A whole genome sequencing study of moderate to severe asthma identifies a lung function locus associated with asthma risk

Diana Chang1, Julie Hunkapiller1, Tushar Bhangale1, Jens Reeder3, Kiran Mukhyala1, Jennifer Tom3, Amy Cowgill1, Jan Vogel1, William F. Forrest1, Zia Khan1, Amy Stockwell1, Mark I. McCarthy1, Tracy L. Staton1, Julie Olsson1, Cecile T. J. Holweg1, Dorothy S. Cheung1, Hubert Chen1, Matthew J. Brauer1,2, Robert R. Graham1,2, Timothy Behrens1,2, Mark S. Wilson1, Joseph R. Arron1, David F. Choy1,2, & Brian L. Yaspan1

Genome-wide association studies (GWAS) have identified many common variant loci associated with asthma susceptibility, but few studies investigate the genetics underlying moderate-to-severe asthma risk. Here, we present a whole-genome sequencing study comparing 3181 moderate-to-severe asthma patients to 3590 non-asthma controls. We demonstrate that asthma risk is genetically correlated with lung function measures and that this component of asthma risk is orthogonal to the eosinophil genetics that also contribute to disease susceptibility. We find that polygenic scores for reduced lung function are associated with younger asthma age of onset. Genome-wide, seven previously reported common asthma variant loci and one previously reported lung function locus, near THSD4, reach significance. We replicate association of the lung function locus in a recently published GWAS of moderate-to-severe asthma patients. We additionally replicate the association of a previously reported rare (minor allele frequency < 1%) coding variant in IL33 and show significant enrichment of rare variant burden in genes from common variant allergic disease loci. Our findings highlight the contribution of lung function genetics to moderate-to-severe asthma risk, and provide initial rare variant support for associations with moderate-to-severe asthma risk at several candidate genes from common variant loci.

Asthma is a heterogeneous complex disease characterized by reversible airway obstruction, airway hyperresponsiveness, and variable inflammation. Genome-wide association studies (GWAS) of asthma have identified more than thirty loci associated with asthma susceptibility1-5. Many of these loci point to inflammation mediated by type 2 immunity (e.g. IL13) and are enriched in regions with histone marks indicating enhancers in immune cells2. Though GWAS have successfully uncovered numerous asthma loci, gaps remain in our understanding of the genetics underlying asthma risk.

First, although many studies have been carried out on asthma risk1, only a small fraction specifically focused on severe or uncontrolled asthma patients6-12. While these patients only constitute 5-10% of all asthma patients13,14, they represent more than 50% of healthcare usage (by asthma patients) and have a large unmet medical need13. For these reasons, we focused our study on this asthma subgroup. A recent study (Shrine et al.) carried out a GWAS of over 10,000 patients (including > 5000 cases from the UK Biobank) with moderate-to-severe asthma. While they found that the majority of moderate-to-severe risk loci overlapped with asthma risk loci, the presence of loci associated with only moderate-to-severe asthma suggest some distinct mechanisms underlie mild versus severe asthma. In addition, though it is clinically appreciated that the two key traits underlying asthma and asthma severity—lung function and eosinophilic inflammation—likely represent distinct pathways, how the genetics contributing to these traits overlap and influence asthma is yet to be fully explored.

Second, though the majority of asthma GWAS have focused on common variants, several rare variants contributing to asthma risk have recently been identified. In particular, previous reports in mild to severe asthma

1Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA. 2Present address: Maze Therapeutics, 131 Oyster Point Blvd STE 200, South San Francisco, CA 94080, USA. 3email: choy.david@gene.com; yaspan.brian@gene.com
have identified rare coding variants in IL33\(^{16}\) and GSDMB\(^{17}\), though their contribution to moderate-to-severe asthma risk have not yet been tested. We therefore carried out a whole-genome sequencing study (WGS) on moderate-to-severe asthma patients. We aim to identify common and rare genetic contribution to disease risk, and to compare and contrast the role of eosinophil and lung function genetics in moderate-to-severe asthma.

**Results**

**Common genetics of asthma risk in moderate-to-severe patients.** In this study, we first analyzed common variants obtained from whole-genome sequences of 3181 moderate-to-severe asthma cases and 3590 non-asthma controls of majority European ancestry (fraction of European ancestry > 0.85 as estimated by Admixture\(^{18}\)). Asthma cases were derived from ten studies, eight of which were clinical trials (see "Methods"). Healthy control samples with comparable sequencing data were unavailable as sequenced samples were mainly from clinical trials. We thus obtained disease controls from participants in non-asthma clinical trials that were sequenced and processed with the same informative pipeline to minimize batch effects that may be introduced via differing sequencing technologies and bioinformatics processing pipelines (see "Methods"). These non-asthma disease controls comprised 1140 and 2450 patients from clinical trials of age-related macular degeneration (AMD) or rheumatoid arthritis (RA), respectively. Though similar pathways may contribute to both asthma and RA (e.g. IL6\(^{19}\) and Th1 cells\(^{20}\)), there is minimal genetic correlation between the two traits (\(r_g = 0.08, P = 0.07\)). Additional RA and AMD are not genetically correlated (\(r_g = 0.035, P > 0.3\)) and share few genetic loci (\(n = 2\)) suggesting limited confounding was introduced by the use of RA and AMD as controls in this study (see Supplementary Note S1). Furthermore, to remove potential association signals originating from our controls, we applied a differential-effects test to remove variants with effect sizes that were significantly different (\(P < 0.01\)) when comparing RA cases to AMD controls versus comparing the same asthma cases to RA controls (see "Methods"). This test successfully filtered known AMD and RA loci (e.g. the associations of the CFH and ARMS2/HTRA1 loci with AMD) (Supplemental Fig. S1).

Variance were removed for failing an allelic depth balance test, Hardy–Weinberg equilibrium, and/or having high missingness rates (see "Methods"). After variant filtering, there were 7,165,996 common variants filtered known AMD and RA loci (e.g. the associations of the CFH and ARMS2/HTRA1 loci with AMD) (Supplemental Fig. S1). We used a previously reported, independent study, by Shrine et al. 2019 moderate-to-severe asthma risk stage 1 GWAS\(^{21}(5135 cases and 25,675 controls) to replicate variants discovered in our study. The use of this independent study to replicate our findings further generalize our findings beyond cases studied in a clinical trial setting.

We first estimated the narrow-sense heritability (\(h^2\)) of moderate-to-severe asthma risk in our study using the LD-score regression framework\(^{22}\). We assumed a prevalence of 0.0084 for moderate-severe asthma by using a prevalence of 0.084 for all asthma and assuming moderate-severe asthma patients account for 10\% of all asthma patients\(^{13}\). Using the prevalence above, we estimated that the \(h^2\) for our moderate-to-severe asthma risk study is 0.29 (s.e. = 0.045) (estimated \(h^2\) assuming a range of other prevalences are available in Supplemental Table S1). We further estimated the genetic correlation between moderate-to-severe asthma risk as defined in our study and in the previously published Shrine et al. study\(^{20}\) to be 0.54 (s.e. = 0.096). This is lower than the estimated genetic correlation between moderate-to-severe asthma risk as defined in our study and general asthma risk as defined in Demenais et al.\(^2\) (\(r_g = 0.71, s.e. = 0.03\)) though the 95\% confidence intervals of these estimates overlap.

Eight regions reached genome-wide significance (\(P < 5 \times 10^{-8}\)), seven of which were within 1 Mb of regions previously reported to be associated with asthma (Fig. 1, Table 1). An additional 16 previously reported asthma associations\(^2\) (Supplemental Table S2) and 34 previously reported allergic disease associations\(^2\) (Supplemental Table S3) showed at least nominal evidence for significance (\(P < 0.05\)) in this study. The previously reported moderate-to-severe asthma risk locus near MUC5AC showed modest levels of significance with the same direction of effect (rs11603634, \(P = 2.99 \times 10^{-3}\), OR\(_G\) = 1.11).

The genome-wide significant association that did not map to any previously reported asthma loci maps to chromosome 15 near the gene THSD4 (rs11631778, OR\(_{G,allelic} = 1.23, P = 3.54 \times 10^{-4}, \text{MAF}_{	ext{cases}} = 0.35\) (Table 1)). The MAF of this variant is higher in our cases as compared to the control samples in our study (\(\text{MAF} = 0.31\)) as well as in gnomAD (v2.1.1) (\(\text{MAF}_{\text{non-Finnish-EUR}} = 0.32\) and UK Biobank participants over the age of 50 with no documented respiratory disorders (\(\text{MAF} = 0.33\)).

We were able to replicate association of this THSD4 variant with moderate-to-severe asthma risk in the Shrine et al. study\(^{20}\) (\(P = 0.0079,\ OR\(_{G,allelic} = 1.06\)), but were unable to replicate this association (using the proxy SNP in high LD with our lead variant—rs11853359, \(r^2 = 0.93\)) in studies that did not enrich for moderate-to-severe asthma patients (European subset of the Demenais et al. study\(^2\), and GWAS of the asthma Phecode in the UK Biobank\(^{22}(P > 0.45\)).

Conditioning on the index variant in the region (rs11631778) did not reveal any independent associations passing a genome-wide significant threshold of \(5 \times 10^{-8}\). Applying FINEMAP\(^{24}\) to the region further supported a single causal signal of association with asthma at this locus and found 10 variants in the 95\% credible set (Supplemental Table S4), with the lead SNP having a posterior inclusion probability of 0.23. Variants in the 95\% credible set overlapped with enhancers and histone marks in lung tissue and lung-related cell-types, and were also associated with expression of THSD4 in lung samples from GTex (Supplemental Table S5). To confirm that the lung eQTL and asthma risk association point to the same underlying causal variant, we carried out colocalization analysis via the coloc package in R (see "Methods"). We found there was a high probability of colocalization between the lung eQTL and the asthma risk association (\(\text{probability}_{\text{colocalization}} = 0.99\)) (Supplemental Fig. S4).

Multiple traits can contribute to asthma pathology with lung function (which can be viewed as a proxy for structural changes in the airway leading to variable airflow limitation) and eosinophilic inflammation being two of the major traits\(^{25}\). It is now appreciated clinically that these two traits may not be causally linked\(^{27}\) and we set
out to test whether genetic analyses further support this distinction. The novel moderate-to-severe asthma locus uncovered above is likely contributing to asthma via lung function as the lead SNP, rs11631778, is in high linkage disequilibrium (LD) ($r^2 = 0.95$) with a SNP (rs1441358) associated with increased COPD risk$^{28}$, and reduced lung function$^{29}$, but is not associated with eosinophil blood count ($P > 0.05$)$^{30}$. We next sought to investigate whether this distinction between the two traits extended beyond the THSD4 locus.

We estimated the genetic correlation between lung function measures and blood eosinophil counts, and asthma risk. We confirmed previous reports$^{31}$ that blood eosinophil cell count and asthma risk are genetically correlated ($r_g = 0.30$, Supplemental Table S6) by applying LD-score regression$^{21}$ to a GWAS of blood eosinophil cell counts in the INTERVAL study$^{30}$ and the Demenais et al. asthma risk GWAS$^{2}$. We found this genetic correlation was also present between moderate-to-severe asthma risk and blood eosinophil cell count ($r_g = 0.28$). We found a similar genetic correlation when using moderate-to-severe asthma risk as defined in Shrine et al.$^{8}$ (Supplemental Table S6). Next, we used a recent meta-analysis of lung function traits carried out on the UK Biobank and SpiroMeta cohorts$^{29}$ of four lung function measures: FEV₁ (forced expiratory volume in 1 s), FVC (forced vital capacity), PEF (peak expiratory flow) and FEV₁/FVC$^{21}$ to calculate the genetic correlation between these traits and asthma risk. We found inverse genetic correlation between overall asthma risk and lung function (e.g. higher asthma risk was genetically correlated with lower lung function) ($r_g \leq -0.21$ for all lung function traits, Supplemental Table S6). As with overall asthma risk, we found that moderate-to-severe asthma risk was

**Table 1.** Sentinel variants in eight regions significantly associated with moderate-to-severe asthma risk in this GWAS of 3181 cases and 3590 controls (OR = odds ratio for coded allele, SE = standard error, $P_{DE} = p$-value for the differential effects test, RAF = risk allele frequency for AMD and RA controls, respectively).

| SNP         | CHR | BP (GRCh38) | Nearest gene | Risk/non-risk allele | OR   | SE   | P-value          | $P_{DE}$ | RAF_case | RAF_control |
|-------------|-----|-------------|--------------|----------------------|------|------|-----------------|---------|----------|-------------|
| rs139210940 | 2   | 102,265,885 | IL1RL2       | AT/A                 | 1.339| 0.051| $8.08 \times 10^{-9}$ | 0.333   | 0.879    | 0.847, 0.843 |
| rs10455025  | 5   | 111,069,301 | TSLP         | G/A                  | 1.304| 0.037| $4.36 \times 10^{-11}$ | 0.856   | 0.387    | 0.34, 0.326 |
| rs17205170  | 6   | 32,634,706  | HLA-DQA1     | G/T                  | 1.449| 0.046| $7.92 \times 10^{-14}$ | 0.168   | 0.84     | 0.797, 0.779 |
| rs2875584   | 6   | 90,240,909  | BACH2        | C/T                  | 1.238| 0.038| $1.57 \times 10^{-8}$ | 0.351   | 0.706    | 0.666, 0.653 |
| rs7130588   | 11  | 76,559,639  | C11orf30     | G/A                  | 1.242| 0.036| $2.46 \times 10^{-9}$ | 0.04    | 0.386    | 0.358, 0.332 |
| rs2104047   | 14  | 68,287,700  | RAD51B       | T/C                  | 1.245| 0.039| $1.28 \times 10^{-4}$ | 0.134   | 0.312    | 0.262, 0.274 |
| rs11631778  | 15  | 71,314,041  | THSD4        | G/A                  | 1.225| 0.037| $3.54 \times 10^{-8}$ | 0.436   | 0.354    | 0.305, 0.311 |
| rs7216588   | 17  | 39,913,818  | GSDMB        | T/C                  | 1.263| 0.035| $1.91 \times 10^{-11}$ | 0.442   | 0.545    | 0.479, 0.494 |

**Figure 1.** Manhattan plot of common variants (MAF > 1%) associated with asthma risk in this GWAS of 3,181 moderate-to-severe asthma patients and 3590 non-asthma controls. The $-\log_{10}$ two-sided $p$-values are displayed. Variants with $P < 0.01$ for the differential effects test (See “Methods”) are filtered out. See Supplemental Fig. S2 for a corresponding plot which includes all variants. The black line indicates the genome-wide significance threshold of $5 \times 10^{-8}$. Previously reported loci are labeled in black, the novel locus is labeled in blue.
also inversely correlated with all lung function measures ($t_r < -0.16$) (Supplemental Table S6). This inverse genetic correlation was replicated in the moderate-to-severe asthma risk published by Shrine et al. (Supplemental Table S6).

We next explored whether lung function genetics overlap with the genetics of blood eosinophil counts or whether they represent independent pathways that may contribute to asthma pathology. While asthma risk shows evidence of shared genetics with both lung function measures and eosinophils, we found that the sharing between eosinophils and lung function measures was low ($-0.086 < t_r < -0.039$), suggesting they represent orthogonal axes contributing to overall asthma risk pathology (Supplemental Table S6). Given the distinct variants contributing to these two traits, we next investigated whether these variants influenced the trajectory of asthma—specifically we asked if and how these two axes impacted age of onset in asthma patients. While several risk factors (e.g. allergic sensitization) contribute to childhood onset asthma, proposed risk factors for adult onset asthma include upper respiratory tract infections, exposure to pollutants, hormonal factors, and obesity. Impaired lung function is also observed to be lower in adult onset asthma as compared to childhood onset asthma, though other studies suggest limited impact of lung function on age of onset. Therefore, we hypothesized that genetic variants associated with lung function would likely be associated with asthma age of onset.

To test this hypothesis, we generated polygenic scores (PS) in a subset of our patients for which age of onset data were available ($N = 1,456$). To briefly summarize polygenic scores, a polygenic score for each individual is calculated by how many phenotype-associated alleles an individual carried weighted by the allele's effect on that phenotype. A large polygenic score can be interpreted as someone with a higher likelihood, based upon their genetics, to have that phenotype (e.g. asthma) compared to someone with a low polygenic score. We scored individuals using publicly available GWAS summary statistics for allergic disease, asthma risk, blood eosinophil count, and lung function measures (see “Methods”). We created a binary phenotype, binning patients into whether they had childhood onset asthma (age onset ≤ 12, $N = 665$) or adult onset asthma (age ≥ 25, $N = 791$) and regressed this binary phenotype on each PS. As expected, we found that increased PSs for allergic disease and asthma were associated ($P ≤ 1.09 \times 10^{-9}$) with childhood onset asthma ($OR_{allergic\_disease} = 0.34$, $OR_{asthma} = 0.38$) and regressed this binary phenotype on each PS. As expected, we found that increased PSs for allergic disease and asthma were associated ($P ≤ 9.06 \times 10^{-4}$) with age of onset, where a higher lung function PS was associated with lower age of onset ($P = 0.004$) (Fig. 2, Supplementary Table S7).

**Rare coding variants associated with moderate-to-severe asthma risk.** After variant level QC filters as described above, we were left with 3,600,569 rare (MAF < 1%) exonic variants. Previous studies reported the minor allele of a rare loss of function variant in IL33 (rs146597587-C) was associated with reduced asthma risk and the minor allele of a rare missense variant in GSDMB (rs124500991-C) was associated with increased risk of asthma. We were able to replicate the association of the IL33 variant in our study ($OR_{C-allele} = 0.37$, $P$-value $= 0.025$, MAF $= 1.93 \times 10^{-4}$), but not the GSDMB variant ($OR_{C-allele} = 1.32$, $P = 0.79$, MAF $= 2.32 \times 10^{-4}$). Genome-wide, no individual rare coding variants were significant after correcting for the number of rare coding variants tested (strict Bonferroni cut-off of correcting for 3,600,569 rare coding variants, $P < 1.39 \times 10^{-8}$).

To improve power (especially for genes with extensive allelic heterogeneity), we further aggregated rare variants into gene coding regions. To do so we employed the rare variant burden test as this was compatible with the differential-effects test used in this study to flag results that may be significant due to the use of disease-controls. Burden tests are well powered for scenarios where there is a large fraction of causal variants with a similar direction of effect. On the other hand, we will have reduced power to detect genes with a small fraction of causal variants with opposing directions of effect. In the burden test, we coded individuals as 0 or 1 dependent on whether the individual carried a rare allele in that gene. We carried out two sets of burden tests. In the first, we only considered variants with a predicted high impact on protein function (loss-of-function test). In the second, we considered variants in the first test as well variants with a moderate predicted impact that had a PolyPhen score (probability of the variant being deleterious) > 0.5 (loss-of-function and moderate impact variant test) (see “Methods”). For each burden test, only genes with ≥ 2 variants and ≥ 5 carriers of rare variants satisfying the criteria above were considered for testing. 316 genes passed our filtering criteria for our first burden test of loss-of-function variants, and 15,836 genes passed our criteria for having loss of function variants and/or moderate impact variants with PolyPhen > 0.5.

Overall, no genes were significant after correcting for the number of genes tested exome-wide (Supplemental Figs. S5, S6). We next sought to test whether collectively they were enriched for association of candidate genes from common-variant allergic diseases, moderate-to-severe asthma and/or lung function GWAS. We investigated three gene-sets corresponding to 132 allergic disease candidate genes, 17 moderate-to-severe asthma candidate genes, and 68 lung function candidate genes for common variant risk loci previously identified. A majority of these candidate genes reported by the original studies were supported with either coding level or eQTL support linking genes to the associated loci (further details can be found in the “Methods” as well as in the original publications). We found significant enrichment for association of genes within the allergic disease gene-set for loss-of-function and moderate impact rare variant burden ($P = 0.004$) (see “Methods”). We also found nominal significance for enrichment of the lung function gene-set ($P = 0.031$) though this did not pass multiple testing correction for these gene-sets and two rare variant burden masks (Supplemental Table S8).
Within the allergic diseases and moderate-to-severe asthma gene-sets, FLG was nominally significant in the loss-of-function burden test (P = 0.024, OR = 1.473), and eight additional genes passed nominal significance in the high or moderate (PolyPhen > 0.5) impact burden test (RERE, IQCB1, PPP2R3C, PITPNM2, DYNAP, TSLP, EAF2, RASA2) (Supplemental Tables S9 and S10). Of the lung function candidate genes, rare variant burden in four genes (MAPT, CFDP1, EML3, and LTBP4) was nominally associated with asthma risk in our study (Supplemental Tables S9 and S10).

To investigate whether any of these nominally associated genes in our study had rare variant support from an independent cohort we turned to the partially released whole-exome sequencing (WES) data for 200K UK Biobank participants. Of the 6312 participants with moderate-to-severe asthma used in the above common variant analysis, WES data were available for 3418 samples. We compared these to 111,261 control participants with WES data available and no respiratory phenotypes (see “Methods”) and carried out rare variant burden tests as described above for the 13 candidate genes with nominal significance for association with asthma in our study. We used METAL to carry out a meta-analysis between the UK Biobank data and our study for these genes. Of the 13 candidate genes, only the loss-of-function association for FLG became more significant in the meta-analysis (P = 0.0013, OR = 1.30) (Supplemental Tables S9 and S10).

**Discussion**

In this study we present a sequencing cohort of patients with moderate to severe asthma. Overall, eight regions reached genome-wide significance in our study, seven of which overlap previously reported asthma risk variants. This is consistent with findings from a previous GWAS of moderate-to-severe asthma that reported significant overlap between variants associated with moderate-to-severe asthma and asthma risk. One possible explanation for this is that the genetic contribution to asthma severity is modest, suggesting a larger role for environmental factors. Additionally, the identification of new genes in our study suggests the potential for further genetic insights into the complex pathophysiology of asthma.
factors. The remaining genome-wide significant locus in our study mapped to a region containing THSD4 and only replicated in a study of moderate-to-severe asthma risk. Colocalization analysis supports THSD4 as the candidate gene for this association signal. THSD4, thrombospondin type 1 domain containing 4, is an extracellular matrix protein that is involved in microfibril formation, and may contribute to the structural integrity of the lungs. This locus has previously been associated with lung function and COPD risk and adds to the growing genetic support for the role of lung function determinants in risk of moderate-to-severe asthma.

Clinically, lung function (as a proxy for structural changes to the airway) and eosinophilic inflammation are key components of asthma pathology. Asthma risk loci contributing via the eosinophilic axes have been well appreciated and there is growing support for lung function genetics as well. In this study, we show that while both axes (blood eosinophil cell count and lung function traits) are genetically correlated with asthma risk, the low genetic correlation between eosinophil counts and lung function measures support these as orthogonal axes that contribute to asthma pathology. In other words, the pathways that underlie the eosinophilic axis of asthma biology are likely distinct from those underlying lung function determinants. Furthermore, our polygenic scoring analyses highlight a greater contribution of lung function genetics to moderate-to-severe asthma age of onset. Though these pathways may be distinct, it does not rule out that for any one patient both (or neither) pathways may be at play and are contributing to disease risk. Indeed, though both traits show genetic correlation with asthma risk, the correlation is moderate, highlighting the need to uncover additional axes underlying asthma biology.

From the whole-genome sequencing data we were able to assess the rare variant contribution to moderate-to-severe asthma risk. We replicated the association of a previously reported rare variant in IL33 with asthma risk and found significant enrichment of rare variant burden in candidate genes from common variant allergic disease loci. In a meta-analysis between this study and the partial release of WES data from the UK Biobank and asthma to moderate-to-severe asthma risk. We replicated the association of a previously reported rare variant in IL33 with asthma.

Though these pathways may be distinct, it does not rule out that for any one patient both (or neither) pathways may be at play and are contributing to disease risk. Indeed, though both traits show genetic correlation with asthma risk, the correlation is moderate, highlighting the need to uncover additional axes underlying asthma biology.

There are several limitations to our study. First, we utilized non-asthma disease samples as controls (see “Methods”) and second, our study samples are mostly derived from trial participants which may introduce biases based on various enrollment criteria (see “Methods”). To address the former, we flagged and removed any variants or results that failed to pass the differential effects test throughout this study (see “Methods” and Supplementary Note S1). To address the use of clinical trial samples, we replicated results (when possible) in external independent cohorts that did not have clinical trial-based enrollment criteria. Despite the use of clinical trial asthma participant cases and non-asthma disease controls, we were able to replicate many previously reported common variant asthma risk associations with similar directions of effect.

In summary, we carried out a whole-genome sequencing analysis of moderate-to-severe asthma. We discovered and replicated a common variant association that overlaps a COPD risk and lung function locus. We further provide genetic support for a role of lung function in both moderate to severe asthma risk and age of onset. Finally, our rare variant analyses replicated a previous association in IL33 and suggest some asthma common variant loci may contain additional rare variant support.

Methods

Cohort description. DNA was derived from moderate to severe asthma patients participating in the clinical studies of omalizumab (NCT00252135 [EXCELS], NCT00314575 [EXTRA] and NCT00813748 [X-PAND], lebrikizumab (NCT01545440 [LUTE], NCT01545453 [VERSE], NCT00930163 [MILLY], NCT01865275 [LAVOLTA II], NCT01868061 [LAVOLTA I]), and an additional asthma observational (NCT00091767 [TENOR II]) study and a smaller biomarker study (BOBCAT). Several of these studies (BOBCAT, EXTRA, LUTE, VERSE, MILLY, LAVOLTA I and LAVOLTA II) had inclusion criteria requiring pre-bronchodilator FEV1% pre-

Batch effects may be introduced if cases and controls are derived from differing sequencing methods and informatics pipelines. We therefore compared our asthma cases to 1140 and 2450 controls without asthma derived from clinical trial cohorts of AMD and RA, respectively that were generated with largely the same sequencing platforms (see Supplementary Note S1 and Supplemental Tables S11 and S13), protocols and also analyzed with the same downstream bioinformatics pipeline.

Data generation and quality control. Samples were sequenced to an average read depth of 30 x using the illumina HiSeq platform. Reads were aligned using BWA (version 0.7.9a-r786) to the GRCh38 reference genome (GCA_000001405.15) including alternate assemblies. In regions with alternate assemblies we followed the same alignment and variant calling procedures below, but used an alternate-assembly aware version of BWA (version 0.7.11) to properly handle alignment of reads to reference and alternate-assemblies. After alignment, we followed the GATK best practices guidelines to jointly call variants from WGS data using the Sentieon Genomics pipeline (version 201611.01). While we were able to jointly genotype all samples for exonic variants for rare variant analyses, due to memory and computing time requirements we were unable to jointly call whole-genome data. Therefore, we carried out several batches of joint-genotyping: one batch for asthma cases, and one for each
control disease (AMD and RA). We used a merged VCF from these three joint-genotyping runs as input for our whole-genome common variant (MAF ≥ 1%) analyses, without imputing or filling in genotypes missing between batches. We ran a fourth joint-genotyping run on case and control samples together in the exonic regions as input for our rare variant genic analyses.

We filtered variants that did not pass GATK variant quality recalibration threshold of 99% sensitivity and set any genotypes to missing where the genotype quality score was < 20. We further removed SNPs with a missingness rate > 0.05, a Hardy–Weinberg equilibrium p-value < 1 × 10⁻⁶, and an allelic depth balance test p-value < 0.01. The allelic depth balance test was carried out by testing for equal allele depth at heterozygote carriers via a binomial test. A total of 7,165,996 common variants and 3,600,569 rare exonic variants passed these variant-level QC metrics.

We estimated ancestry using predefined allele frequencies from reference populations in the 1000 Genomes. This approach has been implemented in iAdmix and the projection function in ADMIXTURE. To minimize confounding due to ancestry, we only retained individuals with fraction of European ancestry > 0.85 for analysis. We further excluded samples with high missingness (missingness > 0.1), relatedness (Z0 ≥ 0.4), excess heterozygosity (≥ 3 standard deviations from the mean), and principal component analysis (PCA) outliers. PCA outliers were defined as at least 6 standard deviations from the mean on any of the top ten principal components, with the outlier removal process iterated five times. Genetic sex was estimated from X-chromosome heterozygosity via PLINK.

**External datasets.** Pre-computed summary statistics for lung function measures, moderate-to-severe asthma risk, asthma risk, and allergic disease risk were downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics). Data for the blood eosinophil count GWAS were downloaded from http://www.bloodcellgenetics.org/. We focused on the INTERVAL study subset of Astle et al. to avoid any confounding from asthma and COPD/bronchiectasis patient samples present in the UKBB and BiLEVE cohorts.

**Moderate to severe asthma age of onset analysis in the UK Biobank.** This research has been conducted using the UK Biobank Resource under Application Number 44257. We defined moderate-to-severe asthma patients and controls in UK Biobank following Shrine et al. (see Supplemental Methods S1). Briefly, moderate-to-severe asthma cases were defined as having doctor diagnosed asthma and did not report having doctor-diagnosed emphysema or chronic bronchitis (UK Biobank data field 6152). In addition, cases were defined by prescriptions (UK Biobank data field 20003) satisfying the British Thoracic Society and British National Formulary guidelines for moderate to severe asthma. For a full list of medications see Supplemental methods of Shrine et al. Controls were defined as individuals with (1) no reported diagnosis of asthma, rhinitis, eczema, allergy, or chronic bronchitis/emphysema (UK Biobank data field 6152), (2) were not taking any medication for lung-related conditions, and (3) did not have asthma (J45-J46) nor COPD/bronchiectasis (J40-J44, J47) ICD10 codes.

Samples were removed if they were related or were not of majority European ancestry (fraction European ancestry > 0.70). In the age of onset analysis, 6312 cases were retained after sample quality control and including only those with age of onset data. We had 9,685,491 common variants (MAF > 1%) from which to construct polygenic scores after filtering on Hardy–Weinberg equilibrium (HWE) (P < 1 × 10⁻⁶), and missingness (missingness > 5%). For the rare variant meta-analysis at candidate genes of common variant loci, we subset to individuals with WES data based upon the October 2020 release of UK Biobank WES data. After sample level QC as described above, we retained 3418 moderate-to-severe asthma cases and 111,261 controls for analysis. We retained 1,759,922 exonic variants with MAF < 0.01, P_HWE > 1 × 10⁻⁶ and missingness > 0.2 as input to our gene-level burden tests.

**Statistical analysis.** For single variant association analysis, asthma risk was regressed on genotype assuming an additive genetic effects model. Single variant analyses were carried out in PLINK (version 1.9). For rare variant (MAF < 1%) burden tests, rare variants were aggregated into genic units. Specifically, individuals were scored for their rare allele carrier status and risk status was regressed on this binary rare allele carrier status. To enrich for rare variants with impact on protein function, we only considered variants with a “HIGH” (e.g. frameshift, splice acceptor, stop gained, etc.,...) and “MODERATE” (e.g. missense variant, inframe deletion or insertion, splice region variant, etc..) predicted impact as annotated with SNPEff (version 4.3q). We carried out two different sets of burden tests: (1) a stricter set consisting of high confidence variants predicted to have a high impact (as predicted by Loss-of-function transcript effect estimator, LOFTEE version 4.3q) and (2) all SNPEff predicted high impact variants as well as moderate impact variants with a PolyPhen score > 0.5.

All association analyses were corrected for genetic sex, and the top five PCs. We used FINEMAP (version 1.3.1) with default settings to estimate 95% credible sets for genome-wide significant associations in our study. Chromatin annotation of variants in the 95% credible set was carried out via haplor (version 3.0.1) to query HaploReg v4.1.

In this WGS-based study, as non-asthma controls samples consisted of AMD and RA patients, we implemented a differential effects test to help filter or flag associations that likely originated from the AMD or RA samples. The intuition motivating the test is that an asthma association will have a similar effect size regardless of the control cohort used in the GWAS (see caveats below). For each variant and association test, we therefore tested whether the effect size comparing asthma-cases to AMD-controls (ORasthma_AMD) was significantly different than...
the effect size when comparing the same asthma-cases to RA-controls (OR\(_{\text{asthma,RA}}\)). Formally, assuming the effect sizes \(\beta = \log(\text{OR})\) are normally distributed, the difference between the effect sizes are also normally distributed:

\[
N(\beta_{\text{asthma,AMD}} - \beta_{\text{asthma,RA}}, \sigma^2_{\text{asthma,AMD}} + \sigma^2_{\text{asthma,RA}} - 2 * \text{corr}(\beta_{\text{asthma,AMD}}, \beta_{\text{asthma,RA}}) * \sigma_{\text{asthma,AMD}} * \sigma_{\text{asthma,RA}})
\]

where \(\sigma^2_{\text{asthma,AMD}}\) and \(\sigma^2_{\text{asthma,RA}}\) are the variances for the respective effect sizes and \(\text{corr}(\beta_{\text{asthma,AMD}}, \beta_{\text{asthma,RA}})\) was previously derived\(^6\) and is approximated as:

\[
\text{corr}(\beta_{\text{asthma,AMD}}, \beta_{\text{asthma,RA}}) \approx \frac{n_{k0} \sqrt{n_{k0} n_{nl0}} + n_{k1} \sqrt{n_{k1} n_{nl1}}}{\sqrt{n_k n_l}}
\]

where \(n_{k0}, n_{k1},\) and \(n_k\) are the number of controls, the number of cases and the total number of samples in the asthma versus AMD association and \(n_{l0}, n_{l1}\) and \(n_l\) correspond to similar numbers for the asthma versus RA association. We observed that a conservative threshold of \(P < 0.01\) filtered association signals known to be originating from our non-asthma controls, we therefore set this as our threshold for filtering suggestive association signals originating from one of the controls (and not our asthma cases). We note that this test cannot distinguish between variants associated with increased asthma risk from variants associated with both AMD and RA. In addition, pleiotropic variants between asthma and either control will have reduced significance in the case of similar effect and increased significance in the case of opposite effects.

We employed co-localization analysis to provide further support of a shared causal variant between two association signals. Co-localization analysis was carried out in R with the \textit{coloc} software\(^{25}\). Input statistics included p-values generated in this study (after removing variants with a significant differential effects p-value, see above) and eQTL p-values generated by GTEx\(^{7}\) (v8). We used MAFs from our asthma GWAS study as additional input into the \textit{coloc} (version 3.1) software. We used a 1 megabase window around the index variant in the colocalization analysis.

SNP-based heritability of common variants and genetic correlation between traits was estimated using LD-score regression (version 1.0.1)\(^{21}\). Asthma risk statistics from this study were used as input after filtering variants failing the differential effects test (see above). We additionally used publicly available summary statistics for lung function\(^{29}\), blood eosinophil count\(^{30}\), and a recently reported moderate-to-severe asthma risk GWAS\(^8\). We used pre-computed LD-scores from the European subset of 1000 Genomes\(^{68}\). We further filtered our input variants to those available in HapMap3 as recommended. We assumed a prevalence of 0.084 for asthma (https://www.cdc.gov/nchs/products/databriefs/db94.htm). We further provide \(h^2\) estimates for a range of prevalences in the Supplementary Note S1.

Age of onset data were available for the EXCELS, TENOR II, EXTRA and Q4458G cohorts, for a total of 1456 subjects. In the UK Biobank cohort, age of onset data were available for 5362 subjects (see above). We dichotomized age of onset into childhood onset asthma (≤ 12 years of age) and adult onset asthma (≥ 25 years of age)\(^{69}\). This resulted in 665 childhood onset and 791 adult onset asthma subjects in our study, and 1261 childhood onset and 4101 adult onset asthma in the UK Biobank study.

We used four publicly available summary statistics to create polygenic scores (PS) for asthma\(^2\), allergic disease\(^22\), lung function\(^{29}\) and blood eosinophil count\(^{30}\). For each PS, we first subset out variants that were present in both our study and the public dataset. We restricted our analyses to variants with MAF > 1% as measured in the European ancestry subset of the 1000 Genomes cohort. We additionally filtered out any variants which could have strand ambiguity (A/T, C/G), and any variants in the HLA region. We next clumped common variants using PLINK1.9 to find independently associated variants that were associated with the scoring trait at \(P < 5 \times 10^{-8}\). Independent variants were defined as having pairwise low LD (\(r^2 < 0.05\)) and were at least 1 Mb apart. Finally, individuals were scored in PLINK based on their genetic risk for each trait using the log(odds-ratio) of each SNP. See Supplementary Note S1 for an evaluation of how these PS performed.

Association analyses between risk scores and phenotypes were carried out via logistic regression in R (version 3.4.3). We corrected for genetic sex and the first five principal components.

Candidate genes for the candidate gene rare variant analysis were identified from previous studies. Specifically, we obtained 139 genes from Ferreira et al. study with either eQTL or coding level evidence linking a gene to the sentinel variant (see Supplementary Tables S12 and S14 in the original publication)\(^{22}\). We obtained 17 candidate genes from the moderate-to-severe asthma GWAS study that had eQTL support linking the gene to the sentinel variant (see supplementary Table S8 in the original publication)\(^6\). Finally, we obtained 73 lung function related candidate genes implicated by coding, eQTL or pQTL data from a recent lung function GWAS (see Table 1 in the original publication)\(^{29}\).

Permutation gene-set enrichment analysis was used to test for significant enrichment of rare variant burden in the three candidate gene gene-sets from common variant allergic disease, moderate-to-severe asthma, and lung function loci. Input genes with a p-value < 0.01 in the differential effects test were excluded from this analysis. We compared the observed sum of the -\(\log_{10}(p\text{-values})\) for all genes in the gene-set to the observed sum of 5000 random samples from all genes tested. To ensure no biases were introduced by the number of rare variants tested in each gene, random genes were sampled to match the distribution of rare variants in the genes from each gene-set. A permutation-based p-value was calculated as the fraction of sums derived from random samples that were as or more significant than the observed sum.

**Ethics approval.** All research in this study was conducted in accordance with the Declaration of Helsinki. The data used in this study were generated from clinical trial participants who signed informed consent forms approved by the ethics committee or IRB responsible for the country or site where the trial’s participants donated
samples for research. Informed consent included use of these data for genetics research. Before execution of the study, an internal Genentech team of informed consent form experts reviewed the forms from all the studies to ensure appropriate use of the samples. The list of these ethics committee and/or IRBs is available in the Supplementary Note S1.

Data availability
The moderate-to-severe asthma risk summary statistics generated during this study are available from the corresponding author on reasonable request.

Received: 17 November 2021; Accepted: 23 March 2022
Published online: 02 April 2022

References
1. 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. https://doi.org/10.1038/cti.2017.54 (2017).
2. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. 50, 42–53. https://doi.org/10.1038/s41588-017-0014-7 (2018).
3. Kim, K. W. & Ober, C. Lessons learned from GWAS of asthma. Allergy Asthma Immunol. Res. 11, 170–187. https://doi.org/10.4168/aair.2019.11.2.170 (2019).
4. 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. https://doi.org/10.1038/s41588-018-0121-0 (2018).
5. Dahlin, A. et al. Large-scale, multietnic genome-wide association study identifies novel loci contributing to asthma susceptibility in adults. J. Allergy Clin. Immunol. 143, 1633–1635. https://doi.org/10.1016/j.jaci.2018.11.037 (2019).
6. Wan, Y. I. et al. Genome-wide association study to identify genetic determinants of severe asthma. Thorax 67, 762–768. https://doi.org/10.1136/thoraxjnl-2011-201262 (2012).
7. Li, X. et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. J. Allergy Clin. Immunol. 125, 328–335. https://doi.org/10.1016/j.jaci.2010.09.118 (2010).
8. Shrieve, N. et al. Moderate-to-severe asthma in individuals of European ancestry: A genome-wide association study. Lancet Respir. Med. 7, 20–34. https://doi.org/10.1016/S2213-2600(18)30389-8 (2019).
9. Herrera-Luis, E. et al. Genome-wide association study reveals a novel locus for asthma with severe exacerbations in diverse populations. Pediatr. Allergy Immunol. 32, 106–115. https://doi.org/10.1111/pai.13357 (2021).
10. Yan, Q. et al. A genome-wide association study of severe asthma exacerbations in Latino children and adolescents. Eur. Respir. J. 57, https://doi.org/10.1183/13993003.02693-2020 (2021).
11. Ahluwalia, T. S. et al. FUT2-ABO epistasis increases the risk of early childhood asthma and Streptococcus pneumoniae respiratory illnesses. Nat. Commun. 11, 6398. https://doi.org/10.1038/s41467-020-19814-6 (2020).
12. Bonnellyke, K. et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat. Genet. 46, S1–53. https://doi.org/10.1038/ng.2830 (2014).
13. Chung, K. F. et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. Eur. Respir. J. 43, 343–373. https://doi.org/10.1183/09031936.0020213 (2014).
14. Chanze, P. et al. Severe asthma in adults: What are the important questions?. J. Allergy Clin. Immunol. 119, 1337–1348. https://doi.org/10.1016/j.jaci.2006.11.702 (2007).
15. Nunes, C., Pereira, A. M. & Morais-Almeida, M. Asthma costs and social impact. Asthma Res. Pract. 3, 1. https://doi.org/10.1186/ s40773-016-0029-3 (2017).
16. Smith, D. et al. Role of IL-6 in asthma and other inflammatory pulmonary diseases. Immunol. Invest. 47, 291–295. https://doi.org/10.1038/ng.3211 (2015).
17. Iqbal, Z. et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nat. Commun. 6, 5965. https://doi.org/10.1038/ncomms9695 (2015).
18. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109 (2009).
19. Rincon, M. & Irvin, C. G. Role of IL-6 in asthma and other inflammatory pulmonary diseases. Int. J. Biol. Sci. 8, 1281–1290. https://doi.org/10.7150/ijbs.4874 (2012).
20. Durrant, D. M. & Metzger, D. W. Emerging roles of T helper subsets in the pathogenesis of asthma. Immunol. Invest. 39, 526–549. https://doi.org/10.3109/08820331036159489 (2010).
21. Bulik-Sullivan, B. K. et al. LD Score regression confounds disentangling from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295. https://doi.org/10.1038/ng.3211 (2015).
22. Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat. Genet. 49, 1752–1757. https://doi.org/10.1038/ng.3985 (2017).
23. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341. https://doi.org/10.1038/s41588-018-0184-y (2018).
24. Benner, C. et al. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501. https://doi.org/10.1093/bioinformatics/btw018 (2016).
25. Giambratolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383. https://doi.org/10.1371/journal.pgen.1004383 (2014).
26. Pavord, I. D. et al. After asthma: Redefining airways diseases. Lancet 391, 350–400. https://doi.org/10.1016/S0140-6736(17)30879-6 (2018).
27. Haldar, P. et al. Mepolizumab and exacerbations of refractory eosinophilic asthma. N. Engl. J. Med. 360, 973–984. https://doi.org/10.1056/NEJMoa0808991 (2009).
28. Hobbs, B. D. et al. General risk loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. Nat. Genet. 49, 426–432. https://doi.org/10.1038/ng.3752 (2017).
29. Shrieve, N. et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. 51, 481–493. https://doi.org/10.1038/s41588-018-0321-7 (2019).
30. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429. https://doi.org/10.1016/j.cell.2016.10.042 (2016).
31. Gudbjartsson, D. F. et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat. Genet. 41, 342–347. https://doi.org/10.1038/ng.323 (2009).
32. de Nijs, S. B., Venekamp, L. N. & Bel, E. H. Adult-onset asthma: Is it really different?. Eur. Respir. Rev. 22, 44–52. https://doi.org/10.1183/16000617.0007112 (2013).
Acknowledgements
We thank all of our Genentech colleagues involved in the Human Genetics Initiative with particular thanks to Suressh Selvaraj, Delphine Lagarde, Ward Ortmann, V. Alejandro Iglesias, and Biosample Management Repository; Slaton Lipscomb, Craig Amundsen, Ryan Lara, and Genentech IT; Sarah Malbon, Karine Brule, Kelly Mewes, Anja Dylst, Alex Hughes, and Biometrics; members of the Asthma Analysis team and study teams.

Author contributions
D.C., T.B., J.R., K.M., J.T., A.C., J.V., Z.K., and A.S. contributed analyses for this manuscript. D.C., B.L.Y., and D.F.C. wrote the main manuscript text. All authors reviewed the manuscript.

Competing interests
All authors are/were employees of Genentech (an affiliate of Roche) when this study was conducted.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09447-8.

Correspondence and requests for materials should be addressed to D.F.C. or B.L.Y.

Reprints and permissions information is available at 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022