

ARTICLE

<https://doi.org/10.1038/s41467-019-14144-8>

OPEN

Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis

Thorunn A. Olafsdottir^{1,2}, Fannar Theodors¹, Kristbjorg Bjarnadottir¹, Unnur Steina Bjornsdottir^{3,4}, Arna B. Agustsdottir¹, Olafur A. Stefansson¹, Erna V. Ivarsdottir^{1,5}, Jon K. Sigurdsson¹, Stefania Benonisdottir¹, Gudmundur I. Eyjolfsson⁶, David Gislason^{4,7}, Thorarinn Gislason^{2,8}, Steinunn Guðmundsdóttir¹, Arnaldur Gylfason¹, Bjarni V. Halldorsson^{1,9}, Gisli H. Halldorsson¹, Thorhildur Juliusdottir¹, Anna M. Kristinsdottir¹, Dora Ludviksdottir^{2,7}, Bjorn R. Ludviksson^{2,10}, Gisli Masson¹, Kristjan Norland¹, Pall T. Onundarson^{2,11}, Isleifur Olafsson¹², Olof Sigurdardottir^{2,13}, Lilja Stefansdottir¹, Gardar Sveinbjornsson¹, Vinicius Tragante^{1,14}, Daniel F. Gudbjartsson^{1,5}, Gudmar Thorleifsson¹, Patrick Sulem¹, Unnur Thorsteinsdottir^{1,2}, Gudmundur L. Norddahl¹, Ingileif Jonsdottir^{1,2*} & Kari Stefansson^{1,2*}

Asthma is one of the most common chronic diseases affecting both children and adults. We report a genome-wide association meta-analysis of 69,189 cases and 702,199 controls from Iceland and UK biobank. We find 88 asthma risk variants at 56 loci, 19 previously unreported, and evaluate their effect on other asthma and allergic phenotypes. Of special interest are two low frequency variants associated with protection against asthma; a missense variant in *TNFRSF8* and 3' UTR variant in *TGFBR1*. Functional studies show that the *TNFRSF8* variant reduces *TNFRSF8* expression both on cell surface and in soluble form, acting as loss of function. eQTL analysis suggests that the *TGFBR1* variant acts through gain of function and together with an intronic variant in a downstream gene, *SMAD3*, points to defective TGFβR1 signaling as one of the biological perturbations increasing asthma risk. Our results increase the number of asthma variants and implicate genes with known role in T cell regulation, inflammation and airway remodeling in asthma pathogenesis.

¹deCODE genetics/Amgen, Inc., Reykjavik, Iceland. ²Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. ³Department of Medicine, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ⁴The Medical Center Mjodd, Reykjavik, Iceland. ⁵School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. ⁶The Laboratory in Mjodd, RAM, Reykjavik, Iceland. ⁷Department of Respiratory Medicine and Sleep, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ⁸Department of Sleep, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ⁹School of Science and Engineering, Reykjavik University, Reykjavik, Iceland. ¹⁰Department of Immunology, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ¹¹Department of Laboratory Hematology, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ¹²Department of Clinical Biochemistry, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland. ¹³Department of Clinical Biochemistry, Akureyri Hospital, Akureyri, Iceland. ¹⁴Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands. *email: ingileif.jonsdottir@decode.is; kari.stefansson@decode.is

Asthma is one of the most common chronic diseases and has a substantial impact on the quality of life of both children and adults. Currently, it is estimated that around 300 million people have asthma worldwide and those numbers are predicted to rise in the coming years¹. Asthma is a syndrome with heterogeneous pathophysiology and different asthma phenotypes differ in age of onset, environmental risk factors, clinical presentation, prognosis, and response to therapies². To further characterize the distinct mechanistic pathways underlying the disease, asthma endotypes have been defined and can broadly be regarded as type 2 (T2) high or T2 low. T2 high asthma is characterized by increased activation of T helper cells of type 2 (Th2), innate lymphoid cells of type 2 (ILC2), and eosinophils. In contrast to T2 high, T2 low endotype is less well defined and is typically characterized by the absence of markers of T2 high disease and has rather been linked with activation of neutrophils, Th1 and/or Th17 cells³. This heterogeneity may explain why fewer asthma loci have been identified through genome-wide association studies (GWAS) than in other diseases of similar prevalence⁴. In total, 44 sequence variants (36 in Europeans and 8 in other ancestries) have been reported to associate with asthma in 27 independent GWAS^{5–7}. We have previously discovered common sequence variants in *IL33* and its receptor *IL1RL1* conferring risk of asthma⁸, followed by an identification of a rare loss of function variant in *IL33* that protects against asthma⁹, thereby supporting its relevance as pharmacological target for asthma.

Comorbidity between asthma and other allergic diseases (especially allergic rhinitis and atopic dermatitis) has been reported¹⁰ and recent publications have focused on shared risk variants and genetic links between these traits^{6,11}.

Here we describe a large meta-analysis of asthma and report 88 independent associations at 56 loci. We perform a series of functional analysis to explore the biological effect of a low frequency missense variant in a gene of the tumor necrosis receptor family, *TNFRSF8*. We also report a low frequency 3 prime untranslated region (UTR) variant in *TGFBR1* that changes a microRNA (miR) recognition site and associates with increased *TGFBR1* expression in blood. Further, we report evidence of a single candidate gene for 8 of the 19 previously unreported asthma variants by extensive study of coding variants, expression quantitative trait loci (eQTLs) as well as enhancer and promoter signals. Lastly, we investigate association of the asthma variants with asthma sub-phenotypes (early-/late-onset and allergic asthma) as well as related traits (eosinophil count and allergic diseases).

Results and discussion

Genome-wide meta-analysis. We performed a meta-analysis combining asthma GWAS results from Iceland ($n = 16,247$ cases, $n = 346,486$ controls) and the UK biobank ($n = 52,942$ cases, $n = 355,713$ controls). The asthma phenotype in both sets was based on physician diagnoses and/or self-reported doctor's diagnosis of asthma. Association between genotype and phenotype was tested by logistic regression, assuming a multiplicative model (see methods). We used genome-wide significant thresholds dependent on variant annotation and found association with asthma at 55 loci (>1 Mb apart) in addition to the extended HLA region (chr6: 25,000,000–35,000,000, build hg38). Conditional analysis revealed one or more secondary signals ($P < 5 \times 10^{-8}$) at 17 of the 56 loci thus yielding 88 independent signals (Fig. 1, Supplementary Data 1). Based on number of cases and controls, we note that the effective sample size is four times bigger for the UK biobank samples ($N = 84,396$) than the Icelandic ($N = 22,689$), hence most of the power comes from the UK biobank samples.

The estimated genomic inflation in the two studies (estimated using LD-score regression)¹² is 1.368 for the Icelandic dataset and 1.092 for the UK biobank dataset, reflecting the relatedness of individuals in Iceland which is also the most likely explanation of genomic inflation in UK biobank where 30% of the participants have a relative (third degree or closer) in the UK biobank dataset¹³.

We note that only 4 out of the 88 variants showed significant heterogeneity in effect sizes based on the number of variants tested ($P < 0.05/8 = 5.62 \times 10^{-4}$) and 28 showed nominal heterogeneity ($P < 0.05$) between the two sample sets demonstrating a good consistency of effect for vast majority of the variants. The observed heterogeneity of the effect estimates is likely due to difference in the phenotype definition and/or ascertainment between the studies, coherent with retrospective studies as reported here. Further, 85 out of the 88 variants ($P < 1 \times 10^{-16}$) had effect size in the same direction in the two cohorts and 44 were at least nominally significant in Iceland showing good replication of effects (Supplementary Data 1).

Of the 88 independent signals, 47 are at 24 previously reported asthma loci, 22 variants at 16 loci have previously been reported for a combined allergic phenotype¹¹ and 19 variants at 16 loci are previously unreported asthma signals. We note that while this manuscript was under consideration two independent reports published association of asthma with 9 of those 19 variants as indicated in Table 1^{14,15}.

Out of the 47 variants at previously reported loci, 24 were represented by previously reported variants ($r^2 \geq 0.2$; Supplementary Data 2) and 23 by previously unreported signals at those loci ($r^2 < 0.2$; Supplementary Data 3). Furthermore, we replicated 31 out of the 36 reported European asthma loci^{7,16} when adjusting for the 36 variants tested ($P < 0.05/36 = 1.4 \times 10^{-3}$). For two of five non-replicating variants we found other genome wide significant (GWS) variants at the respective loci (Supplementary Data 4).

Forty-one of our 88 variants at 32 loci have not been previously associated with asthma, although 22 variants at 16 loci have previously been reported for a combined allergic phenotype of asthma, hay fever and eczema¹¹ (Supplementary Data 5).

Altogether, we identified 19 independent signals at 16 loci not previously reported to associate with asthma or the combined allergic phenotype¹¹ (Table 1). Sixteen of the previously unreported variants were common (Effect Allele Frequency (EAF) $\geq 5\%$) and three were low frequency variants. None of those variants showed significant ($P < 0.05/88 = 5.62 \times 10^{-4}$) heterogeneity of effects in the two sample sets (Table 1).

Loss of function variant in *TNFRSF8*. One of the low frequency variants (EAF = 1.2% (ICE)/1.5% (UK)) rs2230624_A is a missense variant p.Cys273Tyr in *TNFRSF8* (alias *CD30*) that associates with reduced asthma risk. The p.Cys273Tyr variant had the greatest protective effects (OR = 0.82, $P = 8.27 \times 10^{-13}$) of the 19 previously unreported variants and in fact there were only 4 out of the 88 asthma variants with greater effect on asthma risk; rs72782676 (intergenic on chr10, OR = 0.63), rs149045797 (intron variant at *IL33* locus, OR = 0.65), rs12722502 (intron variant at *IL2RA* locus, OR = 0.80 and rs61816761 (stop-gained variant at *FLG* locus, OR = 1.24). The rs2230624_A variant is a singleton (all LD < 0.4; Fig. 2a) and is predicted by PROSITE database¹⁷ to disrupt a disulfide bond between Cys in positions 273 and 259 in the extracellular domain of the protein. CD30 is expressed on the surface of activated lymphocytes and eosinophils and has been implicated in activation, proliferation and apoptosis via NF κ B activation^{18–21}. p.Cys273Tyr has been reported to associate with reduced eosinophil count²² and reduced mosquito

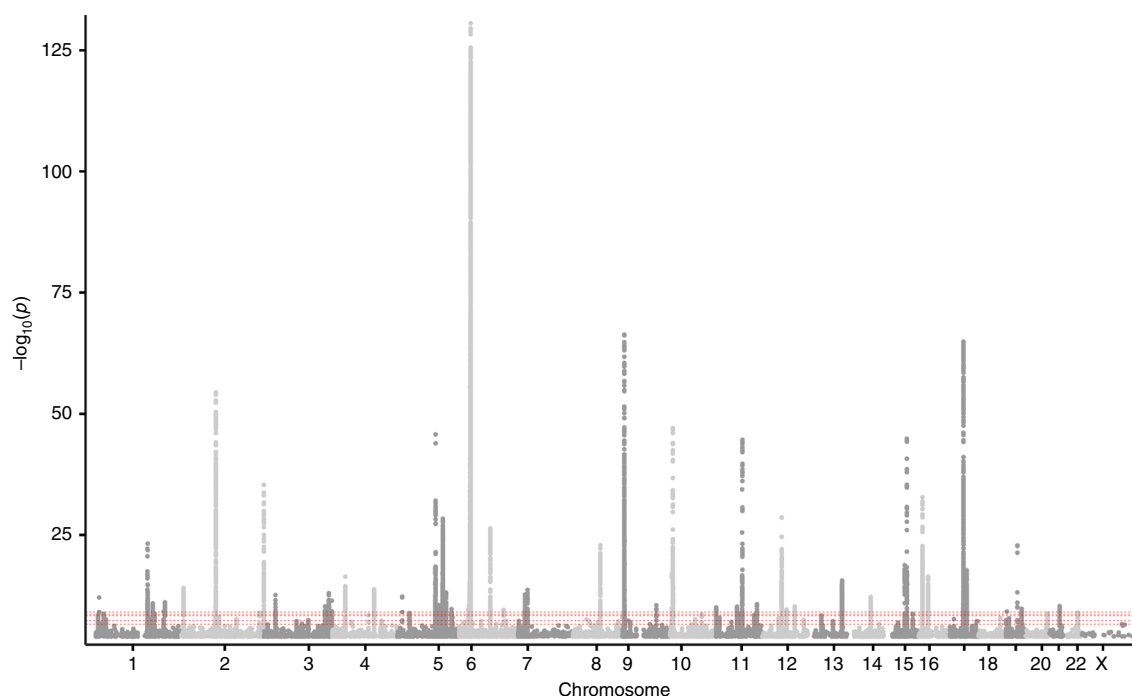


Fig. 1 Sequence variants associating with asthma. Manhattan plot for the Iceland-UK BB meta analyses of asthma ($N_{\text{cases}} = 69,189$). In all, 56 regions harbor genome-wide significant signals. Variants are plotted by chromosomal position (x-axis) and $-\log_{10}P$ values (y-axis). Dotted line indicate the different P value thresholds applied based on variant annotation. The adjusted significance thresholds are represented by horizontal dashed line from bottom to the top in the following order: 2.6×10^{-7} for variants with high impact ($N = 8,464$), 5.1×10^{-8} for variants with moderate impact ($N = 149,983$), 4.6×10^{-9} for low-impact variants ($N = 2,283,889$), 2.3×10^{-9} for other variants in DNase I hypersensitivity sites ($N = 3,913,058$) and 7.9×10^{-10} for all other variants ($N = 26,108,038$).

bite size²³. Increased soluble CD30 (sCD30) in serum has been associated with increased severity of asthma in children²⁴ and CD30 knock-out mice are protected against asthma²⁵. Therefore, we postulate that the missense p.Cys273Tyr in *TNFRSF8* that associates with decreased asthma risk reduces the function of CD30. In order to investigate this, we over-expressed wild-type (WT) CD30 or the p.Cys273Tyr variant in HeLa cells and compared levels of CD30 protein generated. CD30 is expressed as a precursor protein that undergoes post-translational modification, that turns it into the mature form of the protein^{26,27}. Lysates from cells expressing the p.Cys273Tyr variant had higher ratio of the precursor form to the mature form than that observed in cells expressing the WT CD30 ($P = 4.2 \times 10^{-7}$, two-tailed paired t -test; Fig. 2b and Supplementary Fig. 1). Moreover we observed both a lower cellular surface expression on p.Cys273Tyr than WT CD30 cells ($P = 7.4 \times 10^{-5}$, two-tailed paired t -test, Fig. 2c) and a significantly lower amount of sCD30 in the culture supernatants of cells expressing the p.Cys273Tyr variant ($P = 5.5 \times 10^{-4}$, two-tailed paired t -test, Fig. 2d). In line with this, cell surface expression of CD30 was significantly lower on in vitro stimulated peripheral blood mononuclear cells (PBMCs) derived from p.Cys273Tyr heterozygote carriers ($P = 1.5 \times 10^{-2}$, two-tailed Wilcoxon matched-pairs signed rank test; Fig. 2e). We recruited six p.Cys273Tyr homozygotes and matched non-carriers and measured CD30 surface expression both on stimulated PBMCs and Epstein-Barr virus transformed lymphoblasts. Surface expression of CD30 tended to be lower on EBV transformed lymphoblasts from homozygous carriers than from matched non-carriers ($P = 0.063$, two-tailed Wilcoxon matched-pairs signed rank test), whereas no difference was detected when comparing stimulated PBMCs from homozygous carriers to matched non-carriers ($P = 0.22$, two-tailed Wilcoxon matched-pairs signed rank test) (Supplementary Fig. 2). However, these results warrant

further investigation with larger number of homozygotes especially because of the inherent variance of the PBMC stimulation assay. Together our data suggest that the missense variant, leading to the disruption of a disulfide bond between Cys in positions 273 and 259, reduces trafficking of the protein to the cell surface. Furthermore, reduced levels of sCD30 were observed in the cell culture supernatant of in vitro stimulated PBMCs from heterozygote carriers ($P = 3.1 \times 10^{-5}$, two-tailed Wilcoxon matched-pairs signed rank test) compared with those from non-carriers (Fig. 2f). This is consistent with reduced surface-expression although an effect of the variant on the CD30 shedding itself cannot be excluded²⁸.

Gain of function variant in *TGFBRI*. Another low frequency variant of special interest, is a 3 prime UTR variant, rs41283642_T (EAF = 1.9% (ICE)/3.4% (UK)), in *TGFBRI* that associated with reduced asthma risk ($P = 2.16 \times 10^{-10}$, OR = 0.89). eQTL analysis showed that in blood the variant associated with 15.2% increase in *TGFBRI* expression per allele ($P = 7.24 \times 10^{-101}$, effect = 1.18 SD, which is the most significant eQTL for *TGFBRI*; Fig. 3a, b, Supplementary Data 6). rs41283642_T changes a recognition site for microRNA (miR) miR-142-3p (Fig. 3c). miR-142-3p has previously been shown by others²⁹ to bind to the 3 prime UTR site in *TGFBRI* and thereby repress *TGFBRI* expression at both RNA and protein levels. Interestingly, miR-142-3p is one of only three miRs reported to have an increased expression in severe asthmatic lungs³⁰. Together these data suggest that rs41283642_T reduces the binding of miR-142-3p to *TGFBRI* transcripts leading to increased *TGFBRI* expression and through that protection against asthma.

Unreported common asthma risk variants. Four of the 16 previously unreported common variants are missense, rs2228552_T in

Table 1 Previously unreported loci associated with asthma in Iceland - UK biobank meta analyses.									
Lead SNP	chr:Position	Allele (EA/OA)	EAF % (ICE/UK BB)	Gene	Coding effect	Coding change	OR (95% CI)	P	P _{het}
rs2230624 ^a	chr1:12115601	A/G	12.1/15	TNFRSF8	Missense	Cys273Tyr/Cys162Tyr	0.82 (0.78-0.87)	8.3E-13	0.14
rs2228552 ^a	chr1:31699894	T/G	68.3/64.3	COL16A1	Missense	Thr62Lys/Thr52Met	1.04 (1.02-1.05)	2.6E-08	0.86
rs22296618	chr1:198697103	G/A	14.4/13.5	PIPRC ^a	Downstream	—	0.94 (0.93-0.96)	6.9E-10	0.36
rs7626218 ^a	chr3:17134250	T/A	41.1/39.5	TBLXRT ^a	Intron	—	0.96 (0.95-0.97)	5.9E-11	0.14
rs34712979	chr4:105897896	A/G	24.4/25.8	NPNT	Splice region	—	1.04 (1.03-1.06)	3.9E-09	0.85
rs11746314 ^a	chr5:157325949	G/A	3.6/5.8	CYFIP2 ^a	Intron	—	1.09 (1.06-1.12)	2.1E-10	0.832
rs3813308	chr5:119355086	G/C	47.2/43.5	TNFAIP8	5 prime UTR	—	1.04 (1.03-1.06)	2.8E-11	0.8
rs1800797	chr7:22726602	G/A	52.2/58.0	IL6 ^a	Upstream	—	1.04 (1.03-1.05)	2.9E-10	0.66
rs34173062	chr8:144103704	A/G	66/72	SHARPIN	Missense	Ser17Phe	1.08 (1.05-1.10)	1.0E-08	0.28
rs41283642 ^a	chr9:99153605	T/C	19/3.4	TGFB1	3 prime UTR	—	0.89 (0.86-0.83)	2.16E-10	0.05
rs12788104 ^b	chr11:1129831	G/A	73.4/68.7	MUC2 ^c	Intergenic	—	1.04 (1.03-1.06)	1.7E-10	0.09
rs174562 ^a	chr11:61817672	G/A	38.8/34.6	FADS1 ^c	Upstream	—	0.96 (0.95-0.97)	6.3E-11	0.31
rs7961712 ^a	chr12:94211187	A/G	86.3/85.0	PLXNCT ^c	Intron	—	1.06 (1.04-1.08)	5.8E-11	0.45
rs34939984	chr16:27234391	T/C	40.1/35.9	NSMCE1 ^c	Intron	—	1.04 (1.03-1.06)	5.2E-11	0.41
rs3024664	chr16:27360103	C/T	93.5/93.9	IL4R ^c	Intron	—	1.12 (1.09-1.15)	4.4E-17	0.23
rs6498021	chr16:27403057	G/T	85.3/86.0	IL21R ^c	Intron	—	0.95 (0.93-0.97)	2.8E-08	0.30
rs179771	chr16:27406423	C/G	53.9/48.8	IL21R ^c	Intron	—	1.04 (1.03-1.05)	3.23E-09	0.08
rs117552144	chr19:3136093	T/C	5.3/6.7	GNAI5	5 prime UTR	—	1.09 (1.06-1.11)	6.7E-10	0.74
rs8103278 ^a	chr19:45867123	A/G	31.3/35.3	SYMPK ^c	Upstream	—	0.96 (0.95-0.97)	1.8E-10	0.84

Results are shown for the combined sample-sets of Iceland and the UK BB. P_{het} represents the statistical heterogeneity between the two sample-sets. Gene indicates the most likely candidate gene at the locus as described in Supplementary Data 7
EA effect allele, OA other allele, EAF effect allele frequency ICE Iceland, UK BB UK biobank, OR odds ratio
^a loci reported in Johansson et al.¹⁵, while this paper was in review
^b loci reported in Shrine et al.¹⁴, while this paper was in review
^c The closest gene is indicated for loci where our analysis do not pinpoint the most likely gene candidate

COL16A1 (Thr62Lys/Thr52Met) and rs34173062_A in SHARPIN (Ser17Phe) (Table 1) or highly correlated ($r^2 > 0.8$) with coding variants in NSMCE1 (rs34939984) or SYMPK, SIX5 and BHMGI (rs8103278) (Supplementary Data 7). COL16A1 encodes the α -chain of type XVI collagen, involved in integrity of the extracellular matrix in association with fibril forming collagens type I and II³¹. Increased local collagen type I production has been reported in asthma patients, possibly mediated by TGF β secreting eosinophils³². The rs2228552_T associated with increased asthma risk ($P = 2.6 \times 10^{-8}$, OR = 1.04) and was highly correlated ($r^2 > 0.8$) with the top eQTL for increased COL16A1 expression in fibroblasts (Supplementary Data 8). Thus the COL16A1 association with asthma may be through its role in airway remodeling. SHARPIN encodes a component of the LUBAC complex, which plays a role in NF κ B activation and regulation of inflammation³³. The SHARPIN missense variant, Ser17Phe, associated with increased risk of asthma ($P = 1.00 \times 10^{-8}$, OR = 1.08) and increased eosinophil count in blood²². Furthermore, homozygous loss of function mutations in Sharpin mice induce extensive eosinophilic inflammation in multiple organs, including lung, esophagus and skin^{34,35}.

We found four independent signals at the IL4 receptor alpha (IL4R α)/IL21R locus, including the intronic variant rs6498021_G in IL21R that was highly correlated ($r^2 > 0.8$) with the top eQTL signal for decreased IL21R expression in blood (Supplementary Data 6). Polymorphisms at the IL4R α locus have been associated with asthma in candidate gene studies but not in GWAS³⁶⁻³⁹. Biologicals targeting IL4R α show promise in the treatment of persistent asthma⁴⁰ but the role of IL21R in asthma has been less studied. Other loci of interest include NPNT, TNFAIP8, FADS1, GNAI5, and IL6. Taken together, based on analysis of coding variants, eQTLs in different tissues, enhancer and promoter signals in CD4⁺ Th cells we found evidence of a single candidate gene for 8 out of the 19 previously unreported asthma variants (Supplementary Data 9).

Signals at known asthma loci. Among the 24 previously identified loci were IL1RL1/IL18R1, GATA3, TLR1, TSLP, HLA, IL4/IL5/IL13, RORA, GSDMB, SMAD3, and IL33^{9,41}. Our top signal at the SMAD3 locus, rs72743461_A, correlated with reported asthma signals^{16,42,43} (Supplementary Data 2) but we further identified 3 secondary signals (Supplementary Data 3), including rs117683492_A, that associated with increased asthma risk and was the top eQTL reducing SMAD3 expression in blood (Supplementary Tables 1 and 6). SMAD3 is a signaling molecule downstream of TGF β R and Smad3 KO mice show defective TGF β -mediated repression of cytokine production and cell proliferation⁴⁴. Together, the TGFB1 3 prime UTR variant rs41283642_T described above and rs117683492_A at the SMAD3 locus point to increased risk of asthma in individuals with defective TGF β R1 signaling in line with its known role in immunosuppression⁴⁵. rs6926894 is the top variant in the HLA region ($P = 2.5 \times 10^{-126}$, OR = 1.16). Given the complicated LD in the HLA region we tested association of HLA alleles with asthma in the Icelandic dataset finding strongest association with HLA-DRB1*04 ($P = 2.06 \times 10^{-25}$) that largely explains the rs6926894 association (Supplementary Data 10).

Association of asthma variants with eosinophil count. We have previously used the well-established link between eosinophilic airway inflammation and asthma to discover common sequence variants in IL1RL1 and IL33 that associate strongly with blood eosinophil counts and risk of asthma⁸. Therefore, we tested the association of our 88 asthma variants with eosinophil blood counts in Iceland and UK biobank. Of those, 69 associate significantly with eosinophil count after adjusting for the number of

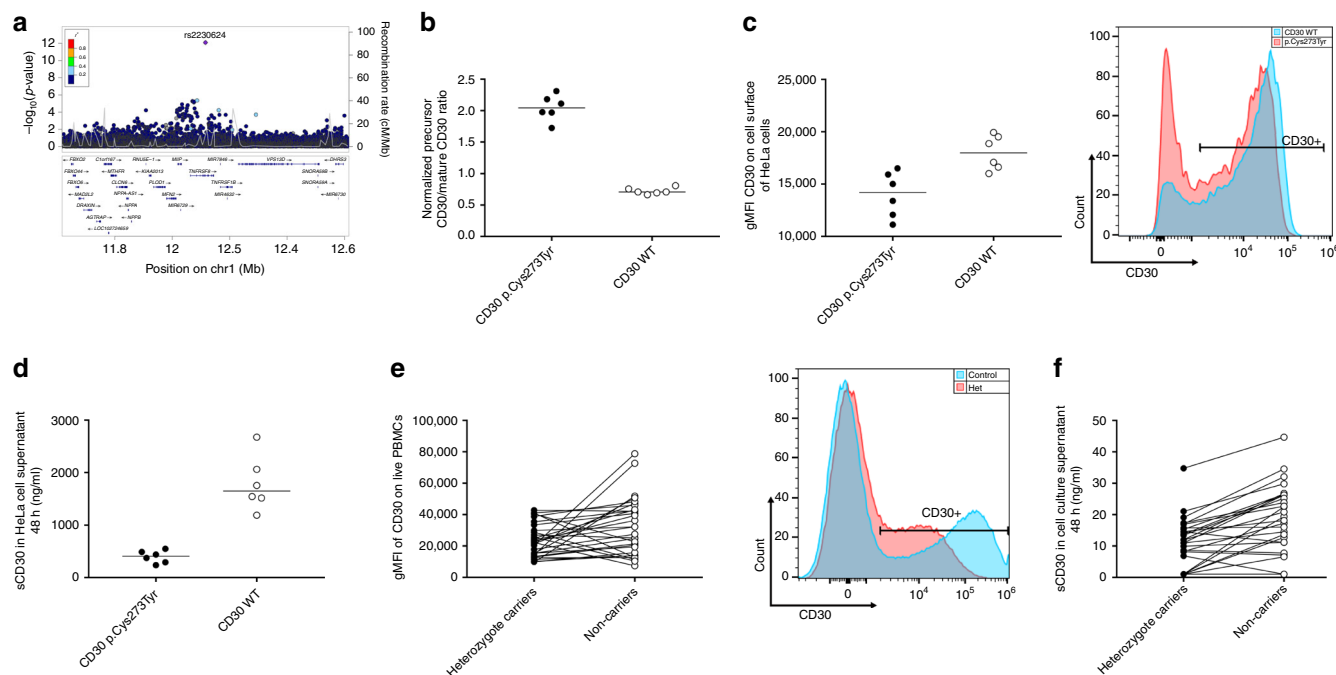


Fig. 2 Variant associating with reduced asthma risk affects protein expression and shedding of CD30. **a** Locus plots showing association with asthma where the lead variant p.Cys273Tyr (rs2230624) is colored in purple. Other variants are colored by degree of correlation (r^2) with the lead variant. **b** Protein simple WES analysis of CD30 expression in cell lysates from HeLa cells overexpressing CD30 wild-type or variant p.Cys273Tyr CD30; graph showing ratio of CD30 precursor/mature intensity. **c** Surface expression of CD30 on HeLa cells over-expressing wild-type or p.Cys273Tyr CD30 measured by flow cytometry, displayed as geometric mean of fluorescence intensity (gMFI). Histogram showing a representative CD30 surface expression p. Cys273Tyr (red) and CD30 WT (blue). **d** sCD30 levels (ng/ml) in cell culture supernatant from HeLa cells over-expressing wild-type or p.Cys273Tyr CD30. **e** Surface expression of CD30 on stimulated PBMCs from heterozygous p.Cys273Tyr carriers and age and gender matched non-carriers, measured by flow cytometry, displayed as gMFI. Histogram showing a representative CD30 surface expression from one pair of PBMCs in heterozygous CD30 p.Cys273Tyr carrier (red) and age and gender matched non-carrier (blue). **f** sCD30 levels (ng/ml) in cell culture supernatant of PBMCs from p.Cys273Tyr heterozygotes and non-carriers. The dots in plots **b–d** represent individual experiments. The dots in plots **e** and **f** represent individual donors. Lines in panel **b–d** indicate median level. In plots **e** and **f** the line between carriers and non-carriers indicated age and gender matched pairs. Two-tailed Wilcoxon matched-pairs signed rank test was used to test for significant differences in PBMCs. Two-tailed paired t -test was used to test for significant differences in HeLa cells. Source data are provided as a Source Data file.

variants tested, with the asthma risk allele associating with higher eosinophil count for 67 of the variants ($P < 0.05/93 = 5.4 \times 10^{-4}$; Supplementary Data 11). Among the 20 variants that did not associate with eosinophil count and deviate from the overall correlation between asthma risk and eosinophil count are variants at the filaggrin family member *FLG/HRNR* locus, suggesting their role in non-eosinophilic asthma (Supplementary Fig. 3).

Effect of asthma variants on different asthma sub-phenotypes.

Given the known heterogeneity of asthma we tested the effect of the asthma variants on three asthma sub-phenotypes, allergic asthma (AA), early (EOA) and late (LOA) onset asthma. Overall, the effects of the 88 asthma variants were similar on the sub-phenotypes although for some of the variants (e.g at *FLG/HRNR* locus) the effect was stronger on EOA than LOA (Fig. 4a). Further, the effect of *IL33* (rs149045797) variant was largest on EOA ($OR = 0.53$, $P = 3.8 \times 10^{-9}$) followed by LOA ($OR = 0.71$, $P = 5.1 \times 10^{-9}$) with smallest effect on AA ($OR = 0.93$, $P = 0.70$; Fig. 4 and supplementary Data 12). This variant is fully correlated ($r^2 = 1$) with a loss of function variant (rs146597587) previously reported to associate with reduced *IL33* mRNA expression and asthma risk and in line with the results reported here, the largest effect was observed in young children who were hospitalized at least 4 times before the age of 6 years due to asthma exacerbation ($OR = 0.24$, $P = 0.04$)⁹. However, 58% of those young children had atopic diagnosis before 6 years of age⁴⁶ indicating that the *IL33* association is driven by severe asthma rather than non-AA

exclusively. We note that we lack information on AA for a large part of our samples, especially in the UK biobank samples where only 542 (1% of all asthma cases) have this diagnosis and there is an overlap both with the early and late onset asthma cases. Therefore, the *IL33* association with asthma severity in AA and non-AA warrants further investigation.

Conversely, the effect of the *FLG* variant was stronger on AA than on LOA and the effect of a variant at the *GATA3* locus was stronger on AA than on either EOA or LOA (Fig. 4c). These results point to a partly distinct genetic architecture behind the three asthma phenotypes tested in our analysis in line with the clinical heterogeneity. Our results show that certain sequence variants have stronger effect in EOA than LOA indicating that genetics may play a larger role in EOA.

Genetic correlation between asthma and allergic phenotypes.

Recently, 136 variants were reported to associate with a combined allergic phenotype (asthma, hay fever and/or eczema)¹¹. The overall effects of these 136 allergy variants correlated well with the effects in our asthma meta-analysis, although several markers (e.g. *TSLP*, *GATA3*, *SMARCE1*) demonstrated weak correlation (Supplementary Fig. 4). This is in line with reported genetic correlation between asthma and allergic diseases (hay fever/allergic rhinitis or eczema)⁶.

We also tested the 88 independent asthma variants for association with different allergic phenotypes, (Supplementary Table 1). Interestingly, the effects of these variants on asthma

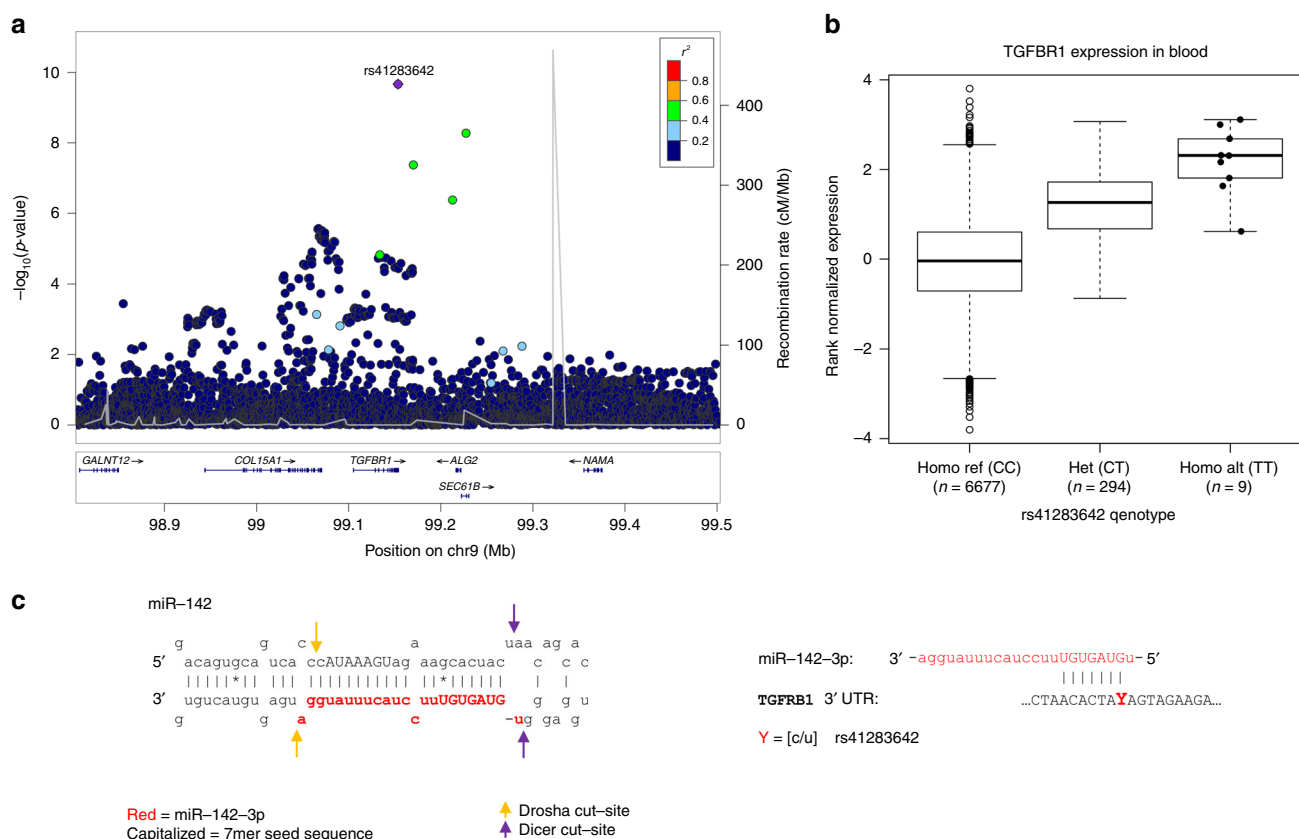


Fig. 3 Variant in the 3'UTR of TGFBR1 increases TGFBR1 expression in blood and associates with reduced asthma risk. a Locus plot showing association with asthma where the lead variant rs41283642 is colored in purple. Other variants are colored by degree of correlation (r^2) with the lead variant. **b** Box plot showing the expression of TGFBR1 in blood of rs41283642_T non-carriers (CC, $N = 6677$), heterozygotes (CT, $N = 294$) and homozygotes (TT, $N = 9$) based on RNA-Seq data; $P = 7.24 \times 10^{-101}$, effect = 1.18 S.D. corresponding to 15.2% increased expression of TGFBR1 per allele. The bottom and top of the boxes correspond to the 25th (Q1) and 75th (Q3) percentiles, the line inside the box corresponds to the median, and the whiskers are located at max(min(Expression), Q1 - 1.5 IQR) and min(max(Expression), Q3 + 1.5 IQR), respectively (where IQR is the interquartile range = Q3 - Q1). **c** Schematic representation of the primary miR-142 transcript, derived from MIR142 located on chromosome 17q22. The 3' end of miR-142 forms miR-142-3p which is known to targets multiple mRNA transcripts including TGFBR1. The predicted binding site for miR-142-3p at the 3'UTR of TGFBR1 is shown wherein Y represents the location of the rs41283642. The reference allele is C, forming a 7mer perfect complementary sequence to the seed sequence of miR-142-3p whereas the protective asthma T_allele (U in RNA), disrupts the 7mer binding site. Source data are provided as a Source Data file.

correlated significantly with their effects on allergic rhinitis (AR; $P = 3.1 \times 10^{-13}$, $r^2 = 0.68$; Supplementary Fig. 5a) but not with atopic dermatitis (AD; $P = 9.80 \times 10^{-2}$, $r^2 = 0.18$; Supplementary Fig. 5b). Out of the 88 asthma loci, two variants at the same locus, a loss of function variant in *FLG* (rs61816761) and intergenic variant (rs12123821, close to *HRNR*) had greatest effect on AD (Supplementary Data 12). Further, significant correlation was observed between effects of the asthma variants on asthma and on nasal polyps (NP) as well as on chronic rhinosinusitis (CRS) with NP (CRS_{NP}) but not with CRS without NP (CRS_{NP}) in agreement with differences in the pathogenesis of these closely related phenotypes⁴⁷ (Supplementary Fig. 5c–f). Variants at *IL33*, *GATA3*, *BACH2*, *HLA* and *TSLP* loci significantly associated with >3 asthma/allergic phenotypes ($(P < 0.05/(88 \times 9) = 5.7 \times 10^{-5})$, adjusted for number of variants and phenotypes tested; Supplementary Data 12), most of which are important for generation and function of ILC2 and Th2⁴⁸. Taken together, the correlation of effects of the 88 asthma variants was stronger with respiratory allergic phenotypes of T2 than with AD or the Th1 driven CRS_{NP}. However, we observed substantial overall genetic correlation between asthma and AD as well as between asthma and allergic rhinitis indicating that other variants than the 88 reported here play a significant role in the genetic link between

asthma and AD (Supplementary Table 2). Since this is a retrospective study, not designed to study co-morbidities of asthma and allergic diseases, we cannot exclude that some of the overlapping association of certain sequence variants with e.g. asthma and AD might be due to overlap of the two phenotypes. However, only 980 out of 16,247 asthma cases in Iceland had the AD diagnosis (6%) while AD was hardly reported as a single disease in UK biobank, making it impossible to accurately define asthmatics without AD.

Pathway analysis. Pathway analysis was performed using DEPICT⁴⁹ to search for the biological connectivity between the asthma association signals. DEPICT prioritized 71 (FDR < 0.05) known genes (Supplementary Data 13) and identified 787 significantly (FRD < 0.01) enriched gene sets, the majority involved in T cell biology, mainly implicating CD4⁺ T cells, regulation of their activation, responses and physiology (Supplementary Data 14).

In summary, our study considerably expands the number of asthma susceptibility loci and confirms many previous findings. The results highlight the role of Th cells in asthma in line with imbalanced T cell regulation reported to play a critical role in

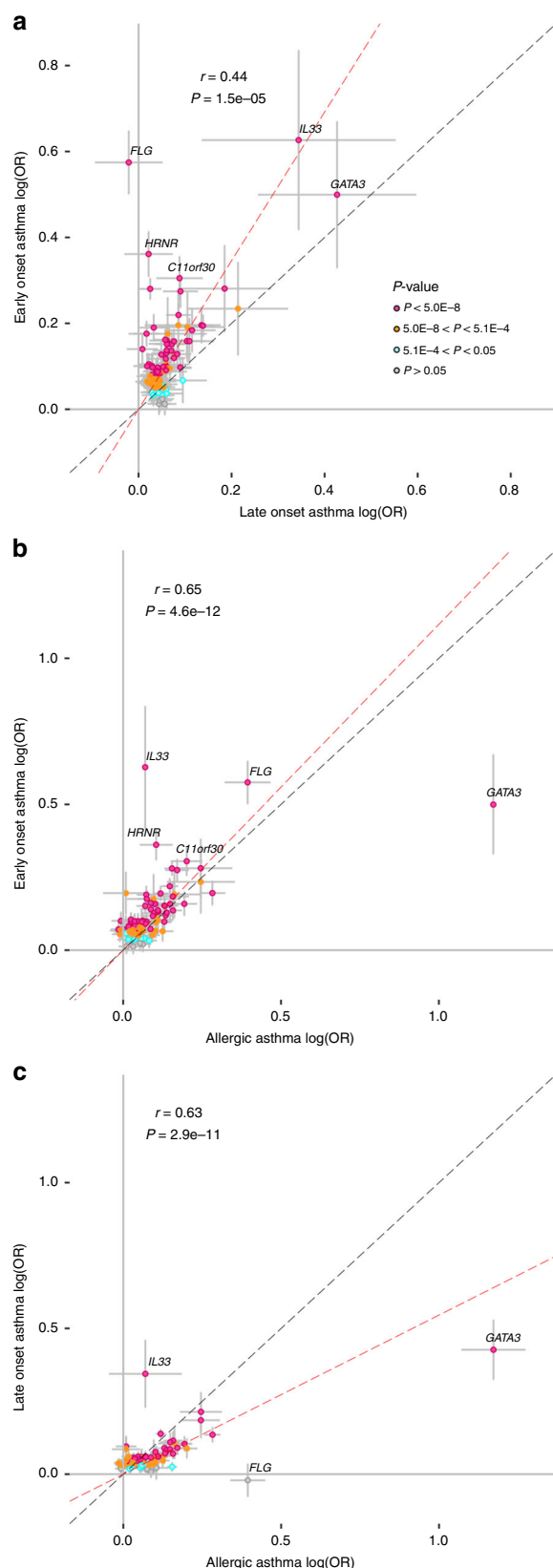


Fig. 4 Effect of asthma associating variants in asthma sub-phenotypes.

Early onset asthma (EOA), Late Onset Asthma (LOA) and Allergic Asthma (AA). The x-axis and the y-axis show the logarithm of the estimated odds ratios for **a** EOA and LOA, **b** AA and EOA, and **c** AA vs LOA, respectively. All effects are shown for the asthma risk increasing allele based on meta-analysis of the Icelandic and the UK BB sample sets. Error bars represent 95% confidence intervals. The red line represents results from a simple linear regression through the origin using MAF (1-MAF) as weights and the gray line indicates the reference line with slope = 1. The weighted correlation coefficients (r) and P values (t -test) are shown in the graphs. Source data are provided as a Source Data file.

Methods

Study sample sets for asthma, allergy phenotypes, and eosinophil count.

Asthma in the UK biobank was defined as ICD10 diagnoses in fields 41202 or 41204, including anyone of J45.0, J45.1, J45.8, J45.9, and J46 and/or self-reported by the non-cancer illness code, self-reported during verbal interview (data-field 20002) with a code for asthma (1111).

Icelandic asthma patients over 18 years of age were recruited who attended an asthma clinic or emergency room at the National University Hospital of Iceland or the Icelandic Medical Center (Laeknasetrid) during the years 1977 to 2017. Asthma diagnosis was based on a combination of physician's diagnosis and ICD10 diagnosis, including anyone of J45.0, J45.1, J45.8, J45.9, and J46 and/or self-reported by a positive reply to the question: has a doctor confirmed your asthma diagnosis. Atopy status determined by skin prick testing and age of onset was available for part of the asthma cohort both in Iceland and UK biobank. Early onset was defined as first diagnosis ≤ 17 years of age and late onset as first diagnosis > 17 years of age. Late onset asthma diagnosis from Iceland contains both where the first diagnosis was at or after 18 years of age and where the first diagnosis is unknown. In total we had 16,247 asthma cases and 346,486 controls from Iceland and 52,947 cases and 355,713 controls from the UK biobank. Estimating the effective number of cases as $2 \times N_a \times N_c / (N_a + N_c) / GC$, where N_a and N_c are the number of cases and controls, respectively, and GC is the genomic adjustment factor, estimated using LD-score regression¹², we can calculate the effective sample size for Iceland to be $N_{eff} = 22,689$ and UK biobank to be $N_{eff} = 84,396$. $GC = 1.368$ in Iceland and $GC = 1.092$ in UK biobank. Characteristics of Icelandic and UK biobank asthma cohorts is further given in Supplementary Table 3.

Allergic rhinitis combines doctoral diagnosis of allergic rhinitis from Iceland and a questionnaire data from UK biobank on hayfever or allergic rhinitis (Non-cancer-illness code 1387). As only 42 individuals had the ICD10 code for Atopic dermatitis (L20) in UK biobank we ran meta-analysis with our Icelandic list ($N = 8325$) derived both from physician's diagnosis and ICD10 code (L20) together with published meta-analysis on AD⁵² downloaded from the GRASP database [<https://grasp.nhlbi.nih.gov/FullResults.aspx>], to study this phenotype. Other allergy diagnosis were based on ICD10 codes from UK biobank and either physician's diagnosis or ICD10 codes from Iceland. The allergy phenotypes used were: Nasal polyps (ICD10:J33), chronic sinusitis (ICD10:J32), chronic sinusitis with nasal polyps (combined: ICD10:J32 and J33), and chronic sinusitis without nasal polyps (ICD10:J32 without ICD10:J33). Number of genotyped individuals in each cohort are listed in Supplementary table 1. Icelandic controls were participants from various deCODE genetics programs.

We obtained eosinophil counts from three of the largest laboratories in Iceland (measurements performed between the years 1993 and 2015). The circulating eosinophil counts were standardized to a standard normal distribution using quantile-quantile standardization and then adjusted for sex, year of birth, and age at measurement, as previously described^{8,9}. A total of 251,307 Icelanders with eosinophil counts were included in the study. In the UK biobank we had eosinophil counts for a total of 396,822 individuals that were adjusted for sex, year of birth, age at measurement, and the first 40 principal components, then inverse normal transformed.

All participating individuals who donated blood signed informed consent. The personal identities of participants were encrypted using a third-party system approved and monitored by the Icelandic Data Protection Authority⁵³. The study was approved by the National Bioethics Committee in Iceland (Approval no. VSN 14-099).

Whole-genome sequencing and imputation. The GWASs in Iceland were performed with 32.5 million markers identified through whole-genome sequencing of 15,520 Icelanders to an average genome-wide coverage of 34X and subsequently imputed into 151,677 chip-typed individuals, as well as their first and second degree relatives. The imputation has been extensively described in recent publications⁵⁴. Genotyping of UK biobank samples was performed using a custom-made Affymetrix chip, UK BiLEVE Axiom⁵⁵, and with the Affymetrix UK Biobank Axiom array⁵⁶. Imputation was performed by the Wellcome Trust Centre for Human Genetics, using the Haplotype Reference Consortium (HRC) and the

asthma pathogenesis^{50,51}. We show evidence that two low frequency variants in *TNFRSF8* and *TGFBR1* associate with decreased asthma risk through loss of function and gain of function, respectively. Other susceptibility loci point to genes involved in inflammation and airway remodeling.

UK10K haplotype resources. This yields a total of 96 million imputed variants, however only 27 million variants imputed using the HRC reference set passed the quality filters used in our study.

Association testing and meta-analysis of disease phenotypes. Logistic regression was used to test for association between variants and disease, assuming a multiplicative model, treating disease status as the response and expected genotype counts from imputation as covariates. We used LD score regression¹² to account for inflation in test statistics due to cryptic relatedness and stratification. We chose to include related individuals in the analysis as removing related individuals in the heavily genotyped Icelandic population would lead to the removal of most participants. Including related individuals while accounting for genomic control has proven to be a robust method^{9,57,58} that does not create false positives as observed in the QQ plots (Supplementary Fig. 6; especially stratified on frequency). Software developed at deCODE genetics⁵⁷ was used to test for association in both populations. Sex, age, and county of origin were included in the logistic regression for the Icelandic dataset, and sex, age and the first 40 principal components for the UK biobank dataset. Imputation information criteria was set so that variants below 0.8 were excluded from the analysis.

Variants were mapped to NCBI Build38 positions and subsequently variants in the UK biobank imputation dataset were matched to the variants in the Icelandic dataset based on allele variation. Mantel-Haenszel model⁵⁹ was used to combine the results from the different study groups in which the groups were assumed to have a common OR but were allowed to have different population frequencies for alleles and genotypes. A likelihood ratio test was used to test heterogeneity by comparing the null hypothesis of the effect being the same in all populations to the alternative hypothesis of each population having a different effect. Heterogeneity was quantified by I^2 statistics which lies between 0 and 100% and describes the proportion of total variation in study estimates that is due to heterogeneity.

Variants were split into five classes based on their genome annotation and associations were considered significant if the p -value in the combined dataset was below a weighted genome-wide significance threshold based on variant annotation. With 32,463,443 sequence variants being tested the weights given in Sveinbjornsson et al. were rescaled to control the family-wise error rate (FWER)⁶⁰. The adjusted significance thresholds are 2.6×10^{-7} for variants with high impact ($N = 8,464$), 5.1×10^{-8} for variants with moderate impact ($N = 149,983$), 4.6×10^{-9} for low-impact variants ($N = 2,283,889$), 2.3×10^{-9} for other variants in DNase I hypersensitivity sites ($N = 3,913,058$) and 7.9×10^{-10} for all other variants ($N = 26,108,038$). Approximate conditional analyses (COJO), implemented in the GCTA software⁶¹ were performed on lead variants, defined by lowest P value at each genomic region (locus), to identify possible secondary signals. LD between variants was estimated using a set of 8700 whole-genome -sequenced Icelandic individuals. The analysis was restricted to variants within 1 Mb from the index variants and that were present in both the Icelandic and the UK biobank datasets. Based on the number of variants tested within the 55 loci (excluding the HLA region, $N = 3,993,179$), we chose a conservative P value threshold of $<5 \times 10^{-8}$ for secondary signals.

Association testing and meta-analysis of eosinophil counts. Linear mixed model implemented by BOLT-LMM⁶² was used to test for association between sequence variants and eosinophil counts, assuming an additive genetic model. We assume that the quantitative measurements follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance-covariance matrix proportional to the kinship matrix⁶³. We used LD score regression¹² to account for inflation in test statistics due to cryptic relatedness and stratification. To combine the deCODE and UK biobank results, we used a fixed-effect inverse variance method based on effect estimates and standard errors⁵⁹. We used a likelihood-ratio test to compute all P values.

Genetic correlation. The cross-trait LD score regression method¹² was used to estimate the genetic correlation between pairs of traits using the summary statistics from the Icelandic and UK biobank datasets. Results for about 1.2 million variants, well imputed in both datasets, were used in this analysis. and pre-computed LD scores for European populations (downloaded from [https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2]) were used for LD information. Genetic correlation between Icelandic GWAS summary statistic was calculated for one trait and the UK biobank GWAS summary statistic for the other traits, and the vice versa, and those results subsequently meta-analyzed, to avoid bias due to overlapping samples. For AD, we also calculated genetic correlation based on the external meta-analysis dataset downloaded from the GRASP database [<https://grasp.nhlbi.nih.gov/FullResults.aspx>].

Cis-eQTL analysis. Cis-eQTL effects were analyzed using RNA sequencing data from Icelandic samples. The generation of poly(A) + cDNA sequencing libraries, RNA sequencing, and data processing were carried out as described before^{64,65}. Two tissue types were available for this analysis: whole blood and adipose tissue. In total, whole blood from 7,007 individuals and adipose tissue from 686 individuals were used. To estimate the association between sequence variant and gene expression, a generalized linear regression assuming additive genetic effect was

used on rank-transformed gene expression estimates. GTEx eQTL's for relevant tissue types (whole-blood, lung, esophagus mucosa, spleen, lymphocytes and fibroblasts) were intersected with lead GWAS association variants, and those in LD ($r^2 > 0.80$) with the lead variant. We then report eQTLs where the top eQTL signal is in LD with lead GWAS association variants.

Enhancer and transcription start site analysis. Variants in linkage disequilibrium (LD) with the lead variants at the previously unreported loci were identified on the basis of in-house genotype data using $r^2 > 0.8$ for pairwise comparison of the nearest 100,000 variants to define an LD class. These variants were then annotated, see results in Supplementary Data 9. Chromatin states for CD4 + T-cells were obtained through ChromHMM analyses of available epigenome data from NIH Roadmap Epigenome Mapping Consortium for 11 histone marks analyzed by ChIP-seq, together with open chromatin regions analysed by DNase-seq, integrated into 25 discrete chromatin states⁶⁶ downloaded from [<http://www.roadmapepigenomics.org>]. The association variants were then annotated for chromatin states involving enhancers (EnhA1, A2, W1, or W2) or DHS sites (DNA hypersensitivity sites, reflecting open chromatin configuration). To identify enhancer-gene targets, we made use of the joint effect of multiple enhancers (JEME) resource⁶⁷. Variants within or proximal (± 25 bp) to transcriptional start sites were annotated on the basis of CAGE sequencing data derived from the Fantom5 project⁶⁸ downloaded from [fantom.gsc.riken.jp/5/data/]. 3'UTR variants in predicted miRNA target sites were annotated by making use of data derived from Targetscan v7.2⁶⁹ targetscan.org.

Generation of CD30 variants. Full-length CD30 cDNA (NM_001243) in pCMV6-Entry Myc-DDK tagged mammalian expression vector was obtained from Origene (RC219819). Cys273Tyr mutant plasmid was generated using Q5 Site-directed mutagenesis kit (New England BioLabs, E0554S) with mutagenesis primers F-5'-AAGACGCCATATGCATGGAAC-3' and R-5'-CTCCACAAGGTCATCTCG-3' (Supplementary Table 4). Plasmids were transformed into NEB Stable Competent E.coli (New England BioLabs, C3040H) and spread on LB agar plates containing 25 μ g/ml kanamycin. Colonies were expanded in LB medium containing 25 μ g/ml kanamycin. Plasmids were purified using Qiagen plasmid maxi kit (Qiagen, 12163), following the manufacturer's protocol. The sequences of WT and Cys273Tyr plasmids were confirmed by Sanger sequencing.

Over expression of CD30 and CD30 Cys273Tyr in Hela cells. One day prior to transfection 0.1×10^6 /ml Hela cells (Public Health England 93021013) were seeded in DMEM medium (ThermoFisher 11995-065) supplemented with 10% fetal bovine serum (ThermoFisher 10500-064) and 50 units/ml penicillin and 50 μ g/ml streptomycin (ThermoFisher 15070-063). Cells were incubated at 37 °C with 5% CO₂ in a humidified incubator.

After 24 hours cells were transfected using FuGENE®HD reagent (Promega E2312) following manufacture's protocol. In short DNA was diluted in OptiMem medium (ThermoFisher 31985-047) and FuGENE®HD reagent was added at a FuGENE®HD reagent: DNA ratio 3:1.

48 hours after transfection media was removed from cells and they were washed once with PBS. Cells were dislodged from wells by incubation for 5 min at 37 °C using PBS + 2 μ M EDTA. Cells were collected and spun down at 300 $\times g$ for 5 min, resuspended in 1 ml PBS + 2% FBS and spun down again. Cell pellet was snap frozen and kept in -80 °C for subsequent lysis preparation.

Hela lysis preparation. Cells were lysed using 100 μ l of RIPA buffer with 1:100 Halt protease and inhibitor cocktail (Thermo Scientific 78430). Lysates were kept on ice for 10 min and subsequently spun down at 4 °C for 15 min at 14,000 $\times g$. Total protein concentration in lysates was estimated using the Pierce BCA protein assay kit (ThermoFisher 23227).

EBV cell culture. EBV transformed lymphoblasts from homozygous carrier and non-carrier of Cys273Tyr were cultured in RPMI1640 (ThermoFisher 61870-036) supplemented with 10% fetal bovine serum (ThermoFisher 10500-064), 50 units/ml penicillin, 50 μ g/mL streptomycin (ThermoFisher 15070-063) and 20 mM HEPES (ThermoFisher 15630-056). Cells were incubated at 37 °C and 5% CO₂ in a humidified incubator. The cells were seeded at 5×10^5 cells/ml in fresh media 2 days prior to harvesting.

Simple Western by size quantification of CD30 signal in Hela. Total of 0.05 mg/ml Hela lysate was run in WES from Protein Simple following manufacture's protocol using WES 12-230 kDa separation module (Biotechnie SM-W002-1). Antibody used: Purified mouse anti human CD30 (clone BerH8) (BD Biosciences cat: 555827) 1:100 dilution. Precursor CD30 signal was quantified as a 143 kDa peak and mature CD30 signal as a 207-kDa peak. As a normalization for total protein we used WES Biotin detection module (Biotechnie DM-004) and followed manufacture's protocol. CD30 precursor and mature signals were normalized to total protein.

Flow cytometry analysis on Hela, EBV, and stimulated PBMCs. EBV cells were harvested, counted and diluted to 1×10^6 cell/ml plated in a 96 well V bottom plate. Cells were then stained with primary antibodies against PE/Cy7 anti-human CD30 (Biolegend 333918), clone BY88 reported to bind at similar amino acid residues as Ber-H6⁷⁰ that has been shown to bind epitope at the N-terminus of the peptide chain up to amino acid residue 93⁷¹ thus unaffected by a possible effect of our variant at amino acid 273. The cells were analyzed for expression of CD30 by FACS (gating strategies shown in Supplementary Fig. 7). Cryopreserved PBMC were thawed and plated for culture in RPMI1640 (ThermoFisher 61870-036) supplemented with 10% fetal bovine serum (ThermoFisher 10500-064), 50 units/mL penicillin, 50 µg/mL streptomycin (ThermoFisher 15070-063), 20 mM HEPES (ThermoFisher 15630056) and 30 IU/ml IL-2 (R&D Systems AFL202) in a 24-well TCR plate and rested overnight. Cells were then stimulated for 48 hours with anti CD3/CD28 dynabeads (ThermoFisher 11131D). At 48 hours cells were harvested and stained with directly PE conjugated antibody (Biolegend 333906clone: BY88) and analyzed for expression of CD30 by FACS. Two-tailed Wilcoxon matched-pairs signed rank test was used to test for significant differences in cell surface expression of CD30 on PBMCs and EBVs. Two-tailed paired t-test was used to test for significant differences in cell surface expression of CD30 on HeLa cells.

Soluble CD30 measured by ELISA. Soluble CD30 was measured by sandwich ELISA (ThermoFisher BMS240INST) in a 1/4 dilution and according to manufacturer's instructions. Two-tailed Wilcoxon matched-pairs signed rank test was used to test for significant differences in levels of sCD30 from PBMC cultures. Two-tailed paired t-test was used to test for significant differences in levels of sCD30 from HeLa cell cultures.

Pathway analysis. We used DEPICT⁴⁹ to (1) prioritize candidate causal genes at associated loci and (2) highlight enriched pathways where genes at associated loci are highly expressed. DEPICT uses gene expression data derived from a panel of 77,840 mRNA expression arrays together with 14,461 existing gene sets based on molecular pathways derived from experimentally verified protein-protein interactions⁷², genotype-phenotype relationships from the Mouse Genetics Initiative⁷³, Reactome pathways⁷⁴, KEGG pathways⁷⁵, and Gene Ontology (GO) terms⁷⁶. DEPICT reconstitutes these 14,461 gene sets by calculating for each gene the probability of membership in each gene set, based on similarities across the expression data. Using these membership probabilities and a set of trait-associated loci, DEPICT tests whether any of the 14,461 reconstituted gene sets are enriched for genes at the trait-associated loci, and prioritizes genes that share predicted functions with genes at other trait-associated loci. We ran DEPICT using all 88 asthma-associated variants.

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

Data availability

The sequence variants from the Icelandic population whole-genome sequence data have been deposited at the European Variant Archive under accession code PRJEB15197. The GWAS summary statistics are available at [<https://www.decode.com/summarydata>]. The authors declare that the data supporting the findings of this study are available within the article, its Supplementary Information file, and upon reasonable request. The source data underlying Figs. 1–4 and supplementary Figs. 1–5 are provided as Source Data file.

Code availability

We used publicly available software (URLs listed below) in conjunction with the above described algorithms in the sequencing processing pipeline (Whole-genome sequencing, Association testing, RNA-seq mapping and analysis): BWA 0.7.10 mem [<https://github.com/lh3/bwa/>]; GenomeAnalysisTKLite 2.3.9 [<https://github.com/broadgsa/gatk/>]; Picard tools 1.117 [<https://broadinstitute.github.io/picard/>]; SAMtools 1.3 [<http://samtools.github.io/>]; Bedtools v2.25.0-76-g5e7c696z [<https://github.com/arq5x/bedtools2/>]; Variant Effect Predictor [<https://github.com/Ensembl/ensembl-vep/>]; BOLT-LMM [<https://data.broadinstitute.org/alkesgroup/BOLT-LMM/downloads/>]; GTEX data [<http://www.gtexportal.org/>]; GWAS catalog [<https://www.ebi.ac.uk/gwas/home>]; Uniprot [<https://www.uniprot.org/>]; Roadmap epigenomics project [<http://www.roadmapepigenomics.org/>]; Fantom5 [fantom.gsc.riken.jp/5/data]

Received: 21 December 2018; Accepted: 5 December 2019;
Published online: 20 January 2020

References

- Dharmage, S. C., Perret, J. L. & Custovic, A. Epidemiology of asthma in children and adults. *Front. Pediatr.* **7**, 246 (2019).
- Lambrecht, B. N. & Hammad, H. The immunology of asthma. *Nat. Immunol.* **16**, 45–56 (2015).
- Kuruvilla, M. E., Lee, F. E. & Lee, G. B. Understanding asthma phenotypes, endotypes, and mechanisms of disease. *Clin. Rev. Allergy Immunol.* **56**, 219–233 (2019).
- Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
- Vicente, C. T., Revez, J. A. & Ferreira, M. A. R. Lessons from ten years of genome-wide association studies of asthma. *Clin. Transl. Immunol.* **6**, e165 (2017).
- Zhu, Z. et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.* **50**, 857–864 (2018).
- Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).
- Gudbjartsson, D. F. et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41**, 342–347 (2009).
- Smith, D. et al. A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. *PLoS Genet.* **13**, e1006659 (2017).
- Pinart, M. et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitized and non-IgE-sensitized children in MeDALL: a population-based cohort study. *Lancet Respir. Med.* **2**, 131–140 (2014).
- Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
- Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
- Shrine, N. et al. Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. *Lancet Respir. Med.* **7**, 20–34 (2019).
- Johansson, A., Rask-Andersen, M., Karlsson, T. & Ek, W. E. Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Hum. Mol. Genet.* **1–20** (2019).
- Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
- Sigrist, C. J. et al. New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–7 (2013).
- Chiarle, R. et al. CD30 in normal and neoplastic cells. *Clin. Immunol.* **90**, 157–164 (1999).
- Hogan, L. E. et al. Increased HIV-1 transcriptional activity and infectious burden in peripheral blood and gut-associated CD4+ T cells expressing CD30. *PLoS Pathog.* **14**, e1006856 (2018).
- Wright, C. W., Rumble, J. M. & Duckett, C. S. CD30 activates both the canonical and alternative NF-kappaB pathways in anaplastic large cell lymphoma cells. *J. Biol. Chem.* **282**, 10252–10262 (2007).
- Berro, A. I., Perry, G. A. & Agrawal, D. K. Increased expression and activation of CD30 induce apoptosis in human blood eosinophils. *J. Immunol.* **173**, 2174–2183 (2004).
- Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 e19 (2016).
- Jones, A. V. et al. GWAS of self-reported mosquito bite size, itch intensity and attractiveness to mosquitoes implicates immune-related predisposition loci. *Hum. Mol. Genet.* **26**, 1391–1406 (2017).
- Heshmat, N. M. & El-Hadidi, E. S. Soluble CD30 serum levels in atopic dermatitis and bronchial asthma and its relationship with disease severity in pediatric age. *Pediatr. Allergy Immunol.* **17**, 297–303 (2006).
- Polte, T., Behrendt, A. K. & Hansen, G. Direct evidence for a critical role of CD30 in the development of allergic asthma. *J. Allergy Clin. Immunol.* **118**, 942–948 (2006).
- Nawrocki, J. F., Kirsten, E. S. & Fisher, R. I. Biochemical and structural properties of a Hodgkin's disease-related membrane protein. *J. Immunol.* **141**, 672–680 (1988).
- Froese, P. et al. Biochemical characterization and biosynthesis of the Ki-1 antigen in Hodgkin-derived and virus-transformed human B and T lymphoid cell lines. *J. Immunol.* **139**, 2081–2087 (1987).
- McDermott, M. F. et al. Germline mutations in the extracellular domains of the 55 kDa TNF receptor, TNFR1, define a family of dominantly inherited autoinflammatory syndromes. *Cell* **97**, 133–144 (1999).
- Lei, Z. et al. MiR-142-3p represses TGF-beta-induced growth inhibition through repression of TGFbetaR1 in non-small cell lung cancer. *FASEB J.* **28**, 2696–2704 (2014).
- Maes, T. et al. Asthma inflammatory phenotypes show differential microRNA expression in sputum. *J. Allergy Clin. Immunol.* **137**, 1433–1446 (2016).
- Grassel, S. & Bauer, R. J. Collagen XVI in health and disease. *Matrix Biol.* **32**, 64–73 (2013).

32. Nomura, A. et al. Increases in collagen type I synthesis in asthma: the role of eosinophils and transforming growth factor-beta. *Clin. Exp. Allergy* **32**, 860–865 (2002).
33. Wang, Z., Potter, C. S., Sundberg, J. P. & Hogenesch, H. SHARPIN is a key regulator of immune and inflammatory responses. *J. Cell. Mol. Med.* **16**, 2271–2279 (2012).
34. Hogenesch, H. et al. A spontaneous mutation characterized by chronic proliferative dermatitis in C57BL mice. *Am. J. Pathol.* **143**, 972–982 (1993).
35. Seymour, R. E. et al. Spontaneous mutations in the mouse Sharpin gene result in multiorgan inflammation, immune system dysregulation and dermatitis. *Genes Immun.* **8**, 416–421 (2007).
36. Rosa-Rosa, L., Zimmermann, N., Bernstein, J. A., Rothenberg, M. E. & Khurana Hershey, G. K. The R576 IL-4 receptor alpha allele correlates with asthma severity. *J. Allergy Clin. Immunol.* **104**, 1008–1014 (1999).
37. Wenzel, S. E. et al. IL4R alpha mutations are associated with asthma exacerbations and mast cell/IgE expression. *Am. J. Respir. Crit. Care Med.* **175**, 570–576 (2007).
38. Zhang, H. et al. Association of IL4R gene polymorphisms with asthma in Chinese populations. *Hum. Mutat.* **28**, 1046 (2007).
39. Al-Muhsen, S. et al. IL-4 receptor alpha single-nucleotide polymorphisms rs1805010 and rs1801275 are associated with increased risk of asthma in a Saudi Arabian population. *Ann. Thorac. Med.* **9**, 81–86 (2014).
40. Fajt, M. L. & Wenzel, S. E. Development of new therapies for severe asthma. *Allergy Asthma Immunol. Res.* **9**, 3–14 (2017).
41. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
42. Moffatt, M. F. et al. A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).
43. 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 (2014).
44. Takimoto, T. et al. Smad2 and Smad3 are redundantly essential for the TGF-beta-mediated regulation of regulatory T plasticity and Th1 development. *J. Immunol.* **185**, 842–855 (2010).
45. Yoshimura, A. & Muto, G. TGF-beta function in immune suppression. *Curr. Top. Microbiol. Immunol.* **350**, 127–147 (2011).
46. Bonnelykke, K. et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* **46**, 51–55 (2014).
47. Van Crombruggen, K., Zhang, N., Gevaert, P., Tomassen, P. & Bachert, C. Pathogenesis of chronic rhinosinusitis: inflammation. *J. Allergy Clin. Immunol.* **128**, 728–732 (2011).
48. van Rijt, L., von Richthofen, H. & van Ree, R. Type 2 innate lymphoid cells: at the cross-roads in allergic asthma. *Semin. Immunopathol.* **38**, 483–496 (2016).
49. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
50. Ling, M. F. & Luster, A. D. Allergen-specific CD4(+) T cells in human asthma. *Ann. Am. Thorac. Soc.* **13**(Suppl 1), S25–S30 (2016).
51. Tumes, D. J. et al. Epigenetic regulation of T-helper cell differentiation, memory, and plasticity in allergic asthma. *Immunol. Rev.* **278**, 8–19 (2017).
52. 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).
53. Gulcher, J. R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
54. Gudbjartsson, D. F. et al. Sequence variants from whole genome sequencing a large group of Icelanders. *Sci. Data* **2**, 150011 (2015).
55. 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. *Lancet Respir. Med.* **3**, 769–781 (2015).
56. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, 26 (2017).
57. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
58. Kristjansson, R. P. et al. A loss-of-function variant in ALOX15 protects against nasal polyps and chronic rhinosinusitis. *Nat. Genet.* **51**, 267–276 (2019).
59. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.* **22**, 719–748 (1959).
60. Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
61. 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, S1–3 (2012).
62. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
63. Benonisdottir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).
64. Oskarsson, G. R. et al. A truncating mutation in EPOR leads to hyporesponsiveness to erythropoietin with normal haemoglobin. *Commun. Biol.* **1**, 49 (2018).
65. Kehr, B. et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
66. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
67. Cao, Q. et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
68. The, F. C. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462 (2014).
69. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, 1–38 (2015).
70. Leca, G., Vita, N., Maiza, H., Fasseu, M. & Bensussan, A. A monoclonal antibody to the Hodgkin's disease-associated antigen CD30 induces activation and long-term growth of human autoreactive gamma delta T cell clone. *Cell Immunol.* **156**, 230–239 (1994).
71. Falini, B. et al. CD30 (Ki-1) molecule: a new cytokine receptor of the tumor necrosis factor receptor superfamily as a tool for diagnosis and immunotherapy. *Blood* **85**, 1–14 (1995).
72. Lage, K. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
73. Bult, C. J. et al. Mouse genome informatics in a new age of biological inquiry. In *Proc. IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, Arlington, VA, USA. <https://doi.org/10.1109/BIBE.2000.889586>, 29–32 (2000).
74. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
75. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
76. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).

Acknowledgements

We thank the individuals who participated in this study and the staff at the Icelandic Patient Recruitment Center and the deCODE genetics core facilities. Further to all our colleagues who contributed to the data collection and phenotypic characterization of clinical samples as well as to the genotyping and analysis of the whole-genome association data. This research has been conducted using the UK biobank Resource under Application Number '24711'.

Author contributions

T.A.O., G.L.N., D.F.G., P.S., U.T., G.T., I.J. and K.S. designed the study and interpreted the results. T.A.O., U.S.B., G.I.E., D.G., T.G., D.L., B.R.L., P.T.O., I.O., O.S. and I.J. carried out the subject ascertainment and recruitment. T.A.O., O.A.S., S.B., J.K.S., E.V.I., B.V.H., G.H.H., G.M., K.N., T.J., G.S., V.T., D.F.G., P.S., U.T., G.T. and I.J. performed the sequencing, genotyping and expression analyses. T.A.O., F.T., K.B., A.B.A., A.M.K., S.G., U.T., G.L.N. and I.J. planned and performed the functional lab work. T.A.O., O.A.S., S.B., J.K.S., E.V.I., B.V.H., A.G., G.H.H., G.M., K.N., T.J., L.S., G.S., V.T., D.F.G., P.S., U.T., G.T. and I.J. performed the statistical and bioinformatics analyses. T.A.O., O.A.S., D.F.G., P.S., U.T., I.J. and K.S. drafted the manuscript. All authors contributed to the final version of the paper.

Competing interests

Authors affiliated with deCODE genetics/Amgen Inc. declare competing interests as employees. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-14144-8>.

Correspondence and requests for materials should be addressed to I.J. or K.S.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

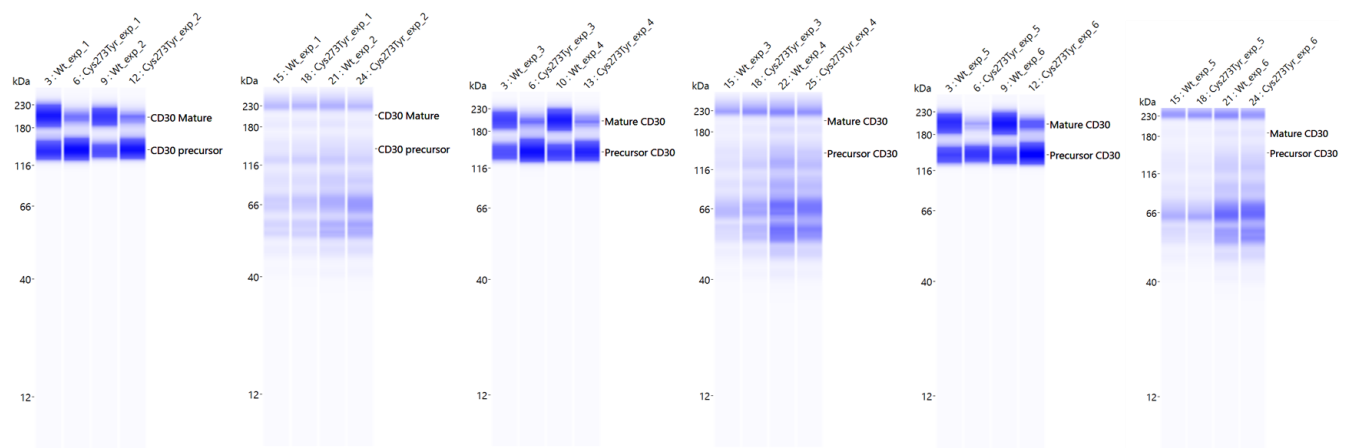
© The Author(s) 2020

Supplementary information

Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis

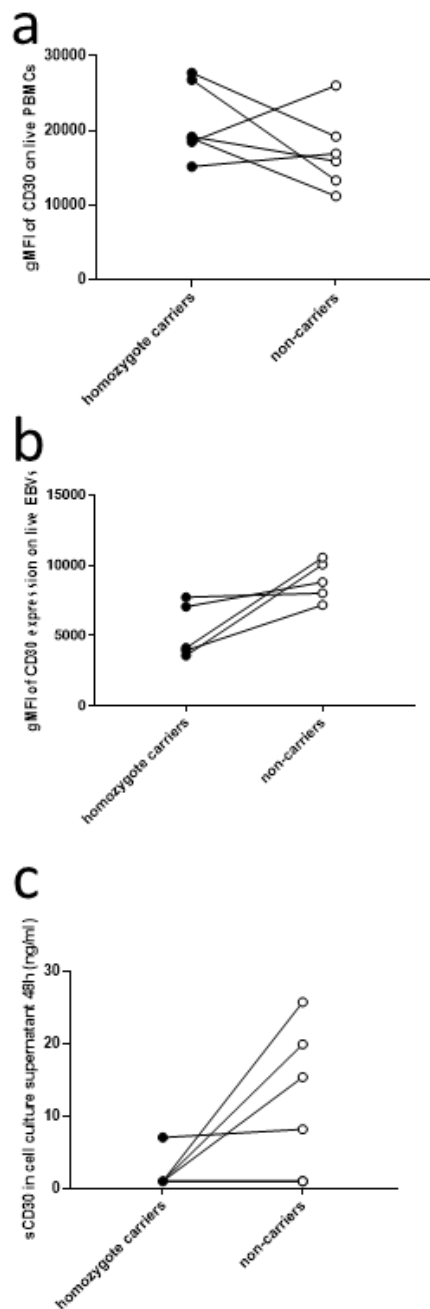
Thorunn A Olafsdottir^{1,2}, Fannar Theodors¹, Kristbjorg Bjarnadottir¹, Unnur Steina Bjornsdottir^{3,4}, Arna B Agustsdottir¹, Olafur A Stefansson¹, Erna V Ivarsdottir^{1,5}, Jon K Sigurdsson¹, Stefania Benonisdottir¹, Gudmundur I Eyjolfsson⁶, David Gislason^{4,7}, Thorarinn Gislason^{2,8}, Steinunn Guðmundsdóttir¹, Arnaldur Gylfason¹, Bjarni V Halldorsson^{1,9}, Gisli H Halldorsson¹, Thorhildur Juliusdottir¹, Anna M Kristinsdottir¹, Dora Ludviksdottir^{2,7}, Bjorn R Ludviksson^{2,10}, Gisli Masson¹, Kristjan Norland¹, Pall T Onundarson^{2,11}, Isleifur Olafsson¹², Olof Sigurdardottir^{2,13}, Lilja Stefansdottir¹, Gardar Sveinbjornsson¹, Vinicius Tragante^{1,14}, Daniel F Gudbjartsson^{1,5}, Gudmar Thorleifsson¹, Patrick Sulem¹, Unnur Thorsteinsdottir^{1,2}, Gudmundur L Norddahl¹, Ingileif Jonsdottir^{1,2} and Kari Stefansson^{1,2}

1. deCODE genetics/Amgen, Inc., Reykjavik, Iceland
2. Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland
3. Department of Medicine, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
4. The Medical Center Mjódd, Reykjavik, Iceland
5. School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
6. The Laboratory in Mjódd, RAM, Reykjavik, Iceland
7. Department of Respiratory Medicine and Sleep, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
8. Department of Sleep, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
9. School of Science and Engineering, Reykjavik University, Reykjavik, Iceland
10. Department of Immunology, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
11. Department of Laboratory Hematology, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
12. Department of Clinical Biochemistry, Landspítali, The National University Hospital of Iceland, Reykjavik, Iceland
13. Department of Clinical Biochemistry, Akureyri Hospital, Akureyri, Iceland
14. Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands.



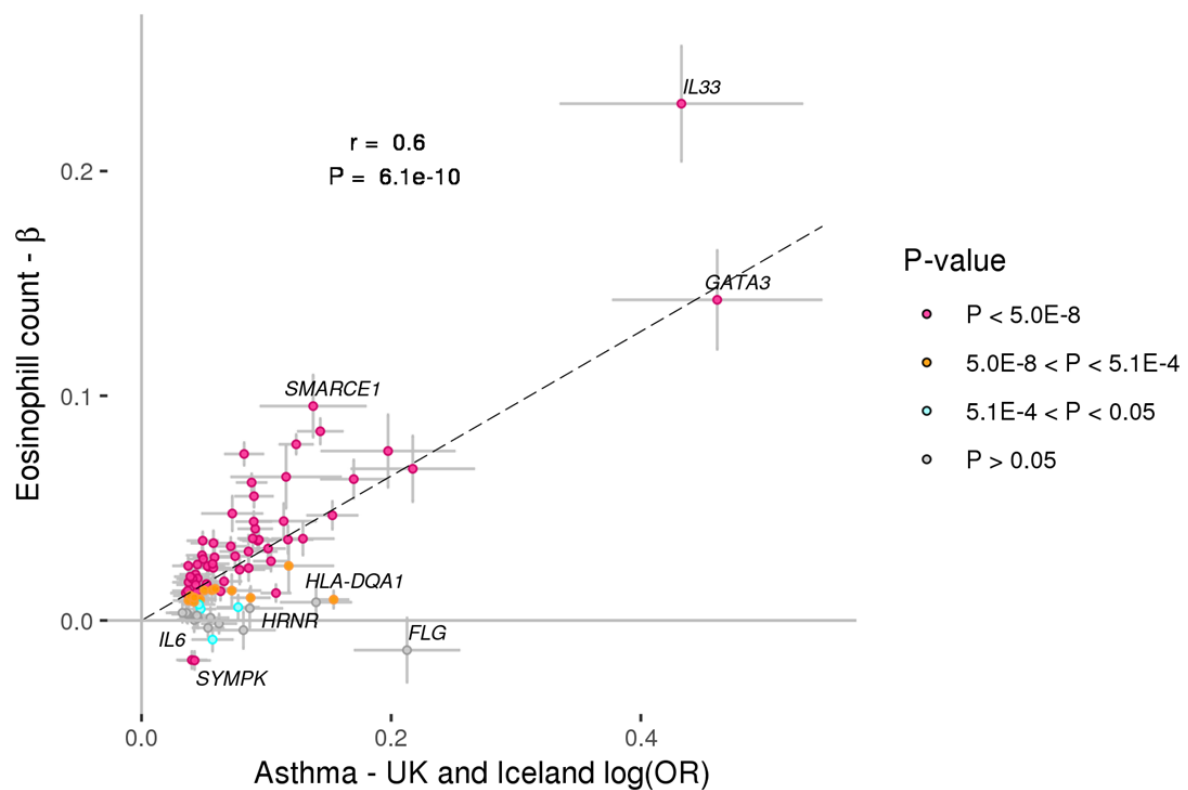
Supplementary figure 1: CD30 expression in HeLa cells overexpressing wild-type or variant p.Cys273Tyr CD30.

Protein simple WES analysis of CD30 expression in cell lysate from HeLa cells overexpressing wild-type or variant p.Cys273Tyr CD30. One representative lane is shown for each triplicate of lysate run and corresponding total protein analysis of those samples. CD30 and total protein quantifications were run in separate wells from the same lysis preparation (N=6).

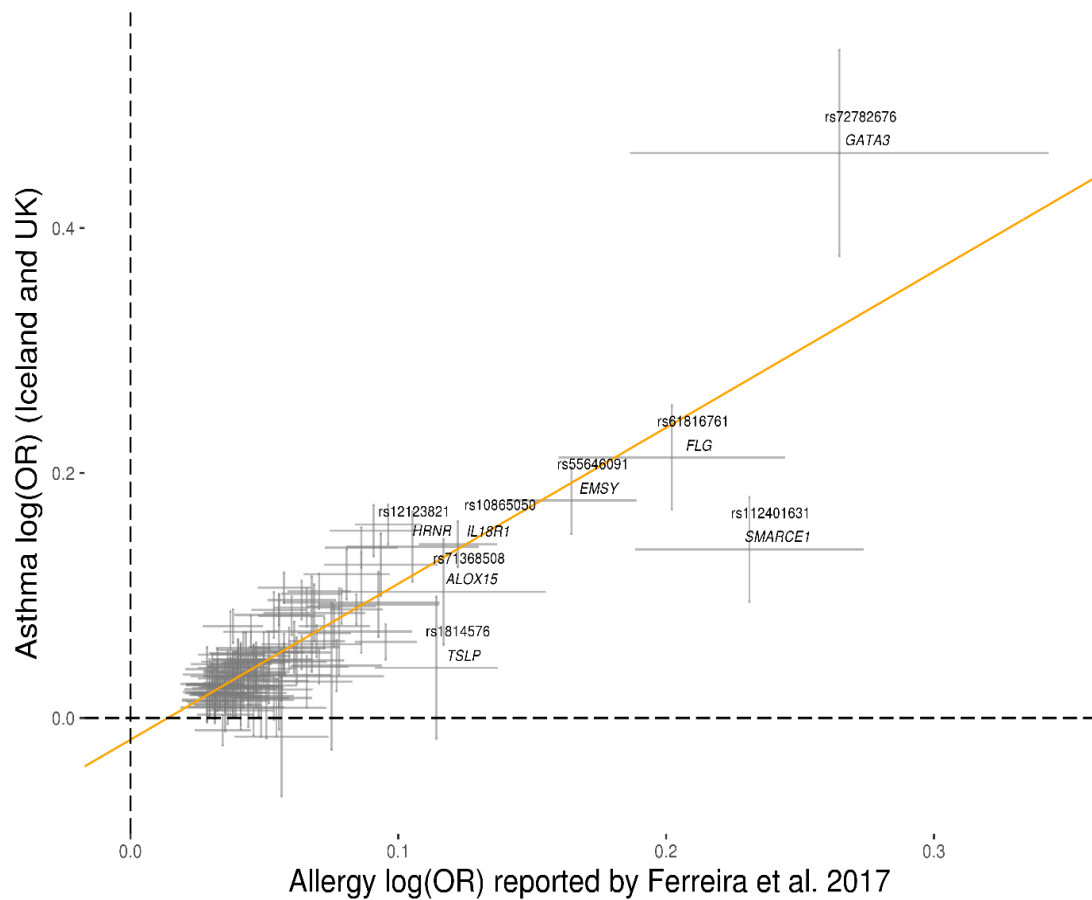


Supplementary figure 2: Surface expression of CD30 on PBMC and EBV transformed lymphoblasts from homozygous p.Cys273Tyr carriers age and gender matched non-carriers.

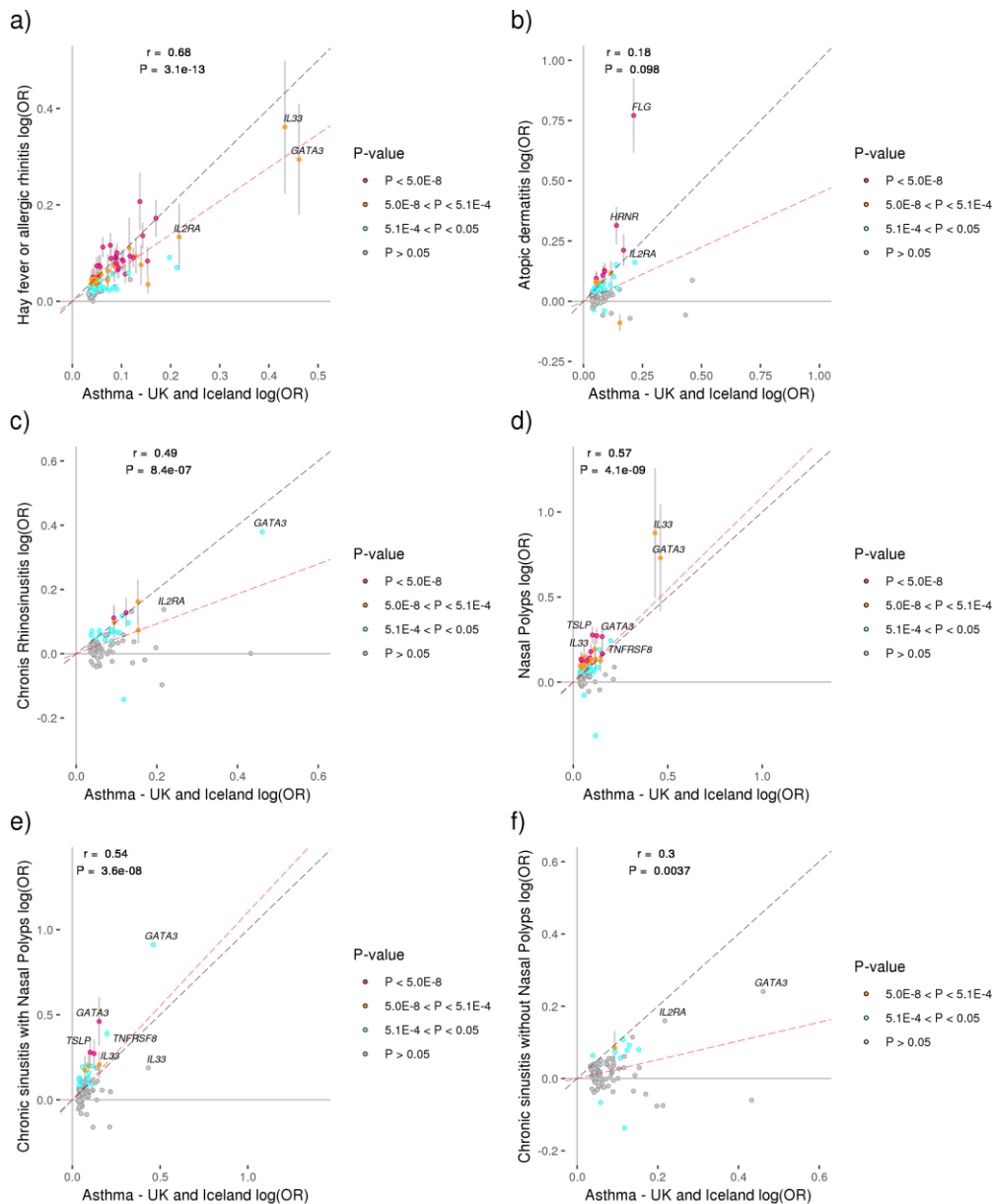
(a) Surface expression of CD30 on stimulated PBMCs from homozygous carriers of p.Cys273Tyr and age and gender matched non-carriers measured by flow cytometry as gMFI. (b) Surface expression of CD30 on EBV transformed lymphoblasts from a homozygous carriers and a non-carrier of p.Cys273Tyr. (c) Soluble CD30 (sCD30) levels (ng/ml) in activated PBMC culture supernatant from homozygotes and non-carriers of p.Cys273Tyr.



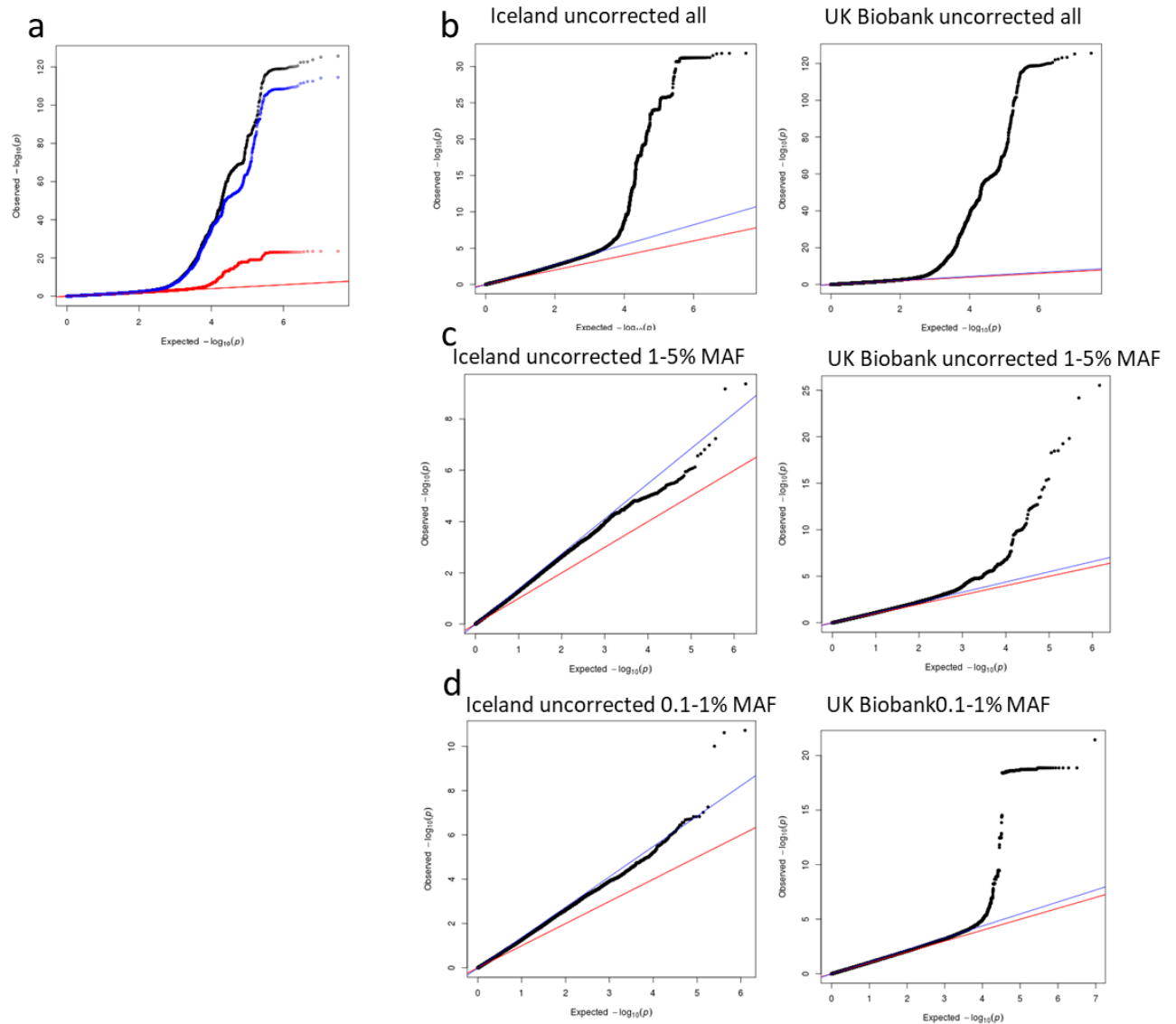
Supplementary figure 3: Effects of the 88 asthma variants on asthma and eosinophil count. The x-axis of the scatter plot shows the logarithm of the asthma odds ratios found in Icelandic and UK meta-analysis and the y-axis shows the age adjusted effect (in SD) on Eosinophil count in a combined data set of Iceland and UK. All effects are shown for the asthma risk increasing allele. Error bars represent 95% confidence intervals. The line represents results from a simple linear regression through the origin using MAF (1-MAF) as weights and the weighted correlation coefficient (r) and P -value (t-test) are shown in the graph.



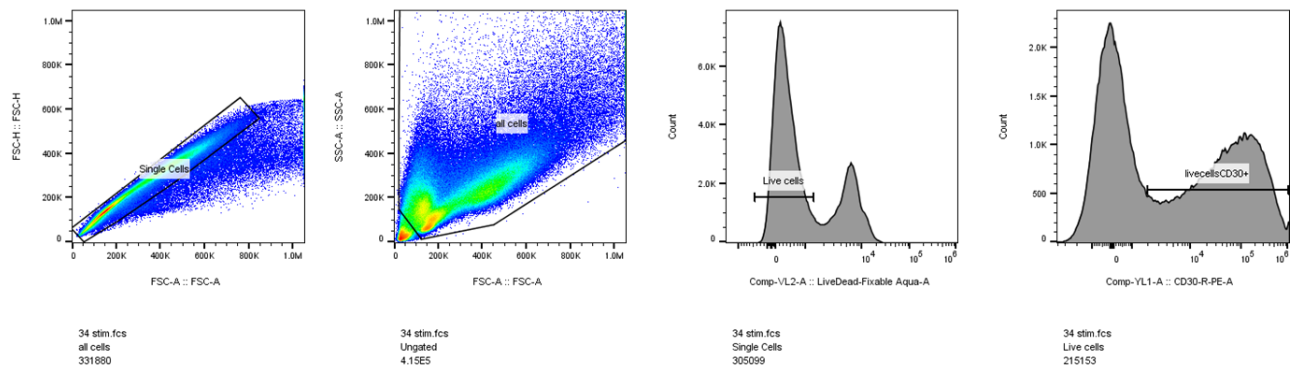
Supplementary figure 4: Effects of reported allergy variants (Ferreira et al. 2017) on asthma and allergy. Scatter plot showing 134 previously reported allergy SNPs. The x-axis shows the logarithm of their allergy odds ratios reported by Ferreira et al. 2017 and the y-axis shows the logarithm of estimated odds ratios for asthma in a combined data set of Iceland and UK (Cases: $N_{\text{Iceland}}=16,247$; $N_{\text{UK}}=52,972$). All effects are shown for the allergy risk increasing allele. Variants associated with the largest allergy risk are labeled. Error bars represent 95% confidence intervals. The orange solid line, $y = -0.02 + 1.27x$, represents results from a simple linear regression using $\text{MAF}(1-\text{MAF})$ as weights with $R = 0.77$ ($P = 6.0 \times 10^{-28}$; t-test).



Supplementary Figure 5. Effects of the 89 asthma variants on asthma and various allergic phenotypes. The x-axis of the scatter plot shows the logarithm of the asthma odds ratios found in Icelandic and UK meta-analysis and the y-axis shows the logarithm of estimated odds ratios for (a) hay fever or allergic rhinitis, (b) Atopic dermatitis, (c) Chronic rhinosinusitis, (d) Nasal Polyps, (e) Chronic rhinosinusitis with Nasal Polyps and (f) Chronic rhinosinusitis with Nasal Polyps in a combined data set of Iceland and UK. Effects of all allergic phenotypes are based on Icelandic and UK meta-analysis except for Atopic dermatitis that is based on Icelandic and external data meta-analysis (Paternoster et. al, 2015) as described in materials and methods. All effects are shown for the asthma risk increasing allele. Error bars represent 95% confidence intervals. The line represents results from a simple linear regression through the origin using MAF (1-MAF) as weights and the weighted correlation coefficient (r) and P -value (t-test) are shown in the graphs.



Supplementary Figure 6. Quantile-quantile plot (QQ-plot) showing chi-square statistics for (a) all variants corrected for genomic control, (b) all variants not corrected for genomic control, (c) low frequency variants (1%-5%) not corrected for genomic control and (d) rare variants (0.1%-1%) not corrected for genomic control. The different colors in (a) represent variants found in Iceland (red), the UK (blue), and meta-analysis of the two cohorts (black). The red diagonal line represents expected distribution assuming no inflation of the chi-square statistics. The blue line in figures b-d represent the genomic correction factor; 1.097 for UK Biobank and 1.370 for Iceland.



Supplementary figure 7: Gating strategy used for facs analysis to determine geometric mean fluorescence intensity of CD30 positive live cells presented in Fig. 2c,e and Supplementary Fig. 2a,b.

Supplementary table 1. Correlation of the effect of asthma associating variants in other asthma/allergy phenotypes. All the correlation is performed on meta-analysis of Icelandic and UK data (number of participants in each country for the different phenotypes is shown in the table). The table depicts the weighted correlation between effects of 88 independent asthma variants and the 11 phenotypes shown in the first column. MAF(1-MAF) were used as weights when computing the correlation.

Phenotype	P	wCorr	Iceland (n)	UK (n)	External****
Asthma ICD10 codes diagnosis ⁺	5.3E-81	0.97	13190	26152	-
Asthma self-reported ⁺	3.30E-120	0.99	6281	47882	-
Early-Onset Asthma [*]	1.4E-27	0.86	4394	12723	-
Late-Onset Asthma ^{**}	1.1E-22	0.82	12329	29731	-
Allergic Asthma	8.4E-19	0.77	3572	542	-
Hay fever or allergic rhinitis ^{***}	3.1E-13	0.68	7460	24804	-
Nasal Polyps	4.10E-09	0.58	1874	3191	-
Chronic sinusitis with Nasal Polyps	3.6E-08	0.54	906	672	-
Chronic Rhinosinusitis	8.4E-07	0.49	3886	2420	-
Chronic sinusitis without Nasal Polyps	3.70E-03	0.3	3061	1748	-
Atopic dermatitis	0.32	0.11	8325	-	21399

⁺There is an overlap between the ICD10 codes diagnosis and Astma self reported lists both in Iceland (N=3224) and in the UK (N=21092)

^{*}Early onset is defined as the first diagnosis < 18 years of age

^{**}Late onset is defined as the first diagnosis ≥ 18 years of age. Late onset asthma diagnosis from Iceland contains both where the first diagnosis was at or after 18 years of age and where the first diagnosis is unknown

^{*}Hayfever or allergic rhinitis combines doctoral diagnosis of allergic rhinitis from Iceland and a questionnaire data from UK Biobank on hayfever or allergic rhinitis (Non-cancer-illness code 1387)

^{****}Due the few ICD10 code (L20) defined Atopic dermatitis cases reported in UKB, the AD association is based on cases defined in Iceland (N=8,325) and the largest GWAS on AD to date (N=21,399; PMID:26482879)

Supplementary Table 2. Genetic correlation calculated between asthma and allergic phenotypes in Iceland and UK biobank

Phenotype (N Iceland)	Phenotype (N UKB)	rg	se	P
Asthma (16,247)	Asthma (52,942)	0.8899	0.097	4.60E-20
Hay fever or allergic rhinitis (7,460)	Hay fever or allergic rhinitis (24,804)	0.9874	0.1266	6.30E-15
Nasal Polyps (1,874)	Nasal Polyps (3,191)	1.0663	0.3301	0.0012
Chronic sinusitis with Nasal Polyps (906)	Chronic sinusitis with Nasal Polyps (672)	0.4068	0.3035	0.1801
Chronis Rhinosinusitis (3,886)	Chronis Rhinosinusitis (2,420)	1.0763	0.6307	0.0879
Chronic sinusitis without Nasal Polyps (3,061)	Chronic sinusitis without Nasal Polyps (1,748)*	-	-	-
Atopic dermatitis (8,325)	Atopic dermatitis (external)	0.9135	0.2522	0.0003
Asthma (16,247)	Hay fever or allergic rhinitis (24,804)	0.4244	0.0936	5.83E-06
Asthma (16,247)	Nasal Polyps (3,191)	0.5442	0.1286	2.32E-05
Asthma (16,247)	Chronic sinusitis with Nasal Polyps (672)	0.3053	0.172	0.076
Asthma (16,247)	Chronis Rhinosinusitis (2,420)	0.3802	0.2545	1.35E-01
Asthma (16,247)	Chronic sinusitis without Nasal Polyps (1,748)*	-	-	-
Asthma (16,247)	Atopic dermatitis (external; N=21,399)	0.3488	0.1333	0.0089
Hay fever or allergic rhinitis (7,460)	Asthma (52,942)	0.7985	0.1033	1.07E-14
Nasal Polyps (1,874)	Asthma (52,942)	0.4312	0.1414	0.0023
Chronic sinusitis with Nasal Polyps (906)	Asthma (52,942)	0.5248	0.193	0.0065
Chronis Rhinosinusitis (3,886)	Asthma (52,942)	0.4733	0.196	0.016
Chronic sinusitis without Nasal Polyps (3,061)	Asthma (52,942)	0.2662	0.1375	0.0529
Atopic dermatitis (8,325)	Asthma (52,942)	0.5617	0.1401	6,07E-05
Atopic dermatitis (external; N=21,399)	Asthma (52,942)	0.4576	0.0671	9,38E-12

* Chronic Sinusitis without Nasal polyps list from the UKB did not yield any results above noise level.

Supplementary table 3: Characteristics of Icelandic and UKB asthma cohorts.

Cohorts	N	Yob (SD)	Male N (%)
Iceland cases	16,247	1965.4 (26.5)	6604 (40.6)
Iceland controls	346,486	1968.7 (29.0)	176,981 (51.1)
UK cases	52,942	1951.6 (8.2)	22,520 (42.5)
UK controls	355,713	1951.1 (8.0)	165,199 (46.4)

In the UK individuals aged between 40 and 69 were recruited whereas there was no age restriction in the Icelandic recruitment.

Supplementary table 4: List of primers used in the study

Primer	Sequence
CD30_C273T_F	5'AAGACGCCATATGCATGGAAC'3
CD30_C273T_R	5'CTCCACAAGGTCATCTCG'3

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a ☐ Confirmed

☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on [graphics for biologists](#) contains an article on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

The "Whole-genome sequencing and imputation" subsection in materials and methods section contains the following information:
"The GWASs in Iceland were performed with 32.5 million markers identified through whole-genome sequencing of 15,520 Icelanders to an average genome-wide coverage of 34X and subsequently imputed into 151,677 chip-typed individuals, as well as their first and second degree relatives. The imputation has been extensively described in recent publications⁴⁵. Genotyping of URB samples was performed using a custom-made Affymetrix chip, UK BiLEVE x10m46, and with the Affymetrix UK Biobank Axiom array⁴⁷. Imputation was performed by the Wellcome Trust Centre for Human Genetics, using the HaploType Reference Consortium (HRC) and the UK10K haplotype resources. This yields a total of 96 million imputed variants, however only 27 million variants imputed using the HRC reference set passed the quality filters used in our study."

Data analysis

The "Code availability" subsection in materials and methods section contains the following information:
"We used publicly available software (URLs listed below) in conjunction with the above described algorithms in the sequencing processing pipeline (Whole-genome sequencing, Association testing, RNA-seq mapping and analysis):
BWA 0.7.10 mem, <https://github.com/hb3/bwa>
GenomeAnalysisTKLite 2.3.9, <https://github.com/broadgenomics/gatk/> 397
Picard tools 1.117, <https://broadinstitute.github.io/picard/> 398
SAMtools 1.3, [http://samtools.github.io/](http://samtools.github.io/399) 399
Bedtools v2.25.0-76-g5e7c696a, <https://github.com/arq5x/bedtools2/> 400
Variant Effect Predictor <https://github.com/Ensembl/ensembl-vep> 401
BOLT-LMM <https://data.broadinstitute.org/alkesgroup/BOLT-LMM/downloads/>"

Further, R was extensively used to analyze data and create plots.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Hela (Human cervix epitheloid carcinoma), Hela (ECACC 93021013) Lot: 14A021

Authentication

Cells were authenticated by ECACC. See certificate of analysis here:
https://www.ghe-culturecollection.org.uk/Ccfs.do?mode=view&collection=ecacc_gc&refid=93021013&formatid=18&lotno=14A021&searchRef=9ftr6k318t4pd

Mycoplasma contamination

Cells tested negative for mycoplasma.

Commonly misidentified lines (See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

The Icelandic data is based on whole-genome sequence data from the whole blood of 15,220 Icelanders participating in various disease projects at deCODE genetics. In addition, 151,677 Icelanders have been genotyped using Illumina SNP chips and genotype probabilities for untyped relatives have been calculated based on Icelandic genealogy. The UK data is from the UK Biobank, a large prospective cohort study of approx. 500,000 individuals from across the United Kingdom, aged between 40-69 at recruitment. As only 42 individuals had the ICD10 code for Atopic dermatitis (L20) in UKB we performed meta-analysis with our Icelandic list together with the largest published meta-analysis on AD to date downloaded from the GRAASP database (<https://grasp.nhlbi.nih.gov/FullResults.aspx>), to study this phenotype. For further details, see subsection "Study sample sets for asthma and allergy phenotypes" in Materials and methods.

Recruitment

Asthma in the UK Biobank was defined as ICD10 diagnoses in fields 41202 or 41204, including anyone of J45.0, J45.1, J45.8, J45.9 and J46 and/or self-reported by the non-cancer illness code, self-reported during verbal interview (data-field 20002) with a code for asthma (11111).
Icelandic asthma patients over 18 years of age were recruited who attended an asthma clinic or emergency room at the National University Hospital of Iceland or the Icelandic Medical Center (Læknisstofud) during the years 1977 to 2017. Asthma diagnosis was based on a combination of physician's diagnosis and ICD10 diagnosis, including anyone of J45.0, J45.1, J45.8, J45.9 and J46 and/or self-reported by a positive reply to the question: "Has a doctor confirmed your asthma diagnosis?". Atopy status determined by skin prick testing and age of onset was available for part of the asthma cohort both in Iceland and UK. Early onset was defined as first diagnosis <17 years of age and late onset as first diagnosis >17 years of age.
Allergic rhinitis combines doctoral diagnosis of allergic rhinitis from Iceland and a questionnaire data from UK Biobank on hayfever or allergic rhinitis (Non-cancer-illness code 1387). As only 42 individuals had the ICD10 code for Atopic dermatitis (L20) in UKB we ran meta-analysis with our Icelandic list (N=8325) derived both from physician's diagnosis and ICD10 code (L20) together with the largest published meta-analysis on AD to date downloaded from the GRAASP database (<https://grasp.nhlbi.nih.gov/FullResults.aspx>), to study this phenotype. Other allergy diagnosis were based on ICD10 codes from UKB and either physician's diagnosis or ICD10 codes from Iceland. The allergy phenotypes used were: Nasal polyps (ICD10-J33), chronic sinusitis (ICD10-J32), chronic sinusitis with nasal polyps (combined: ICD10-J32 and J33) and chronic sinusitis without nasal polyps (ICD10-J32 without ICD10-J33). Number of genotyped individuals in each cohort are listed in Supplementary table 12. Icelandic controls were participants from various deCODE genetics programs.

Ethics oversight

All participating individuals who donated blood signed informed consent. The personal identities of participants were encrypted using a third-party system approved and monitored by the Icelandic Data Protection Authority. The study was approved by the National Bioethics Committee in Iceland.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

EBV cells were harvested from culture, counted and diluted to 1x10⁶ cell/ml plated in a 96 well V bottom plate. Cells were then stained with primary antibodies against CD30 (Biologend 333918) and analyzed for expression of CD30 by FACS. Cryopreserved PBMC were thawed and plated for culture in RPMI1640 (ThermoFisher 61870-036) supplemented with 10% fetal bovine serum (ThermoFisher 10500-064), 50 units/mL penicillin, 50 µg/mL streptomycin (ThermoFisher 15070-063), 20 mM HEPES

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence variants passing GATK filters have been deposited in the European Variation Archive, accession number PRJEB15197 (<https://www.ebi.ac.uk/ena/data/view/PRJEB15197>)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](#)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The sample size corresponds to all available data from Iceland and UK biobank

Data exclusions

No data was excluded from the analyses

Replication

We performed GWAS studies in two independent populations and combined the results. Results are presented for the populations independently and combined and heterogeneity of effects between popus are assessed.

Randomization

No randomizations was used.

Blinding

Not relevant for this study, as this is a case-control GWAS study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study; if you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems	Methods
n/a ☐ Involved in the study	n/a ☐ Involved in the study
☒ Antibodies	☒ ChIP-seq
☒ Eukaryotic cell lines	☒ Flow cytometry
☒ Palaeontology	☒ MRI-based neuroimaging
☒ Animals and other organisms	
☒ Human research participants	
☒ Clinical data	

Antibodies

Antibodies used	TNFRSF8/CD30 (E1A6Y) Rabbit mAb. Cell signaling Cat: 95620 Lot:1 Biologend 333918, Biologend 333906, Biologend 300318, Biologend 300530, Biologend 353210, Biologend 304112
Validation	Monoclonal antibody is produced by immunizing animals with a synthetic peptide corresponding to residues surrounding Leu500 of human TNFRSF8/CD30 protein. TNFRSF8/CD30 (E1A6Y) Rabbit mAb recognizes endogenous levels of total TNFRSF8/CD30 protein. Validated for Western blot by manufacturer by Western blot analysis of extracts from 293T cells, mock transfected (-) or transfected with a construct expressing Myc/DDK-tagged full-length human TNFRSF8/CD30 protein (hCD30-Myc/DDK+), using TNFRSF8/CD30 (E1A6Y) Rabbit antibody. See certificate of analysis here: https://media.cellsignal.com/coa/95620/1/95620-lot-1-coa.pdf Each lot of the Biologend antibodies used is quality control tested by immunofluorescent staining with flow cytometric analysis as by the manufacturers as stated on their website.

	(ThermoFisher 15630056) and 30 IU/ml IL-2 (R&D Systems AFL202) in a 24 well TCR plate and rested overnight. Cells were then stimulated for 48 hours with anti CD3/CD28 dynabeads (ThermoFisher 111310). At 48 hours cells were harvested and stained with directly conjugated antibodies (Biologend 333906, Biologend 300318, Biologend 300530, Biologend 353210, Biologend 304112) and analyzed for expression of CD30 by FACS. Wilcoxon signed rank test was used to test for significant differences in cell surface expression of CD30.
Instrument	Thermo Fisher Life technologies Attune NxT
Software	FlowJo v10
Cell population abundance	Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.
Gating strategy	Cells were first gated by size with FSC Area/SSC Area and doublets were excluded by gating at FSC Height/FSC Area. Then cells were gated for CD30 expression.
☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

In the manuscript "Eighty-nine asthma risk variants highlight the role of T cell regulation, inflammation and airway remodeling in asthma pathogenesis" Olafsdottir et al. used two large study populations from deCODE AND UK Biobank in (UKB) to conduct a meta-GWAS on asthma. They report 89 independent signals at 56 loci of which 21 signals at 17 loci represented novel asthma loci. The respective candidate genes were identified based on functional annotations, eQTL data, and regulatory sites covered by the associated SNPs. In addition, for two variants genotype-specific expression data are generated. The reported genes are involved in T-cell signaling, innate immunity and airway remodeling.

Major comment

The manuscript is well written and the methods are sound. Major assets of the study are the large data sets, providing power for discovery of novel loci. The phenotype definitions are appropriate and the replication of a large number of known asthma/allergy loci points to the validity of the approach. Nonetheless several points need to be clarified.

For imputation, the authors used different methods and different reference sequences. They should comment on this and whether it might have introduced a bias. They do not report any QQ plots, neither for the individual populations nor for the combined analysis. Please provide the plots and information on genomic inflation factors.

Essentially, genome-wide significant association signals are derived from UKB, the majority of loci does not replicate in the Icelandic population. Please comment on this.

What is the advantage of the Mantel-Haenszel model used for meta analysis. It assumes homogeneous ORs in the study populations but some loci show heterogeneity. How does this affect the validity of the results? Are the results different from the results calculated with METAL?

Additional points

On page 3, how was the conditional analysis performed? If conditioning on the lead variant identified more than one additional genome-wide significant signal, how was the number of independent signals determined? This should be explained in the methods section in more detail.

On page 5, replace IL21 with IL21R

On page 7, the effect of the IL33 variant was not only greater in EOA but also in LOA than in AA. Accordingly, this might be a locus specific for non-allergic asthma.

On page 7, the correlation ρ and r^2 values for asthma vs allergic rhinitis and allergic dermatitis do not agree with supplementary figure 5a and 5b respectively.

On page 20 (methods), please specify the Biolegend antibodies.

In figure 2b and c, please explain the number of dots.

In figure 3, please explain the sequences in more detail. Why are some bases on top of the 5'-sequence or below the 3' sequence? Why do the red sequences on the left and on the right differ?

In table 1, please check the overlap with Ferreira et al. (Nat Genet. 2017 Dec;49(12):1752-1757) and with Shrine et al. (Lancet Respir Med. 2019 Jan;7(1):20-34).

On page 16, the p-value for the secondary signal should not be based on the number of loci ($n=56$). Please change this statement. The applied threshold is correct.

In supplementary fig 1, please show the whole size range and also the loading control.

In supplementary fig 4, delete "Ref" in the header or add the reference.

In supplementary fig 5, the scaling of both axes should be identical.

In supplementary tab 1, what was conditioned on in the combined analysis? Comment on heterogeneity of effects? Which threshold was used and why?

At the IL13 locus, would conditioning on rs848 eliminate the other 2 signals? Could this be the causal variant?

In supplementary tab 2, please add the risk allele for the previously reported SNPs in order to compare the effect direction.

In supplementary tab 3, what was conditioned on?

In all tables ordered according to the SNP positions, please check the order of SNPs.

In supplementary tab 5, the confidence interval is missing, rsq values are missing, p-values from Ferreira et al are missing.

Why are p-values from Ferreira et al different for the same SNP (e.g. rs73205303)?

Novel loci are not associated in the Icelandic population, please comment.

In supplementary tabs 7 to 9, why are only novel SNPs included? Since previous studies performed different functional analyses, the authors should report the results for all independent signals.

In supplementary tab 8, please report the effect alleles. Why is whole blood from GTEx not included. It would be interesting to have a comparison of the results with the deCODE data (suppl tab 6).

In supplementary tab 9, why were transcription factor binding sites not included?

In supplementary tab 10, for rs6926894, the OR of UKB is reported instead of deCODE, the CI is missing.

In supplementary tab 12, the number of UKB asthma cases does not agree with page 2, why?

Reviewer #2:

Remarks to the Author:

The authors carried out the largest GWAS meta-analysis of asthma to date with over 69,000 cases and over 700,000 controls. In total 89 independent asthma variants were identified. Of these:

- 48 signals in 24 known asthma loci: 24 new/24 previously reported
- 20 signals in 15 known 'combined allergic phenotype (asthma/hay fever/eczema)' loci
- 21 signals in 17 new asthma loci

They also quantify the correlation of the effects with other asthma and allergic phenotypes and perform functional and eQTL analyses to investigate the implicated genes and their functional roles.

This work is novel in that it is the largest ever meta-analysis; the major claims are the discovery of 21 new independent asthma loci, of which 18 are common ($MAF \geq 5\%$) and 3 uncommon variants, 2 of the uncommon variants show a 'small' protective effect for asthma. Most of the genes implicated in the reported associations with asthma are involved in known T cell regulation, inflammation and airway remodeling.

I praise the authors for the detail of the data reported in the supplementary tables. These data will be essential for other researchers to be able to use the results of this meta-analysis. For instance, complete GWAS summary statistics are needed to perform Mendelian Randomization studies that I'm sure will follow these findings.

Comments

1. In the main text the authors state that the asthma phenotype used for this meta-analysis includes both definitions: doctor ever diagnose asthma and/or self-reported asthma. It is unclear whether the self-reported asthma is self-reported doctor diagnosis of asthma. Can the authors comment on the validity of the cases? What proportion of cases were defined from a doctor diagnoses, self-reported asthma and self-reported doctor diagnosis? Could there be an over-reporting of asthma cases (false positives) and what would the implications be?
2. Main results are described in page 3. It took this reviewer several reads to have a clear understanding of the old/new findings. A suggested opening sentence would be: "Conditional analysis yield 89 independent signals: 48 of these signals in 24 known asthma loci (50% of the signals are new), 20 in 15 known combined (asthma-eczema-rhinitis) phenotype loci and 21 new signals in 17 new asthma loci." And then expand from here..
3. The authors don't seem to provide a description of the studied populations. How do the Iceland and UK populations compare in terms of general demographics?
4. Could the author comment on the magnitude and clinical implications of the protective effects? Comb OR (95%CI) = 0.82 (0.78, 0.87) for variant near TNFRSF8, Comb OR (95%CI) = 0.89 (0.86, 0.92) for variant near TGFBR1.
5. Table 1. Confidence intervals are not provided. Can the authors add the corresponding 95 % CIs (as in Suppl Table 1) or the standard errors so the effects can be fully interpreted together with the p-values?
6. The authors have investigated the effect of the asthma variants on early onset (first diagnosis ≤ 17 years) and late onset (first diagnosis > 17 years) asthma and found that for some variants the effect was stronger on early onset asthma compared to late onset asthma. Do the authors think that these results support the notion that early and late onset asthma have partly distinct genetic architectures? And if so, what implications can be drawn?

Reviewer #3:

Remarks to the Author:

The findings might be interesting but the experimental part that meant to support genetic findings is weak and lacks appropriate quality controls.

Figure 2 b and c are underpowered. There is also a lack of evaluation of soluble CD30 in supernatants from Hela cells transfected with different CD30 variants.

Figure 2 d and e. There is no mention of statistical test used to test difference in CD30 surface expression and soluble levels in d and e. Authors linked heterozygous carriers with age and gender matched non carrier controls suggesting that paired t test was used. I am not convinced that this is appropriate. The data in d might be skew by two pairs.

The representative flow cytometry plots of CD30 surface expression for c-d are required.

Figure 3. The conclusion that rs41283642 increases TGFb expression is not supported experimentally. Evaluation of surface TGFbR1 expression in people with different variants is recommended. Also, in order to claim that rs41283642 increases TGFb expression because of a

lack of negative regulation by microRNA needs to be tested experimentally.

Supplementary Figure 1 does not have a loading control

Supplementary Figure 2 is underpowered (n=3)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

In the manuscript “Eighty-nine asthma risk variants highlight the role of T cell regulation, inflammation and airway remodeling in asthma pathogenesis” Olafsdottir et al. used two large study populations from deCODE AND UK Biobank in (UKB) to conduct a meta-GWAS on asthma. They report 89 independent signals at 56 loci of which 21 signals at 17 loci represented novel asthma loci. The respective candidate genes were identified based on functional annotations, eQTL data, and regulatory sites covered by the associated SNPs. In addition, for two variants genotype-specific expression data are generated. The reported genes are involved in T-cell signaling, innate immunity and airway remodeling.

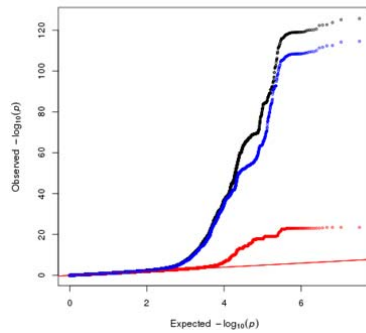
Major comment

The manuscript is well written and the methods are sound. Major assets of the study are the large data sets, providing power for discovery of novel loci. The phenotype definitions are appropriate and the replication of a large number of known asthma/allergy loci points to the validity of the approach. Nonetheless several points need to be clarified.

1. For imputation, the authors used different methods and different reference sequences. They should comment on this and whether it might have introduced a bias. They do not report any QQ plots, neither for the individual populations nor for the combined analysis. Please provide the plots and information on genomic inflation factors.

Response:

Both datasets are imputed based on very large reference sets, the Icelandic population specific and the HRC panel, for UK biobank, that includes a large number of UK individuals. This results in high quality imputation for both datasets, better than could be obtained by using for both populations common publicly available reference set, such as the 1000 genomes training set. There can still be difference in the imputation quality for individual markers between the datasets, primarily due to the different chip genotype platforms used rather than the different reference set used in the imputation. This should, however, not lead to false positive results as low imputation quality, unless correlated with affection status, would simply reduce the power to detect association. All novel associated variants we report have imputation info > 0.87 in both datasets. We have added the QQ plot (shown below) as supplementary figure 6 and information on the genomic inflation factors, estimated using LD-score regression (Bulik-Sullivan, B.K., Nat Gen, 2015) has been added to the materials and methods chapter as described in next response. Note that the large inflation factors for the Icelandic dataset are due to the relatedness of the individuals in the cohort (the Icelandic dataset includes majority of the Icelandic population).



Supplementary Figure 6. Quantile-quantile plot (QQ-plot) showing chi-square statistics corrected for genomic control and showing variants found in UK and UK cohorts for the GWAS of Asthma in Iceland (red) the UK (blue) and meta-analysis of the two cohorts (black). The red diagonal line represents expected distribution assuming no inflation of the chi-square statistics.

2. Essentially, genome-wide significant association signals are derived from UKB, the majority of loci does not replicate in the Icelandic population. Please comment on this.

Response: We note that the effective sample size is four times bigger for the UK Biobank samples than for the Icelandic samples, hence most of the power in the analysis comes from the UK samples. We have added a description of this in the Material and Methods section page 15 "In total we had 16,247 asthma cases and 346,486 controls from Iceland and 52,942 cases and 355,713 controls from the UK Biobank. Estimating the effective number of cases as $2 \times N_a \times N_c / (N_a + N_c) / GC$, where N_a and N_c are the number of cases and controls, respectively, and GC is the genomic adjustment factor, estimated using LD-score regression (Bulik-Sullivan, B.K., Nat Gen, 2015), we can calculate the effective sample size for Iceland to be $N_{eff}=22,689$ and UK Biobabank to be $N_{eff}=84,396$, $GC=1.368$ in Iceland and $GC=1.092$ in UK Biobank". We further note there is no significant heterogeneity ($P < 0.05/88 = 5.62E-04$) in effect sizes between the two sample sets for vast majority of the variants, hence the lack of significance in the Icelandic data simply reflects a lower power to detect association. We have added a sentence on this on page 3 "Based on number of cases and controls, we note that the effective sample size is four times bigger for the UK samples ($N=84,396$) than the Icelandic ($N=22,689$), hence most of the power comes from the UK samples. We note that only 4 out of the 88 variants show significant heterogeneity based on the number of variants tested ($P < 0.05/8 = 5.62 \times 10^{-4}$) and 28 show nominal heterogeneity ($P < 0.05$) in effect sizes between the two sample sets demonstrating a good consistency of effect for vast majority of the variants in the Icelandic samples. Further, 85 out of the 88 variants ($P < 1 \times 10^{-16}$) have effect size in the same direction in the two cohorts and 44 are at least nominally significant in Iceland showing a good replication of effects (Supplementary Table 1)"

3. What is the advantage of the Mantel-Haenszel model used for meta analysis. It assumes homogeneous ORs in the study populations but some loci show heterogeneity. How does this affect the validity of the results? Are the results different from the results calculated with METAL?

Response: The primary justification for using a fixed effect meta-analysis is that in majority of cases we do not detect a significant heterogeneity in the effect estimates, i.e. only 4 out of 88 variants have different effect estimates given the number of variants tested ($P < 0.05/88 = 5.62 \times 10^{-4}$). However, to investigate whether this assumption affects our result we have re-evaluated all the asthma variants using a sample size weighted meta-analysis using METAL. All of the lead variants at the reported 56 loci are GWAS significant even when we do not assume homogeneous OR in the study populations. The choice of Mantel-Haenszel model does therefore not affect the validity of the results. We have added the P values from the METAL analysis in supplementary Table 1.

Additional points

4. On page 3, how was the conditional analysis performed? If conditioning on the lead variant identified more than one additional genome-wide significant signal, how was the number of independent signals determined? This should be explained in the methods section in more detail.

Response: We have added more information on the conditional analysis in the material and methods section page 17 „Approximate conditional analyses (COJO), implemented in the GCTA software⁵¹ were performed on lead variants, defined by lowest P value at each genomic region (locus), to identify possible secondary signals. LD between variants was estimated using a set of 8700 whole-genome –sequenced Icelanders. The analysis was restricted to variants within 1 Mb from the index variants that were present in both the Icelandic and the UK datasets. Based on the number of variants tested within the 55 loci (excluding the HLA region, $N = 3,993,179$), we chose a conservative P value threshold of $< 5 \times 10^{-8}$ for secondary signals.” In the review process we realized that for some of the loci a P value cut-off of $< 1 \times 10^{-7}$ was used. Re-evaluating those secondary signals with the cut-off $< 5 \times 10^{-8}$ did not change anything except for one secondary signal (rs3781094) at the GATA3 locus that did not reach this threshold and was therefore removed from supplementary table 1. The total number of variants reaching our threshold therefore, changed from 89 to 88.

5. On page 5, replace IL21 with IL21R

Response: This has been corrected

6. On page 7, the effect of the IL33 variant was not only greater in EOA but also in LOA than in AA. Accordingly, this might be a locus specific for non-allergic asthma.

Response: We agree that in our data the effect of the IL33 variant does not seem to be driven by allergic asthma. It has the greatest effect on EOA, followed by LOA whereas smallest effect, that does not reach nominal significance, on AA. However, we lack information on AA, especially in the UK samples where only 542 (1% of all asthma cases) have this diagnosis and there is overlap both with the early and late onset asthma cases. Therefore, this warrants further investigation. We have therefore, added this to page 7-8 “Further, the effect of IL33 (rs149045797) variant was largest on EOA (OR=0.53, $P = 3.8 \times 10^{-9}$) followed by LOA (OR=0.71, $P = 5.1 \times 10^{-9}$) with smallest effect on AA (OR=0.93, $P = 0.70$; Figure 4 and supplementary table 13). This variant is fully correlated ($r^2 = 1$) with a loss of function variant (rs146597587) previously reported to associate with reduced IL33 mRNA expression and asthma risk and in line with the results reported here, the largest effect was observed in young children that were hospitalized at least 4 times before the age of 6 years due to asthma exacerbation (OR=0.24, $P = 0.04$)³³. However, 58% of those young children had atopic diagnosis before 6 years of age⁴⁰ indicating that the IL33 association is driven by severe asthma rather than non-AA exclusively. We note that we lack information on AA for a large part of our samples, especially in the

UK samples where only 542 (1% of all asthma cases) have this diagnosis and there is overlap both with the early and late onset asthma cases. Therefore, the IL33 association with asthma severity in AA and non-AA warrants further investigation.”

7. On page 7, the correlation p and rsq values for asthma vs allergic rhinitis and allergic dermatitis do not agree with supplementary figure 5a and 5b respectively.

Response: This has been corrected.

8. On page 20 (methods), please specify the Biolegend antibodies.

Response: This has been added

9. In figure 2b and c, please explain the number of dots.

Response: The following has been added in the legend to Figure 2 “The dots in plots b-d represent individual experiments. The dots in plots e and f represent individual donors. Lines in panel b-d indicate median level”.

10. In figure 3, please explain the sequences in more detail. Why are some bases on top of the 5'-sequence or below the 3' sequence? Why do the red sequences on the left and on the right differ?

Response: Although the question is not entirely clear, we assume that the reviewer is referring to the left-hand side of the figure showing the primary miR-142. This figure shows the folded transcript hence, the sequence is shown at the RNA level. The figure legend additionally clearly states that this is the „transcript“. As to the latter part of the question, there was a missing „c“ on the 3p sequence for which we apologize and we have now corrected this mistake.

11. In table 1, please check the overlap with Ferreira et al. (Nat Genet. 2017 Dec;49(12):1752-1757) and with Shrine et al. (Lancet Respir Med. 2019 Jan;7(1):20-34).

Response: We have carefully checked the overlap with Ferreira et al which is reported in supplementary Table 5. The Shrine paper was not published when we submitted our manuscript but we observe that two of the novel signals claimed in their publication are represented at 2 of the novel loci reported by us. rs12788104 (chr11:1129831) reported by us at the MUC2/MUC5AC locus has low correlation with the variant reported in Shrine et al as a novel variant at MUC2/MUC5AC locus rs11603634 (chr11:1142570; $r^2=0.35$). The rs11603634 does not reach significance in our study ($P=1.20E-05$, effect=1.03). Our secondary signal at the KIAA1109 locus (rs17454584) is highly correlated with the novel signal Shrine reported at this locus rs72687036 ($r^2=0.95$) whereas our primary signal rs45613035 is not ($r^2=0.02$). The variant at the GATA3 locus does not correlate with any of the 7 independent variants we report at this locus. The discrepancy between the two studies could be explained by the different selection criteria where we included all asthma cases whereas they only selected moderate-severe asthma subjects from UK Biobank under the following criteria; Asthma – doctor diagnosis (data field 6152), Exclude doctor diagnosed Emphysema/chronic bronchitis (data field 6152) and moderate-severe asthma based on British Thoracic Society (BTS) guidelines (2014) and the British National Formulary (September 2015) (data field 20003).

We have marked these variants in Table 1, and added the following text below the table “ + loci reported in Shrine et al.⁴⁶ while this paper was in review. The novel variant reported by Shrine et al.

at MUC5AC/MUC2 locus (rs11603634) is weakly correlated with a variant we report here at this locus (rs12788104; $r^2=0.35$) and the reported variant at KIAA1109 locus (rs72687036) is highly correlated with a secondary signal (rs17454584) we report here at that locus. The discrepancy between the two studies could be explained by the different definitions of the asthma phenotypes; we included all asthma cases whereas they only selected moderate-severe asthma cases.

12. On page 16, the p-value for the secondary signal should not be based on the number of loci (n=56). Please change this statement. The applied threshold is correct.

Response: The reviewer is correct that it is based on the number of variants but not loci. We have adjusted the text on page 17 related to this “Based on the number of variants tested within the 55 loci (excluding the HLA region, $N=3,993,179$, we chose a conservative P value threshold of $<5 \times 10^{-8}$ for secondary signals.”

13. In supplementary fig 1, please show the whole size range and also the loading control.

Response: In combination with an increased number of experiments ($n=6$ now instead of $n=3$ before), we have added the relevant loading controls to Supplementary Figure 1. As we are using WES (Protein simple, BioTechne) for our analyses instead of conventional Western blot, CD30 and total protein quantifications were run in separate wells from the same lysis preparation. For the normalization calculation we followed instructions from Protein simple for the total protein module (Biotechne DM-TP01). Precursor and mature CD30 were normalized to total protein.

14. In supplementary fig 4, delete “Ref” in the header or add the reference.

Response: This has been corrected by adding the reference (Ferreira et al. 2017). Furthermore, the secondary header was removed “Reported allergy variants and risk of asthma”

15. In supplementary fig 5, the scaling of both axes should be identical.

Response: This has been changed in a new supplementary fig 5.

16. In supplementary tab 1, what was conditioned on in the combined analysis? Comment on heterogeneity of effects? Which threshold was used and why?

At the IL13 locus, would conditioning on rs848 eliminate the other 2 signals? Could this be the causal variant?

Response: More information regarding the conditional analysis has been added to the material and methods section as described in a previous response regarding the conditional analysis. In the conditional analysis we start with conditioning on the most significant signal at this locus (in the case of IL13 locus it was the rs848) and continue with step-wise regression using all variants reaching the $P < 5 \times 10^{-8}$ cut-off. In the end, variants reaching the P value cutoff are then used to condition on the top signal so that all variants reaching a given P value cut-off are used to condition on. That means that the rs848 did not reach the $P < 5 \times 10^{-8}$ cut-off after conditioning on rs3749833 and rs60156285 whereas those two signals reached the $P < 5 \times 10^{-8}$ cut-off (and in fact got stronger P values) when conditioned on rs848.

17. In supplementary tab 2, please add the risk allele for the previously reported SNPs in order to compare the effect direction.

Response: The table has been modified to report the risk allele based on information found in GWAS catalog.

18. In supplementary tab 3, what was conditioned on?

Response: This is the same conditional analysis as reported in supplementary table 1 and described above. A few inconsistencies were observed between the P values in the two tables and this has been corrected.

19. In all tables ordered according to the SNP positions, please check the order of SNPs.

Response: This has been checked and corrected.

20. In supplementary tab 5, the confidence interval is missing, rsq values are missing, p-values from Ferreira et al are missing.

Response: The CI, risk allele, p values and OR from the Ferreira paper have been added to supplementary table 5. The r^2 values were previously given in column Q but are now after the addition of columns in column W. This has been clarified by adding this header on top of column S " r^2 between Asthma lead SNP and SNP reported in Ferreira". The conditional analysis has been moved from the last column to column N.

21. Why are p-values from Ferreira et al different for the same SNP (e.g. rs73205303)?

Novel loci are not associated in the Icelandic population, please comment.

Response: We think this might be a misunderstanding due to the lack of a descriptive header for the r^2 column in the previous version. Indeed the p-value for the same SNP is always the same in Ferreira but we test if they correlate with all the significant markers we find at the previously reported Ferreira loci. For example, rs73205303 reported in Ferreira, is highly correlated ($r^2=0.97$) with our lead marker at that locus (rs11088309) but not ($r^2=0.02$) with the secondary signal (rs8129030) we report at that locus. With regard to the lack of significant association in the Icelandic dataset we refer to response no 2.

22. In supplementary tabs 7 to 9, why are only novel SNPs included? Since previous studies performed different functional analyses, the authors should report the results for all independent signals.

Response: While we agree with the reviewer that this would be interesting we also note that this requires a lot of work and the previously reported variants are not the main scope of this manuscript. Furthermore, the information provided in supplementary tables 7 to 9 can all be retrieved from publicly available databases and is therefore available for interested readers to look into for all variants presented in this manuscript. On the other hand we screened all the 88 variants for our in-house blood and adipose eQTL data as presented in supplementary table 6.

23. In supplementary tab 8, please report the effect alleles. Why is whole blood from GTEx not

included. It would be interesting to have a comparison of the results with the deCODE data (suppl tab 6).

Response: The effect alleles have been added. The deCODE whole blood data is based on much larger dataset (n=7007) than GTEx whole blood samples (n=369 genotyped individuals) and therefore the deCODE data is much more informative. Additionally, a comparison of two datasets is confounded by multiple factors, i.e. sequencing depth, RNA extraction and library preparation protocols, which is beyond the scope of our current study.

24. In supplementary tab 9, why were transcription factor binding sites not included?

Response: We have now included transcription factor binding sites in supplementary table 9. We identified minor alleles that disrupt a binding site for a given TF, using motifs from HOCOMOCO v11, and report binding sites supported by a chip-seq peak found in GTRD database.

25. In supplementary tab 10, for rs6926894, the OR of UKB is reported instead of deCODE, the CI is missing.

Response: This has been corrected and CI added.

26. In supplementary tab 12, the number of UKB asthma cases does not agree with page 2, why?

Response: Unfortunately, we do not have information on age of onset or allergic asthma for all cases, in either the UKB or Icelandic samples.

Reviewer #2 (Remarks to the Author):

The authors carried out the largest GWAS meta-analysis of asthma to date with over 69,000 cases and over 700,000 controls. In total 89 independent asthma variants were identified. Of these:

- 48 signals in 24 known asthma loci: 24 new/24 previously reported
- 20 signals in 15 known 'combined allergic phenotype (asthma/hay fever/eczema)' loci
- 21 signals in 17 new asthma loci

They also quantify the correlation of the effects with other asthma and allergic phenotypes and perform functional and eQTL analyses to investigate the implicated genes and their functional roles.

This work is novel in that it is the largest ever meta-analysis; the major claims are the discovery of 21 new independent asthma loci, of which 18 are common (MAF $\geq$ 5%) and 3 uncommon variants, 2 of the uncommon variants show a 'small' protective effect for asthma. Most of the genes implicated in the reported associations with asthma are involved in known T cell regulation, inflammation and airway remodeling.

I praise the authors for the detail of the data reported in the supplementary tables. These data will be essential for other researchers to be able to use the results of this meta-analysis. For instance, complete GWAS summary statistics are needed to perform Mendelian Randomization studies that I'm sure will follow these findings.

Comments

1. In the main text the authors state that the asthma phenotype used for this meta-analysis includes both definitions: doctor ever diagnose asthma and/or self-reported asthma. It is unclear whether the self-reported asthma is self-reported doctor diagnosis of asthma.

Can the authors comment on the validity of the cases? What proportion of cases were defined from a doctor diagnoses, self-reported asthma and self-reported doctor diagnosis? Could there be an over-reporting of asthma cases (false positives) and what would the implications be?

Response: We thank the reviewer for raising this important topic. We have now added the number of cases for self-reported asthma and doctors diagnoses in a supplementary table 12 (note that there is an overlap between the two diagnosis both in Iceland (N=3224) and in the UK (N=21092). All of the self-reported asthma is self-reported doctors diagnosis as described in materials and methods “self-reported by a positive reply to the question: “Has a doctor confirmed your asthma diagnosis”. We have added this into the main text on page 2 “The asthma phenotype in both sets was based on physician diagnoses and/or self-reported doctor’s diagnosis of asthma” In supplementary table 13 we report OR and P value for all the 88 variants for ICD10 codes and self-reported cohorts separately. Similar effect size is observed for all variants and therefore, we observe no evidence for a bias between the two asthma sample sets.

2. Main results are described in page 3. It took this reviewer several reads to have a clear understanding of the old/new findings. A suggested opening sentence would be: “Conditional analysis yield 89 independent signals: 48 of these signals in 24 known asthma loci (50% of the signals are new), 20 in 15 known combined (asthma-eczema-rhinitis) phenotype loci and 21 new signals in 17 new asthma loci.” And then expand from here.

Response: We thank the reviewer for this suggestion and have modified the text on page 3 in the following way: “Of the 88 independent signals, 47 are at 24 previously reported asthma loci, 20 variants at 15 loci have previously been reported for a combined allergic phenotype⁷ and 21 variants at 17 loci are novel asthma signals.”

3. The authors don’t seem to provide a description of the studied populations. How do the Iceland and UK populations compare in terms of general demographics?

Response: Information on year of birth and sex for the two samples has been added in supplementary table 17. The age range is wider in the Icelandic cohort mean Year of birth 1965.4 (SD:26.5) in Iceland and 1951.6 (SD: 8.2) in the UK whereas the gender ratio is similar 40.6% males in Iceland and 42.5% in the UK. The different age range is explained by the way recruitment was conducted in the UK biobank where all subjects were aged between 40 and 69 at recruitment whereas there were no age restrictions in the Icelandic recruitment.

4. Could the author comment on the magnitude and clinical implications of the protective effects? Comb OR (95%CI) = 0.82 (0.78, 0.87) for variant near TNFRSF8, Comb OR (95%CI) =0.89 (0.86, 0.92) for variant near TGFB1.

Response: *TNFRSF8* variant has the greatest protective effects of the 21 novel variants reported here. Given the heterogeneity of asthma this can be viewed as a considerable effect on asthma risk and in fact there are only four variants affecting asthma risk more than the *TNFRSF8* variant in our dataset (rs72782676 (intergenic on chr10, $OR_{comb}: 0.63(0.58-0.69)$), rs149045797 (IL33, $OR_{comb}: 0.65(0.59-0.72)$), rs12722502 (IL2RA, $OR_{comb}: 0.80(0.77-0.85)$) and rs61816761 (FLG, 1.24 (1.19-1.29))). Further evaluation of the effect size of those variants in different asthma endotypes will be needed to fully evaluate the clinical implication of those variants.

We have added a sentence on this on page 4: “The p.Cys273Tyr variant had the greatest protective effects ($OR=0.82$, $P=8.27 \times 10^{-13}$) of the 21 novel variants reported here and in fact there were only 4 out of the 88 asthma variants with greater effect on asthma risk; rs72782676 (intergenic on chr10, $OR=0.63$), rs149045797 (intron variant at *IL33* locus, $OR=0.65$), rs12722502 (intron variant at *IL2RA* locus, $OR=0.80$) and rs61816761 (stop-gained variant at *FLG* locus, $OR=1.24$).”

5. Table 1. Confidence intervals are not provided. Can the authors add the corresponding 95 % CIs (as in Suppl Table 1) or the standard errors so the effects can be fully interpreted together with the p-values?

Response: The 95% CI have been added into Table 1. In order to facilitate this we had to remove the column showing I^2 that can now only be observed in supplementary table 1.

6. The authors have investigated the effect of the asthma variants on early onset (first diagnosis ≤ 17 years) and late onset (first diagnosis > 17 years) asthma and found that for some variants the effect was stronger on early onset asthma compared to late onset asthma. Do the authors think that these results support the notion that early and late onset asthma have partly distinct genetic architectures? And if so, what implications can be drawn?

Response: The reviewer correctly pointed out that these results were not discussed further in the manuscript.

We have added a sentence on this on page 8. “These results point to a partly distinct genetic architecture behind the three asthma phenotypes tested in our analysis in line with the clinical heterogeneity of the different asthma phenotypes. Our results show that certain sequence variants have stronger effect in EOA than LOA indicating that genetics may play a larger role in EOA.”

Reviewer #3 (Remarks to the Author):

The findings might be interesting but the experimental part that meant to support genetic findings is weak and lacks appropriate quality controls.

Comments:

1. Figure 2 b and c are underpowered.

Response: We agree with the reviewer on this point and to that end, we have now increased the number of experiments to six. Further, we optimized our WES protocol so that we could better define precursor and mature forms of the CD30 protein. The results from these six experiments are consistent with, and confirmed our previous results showing that in whole cell lysate from HeLa cells expressing the p.Cys273Tyr variant we observe higher ratio of precursor/mature CD30 protein, compared to cells expressing the WT version ($P = 4.2 \times 10^{-7}$, Fig. 2b). Further, by increasing the number of the HeLa cell experiments we observe significantly lower CD30 surface expression on p.Cys273Tyr cells than the WT ($P = 7.4 \times 10^{-5}$, Fig. 2c) in line with what we observe on stimulated PBMCs. We have therefore, added information on this on page 4 in the following way: "CD30 is expressed as a precursor protein that undergoes post-translational modification, which results in the mature form of the protein^{18,19}. Lysates from cells expressing the p.Cys273Tyr variant had higher ratio of the precursor form compared to the mature form than that observed in cells expressing the WT CD30 ($P = 4.2 \times 10^{-7}$ 1.2×10^{-4} ; Fig.2b and Supplementary fig. 1). Moreover we observed a lower surface expression on p.Cys273Tyr cells than WT CD30 ($P = 7.4 \times 10^{-5}$, Fig. 2c)."

2. There is also a lack of evaluation of soluble CD30 in supernatants from HeLa cells transfected with different CD30 variants.

Response: To address this important point we measured the amount of soluble CD30 in the supernatants of HeLa cells expressing either WT or p.Cys273Tyr CD30. These experiments showed a markedly lower amounts of soluble CD30 generated from cells expressing the p.Cys273Tyr variant CD30 ($N = 6$, $P = 5.5 \times 10^{-4}$), consistent with our observation of lower soluble CD30 concentration in supernatants of PBMC from heterozygous carriers of p.Cys273Tyr. We have added a sentence on this on page 4: "and a significantly lower amount of sCD30 in the cell culture supernatant of cells expressing the p.Cys273Tyr variant ($P = 5.5 \times 10^{-4}$, Fig. 2d)."

3. Figure 2 d and e. There is no mention of statistical test used to test difference in CD30 surface expression and soluble levels in d and e. Authors linked heterozygous carriers with age and gender matched non carrier controls suggesting that paired t test was used. I am not convinced that this is appropriate. The data in d might be skew by two pairs.

Response: For the PBMC analyses we used two-tailed Wilcoxon matched-pairs signed rank test to test for significant differences in cell surface and soluble expression between p.Cys273Tyr and WT CD30. We have added further description to this in the material and methods chapter page 21.

4. The representative flow cytometry plots of CD30 surface expression for c-d are required.

Response: We have now incorporated into Figure 2 representative flow cytometry plots (histograms) depicting the expression levels of CD30 for HeLa cells expressing WT or p.Cys273Tyr CD30, and equivalent analysis of PBMC from a non-carrier and p.Cys273Tyr heterozygous carrier.

5. Figure 3. The conclusion that rs41283642 increases TGFb expression is not supported experimentally. Evaluation of surface TGFbR1 expression in people with different variants is recommended. Also, in order to claim that rs41283642 increases TGFb expression because of a lack of negative regulation by microRNA needs to be tested experimentally.

Response: As stated in the manuscript we do observe a significantly increased ($P=7.24 \times 10^{-101}$) expression of TGFBR1 mRNA in RNA sequencing data from peripheral blood cells from heterozygous carriers of rs41283642-T located within an micro RNA binding site for miR-142-3p in the 3' UTR of TGFBR1. The increase in TGFBR1 expression is substantial or 15.2% per rs41283642-T allele demonstrating a strong link between the rs41283642-T and TGFBR1 expression. Furthermore, others (Zhe Lei et al. 2014) have shown that the microRNA miR-142-3p regulates TGFBR1 expression by binding to the 3'UTR of TGFBR1 and suppressing its expression. This was demonstrated both at the RNA and protein level. The asthma associated variant alters the binding site of miR-142-3p in the 3'UTR of TGFBR1. Collectively these data provide strong evidence that rs41283642-T leads to increased expression of TGFBR1. We have described in more detail the results presented by Zhe Lei et al, in the result section of the revised manuscript (page 5), to provide further support for the increase in expression of TGFBR1 because of lack of negative regulation by miR-142-3p:

„Expression quantitative trait locus (eQTL) analysis showed that the variant associates with 15.2% increasing TGFBR1 expression in blood per allele ($P=7.24 \times 10^{-101}$, effect=1.18 SD, which is the most significant eQTL for TGFBR1; Fig. 3a-b, Supplementary Table 6).rs41283642 T changes a recognition site for microRNA (miR) miR-142-3p (Fig. 3c). MiR-142-3p has previously been shown by others¹⁹ to bind to the 3'UTR site in TGFBR1 and repress TGFBR1 expression at both RNA and protein levels. Interestingly, miR-142-3p is one of only three miRs reported to have an increased expression in severe asthmatic lungs²⁰. Together these data suggest that rs41283642 T reduces the binding of miR-142-3p to TGFBR1 transcripts leading to increased TGFBR1 expression and through that protection against asthma“

6. Supplementary Figure 1 does not have a loading control

Response: In combination with an increased number of experiments, we have now added the relevant loading controls to Supplementary Figure 1. As we are using WES (Protein simple, BioTechne) for our analyses instead of conventional Western blot, CD30 and total protein quantifications were run in separate wells from the same lysis preparation. For the normalization calculations we followed instructions from Protein simple for the total protein module (Biotechne DM-TP01). Precursor and mature CD30 were normalized to total protein.

7. Supplementary Figure 2 is underpowered (n=3)

Response: The experiments displayed in Supplementary Figure 2 involve primary cells donated by homozygous carriers of the p.Cys273Tyr variant. We know of sixteen homozygous carriers in Iceland that are alive and we are in the unique position to be able to recruit them based on their genotype. We attempted to recruit all sixteen homozygotes and at the time of submission, we had recruited four. In the meantime, we were able to recruit two additional homozygotes. Since we are dependent on the voluntary participation of individuals we contact, we currently have no means to increase the number of samples beyond these six. The current number of samples is six, and we have updated the Supplementary Figure 2 accordingly.

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

The authors have addressed most of the comments in an accurate manner and the manuscript has improved considerably.

However, there are still some concerns regarding genomic inflation which seems to be a major problem in both studies according to Supplementary Figure 6. The authors should present the results on genomic inflation in more detail in the results section (genomic inflation factors for both studies) and comment on this in the discussion section. What is the inflation factor in UKB and what are potential reasons for genomic inflation in UK Biobank? Have previous studies reported similar inflation in UKB? If population structure/relatedness is a major concern in the Icelandic study, why have the authors not used a method that effectively controls for these factors (eg Chen et al, AJHG 2016, PMID:27018471)

In general, the authors should be more careful in the interpretation of the data. Acknowledging the strengths of their study, they should also recognize potential weaknesses. Out of 88 loci, 29 loci show nominally significant heterogeneity between both studies which is well beyond the expected number. Likewise, only 11 loci reach the Bonferroni corrected p-value for association with asthma in the replication study. It is true that the replication study has less power, but this is attributed to the study design. A combined analysis of both data sets might have been the better choice. All these points need to be discussed in the manuscript, not in materials and methods. Non-replicating loci with an opposite effect direction in the replication study could be assessed as potentially new loci. Finally, loci which are strongly associated with overlapping allergic phenotypes such as eczema could have been identified due to this concomitant allergic disease since allergic co-morbidities were not excluded from the study. This should also be included in the discussion.

Minor points

Table 1 still contains known allergy loci ([ADAD1], IL2-[]-IL21) reported by Ferreira et al (Nat Genet. 2017 Dec;49(12):1752-1757).

In Supplementary Table 2, for SNPs in high LD with previously reported SNPs or for identical SNPs, the authors should clearly state whether the same risk allele replicates or not (additional column). If no data are available in the GWAS catalog, please also check the original manuscripts (e.g. Pickrell et al report alleles and effect directions in the supplementary data).

In Supplementary Table 7, the authors should add the p-values and OR of the coding SNPs.

Reviewer #2:

Remarks to the Author:

The authors have satisfactorily responded to all my comments.

Reviewer #3:

Remarks to the Author:

The experimental part improved greatly after revision. The data is sound.

I have just one small comment. It would be better for readers if the statistical evaluation is included in the figure 2 legend. Currently it is just mentioned in the text

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors have addressed most of the comments in an accurate manner and the manuscript has improved considerably.

Comment #1

However, there are still some concerns regarding genomic inflation which seems to be a major problem in both studies according to Supplementary Figure 6. The authors should present the results on genomic inflation in more detail in the results section (genomic inflation factors for both studies) and comment on this in the discussion section. What is the inflation factor in UKB and what are potential reasons for genomic inflation in UK Biobank? Have previous studies reported similar inflation in UKB? If population structure/relatedness is a major concern in the Icelandic study, why have the authors not used a method that effectively controls for these factors (eg Chen et al, AJHG 2016, PMID:27018471)

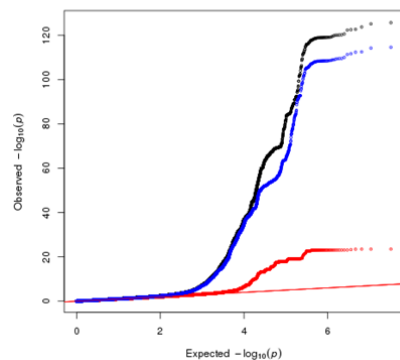
Response:

Regarding supplementary figure 6 then that QQ-plot primarily shows the excess of associations observed in the datasets due to real association signals. The figure also illustrates how much more power is in the UK Biobank dataset than the Icelandic dataset as the excess is much greater for the former. The figure does not indicate any problems in the studies.

The estimated genomic inflation in the two studies (estimated using LD-score regression to reduce the impact of real association signals) is 1.368 for the Icelandic dataset and 1.092 for the UK dataset. For the Icelandic dataset this inflation is due to relatedness of the study individuals. The Icelandic study includes about 360,000 individuals which is almost the whole Icelandic nation, hence every individual has many relatives in the study. This does not indicate any problem with the study and is adjusted for by reducing the test statistic by the inflation factor. For the UK Biobank relatedness is also the most likely cause of the inflation as population structure is adjusted for by including 40 principal components in the analysis. Although not collected as families, 30% of the participants have a relative (third degree or closer) in the UKB dataset and it includes 22,666 sibling pairs (cite: Bycroft C et.al. Nature 11 cot 2018). Large genomic inflation factors have been reported before for GWAS based on UK Biobank data, e.g $\lambda_g = 1.146$ in a study on hair color (cite: PMID: 29662168) and 1.23 in a study on intraocular pressure (cite: <https://doi.org/10.1038/s41588-018-0126-8>).

Methods that attempt to account for relatedness in the case control setting have other problems. In particular, rare variant associations are not properly handled and the association we do uses phenotype information from ungenotyped individuals by inferring genotype information from genotyped relatives (sometimes called familial imputation). This we can do as a very large fraction of the Icelandic population is genotyped and this method can increase the number of cases used in the study by about 75%. The familially imputed genotypes are less informative than imputed genotypes for genotyped individuals and this has to be appropriately accounted for in the association test (cite: <https://doi.org/10.1038/sdata.2015.11>). We have not yet found a method that consistently performs better than the currently applied one. The method we are currently using could potentially be improved to increase power, but it has the benefit of being robust and not produce false positive associations. We note that we are using the bolt method to account for relatedness for quantitative trait associations.

In our previous response we explained how we calculate the effective sample size for each cohort taking into account the genomic adjustment factor (GC), based on LD-score regression for both cohorts ($UK_{GC}=1.092$) and $Iceland_{GC}=1.368$) “In total we had 16,247 asthma cases and 346,486 controls from Iceland and 52,942 cases and 355,713 controls from the UK Biobank. Estimating the effective number of cases as $2 \times N_a \times N_c / (N_a + N_c) / GC$, where N_a and N_c are the number of cases and controls, respectively, and GC is the genomic adjustment factor, estimated using LD-score regression (Bulik-Sullivan, B.K., Nat Gen, 2015), we can calculate the effective sample size for Iceland to be $N_{eff}=22,689$ and UK Biobank to be $N_{eff}=84,396$, $GC=1.368$ in Iceland and $GC=1.092$ in UK Biobank”. This was already added in the Material and Methods section page 15. To further, bring this into the result session we have added a sentence on this on page 3: “ The estimated genomic inflation in the two studies (estimated using LD-score regression)⁷ is 1.368 for the Icelandic dataset and 1.092 for the UK dataset, reflecting the relatedness of individuals in Iceland and is also the most likely explanation of genomic inflation in UK Biobank where 30% of the participants have a relative (third degree or closer) in the UKB dataset⁸. “



Supplementary Figure 6. Quantile-quantile plot (QQ-plot) showing chi-square statistics corrected for genomic control and showing variants found in UK and UK cohorts for the GWAS of Asthma in Iceland (red) the UK (blue) and meta-analysis of the two cohorts (black). The red diagonal line represents expected distribution assuming no inflation of the chi-square statistics.

Comment #2

In general, the authors should be more careful in the interpretation of the data. Acknowledging the strengths of their study, they should also recognize potential weaknesses. Out of 88 loci, 29 loci show nominally significant heterogeneity between both studies which is well beyond the expected number. Likewise, only 11 loci reach the Bonferroni corrected p-value for association with asthma in the replication study. It is true that the replication study has less power, but this is attributed to the study design. A combined analysis of both data sets might have been the better choice. All these points need to be discussed in the manuscript, not in materials and methods. Non-replicating loci with on opposite effect direction in the replication study could be assessed as potentially new loci. Finally, loci which are strongly associated with overlapping allergic phenotypes such as eczema could have been identified due to this concomitant allergic disease since allergic co-morbidities were not excluded from the study. This should also be included in the discussion.

Response:

As clearly stated in the abstract, main text and material and methods, this is a meta-analysis of the two sample sets and therefore, we have indeed performed a combined analysis of both datasets as the reviewer suggested.

Regarding the observed heterogeneity between the effect estimates in the two studies, it is worth noting that these are retrospective studies and the diagnostic criteria, ascertainment and age distribution differs between them. Furthermore, it is likely that environmental exposure differs between the two populations. For these reasons we do not expect the effect estimates to be exactly the same in the two populations and the question if we can detect heterogeneity becomes one of power. Consistent with this scenario, all the variants that show nominally significant heterogeneity are previously published signals, i.e. the signals that have the strongest association. And there is a strong correlation between the strength of the association signals and the significance of the heterogeneity. Thus the observed heterogeneity in the effect estimates probably shows that these two studies are large enough to have good power to detect the real difference in the phenotype definition and/or ascertainment between the studies. This is likely to be true for most meta-analysis of GWAS studies that are being done, where the GWAS studies usually differ in terms of ascertainment and/or phenotype definition; however in most cases the individual studies are not as large as here and the power to detect heterogeneity is much less.

This has been addressed by the following sentence on page 3: „This observed heterogeneity in the effect estimates indicates that these two studies are powered to detect the real difference in the phenotype definition and/or ascertainment between the studies that are coherent with retrospective studies as reported here“.

As pointed out in the result section, 85 out of the 88 variants (which is much more than would be expected randomly, $P < 1 \times 10^{-16}$) have effect size in the same direction in the two sample sets. The three variants that seem to have opposite effect direction are completely flat in Iceland as can be seen in supplementary table 1: rs1684466 (OR=1.00652, $P=6.86E-01$ in Iceland and OR=0.94365, $P=4.79E-15$ in UKB), rs17454584 (OR=0.999, $P=9.56E-01$ in Iceland and OR=1.06375, $P=8.15E-14$ in UKB), rs41284471 (OR=0.9905, $P=6.34E-01$ in Iceland and OR=1.05369, $P=9.98E-10$ in UKB).

The reviewer is correct that we did not exclude allergic co-morbidities from our asthma association studies. However, we note that only 980 individuals of our 16,247 (6%) asthma cases in Iceland have the atopic dermatitis (AD) diagnosis and only approx 100 individuals had the ICD10 code for AD (L20) in the UKB at the time of analysis. However, as this is a retrospective study, not designed to address

co-morbidities of asthma, we cannot exclude that co-morbidities are underreported/diagnosed and that those might affect associations of certain sequence variants with asthma.

This has been addressed by the following paragraph page 9: “Since this is a retrospective study, not designed to study co-morbidities of asthma and allergic diseases, we cannot exclude that some of the overlapping association of certain sequence variants with e.g. asthma and AD might be due to overlap of the two phenotypes. However, only 980 out 16,247 asthma cases in Iceland had the AD dermatitis diagnosis (6%) while AD was hardly reported as a single disease in UKB, making it impossible to accurately define asthmatics without AD.”

Minor points

Comment #3

Table 1 still contains known allergy loci ([ADAD1],IL2-[]-IL21) reported by Ferreira et al (Nat Genet. 2017 Dec;49(12):1752-1757).

Response: The reviewer is right and we apologize for this oversight. We have now moved the ADAD1 locus from Table 1 to supplementary Table 5 where we have summarized other loci reported by Ferreira et al, 2017. We have further, updated the number of novel variants accordingly to 19 at 16 loci and 22 variants at 16 loci that have previously been reported for a combined allergic phenotype

Comment #4

In Supplementary Table 2, for SNPs in high LD with previously reported SNPs or for identical SNPs, the authors should clearly state whether the same risk allele replicates or not (additional column). If no data are available in the GWAS catalog, please also check the original manuscripts (e.g. Pickrell et al report alleles and effect directions in the supplementary data).

Response: Unfortunately, it is not straight forward to understand what the risk allele is from the supplements in Pickrell et al. In the supplementary table for asthma they report the alleles as C/T, G/T etc. which one could believe would be the reference/non-reference allele. That would be in line with their description in the supplementary methods that „Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19.“ However, when we look the SNPs up in ncbi database using GRCh37 (hg19) there is inconsistency for many of the variants with the table reported in Pickrell et al. with regards to which allele is the reference allele. When we blindly follow the instructions that they report the effect of the non-reference allele we do not replicate the allele for any of the variants that are inconsistent with what is reported in the ncbi database. Whereas we replicate the effect for all variants that are in line with the ncbi database e.g. rs1723018 that according to ncbi is an A>G variant and is written as A/G in Pickrell have correlated risk alleles in our data and Pickrell. If we believe that the authors have intended their alleles to be written as reference/non-reference allele and they therefore are always reporting the effect size on the latter allele (whether or not that is the non-reference allele) then we replicate their findings 100% (see Table 1 below). However, it is impossible to tell from the Pickrell paper as it is and the same applies for Wan Yi. 2012 (PMID: 22561531), Torgerson (PMID:21804549) and Almoguera (PMID:27611488) which is probably the reason why the risk allele gets a question

mark in the GWAS catalog. Therefore, we have to stick to what is reported in the GWAS catalog and leave the risk allele with a question mark.

We have added a column for the correlated alleles between SNPs and whether or not we replicate the risk allele for other publications in Supplementary Table 2.

Table 1. Previously reported SNPs from Pickrell_JK, 2016 (PMID:27182965) and correlation of risk alleles with our study.

SNP reported in Pickrell et al	alleles as reported in Pickrell et al	Ref>alt alleles as reported in GRCh37*	OR reported in Pickrell	Risk allele in Pickrell**	Risk allele in our data	Correlated with our allele	Correlated alleles***
rs6683383	A/T	T>A	1.062	A	T	No	T=T; A=G
rs1723018	A/G	A>G	0.945	A	G	Yes	A=G; G=A
rs13412757	A/G	G>A	1.062	A	G	No	G=G; A=A
rs34290285	A/G	G>A	1.107	A	G	No	same variant
rs5743618	A/C	C>A	1.082	A	C	No	same variant
rs2244012	A/G	A>G	1.102	G	T	Yes	A=C; G=T
rs10957978	G/T	G>T	0.934	G	A	Yes	G=A; T=G
rs144829310	G/T	G>T	1.165	T	A	Yes	G=G; T=A
rs12413578	C/T	C>T	0.891	C	C	Yes	same variant
rs7936323	A/G	G>A	0.922	G	T	No	G=G; A=T
rs3784099	A/G	G>A	0.942	G	A	No	G=T; A=A
rs30519068	A/G	G>A	1.10	A	C	No	G=C; A=A
rs56375023	A/G	G>A	0.897	G	A	No	G=C; A=A
rs7203459	C/T	T>C	1.092	C	T	No	T=T; C=A
rs11655198	C/T	C>T	0.847	C	G	Yes	C=G; T=T

*non-reference allele according to GRCh37

** Risk alleles in Pickrell defined from their description in supplementary methodology that "Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19."

*** Correlated alleles according to https://idlink.nci.nih.gov/?var=rs1723018&pop=CEU&r2_d=2&tab=idproxy

Supplementary Table from Pickrell_JK, 2016 (PMID:27182965).

Index SNPs for Strongest Associations

cytoband	assay.name	scaffold	position	alleles	src	pvalue	OR	95% CI	gene.context
17q12	rs11655198	chr17	38026169	C/T	1	1.0×10^{-63}	0.847	[0.831,0.863]	[ZBPB2]
6p21.32	rs3104367	chr6	32603487	C/T	1	1.0×10^{-40}	0.872	[0.855,0.890]	HLA-DRB5---[]-HLA-DQA1
9p24.1	rs144829310	chr9	6208030	G/T	1	1.3×10^{-31}	1.165	[1.136,1.195]	RANBP6---[]-IL33
5q22.1	rs1837253	chr5	110401872	C/T	1	3.3×10^{-31}	0.876	[0.856,0.896]	SLC25A46---[]-TSLP
2q12.1	rs202011557	chr2	102913642	D/I	1	5.1×10^{-31}	0.842	[0.818,0.868]	IL1RL2---[]-IL1RL1
15q22.33	rs56375023	chr15	67448363	A/G	1	2.4×10^{-21}	0.897	[0.877,0.917]	[SMAD3]
11q13.5	rs7936323	chr11	76293758	A/G	1	1.4×10^{-16}	0.922	[0.905,0.940]	C11orf30---[]-LRRRC32
6p21.33	rs2428494	chr6	31322197	A/T	1	1.4×10^{-16}	0.920	[0.902,0.939]	HLA-C---[]-MICA
5q31.1	rs2244012	chr5	131901225	A/G	1	2.1×10^{-16}	1.102	[1.077,1.128]	[RAD50]
2q37.3	rs34290285	chr2	242698640	A/G	1	1.8×10^{-15}	1.107	[1.079,1.135]	[D2HGDH]
16p13.13	rs7203459	chr16	11230703	C/T	1	3.5×10^{-15}	1.092	[1.068,1.117]	[CLEC16A]
1q23.3	rs4233366	chr1	161159147	C/T	1	4.8×10^{-15}	1.090	[1.067,1.114]	B4GALT3---[]ADAMTS4
10p14	rs12413578	chr10	9049253	C/T	1	8.1×10^{-12}	0.891	[0.862,0.921]	GATA3---[]
8q21.13	rs10957978	chr8	81285139	G/T	1	1.1×10^{-11}	0.934	[0.915,0.952]	TPD52---[]---ZBTB10
15q22.2	rs10519068	chr15	61068704	A/G	1	3.8×10^{-11}	1.100	[1.069,1.132]	RORA---[]
4p14	rs5743618	chr4	38798648	A/C	1	3.9×10^{-11}	1.082	[1.057,1.107]	[TLR1]
6q15	rs58521088	chr6	90985198	A/T	1	7.1×10^{-11}	0.934	[0.915,0.954]	[BACH2]
12q13.3	rs3001426	chr12	57509055	C/T	1	1.4×10^{-10}	0.938	[0.919,0.956]	STAT6-[]---LRP1
3q28	rs73196739	chr3	188402471	C/T	1	6.5×10^{-9}	0.922	[0.897,0.948]	[LPP]
1q32.1	rs6683383	chr1	203100504	A/T	1	1.1×10^{-8}	1.062	[1.040,1.084]	[ADORA1]
2p25.1	rs13412757	chr2	8458080	A/G	1	1.3×10^{-8}	1.062	[1.040,1.084]	[]---ID2
1q24.2	rs1723018	chr1	167433420	A/G	1	1.4×10^{-8}	0.945	[0.926,0.963]	[CD247]
14q24.1	rs3784099	chr14	68749927	A/G	1	1.6×10^{-8}	0.942	[0.922,0.961]	[RAD51B]
7q22.3	rs6959584	chr7	105676505	C/T	1	2.0×10^{-8}	1.086	[1.055,1.117]	[CDHR3]
5q31.3	rs200634877	chr5	141529761	D/I	1	2.5×10^{-8}	0.940	[0.919,0.961]	[NDFIP1]
1q25.1	rs6691738	chr1	173152036	G/T	1	2.9×10^{-8}	0.943	[0.923,0.963]	TNFSF18---[]TNFSF4
1p36.22	rs662064	chr1	10557251	C/T	1	3.2×10^{-8}	0.942	[0.922,0.962]	[PEX14]

Comment #5

In Supplementary Table 7, the authors should add the p-values and OR of the coding SNPs.

Response: This has been added with the addition of EA/OA for the coding SNPs and correlation of alleles between the strongest asthma associated variant and the coding variant.

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily responded to all my comments.

Reviewer #3 (Remarks to the Author):

The experimental part improved greatly after revision. The data is sound.

I have just one small comment. It would be better for readers if the statistical evaluation is included in the figure 2 legend. Currently is just mentioned in the text

Response: This has been added.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors have addressed most of the comments in an accurate manner and the manuscript has improved considerably.

Comment #1

However, there are still some concerns regarding genomic inflation which seems to be a major problem in both studies according to Supplementary Figure 6. The authors should present the results on genomic inflation in more detail in the results section (genomic inflation factors for both studies) and comment on this in the discussion section. What is the inflation factor in UKB and what are potential reasons for genomic inflation in UK Biobank? Have previous studies reported similar inflation in UKB? If population structure/relatedness is a major concern in the Icelandic study, why have the authors not used a method that effectively controls for these factors (eg Chen et al, AJHG 2016, PMID:27018471)

Response:

Regarding supplementary figure 6 then that QQ-plot primarily shows the excess of associations observed in the datasets due to real association signals. The figure also illustrates how much more power is in the UK Biobank dataset than the Icelandic dataset as the excess is much greater for the former. The figure does not indicate any problems in the studies.

The QQ plot is not informative, because it shows the GC-corrected p-values. Hence, it lacks essential information on genomic inflation. In order to assess the characteristics of a p-value distribution and thus the validity of potentially true association signals, the authors should present the uncorrected p-values in Supplementary Figure 6.

The estimated genomic inflation in the two studies (estimated using LD-score regression to reduce the impact of real association signals) is 1.368 for the Icelandic dataset and 1.092 for the UK dataset. For the Icelandic dataset this inflation is due to relatedness of the study individuals. The Icelandic study includes about 360,000 individuals which is almost the whole Icelandic nation, hence every individual has many relatives in the study. This does not indicate any problem with the study and is adjusted for by reducing the test statistic by the inflation factor.

Indeed, there seems to be high degree of relatedness/population structure present in the Icelandic data set which accounts for genomic inflation. A high LD-score regression intercept indicates that genomic inflation is not due to true association signals but mainly due to population stratification. This is in contradiction to the first paragraph of the response where the authors state that „the excess of associations observed in the datasets“ is „due to real association signals.“

Genomic inflation needs to be discussed in more detail because it might effect the validity of the association results which is particularly true for the Icelandic data set. This could be a minor issue for common variants but may affect the false positive rate for low frequency and rare variants.

For the UK Biobank relatedness is also the most likely cause of the inflation as population structure is adjusted for by including 40 principle components in the analysis. Although not collected as families, 30% of the participants have a relative (third degree or closer) in the UKB dataset and it includes 22,666 sibling pairs (cite: Bycroft C et.al. Nature 11 cot 2018). Large genomic inflation factors have

been reported before for GWAS based on UK Biobank data, e.g. $\lambda_g = 1.146$ in a study on hair color (cite: PMID: 29662168) and 1.23 in a study on intraocular pressure (cite: <https://doi.org/10.1038/s41588-018-0126-8>).

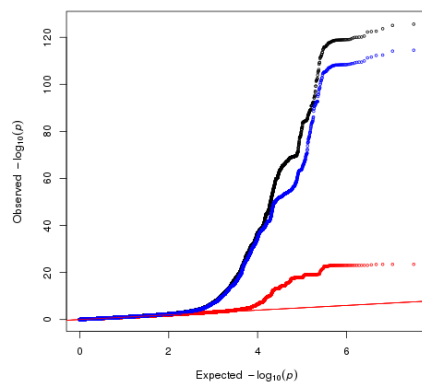
Both UKBB studies referenced here excluded related individuals. The IOP study reported a high lambda but a lower LD-score regression intercept. Likewise, the recently published UKBB GWAS on asthma (Johansson et al. 2019) excluded related individuals yielding an improved LD-score regression intercept of 1.065 for the same data set. Excluding related individuals is a standard QC step if association testing does not account for relatedness. Nonetheless, since the results presented by the authors mainly rely on slightly inflated UKBB data, the analysis should be accepted as it is.

Methods that attempt to account for relatedness in the case control setting have other problems. In particular, rare variant associations are not properly handled and the association we do uses phenotype information from ungenotyped individuals by inferring genotype information from genotyped relatives (sometimes called familial imputation). This we can do as a very large fraction of the Icelandic population is genotyped and this method can increase the number of cases used in the study by about 75%. The familiarly imputed genotypes are less informative than imputed genotypes for genotyped individuals and this has to be appropriately accounted for in the association test (cite: <https://doi.org/10.1038/sdata.2015.11>). We have not yet found a method that consistently performs better than the currently applied one. The method we are currently using could potentially be improved to increase power, but it has the benefit of being robust and not produce false positive associations. We note that we are using the bolt method to account for relatedness for quantitative trait associations.

The use of BOLT is not described in the methods section or in the paper. For which analysis have the authors used BOLT LMM which is only listed in the “Code availability” section? If the familiarly imputed genotypes are used in LMM methods for quantitative traits to account for relatedness why not on qualitative traits? Why is relatedness a problem for quantitative but not for qualitative traits?

Given that the UKBB carries most of the associations on its own, the Icelandic cohort has a lot less weight, and the issue of genomic inflation in the Icelandic set could be left to a proper discussion of the issue. The authors should explain the advantages of their method including of not accounting for relatedness (above paragraph) in detail in the manuscript (discussion section). A discussion on advantages and disadvantages is required (including references).

In our previous response we explained how we calculate the effective sample size for each cohort taking into account the genomic adjustment factor (GC), based on LD-score regression for both cohorts ($UK_{GC}=1.092$) and $Iceland_{GC}=1.368$) “In total we had 16,247 asthma cases and 346,486 controls from Iceland and 52,942 cases and 355,713 controls from the UK Biobank. Estimating the effective number of cases as $2 \times N_a \times N_c / (N_a + N_c) / GC$, where N_a and N_c are the number of cases and controls, respectively, and GC is the genomic adjustment factor, estimated using LD-score regression (Bulik-Sullivan, B.K., Nat Gen, 2015), we can calculate the effective sample size for Iceland to be $N_{eff}=22,689$ and UK Biobank to be $N_{eff}=84,396$, $GC=1.368$ in Iceland and $GC=1.092$ in UK Biobank”. This was already added in the Material and Methods section page 15. To further, bring this into the result section we have added a sentence on this on page 3: “The estimated genomic inflation in the two studies (estimated using LD-score regression)⁷ is 1.368 for the Icelandic dataset and 1.092 for the UK dataset, reflecting the relatedness of individuals in Iceland and is also the most likely explanation of genomic inflation in UK Biobank where 30% of the participants have a relative (third degree or closer) in the UKB dataset⁸.”



Supplementary Figure 6. Quantile-quantile plot (QQ-plot) showing chi-square statistics corrected for genomic control and showing variants found in UK and UK cohorts for the GWAS of Asthma in Iceland (red) the UK (blue) and meta-analysis of the two cohorts (black). The red diagonal line represents expected distribution assuming no inflation of the chi-square statistics.

Again, the authors should use uncorrected p-values for Supplementary Figure 6.

Comment #2

In general, the authors should be more careful in the interpretation of the data. Acknowledging the strengths of their study, they should also recognize potential weaknesses. Out of 88 loci, 29 loci show nominally significant heterogeneity between both studies which is well beyond the expected number. Likewise, only 11 loci reach the Bonferroni corrected p-value for association with asthma in the replication study. It is true that the replication study has less power, but this is attributed to the study design. A combined analysis of both data sets might have been the better choice. All these points need to be discussed in the manuscript, not in materials and methods. Non-replicating loci with on opposite effect direction in the replication study could be assessed as potentially new loci. Finally, loci which are strongly associated with overlapping allergic phenotypes such as eczema could have been identified due to this concomitant allergic disease since allergic co-morbidities were not excluded from the study. This should also be included in the discussion.

Response:

As clearly stated in the abstract, main text and material and methods, this is a meta-analysis of the two sample sets and therefore, we have indeed performed a combined analysis of both datasets as the reviewer suggested.

Regarding the observed heterogeneity between the effect estimates in the two studies, it is worth noting that these are retrospective studies and the diagnostic criteria, ascertainment and age distribution differs between them. Furthermore, it is likely that environmental exposure differs between the two populations. For these reasons we do not expect the effect estimates to be exactly the same in the two populations and the question if we can detect heterogeneity becomes one of power. Consistent with this scenario, all the variants that show nominally significant heterogeneity are previously published signals, i.e. the signals that have the strongest association. And there is a strong correlation between the strength of the association signals and the significance of the heterogeneity. Thus the observed heterogeneity in the effect estimates probably shows that these two studies are large enough to have good power to detect the real difference in the phenotype definition and/or ascertainment between the studies. This is likely to be true for most meta-analysis of GWAS studies that are being done, where the GWAS studies usually differ in terms of ascertainment and/or phenotype definition; however in most cases the individual studies are not as large as here and the power to detect heterogeneity is much less.

This has been addressed by the following sentence on page 3: „This observed heterogeneity in the effect estimates indicates that these two studies are powered to detect the real difference in the phenotype definition and/or ascertainment between the studies that are coherent with retrospective studies as reported here“.

As pointed out in the result section, 85 out of the 88 variants (which is much more than would be expected randomly, $P < 1 \times 10^{-16}$) have effect size in the same direction in the two sample sets. The three variants that seem to have opposite effect direction are completely flat in Iceland as can be seen in supplementary table 1: rs1684466 (OR=1.00652, $P=6.86E-01$ in Iceland and OR=0.94365, $P=4.79E-15$ in UKB), rs17454584 (OR=0.999, $P=9.56E-01$ in Iceland and OR=1.06375, $P=8.15E-14$ in UKB), rs41284471 (OR=0.9905, $P=6.34E-01$ in Iceland and OR=1.05369, $P=9.98E-10$ in UKB).

The reviewer is correct that we did not exclude allergic co-morbidities from our asthma association studies. However, we note that only 980 individuals of our 16,247 (6%) asthma cases in Iceland have the atopic dermatitis (AD) diagnosis and only approx 100 individuals had the ICD10 code for AD (L20) in the UKB at the time of analysis. However, as this is a retrospective study, not designed to address

co-morbidities of asthma, we cannot exclude that co-morbidities are underreported/diagnosed and that those might affect associations of certain sequence variants with asthma.

This has been addressed by the following paragraph page 9: “Since this is a retrospective study, not designed to study co-morbidities of asthma and allergic diseases, we cannot exclude that some of the overlapping association of certain sequence variants with e.g. asthma and AD might be due to overlap of the two phenotypes. However, only 980 out 16,247 asthma cases in Iceland had the AD dermatitis diagnosis (6%) while AD was hardly reported as a single disease in UKB, making it impossible to accurately define asthmatics without AD.”

Minor points

Comment #3

Table 1 still contains known allergy loci ([ADAD1],IL2-[]-IL21) reported by Ferreira et al (Nat Genet. 2017 Dec;49(12):1752-1757).

Response: The reviewer is right and we apologize for this oversight. We have now moved the ADAD1 locus from Table 1 to supplementary Table 5 where we have summarized other loci reported by Ferreira et al, 2017. We have further, updated the number of novel variants accordingly to 19 at 16 loci and 22 variants at 16 loci that have previously been reported for a combined allergic phenotype

Comment #4

In Supplementary Table 2, for SNPs in high LD with previously reported SNPs or for identical SNPs, the authors should clearly state whether the same risk allele replicates or not (additional column). If no data are available in the GWAS catalog, please also check the original manuscripts (e.g. Pickrell et al report alleles and effect directions in the supplementary data).

Response: Unfortunately, it is not straight forward to understand what the risk allele is from the supplements in Pickrell et al. In the supplementary table for asthma they report the alleles as C/T, G/T etc. which one could believe would be the reference/non-reference allele. That would be in line with their description in the supplementary methods that „Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19.“ However, when we look the SNPs up in ncbi database using GRCh37 (hg19) there is inconsistency for many of the variants with the table reported in Pickrell et al. with regards to which allele is the reference allele. When we blindly follow the instructions that they report the effect of the non-reference allele we do not replicate the allele for any of the variants that are inconsistent with what is reported in the ncbi database. Whereas we replicate the effect for all variants that are in line with the ncbi database e.g. rs1723018 that according to ncbi is an A>G variant and is written as A/G in Pickrell have correlated risk alleles in our data and Pickrell. If we believe that the authors have intended their alleles to be written as reference/non-reference allele and they therefore are always reporting the effect size on the latter allele (whether or not that is the non-reference allele) then we replicate their findings 100% (see Table 1 below). However, it is impossible to tell from the Pickrell paper as it is and the same applies for Wan YI. 2012 (PMID: 22561531), Torgerson (PMID:21804549) and Almoguera (PMID:27611488) which is probably the reason why the risk allele gets a question

mark in the GWAS catalog. Therefore, we have to stick to what is reported in the GWAS catalog and leave the risk allele with a question mark.

We have added a column for the correlated alleles between SNPs and whether or not we replicate the risk allele for other publications in Supplementary Table 2.

Table 1. Previously reported SNPs from Pickrell_JK, 2016 (PMID:27182965) and correlation of risk alleles with our study.

SNP reported in Pickrell et al	alleles as reported in Pickrell et al	Ref-alt alleles as reported in GRCh37*	OR reported in Pickrell	Risk allele in Pickrell**	Risk allele in our data	Correlated with our allele	Correlated alleles***
rs6683383	A/T	T>A	1.062	A	T	No	T=T; A=G
rs1723018	A/G	A>G	0.945	A	G	Yes	A=G; G=A
rs13412757	A/G	G>A	1.062	A	G	No	G=G; A=A
rs34290285	A/G	G>A	1.107	A	G	No	same variant
rs5743618	A/C	C>A	1.082	A	C	No	same variant
rs2244012	A/G	A>G	1.102	G	T	Yes	A=C; G=T
rs10957978	G/T	G>T	0.934	G	A	Yes	G=A; T=G
rs144829310	G/T	G>T	1.165	T	A	Yes	G=G; T=A
rs12413578	C/T	C>T	0.891	C	C	Yes	same variant
rs7936323	A/G	G>A	0.922	G	T	No	G=G; A=T
rs3784099	A/G	G>A	0.942	G	A	No	G=T; A=A
rs10519068	A/G	G>A	1.10	A	C	No	G=C; A=T
rs56375023	A/G	G>A	0.897	G	A	No	G=C; A=A
rs7203459	C/T	T>C	1.092	C	T	No	T=T; C=A
rs11655198	C/T	C>T	0.847	C	G	Yes	C=G; T=T

*non-reference allele according to GRCh37

** Risk alleles in Pickrell defined from their description in supplementary methodology that "Throughout the paper we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19."

*** Correlated alleles according to https://idlink.nci.nih.gov/?var=rs1723018&pop=CEU&r2_d=r2&tab=ldproxy

Supplementary Table from Pickrell_JK, 2016 (PMID:27182965).

Index SNPs for Strongest Associations

cytoband	assay.name	scaffold	position	alleles	src	pvalue	OR	95% CI	gene.context
17q12	rs11655198	chr17	38026169	C/T	1	1.0×10^{-63}	0.847	[0.831,0.863]	[ZPBP2]
6p21.32	rs3104367	chr6	32603487	C/T	1	1.0×10^{-40}	0.872	[0.855,0.890]	HLA-DRB5---[]-HLA-DQA1
9p24.1	rs144829310	chr9	6208030	G/T	1	1.3×10^{-31}	1.165	[1.136,1.195]	RANBP6---[]-IL33
5q22.1	rs1837253	chr5	110401872	C/T	1	3.3×10^{-31}	0.876	[0.856,0.896]	SLC25A46---[]-TSLP
2q12.1	rs202011557	chr2	102913642	D/I	1	5.1×10^{-31}	0.842	[0.818,0.868]	IL1RL2--[]--IL1RL1
15q22.33	rs56375023	chr15	67448363	A/G	1	2.4×10^{-21}	0.897	[0.877,0.917]	[SMAD3]
11q13.5	rs7936323	chr11	76293758	A/G	1	1.4×10^{-16}	0.922	[0.905,0.940]	C11orf30--[]--LRRC32
6p21.33	rs2428494	chr6	31322197	A/T	1	1.4×10^{-16}	0.920	[0.902,0.939]	HLA-C--[]--MICA
5q31.1	rs2244012	chr5	131901225	A/G	1	2.1×10^{-16}	1.102	[1.077,1.128]	[RAD50]
2q37.3	rs34290285	chr2	242698640	A/G	1	1.8×10^{-15}	1.107	[1.079,1.135]	[D2HGDH]
16p13.13	rs7203459	chr16	11230703	C/T	1	3.5×10^{-15}	1.092	[1.068,1.117]	[CLEC16A]
1q23.3	rs4233366	chr1	161159147	C/T	1	4.8×10^{-15}	1.090	[1.067,1.114]	B4GALT3--[]ADAMTS4
10p14	rs12413578	chr10	9049253	C/T	1	8.1×10^{-12}	0.891	[0.862,0.921]	GATA3---[]
8q21.13	rs10957978	chr8	81285139	G/T	1	1.1×10^{-11}	0.934	[0.915,0.952]	TPD52---[]---ZBTB10
15q22.2	rs10519068	chr15	61068704	A/G	1	3.8×10^{-11}	1.100	[1.069,1.132]	RORA---[]
4p14	rs5743618	chr4	38798648	A/C	1	3.9×10^{-11}	1.082	[1.057,1.107]	[TLR1]
12q15	rs58521088	chr6	90985198	A/T	1	7.1×10^{-11}	0.934	[0.915,0.954]	[BACH2]
12q13.3	rs3001426	chr12	57509055	C/T	1	1.4×10^{-10}	0.938	[0.919,0.956]	STAT6-[]--LRP1
3q28	rs73196739	chr3	188402471	C/T	1	6.5×10^{-9}	0.922	[0.897,0.948]	[LPP]
1q32.1	rs6683383	chr1	203100504	A/T	1	1.1×10^{-8}	1.062	[1.040,1.084]	[ADORA1]
2p25.1	rs13412757	chr2	8458080	A/G	1	1.3×10^{-8}	1.062	[1.040,1.084]	[]---ID2
1q24.2	rs1723018	chr1	167433420	A/G	1	1.4×10^{-8}	0.945	[0.926,0.963]	[CD247]
14q24.1	rs3784099	chr14	68749927	A/G	1	1.6×10^{-8}	0.942	[0.922,0.961]	[RAD51B]
7q22.3	rs6959584	chr7	105676505	C/T	1	2.0×10^{-8}	1.086	[1.055,1.117]	[CDHR3]
5q31.3	rs200634877	chr5	141529761	D/I	1	2.5×10^{-8}	0.940	[0.919,0.961]	[NDFIP1]
1q25.1	rs6691738	chr1	173152036	G/T	1	2.9×10^{-8}	0.943	[0.923,0.963]	TNFSF18---[]TNFSF4
1p36.22	rs662064	chr1	10557251	C/T	1	3.2×10^{-8}	0.942	[0.922,0.962]	[PEX14]

Comment #5

In Supplementary Table 7, the authors should add the p-values and OR of the coding SNPs.

Response: This has been added with the addition of EA/OA for the coding SNPs and correlation of alleles between the strongest asthma associated variant and the coding variant.

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily responded to all my comments.

Reviewer #3 (Remarks to the Author):

The experimental part improved greatly after revision. The data is sound.

I have just one small comment. It would be better for readers if the statistical evaluation is included in the figure 2 legend. Currently is just mentioned in the text

Response: This has been added.

Additional comment:

A UKBB study on asthma has recently been published (Hum Mol Genet. 2019 Jul 30. pii: ddz175. doi: 10.1093/hmg/ddz175) which should be referenced, and novelty of loci adjusted.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors have addressed most of the comments in an accurate manner and the manuscript has improved considerably.

Comment #1

However, there are still some concerns regarding genomic inflation which seems to be a major problem in both studies according to Supplementary Figure 6. The authors should present the results on genomic inflation in more detail in the results section (genomic inflation factors for both studies) and comment on this in the discussion section. What is the inflation factor in UKB and what are potential reasons for genomic inflation in UK Biobank? Have previous studies reported similar inflation in UKB? If population structure/relatedness is a major concern in the Icelandic study, why have the authors not used a method that effectively controls for these factors (eg Chen et al, AJHG 2016, PMID:27018471)

Response:

Regarding supplementary figure 6 then that QQ-plot primarily shows the excess of associations observed in the datasets due to real association signals. The figure also illustrates how much more power is in the UK Biobank dataset than the Icelandic dataset as the excess is much greater for the former. The figure does not indicate any problems in the studies.

The QQ plot is not informative, because it shows the GC-corrected p-values. Hence, it lacks essential information on genomic inflation. In order to assess the characteristics of a p-value distribution and thus the validity of potentially true association signals, the authors should present the uncorrected p-values in Supplementary Figure 6.

We have added panels to Supplementary Figure 6 with the uncorrected P-values.

The estimated genomic inflation in the two studies (estimated using LD-score regression to reduce the impact of real association signals) is 1.368 for the Icelandic dataset and 1.092 for the UK dataset. For the Icelandic dataset this inflation is due to relatedness of the study individuals. The Icelandic study includes about 360,000 individuals which is almost the whole Icelandic nation, hence every individual has many relatives in the study. This does not indicate any problem with the study and is adjusted for by reducing the test statistic by the inflation factor.

Indeed, there seems to be high degree of relatedness/population structure present in the Icelandic data set which accounts for genomic inflation. A high LD-score regression intercept indicates that genomic inflation is not due to true association signals but mainly due to population stratification. This is in contradiction to the first paragraph of the response where the authors state that „the excess of associations observed in the datasets“ is „due to real association signals.“ Genomic inflation needs to be discussed in more detail because it might effect the validity of the association results which is particularly true for the Icelandic data set. This could be a minor issue for common variants but may affect the false positive rate for low frequency and rare variants.

The reviewer is correct, the high correction factor in the Icelandic dataset is because of relatedness between the individuals. The first paragraph in the response stating that „the excess of associations observed in the datasets“ is „due to real association signals“ was incorrect. We have added QQ plots

for asthma stratified on minor allele frequency to the supplement (Supplementary Figure 6 c-d). These show that there is not a systematic problem with false positive at lower frequencies.

For the UK Biobank relatedness is also the most likely cause of the inflation as population structure is adjusted for by including 40 principle components in the analysis. Although not collected as families, 30% of the participants have a relative (third degree or closer) in the UKB dataset and it includes 22,666 sibling pairs (cite: Bycroft C et.al. Nature 11 cot 2018). Large genomic inflation factors have been reported before for GWAS based on UK Biobank data, e.g $\lambda_g = 1.146$ in a study on hair color (cite: PMID: 29662168) and 1.23 in a study on intraocular pressure (cite: <https://doi.org/10.1038/s41588-018-0126-8>).

Both UKBB studies referenced here excluded related individuals. The IOP study reported a high lambda but a lower LD-score regression intercept. Likewise, the recently published UKBB GWAS on asthma (Johansson et al. 2019) excluded related individuals yielding an improved LD-score regression intercept of 1.065 for the same data set. Excluding related individuals is a standard QC step if association testing does not account for relatedness. Nonetheless, since the results presented by the authors mainly rely on slightly inflated UKBB data, the analysis should be accepted as it is.

As explained later we were missing a sentence on on genomic control with LD score regression in the manuscript. We have added this to the methods pg 21: We used LD score regression¹³ to account for inflation in test statistics due to cryptic relatedness and stratification.

Methods that attempt to account for relatedness in the case control setting have other problems. In particular, rare variant associations are not properly handled and the association we do uses phenotype information from ungenotyped individuals by inferring genotype information from genotyped relatives (sometimes called familial imputation). This we can do as a very large fraction of the Icelandic population is genotyped and this method can increase the number of cases used in the study by about 75%. The familiarly imputed genotypes are less informative than imputed genotypes for genotyped individuals and this has to be appropriately accounted for in the association test (cite: <https://doi.org/10.1038/sdata.2015.11>). We have not yet found a method that consistently performs better than the currently applied one. The method we are currently using could potentially be improved to increase power, but it has the benefit of being robust and not produce false positive associations. We note that we are using the bolt method to account for relatedness for quantitative trait associations.

The use of BOLT is not described in the methods section or in the paper. For which analysis have the authors used BOLT LMM which is only listed in the “Code availability” section?

We use BOLT to perform quantitative trait associations that we used for the association with eosinophil count. We have added a description of this to the methods pg 21.

Association testing and meta-analysis of eosinophil counts

Linear mixed model implemented by BOLT-LMM⁶³ was used to test for association between sequence variants and eosinophil counts, assuming an additive genetic model. We assume that the quantitative measurements follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance–covariance matrix proportional to the kinship matrix⁶⁴. We used LD score regression¹³ to account for inflation in test statistics due to cryptic relatedness and stratification. To combine the deCODE and UK Biobank results, we used a fixed-effect inverse variance method based on effect estimates and standard errors⁶⁰. We used a likelihood-ratio test to compute all *P* values.

Further description of the eosinophil count was added pg. 19

We obtained eosinophil counts from three of the largest laboratories in Iceland (measurements performed between the years 1993 and 2015). The circulating eosinophil counts were standardized to a standard normal distribution using quantile-quantile standardization and then adjusted for sex, year of birth and age at measurement, as previously described^{9,10}. A total of 251,307 Icelandic individuals with eosinophil counts were included in the study. In the UKB we had eosinophil counts for a total of 396,822 individuals that were adjusted for sex, year of birth, age at measurement and the first 40 principal components, then inverse normal transformed.

If the familiarly imputed genotypes are used in LMM methods for quantitative traits to account for relatedness why not on qualitative traits? Why is relatedness a problem for quantitative but not for qualitative traits?

Genomic control, either the older median based adjustment or LD score regression, has proven to be a robust method for accounting for relatedness. However, LMM models are potentially more powerful. The “problem” is the same for both types of analysis and primarily leads to loss of power. Our methods for both do not create false positives.

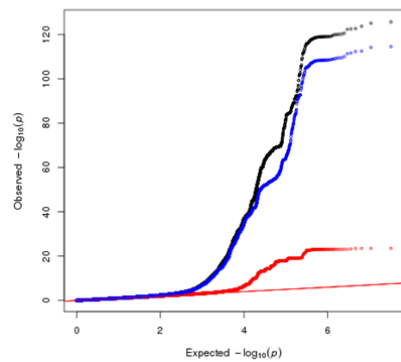
Given that the UKBB carries most of the associations on its own, the Icelandic cohort has a lot less weight, and the issue of genomic inflation in the Icelandic set could be left to a proper discussion of the issue. The authors should explain the advantages of their method including of not accounting for relatedness (above paragraph) in detail in the manuscript (discussion section). A discussion on advantages and disadvantages is required (including references).

We were missing a paragraph on genomic control with LD score regression in manuscript. We have added the following paragraph to the methods pg 20:

We used LD score regression¹³ to account for inflation in test statistics due to cryptic relatedness and stratification. We chose to include related individuals in the analysis as removing related individuals in the heavily genotyped Icelandic population would lead to the removal of most participants. Including related individuals while accounting for genomic control has proven to be a robust method^{10,58,59} that does not create false positives as observed in the QQ plots (supplementary figure 6).

In our previous response we explained how we calculate the effective sample size for each cohort taking into account the genomic adjustment factor (GC), based on LD-score regression for both cohorts ($UK_{GC}=1.092$) and $Iceland_{GC}=1.368$) “In total we had 16,247 asthma cases and 346,486 controls from Iceland and 52,942 cases and 355,713 controls from the UK Biobank. Estimating the effective number of cases as $2 \times N_a \times N_c / (N_a + N_c) / GC$, where N_a and N_c are the number of cases and controls, respectively, and GC is the genomic adjustment factor, estimated using LD-score regression (Bulik-Sullivan, B.K., Nat Gen, 2015), we can calculate the effective sample size for Iceland to be $N_{eff}=22,689$ and UK Biobank to be $N_{eff}=84,396$, $GC=1.368$ in Iceland and $GC=1.092$ in UK Biobank”. This was already added in the Material and Methods section page 15. To further, bring this into the result session we have added a sentence on this on page 3: “ The estimated genomic inflation in the two studies (estimated using LD-score regression)⁷ is 1.368 for the Icelandic dataset and 1.092 for the UK dataset, reflecting the relatedness of individuals in Iceland and is also the most

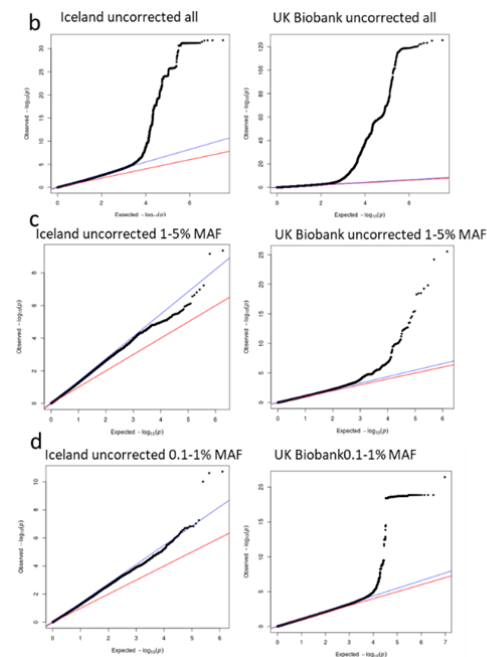
likely explanation of genomic inflation in UK Biobank where 30% of the participants have a relative (third degree or closer) in the UKB dataset⁸.



Supplementary Figure 6. Quantile-quantile plot (QQ-plot) showing chi-square statistics corrected for genomic control and showing variants found in UK and UK cohorts for the GWAS of Asthma in Iceland (red) the UK (blue) and meta-analysis of the two cohorts (black). The red diagonal line represents expected distribution assuming no inflation of the chi-square statistics.

Again, the authors should use uncorrected p-values for Supplementary Figure 6.

Agreed, the following panels have been added to supplementary figure 6.



Comment #2

In general, the authors should be more careful in the interpretation of the data. Acknowledging the strengths of their study, they should also recognize potential weaknesses. Out of 88 loci, 29 loci show nominally significant heterogeneity between both studies which is well beyond the expected number. Likewise, only 11 loci reach the Bonferroni corrected p-value for association with asthma in the replication study. It is true that the replication study has less power, but this is attributed to the study design. A combined analysis of both data sets might have been the better choice. All these points need to be discussed in the manuscript, not in materials and methods. Non-replicating loci with on opposite effect direction in the replication study could be assessed as potentially new loci. Finally, loci which are strongly associated with overlapping allergic phenotypes such as eczema could have been identified due to this concomitant allergic disease since allergic co-morbidities were not excluded from the study. This should also be included in the discussion.

Response:

As clearly stated in the abstract, main text and material and methods, this is a meta-analysis of the two sample sets and therefore, we have indeed performed a combined analysis of both datasets as the reviewer suggested.

Regarding the observed heterogeneity between the effect estimates in the two studies, it is worth noting that these are retrospective studies and the diagnostic criteria, ascertainment and age distribution differs between them. Furthermore, it is likely that environmental exposure differs between the two populations. For these reasons we do not expect the effect estimates to be exactly the same in the two populations and the question if we can detect heterogeneity becomes one of power. Consistent with this scenario, all the variants that show nominally significant heterogeneity are previously published signals, i.e. the signals that have the strongest association. And there is a strong correlation between the strength of the association signals and the significance of the heterogeneity. Thus the observed heterogeneity in the effect estimates probably shows that these two studies are large enough to have good power to detect the real difference in the phenotype definition and/or ascertainment between the studies. This is likely to be true for most meta-analysis of GWAS studies that are being done, where the GWAS studies usually differ in terms of ascertainment and/or phenotype definition; however in most cases the individual studies are not as large as here and the power to detect heterogeneity is much less.

This has been addressed by the following sentence on page 3: „This observed heterogeneity in the effect estimates indicates that these two studies are powered to detect the real difference in the phenotype definition and/or ascertainment between the studies that are coherent with retrospective studies as reported here“.

As pointed out in the result section, 85 out of the 88 variants (which is much more than would be expected randomly, $P < 1 \times 10^{-16}$) have effect size in the same direction in the two sample sets. The three variants that seem to have opposite effect direction are completely flat in Iceland as can be seen in supplementary table 1: rs1684466 (OR=1.00652, $P=6.86E-01$ in Iceland and OR=0.94365, $P=4.79E-15$ in UKB), rs17454584 (OR=0.999, $P=9.56E-01$ in Iceland and OR=1.06375, $P=8.15E-14$ in UKB), rs41284471 (OR=0.9905, $P=6.34E-01$ in Iceland and OR=1.05369, $P=9.98E-10$ in UKB).

The reviewer is correct that we did not exclude allergic co-morbidities from our asthma association studies. However, we note that only 980 individuals of our 16,247 (6%) asthma cases in Iceland have the atopic dermatitis (AD) diagnosis and only approx 100 individuals had the ICD10 code for AD (L20) in the UKB at the time of analysis. However, as this is a retrospective study, not designed to address

co-morbidities of asthma, we cannot exclude that co-morbidities are underreported/diagnosed and that those might affect associations of certain sequence variants with asthma.

This has been addressed by the following paragraph page 9: “Since this is a retrospective study, not designed to study co-morbidities of asthma and allergic diseases, we cannot exclude that some of the overlapping association of certain sequence variants with e.g. asthma and AD might be due to overlap of the two phenotypes. However, only 980 out 16,247 asthma cases in Iceland had the AD dermatitis diagnosis (6%) while AD was hardly reported as a single disease in UKB, making it impossible to accurately define asthmatics without AD.”

Minor points

Comment #3

Table 1 still contains known allergy loci ([ADAD1],IL2-[]-IL21) reported by Ferreira et al (Nat Genet. 2017 Dec;49(12):1752-1757).

Response: The reviewer is right and we apologize for this oversight. We have now moved the ADAD1 locus from Table 1 to supplementary Table 5 where we have summarized other loci reported by Ferreira et al, 2017. We have further, updated the number of novel variants accordingly to 19 at 16 loci and 22 variants at 16 loci that have previously been reported for a combined allergic phenotype

Comment #4

In Supplementary Table 2, for SNPs in high LD with previously reported SNPs or for identical SNPs, the authors should clearly state whether the same risk allele replicates or not (additional column). If no data are available in the GWAS catalog, please also check the original manuscripts (e.g. Pickrell et al report alleles and effect directions in the supplementary data).

Response: Unfortunately, it is not straight forward to understand what the risk allele is from the supplements in Pickrell et al. In the supplementary table for asthma they report the alleles as C/T, G/T etc. which one could believe would be the reference/non-reference allele. That would be in line with their description in the supplementary methods that „Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19.“ However, when we look the SNPs up in ncbi database using GRCh37 (hg19) there is inconsistency for many of the variants with the table reported in Pickrell et al. with regards to which allele is the reference allele. When we blindly follow the instructions that they report the effect of the non-reference allele we do not replicate the allele for any of the variants that are inconsistent with what is reported in the ncbi database. Whereas we replicate the effect for all variants that are in line with the ncbi database e.g. rs1723018 that according to ncbi is an A>G variant and is written as A/G in Pickrell have correlated risk alleles in our data and Pickrell. If we believe that the authors have intended their alleles to be written as reference/non-reference allele and they therefore are always reporting the effect size on the latter allele (whether or not that is the non-reference allele) then we replicate their findings 100% (see Table 1 below). However, it is impossible to tell from the Pickrell paper as it is and the same applies for Wan Yi. 2012 (PMID: 22561531), Torgerson (PMID:21804549) and Almoguera (PMID:27611488) which is probably the reason why the risk allele gets a question

mark in the GWAS catalog. Therefore, we have to stick to what is reported in the GWAS catalog and leave the risk allele with a question mark.

We have added a column for the correlated alleles between SNPs and whether or not we replicate the risk allele for other publications in Supplementary Table 2.

Table 1. Previously reported SNPs from Pickrell_JK, 2016 (PMID:27182965) and correlation of risk alleles with our study.

SNP reported in Pickrell et al	alleles as reported in Pickrell et al	Ref>alt alleles as reported in GRCh37*	OR reported in Pickrell	Risk allele in Pickrell**	Risk allele in our data	Correlated with our allele	Correlated alleles***
rs6683383	A/T	T>A	1.062	A	T	No	T=T; A=G
rs1723018	A/G	A>G	0.945	A	G	Yes	A=G; G=A
rs13412757	A/G	G>A	1.062	A	G	No	G=G; A=A
rs34290285	A/G	G>A	1.107	A	G	No	same variant
rs5743618	A/C	C>A	1.082	A	C	No	same variant
rs2244012	A/G	A>G	1.102	G	T	Yes	A=C; G=T
rs10957978	G/T	G>T	0.934	G	A	Yes	G=A; T=G
rs144829310	G/T	G>T	1.165	T	A	Yes	G=G; T=A
rs12413578	C/T	C>T	0.891	C	C	Yes	same variant
rs7936323	A/G	G>A	0.922	G	T	No	G=G; A=T
rs3784099	A/G	G>A	0.942	G	A	No	G=T; A=A
rs30519068	A/G	G>A	1.10	A	C	No	G=C; A=T
rs56375023	A/G	G>A	0.897	G	A	No	G=C; A=A
rs7203459	C/T	T>C	1.092	C	T	No	T=T; C=A
rs11655198	C/T	C>T	0.847	C	G	Yes	C=G; T=T

*non-reference allele according to GRCh37

** Risk alleles in Pickrell defined from their description in supplementary methodology that "Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19."

*** Correlated alleles according to https://idlink.ncl.nih.gov/?var=rs1723018&pop=CEU&r2_d=2&tab=idproxy

Supplementary Table from Pickrell_JK, 2016 (PMID:27182965).

Index SNPs for Strongest Associations

cytoband	assay.name	scaffold	position	alleles	src	pvalue	OR	95% CI	gene.context
17q12	rs11655198	chr17	38026169	C/T	1	1.0×10^{-63}	0.847	[0.831,0.863]	[ZBPB2]
6p21.32	rs3104367	chr6	32603487	C/T	1	1.0×10^{-40}	0.872	[0.855,0.890]	HLA-DRB5---[]-HLA-DQA1
9p24.1	rs144829310	chr9	6208030	G/T	1	1.3×10^{-31}	1.165	[1.136,1.195]	RANBP6---[]-IL33
5q22.1	rs1837253	chr5	110401872	C/T	1	3.3×10^{-31}	0.876	[0.856,0.896]	SLC25A46---[]-TSLP
2q12.1	rs202011557	chr2	102913642	D/I	1	5.1×10^{-31}	0.842	[0.818,0.868]	IL1RL2---[]-IL1RL1
15q22.33	rs56375023	chr15	67448363	A/G	1	2.4×10^{-21}	0.897	[0.877,0.917]	[SMAD3]
11q13.5	rs7936323	chr11	76293758	A/G	1	1.4×10^{-16}	0.922	[0.905,0.940]	C11orf30---[]-LRRRC32
6p21.33	rs2428494	chr6	31322197	A/T	1	1.4×10^{-16}	0.920	[0.902,0.939]	HLA-C--[]-MICA
5q31.1	rs2244012	chr5	131901225	A/G	1	2.1×10^{-16}	1.102	[1.077,1.128]	[RAD50]
2q37.3	rs34290285	chr2	242698640	A/G	1	1.8×10^{-15}	1.107	[1.079,1.135]	[D2HGDH]
16p13.13	rs7203459	chr16	11230703	C/T	1	3.5×10^{-15}	1.092	[1.068,1.117]	[CLEC16A]
1q23.3	rs4233366	chr1	161159147	C/T	1	4.8×10^{-15}	1.090	[1.067,1.114]	B4GALT3---[]ADAMTS4
10p14	rs12413578	chr10	9049253	C/T	1	8.1×10^{-12}	0.891	[0.862,0.921]	GATA3---[]
8q21.13	rs10957978	chr8	81285139	G/T	1	1.1×10^{-11}	0.934	[0.915,0.952]	TPD52---[]---ZBTB10
15q22.2	rs10519068	chr15	61068704	A/G	1	3.8×10^{-11}	1.100	[1.069,1.132]	RORA---[]
4p14	rs5743618	chr4	38798648	A/C	1	3.9×10^{-11}	1.082	[1.057,1.107]	[TLR1]
6q15	rs58521088	chr6	90985198	A/T	1	7.1×10^{-11}	0.934	[0.915,0.954]	[BACH2]
12q13.3	rs3001426	chr12	57509055	C/T	1	1.4×10^{-10}	0.938	[0.919,0.956]	STAT6-[]---LRP1
3q28	rs73196739	chr3	188402471	C/T	1	6.5×10^{-9}	0.922	[0.897,0.948]	[LPP]
1q32.1	rs6683383	chr1	203100504	A/T	1	1.1×10^{-8}	1.062	[1.040,1.084]	[ADORA1]
2p25.1	rs13412757	chr2	8458080	A/G	1	1.3×10^{-8}	1.062	[1.040,1.084]	[]---ID2
1q24.2	rs1723018	chr1	167433420	A/G	1	1.4×10^{-8}	0.945	[0.926,0.963]	[CD247]
14q24.1	rs3784099	chr14	68749927	A/G	1	1.6×10^{-8}	0.942	[0.922,0.961]	[RAD51B]
7q22.3	rs6959584	chr7	105676505	C/T	1	2.0×10^{-8}	1.086	[1.055,1.117]	[CDHR3]
5q31.3	rs200634877	chr5	141529761	D/I	1	2.5×10^{-8}	0.940	[0.919,0.961]	[NDFIP1]
1q25.1	rs6691738	chr1	173152036	G/T	1	2.9×10^{-8}	0.943	[0.923,0.963]	TNFSF18---[]TNFSF4
1p36.22	rs662064	chr1	10557251	C/T	1	3.2×10^{-8}	0.942	[0.922,0.962]	[PEX14]

Comment #5

In Supplementary Table 7, the authors should add the p-values and OR of the coding SNPs.

Response: This has been added with the addition of EA/OA for the coding SNPs and correlation of alleles between the strongest asthma associated variant and the coding variant.

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily responded to all my comments.

Reviewer #3 (Remarks to the Author):

The experimental part improved greatly after revision. The data is sound.

I have just one small comment. It would be better for readers if the statistical evaluation is included in the figure 2 legend. Currently is just mentioned in the text

Response: This has been added.

Additional comment: A UKBB study on asthma has recently been published (Hum Mol Genet. 2019 Jul 30. pii: ddz175. doi: 10.1093/hmg/ddz175) which should be referenced, and novelty of loci adjusted.

Since this paper came out so late in the review process, we have added a sentence referring to this paper and the one included in our 2nd submission that have reported novel asthma loci while our manuscript was in review and label them in Table 1, accordingly, as we already did for the paper of Shrine et al, (2nd submission), which the reviewers accepted.

Pg. 5: We note that during the final review of this manuscript two independent reports published association of asthma with 9 of those 19 variants as indicated in Table 1^{15,16}.

Title: Supplementary Data 1:

Description: GWS signals associating with asthma in Iceland-UK meta analyses.

Title: Supplementary Data 2:

Description: Known variants ($LD\ r^2 \geq 0.2$) at previously reported loci

Title: Supplementary Data 3:

Description: Unreported variants ($LD\ r^2 \geq 0.2$) at previously reported loci

Title: Supplementary Data 4:

Description: Replication of the 36 previously published asthma loci of European-ancestry in Icelandic UK meta-analysis.

Title: Supplementary Data 5:

Description: Novel asthma variants at previously reported loci reported in Ferreira et al. (PMID:29083406) for a combined phenotype of asthma, hay fever and eczema.

Title: Supplementary Data 6:

Description: Whole blood eQTL variants correlated ($r^2 \geq 0.8$) with asthma lead variants

Title: Supplementary Data 7:

Description: Coding variants strongly correlated ($r^2 > 0.8$) with novel asthma associated sequence variants

Title: Supplementary Data 8:

Description: eQTL variants correlated with novel asthma lead variants found in whole blood, esophagus mucosa, Lung, EBV-transformed lymphocytes and Transformed fibroblasts in the Gtex database.

Title: Supplementary Data 9:

Description: Functional annotations of previously unreported asthma variants

Title: Supplementary Data 10:

Description: Classical HLA alleles correlating with asthma in the Icelandic cohort

Title: Supplementary Data 11:

Description: Effect of genome wide significant Asthma variants on eosinophil count.

Title: Supplementary Data 12:

Description: Effect of the 88 GWAS significant asthma associating variants on asthma subphenotypes and other allergic diseases

Title: Supplementary Data 13:

Description: Genes prioritized with DEPICT pathway analysis

Title: Supplementary Data 14:

Description: Gene sets found enriched in the 88 GWS asthma risk variants using DEPICT pathway analysis