Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology

SUPPLEMENTARY NOTE

Contents **Page**

1. Participating studies

UK Biobank (*n*=138,354)

Sample ascertainment and phenotype definition. The UK Biobank study is a prospective study of 502,682 participants recruited at 22 centers across the UK between 2006 and 2010 [1]. We used data from the UK Biobank Resource under Application Number 10074. We downloaded data-fields approved as part of this application on the 21st of March 2016, and restricted our analysis to a subset of 152,566 individuals with available genotype data at that time.

To classify asthma status, we combined information from three sources: (1) touchscreen questionnaire (data-field 6152); (2) Non-cancer illness code, self-reported during verbal interview (data-field 20002); (3) main (data-field 41202) and secondary (data-field 41204) ICD10 diagnoses. Specifically, inclusion criteria for cases were: (i) a report of "Asthma" in field 6152 and a code for asthma (1111) in field 20002; or (ii) an ICD10 code for asthma in fields 41202 or 41204, including any one of J45.0, J45.1, J45.8, J45.9 and J46. Exclusion criteria for cases were: (i) a report of COPD in fields 6152 or 20002 (code 1112); or (ii) other self-reported respiratory diseases in field 20002 (codes 1113, 1114, 1115 and 1117). Inclusion criteria for controls were no report of asthma in fields 6152, 20002, 41202 and 41204. Exclusion criteria for controls were the same as for cases (no COPD or other self-reported respiratory diseases). According to this definition, we identified 17,456 cases, 129,763 controls and 5,347 with a missing phenotype.

To create the hay fever phenotype, we used the same three sources of information. Specifically, inclusion criteria for cases were: (i) a report of "Hay fever, allergic rhinitis or eczema" in field 6152 and a code for hay fever (1387) in field 20002; or (ii) an ICD10 code for hay fever in fields 41202 or 41204, including any one of J30.0, J30.1, J30.3, J30.4 and J31.0. Inclusion criteria for controls were no report of hay fever in fields 6152, 20002, 41202 and 41204. According to this definition, we identified 8,004 cases, 116,712 controls and 27,850 with a missing phenotype.

The eczema phenotype was created very similarly to the hay fever phenotype. Inclusion criteria for cases were: (i) a report of "Hay fever, allergic rhinitis or eczema" in field 6152 and a code for eczema (1452) in field 20002; or (ii) an ICD10 code for eczema in fields 41202 or 41204, including any one of L20.8, L20.9 or L30.9. Inclusion criteria for controls were no report of eczema in fields 6152, 20002, 41202 and 41204. According to this definition, we identified 3,234 cases, 116,592 controls and 32,740 with a missing phenotype.

For the overall allergic disease phenotype, cases were individuals classified as suffering from at least one condition (asthma and/or hay fever and/or eczema), as described above. Controls were individuals classified as not having suffered from all three conditions. Additional exclusion criteria for controls included codes for anaphylaxis (1374), allergy to food (1385), allergy to drug (1386), allergy to house dust mite (1668), contact dermatitis (1669), allergy to elastoplast (1670) or allergy to nickel (1671) in field 20002. Overall, we identified 44,413 cases, 100,574 controls and 7,579 individuals with a missing phenotype. This sample size was reduced further to 42,246 cases, 96,108 controls and 7,295 missing after exclusions based on ancestry and other genotyping QC filters (see section below).

For the case-only single-disease analyses, we defined three non-overlapping groups of individuals based on the presence of the three individual diseases: (1) asthma-only cases: individuals classified as suffering from asthma but not hay fever and eczema (A+H-E-; n=8,769 after QC); (2) hay fever-only cases: individuals classified as suffering from hay fever but not asthma and eczema (A-H+E-; n=4,838); and (3) individuals classified as suffering from eczema but not asthma and hay fever (A-H-E+; n=1,538).

Age-of-onset information was obtained from two data-fields: 3786 ("Age asthma diagnosed") and 3761 ("Age hay fever, rhinitis or eczema diagnosed"). For each individual, the minimum of these two was taken as the earliest any one allergic disease was first diagnosed. For the analyses of disease-specific age-of-onset, we only considered age-of-onset information for individuals from the three

single-disease groups described above.

Genotyping, quality control and imputation. The procedures used by the UKBiobank for genotyping, QC and imputation have been described in previous publications (eg. [2]). We first downloaded directly genotyped data for 152,566 individuals and 847,442 variants. We then analyzed these data with PLINK v1.90b3.36 [3] to identify individuals of European ancestry as follows. First, we identified common SNPs (minor allele frequency >5%) in low linkage disequilibrium with each other using the –indep-pairwise option, with arguments "500 kb 5 0.1". Second, we restricted the analysis to SNPs also present in the 1000 Genomes Project (release 20101123), with call rate >95% and Hardy-Weinberg equilibrium P-value $>10^{-6}$, and that were not A/T or C/G polymorphisms. Third, we merged the UK Biobank and 1000 Genomes project data (n=1,092, representing 14 ancestry groups) and performed multi-dimensional scaling (MDS) analysis to identify individuals who clustered closely to Europeans of the 1000 Genomes project. Because of computational limitations, the latter step was carried out after splitting the UK Biobank data in two 10 groups of ~15,000 individuals each. To identify individuals clustering with Europeans of the 1000 Genomes project, we calculated the mean for the first and second MDS components based on individuals from the five European ancestry groups (CEU, GBR, FIN, IBS and TSI) of the 1000G project, and then selected all UK Biobank individuals within 5 standard deviations (SD) of those means. With this approach, which is illustrated below for the first group of ~15,000 individuals (boxes show 5 SD from the mean for each of the three major ancestry groups), we identified 146,267 individuals (95.9%) of European ancestry, which were retained for further analyses. Plots for the remaining 9 groups of individuals look identical (not shown). The UKBiLEVE study [4] used a similar approach to define ancestry, with a less stringent cut-off (10 SD from the mean).

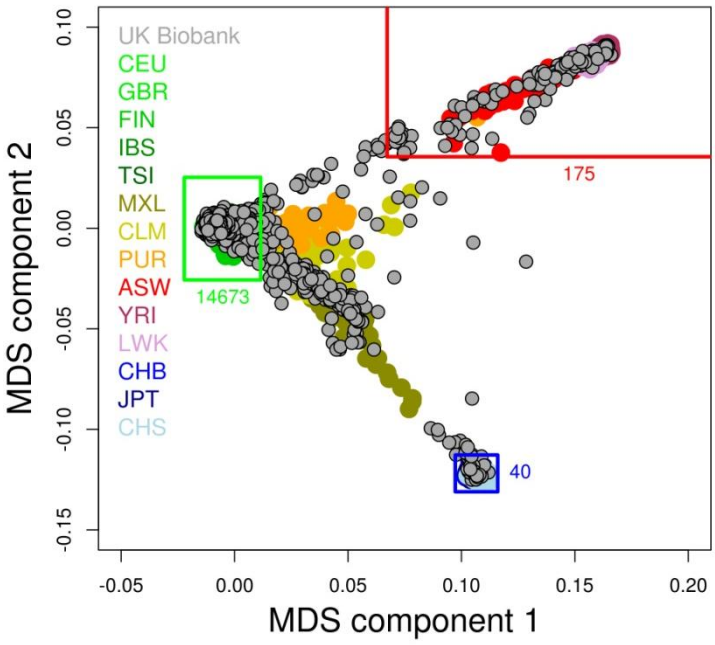

Figure. Multi-dimensional scaling (MDS) analysis of identity-by-state allele sharing between samples of the UK Biobank study (in grey; *n*=15,211, ie. 1/10 of full dataset) and of the 1000 Genomes Project Phase 3 (different colors reflecting 14 ancestry groups). Box edges represent +/- 5 SD from the mean of MDS1 and MDS2, calculated separately for Europeans (green; CEU, GBR, FIN, IBS and TSI), Asians (blue; CHB, JPT and CHS) and Africans (red; ASW, YRI and LWK). The number of UK Biobank individuals inside each box (and so classified in that ancestry group) are indicated.

Lastly, we excluded from analysis an additional 176 individuals with a mismatch between self-reported and genetically-inferred sex, and 442 individuals with outlier heterozygosity (data-field 22010). After these exclusions, data were available for 145,649 individuals, including 42,246 allergic disease cases, 96,108 controls and 7,295 with a missing phenotype. The genotype data was used to impute unmeasured autosomal SNPs based on a combined UK10K and 1000 Genomes Phase 3 reference panel as described by the UK Biobank team (http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=157020); these data were downloaded as part of our approved application. We imputed data on the X-chromosome using the 1000 Genomes Phase 3 Oct2014 release reference panel as follows. First, we applied standard SNP-based QC filters using PLINK to the array data (21,231 SNPs on the X-chromosome), namely exclusion of SNPs with: call rate <95%, Hardy Weinberg equilibrium $P<10^{-6}$, MAF<1% (based on

Europeans only), alleles that were different from 1000 Genomes data, allele frequency differences against 1000 Genomes data (P<0.005), leaving 12,535 SNPs for imputation. We also restricted imputation of the X-chromosome to unrelated individuals (IBD<0.125). Second, we used IMPUTE2 [5] to impute unmeasured variants using the 1000 Genomes Phase 3 Oct2014 release reference panel in males and females jointly with the options –chrX and –Ne 20000. We split the X-chromosome into 77 5Mb-jobs for imputation.

Association analysis. Most (95%) but not all of the 138,354 individuals available for analysis were unrelated to each other (based on an identity by descent cut-off of 0.125). To account for relatedness, the association between the main allergic disease phenotype and imputed SNPs was carried out using the linear mixed model implemented in BOLT-LMM [6], with age, sex and SNP chip included as covariates. We used directly genotype data for 443,239 pruned (r^2<0.9) post-QC (including MAF>1%, call rate >90%, HWE P>0.0001) SNPs to estimate the random effect attributed to variants other than the one being tested. For each SNP, the beta and SE estimated from the linear model were then adjusted using the formulae adj_beta = beta/(mu*(1-mu)) and adj_SE = SE/(mu*(1-mu)), where mu is approximated by the case/control ratio.

For the case-only single-disease analyses, as well as for age-of-onset (which was quantile normalized), association analyses were restricted to unrelated individuals only (IBD<0.125, given very small number of relatives) and performed with SNPTEST, with the options -method expected and -frequentist 1. Sex and age (single-disease only) were included as covariates. Two sets of association analyses were performed for age-of-onset. First, we analysed the earliest age-of-onset reported for any one allergic disease by 35,972 unrelated individuals. Second, we repeated the analysis of age-of-onset in each of three groups: asthma-only cases (n=7,445), hay fever-only cases (n=4,232) and eczema-only cases (n=1,225). Age-of-onset was quantile normalized separately for each group. For a given SNP,

differences in effect size (beta) between groups were quantified using the formula z = sigma / SE_sigma, where sigma = beta_groupA – beta_groupB, and SE_sigma = sqrt(SE_beta_groupA^2 + SE_beta_groupB^2), which follows a normal distribution.

SNPTEST was also used for association analysis of X-chromosome variants, assuming a model of full X inactivation in females (options –method newml –frequentist 1). In this model, male genotypes are encoded as 0 or 1 and females as 0, ½ or 1.

23andMe (*n*=118,269)

Sample ascertainment and phenotype definition. All research participants included in the analyses provided informed consent and answered surveys online according to our human subjects protocol, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). We restricted participants to a set of individuals who have >97% European ancestry, as determined through an analysis of local ancestry. Briefly, our algorithm first partitions phased genomic data into short windows of about 100 SNPs. Within each window, we use a support vector machine (SVM) to classify individual haplotypes into one of 31 reference populations. The SVM classifications are then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments, and gives probabilities for each reference population in each window. Finally, we used simulated admixed individuals to recalibrate the HMM probabilities so that the reported assignments are consistent with the simulated admixture proportions. The reference population data is derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country. A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm [7]. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic

segments identical-by-descent. This level of relatedness (roughly 20% of the genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

The asthma phenotype combines reports of asthma diagnoses from five sources:

1) Your Health Profile Survey: Have you ever been diagnosed or treated for any of the following conditions?' (Asthma: Yes, No, Don't Know); 2) Your Medical History Survey: Have you ever been diagnosed by a doctor with any of the following types of _allergies_?' (Asthma: Yes, No, I don't know); 3) Allergies and Asthma Survey: Have you ever had an asthma attack? (Yes, No, I'm not sure); 4) Asthma Survey: Have you ever been diagnosed by a doctor with asthma or bronchial asthma?' (Yes, No, I'm not sure); and 5) The Roots into the Future intake form Survey: Have you ever been diagnosed or treated for any of the following conditions?' (Asthma: Yes, No, Don't Know). We merged the yes/no responses from these questions, with inconsistent responses scored as missing: cases have at least one positive response and no negative responses, and controls have at least one negative response and no positive responses.

The 'hay fever' phenotype combines answers from three sources: 1) Asthma Survey: Have you ever had any of the following? Please check all that apply: Allergic rhinitis (stuffed or dripping nose caused by allergies) (Yes, No); 2) Quick question used in Research snippet: Have you ever been diagnosed with hay fever (allergic rhinitis)? (Yes, No, I am not sure); and 3) Allergy Survey: A series of questions on allergy symptoms to grasses, trees, weeds, cats, dogs, dust mites and mold. An example one is worded as "What type of reaction did you have after being exposed to trees? Please check all that apply: Itchy or runny nose" (Yes, No). We merged the yes/no responses from the above questions, with inconsistent responses scored as missing: cases have at least one positive response and no negative responses and controls have at least one negative response and no positive responses.

The eczema phenotype comes from nine sources: 1) Your Medical History Survey: Have you ever been diagnosed by a doctor with any of the following autoimmune conditions?

(Eczema)' (Yes, No, I don't know); 2) The Roots into the Future intake form survey: Has a doctor ever told you that you have any of these skin conditions? Please check all that apply. (Eczema, Keloids, Psoriasis, Other skin condition, I'm not sure, None of the above); 3) Allergies and Asthma survey: Did you have any of the following problems as a child (age 17 or younger)? Eczema (atopic dermatitis) (Yes, No, I'm not sure); 4) Asthma survey: Have you ever had any of the following? Please check all that apply: Atopic dermatitis/Eczema (chronic itchy and scaly skin rashes caused by allergies); 5) Question used in Allergies and Inflammatory Bowel Disease Community surveys: Did you have any of these problems before you were 18 years old? Eczema (atopic dermatitis) (Yes, No, I'm not sure); 6) Question used in Research snippet and Your Profile and Health History survey: Have you ever been diagnosed with eczema? (Yes, No, I'm not sure); 7) Your Health and Health History survey: Question 1: What autoimmune diseases have you been diagnosed with? Please check all that apply: Eczema; Question 2: Have you ever been diagnosed or treated for any of the following conditions? An autoimmune disease (a disease in which your immune system attacks part of your body) (Yes, No, I'm not sure); 8) Your Health and Health History survey: Question 1: What skin conditions have you had? Please check all that apply. Eczema (Yes, No, I'm not sure); Question 2: Have you ever been diagnosed or treated for any of the following conditions? A skin condition; and 9) Health Followup Survey: In the last 2 years, have you been newly diagnosed with or started treatment for any of the following conditions? Eczema (Yes, No, I'm not sure). Answers to these questions were combined according to the following steps: Step 1, a combined phenotype is first assigned based on the first unambiguous response from the questions in sources 1 to 6: cases and controls are first assigned based on the question in “Your Medical History”", then for individuals who were not classified, the subsequent questions in sources 2 to 6 are used in the given order. Step 2, another phenotype is assigned based on the two questions in source 7: cases answered "yes" to the first question and controls answered "no" to both questions. Step 3, a third phenotype is assigned based on the two questions in source 8: cases

answered "yes" and controls answered "no" to both questions. Step 4, a fourth phenotype is assigned based on the one question in source 9: cases answered "yes" to the question and controls answered "no" to the question. The final 'eczema' cases are defined as being "case" in at least one of the above four phenotypes, controls are defined as being "control" in at least one of the above four phenotypes. When discordant, we assume the 'case' answer is correct.

Genotyping, quality control and imputation. DNA extraction and genotyping were performed on saliva samples by CLIA-certified and CAP-accredited clinical laboratories of Laboratory Corporation of America. Samples have been genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550+ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The V3 platform was based on the Illumina OmniExpress+ BeadChip, with custom content to improve the overlap with our V2 array, with a total of about 950,000 SNPs. The V4 platform in current use is a fully custom array, including a lower redundancy subset of V2 and V3 SNPs with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. Samples that failed to reach 98.5% call rate were re-analyzed. Individuals whose analyses failed repeatedly were re-contacted by 23andMe customer service to provide additional samples, as is done for all 23andMe customers.

Participant genotype data were imputed against the March 2012 “v3” release of 1000 Genomes reference haplotypes [8]. We phased and imputed data for each genotyping platform separately. First, we used Beagle (version 3.3.1) [9] to phase batches of 8000-9000 individuals across chromosomal segments of no more than 10,000 genotyped SNPs, with overlaps of 200 SNPs. We excluded SNPs with Hardy-Weinberg equilibrium $P<10^{-20}$, call rate < 95%, or with large allele frequency discrepancies compared to European 1000 Genomes reference data. Frequency discrepancies were identified by computing a 2x2 table of allele counts for European 1000 Genomes samples and 2000 randomly sampled 23andMe

customers with European ancestry, and identifying SNPs with a chi squared P<10−15. We imputed each phased segment against all-ethnicity 1000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac2 [10], using 5 rounds and 200 states for parameter estimation. For the non-pseudoautosomal region of the X chromosome, males and females were phased together in segments, treating the males as already phased; the pseudoautosomal regions were phased separately. We then imputed males and females together using minimac, as with the autosomes, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

Association analysis. For case control comparisons, we computed association test results by logistic regression assuming additive allelic effects. For tests using imputed data, we used the imputed dosages rather than best-guess genotypes. We included as covariates age, gender, and the top five principal components to account for residual population structure. The association test P value we report is computed using a likelihood ratio test, which in our experience is better behaved than a Wald test on the regression coefficient. Results for the X chromosome are computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele.

Acknowledgements. We thank the research participants and employees of 23andMe for making this work possible. We particularly thank the following members of the 23andMe Research Team: Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, Bethann S. Hromatka, Karen E. Huber, Aaron Kleinman, Nadia K. Litterman, Matthew H. McIntyre, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A.M. Northover, Steven J. Pitts, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Joyce Y. Tung, Vladimir Vacic, and Catherine H. Wilson.

GERA (*n*=51,218)

Sample ascertainment and phenotype definition. The Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort includes 110,266 adult members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC) and has been described in detail [11]. In this study, we focused on subjects who were at least 18 years of age at time of the survey who were of non-Hispanic white race/ethnicity. All study procedures were approved by the Institutional Review Board of the Kaiser Foundation Research Institute.

Allergic cases in the GERA cohort were identified from clinical diagnoses captured within the Kaiser EHR system using Internal Classification of Disease, Ninth Edition(ICD9) codes. Asthma was defined using ICD9 codes 493, 493.01, 493.1, 493.11, 493.82, 493.9, 493.91, and 493.92. Hayfever was defined using ICD9 codes 477, 477.1, 477.2, 477.8, and 477.9. Eczema was defined with ICD9 code 691.8. Cases and controls were exluded from the analysis if ever recorded with ICD9 codes related to COPD.

Genotyping, quality control and imputation. DNA samples were extracted from Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada) at KPNC and genotyped at the Genomics Core Facility of the University of California, San Francisco (UCSF) Affymetrix Axiom arrays (Affymetrix, Santa Clara, CA, USA). The design details and genome-wide coverage of those arrays have been previously described [12,13]. High genotype quality control (QC) procedures for the GERA cohort were performed on an array-wise basis as described in detail elsewhere[11]. Using strict QC criteria, including, initial genotyping call rate $\geq$ 97%, allele frequency difference ($\leq$ 0.15) between males and females for autosomal markers, and genotype concordance rate (> 0.75) across duplicate genetic markers, around 94% of samples and more than 98% of genetic markers assayed passed QC procedures [11]. Prior to imputation, we additionally excluded genetic markers with a minor allele frequency (MAF) < 1%, or a

genotype call rate < 90%. Pairwise genetic relatedness was assessed using KING [14]; subjects with kinship coefficients >= 0.177 were excluded.

Following the pre-phase of the genotypes with Shape-IT v2.5 [15], genetic markers were imputed from the cosmopolitan reference panel 1000 Genomes Project (phase I integrated release) using IMPUTE2 v2.3.1 [5]. Herein, we reported imputed markers with info-metric $r^2 \geq 0.3$ and MAF $\geq 1\%$; all reported genotyped markers exceeded a genotype call rate $\geq 98\%$, and a p-value ≥ 0.001 for Hardy-Weinberg equilibrium deviation.

Association analysis. Association analyses were conducted using PLINK10 v1.90b3.39. We assessed single-marker associations with allergic phenotypes using logistic regression. We assumed an additive genetic model using allele counts (i.e., 0, 1, or 2 copies of the minor allele) for typed markers or additive dosages for imputed markers, adjusting for age at survey, sex and 10 ancestry principal components (PCs). To calculate the PCs, we used Eigenstrat v4.2 [16] as previously described [17].

Acknowledgments and funding. This work was supported by NIH postdoctoral training grant CA112355.

CATSS (*n*=11,068)

The Child and Adolescent Twin Study in Sweden (CATSS) is the first of three cohorts from the Swedish Twin Registry (STR) that contributed to the meta-analysis. The other two were TwinGene (born 1911-1958) and Screening Across the Lifespan Twin Study: the Younger (SALTY, born 1943-1958). A detailed summary of the collection of biosamples within these three cohorts is available in previous publications [18-21] and briefly outlined below. Genotyping was dependent on biosample availability and quality of extracted DNA. As the cohorts include both identical (monozygotic, MZ)

and fraternal (dizygotic, DZ) twin pairs, genotyping was carried out for only one of the twins within MZ pairs and imputed in the other. In DZ pairs with available biosamples, both were genotyped when possible. The phenotype definitions within the STR cohorts were created using a combination of questionnaire data; population-based register data sources; the Swedish National Patient Register from 1987, and the Swedish Prescribed Drug Register from July 2005. All studies were approved by the Regional Ethical Review Board in Stockholm, Sweden and all participants gave informed consent. Here, we summarise methods for CATSS, with the other two studies described separately below.

Sample ascertainment and phenotype definition. The CATSS is an ongoing longitudinal twin study targeting all twins born from 1992 and living in Sweden. Since 2004, twins are invited to participate in CATSS following their ninth birthday. During the first three years of data collection, twelve-year-old twins were also invited. Participation in CATSS starts with a parental telephone interview on the children's health, perinatal factors, living situation. A module including questions regarding the twin pair's physical similarities is the basis for an algorithm-based assessment of zygosity. Since 2008, twins have also been offered a DNA-based zygosity test using the saliva samples collected by mail in connection with invitation to the study. DNA from saliva is then stored in the biobank of Karolinska Institutet. To date, approximately n=29100 twins have participated in CATSS-9/12.

Asthma was defined as: 1) Answered YES to the question "You have said that X have or have had asthma, did X get it diagnosed by a doctor?"; or 2) Received at least one of the following diagnoses in the Swedish National Patient Register: ICD-10 J45 or J46 and ICD-9 493; or 3) Had at least two dispenses for any of the following prescriptions in the Swedish Prescribed Drug Register: R03AC, R03AK, R03BA or R03DC.

Hayfever was defined if the study participant had answered YES to at least one of the following: 1) "Has the child been diagnosed with hay fever by a doctor?"; or 2) "Has the child been

diagnosed with pollen allergy by a doctor?"; or 3) "Has the child been diagnosed with fur allergies by a doctor?".

Eczema was defined if the study participant had answered YES to "Has the child been diagnosed with atopic dermatitis or atopic eczema by a doctor?".

Exclusion for the controls included an affirmative response to any of the above criteria; if they had any dispense of R03AC, R03AK, R03BA or R03DC; if they had answered that the child had food allergies like celiac disease or lactose intolerant; or if they answered yes to any of the following: Has or has the child had asthma, hay fever or eczema.

Genotyping, quality control and imputation. Saliva samples from the CATSS and SALTY (described below) cohorts were analysed jointly. DNA was extracted using either the Chemagic STAR instrument from Hamilton Robotics, with magnet bead purification kits from Chemagen, or the Puregene extraction kit (Gentra systems, Minneapolis, USA). Genotyping was performed in 18 batches at the SNP&SEQ Technology Platform in Uppsala, Sweden, using the Illumina PsychArray bead chip. Genotype calls from the zCall algorithm for rare variants were combined with those from the Illumina GenCall algorithm to increase sensitivity at low minor allele frequencies. After initial intensity-level quality control, 18,193 samples remained. Additional QC filtering was applied as follows: SNPs with missingness > 2%, SNPs with more than 10% discordant genotypes across replicates or MZ pairs, SNPs out of Hardy-Weinberg equilibrium (exact test P-value < 10-6), SNPs with clear batch effects or absolute MAF difference from 1000 Genomes European samples > 10%, Y-chromosome and mitochondrial SNPs, and SNPs with minor allele count <= 1 were all excluded. Further, individuals with missingness > 2%, individuals with deviant autosomal heterozygosity (autosomal inbreeding coefficient F outside [-0.02, 0.02]), individuals showing excessive mean relatedness to the rest of the sample (mean relatedness > 6 s.d. above the sample mean), individuals where genotype-based sex did

not match phenotype information, and individuals identified as non-European ancestral outliers were excluded. Non-genotyped monozygous twins in the study were imputed from their genotyped twin, resulting in a total of 21,752 individuals with genotypes. The actual number of samples used for the present analyses was limited by the available phenotype data, and is presented in **Supplementary Table 2**. Genotypes were imputed to the 1000 genomes phase 3 version 5 reference panel using Shapeit v2.r790 [15] and Minimac version 1.0.13 [10].

Association analysis. Analyses were performed on imputed genotype dosages using RAREMETALWORKER version 4.13.8. The software implements a linear mixed model association test, with the kinship matrix constructed from provided zygosity information. Effect estimates and standard errors from the linear mixed model were transformed in order to be comparable to those from logistic models, as described above for the UKBiobank study. Analyses were adjusted for sex and the first four ancestry principal components.

NTR (*n*=10,242)

Sample ascertainment and phenotype definition. Genotyped participants within the Netherlands Twin Register (NTR) represent a cross-section of all NTR participants. The NTR is a longitudinal twin-family study with no other selection criteria than being a multiple or one of their family members. The phenotype data analysed in this study were obtained via surveys sent out to all NTR participants [22,23] and/or collected during an interview in biobank studies [24,25]. As the NTR collects information in families, the genotyped population included related individuals and MZ twins.

Asthma was defined based on nine NTR surveys sent out between 1991 and 2010, which included the following questions: “Has a doctor ever concluded that you suffer from asthma or bronchitis?” and “Have you ever been prescribed medication for asthma or bronchitis?”. Subjects could

answer yes or no to these questions. In addition, in the biobank projects conducted between 2004 and 2011, participants were asked in an interview to indicate any chronic diseases they suffered from and report their medication use. The presence of asthma was indicated when: 1) as part of the biobank project, asthma and/or asthma-specific medication was reported, and/or 2) the participant responded "yes" to one or both questions listed above. In the case of multiple reports, consistency across time was checked. Participants who indicated in the surveys that they never received an asthma diagnosis and never had asthma medication prescribed and who did not report the presence of asthma or asthma medication use in the biobank project were classified as controls.

The same nine surveys as used for asthma also included the questions "Has a doctor ever concluded that you suffer from allergy or hay fever?" and "Have you ever been prescribed medication for allergy or hay fever?". Hay fever was defined in a similar way based on biobank reports of disease and medication and on the survey answers to the question whether they were ever diagnosed with hay fever and/or ever used medication for hay fever.

Eczema was indicated when participants reported eczema in the biobank project when asked for the presence of chronic diseases or reported eczema in the surveys in response to an open question about the presence of diseases. In the case eczema was not indicated in surveys or biobank projects, participants were considered to be controls.

Genotyping, quality control and imputation. DNA samples were genotyped with the Illumina Human Quad-Beadchip 660K, Illumina Omni 1M (1000K), Perlegen-Affymetrix chip 600K or Affymetrix 6.0 760K arrays. Quality control filters included removal of SNPs if allele frequency difference > 0.10 with GONL project samples; MAF < 0.005; HWE < 10-12; and/or call rate < 0.95. Individual samples were removed from the analysis if heterozygosity -0.075 < F < 0.075 (476 samples removed); genotype call rate < 0.90 (457); DNA sex did not match phenotype (395); IBS status of family relations fail (822);

Mendelian error rate > 5 sd's from mean (4); and Affymetrix CQC < 0.40 (322). Ethnic outliers were identified using PC analysis. Genotypes were imputed to the 1000 genomes phase 3 version 5 reference panel using the Michigan Imputation server.

Association analysis. To account for the presence of relatives, we used the mixed linear model based association analysis implemented in GCTA [26], specifically the --mlma-loco option. We included age, sex and 20 principal components for ancestry as covariates. Beta and SE estimates from the linear model were then adjusted to be comparable to those obtained from logistic models as described above for the UKBiobank study.

Acknowledgments and funding. This study was supported by multiple grants from the Netherlands Organization for Scientific Research (NWO: 016-115-035, 463-06-001, 451- 04-034); ZonMW (31160008, 911-09-032); and NWO 480-15-001/674: Netherlands Twin Registry Repository: researching the interplay between genome and environment; The Amsterdam Public Health Institute (APH) and Neuroscience Campus Amsterdam (NCA); Biomolecular Resources Research Infrastructure (BBMRI–NL, 184.021.007), European Research Council (ERC-230374); Genotyping was made possible by grants from NWO/SPI 56-464-14192, Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health, Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the Avera Institute, Sioux Falls (USA) and the National Institutes of Health (NIH R01 HD042157-01A1, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995).

LifeLines (*n*=8,560)

Sample ascertainment and phenotype definition. The LifeLines Cohort Study is a multi-disciplinary

prospective population-based cohort study examining in a unique three-generation design the health and health-related behaviors of 167,729 persons living in the North of the Netherlands. It employs a broad range of investigative procedures in assessing the biomedical, socio-demographic, behavioral, physical and psychological factors which contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics [2]. Between 2006 and 2013, inhabitants of the northern part of The Netherlands and their families were invited to participate, thereby contributing to a three-generation design. Participants visited one of the LifeLines research sites for a physical examination, including lung function, ECG and cognition tests, and completed extensive questionnaires. Baseline data were collected for 167 729 participants, aged from 6 months to 93 years. Follow-up visits are scheduled every 5 years, and in between participants receive follow-up questionnaires. For genotyping only unrelated (usually the oldest person in the family available at the time of genotyping), Caucasian-ancestry samples were selected.

Asthma was defined based on a self-reported doctor diagnosis of asthma provided in health questionnaires. Hay fever cases were participants who reported ever having had nasal allergies, including hay fever. Eczema cases were those who reported ever having had eczema. Controls were defined as having no asthma, no hay fever, and no eczema, In addition, we excluded controls with COPD (defined as FEV1%FVC < 70% or < LLN); self-reported allergies to dust, animals, pollen, food or medication; and/or self-reported contact allergies.

Genotyping, quality control and imputation. DNA samples were genotyped with the Illumina CytoSNP-12v2 array. SNP filtering was done on minor allele frequency (MAF) above 0.001, Hardy-Weinberg equilibrium (HWE) P-value >1e-4, call rate of 0.95. Sample filtering was done on a call rate of 0.95 and a principal component analysis (PCA) to check for population outliers. Genotypes were pre-phased using SHAPEIT2 and aligned to reference panels using Genotype Harmonizer

[www.molgenis.org/systemsgenetics] in order to resolve strand issues. The samples were imputed using Minimac (v2012.10.3).

Association analysis. Logistic regression analyses were performed on imputed genotype dosages using PLINK version 1.90b3.32. Analyses were adjusted for age and sex.

Acknowledgements and Funding. The LifeLines Biobank initiative has been made possible by funds from FES (Fonds Economische Structuurversterking), SNN (Samenwerkingsverband Noord Nederland) and REP (Ruimtelijk Economisch Programma). The authors wish to acknowledge the services of the LifeLines Cohort Study, the contributing research centres delivering data to LifeLines, and all the study participants.

TwinGene (*n*=5,517)

Sample ascertainment and phenotype definition. The TwinGene project, conducted between 2004 and 2008, is a population-based Swedish study of twins born between 1911 and 1958. The study participants have previously participated in a telephone interview called Screening Across the Lifespan Twin Study (SALT), conducted between 1998 and 2002. To be included in TwinGene, both twins within a pair had to be alive. In total, 12,591 individuals participated by donating blood at the study participants' local healthcare facilities, and by answering questionnaires about life style and health.

Asthma was defined as: 1) answered YES to: "You have said that you have or have had asthma, have you been diagnosed with asthma by a physician at the clinic or hospital?"; or 2). Received at least one of the following diagnoses in the Swedish National Patient Register: ICD-10 J45 or J46 and ICD-9 493; or 3) Had at least two dispenses for any of the following prescriptions in the Swedish Prescribed Drug Register: ATC-codes R03AC, R03AK, R03BA or R03DC. Exclusion: Those who had received

the diagnosis COPD (extracted from the National Patient Register ICD-10 J44 or ICD-9 490).

Hayfever was defined if the study participant had answered YES to at least one of the following: 1) “You have said that you have or have had hay fever, have you been diagnosed with hay fever by a doctor”; or 2) “You have said that you have or have had pollen allergy, have you been diagnosed with pollen allergy by a physician at the clinic or hospital?”; or 3) “You have said that you have or have had fur allergies, have you been diagnosed with fur allergies by a physician at the clinic or hospital?”

Eczema was defined if an answer YES to the following: “You have said that you have or have had eczema, have you been diagnosed with eczema by a physician at the clinic or hospital?” and chosen one of the following to the question “What diagnosis did you get?”: 1) atopic eczema; or 2) atopic dermatitis.

Exclusion for the controls was if any of above criteria were met, if they had any dispense of ATC codes R03AC, R03AK, R03BA or R03DC; if they had celiac disease or was lactose intolerant; if they had COPD (extracted from the National Patient Register ICD-10 J44 or ICD-9 490); or if they answered yes to any other allergy or eczema.

Genotyping, quality control and imputation. DNA was extracted from whole blood using the Puregene extraction kit (Gentra systems, Minneapolis, USA). After excluding subjects with DNA concentration below 20ng/µl and 302 previously genotyped individuals, DNA from 9,896 individual subjects was sent to the SNP&SEQ Technology Platform Uppsala, Sweden for genome-wide genotyping with Illumina OmniExpress bead chip (all available dizygous twins and one twin from each available MZ twin pair). Genotyping results for 9,836 subjects passed the initial technical quality control. Additional QC filtering was applied as follows: SNPs with missingness > 3%, individuals with missingness > 3%, SNPs with minor allele frequency < 1%, SNPs out of Hardy-Weinberg equilibrium (exact test P-value

< 10-7) were excluded. Further, individuals where genotype-based sex did not match phenotype information, individuals with excess autosomal heterozygosity (> 5 standards deviations from population mean), and cryptically related individuals were excluded. 9,617 individuals passed this filtering. The actual number of samples used for the present analyses was limited by the available phenotype data, and is presented in **Supplementary Table 2**. Genotypes were imputed to the 1000 genomes phase 1 version 3 reference panel using minimac (v2012-10-03) and mach1 version 1.0.18.c.

Association analysis. Analyses were performed on imputed genotype dosages using PLINK version 1.90b3.36. Only one individual per twin pair was included. In outcome-discordant twin pairs, cases were preferably selected. In concordant pairs, one individual with available genotype data was selected at random. Analyses were adjusted for sex and the first four ancestry principal components.

ALSPAC (*n*=4,964)

Sample ascertainment and phenotype definition. The Avon Longitudinal Study of Parents and Children (ALSPAC) recruited 15,247 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992, resulting in 14,775 live births and 14,701 children who were alive at 1 year of age. Enrolment is described in more detail in the cohort profile paper [27] and via the website http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/. Biological samples including DNA have been collected for 10,121 of the children from this cohort. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees and written and informed consent was provided by the parents. The children have been followed up with regular questionnaires and clinic visits. For the current study data collected from the questionnaires was used to classify children as allergic disease cases or controls. When the children were approximately 10 years old, parents were asked if a doctor had ever stated that their child has

asthma or eczema. When the children were approximately 14 years old, parents were asked if their child had ever had hay fever. Cases were defined as a positive response to the asthma, or eczema or hay fever questions. Controls were defined as a negative response to all three.

Genotyping, quality control and imputation. DNA samples were genotyped using the Illumina HumanHap550 quad chip genotyping platforms. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corportation of America) using support from 23andMe. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness (>3%) and insufficient sample replication (IBD < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95% or evidence for violations of Hardy-Weinberg equilibrium (P < 5E-7) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD > 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 subjects and 500,527 SNPs passed these quality control filters.

Imputation was carried out on the ALSPAC children in combination with the mothers. We combined 477,482 SNP genotypes in common between the sample of mothers and sample of children. We removed SNPs with genotype missingness above 1% due to poor quality (11,396 SNPs removed) and removed a further 321 subjects due to potential ID mismatches. This resulted in a dataset of 17,842 subjects containing 6,305 duos and 465,740 SNPs (112 were removed during liftover and 234 were out of HWE after combination). We estimated haplotypes using ShapeIT (v2.r644) which utilises

relatedness during phasing. We obtained a phased version of the 1000 genomes reference panel (Phase 1, Version 3) from the Impute2 reference data repository (phased using ShapeIt v2.r644, haplotype release date Dec 2013). Imputation of the target data was performed using Impute V2.2.2 against the reference panel (all polymorphic SNPs excluding singletons), using all 2,186 reference haplotypes (including non-Europeans). This gave 8,237 eligible children and 8,196 eligible mothers with available genotype data after exclusion of related subjects using cryptic relatedness measures described previously.

Association analysis. Genome-wide association analysis was carried out using an additive model in SNPTESTv2.5, with sex as a covariate. For the X chromosome, sex-specific analyses were run, assuming X inactivation in the females. Results from males and females were then meta-analysed with METAL [28] using a fixed-effects model.

Acknowledgements and Funding. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. LP was funded by a UK MRC fellowship award (MR/J012165/1). LP, JZ, EM, CM, KB work in a unit funded by the UK MRC (MC_UU_12013).

SALTY (*n*=4,062)

Sample ascertainment and phenotype definition. The study participants had participated in a telephone interview called Screening Across the Lifespan Twin Study (SALT), conducted between 1998 and 2002. The target population for SALTY was the younger part of the SALT cohort born between 1943 and 1958. The data collection consisted of three parts: (1) an extensive self-report paper-questionnaire; (2) saliva collection for DNA extraction; and (3) a request to participate in an Internet-based investigation. Saliva samples were collected from the study participants either by mail in connection with invitation to the study. Some of the participants in SALTY were also prior participants of TwinGene – if they had already provided a blood sample they were not also asked to provide saliva. Case-control definition for asthma, hay fever and eczema followed the same approach described for the TwinGene study described above.

Genotyping, quality control and imputation. As described above for the CATSS study.

Association analysis. As described above for the CATSS study.

AAGC (*n*=2,435)

Sample ascertainment and phenotype definition. A total of 2,669 physician-diagnosed asthmatics and 4,528 persons without asthma of European ancestry were genotyped for a recent GWAS of asthma described in detail elsewhere [29]. For the present study, we selected the subset of participants for whom hayfever and eczema information was also available, including 1,975 individuals with a history of asthma, hayfever and/or eczema (as cases) and 460 individuals without a history of any of these three diseases (as controls). Cases and controls were drawn from three studies: Queensland Institute of Medical Research (QIMR: 1,008 cases, 432 controls), the Lung Institute of Western Australia (LIWA: 599 cases, 28 controls), and the Tasmanian Longitudinal Health Study (TAHS: 368 cases). The

questionnaire items used to classify hay fever in each of these studies have been described previously [30]. To classify eczema in the QIMR study, X different items were used: 1) Canberra sub-study: "How often have you had Eczema? Cases: "Only as a child", "Rarely" or "Quite often". Controls: "Never". 2) ALC1 sub-study: How often have you had Eczema [before and after age 14]? Cases: "Sometimes" or "Often", before or after age 14. Controls: "Never", before and after age 14. 3) Asthma sub-study: "Have you ever had eczema? Cases: "Yes". Controls: "No". 4) Eczema sub-study: Has a doctor ever diagnosed you as suffering from Eczema? Cases: "Yes". Controls: "No". 5) Adolescent sub-study: "Have you ever suffered from eczema?" Cases: "Yes". Controls: "No". In the LIWA and TAHS studies, eczema was classified based on the question "Have you ever had eczema?".

Genotyping, quality control and imputation. AAGC. All 2,435 samples were genotyped with the Illumina 610K array. SNPs were excluded from analysis if the call rate was <95%, minor allele frequency (MAF) < 0.01 and Hardy-Weinberg equilibrium test P-value < 10-6. SNPs passing QC were then used to impute with Impute2 5.7 million variants with a MAF ≥ 0.01 and information > 0.3 using the combined 1000 Genomes Project and HapMap 3 phased data (all ancestral groups) as reference panels. X chromosome SNPs (N =140,388 with MAF ≥ 0.01) were imputed with r^2 > 0.3 with MACH/minimac using the 1000G Phase I Integrated Release Version 3 (N = 584 European haplotypes) release of the 1000 Genomes Project. All subjects were confirmed to be unrelated and of European ancestry through the analysis of genome-wide allele sharing.

Association analysis. SNPs were tested for association with case-control status using logistic regression in SNPTEST, with the options -method expected and -frequentist 1. Sex and age were included as covariates. X-chromosome SNPs were analysed with logistic regression under an additive model (dosage for imputed variants) using MACH2DAT and assuming a dosage compensation model, ie.

equating hemizygous males to homozygous females, such that the allelic dosage extremes for males were 0 (if A/-, as for AA females) and 2 (if B/-, as for BB females). Males and females were analysed separately, and results then meta-analysed with METAL [28] using a fixed-effects model.

Acknowledgments and funding. We thank all participants of the QIMR, TAHS and LIWA studies, who made this project possible. The AAGC was funded by a grant from the NHMRC (project ID 613627). M.A.F. was supported by a Research Fellowship from NHMRC (APP1124501).

GENEVA (*n*=2,633)

Sample ascertainment and phenotype definition. Atopic dermatitis patients were recruited from tertiary dermatology clinics based at three centers (Technische Universität München, as part of the GENEVA study, University of Kiel, University of Bonn) and diagnosed on the basis of skin examination by experienced dermatologists according to standard criteria [31]. Comorbidities of asthma and hayfever were recorded during physical examination. Individuals from the population-representative PopGen biorepository [32] and the population-based KORA F4 study [33] were comprehensively characterized for allergic phenotypes. AD, asthma and hayfever were defined based on a questionnaire.

Genotyping, quality control and imputation. Individuals were genotyped using the Affymetrix 6.0 or Illumina 550k array. Samples with extensive missing rate (>5%), excess of heterozygosity or homozygosity, and/or ambiguous sex were excluded. We examined allele sharing between individuals and excluded close relatives (PI_HAT>0.1875, which is the IBD expected between third- and second degree relatives) as well as outliers of unusual ancestry by MDS analysis. SNPs were excluded if genotyping rate was <95%, minor allele frequency was <1%, and/or Hardy-Weinberg equilibrium P<10-8. Imputation was carried out separately by array type. Pre-phasing was carried out with

SHAPEIT [15] and imputation with IMPUTE2 [5] using phase I 1000 Genomes reference panel (integrated variant set of all populations, release March 2012). Post imputation SNPs with low imputation quality (info score<0.4), call rate <95%, deviation from Hardy-Weinberg equilibrium (P<10-8) or minor allele frequency <5% were excluded.

Association analysis. Association analysis was carried out with SNPTEST using a frequentist approach with allele dosages (option –method expected) to account for imputation uncertainty adjusted. Sex was included as a covariate. For the X-chromosome, males and females were analysed separately assuming full X inactivation in females, and results then meta-analysed with METAL [28] using a fixed-effects model.

Acknowledgments and funding. The project received infrastructure support through the DFG Clusters of Excellence "Inflammation at Interfaces" (grants EXC306 and EXC306/2), and was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (sysINFLAME, grant # 01ZX1306A), and the PopGen 2.0 network (01EY1103). The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

GENUFAD-SHIP-1 (*n*=1,781)

Sample ascertainment and phenotype definition. Atopic dermatitis patients were recruited at Charité Universitätsmedizin Berlin, Germany for the extended GENUFAD study (GEnetic analysis of NUclear

Families with Atopic Dermatits) and have been described previously (Berlin cases from Set 1 [34,35]). Children of the extended GENUFAD study were recruited through early onset eczema (<2years) with moderate to severe disease expression, which was diagnosed by a physician according to standard criteria [36]. In addition, asthma status and hay fever status were determined at recruitment by a physician. A total of 417 atopic dermatitis patients were included in the present study. Controls originated from the population-based Study of Health in Pomerania (SHIP) [37], which recruited individuals in the north-eastern part of Germany. Only those individuals were included as controls who at recruitment had confirmed in an interview not to have eczema, asthma, or hay fever. The SHIP set was split for two case-control studies by a random function. 1364 unrelated individuals were included in SHIP-1.

Genotyping, quality control, and imputation. All cases and controls were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0. Individuals with a call rate < 0.95 were excluded from the study. In addition, samples were excluded when the gender estimated from X-chromosome heterozygosity (PLINK –sex check) did not match the clinical records. European ancestry of all participants was confirmed by performing a principal component analysis using EIGENSTRAT (SMARTPCA) [38]. Relatedness between individuals was determined by using GRR and Structure.

Before imputation, SNPs were filtered out according to the following criteria: i) low call rate (< 0.95 in cases or controls); ii) low allele frequency (MAF < 0.01 in cases or controls); iii) genotypes out of Hardy-Weinberg equilibrium ($p < 0.00001$ in cases or $p < 0.0005$ in controls). In addition, SNPs with a call rate lower than 0.99 were excluded if having MAF < 0.05 or if they were out of Hardy-Weinberg equilibrium ($p < 0.001$). Only SNPs fulfilling the above mentioned QC were used in subsequent steps. Genotypes from cases and controls were recoded to the plus strand according to the 1000Genomes phase1 / GIANT data using the –flip command from PLINK [3]. Additional markers were deleted if: i) 3

alleles were detected ii) the allele frequencies in the SHIP control population differed more than 0.1 compared with the frequency in 379 Europeans available from the 1000 Genomes project. After filtering, 643,539 SNPs remained in the analysis. All samples were imputed together with Mach/Minimac {Howie, 2012; Li, 2010) using the 1000Genomes phase1 / GIANT reference panel.

Association analysis. Association between SNP dosage and disease status was tested with logistic regression using Mach2dat v1.0.23, with sex and two ancestry principal components (PCs) included as covariates.

Acknowledgements and funding. We thank all individuals and families for their participation in this study. We thank all physicians and nurses involved in patient recruitment for their valuable contribution to the study. We are grateful to the laboratory technicians C. Flachmeier and T. Thuss for their work. The study was funded by the German Ministry of Education and Research (BMBF) through the Clinical Research Group for Allergy at Charité Berlin, the National Genome Research Network (NGFN). The SHIP authors are grateful to Mario Stanke for the opportunity to use his server cluster for SNP imputation. We thank all staff members and participants of the SHIP studies, as well as all of the genotyping staff for generating the SHIP SNP data set. SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network ‘Greifswald Approach to Individualized Medicine (GANI_MED)’ funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data were supported by the Federal Ministry of Education and Research (grant 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany, and the Federal State of Mecklenburg–West Pomerania. The

University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

GENUFAD-SHIP-2 (*n*=1,735)

Sample ascertainment and phenotype definition. All atopic dermatitis patients were recruited at Charité Universitätsmedizin Berlin for the GENUFAD study (GEnetic analysis of NUclear Families with Atopic Dermatitis) and have been described in a previous GWAS (set 2 in original study {Esparza-Gordillo, 2009}). 270 German families were recruited through two affected siblings with an age of onset below two years of age and moderate to severe disease expression. AD was diagnosed by a physician according to standard criteria [36]. In addition, asthma status and hay fever status were determined at recruitment. One affected child was selected from each family. A total of 262 unrelated atopic dermatitis patients were included in the present study. Controls originated from the population-based Study of Health in Pomerania (SHIP) [37], which recruited individuals in the north-eastern part of Germany. Only those individuals were included as controls who at recruitment had confirmed in an interview not to have eczema, asthma, or hay fever. The SHIP set was split for two case-control studies by a random function. 1473 unrelated individuals were included in SHIP-2.

Genotyping, quality control, and imputation. All cases were genotyped with Affymetrix 500K array and only samples with a high call rate (>0.95) were included in the analysis. Controls were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0 and were excluded if the call rate was <0.96. From both sets, cases and controls, samples were excluded if the gender estimated from X-chromosome heterozygosity (PLINK –sex check) did not match the clinical records. European ancestry of all participants was confirmed by performing a principal component analysis using EIGENSTRAT (SMARTPCA) [38]. Relatedness between individuals was determined by using GRR and Structure. SNPs

from the 500K array were filtered as previously described [35] according to the following criteria: i) low call rate (< 0.95); ii) low allele frequency (MAF < 0.01); iii) Mendelian errors in 5 or more families; iv) unlikely genotypes in more than 5 families (double recombinants as detected by Merlin [39]; v) founder genotypes out of Hardy-Weinberg equilibrium ($p < 0.00001$). In addition, SNPs with a call rate lower than 0.99 were excluded if having MAF < 0.05 or if they were out of Hardy-Weinberg equilibrium ($p < 0.001$). SNPs on the Human SNP Array 6.0 were excluded if having: i) low call rate (< 0.97), ii) low allele frequency (MAF < 0.01), iii) genotypes out of Hardy-Weinberg equilibrium ($p < 0.0005$). In addition, SNPs with a call rate lower than 0.99 were excluded if having MAF < 0.05 or if they were out of Hardy-Weinberg equilibrium ($p <0.001$). Only SNPs fulfilling the above mentioned QC and overlapping between both sets were used in subsequent steps, non-overlapping SNPs were excluded. Genotypes from cases and controls were recoded to the plus strand using the –flip command from PLINK [3] and merged. Additional markers were deleted if: i) 3 alleles were detected ii) the allele frequencies in the SHIP control population differed in more than 0.1 to the 379 Europeans available from the 1000 Genomes project. After filtering, 345407 SNPs remained in the analysis. All samples were imputed together with Mach/Minimac {Howie, 2012; Li, 2010) using the 1000Genomes phase1 / GIANT reference panel.

Association analysis. Association between SNP dosage and disease status was tested with logistic regression using Mach2dat v1.0.23, with sex and two ancestry principal components (PCs) included as covariates.

Acknowledgements and funding. We thank all individuals and families for their participation in this study. We thank all physicians and nurses involved in patient recruitment for their valuable contribution to the study. We are grateful to the laboratory technicians C. Flachmeier and T. Thuss for their work.

The study was funded by the German Ministry of Education and Research (BMBF) through the Clinical Research Group for Allergy at Charité Berlin, the National Genome Research Network (NGFN). The SHIP authors are grateful to Mario Stanke for the opportunity to use his server cluster for SNP imputation. We thank all staff members and participants of the SHIP studies, as well as all of the genotyping staff for generating the SHIP SNP data set. SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data were supported by the Federal Ministry of Education and Research (grant 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany, and the Federal State of Mecklenburg–West Pomerania. The University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

HUNT (*n*=19,564)

Sample ascertainment and phenotype definition. The Nord-Trøndelag Health Study (HUNT) is a series of population-based health studies conducted in the county of Nord-Trøndelag, Norway. The study was carried out at three different time points over approximately 20 years (HUNT1 [1984-1986], HUNT2 [1995-1997] and HUNT3 [2006-2008]]){Krokstad, 2013}. At each study, the entire adult population (≥ 20 years) was invited to participate by completing questionnaires, attending clinical examinations and interviews. Participation rates have generally been high: 89.4% (*n*=77,212), 69.5% (*n*=65 237) and 54.1% (*n*=50 807) in HUNT1, HUNT2 and HUNT3, respectively [40]. Taken together, the health studies included information from over 120,000 different individuals from Nord-Trøndelag. Biological samples

including DNA have been collected for approximately 70,000 participants. The HUNT Study has been described in more detail elsewhere [40]. For the present study, we included participants from HUNT2 and HUNT3 as information on allergic disease was not available in HUNT1. Asthma cases were defined by a positive response to the question "Have you been diagnosed as having asthma by a doctor?" in HUNT2 and/or "Has a doctor diagnosed you as having asthma?" in HUNT3. Participants with inconsistent responses and those also reporting self-reported ever or self-reported doctor-diagnosed COPD were excluded. Hay fever cases were defined by a positive response to the questions "Do you have hay fever or nasal allergies?" and/or "Do you have or have you had hay fever or nasal allergies?" in HUNT3. Participants with inconsistent responses were excluded. Eczema cases were defined by a positive response to the questions "Have you had or do you have any of the following diseases: Eczema on hands" in HUNT3.

Genotyping, quality control and imputation. In total, DNA from 71,860 HUNT samples was genotyped using one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). Samples that failed to reach a 99% call rate, had contamination > 2.5% as estimated with BAF Regress [41], large chromosomal copy number variants, lower call rate of a technical duplicate pair and twins, gonosomal constellations other than XX and XY, or whose inferred sex contradicted the reported gender, were excluded. Samples that passed quality control were analysed in a second round of genotype calling following the Genome Studio quality control protocol described elsewhere [42]. Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; http://genome.ucsc.edu) using BLAT [43]. PLINK v1.90 [44] was then used to exclude variants if their probe sequences could not be perfectly mapped, cluster

separation was < 0.3, Gentrain score < 0.15, showed deviations from Hardy Weinberg equilibrium in unrelated samples of European ancestry with p-value < 0.0001), had a call rate < 99%, or another assay with higher call rate genotyped the same variant. Ancestry of all samples was inferred by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/) [45,46], using PLINK. Recent European ancestry was defined as samples that fell into an ellipsoid spanning exclusively European populations of the HGDP panel. The different arrays were harmonized by reducing to a set of overlapping variants and excluding variants that showed frequency differences > 15% between data sets, or that were monomorphic in one and had MAF > 1% in another data set. The resulting genotype data were phased using Eagle2 v2.3 [47]. Imputation was performed on the samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3)[48] and the Haplotype Reference Consortium reference panel (release version 1.1) [49]. A maximal set of relatively unrelated individuals (kinship coefficient < 0.0884) was chosen using KING [14] and FastIndep [50]. After restricting to the above and those with available phenotype information, 19,564 individuals were included in the analysis sample.

Association analysis. Association analysis were conducted using EPACTS-3.3. The SNP-phenotype associations were modeled using the Wald statistic in logistic regression, assuming an additive genetic model for genotyped markers and imputed genotypes. Models were adjusted for sex, birth year, genotyping batch and four principal components (PCs). PCs were computed using PLINK. Additional filters applied to the analysis included MAF $\geq$ 0.5% and imputation $r^2 \geq 0.3$.

Acknowledgements and Funding. The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine, NTNU, Norwegian University of

Science and Technology), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. We are grateful for the contributions from Anne Heidi Skogholt, He Zhang and Hyun Min Kang. We would also like to acknowledge the support given to us by the Genotyping core and Jin Chen. B.M.B. received a research grant from the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology. J.B.N. was supported by personal grants from the Danish Heart Foundation and the Lundbeck Foundation.

2. Contributors to the 23andMe Research Team

Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, David A. Hinds, Bethann S. Hromatka, Karen E. Huber, Aaron Kleinman, Nadia K. Litterman, Matthew H. McIntyre, Joanna L. Mountain, Carrie A.M. Northover, J. Fah Sathirapongsasuti, Olga V. Sazonova, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Vladimir Vacic, Catherine H. Wilson, Steven J. Pitts

23andMe, Inc., Mountain View, CA, USA

3. Collaborators of the Australian Asthma Genetics Consortium

Dale R. Nyholt[a], John Beilby[b-d], Loren Price[e], Faang Cheah[e], Desiree Mészáros[f], Scott D. Gordon[a], Melissa C. Southey[g], Margaret J. Wright[a], James Markos[h], Li P. Chung[e], Anjali K. Henders[a], Graham Giles[i], Suzanna Temple[e], John Whitfield[a], Brad Shelton[e], Chalermchai Mitrpant[e], Minh Bui, PhD,[j] Mark Jenkins[j], Haydn Walters[f], Michael J. Abramson[k], Michael Hunter[l,d], Bill Musk [l,d,m,n], Matthew A Brown[o], Shyamali C. Dharmage[j], Jennie Hui[d,l,p,q], Svetlana Baltic[e], Peter Le Souëf[r], Grant W. Montgomery[a], Colin F. Robertson[s], Alan James[d,m,t], Guy Marks[u]

[a] QIMR Berghofer Medical Research, Brisbane, Australia.

[b] PathWest Laboratory Medicine of Western Australia (WA), Nedlands, Australia.

[c] School of Pathology and Laboratory Medicine, The University of WA, Nedlands, Australia.

[d] Busselton Population Medical Research Foundation, Sir Charles Gairdner Hospital, Perth, Australia.

[e] Lung Institute of WA and Centre for Asthma, Allergy and Respiratory Research, University of WA, Perth, Australia.

[f] Menzies Research Institute, Hobart, Australia.

[g] Department of Pathology, The University of Melbourne, Melbourne, Australia.

[h] Launceston General Hospital, Launceston, Australia.

[i] Cancer Epidemiology Centre, The Cancer Council Victoria, Melbourne, Australia.

[j] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Australia.

[k] Department of Epidemiology & Preventive Medicine, Monash University, Melbourne, Australia

[l] School of Population Health, The University of WA, Nedlands, Australia

[m] School of Medicine and Pharmacology, University of Western Australia, Nedlands, Australia

[n] Department of Respiratory Medicine, Sir Charles Gairdner Hospital, Perth, Australia

[o] The University of Queensland Diamantina Institute, Translational Research Institute, Australia

[p] PathWest Laboratory Medicine of Western Australia (WA), Nedlands 6009, Australia

[q] School of Pathology and Laboratory Medicine, The University of WA, Crawley 6009, Australia

[r] School of Paediatrics and Child Health, Princess Margaret Hospital for Children, Subiaco 6008, Australia

[s] Respiratory Medicine, Murdoch Children's Research Institute, Melbourne 3052, Australia

[t] Department of Pulmonary Physiology and Sleep Medicine, West Australian Sleep Disorders Research Institute, Nedlands 6009, Australia

[u] Woolcock Institute of Medical Research, University of Sydney, Sydney 2037, Australia

4. BIOS Consortium (Biobank-based Integrative Omics Study)

Management Team Bastiaan T. Heijmans (chair)[1], Peter A.C. 't Hoen[2], Joyce van Meurs[3], Aaron Isaacs[4], Rick Jansen[5], Lude Franke[6].

Cohort collection Dorret I. Boomsma7, René Pool7, Jenny van Dongen7, Jouke J. Hottenga7 (Netherlands Twin Register); Marleen MJ van Greevenbroek[8], Coen D.A. Stehouwer[8], Carla J.H. van der Kallen[8], Casper G. Schalkwijk[8] (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga[6], Lude Franke[6], Sasha Zhernakova[6], Ettje F. Tigchelaar[6] (LifeLines Deep); P. Eline Slagboom[1], Marian Beekman[1], Joris Deelen[1], Diana van Heemst[9] (Leiden Longevity Study); Jan H. Veldink[10], Leonard H. van den Berg[10] (Prospective ALS Study Netherlands); Cornelia M. van Duijn[4], Bert A. Hofman[11], Aaron Isaacs[4], André G. Uitterlinden[3] (Rotterdam Study).

Data Generation Joyce van Meurs (Chair)[3], P. Mila Jhamai[3], Michael Verbiest[3], H. Eka D. Suchiman[1], Marijn Verkerk[3], Ruud van der Breggen[1], Jeroen van Rooij[3], Nico Lakenberg[1].

Data management and computational infrastructure Hailiang Mei (Chair)[12], Maarten van Iterson[1], Michiel van Galen[2], Jan Bot[13], Dasha V. Zhernakova[6], Rick Jansen[5], Peter van 't Hof[12], Patrick Deelen[6], Irene Nooren[13], Peter A.C. 't Hoen[2], Bastiaan T. Heijmans[1], Matthijs Moed[1].

Data Analysis Group Lude Franke (Co-Chair)[6], Martijn Vermaat[2], Dasha V. Zhernakova[6], René Luijk[1], Marc Jan Bonder[6], Maarten van Iterson[1], Patrick Deelen[6], Freerk van Dijk[14], Michiel van Galen[2], Wibowo Arindrarto[12], Szymon M. Kielbasa[15], Morris A. Swertz[14], Erik. W van Zwet[15], Rick Jansen[5], Peter-Bram 't Hoen (Co-Chair)[2], Bastiaan T. Heijmans (Co-Chair)[1].

1. Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

2. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

3. Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands

4. Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands

5. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

6. Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands

7. Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

8. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands

9. Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands

10. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

11. Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands

12. Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands

13. SURFsara, Amsterdam, the Netherlands

14. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

15. Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

5. Collaborators of the LifeLines Cohort Study

U. Bultmann[1], J.M. Geleijnse[2], P. van der Harst[3], S. Mulder[4], J.G.M. Rosmalen[5], E.F.C. van Rossum[6], H.A. Smit[7], M.A Swertz[8,9], E.A.L.M. Verhagen[10], B.Z. Alizadeh[11], H.M. Boezen[11], L. Franke[8], P. Deelen[8,9], G. Navis[12], M. Rots[13], H. Snieder[11], F. van Dijk[8,9] B.H.R. Wolffenbuttel[14], C. Wijmenga[8]

[1]University of Groningen, University Medical Center Groningen, Department of Social Medicine, Groningen, The Netherlands

[2]Wageningen University, Department of Human Nutrition, Wageningen, The Netherlands

[3]University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands

[4]Lifelines Cohort Study, Groningen, The Netherlands

[5]University of Groningen, University Medical Center Groningen, Interdisciplinary Center of Psychopathology of Emotion Regulation (ICPE), Department of Psychiatry, Groningen, The Netherlands

[6]Erasmus Medical Center, Department of Endocrinology, Rotterdam, The Netherlands

[7]University Medical Center Utrecht, Department of Public Health, Utrecht, The Netherlands

[8]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands

[9]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands

[10]VU Medical Center, Department of Public and Occupational Health, Amsterdam, The Netherlands

[11]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands

[12]University of Groningen, University Medical Center Groningen, Department of Nephrology,

Groningen, The Netherlands

[13]University of Groningen, University Medical Center Groningen, Department of Medical Biology, Groningen, The Netherlands

[14]University of Groningen, University Medical Center Groningen, Department of Endocrinology, Groningen, The Netherlands

6. Literature supporting a possible role in allergic disease for genes highlighted in Table 1

Gene	References
RERE	51-54
PPP2R3C	55-58
RASA2	59-61
SIK2	62,63
RTF1	64-67
SMARCE1	68
DYNAP	69,70
THEM4	71,72
ARHGAP15	73-76
SENP7	77

7. Procedure used to identify variants reported to be associated with allergic disease in previous GWAS

We downloaded the full NHGRI-EBI GWAS catalog database [78] on August 14, 2017 (file gwas_catalog_v1.0.1-associations_e89_r2017-07-31.tsv). We then identified SNP associations with a $P \leq 5x10^{-8}$ and that were reported for an allergic condition, specifically for which the "MAPPED_TRAIT" variable included the terms "allergic rhinitis", "allergic sensitization", "allergy", "asthma", "eczema" and "atopic march". There were 240 such associations reported in 34 studies. Of these, we excluded 23 associations reported in four studies because they were for asthma age-of-onset (two associations)[79], diisocyanate-induced asthma (eight associations)[80], comparative analyses of psoriasis and atopic dermatitis [AD] (10 SNPs co-associated with psoriasis and AD)[81] and rank-based analyses (three associations)[82]. The remaining 217 associations included 185 unique rs IDs. We then used the --clump procedure in PLINK [3] and genotype data from individuals of European descent from the 1000 Genomes Project [8] (*n*=294, release 20130502_v5a) to reduce this list of 185 SNPs to variants in low linkage disequilibrium (LD) with each other ($r^2<0.05$), which are likely to represent statistically independent associations with allergic disease. After excluding four variants that were not in the 1000 Genomes Project dataset used and were located within 1 Mb of other published allergy risk variants (rs7212938 [in *GSDMA*], rs10056340 [near *TSLP*], rs9273349 and rs3095318 [both in the MHC]), we identified 89 variants in low LD with each other. We then identified the earliest GWAS to report an association with each of these 89 variants (or with a SNP with $r^2>0.05$ with it) and used the year of publication to generate **Supplementary Fig. 2**.

8. Procedure used to identify genes that were unlikely to have been previously implicated in the pathophysiology of allergic disease

We performed the following PubMed query on August 14th, 2017: (asthma OR rhinitis OR eczema OR atopic OR dermatitis OR allergy OR allergi* OR hayfever OR "hay fever") AND (gene1 OR gene1_aliases OR gene2 OR gene2_aliases OR … OR gene132 OR gene132_aliases). The gene symbols approved by the HUGO Gene Nomenclature Committee (HGNC) for each of the target genes, as well as all aliases listed in the Bioconductor package org.Hs.eg.db, were inserted into the second part of that query. The search was restricted to the title and abstract fields. The search results were downloaded as an .xml file and the number of unique articles (based on PMID) co-citing the allergy-related terms and each gene symbol/alias was counted using in-house scripts (results in **Supplementary Table 15**). The articles identified were reviewed to confirm that the gene symbol/alias term was cited to refer to the appropriate gene and not as an abbreviation for an unrelated term (e.g. MBP1 used as abbreviation for major basic protein 1, and not as an alias for the *EFEMP2* gene). When that was the case, the ambiguous alias term was dropped and the search query repeated. To identify genes likely to have been implicated in immune-related processes, we repeated this approach but replaced the first part of the PubMed query with (immune OR immuni* OR immunol*).

9. References

1 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

2 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

3 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:S0002-9297(07)61352-4 [pii] 10.1086/519795 (2007).

4 Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet. Respiratory medicine* **3**, 769-781, doi:10.1016/S2213-2600(15)00283-0 (2015).

5 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

6 Loh, P. R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**, 1385-1392, doi:10.1038/ng.3431 (2015).

7 Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7**, e34267, doi:10.1371/journal.pone.0034267 (2012).

8 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

9 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097, doi:10.1086/521987 (2007).

10 Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-784, doi:10.1093/bioinformatics/btu704 (2015).

11 Kvale, M. N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051-1060, doi:10.1534/genetics.115.178905 (2015).

12 Hoffmann, T. J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89, doi:10.1016/j.ygeno.2011.04.005 (2011).

13 Hoffmann, T. J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-430, doi:10.1016/j.ygeno.2011.08.007 (2011).

14 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

15 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2011).

16 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:ng1847 [pii] 10.1038/ng1847 (2006).
17 Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-1295, doi:10.1534/genetics.115.178616 (2015).
18 Lichtenstein, P. *et al.* The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *J Intern Med* **252**, 184-205 (2002).
19 Lichtenstein, P. *et al.* The Swedish Twin Registry in the third millennium: an update. *Twin Res Hum Genet* **9**, 875-882, doi:10.1375/183242706779462444 (2006).
20 Anckarsater, H. *et al.* The Child and Adolescent Twin Study in Sweden (CATSS). *Twin Res Hum Genet* **14**, 495-508 (2011).
21 Magnusson, P. K. *et al.* The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin Res Hum Genet* **16**, 317-329, doi:10.1017/thg.2012.104 (2013).
22 Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res Hum Genet* **16**, 271-281, doi:10.1017/thg.2012.140 S1832427412001405 [pii] (2013).
23 van Beijsterveldt, C. E. *et al.* The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children. *Twin Res Hum Genet* **16**, 252-267, doi:10.1017/thg.2012.118 (2013).
24 Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* **13**, 231-245, doi:10.1375/twin.13.3.231 (2010).
25 Sirota, M. *et al.* Effect of genome and environment on metabolic and inflammatory profiles. *PLoS One* **10**, e0120898, doi:10.1371/journal.pone.0120898 (2015).
26 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, S361-363, doi:10.1038/ng.2213 [pii] (2012).
27 Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-127, doi:10.1093/ije/dys064 (2013).
28 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:btq340 [pii] 10.1093/bioinformatics/btq340 (2010).
29 Ferreira, M. A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006-1014, doi:10.1016/S0140-6736(11)60874-X (2011).
30 Ferreira, M. A. *et al.* Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *J Allergy Clin Immunol* **133**, 1564-1571, doi:10.1016/j.jaci.2013.10.030 (2014).
31 Williams, H. C. *et al.* The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. I. Derivation of a minimum set of discriminators for atopic dermatitis. *Br J Dermatol* **131**, 383-396 (1994).
32 Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the

analysis of complex genotype-phenotype relationships. *Community genetics* **9**, 55-61, doi:10.1159/000090694 (2006).

33 Wichmann, H. E., Gieger, C., Illig, T. & Group, M. K. S. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67 Suppl 1**, S26-30, doi:10.1055/s-2005-858226 (2005).

34 Marenholz, I. *et al.* The eczema risk variant on chromosome 11q13 (rs7927894) in the population-based ALSPAC cohort: a novel susceptibility factor for asthma and hay fever. *Hum Mol Genet* **20**, 2443-2449, doi:10.1093/hmg/ddr117 (2011).

35 Esparza-Gordillo, J. *et al.* A common variant on chromosome 11q13 is associated with atopic dermatitis. *Nat Genet* **41**, 596-601, doi:ng.347 [pii] 10.1038/ng.347 (2009).

36 J.M., H. & G., R. Diagnostic Features of Atopic Dermatitis. *Acta Derm* **92 (Suppl.)**, 44-47 (1980).

37 Volzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int J Epidemiol* **40**, 294-307, doi:10.1093/ije/dyp394 (2011).

38 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

39 Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101 (2002).

40 Krokstad, S. *et al.* Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol* **42**, 968-977, doi:10.1093/ije/dys095 (2013).

41 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

42 Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nature protocols* **9**, 2643-2662 (2014).

43 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:nature11247 [pii] 10.1038/nature11247 (2012).

44 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

45 Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* **46**, 409-415, doi:10.1038/ng.2924 (2014).

46 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104, doi:10.1126/science.1153717 (2008).

47 Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *bioRxiv*, doi:http://dx.doi.org/10.1101/052308 (2016).

48 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet*, doi:10.1038/ng.3656 (2016).

49 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

50 Abraham, K. J. & Diaz, C. Identifying large sets of unrelated individuals and unrelated markers. *Source Code Biol Med* **9**, 6, doi:10.1186/1751-0473-9-6 (2014).

51 Vilhais-Neto, G. C. *et al.* Rere controls retinoic acid signalling and somite bilateral

symmetry. *Nature* **463**, 953-957, doi:10.1038/nature08763 (2010).
52 Chen, X. *et al.* Retinoids accelerate B lineage lymphoid differentiation. *J Immunol* **180**, 138-145 (2008).
53 Ueki, S. *et al.* Retinoic acids are potent inhibitors of spontaneous human eosinophil apoptosis. *J Immunol* **181**, 7689-7698 (2008).
54 Ueki, S. *et al.* Retinoic acids up-regulate functional eosinophil-driving receptor CCR3. *Allergy* **68**, 953-956, doi:10.1111/all.12175 (2013).
55 Ruvolo, P. P. The broken "Off" switch in cancer signaling: PP2A as a regulator of tumorigenesis, drug resistance, and immune surveillance. *BBA clinical* **6**, 87-99, doi:10.1016/j.bbacli.2016.08.002 (2016).
56 Woetmann, A. *et al.* Protein phosphatase 2A (PP2A) regulates interleukin-4-mediated STAT6 signaling. *J Biol Chem* **278**, 2787-2791, doi:10.1074/jbc.M210196200 (2003).
57 Apostolidis, S. A. *et al.* Phosphatase PP2A is requisite for the function of regulatory T cells. *Nature immunology* **17**, 556-564, doi:10.1038/ni.3390 (2016).
58 Long, L. *et al.* Recruitment of phosphatase PP2A by RACK1 adaptor protein deactivates transcription factor IRF3 and limits type I interferon signaling. *Immunity* **40**, 515-529, doi:10.1016/j.immuni.2014.01.015 (2014).
59 King, P. D., Lubeck, B. A. & Lapinski, P. E. Nonredundant functions for Ras GTPase-activating proteins in tissue homeostasis. *Science signaling* **6**, re1, doi:10.1126/scisignal.2003669 (2013).
60 Zhang, J., Guo, J., Dzhagalov, I. & He, Y. W. An essential function for the calcium-promoted Ras inactivator in Fcgamma receptor-mediated phagocytosis. *Nature immunology* **6**, 911-919, doi:10.1038/ni1232 (2005).
61 Blanc, L. *et al.* Critical function for the Ras-GTPase activating protein RASA3 in vertebrate erythropoiesis and megakaryopoiesis. *Proc Natl Acad Sci U S A* **109**, 12099-12104, doi:10.1073/pnas.1204948109 (2012).
62 Clark, K. *et al.* Phosphorylation of CRTC3 by the salt-inducible kinases controls the interconversion of classically activated and regulatory macrophages. *Proc Natl Acad Sci U S A* **109**, 16986-16991, doi:10.1073/pnas.1215450109 (2012).
63 Sakamaki, J. *et al.* Role of the SIK2-p35-PJA2 complex in pancreatic beta-cell functional compensation. *Nature cell biology* **16**, 234-244, doi:10.1038/ncb2919 (2014).
64 Liu, L. *et al.* A whole genome screen for HIV restriction factors. *Retrovirology* **8**, 94, doi:10.1186/1742-4690-8-94 (2011).
65 Raposo, R. A. *et al.* Effects of cellular activation on anti-HIV-1 restriction factor expression profile in primary cells. *J Virol* **87**, 11924-11929, doi:10.1128/JVI.02128-13 (2013).
66 Parnas, O. *et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686, doi:10.1016/j.cell.2015.06.059 (2015).
67 Marazzi, I. *et al.* Suppression of the antiviral response by an influenza histone mimic. *Nature* **483**, 428-433, doi:10.1038/nature10892 (2012).
68 Chi, T. H. *et al.* Reciprocal regulation of CD4/CD8 expression by SWI/SNF-like BAF complexes. *Nature* **418**, 195-199, doi:10.1038/nature00876 (2002).
69 Kunoh, T. *et al.* A novel human dynactin-associated protein, dynAP, promotes activation of

Akt, and ergosterol-related compounds induce dynAP-dependent apoptosis of human cancer cells. *Molecular cancer therapeutics* **9**, 2934-2942, doi:10.1158/1535-7163.MCT-10-0730 (2010).

70 Cantrell, D. Protein kinase B (Akt) regulation and function in T lymphocytes. *Seminars in immunology* **14**, 19-26, doi:10.1006/smim.2001.0338 (2002).

71 Zhao, H. *et al.* Correlation of structure and function in the human hotdog-fold enzyme hTHEM4. *Biochemistry* **51**, 6490-6492, doi:10.1021/bi300968n (2012).

72 Wang, Q. *et al.* Vitamin D inhibits COX-2 expression and inflammatory response by targeting thioesterase superfamily member 4. *J Biol Chem* **289**, 11681-11694, doi:10.1074/jbc.M113.517581 (2014).

73 Seoh, M. L., Ng, C. H., Yong, J., Lim, L. & Leung, T. ArhGAP15, a novel human RacGAP protein with GTPase binding property. *FEBS Lett* **539**, 131-137 (2003).

74 Arbibe, L. *et al.* Toll-like receptor 2-mediated NF-kappa B activation requires a Rac1-dependent pathway. *Nature immunology* **1**, 533-540, doi:10.1038/82797 (2000).

75 Juncadella, I. J. *et al.* Apoptotic cell clearance by bronchial epithelial cells critically influences airway inflammation. *Nature* **493**, 547-551, doi:10.1038/nature11714 (2013).

76 Pedersen, E. *et al.* RAC1 in keratinocytes regulates crosstalk to immune cells by Arp2/3-dependent control of STAT1. *J Cell Sci* **125**, 5379-5390, doi:10.1242/jcs.107011 (2012).

77 Cui, Y. *et al.* SENP7 Potentiates cGAS Activation by Relieving SUMO-Mediated Inhibition of Cytosolic DNA Sensing. *PLoS pathogens* **13**, e1006156, doi:10.1371/journal.ppat.1006156 (2017).

78 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

79 Forno, E. *et al.* Genome-wide association study of the age of onset of childhood asthma. *J Allergy Clin Immunol* **130**, 83-90 e84, doi:10.1016/j.jaci.2012.03.020 (2012).

80 Yucesoy, B. *et al.* Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma. *Toxicol Sci* **146**, 192-201, doi:10.1093/toxsci/kfv084 (2015).

81 Baurecht, H. *et al.* Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. *Am J Hum Genet* **96**, 104-120, doi:10.1016/j.ajhg.2014.12.004 (2015).

82 Ding, L. *et al.* Rank-based genome-wide analysis reveals the association of ryanodine receptor-2 gene variants with childhood asthma among human populations. *Hum Genomics* **7**, 16, doi:10.1186/1479-7364-7-16 (2013).