

Neonatal genetics of gene expression reveal the origins of autoimmune and allergic disease risk

Qin Qin Huang^{1,2,*,#}, Howard H. F. Tang^{1,3}, Shu Mei Teo^{1,4}, Scott C. Ritchie^{1,4}, Artika P. Nath^{1,4}, Marta Brozynska^{1,4}, Agus Salim^{5,6}, Andrew Bakshi⁷, Barbara J. Holt⁸, Danny Mok⁸, Chiea Chuen Khor⁹, Peter D. Sly¹⁰, Patrick G. Holt^{8,10}, Kathryn E. Holt^{11,12}, Michael Inouye^{1,2,4,13,*}

¹ Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria 3004, Australia

² Department of Clinical Pathology, University of Melbourne, Parkville, Victoria 3010, Australia

³ School of BioSciences, The University of Melbourne, Parkville, Victoria 3010, Australia

⁴ Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, United Kingdom

⁵ Baker Heart and Diabetes Institute, Melbourne, Victoria 3004, Australia

⁶ Department of Mathematics and Statistics, School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, Victoria 3086, Australia

⁷ Monash Biomedicine Discovery Institute, Prostate Cancer Research Group, Department of Anatomy and Developmental Biology, Monash University, Clayton, Victoria 3800, Australia

⁸ Telethon Kids Institute, The University of Western Australia, Perth, Western Australia 6009, Australia

⁹ Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore

¹⁰ Child Health Research Centre, The University of Queensland, Brisbane, Queensland 4101, Australia

¹¹ Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia

¹² The London School of Hygiene and Tropical Medicine, London WC1E 7TH, United Kingdom

¹³ The Alan Turing Institute, London, United Kingdom

* Corresponding author: QQH (qinqin.huang@baker.edu.au); MI (minouye@baker.edu.au; mi336@medschl.cam.ac.uk)

Current address: Department of Human Genetics, Wellcome Sanger Institute, Cambridge, United Kingdom

Abstract

Chronic immune-mediated diseases of adulthood often originate in early childhood. To investigate genetic associations between neonatal immunity and disease, we collected cord blood samples from a birth cohort and mapped expression quantitative trait loci (eQTLs) in resting monocytes and CD4⁺ T cells as well as in response to lipopolysaccharide (LPS) or phytohemagglutinin (PHA) stimulation, respectively. *Cis*-eQTLs were largely specific to cell type or stimulation, and response eQTLs were identified for 31% of genes with *cis*-eQTLs (eGenes) in monocytes and 52% of eGenes in CD4⁺ T cells. We identified *trans*-eQTLs and mapped *cis* regulatory factors which act as mediators of *trans* effects. There was extensive colocalisation of causal variants for cell type- and stimulation-specific neonatal *cis*-eQTLs and those of autoimmune and allergic diseases, in particular *CTSH* (Cathepsin H) which showed widespread colocalisation across diseases. Mendelian randomisation showed causal neonatal gene transcription effects on disease risk for *BTN3A2*, *HLA-C* and many other genes. Our study elucidates the genetics of gene expression in neonatal conditions and cell types as well as the aetiological origins of autoimmune and allergic diseases.

Introduction

Infancy is a critical period during which physiological and developmental changes impact the pathogenesis of conditions later in life^{1,2}. Many complex diseases, in particular immune and respiratory conditions, are partially determined by genetic predisposition and early-life environment exposures, such as microbes or allergens³⁻⁵. Yet, despite increasing evidence of its importance, little is known about early life genetic regulation of gene expression and its relevance to the genetic predisposition for diseases which arise later in life.

Expression quantitative trait loci (eQTL) studies have provided insights into the gene regulatory effects of genetic variants and their relationship with complex disease⁶⁻⁹. The majority of eQTLs have been identified in adult tissues, and exploration of eQTLs in perinatal tissues has only recently commenced: for example, eQTLs identified in foetal placentas¹⁰ and foetal brains¹¹ are enriched for genetic variants associated with development (e.g. adult height) and neuropsychiatric (e.g. schizophrenia) traits, respectively. In addition to genetic variation, disease development is influenced by individual response to external stimuli. Understanding the interaction between eQTLs and external stimuli can give insights into the condition(s), whether they be cell type, microbe or temperature, under which genetic variants may exert their effect on disease. Previous studies have investigated response eQTLs (reQTLs), eQTLs with genetic effects modified by external stimulation, in CD14⁺ monocytes^{12,13}, macrophages¹⁴, dendritic cells^{15,16}, and CD4⁺ T cells¹⁷. However, reQTL studies to date have largely been performed using samples from adult tissues and cell types and thus our knowledge of the genetic landscape of neonatal gene expression responses to common immune-mediated stimuli is limited.

Here, we characterise the genetics of gene expression in the innate and adaptive arms of the neonatal immune system using purified cord blood samples from 152 neonates¹⁸⁻²³. In these samples, we mapped *cis*- and *trans*-eQTLs of monocytes and CD4⁺ T cells as well as reQTLs for monocytes stimulated with lipopolysaccharide (LPS; a component of bacterial cell walls) and CD4⁺ T cells stimulated with phytohemagglutinin (PHA; a pan-T cell mitogen). Using mediation analysis, putative *trans* gene regulation was investigated to identify *cis* regulatory mechanisms. We explored the shared genetic basis of neonatal eQTLs and reQTLs with common autoimmune and allergic diseases. Finally, we used Mendelian randomisation to uncover causal effects of neonatal monocyte and CD4⁺ T gene expression on disease risk, and to highlight the potential importance of the perinatal period in understanding the origins of immune-mediated disease.

Results

Genetics of neonatal gene expression in innate and adaptive immunity

We performed eQTL analysis on *in vitro* cultures of resting and stimulated neonatal immune cells from 152 neonates of the Childhood Asthma Study (CAS) cohort (**Figure 1**). The cell cultures comprised four experimental conditions: resting monocytes, LPS-stimulated monocytes, resting T cells, and PHA-stimulated T cells. The total number of samples available for eQTL analysis was 116 for resting monocytes, 125 for LPS-stimulated monocytes, 126 for resting T cells, and 127 for PHA-stimulated T cells.

To identify *cis*-eQTLs, we applied a hierarchical procedure to correct for multiple testing within each experimental condition at 5% false discovery rate (FDR; **Methods**). Stimulated cells yielded a larger number of *cis*-eQTLs and associated genes (eGenes) than resting cells (1,347 vs. 971 eGenes in PHA-stimulated vs. resting T cells, respectively; 376 vs. 136 in LPS-stimulated vs. resting monocytes, respectively; **Figure 2A**, **Table S1–4**). To investigate the differences in numbers of eGenes between conditions, we repeated the analysis controlling for differences in sample size (randomly sampling 116 samples in each condition). This yielded similar results to the numerical distribution of *cis*-eGenes: 1,231, 900, and 350 in PHA-stimulated T cells, resting T cells, and LPS-stimulated monocytes, respectively. The lower number of eQTLs in monocytes may be explained by fewer genes being expressed (**Figure S1**).

For eGenes with eQTLs in multiple experimental conditions, we performed conditional analysis to distinguish whether these were independent or shared signals between conditions (**Methods**). The majority (74%) of eQTL signals were specific to one cell type or stimulatory condition (**Figure 2B**), consistent with previous observations¹². We observed a majority of *cis*-eQTL effects after stimulation: 60% (262 of 376) of eGenes in LPS-stimulated monocytes and 58% (778 of 1,347) in PHA-stimulated T cells. Using a two-step conditional analysis (**Methods**), PHA-stimulated T cells had the largest number of eGenes (6.3%; **Table S5**) with multiple independent eQTL signals. *GARFIELD* enrichment analysis²⁴ showed that the *cis*-eSNPs were enriched in 3' untranslated regions (UTR), 5' UTR, and exon regions (**Figure S2**), consistent with known mechanisms of *cis*-eQTLs.

In comparing our resting and LPS-stimulated monocytes to those from adults in Fairfax *et al.*¹², we found that approximately half of the *cis*-eQTLs from neonatal monocytes (including variants in high LD, $r^2 \geq 0.8$) replicated in adult monocytes (**Table S6**). Similarly, we compared resting and stimulated neonatal T cells to adults T cells of the DICE study⁷: although the stimuli were different, 32% and 18% of *cis*-eQTLs from resting and PHA-stimulated CD4⁺ T cells in neonates replicated in resting and anti-CD3/CD28-stimulated CD4⁺ T cells, respectively.

Genetics of neonatal gene expression in response to stimuli

To quantify how genetic regulation of gene expression is altered by external stimuli, we identified response eQTLs (reQTLs) and response eGenes (reGenes) by performing interaction tests on the top eSNPs of each eGene in monocytes and T cells separately, and controlling FDR at 5% using permutation-adjusted P-values (**Methods**). In monocytes, we identified 125 significant reQTLs involving 125 unique reGenes (31% of 398 monocyte eGenes); in T cells, we identified 956 reQTLs involving 918 unique reGenes (52% of 1,749 T cell eGenes), among which 38 reGenes had distinct *cis*-eQTLs in two conditions where both eQTLs were reQTLs (**Table S7–8**). Consistent with our findings for *cis* eQTLs and eGenes, the number of reQTLs and proportion of reGenes were greater in stimulated compared to resting conditions.

For two reQTLs, the direction of eQTL effect changed between conditions (**Figure 2C**). The 'C' allele of the top eSNP (rs5751775) for *DDT* (D-dopachrome tautomerase), a gene functionally related to the inflammatory cytokine *MIF* (migration inhibitory factor), increased *DDT* transcription in resting T cells but decreased expression after PHA stimulation (**Figure S3A, Table S8**). Similarly, the 'T' allele of the top eSNP (rs13068288) for *ZNF589* increased transcription of *ZNF589* in resting T cells but reduced expression after PHA stimulation (**Figure S3B, Table S8**).

Disentangling *trans* and *cis* effects using mediation analysis

We identified 25 *trans*-eQTLs in T cells (10 in resting, 15 in PHA-stimulated), and one *trans*-eQTL in monocytes (**Figure 3A, Table S9**) at a genome-wide FDR of 5%. Notably, the *trans*-eQTL for *MYH10*, a component of myosin heavy chain which regulates cytokinesis, was shared across all four experimental conditions; furthermore, the same eQTL was associated with multiple *trans*-eGenes in T cells: *MIR130A* and *STX1B* in resting and stimulated T cells, and *IP6K2* and *MIR1471* in resting T cells only (**Figure 3A**).

Consistent with previous reports that *trans*-eQTLs are enriched for *cis*-eQTLs²⁵, we found that multiple *trans*-eQTLs (the single *trans*-eQTL in monocytes, 6 of the 10 in resting T cells, and 3 of the 15 in PHA-stimulated T cells) were significantly associated with local genes in *cis* (**Figure 3A, Table S9**). The *trans*-eQTL for *MYH10* was also a *cis*-eQTL for *RPS26* (part of the 40S subunit of the ribosome) in all conditions except PHA-stimulated T cells. The top eSNP (rs1131017) for *RPS26*, located in its 5' UTR, was also a *cis*-eQTL for *SUOX* (sulphite oxidase, a homodimer in the intermembrane space of mitochondria) in resting and stimulated T cells. Separately, the *trans*-eQTL (rs687492) for microRNA *MIR330* was also a *cis*-eQTL for the long non-coding RNA *SNHG8* in resting and stimulated T cells.

Mediation analysis revealed that the *trans*-eQTL effects of rs687492 on *MIR330* were *cis* mediated through *SNHG8* (**Figure 3B–C, Table S10, Methods**), indicating a potential pathway containing this

lncRNA-miRNA cross-talk. Furthermore, mediation analysis also revealed the regulatory logic of the *cis*-eQTL (rs1131017) for *SUOX* and *RPS26*, revealing that rs1131017's *trans*-effects on *MYH10*, *MIR130A*, *IP6K2* were mediated through *RPS26* and not *SUOX* in resting T cells (**Figure 3D**, **Table S10**).

Genetic overlap with immune-mediated diseases

To investigate the genetic overlap between neonatal gene expression and disease, we used a multi-pronged approach. First, we performed enrichment analyses to test for significant overlaps between the *cis*-eQTLs and variants associated with immune-mediated disease in genome-wide association studies (**Methods**). We found widespread enrichment amongst *cis*-eQTLs for genetic variants associated with diseases such as allergic disease (asthma, hay fever, or eczema) and inflammatory bowel disease (**Figure S4**).

Second, we performed colocalisation analysis²⁶ to identify variants sharing regulatory (eQTL) and disease-associated (GWAS) signals (**Methods**). In total, we observed 68 colocalisations, involving 5, 9, 15, and 17 independent *cis*-eQTLs in resting monocytes, LPS-stimulated monocytes, resting T cells, and PHA-stimulated T cells, respectively (**Figure 4A**, **Table S11**). Our analysis revealed widespread colocalisation of the *cis*-eQTL for *BACH2* in resting T cells with variants for autoimmune thyroid disease, celiac disease, multiple sclerosis, rheumatoid arthritis, and type 1 diabetes (**Figure S5**). *BACH2* encodes a transcriptional repressor that restrains terminal differentiation and promotes the development of memory lymphocytes including CD8⁺ T cells²⁷ and B cells²⁸. At the *BACH2* locus, the 'A' allele at the top eSNP (rs72928038) was associated with decreased *BACH2* expression and increased risk of the above diseases. This was consistent with previous studies which showed mutations and loss-of-function variants of *BACH2* resulted in immunodeficiency and disruption to regulatory T cells function with subsequent autoimmunity^{29,30}.

We found 17 colocalisations of reQTLs and disease variants, in total involving 12 reQTLs: one monocyte reQTL specific to LPS stimulation (eQTL for *CTSH*), and 11 reQTLs in T cells, among which 8 were specific to PHA stimulation. Notably, the reQTL for *IL13* in PHA-stimulated T cells colocalised with GWAS hits associated with asthma, allergic sensitisation, and allergic disease (**Figure 4**, **Table S11**). The 'T' allele of the top eSNP (rs1295686) was associated with greater *IL13* expression in PHA-stimulated T cells as well as increased risk of all three diseases (**Figure 4B**). rs1295686 is intronic to *IL13* and in strong LD ($r^2 > 0.98$) with four other eSNPs, including an Gln144Arg missense SNP (rs20541) in *IL13*. At the *CCL20* locus, the 'A' allele of the *cis*-eQTL/reQTL (rs13034664) in PHA-stimulated T cells was associated with lower *CCL20* expression as well as increased risk of childhood-onset asthma (**Figure 4A**, **Figure S6**, **Table S11**). *CCL20* is part of the *CCR6-CCL20* receptor-ligand axis, a key driver of dendritic cell chemotaxis³¹.

Our analyses uncovered complex condition-specific colocalisations at multiple loci. A T cell ubiquitin ligand that regulates apoptosis (*UBASH3A*) had two independent *cis*-eQTLs in resting and stimulated T cells, which were also reQTLs. However, only the *cis*-eQTL in resting T cells (rs1893592) colocalised with celiac disease, rheumatoid arthritis, and primary sclerosing cholangitis (PSC) (**Figure 4, Table S11**). *Cis*-eQTLs for *CTSH*, which encodes the lysosomal cysteine proteinase cathepsin H, colocalised with signals for different diseases in a cell type and condition specific manner (**Figure S7, Table S11**). *Cis*-eQTLs for *CTSH* in resting monocytes and resting T cells both colocalised with GWAS hits for celiac disease, narcolepsy, and type 1 diabetes. On the other hand, *cis*-eQTLs for *CTSH* in LPS-stimulated monocytes and PHA-stimulated T cells both colocalised with primary biliary cirrhosis (PBC).

Causal effects of condition-specific gene expression on immune-mediated diseases

To identify putative causal effects of neonatal gene expression on risk of autoimmune and allergic disease, we performed two-sample Mendelian randomisation (MR) analysis using *cis*-eQTLs as genetic instruments, the neonatal *cis*-eGene as exposure, and disease as outcome (**Methods**). We tested the 52 eGenes which had 3 or more genetic instruments available and the diseases above, for which we had GWAS summary statistics available. We considered genes for which at least three of four MR methods (inverse variance weighted, weighted median, weighted mode, and MR Egger) were in agreement in detecting significant causal effects (P-value ≤ 0.05) on a disease without significant pleiotropic effects (**Table S12**).

In our MR analysis, we found multiple conditions where neonatal gene expression had a causal effect on multiple diseases (**Figure 5**), including *BTN3A2* (butyrophilin subfamily 3 member A2), *HLA-C* (major histocompatibility complex class I molecule), *MICB* (ligand for an activatory receptor expressed on natural killer cells, CD8⁺ $\alpha\beta$ T cells, and $\gamma\delta$ T cells), *ZNRD1* (RNA polymerase 1 subunit), and *SLC22A5* (carnitine transporter) (**Table S12**). *BTN3A2* had a relatively large number of genetic instruments for resting (7 to 8) and stimulated (3 to 5) T cells, and the causal estimates were similar between these two conditions (**Figure S8–9**). In resting T cells, increased expression of *BTN3A2* was causally associated with decreased risk of asthma (weighted mode causal estimate = -0.056 log odds decrease per s.d. increase in *BTN3A2*), both childhood- and adult-onset asthma (-0.047 and -0.039, respectively), allergic rhinitis (-0.044), PSC (-0.440), and systemic lupus erythematosus (SLE; -0.256). Conversely, increased *BTN3A2* expression was associated with increased risk of inflammatory bowel disease (IBD; 0.025), including Crohn's disease (0.053), as well as risk of PBC (0.129), where PBC variants also showed colocalisation with *BTN3A2* eQTLs (**Figure 4, Figure S10**). Expression of *HLA-C* in T cells showed strong causal association with autoimmunity, in particular positive causal effects on psoriasis, SLE, PSC, multiple sclerosis, IBD, and ulcerative colitis; and negative causal effects on juvenile idiopathic arthritis, PBC, and rheumatoid arthritis (**Figure 5**).

Discussion

In this study we investigated the genetic regulation of gene expression in the innate and adaptive arms of the neonatal immune system and its relationship to the genetic basis of autoimmune and allergic diseases. In this context, we illustrated that the genetics of gene expression in neonates is strongly specific to cell type and stimulatory condition. We described regulatory mechanisms of eQTLs whose *trans* effects are mediated via gene expression in *cis*. In exploring the potential early life origins of disease, our analyses showed extensive genetic overlap of genetic variants associated with immune-mediated diseases and those with effects on gene expression in neonatal immune cells. Finally, Mendelian randomisation showed that myriad changes in gene expression at birth had potentially causal effects on autoimmune and allergic disease risk.

We observed stimuli changing the direction of eQTL effects in resting and PHA-stimulated T cells. This was the case for the 'C' allele of reQTL (rs5751775) and *DDT*, a cytokine structurally and functionally related to *MIF*, a critical regulator of both innate and the adaptive immune response^{32,33}. It is known that eQTL effects can change direction in immune cells³⁴. In the Genotype-Tissue Expression (GTEx) Project⁶, the direction of eQTL effect for the 'C' allele at rs5751775 was also variable across liver, pancreas, stomach, testis, brain, and muscle tissues.

Our results revealed a *trans* mediation role for *RPS26*, rather than *SUOX*, consistent with previous studies linking *RPS26* to *IP6K2*^{35,36}. *RPS26* and *SUOX* were both identified as significant mediators in multiple tissues in the GTEx dataset; however, within the same tissue, the *trans*-associations mediated through *SUOX* were not identical to those through *RPS26*, indicating distinct effects of *SUOX* and *RPS26* on distant genes³⁶. *RPS26* encodes a ribosomal subunit protein which, apart from its role in ribosome assembly and translation^{37,38}, is involved in various other cellular processes, including nonsense-mediated mRNA decay³⁹ and p53 transcriptional activity⁴⁰. It is likely that the broad *trans* effects of *RPS26* are related to its ribosomal functions.

At the *CCL20* locus, the reQTL (rs13034664) in PHA-stimulated T cells colocalised with childhood-onset asthma⁴¹ (**Figure 4A, Figure S6**). *CCL20* encodes a C-C chemokine ligand that binds to a G protein-coupled receptor, and elevated *CCL20* expression has been shown in airways of patients with chronic obstructive pulmonary disease (COPD)⁴² and asthma^{43,44}. *CCL20* induces mucin production by binding to its unique receptor (CCR6) in human airway epithelial cells⁴⁵. However, in PHA-stimulated T cells, the 'A' allele of the reQTL (rs13034664), which was linked to increased risk of childhood-onset asthma, was associated with reduced *CCL20* expression (**Figure S6B**). This reQTL and its direction of effect were replicated by others in activated CD4⁺ T cells⁷. T cells themselves respond to *CCL20* via binding to CCR6, and this process is observed during allergen provocation⁴⁶.

On the other hand, the reQTL for *IL13* in neonatal PHA-stimulated T cells appeared to share causal variants with allergic disease and asthma, suggesting that this reQTL may affect allergic disease risk via mechanisms involving T cell activation and interleukin 13 (IL-13). PHA is a pan T cell mitogen⁴⁷, and downstream intracellular signalling may be shared between PHA and allergen-mediated T cell activation. IL-13 is produced by activated CD4⁺ and CD8⁺ T cells⁴⁸ among others, promoting immunoglobulin E (IgE) production in B cells⁴⁹. IL-13 has been shown to induce asthma symptoms including airway hyper-responsiveness, increased total serum IgE, and increased mucus production in murine models⁵⁰. Increased IL-13 expression is observed in sputum and bronchial biopsy in mild⁵¹ and severe⁵² asthma and can serve as a biomarker for severe refractory asthma⁵³. Therapies that target IL-13 (anti-IL-13 antibodies) have been developed, such as lebrikizumab⁵⁴ and tralokinumab⁵⁵; however, they show inconsistent or only modest effects in treating severe asthma exacerbations in phase 3 clinical trials. Our results are consistent with IL-13 as a therapeutic target for asthma and may suggest potentially increased efficacy in genetic subgroups.

We found strong causal effects of neonatal *BTN3A2* expression on various autoimmune and allergic diseases. Butyrophilin (BTN) family members are immunoglobulin-like molecules that act as immune check-point regulators with roles in self-tolerance⁵⁶. Increased *BTN3A2* protein expression is a favourable prognostic biomarker in epithelial ovarian cancer patients, and indicates a higher density of intraepithelial infiltration of T cells⁵⁷. *BTN3* family members *BTN3A1* and *BTN3A3* are proximal to *BTN3A2*. The antigen-presenting *BTN3A1* is critical to human $\gamma\delta$ T cell activation^{58,59}. A recent study showed that *BTN3A2* regulated subcellular localisation of *BTN3A1*, and both were required for T cell activation⁶⁰. Previous MR analysis found *BTN3A2* lung expression had a causal effect on COPD risk⁶¹. Our findings suggest that altered neonatal *BTN3A2* expression, with presumed subsequent dysfunctional immunomodulation, plays a role in the pathogenesis of multiple inflammatory conditions.

In our study, neonatal *HLA-C* expression in both monocytes and T cells was causally associated with multiple autoimmune diseases such as psoriasis, SLE, and primary biliary cirrhosis. *HLA-C* encodes an MHC Class I receptor which presents antigens to CD8⁺ T cells. It is also the major ligand for killer immunoglobulin-like receptors (KIRs), which regulate the activity of natural killer (NK) cells. *HLA-C* is an established locus for psoriasis susceptibility⁶² and the interaction between *HLA-C* and *ERAP1* is associated with psoriasis risk, where *ERAP1* variants only have psoriasis effects in individuals with the *HLA-C* risk allele⁶³. In our analysis of resting neonatal T cells, the largest causal effect of *HLA-C* expression was for psoriasis.

In conclusion, our study shows the remarkable complexity of the genetic regulation of gene expression in the innate and adaptive arms of the immune system at birth, and its potential role in the pathogenesis of autoimmunity and allergic disease.

Methods

Study cohort and RNA sample preparation

The study population is a subset of the Childhood Asthma Study (CAS), a prospective birth cohort of 234 individuals followed from birth to up to 10 years of age¹⁸⁻²³. Cord blood samples were collected for 152 individuals at birth. One million peripheral blood mononuclear cells (PBMCs) from each individual were stimulated with either an innate immune system stimulant (LPS: lipopolysaccharide), or a pan T cell stimulant (PHA: phytohemagglutinin) for 24 hours (**Figure 1**). Unstimulated resting PBMC samples were also available. Non-adherent cells in suspension from resting and PHA-stimulated cultures were removed and purified for CD8⁻/CD4⁺ T cells using Dynabeads (Invitrogen) and stored in RNAProtect Cell reagent (Qiagen). Cells remaining in suspension in resting and LPS-stimulated cultures were aspirated, leaving an enriched population of monocytes and macrophages adhered to the culture wells. These adherent cells were then resuspended and transferred into RNAProtect. All cells in RNAProtect Cell reagent were banked at -80°C.

The cells were thawed and centrifuged briefly for RNA extraction. Reagent was removed and total RNA was extracted from pelleted cells by an established in-house procedure using TRIzol (Life Technologies) in combination with RNEasy MinElute columns (Qiagen). The aqueous phase containing the RNA was then loaded onto an RNeasy MinElute column (Qiagen) to purify and concentrate the RNA. RNA quality was assessed on a Bioanalyzer 2100 using the RNA 6000 Nano kit (Agilent). There were 607 samples (one missing sample) in total from 152 individuals for gene expression profiling.

Gene expression profiling and data processing

Total RNA from four cell culture conditions (resting and LPS-stimulated monocytes, and resting and PHA-stimulated T cells) was quantified with Illumina HumanHT-12 v4 BeadChip gene expression array at the Genome Institute of Singapore. After excluding 31 samples with suspected cross-contamination or insufficient quantity of cDNA, 576 samples were successfully scanned. The raw microarray data and probe detection P-values were exported by the Illumina software *BeadStudio*. We first removed 3 samples with zero intensity for almost all probes including negative controls and probes targeting housekeeping genes (2 resting monocyte samples and 1 LPS-stimulated monocyte sample). We further removed 16 outlier samples (8 resting monocyte samples, 5 LPS-stimulated monocyte samples, and 3 resting T cell samples) with a low number of detectable probes (lying outside median $\pm 2 \times$ inter quartile range). Compared with other samples, these excluded samples had much lower intensity for positive controls including those targeting housekeeping genes.

After quality control, 557 samples remained for normalisation (resting monocytes: 130, LPS-stimulated monocytes: 141, resting T cells: 142, and PHA-stimulated T cells: 144). We performed background correction based on the intensity of the 770 negative control probes on the microarray, and then we performed quantile normalisation and log₂ transformation within each cell type and condition using the *neqc* function from the *limma* R package⁶⁴. We used updated probe annotation data and restricted the analysis to 33,436 reliable probes, excluding unaligned probes and probes aligned to multiple regions that were more than 25 bp apart⁶⁵. Fifteen probes with missing data in ≥ 5 samples were removed. Detectable probes targeting autosomal genes (N = 20,532) were kept, comprising of probes with *BeadStudio* detection P-values ≤ 0.01 in $\geq 2.5\%$ of the samples from a specific condition group, or in $\geq 5\%$ of all samples¹². Gene annotation was obtained from the GENCODE⁶⁶ release 19 (GRCh37 alignment, downloaded in October 2017). Among detectable probes, 19,230 had gene annotation in the GENCODE reference data. For genes that had multiple probes, we kept the probe with the highest mean intensity⁶⁷, resulting in 13,109 autosomal genes. For eQTL analysis, we performed a rank-based inverse normal transformation within each group, so that each gene expression followed a standard normal distribution.

Genotyping and imputation

Genomic DNA was extracted from blood samples collected from 218 individuals. Genotyping was performed with Illumina Omni2.5 BeadChip array, with coverage of approximately 2.5 million markers. Variants with missing call rates $> 1\%$, MAF $< 1\%$, or Hardy-Weinberg Equilibrium (HWE) test P-value $< 1 \times 10^{-6}$ were excluded, and individuals with missing call rates $> 1\%$ were removed. This produced an initial count of 1.4 million SNPs for 215 genotyped individuals. Of these, a total of 135 children also had gene expression data from cord blood (i.e. overlap with the 152 individuals described previously).

We performed genotype imputation using the Michigan Imputation Server⁶⁸ with Haplotype Reference Consortium⁶⁹ (HRC) release r1.1 as the reference panel. After filtering out variants with low imputation accuracy ($R^2 < 0.3$), 12.7 million SNPs remained. For eQTL analysis, we focused on 4.3 million SNPs with MAFs $\geq 10\%$. The MAF cut-off used here was suggested by an eQTL simulation study in order to avoid inflated false positives in low-frequency variants given our limited sample size⁷⁰.

Cis-eQTL mapping and conditional analysis

To identify *cis*-eQTLs within each cell type and treatment group, we performed linear additive regression to model the effect of each SNP located within 1Mb of the transcription start site (TSS) of the corresponding gene using the *Matrix eQTL* R package⁷¹. The sample size for eQTL mapping in each experimental condition was: 116 for resting monocytes, 125 for LPS-stimulated monocytes, 126 for resting T cells, and 127 for PHA-stimulated T cells. Genotype data were recoded as 0, 1, 2 based on the

dosage of the HRC alternative allele. Gender, first three genotype PCs⁷² and first ten PEER⁷³ factors capturing technical variation in transcriptomes were included as covariates in the linear model.

We applied a hierarchical correction procedure to correct for multiple testing⁷⁰. Firstly, nominal P-values for all *cis*-SNPs from *Matrix eQTL* were adjusted by multiplying the number of effective independent SNPs for each gene (local correction), which was estimated by *eigenMT* based on genotype correlation matrix⁷⁴. Secondly, the minimum locally adjusted P-value for each gene was kept and the FDR of significant genes was controlled at 5% using the Benjamini-Hochberg (BH) FDR-controlling procedure (global correction)⁷⁵. Genes with global FDR ≤ 0.05 were considered significant eGenes. Thirdly, to obtain the list of significant eSNPs for each eGene, the locally adjusted minimum P-value corresponding to the global FDR threshold of 0.05 was calculated, and SNPs with a locally adjusted P-value lower than the threshold were considered significant eSNPs.

Next, we performed conditional analyses to identify additional independent eQTL signals for each eGene. The gene-level P-value nominal thresholds calculated in the hierarchical multiple-testing correction (*eigenMT*-BH) were used to determine significant associations: the locally adjusted minimum P-value corresponding to the global FDR threshold of 0.05 multiplied by the number of estimated independent SNPs for each gene. We used a two-step conditional analysis scheme as follows⁷⁶:

(1) *Forward stage*. For each eGene, the number of independent *cis*-eQTL signals was learnt from an iterative procedure. We started from the top SNP with the minimum P-value for the eGene, which was added as a covariate in the linear model to test for *cis*-eQTLs. If any significant SNPs (with P-values smaller than the gene's nominal threshold) were identified, the new top SNP identified in this iteration was added to the list of independent eQTL signals. In the next iteration of eQTL mapping, all previously identified eSNPs were adjusted for as covariates. The forward stage terminated if no additional significant associations were identified.

(2) *Backward stage*. In this stage, the final list of significant SNPs representing each independent eQTL signal was determined. Let the list of independent SNPs for each eGene obtained from the forward stage be $SNP_1, SNP_2, SNP_3, \dots, SNP_M$, where M is the number of independent eQTL signals. Each of the independent eQTL signals was tested separately using a leave-out-one model adjusting for all other SNPs in the list as covariates. For example, when the i^{th} eQTL signal was tested, $SNP_1, \dots, SNP_{i-1}, SNP_{i+1}, \dots, SNP_M$ were added as covariates together with other covariates used in the original eQTL scan. The final set of independent eQTLs comprised of the eSNPs that remained significant in the backward stage.

For genes that had eQTLs in more than one experimental condition, we also applied conditional analyses to identify independent eQTL signals between conditions. More specifically, to determine whether two eQTL signals for an eGene identified in two experimental conditions were independent or the same signal, we adjusted for the top eSNP in one condition by adding it as a covariate in the linear model and performed eQTL scan again in the other condition. If any SNP was significant in the conditional model using the P-value threshold determined by the hierarchical correction procedure (eigenMT-BH), we considered these two eQTLs as independent signals. If none were significant in the conditional model, we considered it as shared eQTL signal between two conditions or lack of power to detect the independent signal.

Replication of *cis*-eQTLs in external datasets

We downloaded summary statistics of significant *cis*-eQTLs from a response eQTL study (Supplemental Table S2 of the Fairfax *et al.* study¹²) and the DICE (database of immune cell expression, eQTLs, and epigenomics) project (URLs)⁷. Fairfax *et al.* mapped *cis*-eQTLs in resting and LPS-stimulated CD14⁺ monocytes (with two different durations of LPS: 2 or 24 hours) obtained from adults aged from 19 to 56 years, with the sample size being 414 for resting monocytes, and 261 and 322 for monocytes treated with LPS for 2 hours and 24 hours, respectively¹². The downloaded *cis*-eQTLs were significant at FDR 5% using pooled BH FDR method (i.e. BH FDR-controlling procedure applied to all tests). In the DICE study, 13 immune cell types were collected from 91 subjects, and CD4⁺ T cells and CD8⁺ T cells also had activated conditions⁷. We downloaded eQTLs (P-value $\leq 1 \times 10^{-4}$) identified in CD14⁺ monocytes, CD4⁺ T cells, and activated CD4⁺ T cells (4-hour treatment with anti-CD3/CD28 antibodies). In each of the four experimental conditions, we focused on top eSNPs and eSNPs in high LD ($r^2 \geq 0.8$) for each eGene, and if any of these eSNPs were significantly associated with the same eGene in the corresponding conditions in the Fairfax *et al.* and the DICE datasets, this eQTL signal was considered as replicated.

Enrichment analysis

We performed enrichment analyses using *GARFIELD* (version 2) to investigate the enrichment patterns of *cis*-eQTLs using predefined features (“annotation data”) such as genic annotations from ENCODE, GENCODE, and Roadmap Epigenomics project provided by this tool²⁴. *GARFIELD* evaluates enrichment using generalised linear regression models that account for allele frequency, distance to the nearest gene TSS, and LD. LD correlation based on the UK10K dataset is also provided by the software. In each experimental condition, we used P-values for all SNPs tested in *cis*-eQTL analysis. If a SNP was tested for association with multiple genes, the smallest P-value was kept. Enrichment odds ratios were calculated at various eQTL significance thresholds: 1×10^{-3} , 1×10^{-4} , ..., 1×10^{-8} .

Response eQTL detection

Response eQTLs (reQTLs) were identified in monocytes and T cells separately. For each cell type, we focused on top eSNPs of eGenes that were significant in either resting or stimulated conditions. For eGenes that were significant in both conditions and for which two top eSNPs were not in high LD ($r^2 < 0.8$), we tested both of the top eSNPs; on the other hand, if the two top eSNPs were in high LD, we tested the more significant one, to reduce tests on redundant SNPs. In monocytes, 417 interaction tests involving 398 eGenes were performed, and 1,959 tests involving 1,749 eGenes were performed in T cells. Gene expression data in two conditions were combined within each cell type, and the following linear mixed-effects model was tested for eGene–top eSNP pairs using the *lmer* function in the *lme4* R package⁷⁷:

$$y_i \sim x_i + c_i + x_i \times c_i + cov_i^1 + \dots + cov_i^{14} + cov_i^1 \times c_i + \dots + cov_i^{14} \times c_i + (1|Subj_i)$$

where y_i indicates the expression level of an eGene for the i^{th} sample, x_i the SNP allele dosage, c_i the condition (resting: 0 and stimulated: 1) in which the gene expression was measured, $cov_i^1, \dots, cov_i^{14}$ the 14 covariates used in the original eQTL mapping (gender, 3 genotype PCs, and 10 PEER factors), and $Subj_i$ the individual from which the i^{th} sample was taken. The term $x_i \times c_i$ models the interaction between the genotype and the condition, and $(1|Subj_i)$ indicates the individual-specific random effect for this paired study design.

We applied permutations to estimate empirical P-values for the interaction term. In each permutation step, the condition variable was shuffled within each individual, and the same linear mixed model was tested to get the permuted statistics for the interaction term. The permutation adjusted P-value for each interaction test was calculated as follows:

$$p = \frac{s + 1}{n + 1}$$

where n was the total number of permutations (1,000) and s was the number of cases where the permuted statistics were more significant than the original observed ones. We added 1 to both the numerator and the denominator to avoid underestimating permutation P-values⁷⁸. BH FDR-controlling procedure was applied to permutation adjusted P-values and significant interactions were identified at 5% FDR.

Trans-eQTL identification

To detect *trans*-acting genetic regulation of gene expression in each condition, we tested for associations between SNPs and genes that were located on different chromosomes using the same linear

model and covariates as in the *cis*-eQTL mapping. We tried the following different approaches to deal with the multiple testing:

(1) *Genome-wide FDR correction*. BH FDR-controlling procedure was applied to nominal P-values from all *trans*-association tests, and significant *trans*-associations were identified at 5% FDR⁶.

(2) *Gene-level FDR correction*. For each gene, P-value of the top SNP was multiplied by 1×10^6 , which was the estimated number of independent SNPs across the genome (calculated as 0.05 divided by the commonly used genome-wide significance threshold of 5×10^{-8}). To control gene-level FDR, BH FDR-controlling procedure was then applied to the minimum adjusted P-values for all genes.

(3) *Gene-level Bonferroni correction*. Bonferroni correction was used to control the gene-level FDR, by using a significance P-value threshold of 3.8×10^{-12} ($5 \times 10^{-8} / 13,109$, where the denominator indicates the number of genes). The Bonferroni correction was extremely conservative because the tests (or genes) were not independent with each other.

In resting monocytes as well as in LPS-stimulated monocytes, one *trans*-eQTL signal was significant in all three methods. At 5% genome-wide FDR level, we observed 10 and 15 eGenes with significant *trans*-eQTLs (*trans*-eGenes) in resting and PHA-stimulated T cells, respectively, corresponding to a nominal P-value threshold of 1.9×10^{-10} in resting T cells and 2.5×10^{-10} in PHA-treated T cells. The number of significant *trans*-eGenes dropped, respectively, to 6 and 8 by using the gene-level FDR correction (corresponding to a nominal P-value threshold of 5.3×10^{-12} in both conditions), and to 5 and 7 by using the gene-level Bonferroni correction. The limited power was the major issue given the sample size; thus we used genome-wide FDR correction, the least conservative method to determine significant *trans*-eQTLs used in the downstream analysis⁶.

Mediation analysis

We hypothesised that *trans*-eQTLs regulated the expression of distant genes through *cis*-mediators, or local genes whose expression was regulated by the same *trans*-eQTLs. To test this hypothesis, we focused on the *trans*-eQTLs that were also associated with adjacent *cis*-eGenes, meaning that the *trans*-eQTLs were also *cis*-eQTLs. For each *trans*-eGene–*cis*-eGene pair, we tested the *trans*-eSNP with the smallest P-value as the exposure, a *cis*-eGene as the mediator, and a *trans*-eGene as the outcome (**Figure 3B**). In total, we tested 14 mediation trios: 1 from resting monocytes, 1 from LPS-stimulated monocytes, 9 from resting T cells, and 3 from PHA-stimulated T cells.

We performed mediation test for the 14 trios using the *mediation* R package⁷⁹. The effect of the exposure on the mediator (α) was estimated in *cis*-eQTL mapping. The effect of the mediator on the

outcome (b) adjusting for the exposure and the effect of the exposure on the outcome (c') adjusting for the mediator were estimated in the following multiple regression:

$$y_i \sim x_i + x_i^{cis} + cov_i^1 + \dots + cov_i^{14}$$

where y_i indicates the value of the outcome (or the expression level of the *trans*-eGene) for the i^{th} sample, x_i the exposure (or the eSNP allele dosage), x_i^{cis} the mediator (or the *cis*-eGene expression), and $cov_i^1, \dots, cov_i^{14}$ the 14 covariates used in eQTL mapping. The estimates of b and c' were beta coefficients for x_i^{cis} and x_i , respectively. The “direct effect” of the exposure on the outcome was quantified as c' , the “indirect effect” of the exposure on the outcome through the mediator was quantified as $a \times b$, and the “total effect” was the sum of the previous two effects⁸⁰. Complete mediation occurs when the direct effect c' is zero after controlling for the mediator, and partial mediation happens when the direct effect is different from zero. To identify significant mediation trios (the null hypothesis $H_0: ab = 0$), we used a nonparametric bootstrap method (10,000 simulations) implemented in the *mediation* R package for variance estimation and P-value calculation. BH FDR-controlling procedure was applied to correct for multiple testing.

Genetic overlap of eQTLs and disease

We downloaded publicly-available GWAS data from the following three resources: ImmunoBase (URLs), LD Hub⁸¹ (URLs), and GWAS Catalog⁸² (URLs). If one disease was investigated in multiple studies, we focused on the most recent one, which usually had the largest sample size. We also downloaded summary statistics of GWAS that were carried out using the ImmunoChip array from the ImmunoBase resource; this array was designed for immunogenetics studies and captured more variants of immune-relevant genetic loci⁸³. We had summary statistics for the following immune-mediated diseases: allergic disease (asthma, hay fever, or eczema)⁸⁴, allergic rhinitis⁸⁵, allergic sensitisation⁸⁵, asthma⁸⁶, childhood-onset asthma⁴¹, adult-onset asthma⁴¹, inflammatory bowel disease (IBD) including its two subtypes – Crohn’s disease and ulcerative colitis⁸⁷, celiac disease^{88,89}, autoimmune thyroid disease⁹⁰, juvenile idiopathic arthritis⁹¹, multiple sclerosis⁹², narcolepsy⁹³, primary biliary cirrhosis (PBC)^{94,95}, primary sclerosing cholangitis (PSC)⁹⁶, psoriasis⁹⁷, rheumatoid arthritis^{98,99}, systemic lupus erythematosus (SLE)¹⁰⁰, and type 1 diabetes^{101,102}. These datasets contained summary statistics obtained using European populations for both significant and non-significant genetic variants, and GRCh37 genomic coordinates were available.

We performed enrichment tests for each of the four sets of significant eQTLs using *GARFIELD*²⁴. Generalised linear models were applied to test for enrichment in eQTLs of variants associated with the above diseases at a significance threshold of 1×10^{-6} . Bonferroni correction was applied to correct for multiple testing, where the number of tests was the number of GWAS datasets (24) multiplied by the

number of eQTL datasets (4), and the Bonferroni-adjusted P-value threshold was 5.2×10^{-4} , adjusting for the 4×24 (96) tests.

Colocalisation of *cis*-eQTLs with disease associations

We applied a Bayesian method implemented in the *coloc* v3.1 R package²⁶ to test whether any of the disease-associated GWAS loci shared the same causal variants with the *cis*-eQTLs. Full summary statistics were required to run the colocalisation analysis using *coloc*. For loci where *cis*-eQTLs were also associated with diseases at a P-value threshold of 1×10^{-6} , colocalisation test was performed on a 400-kb window centered on the top *cis*-eSNP. For each locus, colocalisation test was performed on overlapping SNPs where both eQTL and GWAS summary statistics were available. We excluded regions where not enough SNPs (<25) were available for colocalisation test. As shown in Guo *et al.*¹⁰³, selection of different prior probabilities of a SNP being causal for both of the traits had effects on the posterior support for colocalisation. To be conservative, we used a lower prior probability of 1×10^{-6} instead of the default value of 1×10^{-5} .

For each locus, the Bayesian method assessed the support for the following five exclusive hypotheses: no causal variants for either of the two traits (H_0), a causal variant for gene expression only (H_1), a causal variant for disease risk only (H_2), distinct causal variants for two traits (H_3), and the same shared causal variant for both traits (H_4). The package estimated posterior probabilities (PP_0 , PP_1 , PP_2 , PP_3 , PP_4) to summarise the evidence for the above five hypotheses. High PP_1 or PP_2 and low $PP_3 + PP_4$ indicate a lack of power to identify the causal signals²⁶. We excluded loci where $PP_3 + PP_4 < 0.8$, and focused on loci with strong evidence support for shared causal variants (H_4), i.e. ratio of PP_4 to $PP_3 \geq 5$.

Mendelian randomisation analysis

To investigate causal effects of eGene expression on the above immune-mediated diseases, we performed a two-sample Mendelian randomisation (MR) analysis. Summary statistics from both our eQTL and external GWAS studies were required, including beta coefficient and its standard error, effective allele (based on which the beta was estimated), the other allele, and P-value. For each disease trait, we used *cis*-eSNPs that were also included in the GWAS dataset, and removed ambiguous variants (if any) using the *TwoSampleMR* R package¹⁰⁴. We then selected LD pruned ($r^2 < 0.1$) *cis*-eSNPs as genetic instrumental variables (IVs). We focused on eGenes with at least 3 genetic IVs available, and performed the following MR methods implemented in the *MendelianRandomization* R package¹⁰⁵: inverse variance weighted (IVW), weighted median, weighted mode, and MR Egger. These methods have different assumptions for valid IVs: IVW assumes that all IVs are valid¹⁰⁶; weighted median assumes that valid IVs contribute to more than 50% of the weight¹⁰⁷; weighted mode assumes that the largest group of IVs are valid¹⁰⁸; MR Egger regression, which is the least sensitive, assumes that the

pleiotropic effects of IVs are not correlated with the genetic effects on exposure¹⁰⁹. We excluded the causal associations for which the intercept in the MR Egger method was significantly not equal to 0, indicating significant average pleiotropic effects. Gene expression was considered to have suggestive evidence of causal effects when at least 3 out of the 4 methods provided significant P-value (≤ 0.05). In the **Results** section, we reported the statistics of the weighted mode method, which has the least assumption among all methods except the MR Egger, but more sensitive than MR Egger.

URLs

DICE, <https://www.dice-database.org>

ImmunoBase, <https://www.immunobase.org/>

LD Hub, <http://ldsc.broadinstitute.org/ldhub/>

GWAS Catalog, <https://www.ebi.ac.uk/gwas/>

Funding

This research was supported by the NHMRC of Australia (project grant no. 1049539 to M.I. and K.E.H., Fellowships 1061409 to K.E.H., and 1061435 to M.I.). K.E.H. is supported by a Senior Medical Research Fellowship from the Viertel Foundation of Australia. This research was supported in part by the Victorian Government's Operational Infrastructure Support Program and the UK National Institute of Health Research.

References

1. Postma, D.S., Bush, A. & van den Berge, M. Risk factors and early origins of chronic obstructive pulmonary disease. *Lancet* **385**, 899-909 (2015).
2. Barker, D.J. The origins of the developmental origins theory. *J Intern Med* **261**, 412-7 (2007).
3. Gensollen, T., Iyer, S.S., Kasper, D.L. & Blumberg, R.S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539-44 (2016).
4. Reynolds, L.A. & Finlay, B.B. Early life factors that affect allergy development. *Nat Rev Immunol* **17**, 518-528 (2017).
5. Carraro, S., Scheltema, N., Bont, L. & Baraldi, E. Early-life origins of chronic respiratory diseases: understanding and promoting healthy ageing. *Eur Respir J* **44**, 1682-96 (2014).
6. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
7. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e16 (2018).
8. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895 (2010).
9. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
10. Peng, S. *et al.* Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases. *Hum Mol Genet* **26**, 3432-3441 (2017).
11. O'Brien, H.E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol* **19**, 194 (2018).
12. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
13. Kim-Hellmuth, S. *et al.* Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun* **8**, 266 (2017).
14. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet* **50**, 424-431 (2018).
15. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
16. Barreiro, L.B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A* **109**, 1204-9 (2012).
17. Ye, C.J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
18. Kusel, M.M., de Klerk, N., Holt, P.G. & Sly, P.D. Antibiotic use in the first year of life and risk of atopic disease in early childhood. *Clin Exp Allergy* **38**, 1921-8 (2008).
19. Kusel, M.M. *et al.* Role of respiratory viruses in acute upper and lower respiratory tract illness in the first year of life: a birth cohort study. *Pediatr Infect Dis J* **25**, 680-6 (2006).
20. Kusel, M.M. *et al.* Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. *J Allergy Clin Immunol* **119**, 1105-10 (2007).
21. Kusel, M.M., Keadze, T., Johnston, S.L., Holt, P.G. & Sly, P.D. Febrile respiratory illnesses in infancy and atopy are risk factors for persistent asthma and wheeze. *Eur Respir J* **39**, 876-82 (2012).

22. Teo, S.M. *et al.* The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* **17**, 704-15 (2015).
23. Teo, S.M. *et al.* Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. *Cell Host Microbe* **24**, 341-352 e5 (2018).
24. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat Genet* **51**, 343-353 (2019).
25. Bryois, J. *et al.* Cis and trans effects of human genomic variants on gene expression. *PLoS Genet* **10**, e1004461 (2014).
26. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
27. Roychoudhuri, R. *et al.* BACH2 regulates CD8(+) T cell differentiation by controlling access of AP-1 factors to enhancers. *Nat Immunol* **17**, 851-860 (2016).
28. Shinnakasu, R. *et al.* Regulated selection of germinal-center cells into the memory B cell compartment. *Nat Immunol* **17**, 861-9 (2016).
29. Afzali, B. *et al.* BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency. *Nat Immunol* **18**, 813-823 (2017).
30. Roychoudhuri, R. *et al.* BACH2 represses effector programs to stabilize T(reg)-mediated immune homeostasis. *Nature* **498**, 506-10 (2013).
31. Schutyser, E., Struyf, S. & Van Damme, J. The CC chemokine CCL20 and its receptor CCR6. *Cytokine Growth Factor Rev* **14**, 409-26 (2003).
32. Merk, M., Mitchell, R.A., Endres, S. & Bucala, R. D-dopachrome tautomerase (D-DT or MIF-2): doubling the MIF cytokine family. *Cytokine* **59**, 10-7 (2012).
33. Merk, M. *et al.* The D-dopachrome tautomerase (DDT) gene product is a cytokine and functional homolog of macrophage migration inhibitory factor (MIF). *Proc Natl Acad Sci U S A* **108**, E577-85 (2011).
34. Peters, J.E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS Genet* **12**, e1005908 (2016).
35. Pierce, B.L. *et al.* Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet* **10**, e1004818 (2014).
36. Yang, F., Wang, J., Consortium, G.T., Pierce, B.L. & Chen, L.S. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res* **27**, 1859-1871 (2017).
37. Belyy, A., Levanova, N., Tabakova, I., Rospert, S. & Belyi, Y. Ribosomal Protein Rps26 Influences 80S Ribosome Assembly in *Saccharomyces cerevisiae*. *mSphere* **1**(2016).
38. Ferretti, M.B., Ghalei, H., Ward, E.A., Potts, E.L. & Karbstein, K. Rps26 directs mRNA-specific translation by recognition of Kozak sequence elements. *Nat Struct Mol Biol* **24**, 700-707 (2017).
39. Min, E.E., Roy, B., Amrani, N., He, F. & Jacobson, A. Yeast Upf1 CH domain interacts with Rps26 of the 40S ribosomal subunit. *RNA* **19**, 1105-15 (2013).
40. Cui, D. *et al.* The ribosomal protein S26 regulates p53 activity in response to DNA damage. *Oncogene* **33**, 2225-35 (2014).
41. Pividori, M., Schoettler, N., Nicolae, D.L., Ober, C. & Im, H.K. Shared and distinct genetic risk factors for childhood onset and adult onset asthma. *bioRxiv*, 427427 (2018).

42. Demedts, I.K. *et al.* Accumulation of dendritic cells and increased CCL20 levels in the airways of patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* **175**, 998-1005 (2007).
43. Faiz, A. *et al.* Profiling of healthy and asthmatic airway smooth muscle cells following interleukin-1beta treatment: a novel role for CCL20 in chronic mucus hypersecretion. *Eur Respir J* **52**(2018).
44. Pichavant, M. *et al.* Asthmatic bronchial epithelium activated by the proteolytic allergen Der p 1 increases selective dendritic cell recruitment. *J Allergy Clin Immunol* **115**, 771-8 (2005).
45. Kim, S., Lewis, C. & Nadel, J.A. CCL20/CCR6 feedback exaggerates epidermal growth factor receptor-dependent MUC5AC mucin production in human airway epithelial (NCI-H292) cells. *J Immunol* **186**, 3392-400 (2011).
46. Francis, J.N., Sabroe, I., Lloyd, C.M., Durham, S.R. & Till, S.J. Elevated CCR6+ CD4+ T lymphocytes in tissue compared with blood and induction of CCL20 during the asthmatic late response. *Clin Exp Immunol* **152**, 440-7 (2008).
47. Blomgren, H. & Svedmyr, E. In vitro stimulation of mouse thymus cells by PHA and allogeneic cells. *Cell Immunol* **2**, 285-99 (1971).
48. de Waal Malefyt, R. *et al.* Differential regulation of IL-13 and IL-4 production by human CD8+ and CD4+ Th0, Th1 and Th2 T cell clones and EBV-transformed B cells. *Int Immunol* **7**, 1405-16 (1995).
49. Punnonen, J. *et al.* Interleukin 13 induces interleukin 4-independent IgG4 and IgE synthesis and CD23 expression by human B cells. *Proc Natl Acad Sci U S A* **90**, 3730-4 (1993).
50. Wills-Karp, M. *et al.* Interleukin-13: central mediator of allergic asthma. *Science* **282**, 2258-61 (1998).
51. Berry, M.A. *et al.* Sputum and bronchial submucosal IL-13 expression in asthma and eosinophilic bronchitis. *J Allergy Clin Immunol* **114**, 1106-9 (2004).
52. Saha, S.K. *et al.* Increased sputum and bronchial biopsy IL-13 expression in severe asthma. *J Allergy Clin Immunol* **121**, 685-91 (2008).
53. Tsiligianni, Z. *et al.* Sputum interleukin-13 as a biomarker for the evaluation of asthma control. *Clin Exp Allergy* **46**, 1498 (2016).
54. Hanania, N.A. *et al.* Efficacy and safety of lebrikizumab in patients with uncontrolled asthma (LAVOLTA I and LAVOLTA II): replicate, phase 3, randomised, double-blind, placebo-controlled trials. *Lancet Respir Med* **4**, 781-796 (2016).
55. Panettieri, R.A., Jr. *et al.* Tralokinumab for severe, uncontrolled asthma (STRATOS 1 and STRATOS 2): two randomised, double-blind, placebo-controlled, phase 3 clinical trials. *Lancet Respir Med* **6**, 511-525 (2018).
56. Guo, Y. & Wang, A.Y. Novel Immune Check-Point Regulators in Tolerance Maintenance. *Front Immunol* **6**, 421 (2015).
57. Le Page, C. *et al.* BTN3A2 expression in epithelial ovarian cancer is associated with higher tumor infiltrating T cells and a better prognosis. *PLoS One* **7**, e38541 (2012).
58. Vavassori, S. *et al.* Butyrophilin 3A1 binds phosphorylated antigens and stimulates human gammadelta T cells. *Nat Immunol* **14**, 908-16 (2013).
59. Wang, H. *et al.* Butyrophilin 3A1 plays an essential role in prenyl pyrophosphate stimulation of human Vgamma2Vdelta2 T cells. *J Immunol* **191**, 1029-42 (2013).
60. Vantourout, P. *et al.* Heteromeric interactions regulate butyrophilin (BTN) and BTN-like molecules governing gammadelta T cell biology. *Proc Natl Acad Sci U S A* **115**, 1039-1044 (2018).

61. Lamontagne, M. *et al.* Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Hum Mol Genet* **27**, 1819-1829 (2018).
62. Nair, R.P. *et al.* Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. *Am J Hum Genet* **78**, 827-851 (2006).
63. Genetic Analysis of Psoriasis, C. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet* **42**, 985-90 (2010).
64. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
65. Arloth, J., Bader, D.M., Roh, S. & Altmann, A. Re-Annotator: Annotation Pipeline for Microarray Probe Sequences. *PLoS One* **10**, e0139516 (2015).
66. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).
67. Miller, J.A. *et al.* Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* **12**, 322 (2011).
68. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
69. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
70. Huang, Q.Q., Ritchie, S.C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res* **46**, e133 (2018).
71. Shabalín, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
72. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778 (2017).
73. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-7 (2012).
74. Davis, J.R. *et al.* An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am J Hum Genet* **98**, 216-24 (2016).
75. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289-300 (1995).
76. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat Commun* **8**, 15452 (2017).
77. Bates, D., Machler, M., Bolker, B.M. & Walker, S.C. Fitting linear mixed-effects models using lme4. *J Stat Softw* **67**, 1-48 (2015).
78. Phipson, B. & Smyth, G.K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* **9**, Article39 (2010).
79. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R Package for Causal Mediation Analysis. *J Stat Softw* **59**, 1-38 (2014).
80. Hayes, A.F. & Scharkow, M. The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: does method really matter? *Psychol Sci* **24**, 1918-27 (2013).
81. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).

82. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
83. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13**, 101 (2011).
84. Ferreira, M.A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet* **49**, 1752-1757 (2017).
85. Waage, J. *et al.* Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet* **50**, 1072-1080 (2018).
86. 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).
87. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
88. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295-302 (2010).
89. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
90. Cooper, J.D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum Mol Genet* **21**, 5202-8 (2012).
91. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet* **45**, 664-9 (2013).
92. International Multiple Sclerosis Genetics, C. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45**, 1353-60 (2013).
93. Faraco, J. *et al.* ImmunoChip study implicates antigen presentation to T cells in narcolepsy. *PLoS Genet* **9**, e1003270 (2013).
94. Cordell, H.J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat Commun* **6**, 8019 (2015).
95. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **44**, 1137-41 (2012).
96. Ji, S.G. *et al.* Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet* **49**, 269-273 (2017).
97. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44**, 1341-8 (2012).
98. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
99. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* **44**, 1336-40 (2012).
100. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet* **47**, 1457-1464 (2015).
101. Bradfield, J.P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet* **7**, e1002293 (2011).
102. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* **47**, 381-6 (2015).

103. Guo, H. *et al.* Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum Mol Genet* **24**, 3305-13 (2015).
104. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).
105. Yavorska, O.O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46**, 1734-1739 (2017).
106. Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-65 (2013).
107. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304-14 (2016).
108. Hartwig, F.P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985-1998 (2017).
109. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-25 (2015).

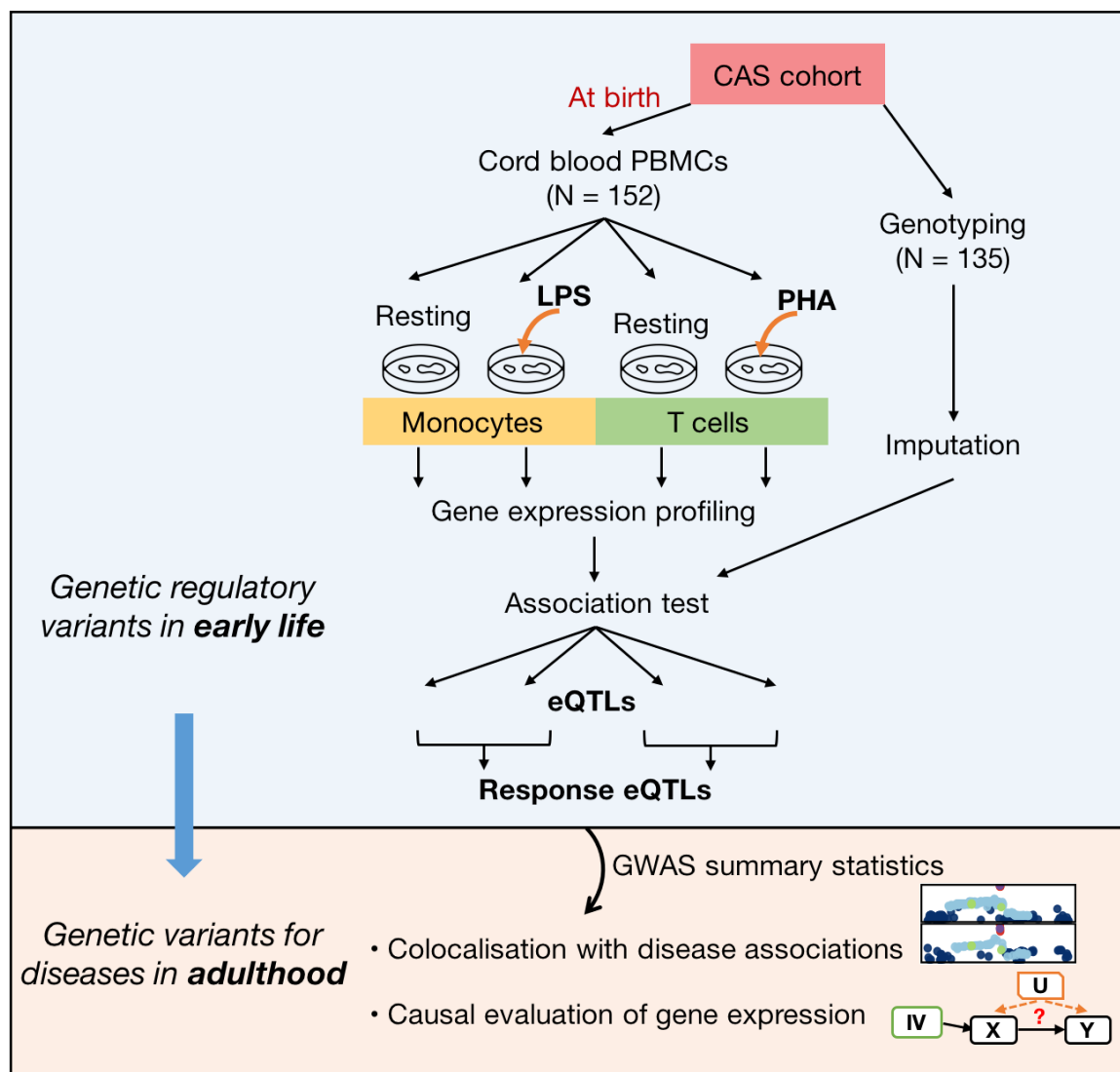


Figure 1: Study design and analysis work flow. Monocytes and T cell cultures were purified from resting and stimulated cord blood samples from the Childhood Asthma Study (CAS) cohort. Gene expression was quantified using a microarray platform. Genotype data are available for a subset of the CAS individuals. eQTLs were identified within each experimental condition. Datasets for resting and stimulated samples were merged to detect response eQTLs within each cell type. Next, we identified genetic loci where neonatal eQTLs and disease associations obtained from external GWAS datasets shared the same causal variants. We investigated the causal effects of gene expression at birth on immune diseases that develop later in life.

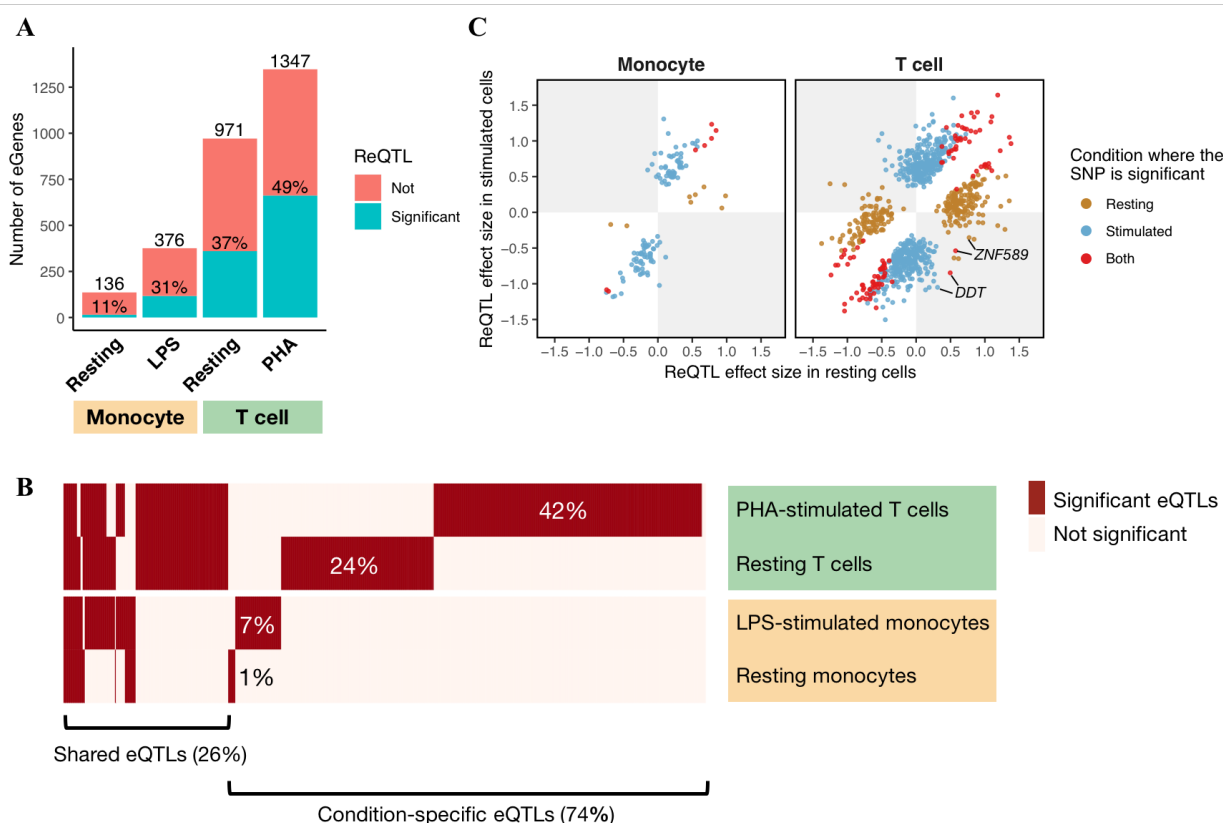


Figure 2: Cis-eQTLs and response eQTLs (reQTLs) in monocytes and T cells. (A) A bar plot shows the number of genes with significant *cis*-eQTLs (eGenes) identified in each cell type and treatment group (on x-axis). Percentages on each bar indicate the proportion of eGenes with significant reQTLs (green). (B) A heatmap shows eQTL sharing across four experimental conditions (rows). Columns in the heatmap are unique eQTL–eGene associations. Significant associations are in red. Percentages labelled on the heatmap show the proportion of unique eQTL–eGene associations that are specific to a certain cell type and stimulatory condition. (C) Two point plots show effect sizes (s.d. change in gene expression per allele) of significant reQTLs in resting (x-axes) and stimulated conditions (y-axes) in two cell populations: monocytes (left) and T cells (right). A gene might have two dots indicating two independent top SNPs (**Methods**). Colours indicate the condition in which the SNP was significant. ReQTLs of *DDT* and *ZNF585* in the grey quadrants (red dots) show opposite directions of eQTL effects across conditions.

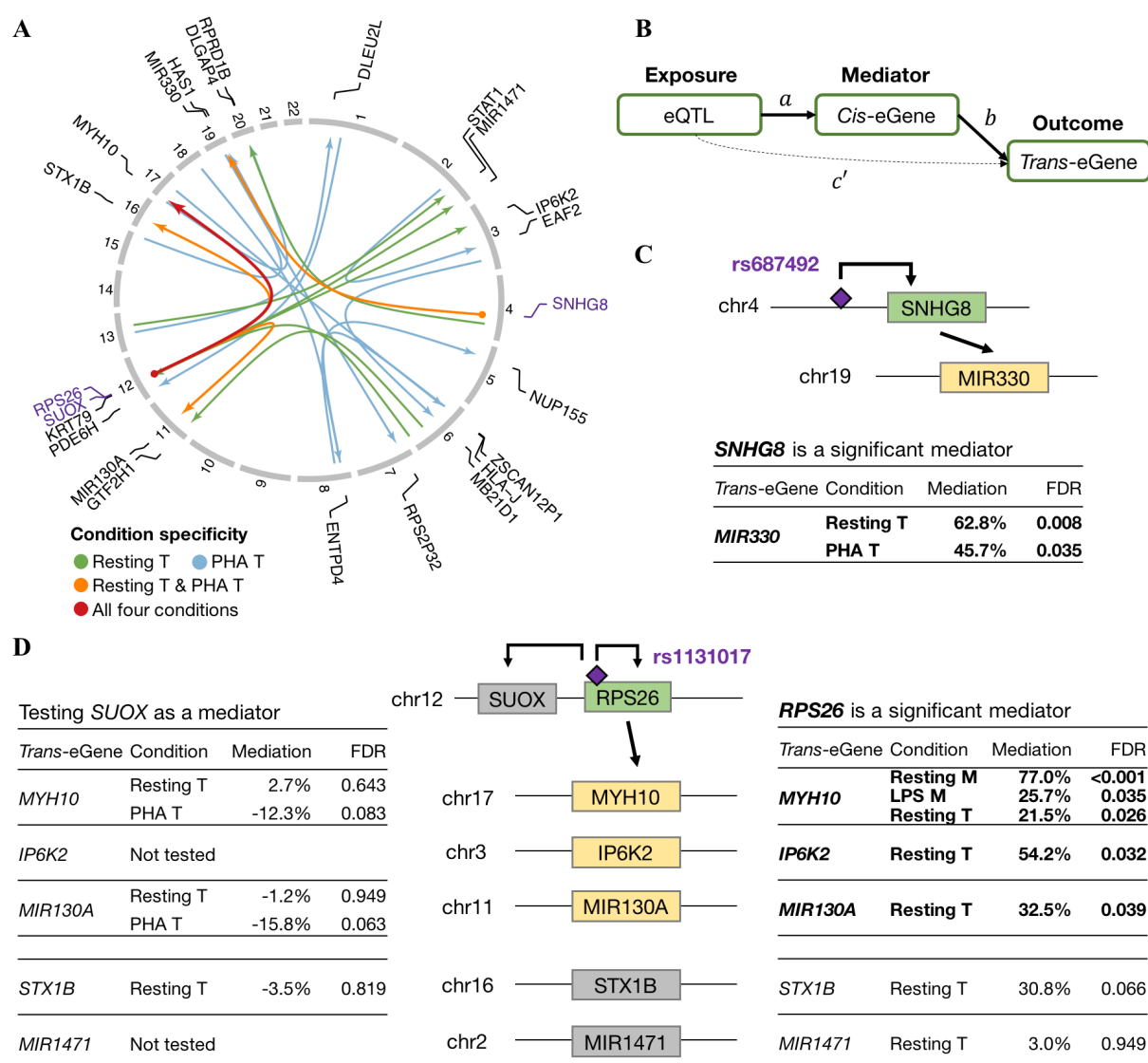


Figure 3: *Trans*-eQTL effects and their *cis*-mediators. (A) A circular plot shows *trans*-eQTL associations in lines, with arrows pointing to *trans*-eGenes with names annotated (in black) outside the rim indicating chromosome numbers. Dots on the other end point indicate nearby genes (names in purple) that are associated with the same loci (*cis*-eQTLs). Colours of the lines indicate the experimental conditions where the *trans*-eQTLs were identified: “Resting T”: resting T cells only, “PHA T”: stimulated T cells only, “Resting T & PHA T”: shared between both conditions of T cells, and “All four conditions”: shared across all four experimental conditions. (B) A diagram demonstrates the mediation analysis model, where effects of *trans*-acting eQTL (exposure) on *trans*-eGene (outcome) are either mediated through a *cis*-eGene (mediator), or through direct effects (Methods). (C) and (D) show two examples of *cis*-eGenes (green), *SNHG8* and *RPS26*, acting as mediators for *trans*-effects (*trans*-eGenes in yellow). Genes that were not significant in mediation analysis are in grey. Tables show statistics of the mediation tests, and the “Mediation” column indicates the proportion of total effects of the eQTL on the *trans*-eGene that was mediated through the *cis*-eGene. Two models involving *SUOX* (D) were not tested because the *trans*-eSNPs of *IP6K2* and *MIR1471* were not significantly associated with *SUOX* (Table S9). Significant mediations (FDR ≤ 0.05) are highlighted in bold.

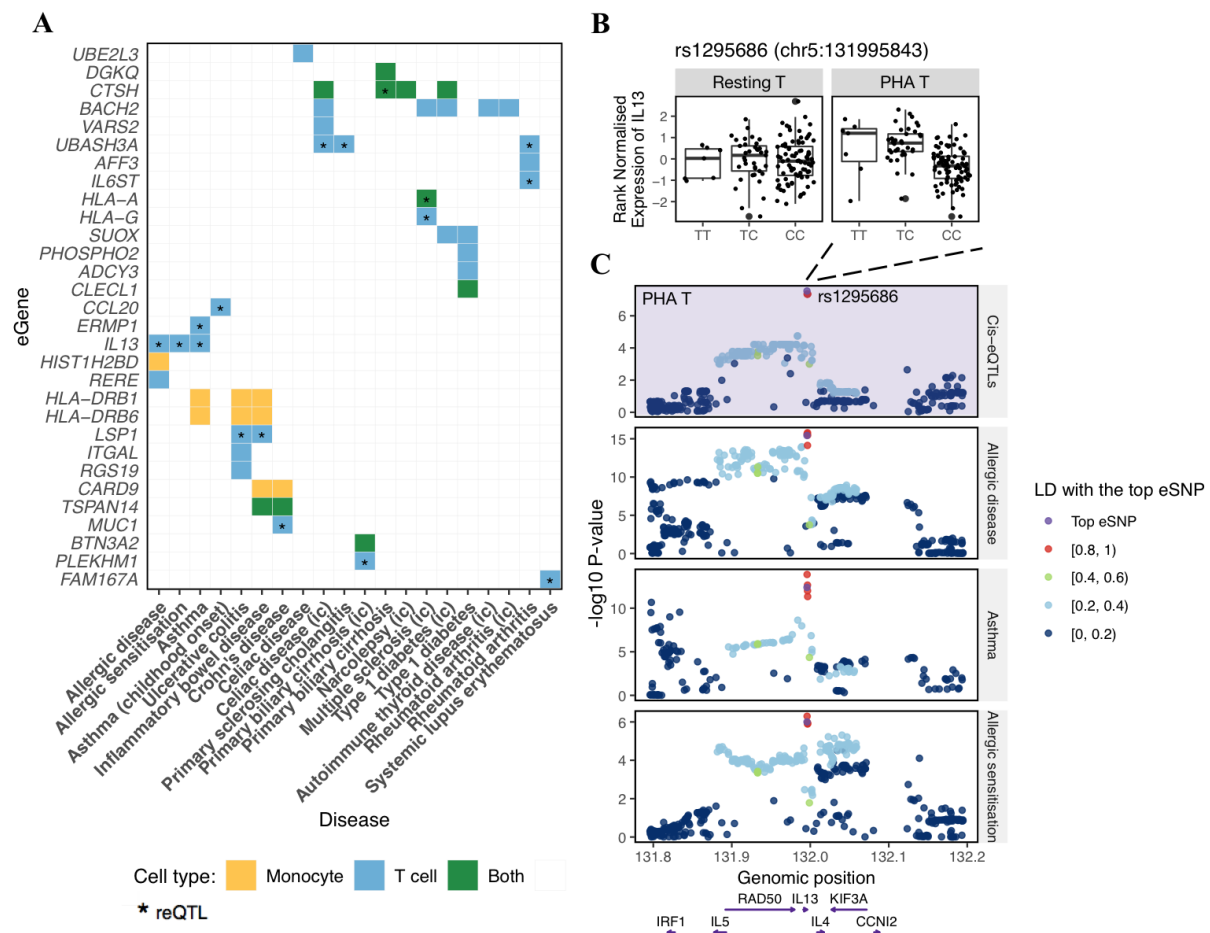


Figure 4: Colocalisation of *cis*-eQTLs with disease associations. (A) A heatmap shows all cases with strong evidence of colocalisation between *cis*-eQTLs of corresponding genes (eGenes) in rows and GWAS hits associated with allergic and autoimmune diseases in columns (“ic” indicates that the study was performed using ImmunoChip array). Colours indicate the cell type where the significant colocalisation was observed. Asterisks indicate that the colocalised eQTLs are response eQTLs (reQTLs). (B) Box plots show the rank-normalised gene expression of *IL13* (y-axes) in resting T cells (left) and in PHA-stimulated T cells (right) stratified by genotypes of the reQTL rs1295686 (x-axes), the top eSNP in PHA-stimulated T cells. In resting T cells, no SNP was significantly associated with *IL13*. (C) Regional plots show eQTL association with gene expression of *IL13* in PHA-stimulated T cells (purple background), and GWAS associations with allergic disease (asthma, hay fever, or eczema), asthma, and allergic sensitisation. The minus log₁₀ P-value is plotted on y-axes for all SNPs located within 200 kb from the top eSNP of *IL13*. Colours of dots indicate the LD correlation with the top eSNP (in purple). Positions of genes located on this locus are shown at the bottom.

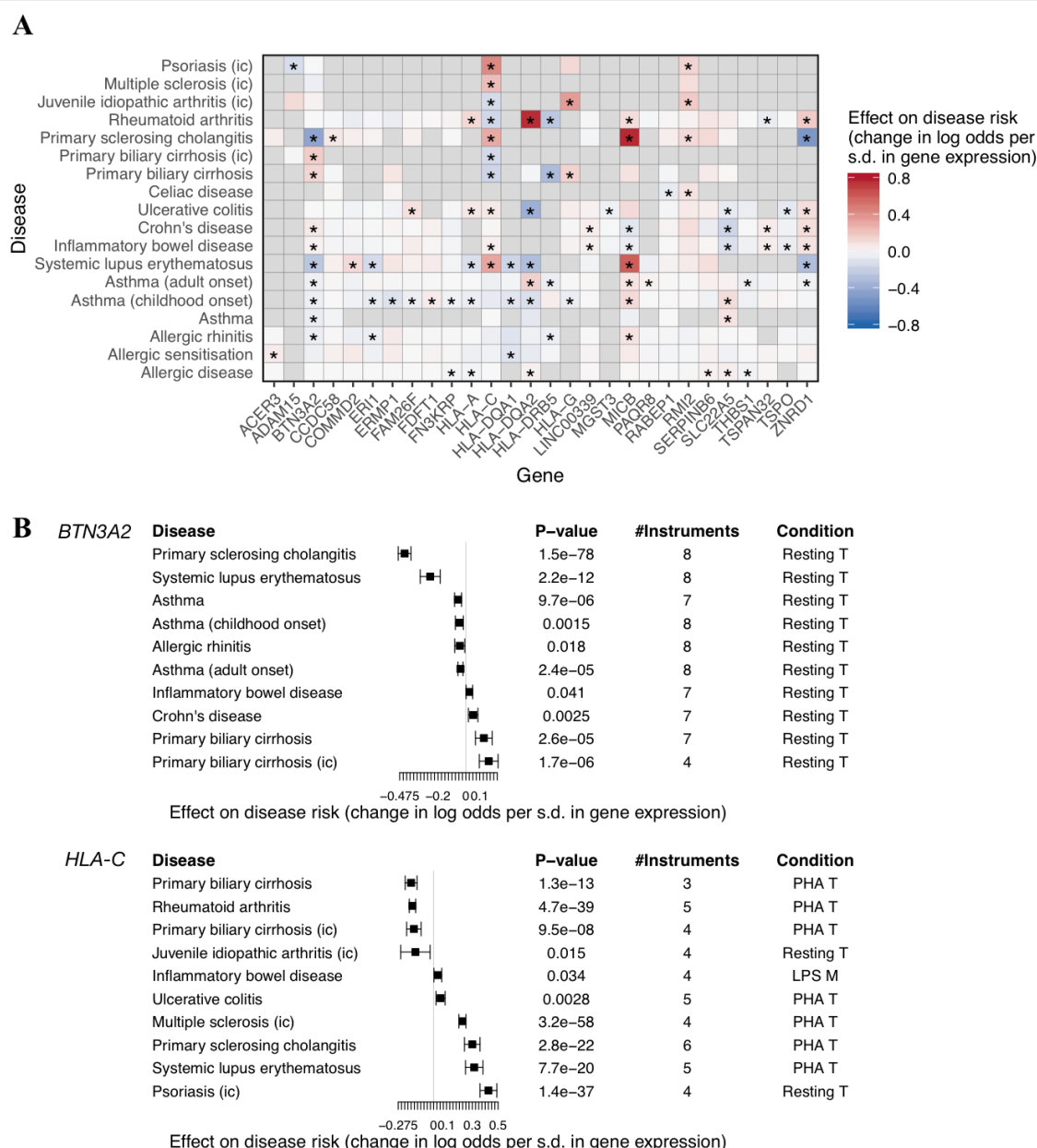


Figure 5: Causal effects of neonatal gene expression on multiple immune-related diseases. (A) The heatmap shows the causal effect size estimated using the weighted mode method in the Mendelian randomisation (MR) analysis. Asterisks indicate significant causal associations (**Methods**). Causal associations with significant pleiotropy was excluded. If a gene was tested using the expression levels in multiple experimental conditions, the one with the highest number of genetic instruments was kept. Statistics of all MR tests are in **Table S12**. Grey indicates the gene-disease pairs that were not tested due to small number of genetic instruments (<3). Positive effect estimates in red indicate that increased gene expression is causally associated with increased disease risk, and negative causal associations are in blue. (B) Forest plots show causal estimates and 95% confidence intervals for neonatal expression of *BTN3A2* and *HLA-C*.