *ASB2* (cg03389164) are associated with eosinophil counts [49, 50] and located within the gene body and promoter, respectively.

Looking at Figs. 2 and 3, it can be seen that not all total effects are significant while the indirect effects are. While significant total effects were a prerequisite of potential mediation in the traditional causal step approach proposed by Baron and Kenny in 1986 [51], it is not a formal requirement in the causal mediation analysis approach we used, but reduces that statistical power to detect indirect effects [50, 51]. While all of our exposures are known risk factors for aeroallergen sensitization, they might not necessarily show significance in our reduced sub-sample. The total effect is defined as the sum of the direct and all indirect effects and we do sometimes observe opposite effect signs for direct and indirect effects (e.g., cg17992705), which can attenuate the total effects.

The present study has multiple strengths: We have objectively measured data on all levels of the analysis for the model in which PRS is the exposure, as neither PRS, DNAm, nor blood-measured aeroallergen sensitization is subject to recall bias. In addition, the LISA study is a well-established prospective German birth cohort with still ongoing follow-up and provides a valuable data source for studying allergic diseases. This also supports the causal interpretation, as the longitudinal succession of measured mediators and outcomes was possible due to the longitudinal design of the study. DNAm is being measured repeatedly at both six and ten years, as well as consecutive time points being used for the definition of exposure, mediator, and outcome. This longitudinal design might also enable future analyses, ideally paired with similar studies with comparable design to reach higher statistical power for epigenome-wide mediation analyses. Further, we applied three different HMA methods complemented with causal mediation analysis to investigate their applicability to the allergic context in contrast to simpler screening methods for reduction of the multiple-testing burden. Each HMA approach is based on different assumptions and uses different strategies to deal with the challenges of multiple testing.

Limitations of the presented study include the small sample size, which might be insufficient to detect all potential mediation effects, especially as effects of single CpGs are rather small. This might also explain why we could not replicate single CpGs in both time windows (A&B) or why we did not find significant gene-units using the gHMA approach. It could also be speculated that single CpGs might be more relevant in relation to allergic sensitization than methylation across a whole gene, as this is the biggest difference between gHMA as a gene-based approach and the others (HIMA and DACT) as CpG-based approaches. Further, applying the PRS as an exposure, we did not check whether there is significant mediation between single SNPs and CpGs, but with the development of relevant methodology [52] this is of great interest for future studies. MRS were further determined according to their cross-sectional prediction accuracy and not optimized according to their performance in a prospective or mediation setting as applied here. Another general issue might be confounding, which is a serious problem in mediation analysis [28]. We adjusted our models based on DAGs to the best of our knowledge, however, unmeasured confounding cannot be ruled out completely in observational studies.

Conclusions

In conclusion, we found indications that DNAm could either be the cause of allergic sensitization or the consequence thereof, depending on the genomic location. The two different sets of DNAm patterns, namely MRS as consequence of sensitization or single CpGs as cause, have differing clinical implications: While MRS might be considered as cross-sectional biomarkers, the single CpGs might be clinically relevant early predictors of sensitization and should be investigated in future studies.

Abbreviations

BAMSE: Barn (=Children) Allergy Milieu Stockholm Epidemiology; BMI: Body Mass Index; CpG: Cytosine and Guanine only separated by their phosphate backbone; CTP: Cell-type proportions; DACT: Divide-Aggregate Composite-Null Test; DAG: Directed Acyclic Graph; DNAm: DNA methylation; EWAS: Epigenome-wide association studies; FDR: False Discovery Rate; gHMA: Gene-based High-dimensional Mediation Analysis; GWAS: Genome-wide association studies; HMA: High-dimensional mediation analysis; IgE: Immunoglobulin E; LISA: Influence of Life-style factors on Development of the Immune System and Allergies in East and West Germany study; MRS: Methylation Risk Score; NO_2: Nitrogen Dioxide; PRS: Polygenic risk score; SES: Socio-economic status; SIS: Sure Independence Screening.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-022-01332-5.

Additional file 1. Supplementary methods, Table S1 and Figures S1-S9.

Additional file 2. Supplementary Tables S2-S14.

Acknowledgements

The authors thank all the families for their participation in the LISA study. Furthermore, we thank all members of the LISA Study Group for their excellent work. The LISA Study group consists of the following: Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology, Munich (Heinrich J, Schnappinger M, Brüske I, Ferland M, Schulz H, Zeller C, Standl M, Thiering E, Tiesler C, Flexeder C); Department of Pediatrics, Municipal Hospital "St. Georg", Leipzig (Borte M, Diez U, Dorn C, Braun E); Marien Hospital Wesel, Department of Pediatrics, Wesel (von Berg A, Berdel D, Stiers G, Maas B); Pediatric Practice, Bad Honnef (Schaaf B); Helmholtz Centre of Environmental Research—UFZ, Department of Environmental Immunology/Core Facility Studies, Leipzig (Lehmann I, Bauer M, Röder S, Schilde M,

Nowak M, Herberth G , Müller J); Technical University Munich, Department of Pediatrics, Munich (Hoffmann U, Paschke M, Marra S); Clinical Research Group Molecular Dermatology, Department of Dermatology and Allergy, Technische Universität München (TUM), Munich (Ollert M, J. Grosch). We further want to thank Nadine Lindemann for her work analyzing the DNA methylation samples and Nicole Gladish for providing her processing scripts for DNA methylation data. We thank all the participants and parents participating in the BAMSE cohort and the staff involved in the study through the years.

Author contributions
AK: Formal analysis, Methodology, Visualization, Writing—original draft preparation, Writing—review & editing. SKM: Formal analysis, Validation, Writing—review and editing. SA: Methodology, Writing—review & editing. DF: Methodology, Writing—review & editing. ET: Supervision, Writing—review & editing. MW: Data Curation, Writing—review & editing. EM: Data Curation, Funding acquisition, Investigation, Resources, Writing—review & editing. AP: Supervision, Resources, Writing—review & editing. MS: Data Curation, Methodology, Resources, Supervision, Writing—review & editing. AH: Conceptualization, Methodology, Supervision, Writing—review & editing. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The LISA study was mainly supported by grants from the Federal Ministry for Education, Science, Research and Technology and in addition from Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef for the first 2 years. The 4 year, 6 year, 10 year and 15 year follow-up examinations of the LISA study were covered from the respective budgets of the involved partners (Helmholtz Zentrum Munich (former GSF), Helmholtz Centre for Environmental Research—UFZ, Leipzig, Research Institute at Marien-Hospital Wesel, Pediatric Practice, Bad Honnef, IUF—Leibniz-Research Institute for Environmental Medicine at the University of Düsseldorf) and in addition by a grant from the Federal Ministry for Environment (IUF Düsseldorf, FKZ 20462296). Further, the 15-year follow-up examination of the LISA study was supported by the Commission of the European Communities, the 7th Framework Program: MeDALL project. The BAMSE study is supported by grants from the Swedish Research Council, the Swedish Heart-Lung Foundation, Karolinska Institutet SFO Epidemiology and Region Stockholm (ALF project, and for cohort and database maintenance). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 949906). AK was supported by a research fellowship (No. 57504619) from the DAAD (German academic exchange service) to conduct this project with AH at Emory University. AH is supported by the HERCULES Center (NIEHS P30ES019776).

Availability of data and materials
Due to data protection reasons, the datasets generated and/or analyzed during the current study cannot be made publicly available. The datasets are available to interested researchers from the corresponding author on reasonable request, provided the release is consistent with the consent given by the LISA and/or BAMSE study participants. Ethical approval might be obtained for the release and a data transfer agreement from the legal department of the Helmholtz Zentrum München and/or Karolinska Institutet must be accepted.

Declarations

Ethical approval and consent to participate.
The LISA study was approved by the local ethics committee (Bavarian Board of Physicians (Reference numbers: 6 years—03166; 10 years—07098)) and written, informed consent was obtained from the parents or legal guardians. The present study uses only data from the Munich study center. The Regional Ethics Board approved the BAMSE study (reference number: 93-189, 98-175, 01-475, 02-420, 2010/1474-31/3, and 2011/2037-32) and informed consent (oral and written) was obtained from the parents or legal guardians.

Consent for publication
Not applicable.

Competing interests
EM has received lecture and advisory board fees from ALK outside the submitted work. The other authors declare that they have no competing interests.

Author details
[1]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany. [2]Institute for Medical Information Processing, Biometry, and Epidemiology; Pettenkofer School of Public Health, LMU Munich, Munich, Germany. [3]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd NE, Atlanta, GA 30322, USA. [4]Division of Metabolic and Nutritional Medicine, Dr. Von Hauner Children's Hospital, University of Munich Medical Center, Munich, Germany. [5]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [6]Research Unit Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany. [7]Chair of Epidemiology, Ludwig-Maximilians University, Marchioninistr. 15, 81377 Munich, Germany. [8]German Center for Lung Research (DZL), Munich, Germany. [9]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA.

Received: 11 May 2022 Accepted: 5 September 2022
Published online: 16 September 2022

References

1. Everson TM, Lyons G, Zhang H, Soto-Ramírez N, Lockett GA, Patil VK, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. Genome Med. 2015;7(1):89.
2. Chen W, Wang T, Pino-Yanes M, Forno E, Liang L, Yan Q, et al. An epigenome-wide association study of total serum immunoglobulin E in Hispanic children. J Allergy Clin Immunol. 2017;140(2):571–7.
3. Peng C, Van Meel ER, Cardenas A, Rifas-Shiman SL, Sonawane AR, Glass KR, et al. Epigenome-wide association study reveals methylation pathways associated with childhood allergic sensitization. Epigenetics. 2019;14(5):445–66.
4. Zhang H, Kaushal A, Merid SK, Melén E, Pershagen G, Rezwan FI, et al. DNA methylation and allergic sensitizations: a genome-scale longitudinal study during adolescence. Allergy. 2019;74(6):1166–75.
5. Reese SE, Xu CJ, den Dekker HT, Lee MK, Sikdar S, Ruiz-Arenas C, et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. J Allergy Clin Immunol. 2019;143(6):2062–74.
6. Xu CJ, Gruzieva O, Qi C, Esplugues A, Gehring U, Bergström A, et al. Shared DNA methylation signatures in childhood allergy: the MeDALL study. J Allergy Clin Immunol. 2021;147(3):1031–40.
7. Kilanowski A, Chen J, Everson T, Thiering E, Wilson R, Gladish N, et al. Methylation risk scores for childhood aeroallergen sensitization: results from the LISA birth cohort. Allergy. 2022;77:2803–17. https://doi.org/10.1111/all.15315.
8. Peng C, Cardenas A, Rifas-Shiman SL, Hivert MF, Gold DR, Platts-Mills TA, et al. Epigenome-wide association study of total serum immunoglobulin E in children: a life course approach. Clin Epigenetics. 2018;10:55.
9. Han L, Kaushal A, Zhang H, Kadalayil L, Duan J, Holloway JW, et al. DNA methylation at birth is associated with childhood serum immunoglobulin E levels. Epigenetics Insights. 2021;14:25168657211008108.
10. Slob EMA, Longo C, Vijverberg SJH, van Beijsterveldt TCEM, Bartels M, Hottenga JJ, et al. Persistence of parental-reported asthma at early ages: a longitudinal twin study. Pediatr Allergy Immunol Off Publ Eur Soc Pediatr Allergy Immunol. 2022;33(3): e13762.
11. Bataille V, Lens M, Spector TD. The use of the twin model to investigate the genetics and epigenetics of skin diseases with genomic, transcriptomic and methylation data. J Eur Acad Dermatol Venereol. 2012;26(9):1067–73.
12. Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev. 2011;242(1):10–30.
13. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet. 2015;47(12):1449–56.

14. Waage J, Standl M, Curtin JA, Jessen LE, Thorsen J, Tian C, et al. Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. Nat Genet. 2018;50(8):1072–80.
15. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet. 2017;49(12):1752–7.
16. Ferreira MAR, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Eleven loci with new reproducible genetic associations with allergic disease risk. J Allergy Clin Immunol. 2019;143(2):691–9.
17. Clark H, Granell R, Curtin JA, Belgrave D, Simpson A, Murray C, et al. Differential associations of allergic disease genetic variants with developmental profiles of eczema, wheeze and rhinitis. Clin Exp Allergy J Br Soc Allergy Clin Immunol. 2019;49(11):1475–86.
18. Spycher BD, Henderson J, Granell R, Evans DM, Smith GD, Timpson NJ, et al. Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. J Allergy Clin Immunol. 2012;130(2):503-509.e7.
19. Sordillo JE, Lutz SM, Jorgenson E, Iribarren C, McGeachie M, Dahlin A, et al. A polygenic risk score for asthma in a large racially diverse population. Clin Exp Allergy J Br Soc Allergy Clin Immunol. 2021;51(11):1410–20.
20. Thacher JD, Gehring U, Gruzieva O, Standl M, Pershagen G, Bauer CP, et al. Maternal Smoking during Pregnancy and Early Childhood and Development of Asthma and Rhinoconjunctivitis—a MeDALL Project. Environ Health Perspect. 2018;126(4): 047005.
21. Janbazacyabar H, van Bergenhenegouwen J, Garssen J, Leusink-Muis T, van Ark I, van Daal MT, et al. Prenatal and postnatal cigarette smoke exposure is associated with increased risk of exacerbated allergic airway immune responses: a preclinical mouse model. Front Immunol. 2021;12: 797376.
22. Grosche S, Marenholz I, Esparza-Gordillo J, Arnau-Soler A, Pairo-Castineira E, Rüschendorf F, et al. Rare variant analysis in eczema identifies exonic variants in DUSP1, NOTCH4 and SLC9A4. Nat Commun. 2021;12(1):6618.
23. Kilanowski A, Thiering E, Wang G, Kumar A, Kress S, Flexeder C, Bauer CP, Berdel D, von Berg A, Bergström A, Gappa M, Heinrich J, Herberth G, Koletzko S, Kull I, Melen E, Schikowski T, Peters A, Standl M. Allergic disease trajectories up to adolescence: Characteristics, early-life and genetic determinants. Allergy. 2022. https://doi.org/10.1111/all.15511.
24. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15(12):503.
25. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat Oxf Engl. 2007;8(1):118–27.
26. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017;13(18):105.
27. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics. 2020;15:1–11.
28. VanderWeele TJ. Mediation analysis: a practitioner's guide. Annu Rev Public Health. 2016;37(1):17–32.
29. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package "dagitty." Int J Epidemiol. 2016;45(6):1887–94.
30. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. **mediation** : *R* Package for Causal Mediation Analysis. J Stat Softw [Internet]. 2014 [cited 2021 Aug 3];59(5). http://www.jstatsoft.org/v59/i05/
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
32. Liu Z, Shen J, Barfield R, Schwartz J, Baccarelli AA, Lin X. Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. 2021 Apr [cited 2021 Aug 13] p. 2020.09.20.20198226. https://www.medrxiv.org/content/10.1101/2020.09.20.20198226v2
33. Efron B. Large-scale simultaneous hypothesis testing. J Am Stat Assoc. 2004;99(465):96–104.
34. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinforma Oxf Engl. 2016;32(20):3150–4.
35. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol. 2008;70(5):849–911.
36. Fang R, Yang H, Gao Y, Cao H, Goode EL, Cui Y. Gene-based mediation analysis in epigenetic studies. Brief Bioinform. 2021;22(3):bbaa113. https://doi.org/10.1093/bib/bbaa113.
37. Hüls A, Robins C, Conneely KN, Edgar R, De Jager PL, Bennett DA, et al. Brain DNA methylation patterns in CLDN5 associated with cognitive decline. Biol Psychiatry. 2021;S0006–3223(21):00084–6.
38. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17(1):61.
39. Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat Genet. 2022;54(1):18–29.
40. Leffler J, Stumbles PA, Strickland DH. Immunological processes driving IgE sensitisation and disease development in males and females. Int J Mol Sci. 2018;19(6):E1554.
41. Melén E, Bergström A, Kull I, Almqvist C, Andersson N, Asarnoj A, et al. Male sex is strongly associated with IgE-sensitization to airborne but not food allergens: results up to age 24 years from the BAMSE birth cohort. Clin Transl Allergy. 2020;10:15.
42. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/
43. Rathod R, Zhang H, Karmaus W, Ewart S, Kadalayil L, Relton C, et al. BMI trajectory in childhood is associated with asthma incidence at young adulthood mediated by DNA methylation. Allergy Asthma Clin Immunol Off J Can Soc Allergy Clin Immunol. 2021;17(1):77.
44. Huang JV, Cardenas A, Colicino E, Schooling CM, Rifas-Shiman SL, Agha G, et al. DNA methylation in blood as a mediator of the association of mid-childhood body mass index with cardio-metabolic risk score in early adolescence. Epigenetics. 2018;13(10–11):1072–87.
45. Li L, Zhang H, Holloway JW, Ewart S, Relton CL, Arshad SH, et al. Does DNA methylation mediate the association of age at puberty with forced vital capacity or forced expiratory volume in 1 s? ERJ Open Res. 2022;8(1):00476–2021.
46. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021;53(9):1311–21.
47. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The polygenic and monogenic basis of blood traits and diseases. Cell. 2020;182(5):1214-1231.e11.
48. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging polygenic functional enrichment to improve gwas power. Am J Hum Genet. 2019;104(1):65–75.
49. Chen MH, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell. 2020;182(5):1198-1213.e14.
50. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415-1429.e19.
51. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173–82.
52. Zhao Y, Li L, Initiative ADN. Multimodal data integration via mediation analysis with high-dimensional exposures and mediators. Hum Brain Mapp. 2022;43(8):2519–33.

Publisher's Note

Part III

Appendix

A Additional manuscript

A.1 Publication: European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation

Title:	European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation
Authors:	Ashley Budu-Aggrey*, Anna Kilanowski*, Maria K Sobczyk, ... , Marie Standl*, Lavinia Paternoster*
Journal:	*Nature Communications*
Year:	2023
Volume:	Online ahead of print
DOI:	10.1038/s41467-023-41180-2
Supplements:	https://www.nature.com/articles/s41467-023-41180-2
Impact factor:	17.694 (Journal Citation Reports^{TM} 2021)
Rank:	6/73 in category MULTIDISCIPLINARY SCIENCES (Journal Citation Reports^{TM} 2021)

* These authors have contributed equally to this work.

nature communications

Article

https://doi.org/10.1038/s41467-023-41180-2

European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation

Received: 20 September 2022

Accepted: 24 August 2023

Published online: 04 October 2023

Check for updates

A list of authors and their affiliations appears at the end of the paper

Atopic dermatitis (AD) is a common inflammatory skin condition and prior genome-wide association studies (GWAS) have identified 71 associated loci. In the current study we conducted the largest AD GWAS to date (discovery N = 1,086,394, replication N = 3,604,027), combining previously reported cohorts with additional available data. We identified 81 loci (29 novel) in the European-only analysis (which all replicated in a separate European analysis) and 10 additional loci in the multi-ancestry analysis (3 novel). Eight variants from the multi-ancestry analysis replicated in at least one of the populations tested (European, Latino or African), while two may be specific to individuals of Japanese ancestry. AD loci showed enrichment for DNAse I hypersensitivity and eQTL associations in blood. At each locus we prioritised candidate genes by integrating multi-omic data. The implicated genes are predominantly in immune pathways of relevance to atopic inflammation and some offer drug repurposing opportunities.

Atopic dermatitis (AD, or eczema) is a common allergic disease, characterised by (often relapsing) skin inflammation affecting up to 20% of children and 10% of adults[1]. Several genome-wide association studies (GWAS) have been performed in recent years, identifying genetic risk loci for AD.

Our most recent GWAS meta-analysis within the EAGLE (EArly Genetics and Lifecourse Epidemiology) consortium, published in 2015 uncovered 31 AD risk loci[2]. Since then, additional GWAS have been published which have confirmed known risk loci[3,4] and discovered novel loci[5]. Five novel loci were identified in a European meta-analysis[6], and variants in 3 genes were implicated in a rare variant study in addition to 5 novel loci[7]. Four novel loci were reported in a Japanese population (and another 4 identified in a trans-ethnic meta-analysis in the same study)[8], giving a total of 71 previously reported AD loci[2–14] (defined as 1 Mb regions) of which 57 have been reported in European ancestry individuals, 18 have been reported in individuals of non-European ancestry and 29 in individuals across multiple ancestry groups (Supplementary Data 1).

The availability of several new large population-based studies has provided an opportunity to perform an updated GWAS of AD, aiming to incorporate data from all cohorts that have contributed to previously published AD GWAS, as well as data from additional cohorts, to present the most comprehensive GWAS of AD to date, including comparison of effects between European, East Asian, Latino and African ancestral groups. In this work we identify novel loci and use multi-omic data to further characterise these associations, prioritising candidate causal genes at individual loci and investigating the genetic architecture of AD in relation to tissues of importance and shared genetic risk with other traits.

Results

European GWAS

The discovery European meta-analysis (N = 864,982; 60,653 AD cases and 804,329 controls from 40 cohorts, summarised in Supplementary Data 2) identified 81 genome-wide significant independent associated loci (Fig. 1a and Supplementary Fig. 1). 52 were at

e-mail: l.paternoster@bristol.ac.uk

https://doi.org/10.1038/s41467-023-41180-2

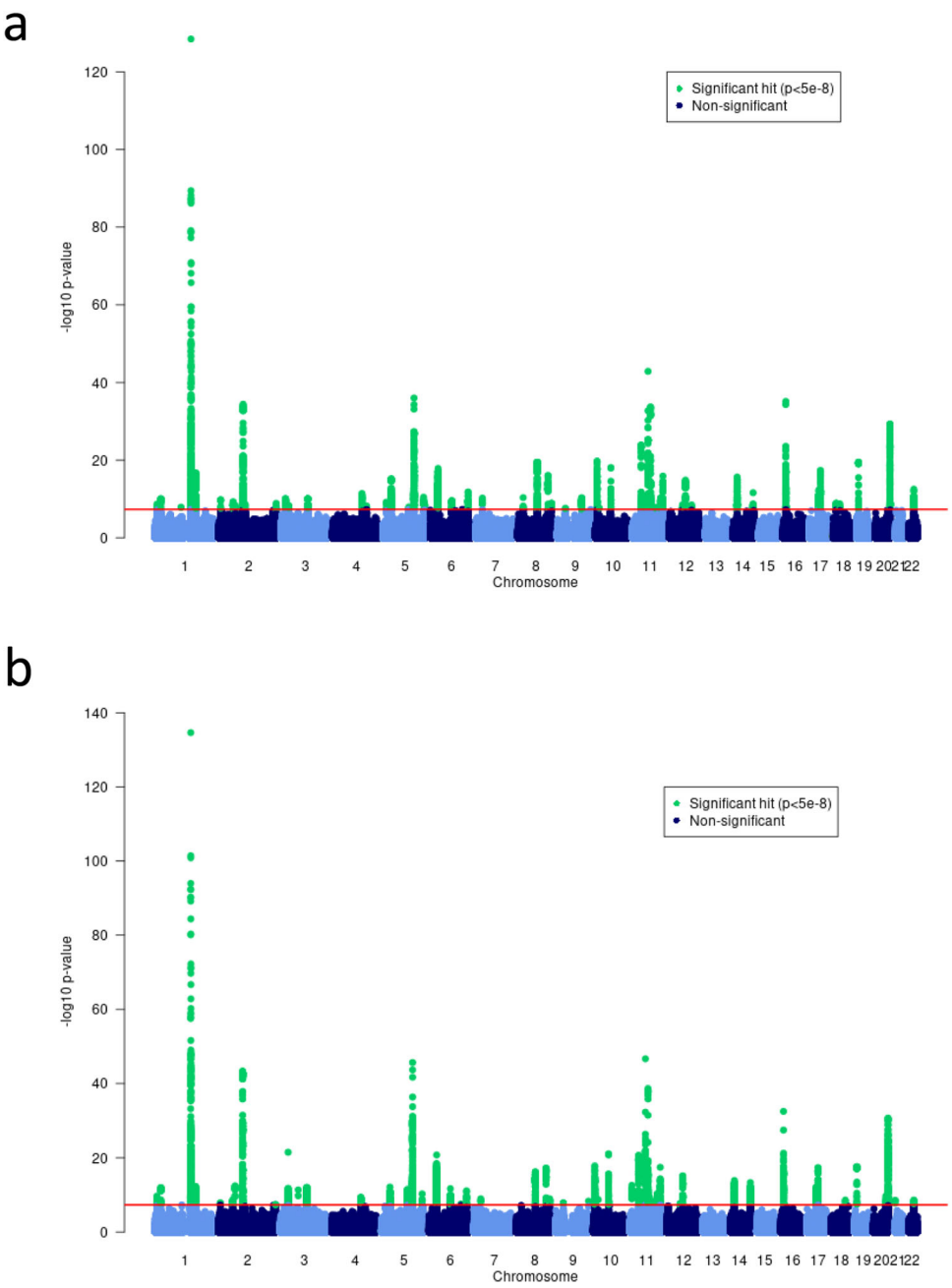


Fig. 1 | Manhattan plots of atopic dermatitis GWAS. (**a**) the European-only fixed effects meta-analysis (n = 864,982 individuals) and (**b**) the multi-ancestry MR-MEGA meta-analysis (n = 1,086,394 individuals). $-\log_{10}$(P-values) are displayed for all variants in the meta-analysis. Variants that meet the genome-wide significance threshold (5×10^{-8}, red line) are shown in green.

previously reported loci (Table 1) and 29 (Table 2) were novel (according to criteria detailed in the methods). All 81 were associated in the European 23andMe replication analysis (Bonferroni corrected $P < 0.05/81 = 6 \times 10^{-4}$), N = 2,904,664, Table 1). There was little evidence of genomic inflation in the individual studies (lambda <1.05) and overall (1.06). Conditional analysis determined 44 additional secondary independent associations ($P < 1 \times 10^{-5}$) across 21 loci (Supplementary Data 3).

The SNP-based heritability (h^2_{SNP}) for AD was estimated to be 5.6% in the European discovery meta-analysis (LDSC intercept=1.042 (SE = 0.011)). This is low in comparison to heritability estimates for twin studies (~80%)[15,16], but comparable with previous h^2_{SNP} estimates for AD in Europeans (5.4%)[6].

Multi-ancestry GWAS

In a multi-ancestry analysis including individuals of European, Japanese, Latino and African ancestry (Supplementary Data 2, N = 1,086,394; 65,107 AD cases and 1,021,287 controls), a total of 89 loci were identified as associated with AD (Fig. 1b and Supplementary Fig. 1). 75 of these were not independent of lead variants identified in the European-only analysis ($r^2 > 0.01$ in the relevant ancestry) and a further 9 showed some evidence for association (Bonferroni corrected $P < 0.05/89 = 5.6 \times 10^{-4}$) in the European analysis, but 5 were not associated ($P > 0.1$) in Europeans (Table 3, Supplementary Data 4).

Of the 14 loci that reached genome-wide significance in the multi-ancestry discovery analysis only (Table 3), 8 replicated in at least one of

https://doi.org/10.1038/s41467-023-41180-2

Table 1 | Genome-wide significant loci in European-only analysis that have been previously reported

			European discovery			Multi-ancestry discovery		23andMe European replication (*N* = 2,904,664)			
Variant	**Chr:position**	**Alleles (EAF)**	**OR (CI)**	***P***	***N* (studies)**	***P***	**N (studies)**	**OR (CI)**	***P***	**Gene**	**Pathway/Function**
rs7542147	1:25294618	C/T (0.49)	1.04 (1.03–1.06)	8.52E−11	860840 (38)	2.4E−09	870216 (42)	1.05 (1.04–1.05)	4.6E−56	*RUNX3*	Versatile transcription factor, incl. T cell differentiation
rs12123821	1:152179152	T/C (0.05)	1.40 (1.35–1.45)	4.05E−90	850727 (29)	2.3E−98	857207 (31)	1.27 (1.25–1.29)	1.4E−228	*FLG*	Skin barrier protein
rs61816766[a]	1:152319572	C/T (0.03)	1.66 (1.58–1.74)	6.44E−89	627936 (20)	1.1E−102	634416 (22)	1.41 (1.39–1.43)	1.4E−228	*FLG*	Skin barrier protein
rs72702900	1:152771963	A/T (0.04)	1.28 (1.24–1.33)	2.98E−46	851612 (29)	3.0E−49	853748 (30)	1.23 (1.22–1.25)	4.2E−163	*FLG*	Skin barrier protein
rs61815704	1:152893891	G/C (0.02)	1.78 (1.67–1.89)	3.21E−71	530473 (19)	9.2E−72	536953 (21)	1.36 (1.34–1.39)	5.5E−212	*S100A9*[b]	TLR4 signalling
rs12133641	1:154428283	G/A (0.39)	1.07 (1.05–1.08)	1.72E−21	857974 (37)	1.8E−22	1079390 (42)	1.04 (1.04–1.05)	3.0E−45	*IL6R*	Cytokine signalling in immune system
rs859723	1:172744543	A/G (0.36)	0.94 (0.93–0.96)	3.74E−14	522713 (37)	2.4E−14	744125 (42)	0.96 (0.96–0.97)	2.2E−39	*TNFSF4*[b]	Cytokine signalling in immune system
rs11811788	1:173150727	G/C (0.24)	1.07 (1.05–1.08)	1.85E−17	859747 (38)	3.1E−16	1081160 (43)	1.04 (1.04–1.05)	1.6E−39	*TNFSF4*	Cytokine signalling in immune system
rs891058	2:8442547	A/G (0.29)	0.96 (0.94–0.97)	1.76E−10	862482 (38)	2.2E−11	1083890 (43)	0.97 (0.97–0.98)	3.0E−18	*ID2*	Transcriptional regulator of many cellular processes
rs112111458	2:71100105	G/A (0.12)	0.94 (0.92–0.96)	5.50E−09	858567 (37)	1.4E−11	1079980 (42)	0.96 (0.95–0.97)	1.3E−21	*CD207*	Dendritic cell function
rs2272128	2:103039929	A/G (0.77)	0.91 (0.90–0.92)	8.14E−35	862259 (39)	3.8E−48	1083670 (44)	0.93 (0.93–0.94)	2.2E−100	*IL18RAP*	Cytokine signalling in immune system
rs4131280	3:18414570	A/G (0.57)	0.96 (0.95–0.98)	1.2E−08	864982 (40)	5.8E−08	1086390 (45)	0.97 (0.97–0.98)	2.2E−19	*SATB1*	Regulates chromatin structure and gene expression
rs13097010	3:18673161	G/A (0.34)	1.05 (1.03–1.06)	9.0E−11	864982 (40)	1.5E−08	1086390 (45)	1.02 (1.01–1.02)	1.4E−07	*SATB1*	Regulates chromatin structure and gene expression
rs35570272	3:33047662	T/G (0.40)	1.04 (1.03–1.05)	5.7E−09	864982 (40)	2.3E−20	1086390 (45)	1.03 (1.03–1.04)	1.6E−26[a]	*GLB1*	Sphingolipid metabolism
rs6808249	3:112648985	T/C (0.54)	0.96 (0.95–0.97)	9.05E−11	859747 (38)	3.8E−12	1081160 (43)	0.97 (0.96–0.97)	4.7E−29	*CD200R1*	Adaptive immune system
rs45599938	4:123386720	A/G (0.35)	1.05 (1.03–1.06)	4.61E−12	859747 (38)	3.7E−10	1081160 (43)	1.05 (1.05–1.06)	1.3E−62	*KIAA1109*	Endosomal transport
rs10214273	5:35883986	G/T (0.27)	0.94 (0.93–0.96)	5.97E−16	863209 (39)	1.8E−14	1084620 (44)	0.93 (0.93–0.94)	2.9E−99	*IL7R*	Cytokine signalling in immune system
rs17132590	5:110331899	C/T (0.10)	1.07 (1.05–1.10)	1.16E−08	525225 (38)	1.7E−08	746637 (43)	1.03 (1.02–1.04)	1.0E−07	*CAMK4*	Immune response, inflammation & memory consolidation
rs4706020	5:130674076	A/G (0.34)	0.95 (0.93–0.96)	1.12E−11	518425 (35)	2.7E−11	527801 (39)	0.98 (0.98–0.99)	6.4E−09	*CDC42SE2*	F-actin accumulation at immunological synapse of T cells
rs4705908	5:131347520	A/G (0.37)	0.95 (0.93–0.96)	6.80E−13	520344 (36)	1.6E−11	529720 (40)	0.98 (0.97–0.98)	8.0E−15	*SLC22A5*	Organic cation transport
rs20541	5:131995964	G/A (0.78)	0.91 (0.89–0.92)	1.00E−36	859747 (38)	8.4E−51	1076820 (42)	0.92 (0.91–0.92)	1.2E−129	*SLC22A5*	Organic cation transport
rs114503346	5:172192350	T/C (0.04)	0.89 (0.86–0.92)	3.62E−11	855569 (33)	1.3E−10	862049 (35)	0.94 (0.93–0.95)	3.2E−17	*ERGIC1*	Transport between endoplasmic reticulum and golgi
rs41293876	6:31466536	C/G (0.14)	0.90 (0.88–0.93)	7.02E−16	645820 (36)	6.5E−18	865966 (40)	0.95 (0.95–0.96)	4.3E−32	*TNF*	Cytokine signalling in immune system
rs12153855	6:32074804	C/T (0.10)	0.92 (0.90–0.94)	1.96E−11	812536 (37)	2.8E−10	821912 (41)	0.96 (0.95–0.97)	2.3E−18	*ATF6B*	Endoplasmic reticulum stress response
rs28383330	6:32600340	G/A (0.13)	0.88 (0.85–0.90)	1.42E−18	625716 (28)	1.8E−17	632956 (31)	0.94 (0.93–0.95)	2.4E−51	*AGER*	Immunoglobulin surface receptor
rs9275218	6:32658933	G/C (0.34)	1.06 (1.04–1.08)	5.36E−10	505320 (34)	1.0E−09	512560 (37)	1.01 (1.01–1.02)	1.0E−04	*HLA-DRA*	Immune response antigen presentation
rs629326	6:159496713	T/G (0.61)	0.95 (0.94–0.97)	1.7E−12	859747 (38)	4.5E−12	1081160 (43)	0.95 (0.95–0.96)	5.4E−61[a]	*TAGAP*[b]	T cell activation
rs952558	8:81288734	T/A (0.62)	0.94 (0.93–0.95)	3.60E−20	862259 (39)	1.3E−19	1083670 (44)	0.97 (0.96–0.97)	2.2E−31	*ZBTB10*	Transcriptional regulation
rs6996614	8:126609868	A/C (0.53)	1.07 (1.05–1.08)	8.48E−17	693031 (37)	1.0E−17	914443 (42)	1.03 (1.02–1.03)	1.5E−19	*TRIB1*	Protein kinase regulation
rs12251307	10:6123495	T/C (0.12)	1.10 (1.08–1.12)	1.98E−20	864982 (40)	8.4E−19	1086390 (45)	1.10 (1.09–1.11)	4.7E−107	*IL2RA*	Cytokine signalling in immune system
rs10796303	10:6627700	C/T (0.66)	0.96 (0.94–0.97)	8.69E−10	856884 (38)	8.5E−10	1078300 (43)	0.97 (0.96–0.97)	5.6E−25	*PRKCQ*	T cell activation
rs10822037	10:64376558	C/T (0.61)	1.06 (1.05–1.08)	8.53E−19	864982 (40)	1.3E−24	1086390 (45)	1.05 (1.04–1.05)	4.0E−55	*ADO*	Taurine biosynthesis
rs10836538	11:36365253	T/G (0.34)	0.96 (0.94–0.97)	9.18E−11	863063 (39)	1.1E−13	1084480 (44)	0.95 (0.95–0.96)	6.2E−55	*PRR5L*	Protein phosphorylation
rs28520436	11:36428447	T/C (0.03)	1.20 (1.16–1.24)	1.22E−24	855865 (29)	4.1E−25	1074380 (32)	1.18 (1.16–1.20)	5.3E−81	*PRR5L*	Protein phosphorylation
rs10791824	11:65559266	G/A (0.58)	1.10 (1.08–1.11)	1.34E−43	864982 (40)	1.2E−51	1086390 (45)	1.07 (1.06–1.07)	1.2E−105	*MAP3K11*	Cytokine signalling in immune system
rs7936323	11:76293758	A/G (0.46)	1.08 (1.07–1.10)	2.07E−34	864982 (40)	1.8E−39	1086390 (45)	1.07 (1.07–1.08)	1.9E−133	*LRRC32*	TGF beta regulation incl. on T cells

https://doi.org/10.1038/s41467-023-41180-2

Table 1 (continued)

			European discovery			Multi-ancestry discovery		23andMe European replication (*N* = 2,904,664)			
Variant	**Chr:position**	**Alleles (EAF)**	**OR (CI)**	***P***	***N* (studies)**	***P***	**N (studies)**	**OR (CI)**	***P***	**Gene**	**Pathway/Function**
rs11236813	11:76343427	C/G (0.10)	0.93 (0.91–0.95)	1.94E−12	864646 (39)	4.8E−12	1086060 (44)	0.95 (0.94–0.96)	2.6E−26	*LRRC32*	TGF beta regulation incl. on T cells
rs10790275	11:118745884	C/G (0.80)	1.06 (1.04–1.07)	5.46E−11	859747 (38)	4.8E−09	1081160 (43)	1.02 (1.02–1.03)	1.0E−10	*DDX6*[b]	mRNA degradation
rs7127307	11:128187383	C/T (0.49)	0.95 (0.93–0.96)	1.29E−16	859747 (38)	1.0E−17	1081160 (43)	0.96 (0.95–0.96)	6.1E−52	*FLI1*	NF-kappaB signalling
rs705699	12:56384804	A/G (0.40)	1.04 (1.03–1.05)	3.31E−09	864982 (40)	6.7E−08	1086390 (45)	1.03 (1.03–1.04)	8.7E−27	*RPS26*	Peptide chain elongation
rs2227491	12:68646521	C/T (0.61)	1.05 (1.04–1.07)	1.46E−15	864982 (40)	1.9E−15	1086390 (45)	1.05 (1.05–1.06)	1.2E−71	*IL22*	Cytokine signalling in immune system
rs2415269	14:35638937	A/G (0.26)	0.94 (0.93–0.96)	2.26E−16	862613 (39)	9.3E−15	1084020 (44)	0.96 (0.96–0.97)	3.8E−32	*SRP54*	Peptide chain elongation
rs4906263	14:103249127	C/G (0.65)	1.06 (1.04–1.07)	2.65E−12	693031 (37)	1.5E−10	702407 (41)	1.04 (1.03–1.04)	2.9E−36	*TRAF3*	Cytokine signalling in immune system
rs2041733	16:11229589	C/T (0.54)	0.92 (0.91–0.93)	7.85E−36	864982 (40)	5.8E−40	1086390 (45)	0.94 (0.94–0.95)	4.2E−95	*RMI2*	DNA repair
rs1358175	17:38757789	T/C (0.63)	1.05 (1.03–1.06)	1.99E−11	864982 (40)	1.4E−14	1086390 (45)	1.03 (1.03–1.04)	1.2E−26	*CCR7*	B and T lymphocyte activation
rs17881320	17:40485239	T/G (0.08)	1.09 (1.07–1.12)	5.34E−13	862032 (38)	2.0E−11	870142 (41)	1.07 (1.06–1.08)	9.8E−39	*STAT3*[b]	Cytokine signalling in immune system
rs4247364	17:43336687	C/G (0.70)	0.96 (0.95–0.98)	4.54E−08	862470 (39)	1.3E−07	1083880 (44)	0.97 (0.97–0.98)	1.7E−17	*DCAKD*[b]	Coenzyme A biosynthetic process
rs56308324	17:45819206	T/A (0.13)	1.06 (1.04–1.08)	4.89E−10	860694 (38)	1.1E−08	1082110 (43)	1.03 (1.02–1.04)	2.6E−11	*TBX21*[b]	Th1 differentiation
rs28406364	17:47454507	T/C (0.38)	1.06 (1.05–1.07)	5.01E−18	864982 (40)	2.3E−18	1086390 (45)	1.04 (1.03–1.04)	1.5E−34	*GNGT2*	G protein signalling
rs2967677	19:8789721	T/C (0.15)	1.08 (1.07–1.10)	3.35E−20	861624 (38)	5.8E−23	1083040 (43)	1.06 (1.05–1.07)	7.5E−49	*CERS4*	Sphingolipid metabolism
rs6062486	20:62302539	A/G (0.69)	1.09 (1.07–1.10)	5.03E−30	782263 (37)	4.4E−32	1003680 (42)	1.07 (1.07–1.08)	4.5E−109	*RTEL1*	DNA repair
rs4821569	22:37316873	G/A (0.53)	1.05 (1.04–1.06)	3.14E−13	863063 (39)	1.6E−11	1084480 (44)	1.04 (1.04–1.05)	5.4E−50	*CSF2RB*	Cytokine signalling in immune system

The lead SNP at each independent locus is displayed, along with the results from the European-only discovery, multi-ancestry discovery and European replication. The top ranked gene from our gene prioritisation is listed, along with a description of the pathway/function of the gene. The evidence implicating each gene is presented in Supplementary Data 11.

Alleles are listed as effect allele/other allele, the effect allele frequency (EAF) in Europeans (average EAF, weighted by the sample size of each cohort).

Association statistics, Odds ratios (with 95% confidence intervals) and (unadjusted, two-sided) *P*-values are displayed for the fixed effects European-only meta-analysis and the replication analysis. *P*-values (unadjusted, two-sided) only are available from the MR-MEGA meta-regression multi-ancestry analysis.

Genome build = GRCh37/hg19.

[a]Imputation batch effect observed in 23andMe data.

[b]One of two or three tied genes at these loci are shown.

https://doi.org/10.1038/s41467-023-41180-2

Table 2 | Novel genome-wide significant loci in European-only analysis

			European Discovery			Multi-ancestry discovery		23andMe European replication (*N* = 2,904,664)			
Variant	**Chr:position**	**Alleles (EAF)**	**OR (CI)**	***P***	**N (studies)**	***P***	**N (studies)**	**OR (CI)**	***P***	**Gene**	**Pathway**
rs301804[b]	1:8476441	G/C (0.30)	1.05 (1.03–1.07)	2.3E−09	698266 (39)	8.5E−09	707642 (43)	1.03 (1.02–1.03)	5.5E−16	*RERE*	Apoptosis
rs61776548	1:12091024	A/G (0.47)	1.04 (1.02–1.05)	4.2E−08	787144 (39)	1.4E−07	1008560 (44)	1.02 (1.01–1.02)	5.6E−09	*TNFRSF1B*	Cytokine signalling in immune response
rs12565349	1:110371629	G/C (0.15)	1.05 (1.03–1.07)	1.3E−08	862259 (39)	1.9E−07	1083670 (44)	1.03 (1.02–1.04)	5.8E−15	*CSF1*	Cytokine signalling in immune response
rs187080438	1:150374354	T/C (0.03)	1.17 (1.11–1.23)	3.7E−10	758729 (20)	2.2E−12	765209 (22)	1.14 (1.12–1.16)	2.0E−41	*CTSS*	Antigen presentation in immune response
rs146527530[b]	1:151059196	G/T (0.02)	1.27 (1.20–1.35)	5.5E−15	744128 (13)	7.4E−19	744128 (13)	1.25 (1.22–1.28)	1.5E−88	*CTSS*	Antigen presentation in immune response
rs115161931[b]	1:151063299	T/C (0.04)	1.18 (1.13–1.23)	1.0E−13	472565 (26)	3.2E−12	479045 (28)	1.09 (1.08–1.11)	2.0E−32	*CTSS*	Antigen presentation in immune response
rs71625130[b]	1:151625094	A/G (0.04)	1.23 (1.18–1.28)	2.4E−27	770827 (25)	7.2E−30	772963 (26)	1.17 (1.16–1.19)	1.7E−89	*RORC*[c]	Cytokine signalling in immune response
rs149199808[b]	1:151626396	T/C (0.03)	1.32 (1.26-1.38)	4.4E−30	756174 (19)	8.7E−34	762654 (21)	1.24 (1.22–1.26)	3.1E−134	*RORC*	Cytokine signalling in immune response
rs821429[b]	1:153275443	A/G (0.96)	0.86 (0.84–0.89)	5.9E−18	852224 (30)	8.2E−16	858704 (32)	0.91 (0.89–0.92)	2.7E−38	*S100A7*	Differentiation regulation incl. in the innate immune system
rs12138773	1:153843489	A/C (0.03)	1.11 (1.07–1.16)	2.3E−08	851937 (28)	1.3E−09	858417 (30)	1.07 (1.05–1.09)	3.5E−16	*S100A12*[c]	Regulation of inflammatory processes and immune response
rs67766926[a,b]	2:61163581	G/C (0.23)	1.05 (1.03–1.06)	5.7E−10	863063 (39)	2.9E−11	1084480 (44)	1.05 (1.04–1.05)	1.2E−41	*AHSA2P*	Protein folding
rs112385344	2:112275538	T/C (0.12)	1.06 (1.04–1.08)	2.8E−09	852837 (34)	3.9E−08	862213 (38)	1.04 (1.03–1.05)	1.5E−18	*MERTK*[c]	Inhibits TLR-mediated innate immune response
rs62193132	2:242788256	T/C (0.46)	1.04 (1.03–1.06)	1.5E−09	832761 (26)	7.1E−08	1052040 (30)	1.03 (1.02–1.03)	1.5E−19	*NEU4*	Sphingolipid metabolism
rs10833[b]	4:142654547	C/T (0.65)	1.04 (1.03–1.06)	7.3E−09	859747 (38)	6.0E−08	1081160 (43)	1.02 (1.02–1.03)	3.4E−15	*IL15*	Cytokine signalling in immune response
rs148161264[b]	5:14604521	G/C (0.04)	1.10 (1.07–1.14)	7.4E−10	850619 (29)	2.0E−08	857099 (31)	1.05 (1.03–1.06)	1.6E−08	*OTULINL*	Endoplasmic reticulum component
rs7701967	5:130059750	A/G (0.31)	0.95 (0.94–0.97)	3.4E−09	520344 (36)	3.6E−09	529720 (40)	0.99 (0.98–0.99)	1.1E−06	*LYRM7*	Mitochondrial respiratory chain complex assembly
rs4532376[b]	5:176774403	A/G (0.30)	1.04 (1.03–1.06)	3.5E−09	859747 (38)	2.3E−09	1081160 (43)	1.03 (1.02–1.03)	1.4E−18	*RGS14*	G-alpha signalling
rs72925996[b]	6:90930513	C/T (0.33)	0.96 (0.94–0.97)	3.2E−10	862259 (39)	5.4E−09	1083670 (44)	0.96 (0.95–0.96)	2.2E−44	*BACH2*	NF-kappaB proinflammatory signalling
rs989437	7:28830498	G/A (0.61)	0.96 (0.95–0.97)	6.1E−11	864982 (40)	1.0E−09	1086390 (45)	0.97 (0.96–0.97)	6.9E−31	*CREB5*[c]	AMPK & ATK signalling
rs34215892	8:21767240	A/G (0.03)	0.87 (0.83–0.90)	4.7E−11	436369 (24)	2.0E−09	442849 (26)	0.89 (0.88–0.91)	1.0E−36	*DOK2*	Immune response IL-23 signalling
rs118162691	8:21767809	A/C (0.05)	0.92 (0.89–0.94)	7.8E−09	856229 (30)	1.8E−07	862709 (32)	0.90 (0.88–0.91)	1.1E−44	*DOK2*	Immune response IL-23 signalling
rs7843258	8:141601542	C/T (0.82)	1.05 (1.04–1.07)	1.5E−09	859747 (38)	3.6E−10	1081160 (43)	1.04 (1.03–1.05)	7.0E−25	*AGO2*	siRNA-mediated gene silencing
rs7857407	9:33430707	A/T (0.40)	1.04 (1.02–1.05)	2.5E−08	864982 (40)	9.0E−09	1086390 (45)	1.03 (1.02–1.03)	5.1E−18	*AQP3*	Aquaporin-mediated transport
rs10988863	9:102331281	C/A (0.21)	0.95 (0.93–0.96)	5.1E−11	862259 (39)	3.0E−09	1083670 (44)	0.97 (0.97–0.98)	1.3E−13	*NR4A3*	Transcriptional activator
rs17368814	11:102748695	G/A (0.13)	0.95 (0.93–0.97)	1.4E−08	858117 (37)	6.8E−07	1078260 (41)	0.95 (0.95–0.96)	1.2E−27	*MMP12*	Extracellular matrix organisation
rs11216206	11:116843425	G/C (0.07)	1.10 (1.07–1.14)	5.5E−10	557183 (35)	2.9E−10	778595 (40)	1.04 (1.03–1.05)	8.5E−15	*SIK3*	LKB1 signalling
rs5005507[b]	12:94611908	C/G (0.74)	1.05 (1.03–1.06)	3.6E−09	859747 (38)	9.6E−08	1081160 (43)	1.03 (1.02–1.04)	2.7E−18	*PLXNC1*	Semaphorin interactions incl. in immune response
rs7147439	14:105523663	A/G (0.73)	0.96 (0.95–0.97)	4.7E−08	781909 (37)	6.6E−07	1003320 (42)	0.97 (0.96–0.97)	4.8E−24	*GPR132*	GPCR signalling
rs2542147	18:12775851	T/G (0.84)	0.95 (0.93–0.96)	1.5E−09	862470 (39)	7.5E−08	1083880 (44)	0.96 (0.95–0.97)	2.6E−26	*PTPN2*	Cytokine signalling in immune response

The lead SNP at each independent locus is displayed, along with the results from the European-only discovery, multi-ancestry discovery and European replication. The top-ranked gene from our gene prioritisation is listed, along with a description of the pathway/function of the gene. The evidence implicating each gene is presented in Supplementary Data 11.

Alleles are listed as Effect allele/other allele, the effect allele frequency (EAF) in Europeans (average EAF, weighted by the sample size of each cohort).

Genome build = GRCh37/hg19.

[a]rs4643526 at the same locus was previously identified in the discovery analysis of Paternoster et al. [2]. However, this association did not replicate in that study.

[b]Whilst not identified in any GWAS AD papers, these loci have previously shown evidence for association with AD in supplementary material of methodological papers[92,93].

[c]One of two or three tied genes at these loci are shown.

https://doi.org/10.1038/s41467-023-41180-2

Table 3 | Additional loci associated with the multi-ancestry analysis

Variant	Chr:position	Alleles (EAF)	Multi-ancestry discovery *N* = 992,907 *P*	European discovery *N* = 864,982 *P*	RIKEN - Biobank Japan *N* = 118,287 *P*	23andMe Latino *N* = 525,348 *P*	23andMe African *N* = 174,015 *P*	23andMe European *N* = 2,904,664 *P*	Known Associations	Novel Associations
rs114059822[a]	1:19804918	T/G (0.03)	8.59E−09	0.25	–	0.07	0.03	0.87	NA	NA
rs9247	2:234113301	T/C (0.21)	1.92E−09	7.32E−08	7.71E−05	1.49E−13	7.23E−03	2.93E−51		**all**[b]
rs9864845	3:112383847	A/G (0.37)	2.17E−12	0.22	3.92E−13	0.75	0.23	0.12	Japanese (Tanaka et al.[8])	
rs34599047	6:106629690	C/T (0.18)	3.32E−08	1.29E−07	0.03	7.18E−04	0.02	3.23E−22		**all**[b]
rs7773987	6:135707486	T/C (0.60)	1.22E−08	9.57E−08	0.15	0.18	1.95E−03	5.93E−13		**European, African**
rs118029610[a]	9:1894613	T/C (0.03)	1.89E−08	2.97E−04	–	0.5	0.31	0.78	NA	NA
rs117137535	9:140500443	A/G (0.03)	1.99E−08	5.50E−08	–	3.99E−07	0.33	9.25E−19	European (Grosche et al.[7])	Latino
rs4312054	11:7977161	G/T (0.43)	3.21E−12	0.86	3.46E−15	0.4	0.33	0.52	Japanese (Tanaka et al.[8])	
rs150113720[a]	11:83439186	G/C (0.02)	5.52E−10	0.40	–	0.1	0.22	0.14	NA	NA
rs115148078[a]	11:101361300	T/C (0.02)	5.91E−09	0.37	–	3.69E−03	0.91	0.89	NA	NA
rs4262739	11:128421175	A/G (0.50)	2.20E−08	6.03E−07	2.28E−03	1.89E−06	0.09	1.45E−36	European & Japanese (Tanaka et al.[8])	Latino
rs1059513	12:57489709	C/T (0.08)	5.15E−09	1.57E−07	0.33	3.06E−04	0.17	6.95E−16	European (Tanaka et al.[8])	Latino
rs4574025	18:60009814	T/C (0.55)	7.00E−10	1.48E−06	2.67E−05	2.59E−04	1.24E−05	2.96E−05	European & Japanese (Tanaka et al.[8])	Latino, African
rs6023002	20:52797237	C/G (0.52)	4.05E−10	2.26E−06	2.82E−07	5.96E−03	0.07	3.22E−28	European & Japanese (Tanaka et al.[8])	Latino

For loci that were associated in the multi-ancestry discovery analysis, but not the European discovery analysis, we show the (unadjusted two-sided) P-values for association across 4 diverse ancestral groups, European, Japanese, Latino and African. Full association statistics (including OR and 95% CI) for each variant can be viewed in Supplementary Data 4 (and results across all cohorts individually are depicted in Supplementary Fig. 2).
Alleles are reported as effect allele/other allele.
Genome build = GRCh37/hg19.
NA indicates finding not replicated and likely to be false-positive in discovery.
Bold is used in the novel column to denote the 3 associations that are entirely novel (i.e. locus has not been associated in any ancestry previously).
– Variant was not available in dataset.
[a]Genome-wide significant loci without replication that are assumed to be false positives in the discovery data.
[b]Whilst not identified in any GWAS AD papers, these loci have previously shown evidence for association with AD in the supplementary material of methodological papers[9,2] or GWAS of combined allergic disease phenotype[5].

the replication samples (of European, Latino and/or African ancestry; Bonferroni corrected $P < 0.05/14 = 3.6 \times 10^{-3}$). Two index SNPs which did not replicate in any of the samples (rs9864845 (near *CCDC80*), rs4312054 (near *NLRP10*)) appear to have been driven by association in the Japanese RIKEN study only (Supplementary Data 4, Supplementary Figs. 2, 3). Whilst the allele frequencies of these index SNPs are similar between Europeans and Japanese (37% vs 42% for rs9864845, 41% vs 46% for rs4312054, Supplementary Data 5), in a multi-ancestry fixed effect meta-analysis at both these loci there were neighbouring (previously reported)[8] SNPs with stronger evidence of association (rs72943976, $P = 2 \times 10^{-9}$ and rs59039403 $P = 2 \times 10^{-35}$, Supplementary Fig. 3), that did show large allele frequencies for Japanese (~34% and 13%, respectively) but <1% in Europeans. A further 4 loci did not replicate, and on closer examination (Supplementary Fig. 2, and MAF in cases <1%), their association in the discovery analysis appeared to be driven by a false positive outlying result in a single European cohort.

Seven of the loci in Table 3 have been previously reported as associated with AD. Two (rs117137535 (near *ARRDC1*)[7] and rs1059513 (near *STAT6*)[8]) were previously only associated with Europeans (and these were variants that were just below the genome-wide significance threshold in our European only analysis). Three (rs4262739 (near *ETS1*), rs4574025 (within *TNFRSF11A*) and rs6023002 (near *CYP24A1*)) were previously associated in Japanese and Europeans[8], while 2 were previously associated only in Japanese[8,10], using the same Japanese data (RIKEN) that we include here. Therefore, in our multi-ancestry analysis (and replication) we identify 3 loci that have not previously been reported in a GWAS of AD of any ancestry (rs9247 (near *INPP5D*), rs34599047 (near *ATG5*) and rs7773987 (near *AHI1*)), all of which are associated in two or more populations in our data (Table 3).

In addition, for 5 loci which had previously been associated with individuals of European and/or Japanese ancestry, we now show evidence that these are also associated with individuals of Latino ancestry and one is also associated in individuals of African ancestry (Table 3).

Comparison of associations between ancestries

Effect sizes of the index SNPs were remarkably similar between individuals of European and Latino ancestry (Supplementary Fig. 4A). There were only two variants with any evidence for a difference (where Latino $P > 5 \times 10^{-4}$ and the 95% confidence intervals didn't overlap), but the plot shows that these were only marginally different and likely to be due to chance. Effect size comparison of the index SNPs between individuals of European and African ancestry showed greater differences (Supplementary Fig. 4B). 17 SNPs showed some evidence for being European-specific in that comparison. The confidence intervals in the Japanese data were much wider but there was weak evidence for one SNP being European-specific and stronger evidence for two SNPs being Japanese-specific (Supplementary Fig. 4C). These were rs4312054 (JAP CI: 0.75-0.84, EUR CI: 0.99-1.01) and rs9864845 (JAP CI: 1.16-1.30, EUR CI: 0.99-1.06), mentioned earlier as the SNPs that appeared to be driven only by Japanese individuals in the multi-ancestry meta-analysis (Supplementary Data 4).

Established associations

A review of previous work in this field (Supplementary Data 1) shows that a total of 202 unique variants (across a much smaller number of loci) have been reported to be associated with AD. We found evidence for all but 7 variants of these being nominally associated in the current GWAS (81% in the European and 96% in the multi-ancestry analysis). Variants we did not find to be associated were either rare variants (MAF < 0.01), or insertion/deletion mutations, which were not included in our analysis.

Genetic correlation between AD and other traits

LD score regression analyses showed high genetic correlation, as expected, between AD and related allergic traits, e.g. asthma (rg=0.53, $P = 2 \times 10^{-32}$), hay fever (rg=0.51, $P = 7 \times 10^{-17}$) and eosinophil count (rg = 0.27, $P = 1 \times 10^{-7}$) (Supplementary Fig. 5 and Supplementary Data 6). In addition, depression and anxiety showed notable

genetic correlation with AD (rg = 0.17, $P = 2 \times 10^{-7}$), a relationship which has been reported previously, but causality has not been established[17]. Furthermore, gastritis also showed substantial genetic correlation (rg = 0.31, $P = 1 \times 10^{-5}$), which may be due to the AD genetic signal including variants with pervasive inflammatory function or the observed correlation could indicate a shared risk locus for inflammation or microbiome alteration in the upper gastrointestinal tract, or it may reflect the use of systemic corticosteroid treatment for atopic disease which in some cases causes gastritis as a side effect.

Tissue, cell and gene-set enrichment

The tissue enrichment analyses using distinct molecular evidence (representing open chromatin and gene expression) both found blood to be the tissue showing strongest enrichment of GWAS loci (Fig. 2). The Garfield test for enrichment of genome-wide loci (with $P < 1 \times 10^{-8}$) in DNase I hypersensitive sites (DHS broad peaks) found evidence of enrichment ($P < 0.00012$) in 41 blood tissue analyses, a greater signal than another tissue or cell type (Fig. 2a and Supplementary Data 7). The strongest enrichment (OR > 5.5 and $P < 1 \times 10^{-10}$) was seen for T-cell, B-cell and natural killer lymphocytes (CD3+, CD4+, CD56+ and CD19+). As expected for AD, Th2 showed stronger enrichment (OR = 4.3, $P = 1 \times 10^{-8}$) than Th1 (OR = 2.3, $P = 2 \times 10^{-4}$). The strongest enrichment in tissue samples representing skin was seen for foreskin keratinocytes (OR = 2.0, $P = 0.008$), but this did not meet a Bonferroni-corrected *P*-value threshold ($0.05/425 = 1 \times 10^{-4}$).

The most enriched tissue type in MAGMA gene expression enrichment analysis was whole blood ($P = 2 \times 10^{-14}$). Others that met our Bonferroni-corrected *P*-value ($P < 0.0009$) were spleen, EBV-transformed lymphocytes, sun-exposed and unexposed skin, small intestine and lung (Fig. 2b and Supplementary Data 8).

DEPICT cell-type enrichment analysis identified a similar set of enriched cell-types: blood, leucocytes, lymphocytes and natural killer cells, but with the addition that the strongest enrichment was seen for synovial fluid ($P = 2 \times 10^{-7}$), which may be due to its immune cell component.

The DEPICT pathway analysis found 420 GO terms with enrichment (FDR < 5%) amongst the genes from our GWAS loci (Supplementary Data 9). The pathway with the strongest evidence of enrichment was 'hemopoietic or lymphoid organ development' ($P = 1 \times 10^{-16}$). All terms with FDR < 5% are represented in Supplementary Fig. 6, where the terms are grouped according to similarity and the parent terms labelled illustrating the strong theme of immune system development and signalling.

Gene prioritisation and biological interpretation in silico

The top genes prioritised using our composite score from publicly available data for each of the established European AD loci are shown in Table 1 and Fig. 3a (and the evidence that makes up the prioritisation scores is shown in Supplementary Fig. 7). The top three prioritised genes at each independent locus are shown in Supplementary Data 10 and a summary of all evidence for all genes reviewed in silico is presented in Supplementary Data 11.

In most cases the top prioritised gene had been implicated (in previous GWAS) or is only superseded marginally by an alternative candidate. One interesting exception is on chromosome 11, where *MAP3K11* (with a role in cytokine signalling – regulating the JNK signalling pathway) is markedly prioritised over the previously implicated *OVOL1*[18] (involved in hair formation and spermatogenesis), although the prioritisation of *MAP3K11* is predominantly driven by TWAS evidence in multiple cell types rather than colocalisation or other evidence.

There are three instances where multiple associations in the region implicate additional novel genes. Two are genes involved in TLR4 signalling: *S100A9* (prioritised in addition to the established *FLG* and *IL6R* on chromosome 1) and *AGER* (prioritised in addition to *HLA-DRA* on chromosome 6). The third has a likely role in T-cell activation: *CDC42SE2* (prioritised in addition to *SLC22A5* on chromosome 5).

The top prioritised gene at each of the novel European loci are shown in Table 2 and Fig. 3b. Many are in pathways already identified by previous findings (e.g. cytokine signalling—specially IL-23, antigen presentation and NF-kappaB proinflammatory response). At one locus, the index SNP, rs34215892 is a missense (Pro274Leu) mutation within the *DOK2* gene, although this mutation is categorised as tolerated or benign by SIFT and PolyPhen. The genes with the highest prioritisation score amongst the novel loci were *GPR132* (total evidence Score=24), *NEU4* (score=22), *TNFRSF1B* (score = 19) and *RGS14* (score=19) and each show biological plausibility as candidates for AD pathogenesis.

GPR132 is a proton-sensing transmembrane receptor, involved in modulating several downstream biological processes, including immune regulation and inflammatory response, as reported previously in an investigation of this protein's role in inflammatory bowel disease[19]. The index SNP at this locus, rs7147439 (which was associated with Europeans, Latinos, Africans, but not Japanese), is an intronic variant within the *GPR132* gene. The AD GWAS association at this locus colocalises with the eQTL association for *GPR132* in several immune cell types (macrophages[20], neutrophils[21], several T-cell datasets[22]) as well as in colon, lung and small intestine in GTEx[23]. *GPR132* has also been shown to be upregulated in lesional and nonlesional skin in AD patients, compared to skin from control individuals[24,25]. OpenTargets and POSTGAP both prioritise *GPR132* for this locus.

The SNP rs62193132 (which showed consistent effects in European, Latino and Japanese individuals, but little evidence for association in African individuals, Supplementary Fig. 2), is in an intergenic region between *NEU4* (~26 kb) and *PDCD1* (~4 kb away) on chromosome 2. *NEU4* was the highest scoring in our gene prioritisation pipeline (score=22). However, *PDCD1* also scores highly (score = 18, Supplementary Data 10). NEU4 is an enzyme that removes sialic acid residues from glycoproteins and glycolipids, whereas PDCD1 is involved in the regulation of T cell function. The AD GWAS association at this locus colocalises with the eQTL for *NEU4* in several monocyte and macrophage datasets[22,26–28] as well as in the ileum, colon and skin[23,29]. The eQTL for *PDCD1* also colocalises in monocytes and macrophages[27,28] as well as T-cells[22], skin and whole blood[23]. In addition to the eQTL evidence, *PCDC1* is upregulated in lesional and non-lesional skin in AD patients compared to skin from control individuals[24,25]. OpenTargets and PoPs prioritise *NEU4*, whilst POSTGAP prioritises *PDCD1* at this locus.

TNFRSF1B is part of the TNF receptor, with an established role in cytokine signalling. rs61776548 (which showed consistent associations across all major ancestries tested) is 136 kb upstream of *TNFRSF1B*, actually within an intron of *MIIP*. *MIIP* encodes Migration and Invasion-Inhibitory Protein, which may function as a tumour suppressor. However, *TNFRSF1B* is a stronger candidate gene since the AD GWAS association at this locus colocalises with the eQTL for *TNFRSF1B* T cells[22,30], macrophages[20], fibrobasts[31] and platelets[29]. Furthermore, *TNFRSF1B* gene expression and the corresponding protein are upregulated in lesional and nonlesional skin compared to controls[24,25,32] and the PoPs method prioritised this gene at this locus.

RGS14 is a multifunctional cytoplasmic-nuclear shuttling protein which regulates G-protein signalling, but whose role in the immune system is yet to be established. rs4532376 is 10.5 kb upstream of *RGS14* and within an intron of *LMAN2*. The AD GWAS association at this locus colocalises with the eQTL for *RGS14* in macrophages[20], CD8 T-cells[22], blood[33] and colon[23]. *RGS14* has also been shown to be upregulated in lesional skin of AD cases compared to skin from control individuals[25] and DEPICT prioritises this gene. However, at this locus *LMAN2* is also a reasonably promising candidate (score=15) based on colocalisation and differential expression evidence (Supplementary Data 11). OpenTargets and POSTGAP prioritise this alternative gene at this locus and

https://doi.org/10.1038/s41467-023-41180-2

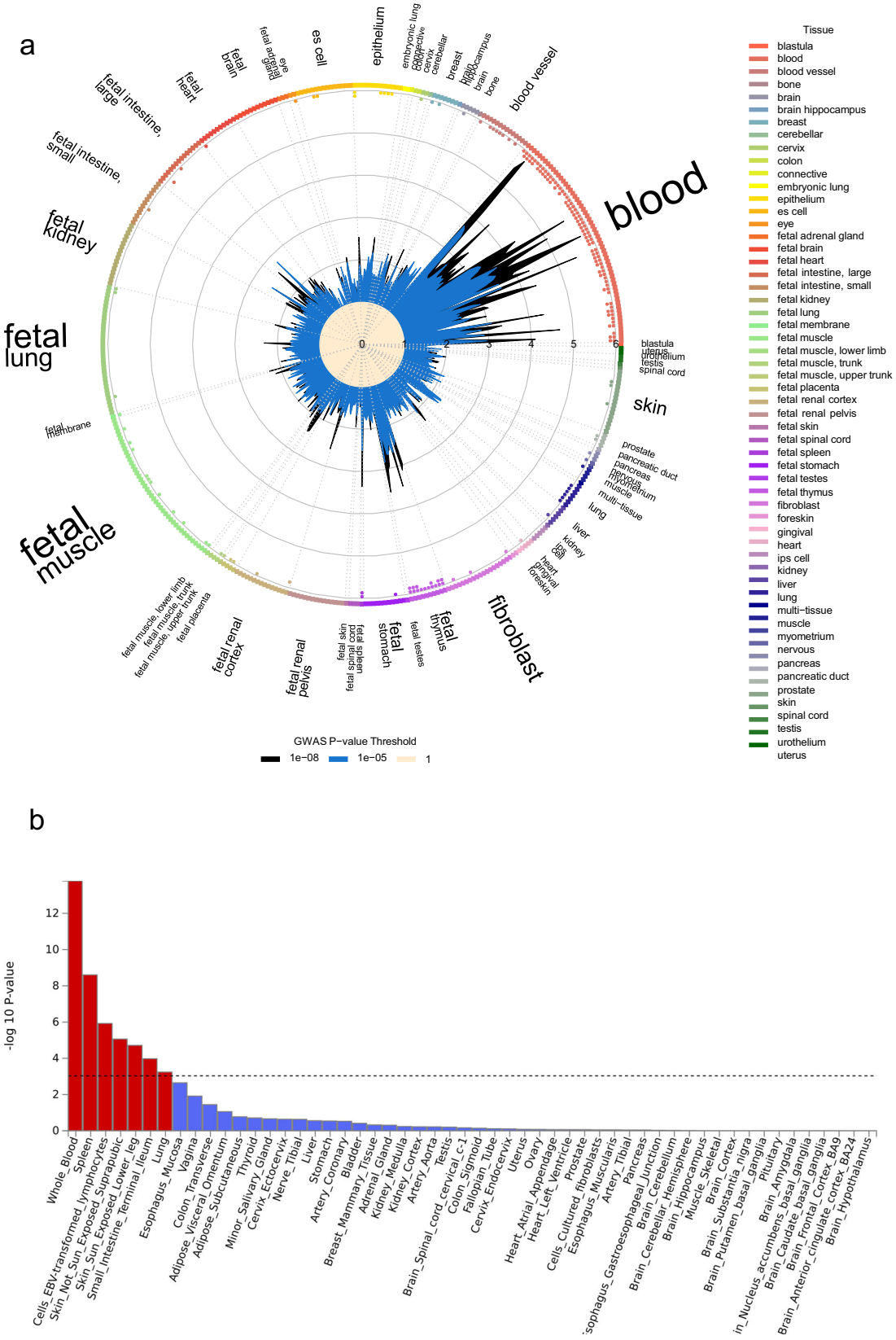


Fig. 2 | Cell type tissue enrichment analysis. a GARFIELD enrichment analysis of open chromatin data. Plot shows enrichment for AD associated variants in DNase I Hypersensitive sites (broad peaks) from ENCODE and Roadmap Epigenomics datasets across cell types. Cell types are sorted and labelled by tissue type. ORs for enrichment are shown for variants at GWAS thresholds of $P < 1 \times 10^{-8}$ (black) and $P < 1 \times 10^{-5}$ (blue) after multiple-testing correction for the number of effective annotations. Outer dots represent enrichment thresholds of $P < 1 \times 10^{-5}$ (one dot) and $P < 1 \times 10^{-6}$ (two dots). Font size of tissue labels corresponds to the number of cell types from that tissue tested. **b** MAGMA enrichment analysis of gene expression data. Plot shows *P*-value for MAGMA enrichment for AD associated variants with gene expression from 54 GTEx ver.8 tissue types. The enrichment $-\log_{10}(P\text{-value})$ for each tissue type is plotted on the *y*-axis. The Bonferroni corrected threshold $P = 0.0009$ is shown as a dotted line and the 7 tissue types that meet this threshold are highlighted as red bars.

https://doi.org/10.1038/s41467-023-41180-2

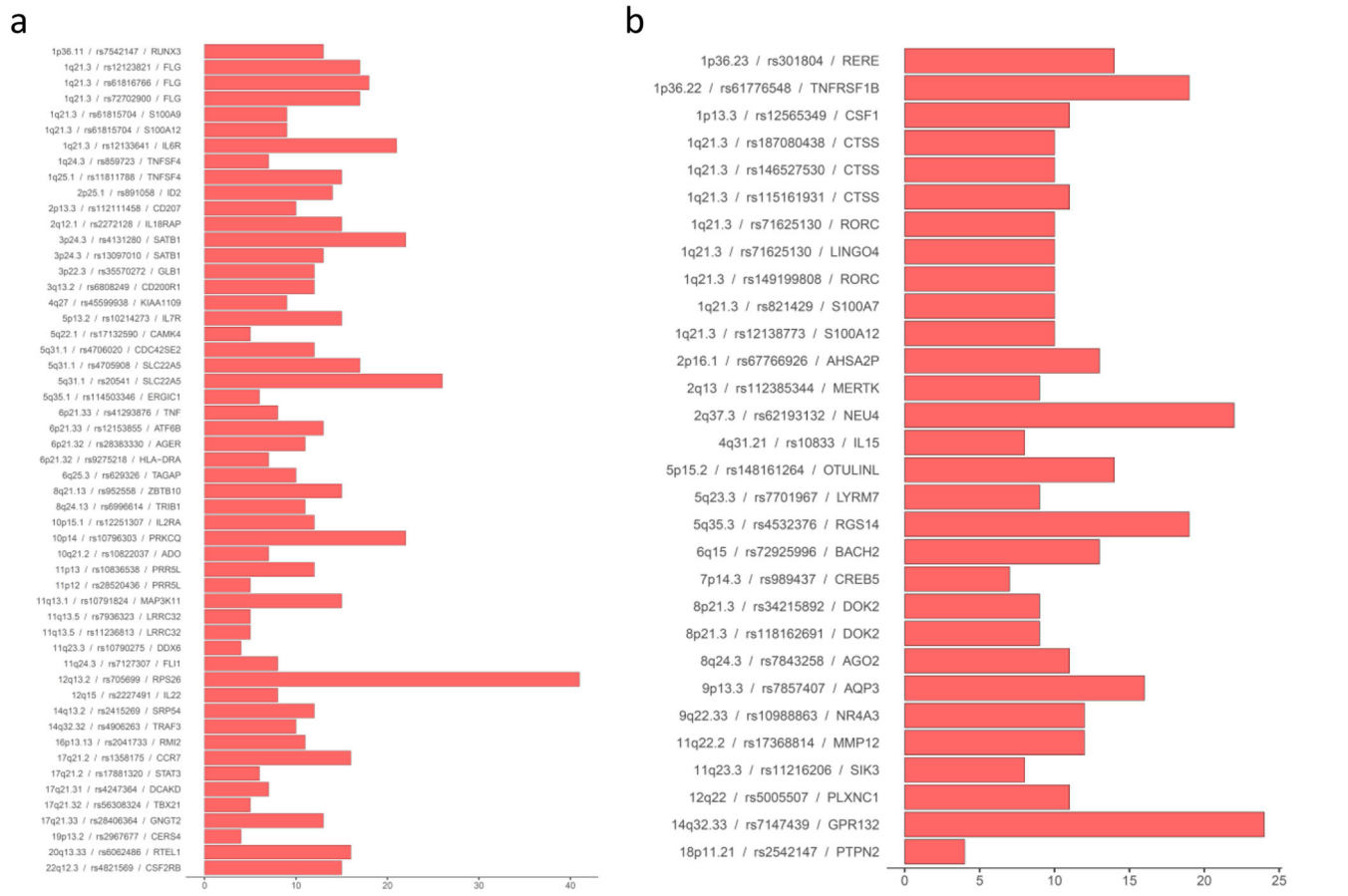


Fig. 3 | Prioritised genes at GWAS loci. Prioritised genes at known (**a**) and novel (**b**) loci. For each independent GWAS locus the top prioritised gene (or genes if they were tied) from our bioinformatic analysis is presented along with a bar representing the total evidence score for that gene. A more detailed breakdown of the constituent parts of this evidence score is presented in Supplementary Fig. 5 and the total evidence scores for the top 3 genes at each locus are presented in Supplementary Data 10. NB. There are some cases of two independent GWAS signals implicating the same gene.

it is possible that genetic variants at this locus influence AD risk through both genetic mechanisms.

We did not include the 3 novel variants from the multi-ancestry analysis in the comprehensive gene prioritisation pipeline because the available resources used predominantly represent European samples only. We did however investigate these variants using Open Targets Genetics, to identify any evidence implicating specific genes at these loci. rs9247 is a missense variant in *INPP5D*, encoding SHIP1, a protein that functions as a negative regulator of myeloid cell proliferation and survival. The *INPP5D* gene has been implicated in hay fever and/or eczema[5] and other epithelial barrier disorders including inflammatory bowel disease. rs7773987 is intronic for *AHI1 (*Abelson helper integration site 1) which is involved with brain development but expressed in a range of tissues throughout the body; single cell analysis in skin shows expression in multiple cell types including specialised immune cells and keratinocytes, but the highest abundance is in endothelial cells (data available from v21.1 proteinatlas.org). The closest genes to rs34599047 are *ATG5* (involved in autophagic vesicle formation) and *PRDM1* (which encodes a master regulator of B cells).

Network analysis

STRING network analysis of the 70 human proteins encoded by genes listed in Tables 1 and 2 showed a protein-protein interaction (PPI) enrichment p-value $< 1 \times 10^{-16}$. The five most highly significant (FDR $P = 1 \times 10^{-9}$) Gene Ontology (GO) terms for biological process relate to immune system activation and regulation (Supplementary Data 12). The network described by the highly enriched term 'Regulation of immune system process' (GO:0002682) is shown in Fig. 4.

Extending the network to include the less well characterised genes/proteins from the multi-ancestry analysis further strengthened this predicted network: The PPI enrichment was again $P < 1 \times 10^{-16}$ and 'Regulation of immune system process' was the most enriched term (FDR $P = 5 \times 10^{-13}$).

Discussion

We present the results of a comprehensive genome-wide association meta-analysis of AD in which we have identified a total of 91 associated loci. This includes 81 loci identified amongst individuals of European ancestry replicated in a further sample of 2.9 million European individuals (as well as many showing replication in data for other ancestries). Of the additional 10 loci identified in a multi-ancestry analysis, 8 replicated in at least one of the populations tested (European, Latino and African ancestry) and a further 2 may be specific to individuals of East Asian ancestry (but require replication).

The majority of the loci associated with AD are shared between the ancestry groups represented in our data, though there were some notable exceptions. We report two previously identified loci with associations that appear to be specific to the Japanese cohort (although driven by just one cohort and still require independent replication). Whilst these have been previously reported[8], this used the same data as examined here. However, rs59039403 within *NLRP10* is a likely deleterious missense mutation at reasonable frequency in Japanese (13%) that is present at a far lower frequency (<1%) in Europeans. Equally, previous further investigation of the association near *CCDC80* found a putative functional variant (rs12637953) that affects the expression of an enhancer (associated with *CCDC80* promoter) in

https://doi.org/10.1038/s41467-023-41180-2

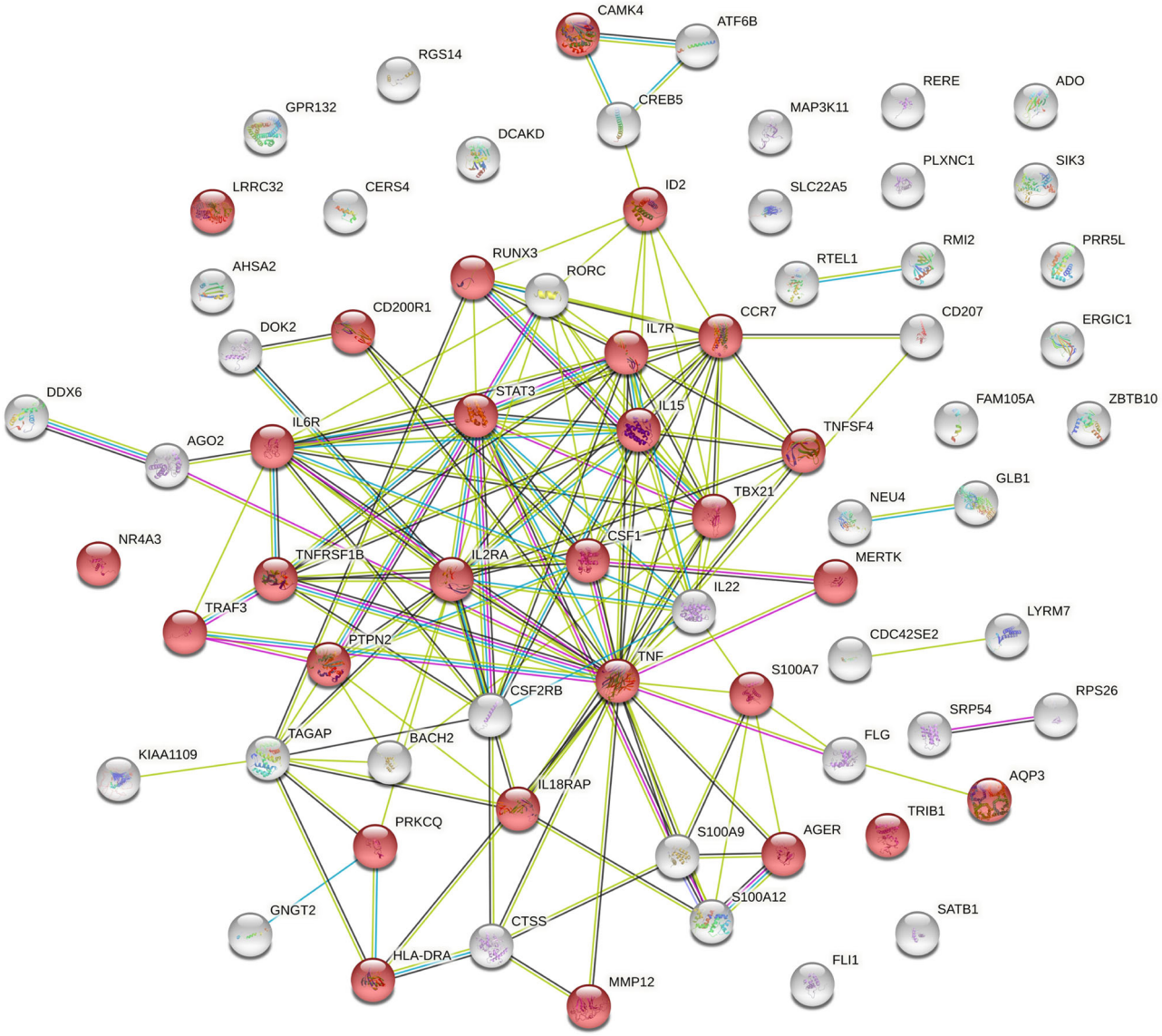


Fig. 4 | Predicted interaction network of proteins encoded by the top prioritised genes from known and novel European GWAS loci. Protein-protein interaction analysis carried out in STRING v11.5; nodes coloured red represent the GO term 'Regulation of immune system process' (GO:0002682) for which 28/1514 proteins are included (FDR $P = 1 \times 10^{-9}$). Full results for all identified pathways are available in Supplementary Data 12.

epidermis and Langerhans cells[8], increasing the evidence that these Japanese-specific loci are real. Furthermore, we have identified several loci with association in Europeans (many of which also showed association in individuals of Japanese or Latino ancestry) but which showed no evidence of association in individuals of African ancestry. It is tempting to speculate, using our knowledge of the differing AD phenotypes between European, Asian and African people[34,35] that the differing genetic associations at some loci may contribute to these clinical observations. rs7773987 within an intron of *AHI1* may, for example, indicate a mechanism contributing to neuronal sensitisation leading to the marked lichenification and nodular prurigo-type lesions[36] that characterise AD in some people of African and European ethnicities[37]. Large-scale population cohorts (as used here) have been useful for identifying associated variants. However, we do note that the variants identified should be further examined with respect to specific aspects of AD (age of onset, severity and longitudinal classes[38]) in future analysis.

The dominance of blood as the tissue showing most enrichment of our GWAS signals in regions of DNAse hypersensitivity and of eQTLs suggests the importance of systemic inflammation in AD and this is in keeping with knowledge of the multisystem comorbidities associated with AD[39]. The dominance of blood also supports the utility of this easily accessible tissue when characterising genetic risk mechanisms, and for the measurement of biomarkers for many of the implicated loci. However, skin tissue also showed enrichment and there are likely to be some genes for which the effect is only seen in skin. For example, we know that two genes previously implicated in AD, *FLG* and *CD207*[2,18] are predominantly expressed in the skin and in our gene prioritisation investigations there was no evidence from blood linking *FLG* to the rs61816766 association and only one analysis of monocytes separated from peripheral blood mononuclear cell (PBMC) samples[28] which implicated *CD207* for the rs112111458 association, amongst an abundance of evidence from skin for both genes playing a role in AD (Supplementary Data 11). So, whilst the enrichment analysis suggests blood as a useful tissue for genome scale studies of AD and a reasonable tissue to include for further investigation at specific loci, it does not preclude skin as the more relevant tissue for a subset of important genes.

At many of the loci identified in this GWAS, our gene prioritisation analysis, as well as the DEPICT pathway analysis, implicated genes from

pathways that are already known to have a role in AD pathology. The overwhelming majority of these are in pathways related to immune system function; STRING network analysis highlighted the importance of immune system regulation, in keeping with an increasing awareness of the importance of balance in opposing immune mechanisms that can cause paradoxical atopic or psoriatic skin inflammation[40]. Whilst our in silico *analyses* cannot definitively identify specific causal genes (rather, we present a prioritised list of all genes at each locus along with the corresponding evidence for individual evaluation), it is of note that for many of the previously known loci (Table 1) our approach identifies genes which have been validated in experimental settings, e.g. *FLG*[41], *TNF*[42] and *IL22*[43]. The individual components of the gene prioritisation analysis have their limitations, particularly the high probability that findings, whilst demonstrating correlation, do not necessarily provide evidence for a causal relationship. This has been particularly highlighted with respect to colocalisation of GWAS and eQTL associations, where high co-regulation can implicate many potentially causal genes[44]. Another limitation is that only cell types (and conditions) that have been studied and made available are included in the in silico analysis, and gaps in the data may prove crucial. However, we believe this broad-reaching review of complementary datasets and methods is a useful initial approach to summarise the available evidence, prioritise genes for follow-up and provide information to inform functional experiments. The best evidence is likely to be produced from triangulation of multiple experiments and/or datasets and we have presented our workflow and findings in a way to allow readers to make their own assessments. Another important limitation of our gene prioritisation, is that we only undertook the comprehensive approach for loci associated in European individuals, given that the majority of datasets used come from (and may only be relevant for) European individuals. Expansion of resources that allow for similarly comprehensive follow-up of GWAS loci in individuals of non-European ancestry are urgently needed[45]. However, we do report some evidence that implicates certain genes at loci from our multi-ancestry analysis, whilst noting that these require further investigation in appropriate samples from representative populations.

Amongst the genes prioritised at the novel loci identified in this study, four are targets of existing drugs (and have the required direction of action consistent with the AD risk allele's direction of effect on the gene expression) as reported by Open Targets[46]: *CSF1* is targeted by a macrophage colony-stimulating factor 1 inhibiting antibody (in phase II trials as cancer therapy but also for the treatment of rheumatoid arthritis and cutaneous lupus); *CTSS* is targeted by a small molecule cathepsin S inhibitor (in phase I-II trials for coeliac disease and Sjogren syndrome); *IL15*, targeted by an anti-IL-15 antibody (in phase II trials for autoimmune conditions including vitiligo and psoriasis); and *MMP12*, targeted by small molecule matrix metalloprotease inhibitors (in phase III studies for breast and lung cancer, plus phase II for cystic fibrosis and COPD)[47]. These may offer valuable drug repurposing opportunities.

We have presented the largest GWAS of AD to date, identifying 91 robustly associated loci, 22 with some evidence of population-specific effects. This represents a significant increase in knowledge of AD genetics compared to previous efforts, taking the number of GWAS hits identified in a single study from 31 to 91 and making available the well-powered summary statistics to enable many future important studies (e.g. Mendelian Randomization to investigate causal relationships). To aid translation we have undertaken comprehensive post-GWAS analyses to prioritise potentially causal genes at each locus, implicating many immune system genes and pathways and identifying potential novel drug targets.

Methods

Appropriate ethical approval was obtained for all cohorts by their ethics committees as detailed in the Supplementary Methods.

Phenotype definition

Cases were defined as those who have "ever had atopic dermatitis", according to the best definition for the cohort, where doctor-diagnosed cases were preferred. Controls were defined as those who had never had AD. Further details on the phenotype definitions for the included studies can be found in Supplementary Methods and Supplementary Data 2.

GWAS analysis and quality control of summary data

We performed genome-wide association analysis (GWAS) for AD case-control status across 40 cohorts including 60,653 AD cases and 804,329 controls of European ancestry. We also included cohorts with individuals of mixed ancestry (Generation R), as well as Japanese (Biobank Japan), African American (SAGE II and SAPPHIRE) and Latino (GALA II) studies, giving a total of 65,107 AD cases and 1,021,287 controls.

Genetic data was imputed separately for each cohort with the majority of European cohorts using the haplotype reference consortium (HRC version r1.1) reference panel[48] (imputed with either the Michigan or Sanger server). 8 European and 2 non-European cohorts instead used the 1000 Genomes Project Phase 1 reference panel for imputation. GWAS was performed separately for each cohort while adjusting for sex and ancestry principal components derived from a genotype matrix (as appropriate in each cohort). Genetic variants were restricted to a MAF > 1% and an imputation quality score > 0.5 unless otherwise specified in the Supplementary Methods. In order to robustly incorporate cohorts with small sample sizes, we applied additional filtering based on the expected minor allele count (EMAC) as previously demonstrated[49]. EMAC combines information on sample size, MAF and imputation quality (2*N*MAF*imputation quality score) and a threshold of >50 EMAC was used to include variants for all cohorts. QQ-plots and Manhattan plots for each cohort were generated and visually inspected as part of the quality control process.

Meta-analysis

For the discovery phase, meta-analysis of the European cohorts was performed with GWAMA[47] for 12,147,822 variants assuming fixed effects, while the multi-ancestry analysis of all cohorts was conducted in MR-MEGA[50] (which models the heterogeneity in allelic effects that is correlated with ancestry). The latter included only 8,684,278 variants as MR-MEGA excludes variants where the number of contributing cohorts is less than 6. $P < 5 \times 10^{-8}$ was used to define genome-wide significance. Clumping was performed (in PLINK 1.90[51]) to identify independent loci. We formed clumps of all SNPs which were ±500kb of each index SNP with a linkage disequilibrium $r^2 > 0.001$. Only the index SNP within each clump is reported. For multi-ancestry index variants within 500 kb of index SNPs identified in the European-only analysis, we considered these to be independent if the lead multi-ancestry SNP was not in LD ($r^2 < 0.01$) with the lead neighbouring European variant. Multi-ancestry fixed effect meta-analysis was also performed for comparison with the MR-MEGA results.

Known/Novel assignment

Novel associations are defined as a SNP that had not been reported in a previous GWAS (Supplementary Data 1), or was not correlated ($r^2 < 0.1$ in the relevant ancestry) with a known SNP from this list. In addition, following the assignment of genes to loci (see gene prioritisation) any locus annotated with a gene that has been previously reported were also moved to the 'known' list. Therefore, some loci which are reported in Open Targets[52,53] (but not reported in a published AD GWAS study) have been classed as novel. These loci are marked as such in Table 2.

Conditional analysis

Conditional analysis was performed to identify any independent secondary associations in the European meta-analysis. Genome-wide

complex trait analysis-conditional and joint analysis (GCTA-COJO[54]) was used to test for independent associations 250 kb either side of the index SNPs using UK Biobank HRC imputed data as the reference. COJO-slct was used to determine which SNPs in the region were conditionally independent (using default $P < 1 \times 10^{-5}$) and therefore represent independent secondary associations. COJO-cond was then used to condition on the top hit in each region to determine the conditional effect estimates.

Replication

The genome-wide index SNPs identified from the European and mixed-ancestry discovery meta-analyses were taken forward for replication in 23andMe, Inc. Individuals of European ($N = 2{,}904{,}664$), Latino ($N = 525{,}348$) and African ancestry ($N = 174{,}015$) were analysed separately. Full details are available in the Supplementary Methods.

LD score regression

Linkage disequilibrium score (LDSC) regression software (version 1.0.1)[55] was used to estimate the SNP-based heritability (h^2_{SNP}) for AD. This was performed with the summary statistics of the European discovery meta-analysis. The h^2_{SNP} was estimated on liability scale with a population prevalence of 0.15 and a sample prevalence of 0.070.

Genetic correlation with other traits was assessed using all the traits available on CTG-VL[56] (accessed on 5th November 2021). We considered phenotypes with p-values below the Bonferroni-corrected alpha threshold (i.e., $0.05/1376 = 4 \times 10^{-5}$) to be genetically correlated with AD (a conservative threshold given the likely correlation between many traits tested).

Bioinformatic analysis

For the following analyses we defined the regions within which the true causal SNP resides to be determined by boundaries containing furthest distanced SNPs with $r^2 >= 0.2$ within ±500kb of the index SNP[18]. We refer to such regions as locus intervals and we used them as input for the analyses described below.

Enrichment analysis

Enrichment of tissues and cell types and gene sets for AD GWAS loci was investigated using DEPICT[57] and GARFIELD (GWAS analysis of regulatory or functional information enrichment with LD correction)[58] ran with default settings, as well as MAGMA v.1.06[59] (using GTEx ver. 8[23] on the FUMA[60] platform). In addition, we used MendelVar[61] run with default settings to check for enrichment of any ontology terms assigned to Mendelian disease genes within the locus interval regions.

By default, MAGMA only assigns variants within genes. DEPICT maps all genes within a given LD ($r^2 > 0.5$) boundary of the index variant. DEPICT gene set enrichment results for GO terms only were grouped (using the Biological Processes ontology) and displayed using the rrvgo package. The default scatter function was adapted to only plot parent terms[62].

Prioritisation of candidate genes

To prioritise candidate genes at each of the loci identified in the European GWAS, we investigated all genes within ±500 kb of each index SNP (selected to capture an estimated 98% of causal genes)[63]. The approach used has been previously described by Sobczyk et al.[18]. For each gene we collated evidence from a range of approaches (as described below) to link SNP to gene, resulting in 14 annotation categories (represented as columns in Supplementary Fig. 7). We summarised these annotations for each gene into a score in order to prioritise genes at each locus. We present the top prioritised gene in the main tables, but strength of evidence varies and so we encourage readers to use our full evaluation (of all the evidence presented in Supplementary Data 11 for all genes at each locus) for loci of interest.

We tested for colocalisation with molecular QTLs, where full summary statistics were available, using coloc[64] method (with betas as input). We used the eQTL Catalogue[65] and Open GWAS[66] to download a range of eQTL datasets from all skin, whole blood and immune cell types as well as additional tissue types which showed enrichment for our GWAS loci, such as spleen and oesophagus mucosa[18]. A complete list of eQTL datasets[20–23,26–31,33,67–71] is displayed in Supplementary Data 13. pQTL summary statistics for plasma proteins[72] were downloaded from Open GWAS. An annotation was included in our gene prioritisation pipeline if there was a posterior probability >95% that the associations from the AD GWAS and the relevant QTL analysis shared the same causal variant.

Additional colocalisation methods were also applied. TWAS (Transcriptome-Wide association Study)-based S-MultiXcan[73] and SMR (Summary-based Mendelian Randomization)[74] were run on datasets available via the CTG-VL platform (including GTEx tissue types and 2 whole blood pQTL[72,75] datasets available for the SMR pipeline). For S-MultiXcan and SMR, we report only results with p-values below the alpha threshold established with Bonferroni correction, as well as no evidence of heterogeneity (HEIDI P-value > 0.05) in SMR analysis.

Genes were also annotated if they were included in any of the globally enriched ontology/pathway terms from the MendelVar analysis described above or if they were identified in direct look-ups of keywords: “skin”, “kera”, “derma” in their OMIM[76] descriptions, or Human Phenotype Ontology[77]/Disease Ontology[78] terms.

We also used machine learning candidate gene prioritisation pipelines – DEPICT[57], PoPs[79], POSTGAP[80] and Open Targets Genetics[53] Variant 2 Gene mapping tool as well as gene-based MAGMA[59] test. We added annotations to genes reported in the top 3 (by each of the pipelines).

We mined the literature for a list of differential expression studies and found 9 RNA-Seq/microarray plus 4 proteomic analyses involving comparisons of AD lesional[25,32,81–84] or AD nonlesional[24,25,32,82,85–87] skin vs healthy controls. Studies with comparisons of AD lesional acute vs chronic[88], blood proteome in AD vs healthy control[32] and *FLG* knockdown vs control in living skin-equivalent[89] were also included. We annotated each gene (including direction of effect, i.e. upregulated/downregulated) with FDR < 0.05 in any dataset.

Lastly, we annotated genes where the index SNP resided within the coding region according to VEP (Variant Effect Predictor)[90] analysis.

For each candidate gene, we established a pragmatic approach to combine all available evidence in order to prioritise which the most plausible candidate gene(s). This prioritisation was carried out as follows:

- The number of annotations (each representing one piece of evidence) were summed across all methods and datasets, to derive a ‘total evidence score’, i.e., if coloc evidence was observed for 5 datasets for a particular gene, this would add 5 to the score for that gene.
- Additionally, to assess if evidence was coming from multiple datasets using the same method, or evidence was coming from diverse approaches, we counted ‘evidence types’, summing up the methods (as opposed to datasets) with an annotation for each gene tested (up to a maximum of 14), i.e., in the same example of coloc evidence observed in 5 datasets, this would add 1 to this measure for this gene. Evidence types are represented by the columns in Supplementary Fig. 7.
- In order to prioritise genes with the most evidence, whilst ensuring there was some evidence of triangulation across methods, at each locus we prioritised the gene with the highest ‘total evidence score’ with a minimum ‘evidence type’ of 3. ‘Evidence type’ was also used to break ties.

Network analysis
Network analysis of the prioritised genes was carried out using standard settings (minimum interaction score 0.4) in STRING v11.5[91].

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
Summary statistics of the GWAS meta-analyses generated in this study have been deposited in the GWAS Catalog under study accession IDs GCST90244787 and GCST90244788. The variant-level data for the 23andMe replication dataset are fully disclosed in the main tables and supplementary tables. Individual-level data are protected and are not available due to data privacy laws, and in accordance with the IRB-approved protocol under which the study was conducted.

Code availability
Code for the bioinformatic analysis is available here: https://github.com/marynias/eczema_gwas_fu/tree/bc4/new_gwas.

References

1. Weidinger, S. et al. Atopic dermatitis. *The Lancet* **387**, 1109–1122 (2016).
2. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456 (2015).
3. Paternoster, L. et al. Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat. Genet.* **44**, 187–192 (2012).
4. Weidinger, S. et al. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum. Mol. Genet.* **22**, 4841–4856 (2013).
5. Johansson, Å., Rask-Andersen, M., Karlsson, T. & Ek, W. E. Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Assoc. Stud. Article* **28**, 4022–4041 (2019).
6. Sliz, E. et al. Uniting biobank resources reveals novel genetic pathways modulating susceptibility for atopic dermatitis. *J. Allergy Clin. Immunol.* **149**, 1105–1112.e9 (2022).
7. Grosche, S. et al. Rare variant analysis in eczema identifies exonic variants in DUSP1, NOTCH4 and SLC9A4. *Nat. Commun.* **12**, 1–11 (2021).
8. Tanaka, N. et al. Eight novel susceptibility loci and putative causal variants in atopic dermatitis. *J. Allergy Clin. Immunol.* **148**, 1293–1306 (2021).
9. Schaarschmidt, H. et al. A genome-wide association study reveals 2 new susceptibility loci for atopic dermatitis. *J. Allergy Clin. Immunol.* **136**, 802–806 (2015).
10. Hirota, T. et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. *Nat. Genet.* **44**, 1222–1226 (2012).
11. Kim, K. W. et al. Genome-wide association study of recalcitrant atopic dermatitis in Korean children. *J. Allergy Clin. Immunol.* **136**, 678–684.e4 (2015).
12. Sun, L.-D. et al. Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population. *Nat. Genet.* **43**, 690–694 (2011).
13. Esparza-Gordillo, J. et al. A functional IL-6 receptor (IL6R) variant is a risk factor for persistent atopic dermatitis. *J. Allergy Clin. Immunol.* **132**, 371–377 (2013).
14. Ellinghaus, D. et al. High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nat. Genet.* **45**, 808–812 (2013).
15. Larsen, F. S., Holm, N. V. & Henningsen, K. Atopic dermatitis. A genetic-epidemiologic study in a population-based twin sample. *J. Am. Acad. Dermatol.* **15**, 487–494 (1986).
16. Schultz Larsen, F. Atopic dermatitis: a genetic-epidemiologic study in a population-based twin sample. *J Am Acad. Dermatol.* **28**, 719–723 (1993).
17. Budu-Aggrey, A. et al. Investigating the causal relationship between allergic disease and mental health. *Clin. Exp. Allergy* **51**, 1449–1458 (2021).
18. Sobczyk, M. K. et al. Triangulating molecular evidence to prioritize candidate causal genes at established atopic dermatitis loci. *J. Invest. Dermatol.* **141**, 2620–2629 (2021).
19. Zeng, Z. et al. Roles of G protein-coupled receptors in inflammatory bowel disease. *World J. Gastroenterol.* **26**, 1242–1261 (2020).
20. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424 (2018).
21. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
22. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715.e16 (2018).
23. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
24. Winge, M. C. G. et al. Filaggrin genotype determines functional and molecular alterations in skin of patients with atopic dermatitis and ichthyosis vulgaris. *PLoS ONE* **6**, e28254 (2011).
25. He, H. et al. Tape strips detect distinct immune and barrier profiles in atopic dermatitis and psoriasis. *J. Allergy Clin. Immunol.* **147**, 199–212 (2021).
26. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
27. Nédélec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669.e21 (2016).
28. Quach, H. et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656.e17 (2016).
29. Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
30. Kasela, S. et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet.* **13**, e1006643 (2017).
31. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
32. Pavel, A. B. et al. The proteomic skin profile of moderate-to-severe atopic dermatitis patients shows an inflammatory signature. *J. Am. Acad. Dermatol.* **82**, 690–699 (2020).
33. Buil, A. et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
34. Nomura, T., Wu, J., Kabashima, K. & Guttman-Yassky, E. Endophenotypic variations of atopic dermatitis by age, race, and ethnicity. *J. Allergy Clin. Immunol. Pract.* **8**, 1840–1852 (2020).
35. Yew, Y. W., Thyssen, J. P. & Silverberg, J. I. A systematic review and meta-analysis of the regional and age-related differences in atopic dermatitis clinical characteristics. *J. Am. Acad. Dermatol.* **80**, 390–401 (2019).
36. Ständer, H. F., Elmariah, S., Zeidler, C., Spellman, M. & Ständer, S. Diagnostic and treatment algorithm for chronic nodular prurigo. *J. Am. Acad. Dermatol.* **82**, 460–468 (2020).

37. Sangha, A. M. Dermatological conditions in SKIN OF COLOR-: managing atopic dermatitis. *J. Clin. Aesthet. Dermatol.* **14**, S20–S22 (2021).
38. Paternoster, L. et al. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. *J. Allergy Clin. Immunol.* **141**, 964–971 (2018).
39. Langan, S. M., Irvine, A. D. & Weidinger, S. Atopic dermatitis. *Lancet* **396**, 345–360 (2020).
40. Al-Janabi, A., Foulkes, A. C., Griffiths, C. E. M. & Warren, R. B. Paradoxical eczema in patients with psoriasis receiving biologics: a case series. *Clin. Exp. Dermatol.* **47**, 1174–1178 (2022).
41. McAleer, M. A. & Irvine, A. D. The multifunctional role of filaggrin in allergic skin disease. *J. Allergy Clin. Immunol.* **131**, 280–291 (2013).
42. Danso, M. O. et al. TNF-α and Th2 cytokines induce atopic dermatitis-like features on epidermal differentiation proteins and stratum corneum lipids in human skin equivalents. *J. Invest. Dermatol.* **134**, 1941–1950 (2014).
43. Gutowska-Owsiak, D., Schaupp, A. L., Salimi, M., Taylor, S. & Ogg, G. S. Interleukin-22 downregulates filaggrin expression and affects expression of profilaggrin processing enzymes. *Br. J. Dermatol.* **165**, 492–498 (2011).
44. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
45. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
46. Ochoa, D. et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).
47. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinform.* **11**, 288 (2010).
48. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
49. Bustamante, M. et al. A genome-wide association meta-analysis of diarrhoeal disease in young children identifies FUT2 locus and provides plausible biological pathways. *Hum. Mol. Genet.* **25**, 4127–4142 (2016).
50. Mägi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
51. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
52. Ghoussaini, M. et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).
53. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
54. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, (2012).
55. Zheng, J. et al. LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
56. Cuéllar-Partida, G. et al. Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses. *bioRxiv* https://doi.org/10.1101/518027 (2019)
57. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
58. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* **51**, 343–353 (2019).
59. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
60. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1–11 (2017).
61. Sobczyk, M. K., Gaunt, T. R. & Paternoster, L. MendelVar: gene prioritization at GWAS loci using phenotypic enrichment of Mendelian disease genes. *Bioinformatics* **37**, 1–8 (2021).
62. Sayols, S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. *MicroPubl. Biol.* https://doi.org/10.17912/micropub.biology.000811 (2023)
63. Stacey, D. et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, e3–e3 (2019).
64. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
65. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
66. Elsworth, B. et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv* https://doi.org/10.1101/2020.08.10.244293 (2020)
67. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
68. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
69. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
70. Lepik, K. et al. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.* **13**, e1005766 (2017).
71. Naranbhai, V. et al. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
72. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
73. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15**, e1007889 (2019).
74. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
75. Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
76. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
77. Köhler, S. et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
78. Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).
79. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* https://doi.org/10.1101/2020.09.08.20190561 (2020)
80. Peat, G. et al. The open targets post-GWAS analysis pipeline. *Bioinformatics* **36**, 2936–2937 (2020).

81. Rojahn, T. B. et al. Single-cell transcriptomics combined with interstitial fluid proteomics defines cell type-specific immune regulation in atopic dermatitis. *J. Allergy Clin. Immunol.* **146**, 1056–1069 (2020).
82. Pavel, A. B. et al. Tape strips from early-onset pediatric atopic dermatitis highlight disease abnormalities in nonlesional skin. *Allergy* **76**, 314–325 (2021).
83. Dyjack, N. et al. Minimally invasive skin tape strip RNA sequencing identifies novel characteristics of the type 2-high atopic dermatitis disease endotype. *J. Allergy Clin. Immunol.* **141**, 1298–1309 (2018).
84. Molin, S. et al. The hand eczema proteome: imbalance of epidermal barrier proteins. *Br. J. Dermatol.* **172**, 994–1001 (2015).
85. Cole, C. et al. Filaggrin-stratified transcriptomic analysis of pediatric skin identifies mechanistic pathways in patients with atopic dermatitis. *J. Allergy Clin. Immunol.* **134**, 82–91 (2014).
86. Ewald, D. A. et al. Meta-analysis derived atopic dermatitis (MADAD) transcriptome defines a robust AD signature highlighting the involvement of atherosclerosis and lipid metabolism pathways. *BMC Med. Genomics* **8**, 60 (2015).
87. Morelli, P. et al. Proteomic analysis from skin swabs reveals a new set of proteins identifying skin impairment in atopic dermatitis. *Exp. Dermatol.* **30**, 811–819 (2021).
88. Tsoi, L. C. et al. Progression of acute-to-chronic atopic dermatitis is associated with quantitative rather than qualitative changes in cytokine responses. *J. Allergy Clin. Immunol.* **145**, 1406–1415 (2020).
89. Elias, M. S. et al. Proteomic analysis of filaggrin deficiency identifies molecular signatures characteristic of atopic eczema. *J. Allergy Clin. Immunol.* **140**, 1299–1309 (2017).
90. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
91. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
92. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
93. Neale, B. GWAS results round 2. http://www.nealelab.is/uk-biobank/ (2018).

Acknowledgements

For this work, A.B.-A., S.J.B. and L.P. were funded by the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 821511 (BIOMAP). The J.U. receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This publication reflects only the author's view and the J.U. is not responsible for any use that may be made of the information it contains. A.B.A., M.K.S., J.L.M., and L.P. and work in a research unit funded by the UK Medical Research Council (MC_UU_00011/1 and MC_UU_00011/4). LP received funding from the British Skin Foundation (8010 Innovative Project) and the Academy of Medical Sciences Springboard Award, which is supported by the Wellcome Trust, The Government Department for Business, Energy and Industrial Strategy, the Global Challenges Research Fund and the British Heart Foundation [SBF003\1094]. S.J.B. holds a Wellcome Trust Senior Research Fellowship in Clinical Science [220875/Z/20/Z]. S.H. is supported by a Vera Davie Study and Research Sabbatical Bursary, NRF Thuthuka Grant (117721), NRF Competitive Support for Unrated Researcher (138072), MRC South Africa under a Self-initiated grant. M.S. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 949906). Thanks to Sergi Sayols (developer of rrvgo), who provided additional code to alter the scatter plot produced by rrvgo to only display parent terms, and to Gibran Hemani (University of Bristol) who provided valuable guidance on the comparison of effects between ancestries. This publication is the work of the authors and LP will serve as guarantor for the contents of this paper. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol—http://www.bristol.ac.uk/acrc/. Individual cohort acknowledgements are in the Supplementary Methods.

Author contributions

Designed and co ordinated the study: M.St., L.P. Performed the meta analysis: A.B.-A., A. Ki. Performed the bioinformatic analysis: M.K.So. Performed the STRING analysis: S.J. Br. Performed statistical analysis within cohorts: A.B.-A., A. Ki, R. Mi, K.R., R. Mä, M.N., N.T., B.M.B., L.F.T., P.S.N., C.F., A.E.O., E.H.L., J.V.T.L., J.B.J., I.M., A.A.-S., A.J., H. Ba., E.R., A. Ku., C.M.G., C.H., C.Q., P.T., E.A., J.F., C.A.W., E.T., B.W., S.K., D.M.K., L.K., J.D., H.Z., C.A., V.U., R.K., A. Sz., A.C.S.N.J., A.G., M.I., M. M.-Nu., T.S.A., M.B., C.G., M.P.Y., D.P.S., N.G., Y.A.L., A.D.I., L.K.W., C.M., S.J. Br. Data acquisition/supported analysis/interpretation of data: A.B.-A., A. Ki, A.R., P.S.N., A.J., C.j.X., S.E.H., J.F., E.T., S.K., H.Z., S.H., T. Ho., E.J., H.C., N.R., P.N., O.A.A., S.J., C.A., T.G., V.U., P.K.E.M., E.G.B., J.P.T., T. Ha., L.L.K., T.M.D., A.Ar., G.H., S.L., M.M. Nö., N.H., M.I., T.S.Ah., J.S., B.C., A.M.M.S., A.E., S.Ar., T.E., L.A.M., A.M., C.T., K.A., M.L., K.H., B.J., D.P.S., Y.A.L., N.P.H., S.W., A.D.I., D.J., T.N., L.D., J.M.V., G.H.K., K.M.G., B.F., C.E.P., P.D.S., P.G.Ho., H. Bi., K.B., J.C., A. Si., T.S., S.J. Br., M.St., L.P. Wrote the paper: A.B.-A., A. Ki, M.K.So., S.J. Br., M.S., L.P. Approved final version of paper: A.B.-A., A. Ki, M.K.So., S.S.S., R. Mi, K.R., A.R., R. Mä, M.N., N.T., B.M.B., L.F.T., P.S.N., C.F., A.E.O., E.H.L., J.V.T.L., J.B.J., I.M., A.A.-S., A.J., K.A.F., H. Ba., E.R., A.C.A., A. Ku., P.M.S., X.C., C.M.G., C.H., C.j.X., C.Q., S.E.H., P.T., E.A., J.F., C.A.W., E.T., B.W., S.K., D.M.K., L.K., J.D., H.Z., S.H., T. Ho., E.J., H.C., N.R., P.N., O.A.A., S.J., C.A., T.G., V.U., R.K., P.K.E.M., A. Sz., E.G.B., J.P.T., T. Ha., L.L.K., T.M.D., A.C.S.N.J., A.G., A.Ar., G.H., S.L., M.M. Nö., N.H., M.I., A.V., M.F., V.B., P.Hy., N.B., D.I.B., J.J.H., M.M.-Nu., T.S.Ah., J.S., B.C., A.M.M.S., A.E., M.B., B.R., S.Ar., C.G., T.E., L.A.M., A.M., C.T., K.A., M.L., K.H., B.J., M.P.Y., D.P.S., N.G., A.L., Y.A.L., N.P.H., S.W., M.R.J., E.M., H.H., A.D.I., D.J., T.N., L.D., J.M.V., G.H.K., K.M.G., S.J. Ba., B.F., C.E.P., P.D.S., P.G.Ho., L.K.W., H. Bi., K.B., J.C., A. Si., C.M., T.S., S.Bu., S.T.W., J.W.H., J.L.M., S.J. Br., M.St., L.P.

Competing interests

K.M.G. has received reimbursement for speaking at conferences sponsored by companies selling nutritional products, and is part of an academic consortium that has received research funding from Abbott Nutrition, Nestec, BenevolentAI Bio Ltd. and Danone. C.G., S.S.S., and 23andMe Research Team are employed by and hold stock or stock options in 23andMe, Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-41180-2.

Correspondence and requests for materials should be addressed to Lavinia Paternoster.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Ashley Budu-Aggrey[1,2,124], Anna Kilanowski[3,4,5,124], Maria K. Sobczyk[1,2], 23andMe Research Team, Suyash S. Shringarpure[6], Ruth Mitchell[1,2], Kadri Reis[7], Anu Reigo[7], Estonian Biobank Research Team, Reedik Mägi[7], Mari Nelis[7,8], Nao Tanaka[9,10], Ben M. Brumpton[11,12,13], Laurent F. Thomas[11,14,15,16], Pol Sole-Navais[17], Christopher Flatley[17], Antonio Espuela-Ortiz[18], Esther Herrera-Luis[18], Jesus V. T. Lominchar[19], Jette Bork-Jensen[19], Ingo Marenholz[20,21], Aleix Arnau-Soler[20,21], Ayoung Jeong[22,23], Katherine A. Fawcett[24], Hansjorg Baurecht[25], Elke Rodriguez[26], Alexessander Couto Alves[27], Ashish Kumar[28], Patrick M. Sleiman[29,30,31], Xiao Chang[29], Carolina Medina-Gomez[32,33], Chen Hu[32,34], Cheng-jian Xu[35,36,37,38], Cancan Qi[35,36], Sarah El-Heis[39], Philip Titcombe[39], Elie Antoun[40,41], João Fadista[42,43,44,45], Carol A. Wang[46,47], Elisabeth Thiering[3,4], Baojun Wu[48], Sara Kress[49], Dilini M. Kothalawala[50,51], Latha Kadalayil[50], Jiasong Duan[52], Hongmei Zhang[52], Sabelo Hadebe[53], Thomas Hoffmann[54,55], Eric Jorgenson[56], Hélène Choquet[57], Neil Risch[54,55], Pål Njølstad[58,59], Ole A. Andreassen[60,61], Stefan Johansson[58,62], Catarina Almqvist[63,64], Tong Gong[63], Vilhelmina Ullemar[63], Robert Karlsson[63], Patrik K. E. Magnusson[63], Agnieszka Szwajda[63], Esteban G. Burchard[65,66], Jacob P. Thyssen[67], Torben Hansen[19], Line L. Kårhus[68], Thomas M. Dantoft[68], Alexander C.S.N. Jeanrenaud[20,21], Ahla Ghauri[20,21], Andreas Arnold[69], Georg Homuth[70], Susanne Lau[71], Markus M. Nöthen[72], Norbert Hübner[20,73], Medea Imboden[22,23], Alessia Visconti[74], Mario Falchi[74], Veronique Bataille[74,75], Pirro Hysi[74], Natalia Ballardini[28], Dorret I. Boomsma[76,77], Jouke J. Hottenga[76], Martina Müller-Nurasyid[78,79,80], Tarunveer S. Ahluwalia[81,82,83], Jakob Stokholm[81,84], Bo Chawes[81], Ann-Marie M. Schoos[81,84], Ana Esplugues[85,86], Mariona Bustamante[87,88,89], Benjamin Raby[90], Syed Arshad[91,92], Chris German[6], Tõnu Esko[7], Lili A. Milani[7], Andres Metspalu[7], Chikashi Terao[9,93,94], Katrina Abuabara[95], Mari Løset[11,96], Kristian Hveem[11,97], Bo Jacobsson[17,98], Maria Pino-Yanes[18,99,100], David P. Strachan[101], Niels Grarup[19], Allan Linneberg[68,102], Young-Ae Lee[20,21], Nicole Probst-Hensch[22,23], Stephan Weidinger[103], Marjo-Riitta Jarvelin[104,105,106], Erik Melén[28], Hakon Hakonarson[29,107,108], Alan D. Irvine[109], Deborah Jarvis[110,111], Tamar Nijsten[34], Liesbeth Duijts[112,113], Judith M. Vonk[36,114], Gerard H. Koppelmann[35,36], Keith M. Godfrey[115], Sheila J. Barton[39], Bjarke Feenstra[43], Craig E. Pennell[46,47], Peter D. Sly[116,117], Patrick G. Holt[118], L. Keoki Williams[48], Hans Bisgaard[81,126], Klaus Bønnelykke[81], John Curtin[119], Angela Simpson[119], Clare Murray[119], Tamara Schikowski[120], Supinda Bunyavanich[121], Scott T. Weiss[90], John W. Holloway[50,91], Josine L. Min[1,2], Sara J. Brown[122], Marie Standl[3,123] & Lavinia Paternoster[1,2,125] ✉

[1]Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, England. [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, England. [3]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany. [4]Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, Munich, Germany. [5]Pettenkofer School of Public Health, Ludwig-Maximilians University Munich, Munich, Germany. [6]23andMe, Inc., Sunnyvale, CA, USA. [7]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia. [8]Core Facility of Genomics, University of Tartu, Tartu, Estonia. [9]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [10]Department of Rheumatology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University (TMDU), Tokyo, Japan. [11]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim 7030, Norway. [12]HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger 7600, Norway. [13]Clinic of Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim 7030, Norway. [14]Department of Clinical and Molecular Medicine, NTNU Norwegian University of Science and Technology, Trondheim, Norway. [15]BioCore - Bioinformatics Core Facility, NTNU, Norwegian University of Science and Technology, Trondheim, Norway. [16]Clinic of Laboratory Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [17]Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [18]Genomics and Health Group, Department of Biochemistry, Microbiology, Cell Biology and Genetics, Universidad de La Laguna, La Laguna, Tenerife, Spain. [19]Novo Nordisk Foundation Center for Basic Metabolic

Research, Faculty of Health and Medical Sciences, University of Copenhagen, København, Denmark. [20]Max-Delbrück-Center for Molecular Medicine, Berlin, Germany. [21]Clinic for Pediatric Allergy, Experimental and Clinical Research Center, Charité-Universitätsmedizin Berlin, Berlin, Germany. [22]Swiss Tropical and Public Health Institute, CH-4123 Basel, Switzerland. [23]University of Basel, CH-4001 Basel, Switzerland. [24]Department of Health Sciences, University of Leicester, Leicester LE1 7RH, UK. [25]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany. [26]Department of Dermatology and Allergy, University Hospital Schleswig-Holstein, Kiel, Germany. [27]School of Biosciences and Medicine, University of Surrey, Guildford, UK. [28]Department of Clinical Science and Education Södersjukhuset, Karolinska Institutet, Solna, Sweden. [29]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [30]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [31]Rhythm Pharmaceuticals, 222 Berkley Street, Boston 02116, USA. [32]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [33]Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [34]Department of Dermatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [35]University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergy, Beatrix Children's Hospital, Groningen, The Netherlands. [36]University of Groningen, University Medical Center Groningen, GRIAC Research Institute, Groningen, The Netherlands. [37]Centre for Individualized Infection Medicine, CiiM, a joint venture between Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover, Germany. [38]TWINCORE, Centre for Experimental and Clinical Infection Research, a joint venture between the Hannover Medical School and the Helmholtz Centre for Infection Research, Hannover, Germany. [39]MRC Lifecourse Epidemiology Centre, University of Southampton, Southampton, UK. [40]Faculty of Medicine, University of Southampton, Southampton, UK. [41]Institute of Developmental Sciences, University of Southampton, Southampton, UK. [42]Department of Bioinformatics & Data Mining, Måløv, Denmark. [43]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark. [44]Department of Clinical Sciences, Lund University Diabetes Centre, Malmö, Sweden. [45]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [46]School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia. [47]Hunter Medical Research Institute, Newcastle, NSW, Australia. [48]Center for Individualized and Genomic Medicine Research (CIGMA), Department of Medicine, Henry Ford Health, Detroit, MI 48104, USA. [49]Environmental Epidemiology of Lung, Brain and Skin Aging, IUF – Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany. [50]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK. [51]NIHR Southampton Biomedical Research Centre, University Hospital Southampton, Southampton, UK. [52]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA. [53]Division of Immunology, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. [54]Institute for Human Genetics, UCSF, San Francisco, CA 94143, USA. [55]Department of Epidemiology and Biostatistics, UCSF, San Francisco, CA 94158, USA. [56]Regeneron Genetics Center, Tarrytown, NY, USA. [57]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. [58]Center for Diabetes Research, Department of Clinical Science, University of Bergen, NO-5020 Bergen, Norway. [59]Children and Youth Clinic, Haukeland University Hospital, NO-5021 Bergen, Norway. [60]NORMENT Centre, Institute of Clinical Medicine, University of Oslo, 0450 Oslo, Norway. [61]Division of Mental Health and Addiction, Oslo University Hospital, 0450 Oslo, Norway. [62]Department of Medical Genetics, Haukeland University Hospital, NO-5021 Bergen, Norway. [63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [64]Pediatric Lung and Allergy Unit, Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden. [65]Department of Medicine, University of California San Francisco, San Francisco, CA, USA. [66]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [67]Department of Dermatology, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark. [68]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Frederiksberg, Denmark. [69]Clinic and Polyclinic of Dermatology, University Medicine Greifswald, Greifswald, Germany. [70]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [71]Department of Pediatric Respiratory Medicine, Immunology, and Critical Care Medicine, Charité-Universitätsmedizin Berlin, Berlin, Germany. [72]Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany. [73]Charite-Universitätsmedizin Berlin, Berlin, Germany. [74]Department of Twin Research & Genetics Epidemiology, Kings College London, London, UK. [75]Dermatology Department, West Herts NHS Trust, Watford, UK. [76]Dept Biological Psychology, Netherlands Twin Register, VU University, Amsterdam, the Netherlands. [77]Institute for Health and Care Research (EMGO), VU University, Amsterdam, the Netherlands. [78]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany. [79]IBE, Faculty of Medicine, LMU Munich, Munich, Germany. [80]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany. [81]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark. [82]Steno Diabetes Center Copenhagen, Herlev, Denmark. [83]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [84]Department of Pediatrics, Slagelse Hospital, Slagelse, Denmark. [85]Nursing School, University of Valencia, FISABIO-University Jaume I-University of Valencia, Valencia, Spain. [86]Joint Research Unit of Epidemiology and Environmental Health, CIBERESP, Valencia, Spain. [87]ISGlobal, Institute for Global Health, Barcelona, Spain. [88]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [89]CIBER Epidemiología y Salud Pública, Madrid, Spain. [90]Channing Division of Network Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, USA. [91]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. [92]David Hide Asthma and Allergy Research Centre, Isle of Wight, UK. [93]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. [94]Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [95]Department of Dermatology, University of California San Francisco, San Francisco, CA, USA. [96]Department of Dermatology, Clinic of Orthopaedy, Rheumatology and Dermatology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [97]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [98]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. [99]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain. [100]Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, San Cristóbal de La Laguna, Santa Cruz de Tenerife, Spain. [101]Population Health Research Institute, St George's, University of London, Cranmer Terrace, London SW17 ORE, UK. [102]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [103]Department of Dermatology, Allergology and Venereology, University Hospital Schleswig-Holstein, Kiel, Germany. [104]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment & Health, School of Public Health,Imperial College London, London, UK. [105]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland. [106]Biocenter Oulu, University of Oulu, Oulu, Finland. [107]Department of Pediatrics, Divisions of Human Genetics and Pulmonary Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [108]Faculty of Medicine, University of Iceland, 101 Reykjavík, Iceland. [109]Department of Clinical Medicine, Trinity College, Dublin, Ireland. [110]Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College London, London, United Kingdom. [111]Medical Research Council and Public Health England Centre for Environment and Health, London, United Kingdom. [112]Department of Pediatrics, division of Respiratory Medicine and Allergology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [113]Department of Pediatrics, division of Neonatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [114]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands. [115]MRC Lifecourse Epidemiology Centre and

NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, UK. [116]Children's Health and Environment Program, Child Health Research Centre, The University of Queensland, South Brisbane 4101 Queensland, Australia. [117]Australian Infectious Diseases Research Centre, The University of Queensland, St Lucia 4072 QLD, Australia. [118]Telethon Kids Institute, University of Western Australia, Perth, WA, Australia. [119]Division of Immunology, Immunity to Infection and Respiratory Medicine, School of Biological Sciences, The University of Manchester, Manchester Academic Health Science Centre, and Manchester University NHS Foundation Trust, Manchester, England. [120]Environmental Epidemiology of Lung, Brain and Skin Aging, Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany. [121]Division of Allergy and Immunology, Department of Pediatrics, and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [122]Centre for Genomics and Experimental Medicine, Institute for Genetics and Cancer, University of Edinburgh, Crewe Road, Edinburgh UK EH4 2XU, Scotland. [123]German Center for Lung Research (DZL), Munich, Germany. [124]These authors contributed equally: Ashley Budu-Aggrey, Anna Kilanowski. [125]These authors jointly supervised this work: Marie Standl, Lavinia Paternoster. [126]Deceased: Hans Bisgaard.
✉e-mail: l.paternoster@bristol.ac.uk

23andMe Research Team

Suyash S. Shringarpure[6] & Chris German[6]

Estonian Biobank Research Team

Reedik Mägi[7], Mari Nelis[7,8], Tõnu Esko[7], Lili A. Milani[7] & Andres Metspalu[7]

B Acknowledgements

During my Ph.D. times, I learned that next to your own motivation and persistence there are two major factors affecting both quality and happiness levels during the doctoral phase: The first one is great supervision and the second the support and contact with other doctoral researchers – luckily I had both during this time!

First and foremost, I would like to thank my direct supervisor Dr. Marie Standl for the amazing supervision, ongoing support and encouragement throughout my whole Ph.D.
Thank you for including me in all these projects, providing insight into so many areas of research and collaborations and for giving me the chance to spend this time at Helmholtz Munich, even though I had never heard the words "Epigenetics" or "DNA methylation" before our first initial call.
Also at Helmholtz Munich, I would like to thank the great working groups of Allergic disease and lung epidemiology and especially Dr. Elisabeth Thiering for all the support a doctoral researcher could wish for.
My gratitude further goes to Dr. Anke Hüls at the Emory University in Atlanta, where I could stay for two months abroad. It was a unique experience right after spending so many months in home office during the pandemic. Thank you very much for the great supervision, not only for the actually planned project but also for the second one and the support over at least half of my Ph.D. time.
I would also like to thank my thesis advisory committee for their helpful feedback during all TAC meetings and especially Dr. Annette Peters as my primary supervisor and institute director.

Coming to the second factor affecting your Ph.D., I want to thank all the other doctoral researchers at the EPI institute – I really loved our lunch breaks, especially after spending way too much time at home during the pandemic.
Many thanks also to all members of the doctoral initiative at Helmholtz Munich, DINI, and the Helmholtz Juniors all over Germany. Being involved in these initiatives really helped me through the pandemic times and I would never want to miss all the amazing people I got to know and am now calling my friends.

Last but not least, I want to thank my parents so much for always supporting me through life and believing in me. I am pretty sure, none of us believed that it would end up in an actual Ph.D. degree in the beginning. I could have never gone all those steps from Wuppertal to London and Munich without knowing that you always had my back and I could 100% rely on you.
And also to my little brother, even though he believes that I am "*only letting the computer calculate some stuff for me*".

- Thank you all so much!

C List of publications

1. Love SB, Kilanowski A, Yorke-Edwards V, Old O, Barr H, Stokes C, et al. Use of routinely collected health data in randomised clinical trials: comparison of trial-specific death data in the BOSS trial with NHS Digital data. Trials. 2021 Sep;22(1):654. Available from: https://doi.org/10.1186/s13063-021-05613-x

2. Kilanowski A, Chen J, Everson T, Thiering E, Wilson R, Gladish N, et al. Methylation risk scores for childhood aeroallergen sensitization: Results from the LISA birth cohort. Allergy. 2022;77(9):2803-17. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15315

3. Kress S, Kilanowski A, Wigmann C, Zhao Q, Zhao T, Abramson MJ, et al. Airway inflammation in adolescents and elderly women: Chronic air pollution exposure and polygenic susceptibility. Science of The Total Environment. 2022 Oct;841:156655. Available from: https://www.sciencedirect.com/science/article/pii/S0048969722037524

4. Kilanowski A, Thiering E, Wang G, Kumar A, Kress S, Flexeder C, et al. Allergic disease trajectories up to adolescence: Characteristics, early-life, and genetic determinants. Allergy. 2022;n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15511

5. Kilanowski A, Merid SK, Abrishamcar S, Feil D, Thiering E, Waldenberger M, et al. DNA methylation and aeroallergen sensitization: The chicken or the egg? Clinical Epigenetics. 2022 Sep;14(1):114. Available from: https://doi.org/10.1186/s13148-022-01332-5

6. Abrishamcar S, Chen J, Feil D, Kilanowski A, Koen N, Vanker A, et al. DNA methylation as a potential mediator of the association between prenatal tobacco and alcohol exposure and child neurodevelopment in a South African birth cohort. Translational Psychiatry. 2022 Sep;12(1):1-12. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41398-022-02195-3

7. Feil D, Abrishamcar S, Christensen GM, Vanker A, Koen N, Kilanowski A, et al.. DNA methylation as a potential mediator of the association between indoor air pollution and neurodevelopmental delay in a South African birth cohort; 2022. Available from: https://www.researchsquare.com

8. Budu-Aggrey A, Kilanowski A, Sobczyk MK, Shringarpure SS, Mitchell R, Reis K, et al.. European and multi-ancestry genome-wide association meta-analysis of atopic dermatitis highlights importance of systemic immune regulation; 2022. Publisher: Cold Spring Harbor Laboratory Press _eprint: https://www.medrxiv.org/content/early/2022/10/10/2022.10.05.22279072.full.pdf. Available from: https://www.medrxiv.org/content/early/2022/10/10/2022.10.05.22279072

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office Medical Faculty
Faculty of Medicine

Affidavit

Kilanowski, Anna

Surname, first name

Address

I hereby declare, that the submitted thesis entitled

Longitudinal assessment of the development of allergic diseases, aeroallergen sensitization and the role of DNA methylation

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

Limburgerhof, 03.10.2023

Place, Date

Anna Kilanowski

Signature doctoral candidate

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office Medical Faculty
Doctoral Office

Confirmation of congruency between printed and electronic version of the doctoral thesis

Doctoral candidate: Anna Kilanowski

Address:

I hereby declare that the electronic version of the submitted thesis, entitled

Longitudinal assessment of the development of allergic diseases, aeroallergen sensitization and the role of DNA methylation

is congruent with the printed version both in content and format.

Limburgerhof, 03.10.2023
Place, Date

Anna Kilanowski
Signature doctoral candidate