Protein-coding variants contribute to the risk of atopic dermatitis and skin-specific gene expression

Sören Mucha[1,*], Hansjörg Baurecht[2,3*], Natalija Novak[4], Elke Rodríguez[2], Saptarshi Bej[5], Gabriele Mayr[1], Hila Emmert[2], Dora Stölzl[2], Sascha Gerdes[2], Frauke Degenhardt[1], Matthias Hübenthal[1,2], Eva Ellinghaus[1], Eun Suk Jung[1,6], Jan Christian Kässens[1], Lars Wienbrandt[1], Wolfgang Lieb[7], Martina Müller-Nurasyid[8], Melanie Hotze[2], Nick Dand[9], Sarah Grosche[10,11], Ingo Marenholz[10,11], Andreas Arnold[12], Georg Homuth[13], Carsten O. Schmidt[14], Ulrike Wehkamp[2], Markus M. Nöthen[15], Per Hoffmann[15], Lavinia Paternoster[16], Marie Standl[17], Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, Klaus Bønnelykke[18], Tarunveer S. Ahluwalia[18,19], Hans Bisgaard[18], Annette Peters[17], Christian Gieger[20], Melanie Waldenberger[20], Holger Schulz[17], Konstantin Strauch[8], Thomas Werfel[21], Young-Ae Lee[10,11], Markus Wolfien[5], Philip Rosenstiel[1], Olaf Wolkenhauer[5], Stefan Schreiber[1,22], Andre Franke[1,†], Stephan Weidinger[2,†], David Ellinghaus[1,23,24,†]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany.

[2]Department of Dermatology, Venereology and Allergy, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.

[3]Institute for Epidemiology and Preventive Medicine, University Hospital Regensburg, Regensburg, Germany

[4]Department of Dermatology and Allergy, University Hospital Bonn, Bonn, Germany.

[5]Department of Systems Biology and Bioinformatics, University of Rostock, Germany.

[6]Department of Internal Medicine and Institute of Gastroenterology, Yonsei University College of Medicine, Seoul, Republic of Korea.

[7]Institute of Epidemiology and Biobank PopGen, Christian-Albrechts-University of Kiel, Kiel, Germany.

[8]Institute of Genetic Epidemiology, Helmholtz Zentrum Munich, Neuherberg, Germany.

[9]School of Basic & Medical Biosciences, Faculty of Life Sciences & Medicine, King's College London, London, UK.

[10]Pediatric Allergology, Experimental and Clinical Research Center, Charité Universitätsmedizin Berlin, Berlin, Germany.

[11]Max-Delbrück-Centrum (MDC) for Molecular Medicine, Berlin, Germany.

[12]Clinic and Polyclinic of Dermatology, University Medicine Greifswald, Greifswald, Germany.

[13]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine and Ernst-Moritz-Arndt-University Greifswald, Greifswald, Germany.

[14]Institute for Community Medicine, Study of Health in Pomerania/KEF, University Medicine Greifswald, Greifswald, Germany.

[15]Institute of Human Genetics, University of Bonn, Bonn, Germany.

[16]Medical Research Council (MRC) Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK.

School of Social and Community Medicine, University of Bristol, Bristol, UK.

[17]Institute of Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany.

[18]Copenhagen Prospective Studies on Asthma in Childhood (COPSAC), Herlev and Gentofte Hospital, Denmark.

[19]Steno Diabetes Center Copenhagen, Gentofte 2820, Denmark.

[20]Research Unit of Molecular Epidemiology and Institute of Epidemiology II, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany.

[21]Departement of Dermatology and Allergy, Division of Immunodermatology and Allergy Research, Hannover Medical School, Hannover, Germany.

[22]First Medical Department, University Hospital Schleswig-Holstein, Kiel, Germany.

[23]Focused Research Unit for Molecular Diagnostic and Clinical Research, Hospital of Southern Jutland, Aabenraa, Denmark.

[24]Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark.

Corresponding author: David Ellinghaus (d.ellinghaus@ikmb.uni-kiel.de), +49 (0) 431 / 500 - 15 131

*These authors contributed equally to this work.

[†]These authors jointly supervised this work.

Abstract

Atopic dermatitis (AD) is the most common inflammatory skin disorder, with 14.91% of risk heritability explained by 31 common genetic variants identified by genome-wide association studies (GWAS). Here, we evaluated common, low-frequency and rare protein-coding variants in 15,574 AD cases and 377,839 controls. 12.56% (s.e. 0.74%) of AD heritability is further explained by rare variation in gene coding sequence. Multi-tissue gene expression profile analysis showed that AD-associated protein-coding variants exert their greatest effect in skin tissues. Protein domain analysis suggests that AD-associated missense variants jointly affect tyrosine phosphorylation and protein interaction sites in DOK2 and CD200R1 that are important for down-regulation of immune receptor signalling. Multiomics-based network analysis combined with whole transcriptome data on lesional, non-lesional and healthy skin revealed *DOK2* as a central hub interacting, among others, with *CD200R1*, *IL6R* and *STAT3*. Our discoveries highlight a potential role of rare coding variants in AD acting independently of common variants.

Introduction

Atopic dermatitis (AD; MIM: 603165) is the most common chronic inflammatory skin disorder affecting 15–20 % of children and 5–10 % of adults (~280 million people worldwide), and the leading cause of the non-fatal disease burden conferred by skin conditions[1]. Given its high genetic heritability (90 % in Europeans), finding causal genes is a crucial step for developing effective preventive and therapeutic approaches for AD. GWAS so far have identified 31 specific genomic regions associated with AD susceptibility[2-8]. The reported susceptibility variants are common (n = 31 with MAF ≥ 5 %) and mostly located in non-coding DNA regions of the genome, have rather small effect sizes (odds ratio (OR) < 1.15) and a largely unclear functional significance[8]. Notable exceptions are low-frequency null mutations in the gene encoding the epidermal structural protein filaggrin (*FLG*), which lead to a reduction in biologically active filaggrin peptides, and a complex perturbation of skin barrier function[9], as well as common missense variants in the genes encoding the T helper 2 (T_h2) signature cytokine IL13 (*IL13*; rs20541)[10] as well as the IL6 receptor (*IL6R;* rs2228145)[11]. Recently, an exome chip based association study of low-frequency variation across all autosomal exons in multiple sclerosis (MS) cohorts of European ancestry led to the detection of low-frequency MS-associated coding variants for four genes that were missed by previous large-scale MS consortium GWAS[12] and that explain another 5% of MS heritability, thus reopening the debate on the contribution of low-frequency and rare variants to disease risk in complex diseases.

To systematically evaluate the contribution of genetic variation to the genetic architecture of AD on the exome-wide scale, particularly protein-altering variants of low or rare-frequency, we profiled 1,913 AD patients and 14,295 controls in two German cohorts using the Illumina HumanExomeBeadchip (exome chip) (see also ***Supplementary Table 1***). The exome chip captures approximately 88 % of low-frequency and rare-coding variants (non-synonymous, splice-site and stop altering, MAFs between 0.01 % and 5 %) present in

Europeans[13]. Suggestive significant novel associations ($p_{exomechip} < 1\times10^{-5}$) were taken forward to replication genotyping in a third German cohort of 1,789 AD cases and 3,272 controls, and a Danish exome chip case-control study of 292 severe AD cases and 650 controls. GWAS association statistics of the EAGLE eczema consortium of 2,298 independent AD cases and 7,802 controls were further used for replication analysis (see also ***Supplementary Table 1***) as well as association summary statistics of 361,132 individuals from UK Biobank with self-reported information (see also ***Supplementary Table 1***). Gene-based tests and Bayesian fine-mapping analysis as well as whole transcriptome RNA-seq, immunohistochemistry and variant protein analyses were conducted to elucidate potential functional consequences of coding variation associated with AD. Finally, we performed multiomics-based network, pathway gene set and gene expression tissue profile analyses, and we quantified the overall contribution of exome chip variation to AD risk.

Results

[Exome chip single-variant and gene-based association analysis]

In our exome chip discovery search for common, low-frequency and rare variant associations, two German AD discovery cohorts were combined via a meta-analysis resulting in single variant association analysis score statistics of 143,884 genotyped and 1,357,289 single nucleotide variants after imputation (see also ***Supplementary Figure 1-2, Supplementary Table 2,*** *see* ***Methods***). 438 and 1,331 SNPs within 8 and 26 loci were identified with genome-wide significance ($p_{exomechip} < 5\times10^{-8}$) and suggestive association ($p_{exomechip} < 1\times10^{-5}$), respectively (see also ***Supplementary Figure 3, Supplementary Table 3***). To identify novel susceptibility variants outside of established AD GWAS loci (see also ***Supplementary Table 4***) eleven suggestively associated SNPs ($p_{exomechip} < 10^{-5}$ with MAF ≥ 1 %) were selected based on LD clumping method (*see* ***Methods***) and carried forward for replication (***Supplementary Table 5***). Using the genome-wide threshold of 5×10^{-8} for the combined analysis of discovery and replication, we identified a novel low frequency missense variant in exon 3 of the gene *Docking protein 2* (*DOK2* at 8p21.3; rs34215892 (p.P274L); $p_{exomechip} = 9.83\times10^{-7}$) which consistently and robustly replicated in three independent cohorts ($p_{German} = 3.75\times10^{-4}$; $p_{Denmark} = 7.60\times10^{-3}$; $p_{EAGLE} = 4.35\times10^{-2}$; $p_{combined} = 2.15\times10^{-10}$; $OR_{combined} = 0.64$; ***Table 1;*** see also ***Supplementary Table 5***). Further, we detected a novel association between AD and a common missense variant at 3q13.2 (rs9865242; p.E312Q; $p_{combined} = 1.17\times10^{-7}$; $OR_{combined} = 1.16$, see also ***Supplementary Table 5***) located in exon 7 of *CD200 Receptor 1* (*CD200R1*) and 266,51 kb upstream of the previously reported intergenic locus 3q13.2 (*CCDC80*, rs12634229) described only in a Japanese AD cohort[5] so far ($r^2_{rs9865242\text{-}rs12634229} = 0.005$). A look-up of association results from UK Biobank for the self-reported broad allergic disease phenotype "AD (eczema), allergic rhinitis and/or hayfever"[14] (*see* ***Methods***) further confirmed association signals for *DOK2* (rs34215892; $p_{UK\text{-}Biobank} = 3.35\times10^{-6}$) and *CD200R1* (rs9865242; $p_{UK\text{-}Biobank} = 1.35\times10^{-8}$) (***Table 1***). Further, we

performed gene-based association analysis, in which we evaluated the cumulative effects of low-frequency and rare variants for each gene from autosomes (*see* ***Methods***). Only *DOK2* met the exome-wide significance threshold (*see* ***Methods***) and exhibited a stronger association signal than compared to single variant analysis ($p_{combined\text{-}DOK2} = 4.23\times10^{-13}$; ***Table 2;*** see also ***Supplementary Table 6***) comprising twelve protein-altering variants of which two were of low-frequency and LD-independent (rs34215892 and rs56094005; 1 % ≤ MAF < 5 %) and ten were rare (MAF < 1 %) (***Figure 1***). We genotyped the second low-frequency variant rs56094005 (p.L138S; $p_{rs56094005\text{-}exomechip} = 4.31\times10^{-3}$) in our replication set, in addition to rs34215892, and successfully confirmed the single SNP ($p_{rs56094005\text{-}replication} = 5.96\times10^{-3}$; see also ***Supplementary Table 5*)** and the aggregated *DOK2* association signal ($p_{DOK2\text{-}replication} = 1.54\times10^{-6}$; see also ***Supplementary Table 6***).

[Bayesian fine-mapping and functional annotation]

As a next step towards understanding the functional causality of the identified AD-associated variants, we carried out Bayesian fine-mapping and functional annotation analysis for loci *CD200R1* and *DOK2* using imputed exome chip data of our discovery cohorts (*see* ***Methods***). Fine-mapping strengthened our hypothesis that lead variants rs9865242 (*CD200R1*) and rs34215892 (*DOK2*) as being most likely causal (in the context of fine-mapping), with posterior probabilities of 97.2 % and 44.3 %, respectively (see also ***Supplementary Figure 2***). rs34215892 overlaps enhancer histone marks and DNase peaks in 15 and 12 different tissues respectively, in each case including the skin, and is predicted to affect protein binding and regulatory motifs (see also ***Supplementary Table 7).*** The second *DOK2* lead variant rs56094005 (***Figure 1***) overlaps promoter histone marks and DNase peaks in ten tissues, in each case including T and B cells, and is in perfect LD ($r^2 = 1$) with an intronic variant rs118162691 ($p_{rs118162691\text{-}exomechip} = 4.17\times10^{-3}$; MAF = 3.9 %) which overlaps enhancer histone marks in nine different tissues including the skin (see also ***Supplementary***

Table 7). The *CD200R1* missense variant rs9865242 has been suggested as a cis-eQTL for *GTPBP8* in whole blood[15]. Ten variants in perfect LD with rs9865242, further overlapping enhancer histone marks in blood predicted to alter regulatory motifs, are located in a conserved region or are found as eQTLs for *CD200R1*[16] (see also ***Supplementary Table 7***).

[Immunohistochemistry and whole transcriptome mRNA-seq data analysis]

Immunohistochemistry was used to determine the location of DOK2 in skin tissue (*see* ***Methods***). It showed strong epidermal staining with clear differences among AD lesional, AD non-lesional and healthy skin (see also ***Supplementary Figure 5***). DOK2 is predominantly expressed in lymphocytes and the increased abundance of DOK2 in lesional AD skin correlates with the degree of lymphocyte infiltration (see also ***Supplementary Figure 5***). Moreover, we observed a significantly increased *DOK2* and *CD200R1* mRNA expression in whole transcriptome mRNA-seq data (*see* ***Methods***) on lesional as compared to non-lesional skin samples of 27 AD patients ($p_{DOK2} = 4.2\times10^{-5}$, $p_{CD200R1} = 2.2\times10^{-5}$) and as compared to skin from 38 healthy individuals ($p_{DOK2} = 8.8\times10^{-11}$, $p_{CD200R1} = 2.2\times10^{-7}$), as well as in AD non-lesional skin compared to healthy skin ($p_{DOK2} = 4.5\times10^{-3}$, $p_{CD200R1} = 2.0\times10^{-2}$) (see also ***Supplementary Figure 6a-b***). We also observed a slightly increased expression of *DOK1*, whose protein is a heterodimeric partner for DOK2, and *CD200* in lesional skin compared to non-lesional ($p_{DOK1} = 1.1\times10^{-1}$, $p_{CD200} = 2.9\times10^{-1}$) or healthy skin ($p_{DOK1} = 3.1\times10^{-4}$, $p_{CD200} = 2.0\times10^{-1}$), as well as non-lesional skin compared to healthy skin ($p_{DOK1} = 4.0\times10^{-3}$, $p_{CD200} = 7.8\times10^{-1}$) (see also ***Supplementary Figure 6c-d***).

[*In silico* variant protein analysis]

To construct a first hypothetical model of whether *CD200R1* and *DOK2* missense lead variants are likely to interfere with functionally active domains on the protein level, we performed extensive literature search and further conducted protein domain analyses of

DOK2 and the CD200/CD200R1 receptor complex (*see* ***Methods***) (***Figure 2***). Members of the DOK adapter protein family act as regulators of cell stimulatory signals by serving as substrates for diverse receptor and cytoplasmic kinases, and the highly similar and interacting members DOK1 and DOK2 are involved in the down-regulation of immune receptor signalling in $CD4^+$ T cells as well as myeloid cells such as macrophages and neutrophils[17]. It is assumed that the inhibitory role of DOK2 is accomplished by recruiting and activating RasGAP to inhibit Ras and thus to suppress pro-inflammatory ERK, JNK and MAPK pathways for the DOK2 response to CD200R1 in human myeloid cells[18]. By means of protein sequence and structural domain analyses we hypothesize that p.P274L (rs34215892; *DOK2*) is likely to interfere with RasGAP binding following the observation that the variant locates within an invariant RasGAP-SH2 binding motif (YxxPxD) since proline side chains confer strong local main chain rigidity compared to other amino acids. Less structural effect was predicted for p.L138S (*DOK2*) but the variant may interfere with DOK1/2-regulation due to its location within the protein interaction site. p.E312Q (rs9865242; *CD200R1*) causes a loss of negative charge in proximity of the NPLY motif needed for DOK2 recruitment and may therefore affect protein-protein contacts and signalling.[18,19]

[AD core network construction and differential gene expression analysis of core genes]

In order to assess possible interactions of genes *CD200R1* and *DOK2* with candidate genes from AD GWAS loci[2-8], we generated an AD core interaction network using network prioritization algorithms which make use of protein-protein, protein-gene, co-expression, and shared protein domains data from public repositories (see also ***Supplementary Figure 7; Supplementary Table 8***; *see* ***Methods***). We identified *DOK2* as a central hub node interacting with *CD200R1* and functionally established AD susceptibility genes *STAT3*[8], *MICB*[20], *CLEC16A*[7] and *IL6R*[11] (***Figure 3a***). We used our aforementioned whole transcriptome mRNA-seq data (*see* ***Methods***) to assess whether expression levels from genes of our AD

core network are differentially expressed in lesional, non-lesional and healthy skin from AD patients and healthy individuals. 22 out of the 30 AD core genes (***Figure 3a***) are significantly up or downregulated (see also ***Supplementary Table 9; Supplementary Figure 8***), with 17 out of 22 directly interacting with *DOK2* (***Figure 3b***), thus indicating the biological importance of *DOK2* in AD pathogenesis.

[Pathway and gene expression tissue specificity analysis for exome chip variants]

To reveal potential differences in terms of biological pathways and involved tissues for exome chip variant (exome chip) and common variants investigated in AD GWAS meta-analysis studies (GWAS) from the EAGLE consortium, we performed gene-set pathway enrichment analysis for curated gene sets and gene ontology (GO) terms as well as tissue-specific gene expression profile analysis[21] for 53 tissue types and 11,688 samples from GTEx release 7 (*see* ***Methods***). The pathway enrichment analysis identified one significant gene set including five genes ($P_{Bonferroni} < 0.05$; *IL4R*, *IL13*, *JAK1*, *JAK2*, *TYK2*) of the *IL13* pathway for the exome chip in comparison to eight blood-cell related gene sets (including regulation of immunoglobulin production, B cell activation and B cell mediated immunity) for the EAGLE GWAS data. Recently, AD was characterized as an IL13-dominant disease based on high-depth RNA-seq transcriptome data of 147 samples from cohorts of AD patients, psoriasis patients, and healthy controls, with IL13 being the most distinctive marker for AD[22]. We hypothesize that low-frequency and rare coding variants in genes of the *IL13* pathway are further likely to be associated with AD. In the tissue specificity analysis, we tested for relationships between tissue-specific gene expression profiles and variant association statistics from exome chip and GWAS, respectively. For the exome-chip data, we observed a Bonferroni-corrected significant association ($P < 9.43\times10^{-4}$) with two skin tissue types (sun exposed and non-sun exposed skin; ***Figure 4a***). In comparison the EAGLE GWAS data revealed a significant association with tissues whole blood and spleen (***Figure 4b***).

[Estimation of liability-heritability from exome chip variants]

To quantify the overall contribution of exome chip variation to AD risk, we used a restricted maximum-likelihood approach to model heritability attributable to genotypic variation across genotyped-only exome chip variants (*see* ***Methods***). We found that common variants from exome chip (LD pruned, MAF ≥ 5 %; *see* ***Methods***) explain 1.04 % of heritability on the observed scale (corresponding to 2.93 % on the liability scale assuming a prevalence of 0.14). By lowering the MAF threshold to 1%, exome chip heritability increased to 1.27 % (3.60 %). Dividing variants into low-frequency (1 % ≤ MAF < 5 %) and rare (MAF < 1 %), heritability was found to be 0.37 % for low-frequency variants (0.94 %) and 4.47 % for rare variants (12.56 %). Furthermore, we estimated the heritability explained by the newly identified lead variants of *DOK2* (rs3215892, rs56094005) and *CD200R1* (rs9865242). Variants rs3215892, rs56094005 and rs9865242 explain approximately 0.015 % (0.041 %), 0.004 % (0.012 %) and 0.013 % (0.035 %) on the observed scale (liability scale), respectively.

Discussion

In conclusion, we analysed the association of AD with common, low-frequency and rare protein-coding variants and implicate two novel genes (*DOK2* and *CD200R1*) contributing to AD risk. So far, 14.91 % (6.95 % excluding *FLG* mutations) of AD liability-scale heritability could be attributed to common lead variants of 31 GWAS loci, estimated from a set of 5,775 clinically diagnosed AD patients and assuming a population prevalence of 0.15[2,8]. Recently, Ferreira *et al.*[14] further reported a genome-wide liability-scale heritability estimate of 9.04 % from 1.2 million HapMap SNPs (assuming a population prevalence of 0.14) and on the basis of self-reported information from questionnaires. Our newly identified variants (rs3215892, rs56094005, rs9865242) explain another 0.088 % of the variance in liability. Interestingly, approximately 12.56 % (s.e. 0.74 %) of liability-scale heritability is estimated to be explained by rare variants (MAF < 1 %), thus highlighting the importance to study rare coding variation in AD. Our results encourage future studies along the same path and highlight the importance to study the impact of protein-coding variants for phenotypically well-defined clinically diagnosed cohorts.

Pathway enrichment analysis of exome chip data revealed that association signals of low-frequency and rare coding variants are enriched in five genes of the *IL13* pathway, all of which are targets for novel systemic AD therapeutics in advanced stages of clinical investigations[23], supporting the pivotal role of type 2 inflammation in AD pathogenesis[1]. Gene expression tissue profile analysis showed that exome chip variants associated with AD cumulatively have a stronger effect on skin tissue gene expression than common GWAS variants associated with AD, as identified in combination with whole transcriptome RNA-sequencing data on lesional, non-lesional and healthy skin tissue as well as tissue samples from GTEx. In accordance with results from the exome chip study for multiple sclerosis[12], we observed that the minor allele of low-frequency lead missense variants in *DOK2* is mostly protective (in context of the odds ratio; ***Table 2***).

The newly discovered genes *DOK2* and *CD200R1* have clear immunological functions, confirming that AD pathogenesis is primarily driven by immune dysregulation. Structural protein domain analysis, topological network and differential gene expression analyses suggest that missense variants in *DOK2* (rs34215892; rs56094005) in combination with the missense AD risk variant in *CD200R1* (rs9865242) together may affect tyrosine phosphorylation sites in DOK2 and CD200R1 (***Figure 2***). *DOK2* belongs to the DOK gene family encoding for seven different DOK proteins (DOK1–7) being involved in signal transduction[24-27]. DOK1–2 are adaptor proteins and mainly expressed in hematopoietic/immune cells[28] and have been implicated to negatively regulate proliferation and constitutive expression of DOK2 [29]. Further studies showed that both DOK1–2 are essential negative regulators of ERK signalling downstream of Toll-like receptor 4[30]. Mice lacking DOK1–3 showed significant defects in immune cell development and in immune responses [31]. Furthermore, DOK1–2 play a role in the maturation of NK cells [32], which have been shown to be reduced in AD. In line with these observations, we observed a significantly increased *DOK2* and *CD200R1* mRNA expression in lesional as compared to non-lesional skin samples of AD patients, and as compared to skin from healthy individuals. Our AD core network (***Figure 3***) further showed that *DOK2* acts as a central hub gene interacting with *CD200R1* as well as several candidate AD GWAS susceptibility genes on the cellular level. Our AD core network revealed 21 genes that are directly functionally related to *DOK2* of which 16 are significantly upregulated (including *DOK1*) and one (*RASGAP*) is significantly downregulated in lesional skin samples of AD patients. Taken into account the reported inflammatory role of both DOK2 and CD200R1, we conclude that signalling through CD200/CD200R1/DOK2 could be an important new regulatory signalling pathway in AD. Extensive functional studies are required to detect all potential causal variants and thus to specify the contribution of the *DOK2* and *CD200R1* to overall disease susceptibility.

Methods

Study samples and genotyping

All cases had been diagnosed with AD by a dermatologist except for UK Biobank cases. All participants provided written informed consent and the study was approved by the ethics boards of the participating institutions, in agreement with the Declaration of Helsinki principles.

German discovery cohort 1. German AD patients ($n = 1{,}056$) were recruited at the Department of Dermatology at Christian-Albrechts-University Kiel, Department of Dermatology and Allergy at the Technical University of Munich, and the Department of Dermatology and Allergy at the University of Bonn. Data from healthy control individuals ($n = 7{,}026$) were obtained from the PopGen biobank[33], the KORA S4 survey (an independent population-based sample from the general population living in the region of Augsburg, southern Germany) [34] and the Heinz-Nixdorf Recall (HNR) cohort[35], Bonn. AD cases as well as controls were genotyped using Illumina HumanExome-12 v1.0 BeadChips (see also ***Supplementary Table 1***).

German discovery cohort 2. German AD patients ($n = 1{,}051$) were recruited from dermatology clinics in Kiel or Hannover (the University of Kiel and Medizinische Hochschule of Hannover). The AD cases were genotyped using Illumina HumanCoreExome-24 v1.0 A or HumanCoreExome-24 v1.1 A BeadChips. Data from healthy control individuals ($n = 8{,}135$) were obtained from the SHIP and SHIP-TREND cohorts (from the Study of Health in Pomerania, a prospective longitudinal population-based cohort study in West Pomerania)[36]. All German controls were genotyped using Illumina HumanExome-12 v1.0 BeadChips (see also ***Supplementary Table 1***).

G*erman replication cohort*. German AD patients ($n = 1{,}789$) were recruited from dermatology clinics in Kiel or Berlin (University Children's Hospital, Charité Universitätsmedizin Berlin, as part of the Genetic Studies in Nuclear Families with AD

(GENUFAD) study). Data from healthy control individuals (n = 3,272) were obtained from University Hospital in Kiel and Lübeck at the Institute of Transfusional Medicine (see also ***Supplementary Table 1***).

Danish replication cohort. All Danish AD cases (n = 292) are hospitalized severe cases from the COPSAC eczema REGISTRY. Healthy control individuals (n = 650) were obtained from the COPSAC2000 and COPSAC2010 birth cohorts[37] in Copenhagen, Denmark. Both, cases and controls were genotyped on the Illumina Infinium OmniExpressExome-8 v1.4 BeadChip (see also ***Supplementary Table 1***).

EAGLE GWAS replication cohorts. We used imputed summary statistics of the EAGLE Eczema Consortium for the discovery cohorts (excluding 23andMe). (https://data.bris.ac.uk/data/dataset/28uchsdpmub118uex26ylacqm) comprising 11,294,660 SNP markers with MAF ≥ 1 % and 10,788 AD cases and 30,047 (see also ***Supplementary Table 1***).[8] For replication analysis, we used only European studies independent from our German discovery cohorts and where AD diagnosis was ascertained by a dermatologist. For the two *DOK2* variants (rs34215892; rs56094005) association statistics were available for 1,881 cases and 6,154 controls in total. For the *CD200R1* variant association statistics for 2,298 cases and 7,802 controls in total were available.

UK Biobank. Since only 33 (primary diagnosis; field 41203) and 50 patients (secondary diagnosis; field 41204) have been diagnosed with "atopic dermatitis" (ICD10 code L30) in UK Biobank, we used questionnaire information from UK Biobank (release of March 2018). Key words "atopic dermatitis" and/or "eczema" are contained in self-reported data-fields" Non-cancer illness code; self-reported: eczema/dermatitis" (data-field 20002) and "Age hay fever, rhinitis or eczema diagnosed" (data-field 3761) and three others (data-fields 6152, 41202, 41203). For the three lead variants (*DOK2*: rs34215892; rs56094005; *CD200R1*: rs9865242), we downloaded imputed summary association statistics (http://www.nealelab.is/uk-biobank/; http://www.nealelab.is/uk-biobank/faq; release March

2018; see also ***Supplementary Table 1***) for (a) 83,407 cases with eczema, allergic rhinitis or/and hayfever (data-field 20002) versus 277,120 controls and (b) 9,312 cases with eczema or dermatitis (data-field 3761) versus 351,820 controls. Summary associations statistics had been generated using a least-squares linear model predicting the phenotype with an additive genotype coding (0, 1, or 2 copies of the minor allele), with sex and the first ten principal components from the UK Biobank sample QC file as covariates.

An overview of the study design is shown in ***Supplementary Figure 1***, and detailed characteristics of the discovery and replication case-control cohorts are provided in ***Supplementary Tables 1***.

Genotype calling, quality control and technical validation

DNA samples of discovery cohorts were genotyped on Illumina HumanExome-12 v1.0, Illumina HumanCoreExome-24 v1.0 A or HumanCoreExome-24 v1.1 A BeadChips. Genotyping calling was performed separately for the three different chips using GenTrain version 2.0 in GenomeStudio Data Analysis software. Sample and marker quality control (QC) was performed with PLINK (v1.9; https://www.cog-genomics.org/plink/1.9)[38]. Sample exclusion criteria were based on call rate (< 98 %), average heterozygosity (±5 s.d. from the mean) and non-European ancestry defined as population outliers from principal component analysis (PCA) using EIGENSTRAT[39] (see also ***Supplementary Figure 9)***. Duplicated and related samples (PI_HAT > 0.8 and PI_HAT > 0.1875), respectively, were removed. SNP markers were excluded due to low call rate (< 95 %), deviations from Hardy-Weinberg equilibrium ($p < 10^{-5}$) and differential missingness between cases and controls ($p_{fisher} < 10^{-50}$). insertions/deletions, duplicated and tri-allelic variants were further removed. After QC, both discovery sets were merged. The resulting data set comprised 143,884 overlapping single nucleotide variants (SNVs) and a total number of 1,913 cases and 14,295 controls (see also ***Supplementary Figure 1, Supplementary Figure 10***). EVOKER (v2.3;

https://sourceforge.net/projects/evoker/)[40] was used to manually inspect intensity cluster plots (see also ***Supplementary Figure 11***) for 6,697 SNVs according to the following criteria: SNVs occurring only in cases, only in controls or occurring more frequent (factor > 10) in cases than controls. If acceptable the genotypes were manually corrected and reintegrated into QCed data sets.

Replication genotyping for the German replication cohort was carried out with TaqMan technology from Applied Biosystems. QC was performed with PLINK[38]. Marker with a call rate of < 95 % or marker deviating from Hardy-Weinberg equilibrium ($p < 10^{-5}$) were excluded.

Technical validation of low-frequency lead variant rs34215892: 79 German AD cases with heterozygotic or homozygotic genotypes for the rare allele of *DOK2* (rs34215892; $MAF_{cases} = 2.5$ %; $n = 79$) of German data sets 1 and 2 were genotyped via TaqMan for technical validation. The validation confirmed genotypes derived from the exome chip.

Genotype imputation and dosage-based association analysis

We used the genotype imputation service provided by the Wellcome Trust Sanger Institute (https://imputation.sanger.ac.uk). The Haplotype Reference Consortium (HRC) created a large reference panel (HRC.r1)[41] consisting of 64,940 haplotypes with 39,635,008 SNPs of predominantly European origin. EAGLE (v2.0.5; https://data.broadinstitute.org/alkesgroup/Eagle)[42] was applied for phasing against the reference panel followed by the imputation using PBWT (Positional Burrows-Wheeler Transform)[43]. For *German discovery cohort 1* 3,942,401 SNVs and for *German discovery cohort 2* 3,746,633 SNVs were imputed. Post-imputation, SNPs with an info score of less than 0.5, as well as tri-allelic variants were removed from each data set. Then for both data sets a fixed-effects meta-analysis was conducted.

Single marker association testing

RAREMETALWORKER was used with the first ten principal components from PCA as covariates to analyse individual imputed studies and to generate association summary statistics that can later be combined across studies. Results were combined by a meta-analysis using RAREMETAL[44]. For the gene-based association testing we used the same approach as for the genotyped variants. EPACTS (v3.2.3; http://genome.sph.umich.edu/wiki/EPACTS) was used to annotate the imputed variants followed by a gene-based analysis for all protein-altering variants.

Due to insufficient statistical power to perform single marker tests for rare variants (power = 15 % to detect an odds ratio of 2 with MAF = 1% at a Bonferroni-corrected significance threshold of $0.05/110{,}903=4.5\text{x}10^{-7}$), single variant association tests were carried out for common and low-frequency variants (MAF ≥ 1%). In the single marker association test, we examined genotyped and imputed variants with MAF ≥ 1 %, Hardy-Weinberg equilibrium ($p < 10^{-6}$) and with consistent direction of effects in across both studies. Following QC, a total number of 1,246,386 SNPs was available for association testing across 1,913 AD cases and 14,295 population controls (see also ***Supplementary Figure 10***). To check for residual population stratification before and after imputation, we generated a set of 'null' SNPs ($n_{genotyped} = 14{,}622$; $n_{imputed} = 39{,}654$) based on the intersection of SNPs available for both data sets after filtering for LD ($r^2 > 0.5$) and allele frequency (MAF > 5 %). For the genotyped and imputed data sets 14,328 and 38,759 'null' variants remained after exclusion of SNPs from the MHC region (chr6:25-34 Mb) as well as from established AD loci (n = 31), respectively. For *German discovery cohort 1* we obtained genomic inflation factor $\lambda_{imputed} = 1.06$ and $\lambda_{1000imputed} = 1.04$ ($\lambda_{genotyped} = 1.07$ and $\lambda_{1000genotyped} = 1.04$), for *German discovery cohort 2* $\lambda_{imputed} = 1.03$ and $\lambda_{1000imputed} = 1.02$ ($\lambda_{genotyped} = 1.05$ and $\lambda_{1000genotyped} = 1.03$), respectively, indicating minimal evidence of residual population stratification *(*see also ***Supplementary Figure 2).***

For each cohort we used RAREMETALWORKER[44] to compute score test statistics as well as the corresponding variance-covariance matrix with the first ten principal components as covariates. The following meta-analysis was performed with RAREMETAL. Odds ratios (ORs) were calculated with PLINK (v1.9)[38] using a logistic regression model (--dosage option) with the first ten principal components from PCA as covariates for each single data set. For exome chip genotyped variants, genome-wide significance threshold of $p < 5\times10^{-8}$ was applied and only variants with consistent direction of effects across both studies were considered to be associated with AD. For variants with $p < 1\times10^{-5}$, genotype cluster plots were manually inspected and assessed with regards to shape and consistency across different chip types. If acceptable, genotypes were corrected using Evoker, reintegrated into the data set and analysis was repeated.

Gene-based association testing

The exome chip genotype data of the full allele spectrum (see also ***Supplementary Figure 10;*** excluding imputed variants) was converted into vcf format with PLINK [38] and variants were functionally annotated with EPACTS for 15,998 genes. The gene group file included protein2altering (non-synonymous, stop-gain and essential splice site; n = 120,086; see also ***Supplementary Table 2***) variants, which was used for the gene-based analysis with RAREMETAL[44]. Score statistics and covariance information were obtained with RAREMETALWORKER[44] as described above. We applied the variance component test SNP-Set Kernel Association Test (SKAT)[45] to model the possibility that minor alleles can be deleterious or protective. Bonferroni correction was applied for the number of genes, resulting in a significance threshold of $\alpha = 0.05/15{,}998 = 3\times10^{-6}$.

SNP selection for replication analysis

For replication genotyping, we selected the most strongly associated SNP (n = 11) with $p < 10^{-5}$ from each associated locus by means of PLINK's clumping procedure (using default settings: $p_1 < 10^{-5}$, $p_2 < 0.05$, $r^2 \geq 0.5$, kb = 250) representing 11 loci (see also ***Supplementary Table 5***), which do not overlap genomic boundaries of the 32 previously reported AD GWAS loci[2-8] (see also ***Supplementary Table 4***).

Bayesian fine-mapping analysis

A Bayesian fine-mapping analysis was carried out using FINEMAP[46] in order to determine a credible set, a minimum set of SNVs containing all causal variants, and to calculate the posterior inclusion probability (PIP) for each SNP as causal in any of the models [47] (see also ***Supplementary Figure 4***). To this end, we extracted regions of interest from the exome chip and calculated the local LD structure in the discovery cohort by LDstore[48], which served together with the exome chip summary statistics as an input for FINEMAP. We set the option --n-causal-snps 1. Analyses carried out with k = 2, 3, 5 (default) and 10 expected causal variants gave similar results.

Variant effect prediction

We predicted the protein-altering effects of variants using SNPNEXUS (http://www.snp-nexus.org/). This software includes two tools: SIFT[49] and PolyPhen[50]. Variants with SIFT score below 0.05 and/or PolyPhen-2 estimated false-positive rate below 0.05 were considered to have a functional effect.

Linkage disequilibrium (LD) estimates

For LD calculations we used the LD link software from NIH with the LDpair option (v3.0, https://analysistools.nci.nih.gov/LDlink/)[51,52]. Pair-wise LD calculations are based on

494 the European population (CEU) comprising 503 samples of the 1000 Genomes Project Phase
495 3[48].
496
497 **Functional variant annotation using HaploReg and RegulomeDB**
498 The HaploReg (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php) is a
499 tool to explore annotations of the non-coding genome with the help of haplotype blocks[53].
500 Using the LD information of the haplotype blocks from the 1000 Genomes Project, linked
501 SNPs and small indels can be visualized along with chromatin state and protein binding
502 annotation from the Roadmap Epigenomics and ENCODE projects including the effect of
503 SNPs on regulatory motifs, the effect of SNPs on expression from eQTL (expression
504 quantitative trait loci) studies and sequence conservation. HaploReg v4 consists of a core set
505 of 52,054,804 variants, which are primarily of single-nucleotide polymorphisms (SNPs) using
506 all refSNP IDs, hg19 positions and alleles from dbSNP release b137[53,54]. In order to annotate
507 variants by their effect on regulatory motifs, a library of position weight matrices (PWMs)
508 was used[55]. The affinity in the reference sequence is higher, if $PWM_{alt} - PWM_{ref}$ is negative[53].
509 A further application of the HaploReg tool is to identify evolutionarily conserved regions
510 based on SiPhy (SIte-specificPHYlogenetic) analysis[56].
511 We used HaploReg (version 4.1) to select the linked SNPs ($r^2 \geq 0.5$ in European
512 populations) with the index SNP from the 1000 Genomes Project, investigate characteristic of
513 transcriptional regulation activity based on the Roadmap Epigenomics data, and predict the
514 target gene according to the eQTL studies.
515 RegulomeDB is a database that annotates SNPs with known and predicted regulatory
516 elements in the intergenic regions of the human genome. These elements include regions of
517 DNase hypersensitivity, binding sites of transcription factors, and promoter regions that have
518 been biochemically characterized to regulation transcription. RegulomeDB uses public data
519 sets from GEO, the ENCODE project, and published literature[57].

520

521 **Immunohistochemistry**

522 The staining procedure was adapted from a protocol previously described[58]. Formalin-
523 fixed and paraffin-embedded skin samples from healthy controls as well as from lesional and
524 non-lesional sites of AD patients were investigated. The 4.0 µm tissue sections were
525 deparaffinised in xylene and dehydrated in graded ethanol solutions. Then tissue sections
526 were heated for 20 min at 80 °C in EDTA/Tris buffer, pH 9.0, for antigen retrieval. To block
527 endogenous peroxidase activity the sections were incubated with 30 ml/l H_2O_2 for 20 min.
528 After incubation with a mouse anti-Dok2 monoclonal antibody (dilution 1:200, Santa Cruz
529 Biotechnology, Inc., Dallas, Texas, US) at 4 °C overnight, staining was performed using the
530 labelled streptavidin-biotin method. For the negative tissue section, the primary antibody was
531 omitted. Examination and judgment of the staining was carried out semi-quantitatively and
532 rated by the degree of infiltration and abundance by an expert dermatopathologist.

533

534 **Differential gene expression analysis**

535 We performed differential gene expression analysis using a recently published RNA-
536 seq study of skin samples from 27 AD patients and 38 healthy controls[22]. Sample preparation
537 and sequencing was described in the original publication[22]. In brief, total RNA was obtained
538 from 5mm skin punch biopsies and sequenced on the Illumina HiSeq2500 with 2x125 bp after
539 preparation using the Illumina Truseq® Stranded total RNA Protocol. After quality control
540 and mapping to the human reference genome (b37) RNA read counts were TMM-normalized
541 (Trimmed mean of M-values normalization) and log2-transformed for analysis. Genes of
542 interest selected by the network analysis were investigated in lesional (L) and non-lesional
543 (NL) skin of AD patients and compared with expression values from skin samples of healthy
544 controls using the Mann-Whitney-U test on a nominal significance threshold.

545

546 ***In silico* structural protein domain analysis**

547 Protein sequence annotations and domain borders indicated in ***Figure 2*** were derived
548 from UniProt (DOK2_HUMAN/O60496, MO2R1_HUMAN/Q8TD46). The 3D structure of
549 CD200/CD200R1 receptor complex was obtained from the PDB (PDB ID 4bfi) and
550 visualized using PyMOL[59,60].

551

552 **Functional AD core disease network**

553 We developed a functional AD core network depicting the most important functional
554 interactions prioritized from candidate genes of 32 known AD GWAS risk loci as well as
555 genes *DOK2* and *CD200R1* identified in this study. In a first step, we used we used
556 Cytoscape[61] in combination with BisoGenet[62] for network construction and searched for
557 topologically important nodes, commonly referred to as hubs. As an input we used 42 genes
558 including the 40 candidate genes from 32 known AD GWAS loci listed in Paternoster *et al.*[8]
559 as well as genes *DOK2* and *CD200R1*. Using BisoGenet, we retrieved information from
560 public databases including DIP[63], BIOGRID[64], HPRD[65], IntAct[66], MINT[67], BIND[68]. The
561 databases consider protein-protein or protein-gene interactions based on experimental
562 evidences. The constructed network had 228 nodes and 562 interactions (see also
563 ***Supplementary Figure 7***) and included genes and proteins that have interactions with at least
564 two of the 42 input genes (or encoded proteins). We then ranked the 228 genes based on a
565 prioritization function (***Equation 1***) of topological properties including the node degree (D),
566 betweenness centrality (BC), closeness centrality (CC) and clustering coefficient (ClC) of a
567 protein or gene (see also ***Supplementary Table 8, Supplementary Figure 7***).

568

569 **Equation 1: Equation for ranking highly connected and most important proteins within**
570 **the constructed AD network:** V(N) denotes the set of nodes in the network N. Here $D_N(p)$,

$BC_N(p)$, $CC_N(p)$ and $ClC_N(p)$ denote network properties Degree, Betweenness Centrality, Closeness Centrality and Clustering coefficient, respectively, of the protein or gene identifier (denoted as p) in N. To give equal importance to each property, for each property, the property value was normalized according to the maximum value (***Equation 1***):

$$\theta(p) = \frac{D_N(p)}{max\{D_{\{v \in V(N)\}}(v)\}} + \frac{BC_N(p)}{max\{BC_{\{v \in V(N)\}}(v)\}} + \frac{CC_N(p)}{max\{CC_{\{v \in V(N)\}}(v)\}} + \frac{ClC_N(p)}{max\{ClC_{\{v \in V(N)\}}(v)\}}$$

In a next step, we constructed an AD core network based on functional features including co-expression, co-localization, genetic interactions, pathway information, physical interactions and shared protein domains using the GeneMANIA[69] App for Cytoscape, and we used the top ten hub nodes from the protein-protein and protein-gene network as input. In brief, GeneMANIA finds functionally similar genes using a wealth of human genomics, transcriptomics and proteomics data by highly weighting protein domain similarity networks and suggesting additional genes with a similar domain structure. GeneMANIA uses hundreds of data sets and hundreds of millions of interactions that have been collected from: GEO, BioGRID, EMBL-EBI, Pfam, Ensembl, NCBI, MGI, I2D, InParanoid and Pathway Commons. Our final AD core disease network (***Figure 3a***) included 30 nodes and was visualized with GeneMANIA

FUMA – Functional Mapping and Annotation of Genome-Wide Association Studies

We used FUMA's[70] SNP2GENE function to compute gene-based p-values (gene analysis) and gene set p-values (gene set analysis) from association summary statistics of the exome chip data and the GWAS EAGLE consortium data, respectively. In FUMA version 1.3.3c, 10,655 gene sets were provided, including 4,738 curated gene sets and 5,917 GO terms from MsigDB v6.1. MAGMA v1.6[21] gene-set (pathway enrichment) analyses used the

full distribution of SNP p-values and Bonferroni correction was applied for tested gene sets. The 1000G phase 3 reference panel[71] was used to calculate LD across SNPs and genes.

FUMA performs MAGMA gene-property (tissue specificity) analyses to test relationships between tissue-specific gene expression profiles and disease-gene associations. The gene-property analysis is based on a multiple regression model[21] where gene-based p-values (excluding genes from the extended MHC region on chr6:25-34 Mb) are converted to gene Z-scores. A one-sided test is performed to test the positive relationship between tissue specificity and genetic association of genes. Normalized gene expression values of 53 tissue types were obtained from GTEx version 7 (http://www.gtexportal.org/home/datasets/GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_rpm.gct.gz), including 11,688 samples and 56,203 genes in total. This set of genes was filtered that the average TPM per tissue was >1 in at least one of the 53 tissues. TPM was winsorized at 50 (replaced TPM > 50 with 50). Then average of log transformed TPM with pseudocount 1 ($\log_2$(TPM+1)) per tissue (for either 53 detail or 30 general tissues) was used as the covariates conditioning on the average across all the tissues. This resulted in 32,335 genes. MAGMA gene-property test was performed by default settings for both exome chip and EAGLE consortium GWAS summary association results.

Estimation of SNP-based heritability

We used the genome-wide complex trait analysis (GCTA) software to calculate the heritability based on a restricted maximum-likelihood (REML) approach to genotypic variation across the exome chip[72]. We estimated the heritability for two LD pruned SNP sets including common SNPs with MAF ≥ 5 % (n = 14,622) and common and low-frequency SNPs with MAF ≥ 1 % (n = 20,918). In addition, we calculated the heritability for sets MAF < 5 % (n = 119,480), low-frequency (1 % ≤ MAF < 5 %, n = 8,705) and rare (MAF < 1 %, n = 110,903) variants separately (see also ***Supplementary Table 2***). For each

622 component we computed genetic relatedness matrices, followed by the calculation of narrow-
623 sense heritability (h^2) with 100 iterations of constrained REML fitting, assuming a disease
624 prevalence of 0.14[14]. We used the first ten principal components from population structure as
625 covariates. Furthermore, we estimated single SNP heritability for both *DOK2* (rs34215892,
626 rs56094005) and *CD200R1* (rs9865242) using the method of So *et al.*[73].
627

References

1. Weidinger, S. & Novak, N. Atopic dermatitis. *Lancet* **387**, 1109-1122 (2016).
2. Schaarschmidt, H. *et al.* A genome-wide association study reveals 2 new susceptibility loci for atopic dermatitis. *J Allergy Clin Immunol* **136**, 802-6 (2015).
3. Esparza-Gordillo, J. *et al.* A common variant on chromosome 11q13 is associated with atopic dermatitis. *Nat Genet* **41**, 596-601 (2009).
4. Sun, L.D. *et al.* Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population. *Nat Genet* **43**, 690-4 (2011).
5. Hirota, T. *et al.* Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. *Nat Genet* **44**, 1222-6 (2012).
6. Paternoster, L. *et al.* Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis. *Nat Genet* **44**, 187-92 (2011).
7. Ellinghaus, D. *et al.* High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nat Genet* **45**, 808-12 (2013).
8. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet* **47**, 1449-1456 (2015).
9. Irvine, A.D., McLean, W.H. & Leung, D.Y. Filaggrin mutations associated with skin and allergic diseases. *N Engl J Med* **365**, 1315-27 (2011).
10. Tsunemi, Y. *et al.* Interleukin-13 gene polymorphism G4257A is associated with atopic dermatitis in Japanese patients. *J Dermatol Sci* **30**, 100-7 (2002).
11. Esparza-Gordillo, J. *et al.* A functional IL-6 receptor (IL6R) variant is a risk factor for persistent atopic dermatitis. *J Allergy Clin Immunol* **132**, 371-7 (2013).
12. International Multiple Sclerosis Genetics Consortium *et al.* Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. *Cell* **175**, 1679-1687 e7 (2018).
13. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
14. Ferreira, M.A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet* **49**, 1752-1757 (2017).
15. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243 (2013).
16. Fehrmann, R.S. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* **7**, e1002197 (2011).
17. Mashima, R., Hishida, Y., Tezuka, T. & Yamanashi, Y. The roles of Dok family adapters in immunoreceptor signaling. *Immunol Rev* **232**, 273-85 (2009).
18. Mihrshahi, R., Barclay, A.N. & Brown, M.H. Essential roles for Dok2 and RasGAP in CD200 receptor-mediated regulation of human myeloid cells. *J Immunol* **183**, 4879-86 (2009).
19. Mihrshahi, R. & Brown, M.H. Downstream of tyrosine kinase 1 and 2 play opposing roles in CD200 receptor signaling. *J Immunol* **185**, 7216-22 (2010).
20. Schrambach, S., Ardizzone, M., Leymarie, V., Sibilia, J. & Bahram, S. In vivo expression pattern of MICA and MICB and its relevance to auto-immunity and cancer. *PLoS One* **2**, e518 (2007).
21. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).
22. Tsoi, L.C. *et al.* Atopic dermatitis is an IL-13 dominant disease with greater molecular heterogeneity compared to psoriasis. *J Invest Dermatol* (2019).

23. Weidinger, S., Beck, L.A., Bieber, T., Kabashima, K. & Irvine, A.D. Atopic dermatitis. *Nat Rev Dis Primers* **4**, 1 (2018).
24. Boulay, I., Nemorin, J.G. & Duplay, P. Phosphotyrosine binding-mediated oligomerization of downstream of tyrosine kinase (Dok)-1 and Dok-2 is involved in CD2-induced Dok phosphorylation. *J Immunol* **175**, 4483-9 (2005).
25. Abramson, J., Rozenblum, G. & Pecht, I. Dok protein family members are involved in signaling mediated by the type 1 Fcepsilon receptor. *Eur J Immunol* **33**, 85-91 (2003).
26. Gerard, A. *et al.* Functional interaction of RasGAP-binding proteins Dok-1 and Dok-2 with the Tec protein tyrosine kinase. *Oncogene* **23**, 1594-8 (2004).
27. Yasuda, T. *et al.* Dok-1 and Dok-2 are negative regulators of T cell receptor signaling. *Int Immunol* **19**, 487-95 (2007).
28. Di Cristofano, A. *et al.* Molecular cloning and characterization of p56dok-2 defines a new family of RasGAP-binding proteins. *J Biol Chem* **273**, 4827-30 (1998).
29. Gugasyan, R. *et al.* Dok-related protein negatively regulates T cell development via its RasGTPase-activating protein and Nck docking sites. *J Cell Biol* **158**, 115-25 (2002).
30. Shinohara, H. *et al.* Dok-1 and Dok-2 are negative regulators of lipopolysaccharide-induced signaling. *J Exp Med* **201**, 333-9 (2005).
31. Mashima, R. *et al.* Dok adaptors play anti-inflammatory roles in pulmonary homeostasis. *Genes Cells* **18**, 56-65 (2013).
32. Celis-Gutierrez, J. *et al.* Dok1 and Dok2 proteins regulate natural killer cell development and function. *EMBO J* **33**, 1928-40 (2014).
33. Nothlings, U. & Krawczak, M. [PopGen. A population-based biobank with prospective follow-up of a control group]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **55**, 831-5 (2012).
34. 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 (2005).
35. Schmermund, A. *et al.* Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. *Am Heart J* **144**, 212-8 (2002).
36. Volzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int J Epidemiol* **40**, 294-307 (2011).
37. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann Allergy Asthma Immunol* **93**, 381-9 (2004).
38. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
39. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
40. Morris, J.A., Randall, J.C., Maller, J.B. & Barrett, J.C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26**, 1786-7 (2010).
41. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
42. Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-6 (2016).
43. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266-72 (2014).
44. Feng, S., Liu, D., Zhan, X., Wing, M.K. & Abecasis, G.R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828-9 (2014).

45. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
46. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-501 (2016).
47. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
48. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am J Hum Genet* **101**, 539-551 (2017).
49. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
50. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
51. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-7 (2015).
52. Machiela, M.J. & Chanock, S.J. LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. *Bioinformatics* **34**, 887-889 (2018).
53. Ward, L.D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* **44**, D877-81 (2016).
54. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
55. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).
56. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).
57. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
58. Huang, J. *et al.* Co-expression and significance of Dok2 and Ras p21 protein activator 1 in breast cancer. *Oncol Lett* **14**, 5386-5392 (2017).
59. DeLano, W.L. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* **12**, 14-20 (2002).
60. DeLano, W.L. The PyMOL Molecular Graphics System. . *Delano Scientific, San Carlos, CA* (2002).
61. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
62. Martin, A. *et al.* BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics* **11**, 91 (2010).
63. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91 (2000).
64. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
65. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501 (2004).
66. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841-6 (2012).
67. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572-4 (2007).

68. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50 (2003).
69. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**, W214-20 (2010).
70. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
71. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
72. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
73. So, H.C., Li, M. & Sham, P.C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet Epidemiol* **35**, 447-56 (2011).
74. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
75. Holmannova, D. *et al.* CD200/CD200R paired potent inhibitory molecules regulating immune and inflammatory responses; Part I: CD200/CD200R structure, activation, and function. *Acta Medica (Hradec Kralove)* **55**, 12-7 (2012).
76. Zhang, S., Cherwinski, H., Sedgwick, J.D. & Phillips, J.H. Molecular mechanisms of CD200 inhibition of mast cell activation. *J Immunol* **173**, 6786-93 (2004).

Acknowledgements

This work 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 01ZX1606A; GB-XMAP grant 01ZX1709) as well as the German ELIXIR node de.NBI (de.STAIR grant 031L0106C). The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany`s Excellence Strategy – EXC 2167-390884018. The project received infrastructure support from the DFG Excellence Cluster No. 306 "Inflammation at Interfaces" and the PopGen Biobank (Kiel, Germany). The KORA research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by 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 LMU innovativ. SHIP (Study of Health in Pomerania) is part of the Community Medicine Research net (CMR) of the University of Greifswald, Germany, which is funded by BMBF (grants 01ZZ9603, 01ZZ0103 and 01ZZ0403) and 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 BMBF (grant 03IS2061A). Exome chip data have been supported by BMBF (grant no. 03Z1CN22) and the Federal State of Mecklenburg-West Pomerania. LP was supported by a UK Medical Research Council fellowship (MR/J012165/1).

Author contributions

A.F., S.W. and D.E conceived and designed the study. A.F. and D.E. supervised the study. W.L., M.M-N., M.M.N., P.H., A.A., G.H., C.O.M., M.S., K.B., T.S.A., H.Bi., D.S., S.G., A.P., C.G., M.Wa., H.S., T.W., E.R., K.S., T.W., R.F-H., N.N., M.Ho., P.R., S.S., Y.-A. L., and S.W. contributed to sample collection and phenotyping. S.M., D.E., E.R. and E.E. executed sample handling, genotyping and/or quality control. U.W. performed evaluation of immunohistochemistry. G.M. performed *in silico* variant protein analysis. S.M. and H.B. performed statistical and bioinformatics analysis for the discovery and the combined meta analyses, while T.S.A. performed statistical analysis for the Danish replication cohort and S.G., I.M. performed statistical analysis for UK Biobank. H.E., F.D., M.Hü., J.C.K., L.W. and D.E. helped with bioinformatic analyses. S.B., M.W., O.W., S.M. and D.E. performed network analysis. S.M., H.B., G.M., E.E., E.S.J., N.D., L.P., S.W., A.F. and D.E. participated in interpretation of results. S.M., H.B. and D.E. wrote the manuscript. All authors critically reviewed and approved the manuscript.

Declaration of Interests

No conflict of interest.

Figure legends

Figure 1 – Exome chip association analysis identified two low-frequency (rs34215892 and rs56094005; 1 % ≤ MAF < 5 %) and ten rare (MAF < 1 %) coding variants contributing to the genome-wide significant association signal at *DOK2* ($p_{DOK2} = 4.23\times10^{-13}$; *Table 2*). Minor allele frequency in AD cases (dark blue) and controls (light blue) are shown from discovery exome chip meta-analysis (see also ***Supplementary Table 1***). 10 out of 12 variants were predicted to be pathogenic (*see **Methods***), suggesting that multiple missense variants contribute to the gene-based association signal. Both low-frequency lead variants are LD-independent ($r^2_{rs56094005\text{-}rs34215892} = 0.0015$). Variant effect predictions (by SIFT and PolyPhen-2) depicted in red (white or grey) represent amino acid substitutions predicted to be potentially damaging (probably not damaging or prediction not possible). SNV – single nucleotide variant; ESS – essential splice site; pos – amino acid position; PP – PolyPhen-2. PH – Pleckstrin-homology domain; PTB – Phosphotyrosine-binding domain; PRR1 – Prolin-rich region 1; PRR2 – Prolin-rich region 2. A – Alanine; D – Aspatic acid; F – Phenylalanine; H – Histidine; L – Leucin; N – Asparagine; P – Proline; R – Arginine; S – Serine; T – Threonine; W – Tryptophan.

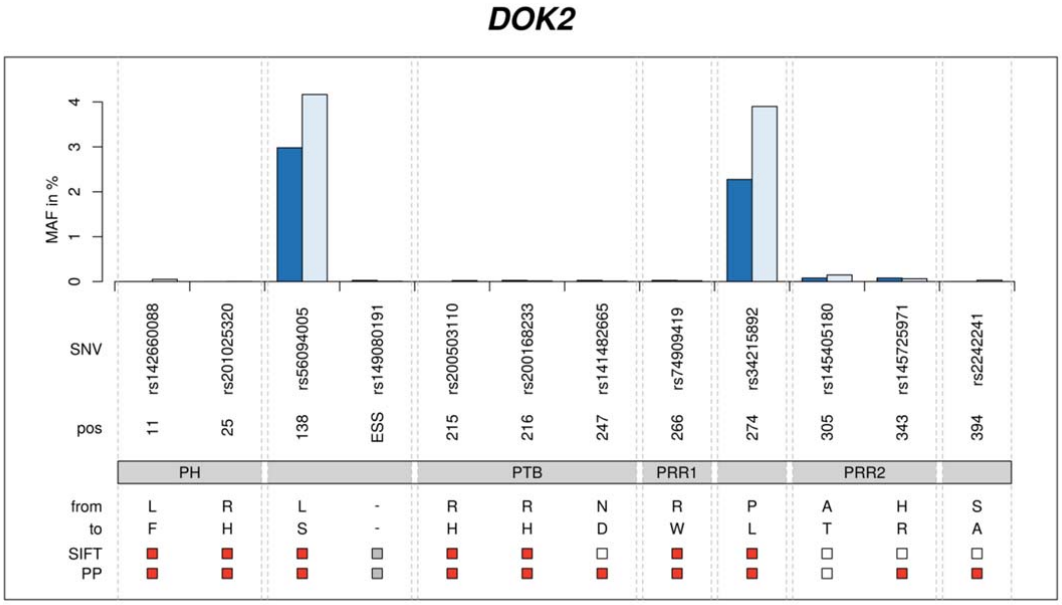


Figure 2 – Hypothetical model constructed from protein sequence and structural domain analysis suggesting that missense lead variants rs56094005, rs34215892 (p.L138S and p.P274L; *DOK2*) and rs9865242 (p.E312Q; *CD200R1*) are located near functionally important tyrosine phosphorylation sites and may interfere with CD200/CD200R1 receptor complex and DOK2 function.

We assume a simplified illustration of DOK2 function in response to CD200R1 in human myeloid cells[74] in which CD200/CD200R1 binding leads to tyrosine phosphorylation of the NPLY motif in the cytoplasmic tail of CD200R1 and the recruitment of the DOK2 adapter protein [18]. DOK2 interaction with CD200R1 leads to DOK2 tyrosine phosphorylation at positions Y271/Y299 (activating RasGAP) and Y139 (activating DOK1), which leads to the recruitment of RasGAP and its subsequent activation[19]. Activated RasGAP inhibits mitogen-activated kinase (MAPK) signalling and subsequenty reduces production of pro-inflammatory cytokines such as TNFα, INFγ, IL1, IL17, IL6, IL8[75,76]. By means of our protein sequence and structure analyses we observed that variant rs56094005 (p.L138S) locates within a linker sequence between the PH and the PTB domain adjacent the Y139 phosphorylation-dependent DOK1 interaction site and may interfere with heterodimerization of DOK1 and DOK2 required for full phosphorylation of the two proteins and signalling[24]. Variant rs34215892 (p.P274L), located in the invariant RasGAP-SH2 binding consensus motif YxxPxD[24], is likely affecting local protein structure conformation due to the unique structural rigidity of the proline side chain and therefore predicted to interfere with RasGAP signalling. Assuming an important stabilizing structural role of the proline within the binding motif, the variant p.P274L is likely to disturb the RasGAP activation by DOK2. The amino acid substitution of E to Q of variant rs9865242 (p.E312Q) causes a loss of positive charge in proximity of the NPLY motif and protein interaction site and may therefore modify protein-protein contacts and signalling.

Interacting CD200 and CD200R1 extracellular domains are visualized as structural models received from the Protein Databank (PDB ID 4bfi), while CD200R1 transmembrane and cytoplasmic tail is shown as blue lines. Variants are highlighted in red to indicate their relative positions within protein domains. The structural effect of protein variants cannot be modelled in 3D due to the lack of structural templates in the PDB for the concerned regions. Interactions between proteins are visualized as solid lines, following events as dashed lines. Cell membranes anchoring the receptors are illustrated in orange-grey. DOK2 domains were abbreviated as PH: Pleckstrin-homology domain, PTB: Phosphotyrosine-binding domain, PRR: Proline-rich region.

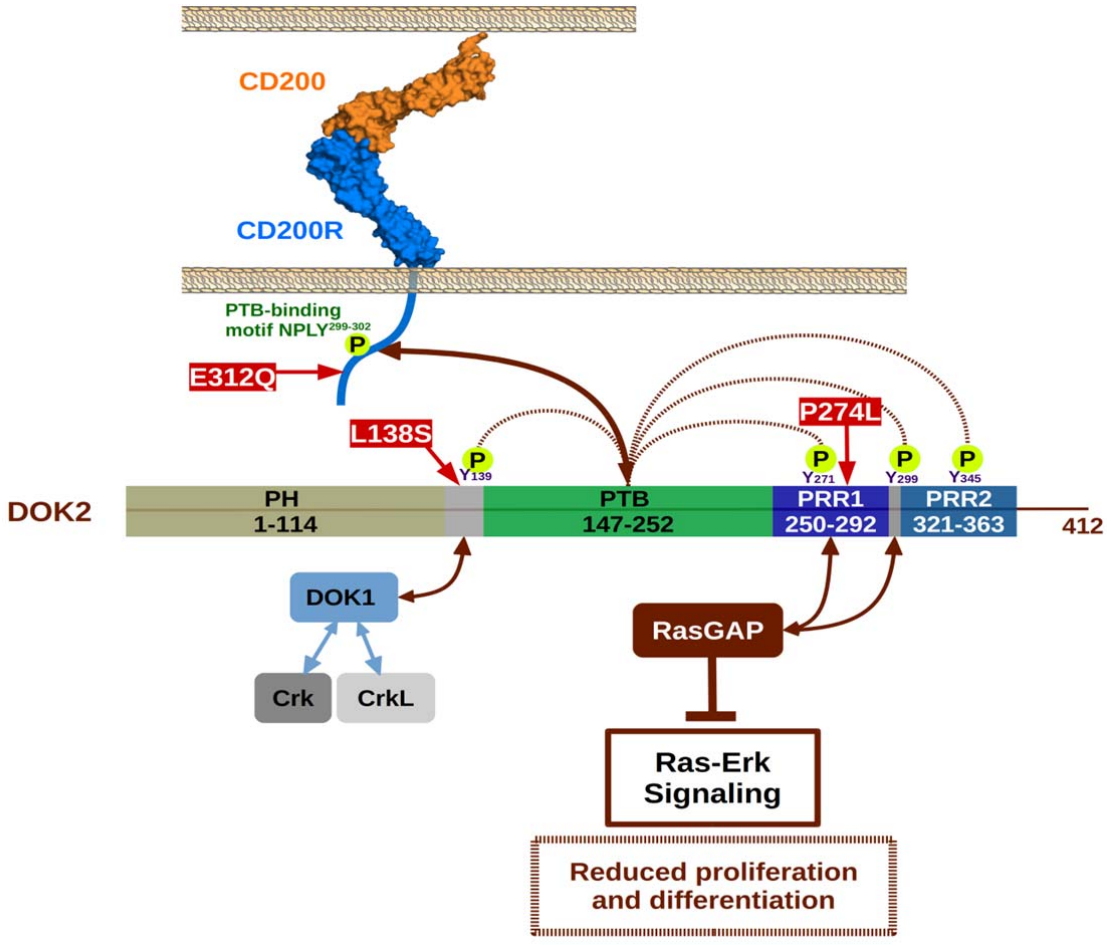


Figure 3 – Multiomics-based network interaction analysis unveiled topologically and functionally important AD susceptibility genes and determined *DOK2* as a central hub node interacting with *CD200R1* and candidate susceptibility genes identified by previous large AD consortium GWAS. (A) We developed a functional AD core network depicting the most important functional interactions prioritized from candidate genes taken from 32 established AD GWAS risk loci[8] as well as genes *DOK2* and *CD200R1* identified in this study. Topological network (step 1) and functional similarity (step 2) algorithms prioritized 30 core genes from initially 228 gene nodes and 562 interactions (see also ***Supplementary Figure 7***) based on protein-protein, protein-gene, co-expression, co-localization, and shared protein domain data sets from public resources (*see* ***Methods***). We identified *DOK2* as a central hub node interacting with *CD200R1* as well as main candidate genes (marked with *) from AD GWAS loci[2-8]. The majority of core genes (n = 22 out of 30; including *DOK2*) shows significantly upregulated and downregulated gene expression levels (depicted by black upwards and downwards arrows) in whole transcriptome mRNA-seq data on lesional and non-lesional skin samples from AD patients and skin samples from healthy individuals (see also ***Supplementary Table 9***). Grey-striped nodes (n = 10, including *DOK2*) represent the most relevant genes from topological network analysis (step 1; *see* ***Methods***), with two light grey-striped nodes (*CLEC16A*, *RAD50*) representing genes not interacting with *DOK2* and seven dark grey-striped nodes (*POLR2A, IL6, INPP5D, TRAF3, MICB, CD200R1, STAT3*) representing genes directly interacting with *DOK2*. Non-striped nodes (n = 20) represent genes additionally added by a functional similarity search (step 2; *see* ***Methods***), with 6 light grey non-striped nodes (*CD200*, *NBN*, *MRE11A*, *MET*, *HUWE1*, *POLR2D*) not directly interacting with *DOK2* and 14 dark grey non-striped nodes (*IL6*, *RASGAP*, *DOK1* among others) directly interacting with *DOK2*. **(B)** 17 out of the significantly up- or downregulated genes (n = 22; including *DOK2*), depicted with log2-transformed gene expression counts, are interacting with *DOK2*.

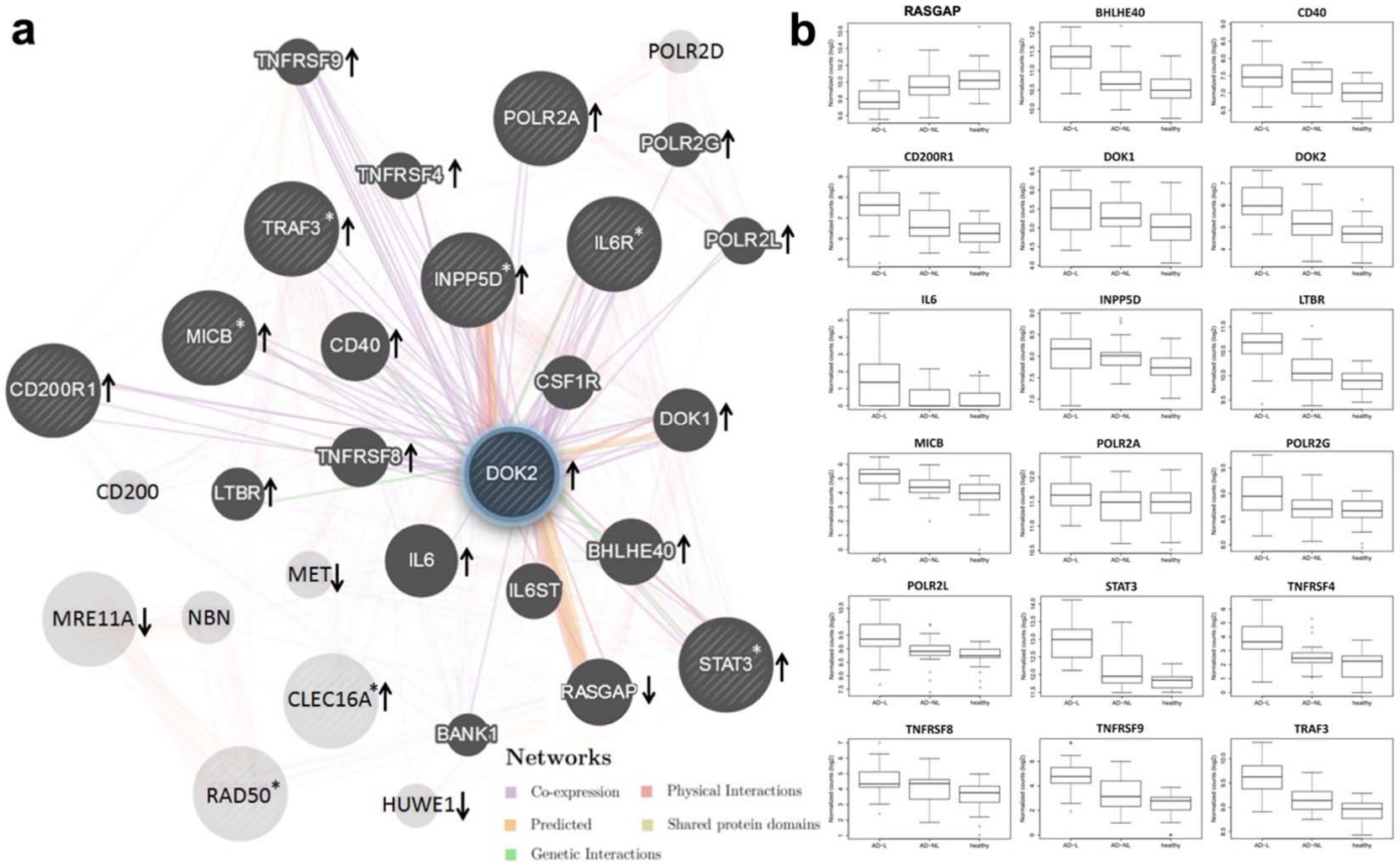
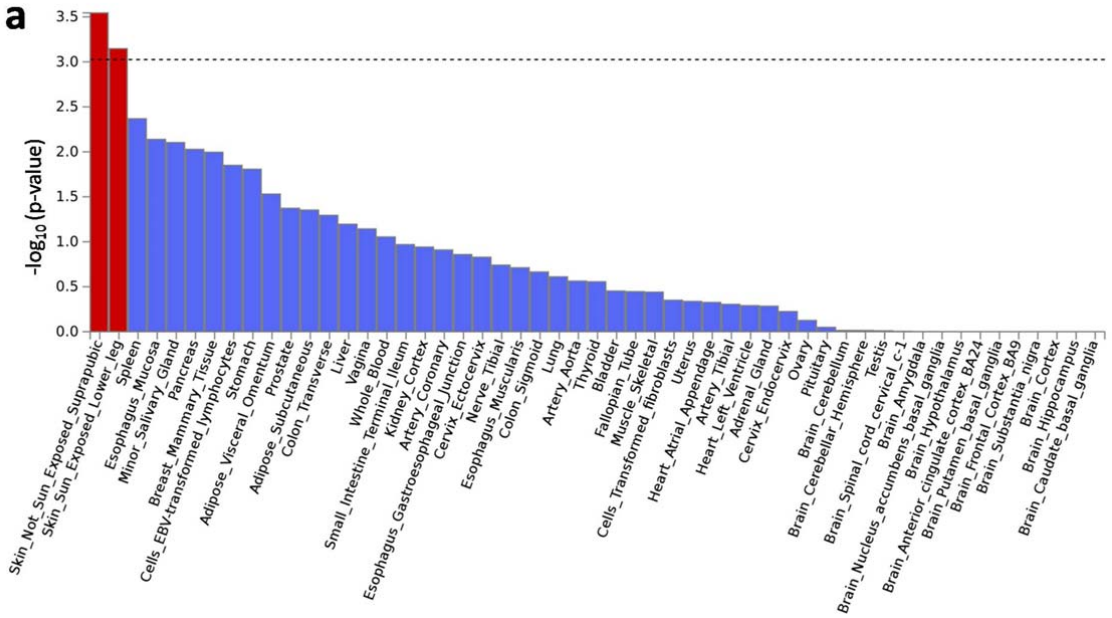
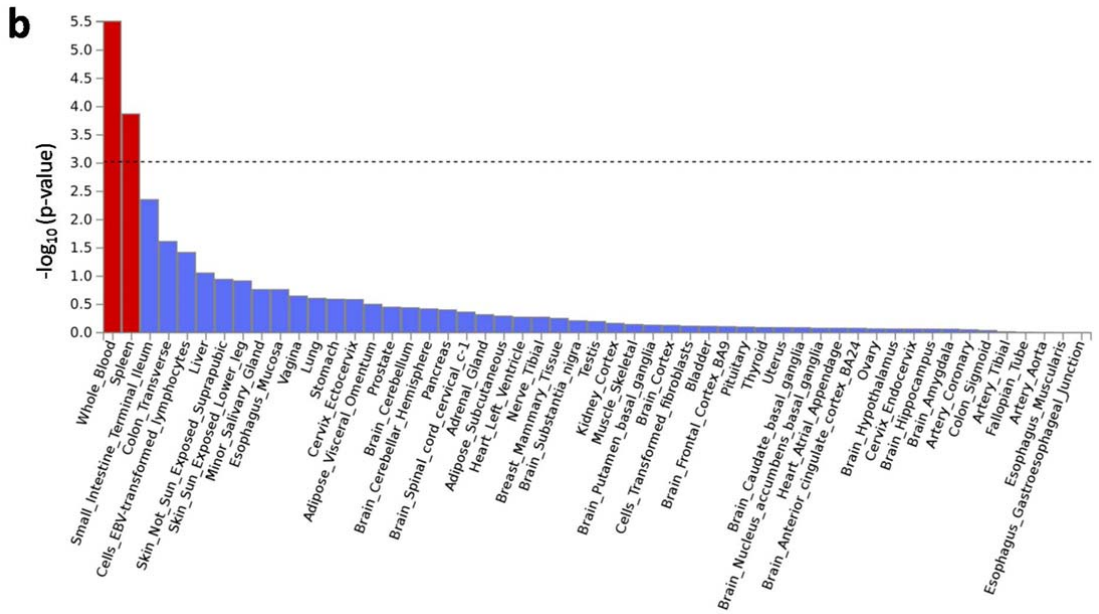

a
TNFRSF9
POLR2D
POLR2A
POLR2G
TNFRSF4
TRAF3*
IL6R*
POLR2L
INPP5D*
MICB*
CD40
CD200R1
CSF1R
DOK1
TNFRSF8
DOK2
CD200
LTBR
IL6
BHLHE40
MET
IL6ST
MRE11A
NBN
STAT3*
CLEC16A*
RASGAP
BANK1
RAD50*
HUWE1
Networks
Co-expression
Physical Interactions
Predicted
Shared protein domains
Genetic Interactions
b
RASGAP
BHLHE40
CD40
CD200R1
DOK1
DOK2
IL6
INPP5D
LTBR
MICB
POLR2A
POLR2G
POLR2L
STAT3
TNFRSF4
TNFRSF8
TNFRSF9
TRAF3
AD-L
AD-NL
healthy

Figure 4 – Tissue specificity analysis using 53 tissue types from GTEx identifies a significant positive relationship between skin tissue gene expression profiles and exome chip association statistics. Exome chip single variant association score statistics (exome chip data set) and GWAS SNP association summary statistics from the EAGLE Eczema Consortium (GWAS data set), respectively, were converted to AD-gene association Z scores using a multiple linear principal components regression model as implemented in MAGMA[21] thereby ensuring that LD between SNPs is accounted for (*see* ***Methods***). We conducted tissue specificity analysis with FUMA[70] to test for a positive relationship between tissue-specific gene expression profiles constructed from 53 tissue types and 11,688 samples from GTEx v7 and AD-gene associations (either from exome chip or GWAS data set) represented by gene-based Z scores (*see* ***Methods***). We used Bonferroni correction $p < 0.05/53 = 9.43\times10^{-4}$ as significance cut-off. (**a)** The exome chip data set revealed a positive relationship between gene expression profiles from skin tissues "sun exposed" and "non-sun exposed" and AD-gene associations. **(b)** In comparison, the GWAS data set showed an association with tissues "whole blood" and "spleen".




b
-log10 (p-value)
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Whole_Blood
Spleen
Small_Intestine_Terminal_Ileum
Colon_Transverse
Cells_EBV-transformed_lymphocytes
Liver
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Minor_Salivary_Gland
Esophagus_Mucosa
Vagina
Lung
Stomach
Cervix_Ectocervix
Adipose_Visceral_Omentum
Prostate
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Pancreas
Brain_Spinal_cord_cervical_c-1
Adrenal_Gland
Adipose_Subcutaneous
Heart_Left_Ventricle
Nerve_Tibial
Breast_Mammary_Tissue
Brain_Substantia_nigra
Testis
Kidney_Cortex
Muscle_Skeletal
Brain_Putamen_basal_ganglia
Brain_Cortex
Cells_Transformed_fibroblasts
Brain_Frontal_Cortex_BA9
Bladder
Pituitary
Thyroid
Uterus
Brain_Caudate_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Heart_Atrial_Appendage
Brain_Anterior_cingulate_cortex_BA24
Ovary
Brain_Hypothalamus
Cervix_Endocervix
Brain_Hippocampus
Brain_Amygdala
Artery_Coronary
Colon_Sigmoid
Artery_Tibial
Fallopian_Tube
Artery_Aorta
Esophagus_Muscularis
Esophagus_Gastroesophageal_Junction

931

932

Tables

Table 1 – Exome chip discovery, replication and UK Biobank single-marker association analysis revealed *DOK2* and *CD200R1* as genome-wide significant AD susceptibility genes.

dbSNP ID AA substitution	Position (chr:bp)	Gene	Variant type	A1/ A2	AF_{cases}	$AF_{controls}$	Discovery $n_{cases}/n_{controls}$	Discovery p-value/OR (95 % CI)	Replication $n_{cases}/n_{controls}$	Replication p-value/OR (95 % CI)	Combined p-value/OR (95 % CI)
rs34215892 P274L	8:21767240	*DOK2*	missense	A/G	0.025	0.039	1,913/14,295	9.83×10^{-7}; 0.61 (0.34–0.86)	3,962/10,076	1.35×10^{-5}; 0.68 (0.56–0.79)	**2.15×10^{-10}; 0.64 (0.53–0.76)**
rs9865242 E312Q	3:112642568	*CD200R1*	missense	C/G	0.457	0.435	1,913/14,295	5.80×10^{-6}; 1.57 (1.49–1.65)	4,349/11,724	3.38×10^{-3}; 1.14 (1.09–1.19)	1.17×10^{-7}; 1.16 (1.13–1.19)
					UK Biobank summary statistics for AD, allergic rhinitis and/or hayfever (self-reported)				**UK Biobank summary statistics for AD or dermatitis (self-reported)**			
dbSNP ID AA substitution	**Position (chr:bp)**	**Gene**	**Variant type**	**A1/ A2**	**AF_{cases}**	**$AF_{controls}$**	**Phenotype 1 $n_{cases}/n_{controls}$**	**p-value**	**AF_{cases}**	**$AF_{controls}$**	**Phenotype 2 $n_{cases}/n_{controls}$**	**p-value**
rs34215892 P274L	8:21767240	*DOK2*	missense	A/G	0.027	0.029	83,407/277,120	3.35×10^{-6}	0.025	0.029	9,312/351,820	2.50×10^{-3}
rs9865242 E312Q	3:112642568	*CD200R1*	missense	C/G	0.474	0.467	83,407/277,120	**1.35×10^{-8}**	0.487	0.468	9,312/351,820	8.33×10^{-8}

AA, amino acid; chr, chromosome of the marker; bp, genomic position from NCBI dbSNP build v150 (genome build hg19); Gene, candidate gene; A1, minor allele; A2, major allele; AF, allele frequency of A1; OR, estimated odds ratio; P-values and ORs were calculated with respect to the minor allele; genome-wide significant p-values ($p < 5\times10^{-8}$) are indicated in bold. All AD cases from discovery and replication panels had been

940 diagnosed with AD (eczema) by a dermatologist. Association results of UK Biobank (self-reported phenotypes; see also ***Supplementary Table 1***)

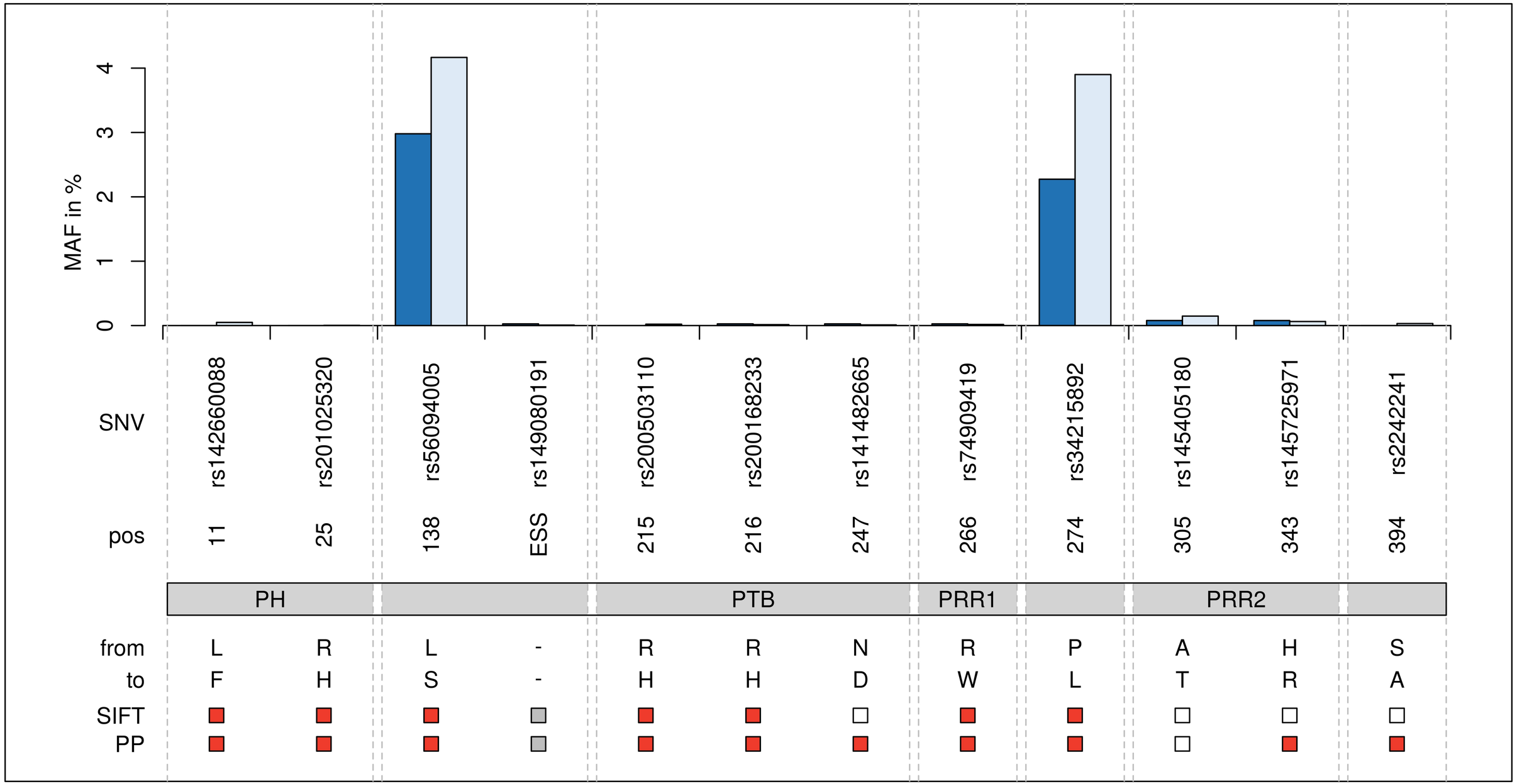
941 are listed separately.

Table 2 – Meta-analysis of gene-based aggregation tests for *DOK2* increased the genome-wide significant association signal by more than two orders of magnitude in comparison to single-marker analysis (*Table 1*), indicating that multiple rare variants (with 10 out of 12 variants predicted to be pathogenic; *Figure 1*) contribute to the association signal.

Gene	dbSNP ID	Position (chr:bp)	AF_{mean}	AA	Prediction	Discovery $n_{cases}/n_{controls}$ $pSKAT_{discovery}$	Replication $n_{cases}/n_{controls}$ $pSKAT_{replication}$	Combined $n_{cases}/n_{controls}$ $pSKAT_{combined}$
DOK2	rs2242241[#]	8:21766881	0.00028	S394A	Tolerated, possibly damaging	1,913/14,295 **2.61×10^{-7}**	3,932/10,076 **1.54×10^{-6}**	5,845/24,371 **4.23×10^{-13}**
	rs145725971[*]	8:21767033	0.00065	H343R	Tolerated, possibly damaging			
	rs145405180[#]	8:21767148	0.00139	A305T	Tolerated, benign			
	rs34215892[#]	8:21767240	0.03708	P274L	Damaging, probably damaging			
	rs74909419[*]	8:21767265	0.00019	R266W	Damaging, probably damaging			
	rs141482665[*]	8:21767322	0.00012	N247D	Tolerated, probably damaging			
	rs200168233[*]	8:21767414	0.00015	R216H	Damaging, probably damaging			
	rs200503110[#]	8:21767417	0.00019	R215H	Damaging, probably damaging			
	rs149080191[*]	8:21768370	0.00009	E.S.S.				
	rs56094005[#]	8:21769432	0.04026	L138S	Damaging, probably damaging			
	rs201025320[#]	8:21770011	0.00003	R25H	Damaging, probably damaging			
	rs142660088[#]	8:21771080	0.00043	L11F	Damaging, probably damaging			

AA, amino acid substitution; E.S.S., essential splice site; Gene, candidate gene; Chr, chromosome of the marker; Bp, genomic position from NCBI dbSNP build v150 (genome build hg19); AF_{mean}, mean allele frequency of minor allele from discovery panels (see also ***Supplementary Table 5***); Prediction, SIFT prediction, PolyPhen-2 prediction; pSKAT, p of Sequence Kernel Association Test; Bonferroni-corrected exome chip significant gene-based p-values ($p_{gene} < 0.05/15{,}998 = 3\times10^{-6}$; 15,998 genes) are indicated in bold. Cohort specific association details are given in ***Supplementary Table 6.*** *, risk variant (in context of the odds ratio); #, protective variant

DOK2


MAF in %
0
1
2
3
4
SNV
rs142660088
rs201025320
rs56094005
rs149080191
rs200503110
rs200168233
rs141482665
rs74909419
rs34215892
rs145405180
rs145725971
rs2242241
pos
11
25
138
ESS
215
216
247
266
274
305
343
394
PH
PTB
PRR1
PRR2
from
L
R
L
-
R
R
N
R
P
A
H
S
to
F
H
S
-
H
H
D
W
L
T
R
A
SIFT
PP

CD200

CD200R

PTB-binding
motif NPLY$^{299\text{-}302}$

P

E312Q

L138S

P
Y_{139}

P274L

P
Y_{271}

P
Y_{299}

P
Y_{345}

DOK2

PH	PTB	PRR1	PRR2	
1-114	147-252	250-292	321-363	412

DOK1

Crk

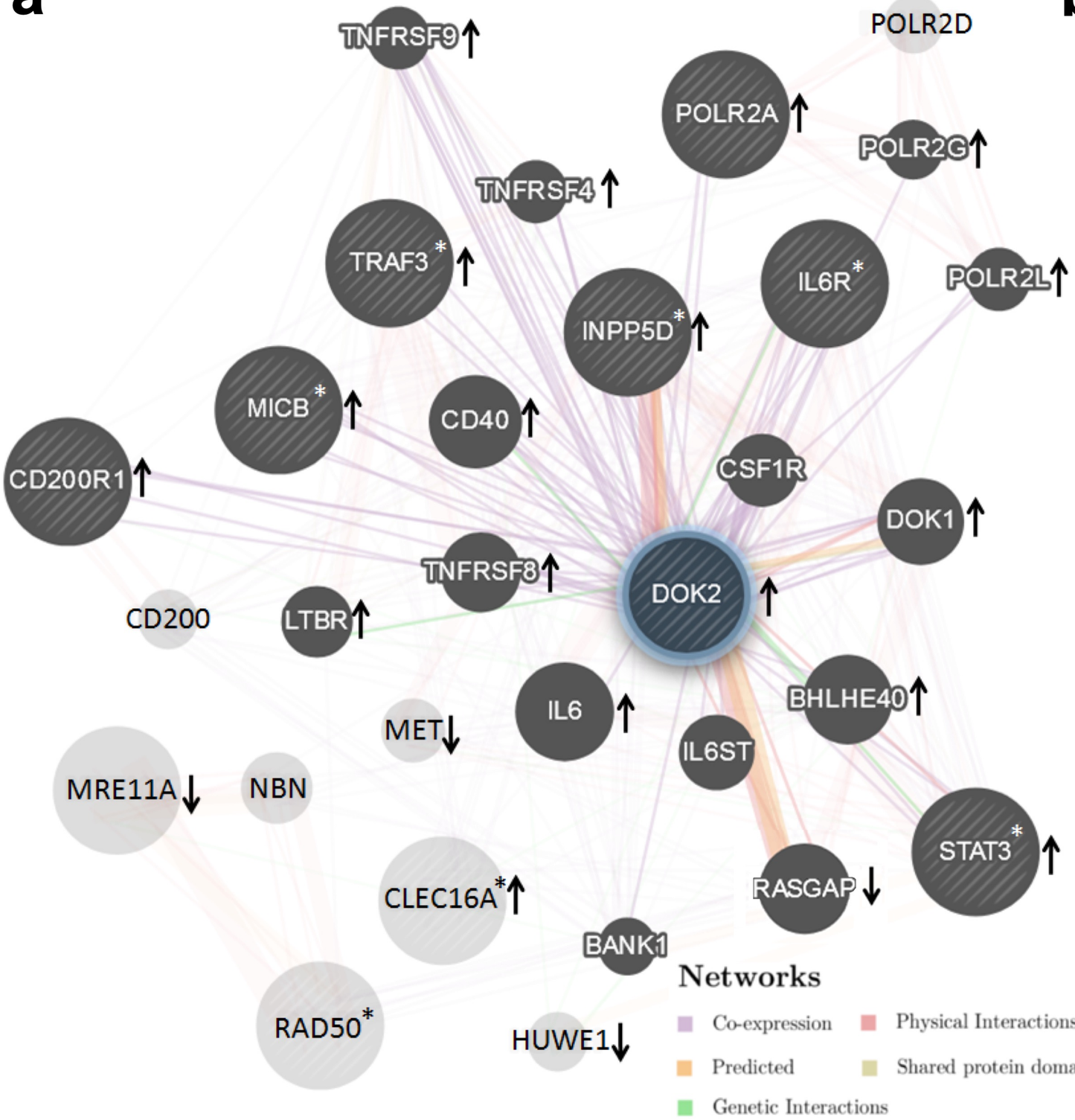
CrkL

RasGAP

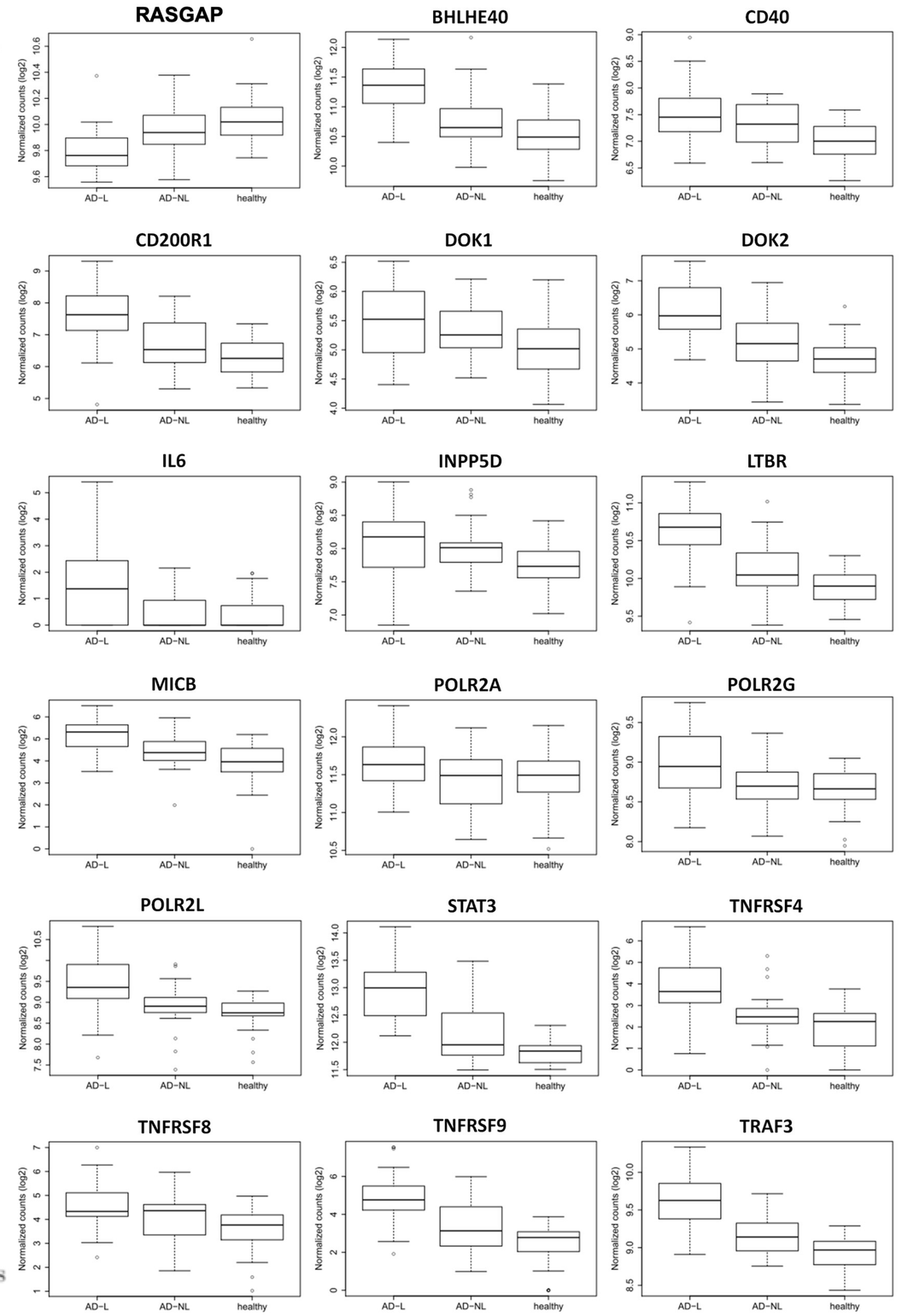
Ras-Erk
Signaling

Reduced proliferation
and differentiation

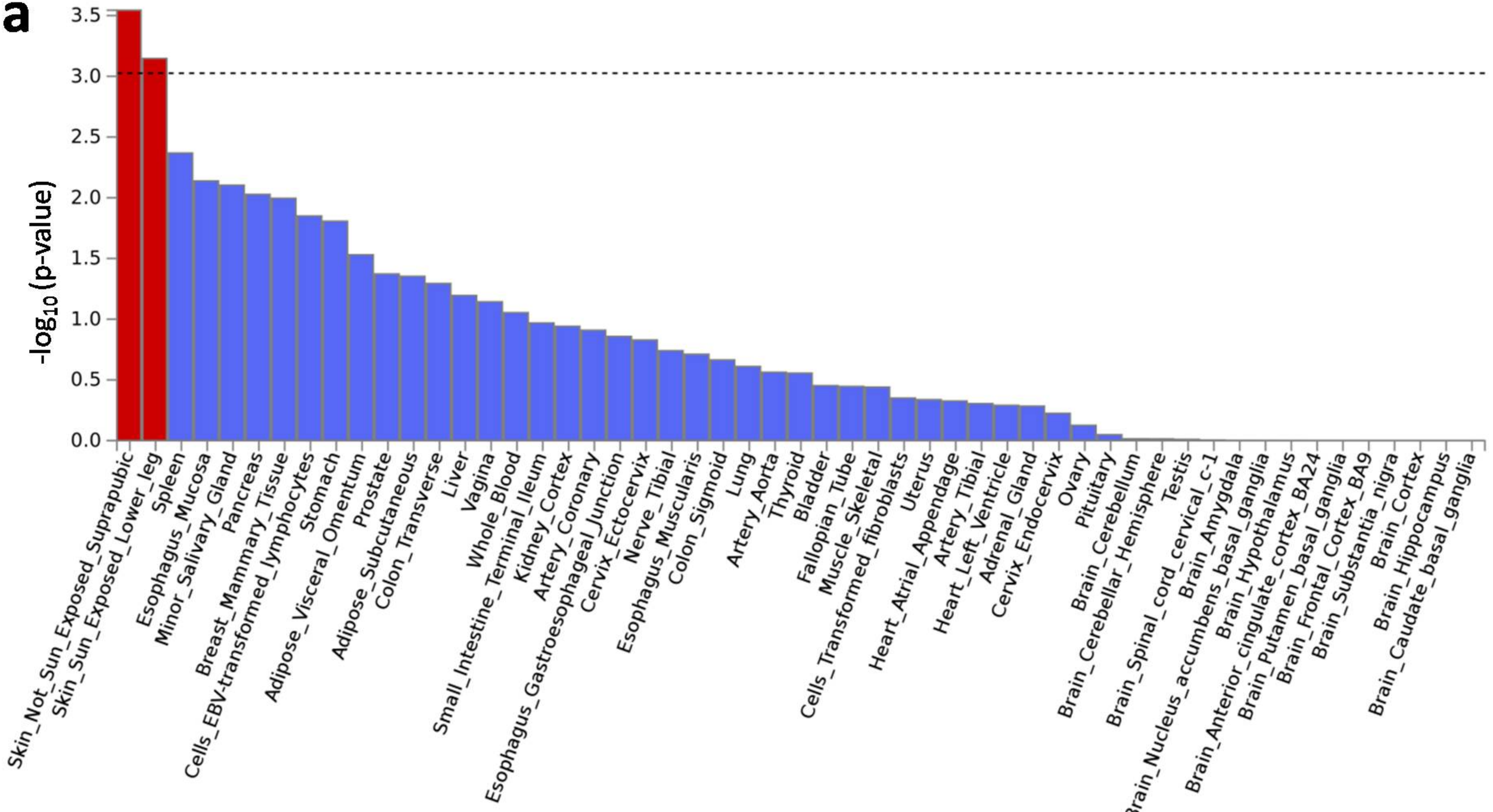
a

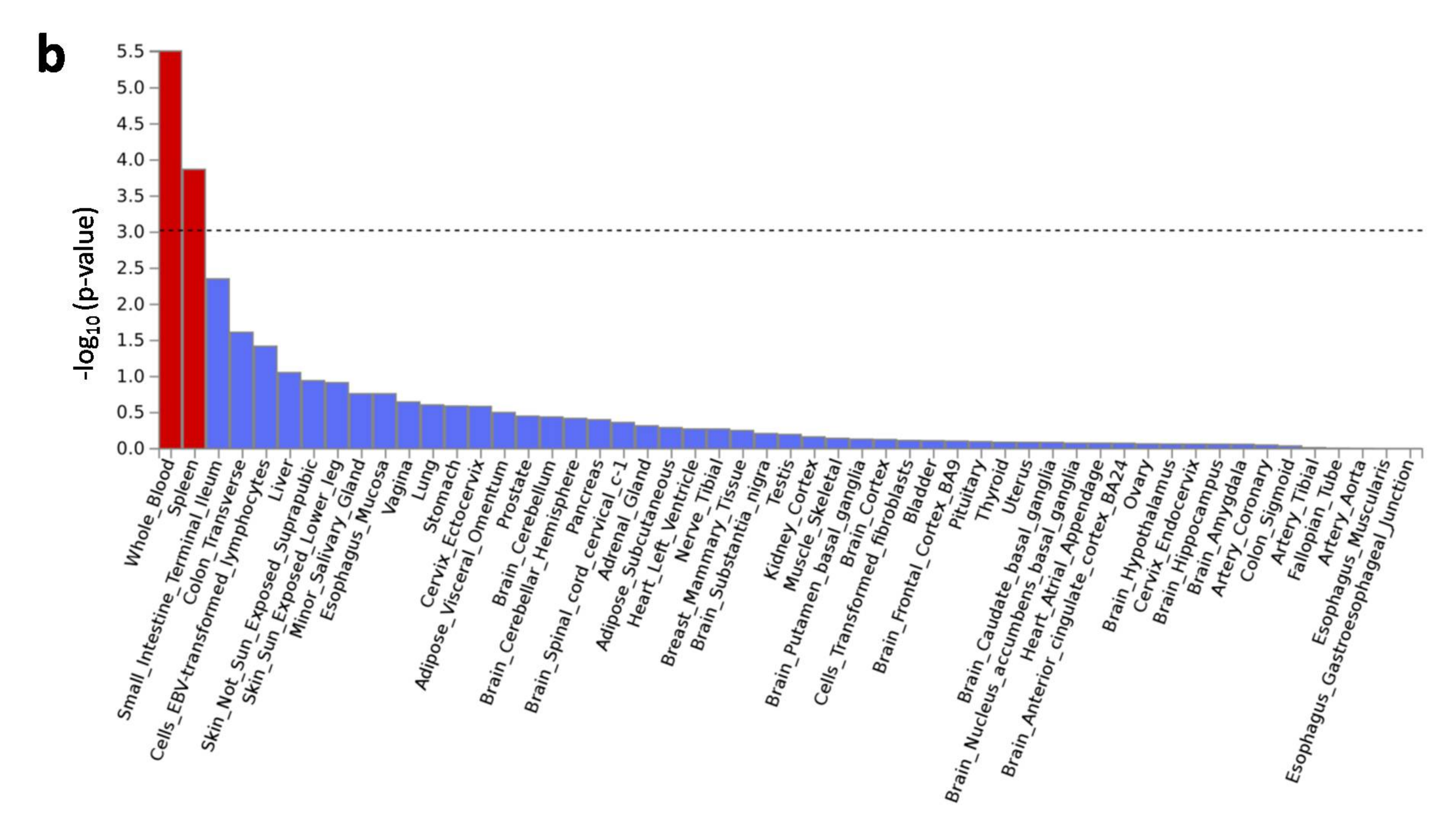

TNFRSF9
POLR2D
POLR2A
POLR2G
TNFRSF4
TRAF3*
IL6R*
POLR2L
INPP5D*
MICB*
CD40
CSF1R
CD200R1
DOK1
TNFRSF8
DOK2
CD200
LTBR
IL6
BHLHE40
MET
IL6ST
MRE11A
NBN
STAT3*
CLEC16A*
RASGAP
BANK1
RAD50*
HUWE1
Networks
Co-expression
Physical Interactions
Predicted
Shared protein domains
Genetic Interactions

b


RASGAP
BHLHE40
CD40
CD200R1
DOK1
DOK2
IL6
INPP5D
LTBR
MICB
POLR2A
POLR2G
POLR2L
STAT3
TNFRSF4
TNFRSF8
TNFRSF9
TRAF3
Normalized counts (log2)
AD−L
AD−NL
healthy

a


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-log10 (p-value)
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Spleen
Esophagus_Mucosa
Minor_Salivary_Gland
Pancreas
Breast_Mammary_Tissue
Cells_EBV-transformed_lymphocytes
Stomach
Adipose_Visceral_Omentum
Prostate
Adipose_Subcutaneous
Colon_Transverse
Liver
Vagina
Whole_Blood
Small_Intestine_Terminal_Ileum
Kidney_Cortex
Artery_Coronary
Esophagus_Gastroesophageal_Junction
Cervix_Ectocervix
Nerve_Tibial
Esophagus_Muscularis
Colon_Sigmoid
Lung
Artery_Aorta
Thyroid
Bladder
Fallopian_Tube
Muscle_Skeletal
Cells_Transformed_fibroblasts
Uterus
Heart_Atrial_Appendage
Artery_Tibial
Heart_Left_Ventricle
Adrenal_Gland
Cervix_Endocervix
Ovary
Pituitary
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Testis
Brain_Spinal_cord_cervical_c-1
Brain_Amygdala
Brain_Nucleus_accumbens_basal_ganglia
Brain_Hypothalamus
Brain_Anterior_cingulate_cortex_BA24
Brain_Putamen_basal_ganglia
Brain_Frontal_Cortex_BA9
Brain_Substantia_nigra
Brain_Cortex
Brain_Hippocampus
Brain_Caudate_basal_ganglia


b
-log10 (p-value)
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Whole_Blood
Spleen
Small_Intestine_Terminal_Ileum
Colon_Transverse
Cells_EBV-transformed_lymphocytes
Liver
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Minor_Salivary_Gland
Esophagus_Mucosa
Vagina
Lung
Stomach
Cervix_Ectocervix
Adipose_Visceral_Omentum
Prostate
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Brain_Spinal_cord_cervical_c-1
Pancreas
Adrenal_Gland
Adipose_Subcutaneous
Heart_Left_Ventricle
Nerve_Tibial
Breast_Mammary_Tissue
Brain_Substantia_nigra
Testis
Kidney_Cortex
Muscle_Skeletal
Brain_Putamen_basal_ganglia
Brain_Cortex
Cells_Transformed_fibroblasts
Bladder
Brain_Frontal_Cortex_BA9
Pituitary
Thyroid
Uterus
Brain_Caudate_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Heart_Atrial_Appendage
Brain_Anterior_cingulate_cortex_BA24
Ovary
Brain_Hypothalamus
Cervix_Endocervix
Brain_Hippocampus
Brain_Amygdala
Artery_Coronary
Colon_Sigmoid
Artery_Tibial
Fallopian_Tube
Artery_Aorta
Esophagus_Muscularis
Esophagus_Gastroesophageal_Junction