Genetic regulation of serum IgA levels and susceptibility to common immune, infectious, kidney, and cardio-metabolic traits

Lili Liu¹, Atlas Khan¹, Elena Sanchez-Rodriguez¹, Francesca Zanoni¹, Yifu Li¹, Nicholas Steers¹, Olivia Balderes¹, Junying Zhang¹, Priya Krithivasan¹, Robert A. LeDesma³, Clara Fischman⁴, Scott J. Hebring⁵, John B. Harley^{6,7}, Halima Moncrieffe⁶, Leah C. Kottyan⁶, Bahram Namjou-Khales⁶, Theresa L. Walunas⁸, Rachel Knevel⁹, Soumya Raychaudhuri⁹, Elizabeth W. Karlson⁹, Joshua C. Denny¹⁰, Ian B. Stanaway¹¹, David Crosslin¹², Thomas Rauen¹³, Jürgen Floege¹³, Frank Eitner¹⁴, Zina Moldoveanu², Colin Reily², Barbora Knoppova², Stacy Hall², Justin T. Sheff², Bruce A. Julian², Robert J. Wyatt¹⁵, Hitoshi Suzuki¹⁶, Jingyuan Xie¹⁷, Nan Chen¹⁷, Xujie Zhou¹⁸, Hong Zhang¹⁸, Lennart Hammarström¹⁹, Alexander Viktorin²⁰, Patrik K. E. Magnusson²⁰, Ning Shang²¹, George Hripcsak²¹, Chunhua Weng²¹, Tatjana Rundek²², Mitchell S. V. Elkind²³, Elizabeth C. Oelsner¹, R. Graham Barr²⁴, Iuliana Ionita-Laza²⁵, Jan Novak², Ali G. Gharavi¹ and Krzysztof Kiryluk¹

¹Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA.

²Department of Microbiology and Medicine, University of Alabama at Birmingham, Birmingham, AL, USA.

³Lewis Thomas Laboratory, Department of Molecular Biology, Princeton University, Princeton, NJ, USA.

⁴Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

⁵Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA.

⁶Center of Autoimmune Genomics and Etiology, Cincinnati Children's Hospital, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA.

⁷US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA.

⁸Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

⁹Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

¹⁰Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA.

¹¹Kidney Research Institute, Division of Nephrology, Department of Medicine, University of Washington, Seattle, WA, USA.

¹²Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, USA.

¹³Department of Nephrology, RWTH University of Aachen, Aachen, Germany.

¹⁴Department of Nephrology, RWTH University of Aachen, Aachen, Germany and Kidney Diseases Research, Bayer Pharma AG, Wuppertal, Germany.

¹⁵Division of Pediatric Nephrology, University of Tennessee Health Sciences Center, Memphis, TN, USA.

¹⁶Department of Nephrology, Juntendo University Faculty of Medicine, Tokyo, Japan.

¹⁷Department of Nephrology, Institute of Nephrology, Shanghai Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

¹⁸Renal Division, Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China.

¹⁹Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden.

²⁰Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

²¹Department of Biomedical Informatics, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA.

²²Department of Neurology and Evelyn F. McKnight Brain Institute, University of Miami, Miami, FL, USA.

²³Department of Neurology, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA.

²⁴Department of Medicine, Vagelos College of Physicians & Surgeons and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA.

²⁵Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA.

Correspondence:

Krzysztof Kiryluk, M.D., M.S.

Associate Professor of Medicine, Department of Medicine, Division of Nephrology

College of Physicians & Surgeons, Columbia University

1150 St Nicholas Ave, Russ Berrie Pavilion #412E, New York, NY 10032

Tel: 212-851-4926

Email: kk473@columbia.edu

Abstract

Immunoglobulin A (IgA) mediates mucosal responses to food antigens and the intestinal microbiome and has a known role in susceptibility to mucosal pathogens, celiac disease, inflammatory bowel disease, and IgA nephropathy. We performed genetic analyses of serum IgA levels in 41,263 individuals of diverse ancestries. We observed unexpected variability in IgA levels across major ancestral populations, with African ancestry being reproducibly associated with higher serum IgA levels compared to other ancestries. The trans-ethnic GWAS analysis identified 20 genome-wide significant loci associated with serum IgA levels, including nine known and 11 novel loci. Systematic co-localization analysis with blood and primary immune cell expression QTLs prioritized candidate genes for 14 of 20 loci. Most GWAS loci encoded genes that produce immune defects and IgA abnormalities when genetically manipulated in mice. We uncovered positive genetic correlations of serum IgA levels with IgA nephropathy, type 2 diabetes and body mass index, as well as negative genetic correlations with celiac disease, inflammatory bowel disease, several infections, and intestinal microbiome diversity. Our findings provide novel insights into the genetic regulation of IgA production and its potential role in human disease.

Introduction

Immunoglobulin A (IgA) provides protection against mucosal infections and contributes to the pathogenesis of autoimmune and inflammatory disorders^{1,2}. Most of the IgA production occurs at the mucosal surfaces along the gastrointestinal and respiratory tracts, but a large portion of circulating IgA is contributed by bone-marrow plasma cells³. IgA neutralizes mucosal pathogens⁴ and enhanced IgA responsiveness has been reported in various respiratory and gastrointestinal infections, including acute SARS-CoV-2 infection⁵⁻⁷. Increased serum IgA levels are a common phenomenon in patients with IgA nephropathy^{8,9}, diabetes¹⁰ and metabolic syndrome¹¹. The serum concentration of IgA can be influenced by a combination of inherited factors and environmental exposures, including age, sex, and lifestyle factors¹¹⁻¹⁴. The heritability of serum IgA levels has been estimated in the range 20-60%¹⁵⁻¹⁸. Several GWAS have investigated genetic determinants of serum IgA levels in individuals from either European or East Asian ancestry, and nine significant GWAS loci have been identified to date^{16,19,20}. Notably, African and other more diverse populations have not been included in prior studies, and limited data exists on the ancestral differences in IgA levels.

In this study, we conducted a multiethnic GWAS meta-analysis for serum IgA levels in 41,263 individuals across 17 international cohorts of diverse ancestries, maximizing power for genetic discovery. We identified novel genetic determinants of serum IgA levels and, through comprehensive functional annotations, we prioritized candidate causal genes at each of the IgA-associated loci. We then investigated the shared genetic architecture between serum IgA levels and other human traits using several approaches, including genome-wide genetic correlation analysis, co-localization of GWAS loci, Mendelian randomization, and phenome-wide association studies. Our results provide new insights into the genetic regulation of serum IgA levels and its role in genetic susceptibility to several human diseases, including immune, infectious, kidney and cardio-metabolic traits.

Results

Ancestral differences in serum levels of IgA. We tested for differences in the distribution of serum IgA levels among 12,260 multiethnic individuals broadly classified into 4 major groups based on genetic ancestry (1,751 African American, 6,791 European, 1,257 East Asian, and 2,461 Latinx or admixed ancestry individuals). First, we performed laboratory measurements of serum IgA levels in all 5,420 participants of the Multi-Ethnic Study of Atherosclerosis (MESA) with available serum, providing us with the largest multiethnic cohort with standardized IgA measurements. For group comparisons, we generated standardized residuals of log-transformed serum IgA levels after adjustment for age and sex. Notably, MESA participants of African ancestry had significantly higher mean age and sex-adjusted serum IgA levels compared to all other ancestries (**Figure 1a**). We next examined the distribution of adjusted IgA levels in the pooled dataset of 6,840 diverse non-MESA participants included in our genetic studies. In this independent dataset, we replicated the strong association of genetic African ancestry with higher IgA levels after age and sex adjustment (**Figure 1b**). The admixture analysis across the MESA cohort confirmed weak, but highly statistically significant positive correlation between African ancestry and age- and sex-adjusted serum IgA levels ($r=0.026$, $P=4.6 \times 10^{-33}$, **Figure 1c**). This correlation remained significant after additional adjustment for body mass index (BMI) and diabetes.

Multiethnic GWAS meta-analysis. Next, we aimed to identify genetic loci controlling serum IgA levels using GWAS. We performed a meta-analysis of 17 diverse ancestry-defined cohorts comprised of 41,263 individuals with serum IgA measurements (35,094 European, 1,751 African American, 1,957 East Asian, and 2,461 Latinx or admixed-ancestry individuals). These cohorts were not ascertained based upon any specific immune or disease phenotype. Within each of the 17 cohorts, IgA phenotypes were defined as standardized residuals of log-transformed IgA levels regressed against age and sex. Each cohort was genotyped with high-density SNP arrays and imputed using the latest genome sequence reference panels (**Table 1 and Online Methods**).

The results of fixed effects meta-analysis are provided in **Figure 2a and Table 2** with regional association plots shown in **Supplementary Figure 1**. We observed minimal genomic inflation of the final meta-analysis summary statistics ($\lambda=1.016$, **Supplementary Figure 2**), confirming negligible effects of population stratification. In total, we identified 20 genome-wide significant independent loci, including nine known and 11 novel loci based on $P < 5 \times 10^{-8}$. Stepwise conditional analyses of the genome-wide significant loci revealed that

five of the 20 loci had at least two independently genome-wide significant variants underlying their association signal (**Table 2 and Supplementary Table 1**). The conditionally independent genome-wide significant SNPs jointly explained approximately 2% of the overall phenotypic variance. In addition, we identified eight suggestive signals with $P < 1 \times 10^{-6}$ (**Supplementary Table 2**), including the *TNFSF13* locus previously reported in a GWAS performed in a Han Chinese cohort¹⁹. Using linkage disequilibrium (LD)-score regression method²¹, we estimated the genome-wide SNP-based heritability of age- and sex-adjusted IgA levels at approximately 7% (95%CI: 2%-11%).

As expected, the effect sizes of independently associated variants were inversely related to their minor allelic frequencies (**Figure 2b**). Two relatively rare genome-wide significant variants exhibited the largest effects, including the previously reported *RUNX3* locus (rs188468174, MAF<1%, Beta=0.88, $P=3.42 \times 10^{-92}$) and the novel *GPATCH2* locus (rs73100295, MAF 1-2%, Beta=0.36, $P=3.91 \times 10^{-8}$). Interestingly, the IgA-increasing alleles at 15 of the 20 loci (75%) were more frequent in African compared to other ancestries, with 10 of 15 (67%) representing ancestral alleles. These findings were consistent with the results of our admixture analysis demonstrating positive correlation between the fraction of African ancestry and IgA levels, leading us to hypothesize that the IgA-increasing alleles could have provided some degree of fitness advantage in Africa, or decreased fitness in non-African environment. At the same time, we detected no clear signature of polygenic adaptation, as there was no significant correlation between African-European risk allele frequency difference and effect size, or risk variant singleton density score (SDS)²² and effect size for the genome-wide significant loci.

Pathway and tissue enrichments. We performed pathway enrichment analyses based on genome-wide summary statistics²³ and identified five significantly enriched pathways at Bonferroni-adjusted $P < 0.05$ (**Figure 4a and Supplementary Table 3**), including cytokine signaling in immune system ($P=1.2 \times 10^{-6}$), signaling by interleukins ($P=1.1 \times 10^{-5}$), cytokine-cytokine receptor interactions ($P=5.2 \times 10^{-6}$), TNFs and their physiological receptors ($P=1.4 \times 10^{-5}$), and IL-6-type cytokine receptor ligand interactions ($P=5.0 \times 10^{-5}$). Using data-driven expression-prioritized integration for complex traits (DEPICT) analysis²⁴, we further prioritized 17 tissues and cell types at FDR<0.05 based on our GWAS results, with the strongest enrichment in bone marrow cells, hematopoietic system, as well as blood and myeloid cells (**Figure 3b and Supplementary Table 4**).

Pleiotropic associations based on GWAS catalogue. We next interrogated the pleiotropic effect of individual loci by annotating lead SNPs and their proxies against the GWAS catalogue database (see **Online Methods**)²⁵. Most of the genome-wide significant loci had previous GWAS associations with immune-mediated disorders, infections, or hematological traits (**Figure 2c and Supplementary Table 5**). In particular, eight non-HLA loci, *SH2B3*, *ANKRD55*, *HDAC7*, *RCOR1/TRAFF3*, *TNFSF4*, *POU2AF1*, *FADS2* and *OVOL1/RELA*, displayed either concordant or opposed effects on 18 different autoimmune and inflammatory disorders, suggesting that genetic regulation of IgA levels may play a pervasive role in the control of autoimmunity and inflammation. The *SH2B3* locus displayed the highest degree of pleiotropy, being associated with 79 different GWAS traits. We also found that the alleles associated with higher serum IgA level at both *SH2B3* and *HORMAD2/LIF* loci were associated with increased risk of tonsillectomy, a procedure frequently performed in the setting of recurrent pharyngitis²⁶. Moreover, concordant effects on high blood pressure were found at three loci, including *SH2B3*, *CTF1* and *HDAC7*, consistent with the epidemiologic association of high blood pressure with increased IgA levels¹¹. Interestingly, the *SH2B3* locus showed concordant risk effects on several cardiovascular traits, including coronary artery disease and hypertension, but had an opposed effect on LDL cholesterol. The effect on BMI was concordant with IgA levels for *SH2B3* and *RCOR1/TRAFF3* loci, also consistent with the reported correlation between higher serum IgA levels and obesity¹¹.

Functional Annotations of GWAS loci. The majority of lead SNPs at genome-wide significant loci map to non-coding (intronic or intergenic) regions (**Supplementary Table 6**). The lead SNPs at five loci had proxies in the 5' or 3'UTR regions, in *RUNX2*, *ELL2*, *TNFSF15*, *TNFSF4*, *TRAFF3*, and *FADS2* genes, and three loci had missense proxies, in *FBXL19*, *ELL2* and *SH2B3* genes. We identified the Thr298Ala missense variant in *ELL2* exon 7 (rs3815768, $P=2.9 \times 10^{-27}$) in linkage disequilibrium with the lead SNP (rs3777175, $P=7.8 \times 10^{-30}$, $r^2=0.69$) and located to the end of the ELL2 domain required for transcriptional elongation²⁷. At the *SH2B3* locus, rs3184504 in tight LD with the top SNP ($r^2=0.94$) is a missense variant in the canonical transcript of *SH2B3* gene, although this variant is predicted to be benign by PolyPhen2. Given that other top signals map to non-coding regions, we evaluated their potential regulatory function by systematic eQTL co-localization analyses using whole blood²⁸ and 13 primary immune cell types²⁹. The co-localization probability (PP4)

exceeded 50% in at least one cell type for 14 of 20 GWAS loci, prioritizing biologically plausible candidate genes at each of these loci (**Figure 3a and Supplementary Table 7**). We will next summarize some of our findings for the top GWAS loci.

***RUNX2* and *RUNX3* loci.** The *RUNX3* locus on chr.1p36.11 represents one of the strongest signals in our meta-analysis with the largest effect size (rs188468174, Beta=0.88, $P=3.42 \times 10^{-92}$). This locus was previously associated with IgA levels²⁰ and IgG glycosylation³⁰. Interestingly, we have also picked up a novel locus with smaller effect on chr.6p21.1 encoding *RUNX2*, a related transcription factor (rs1200427, Beta=0.06, $P=6.85 \times 10^{-14}$). *RUNX* transcription factors are essential regulators of diverse developmental and signaling pathways^{31,32}. Both *RUNX2* and *RUNX3* have previously been linked to retinoic-acid- and TGF- β -induced IgA class switching³³. The novel *RUNX2* locus co-localized with the eQTL for the *RUNX2* gene in whole blood with PP4=0.99; the IgA-decreasing allele at the index SNP (rs1200427-G) was associated with higher mRNA levels of *RUNX2* (**Figure 3a and Supplementary Table 7**). Notably, the lead SNP is in LD with rs1200428 ($r^2=0.51$), the 3'UTR variant in *RUNX2* that has the strongest eQTL effects in blood. Similar phenomenon was previously reported for the *RUNX3* locus, where the minor allele of the top SNP (rs188468174-T) was associated with lower IgA levels and increased mRNA abundance of the long isoform of *RUNX3*²⁰. Therefore, our study solidifies the evidence for a genetic control model of *RUNX* transcription factors, in which increased *RUNX* expression suppresses IgA class switching, reducing circulating IgA levels.

***ELL2* locus.** Another strong signal in our GWAS comes from the previously reported locus on chr.5q15 (rs3777175, Beta=0.08, $P=7.81 \times 10^{-30}$). We identified three independent signals by conditional analysis of this locus, suggesting a complex pattern of association. The co-localization analysis between this region and eQTLs in whole blood suggested *ELL2* (elongation factor for RNA polymerase II 2) as a candidate gene (PP4=0.53, **Figure 3a and Supplementary Table 7**) with the IgA-increasing allele (rs3777175-G) associated with lower expression of *ELL2* ($P=6.76 \times 10^{-24}$). The negative correlation between *ELL2* expression and IgA production has recently been reported in another study³⁴. Jointly, these results prioritize *ELL2* as a causal gene at this locus.

***TMEM258/FADS2* locus:** We detected two independent genome-wide significant signals on chr.11q12.2 (rs968567, Beta=0.12, $P=2.42 \times 10^{-41}$, and rs102275, Beta=0.07, $P=3.60 \times 10^{-29}$) within the *TMEM258-FADS2* locus previously associated with IgA levels²⁰. The top SNP is within the *FADS2* gene, and the IgA-increasing haplotype has previously been associated with ratio of arachidonic to linoleic acid reflective of fatty acid desaturase activity and linked to the risk of coronary artery disease^{35,36}. At the same time, the IgA-increasing haplotype has previously been described as protective from rheumatoid arthritis (**Figure 2c**). The second independent signal centers on the *TMEM258* gene, previously associated with the risk of Crohn's disease³⁷. The co-localization analyses of this locus across multiple immune cell types and whole blood suggested several candidate genes, including *FADS1* and *FADS2* genes (fatty acid desaturases regulating unsaturation of fatty acids, co-localized across most cell types) and the *TMEM258* gene (co-localized specifically in T cells), with the IgA-increasing allele (rs968567-T) associated with higher expression of all three transcripts. Rs968567 falls into a genetically regulated histone-modification peak (H3K4me1) for *FADS2* in T cells, monocytes and neutrophils³⁸. *TMEM258* is a particularly attractive candidate gene at this locus, because it appears to play an important role in the regulation of intestinal inflammation, and mice deficient in *TMEM258* exhibit severe intestinal inflammation in a model of colitis³⁹.

***LITAF, TRAF3, and HDAC7/VDR* loci.** The following three loci exhibit immune cell-type specific co-localization. The locus on chr.16p13.13 (rs113962704, Beta=0.05, $P=1.91 \times 10^{-12}$) encodes lipopolysaccharide-induced TNF-alpha factor (*LITAF*), a transcription factor regulating TNF-alpha expression in intestinal macrophages^{40,41}. Notably, macrophage-specific deficiency of *LITAF* in mice leads to attenuated TNF and IL-6 response upon LPS stimulation⁴². Our locus co-localized with eQTL for *LITAF* in monocytes and the top SNP represented a monocyte-specific hQTL (H3K27ac)³⁸, suggesting that it alters *LITAF* enhancer activity in the monocytic lineage (**Figure 3a and c**). The IgA-increasing allele (rs113962704-T) was associated with higher mRNA expression of *LITAF* in monocytes, supporting the hypothesis that this transcription factor provides a stimulus for IgA production by altering monocyte function.

The locus on chr.14q32.32 (rs12147883, Beta=0.05, $P=5.42 \times 10^{-14}$) contains *TRAF3* encoding TNF Receptor Associated Factor 3, a protein participating in the CD40 signaling, inhibiting non-classical NF-κB signaling, and participating in the regulation of class switch recombination in B cells⁴³⁻⁴⁵. Interestingly, this locus

co-localizes with eQTL for *TRAF3* specifically in T cell lineage, where the IgA-increasing variant is associated with higher *TRAF3* expression. In contrast to its inhibitory functions in B-cells, TRAF3 is known to promote many T-cell effector functions through enhancing signaling by the T-cell receptor-CD28 complex^{46,47}.

The third locus on chr.12q13.11 (rs7487637, Beta=0.05, $P=9.97 \times 10^{-15}$) contains *HDAC7* encoding an important histone deacetylase regulating differentiation and function of T-cells⁴⁸. Accordingly, this GWAS locus co-localizes with eQTL for *HDAC7* specifically in T-cells, where the IgA-increasing allele is associated with higher mRNA levels of *HDAC7*. Notably, *VDR* (encoding Vitamin D Receptor) is another promising candidate gene at this locus previously prioritized as a potential host factor shaping the gut microbiome⁴⁹, however, our top SNP at this locus was not in LD with the top SNP from the microbiome GWAS ($r^2=0.0007$).

TNFSF4/TNFSF18, TNFSF8/TNFSF15, and TNFSF13 loci. We detected three independent loci encoding members of tumor necrosis factor ligand superfamily; *TNFSF4/TNFSF18* on chr.1q25.1 (rs7518129, Beta=0.06, $P=1.06 \times 10^{-16}$, *TNFSF4* is the closest gene), *TNFSF8/TNFSF15* on chr.9q33.1 as previously reported¹⁶ (rs3181356, Beta=0.07, $P=1.13 \times 10^{-18}$, *TNFSF8* is the closest gene), and a suggestive *TNFSF13* locus on chr.17p13.1 (rs3803800, Beta=0.06, $P=9.41 \times 10^{-08}$, also previously associated with IgA levels¹⁹). These loci encode powerful TNFSF cytokines with partially overlapping receptors⁵⁰. *TNFSF8/TNFSF15* and *TNFSF13* loci are associated with the risk of IgA nephropathy (see accompanying manuscript), while *TNFSF4/TNFSF18* locus has previously been associated with the risk of eczema, asthma and narcolepsy, all with concordant effects to IgA levels (**Figure 2c and Supplemental Table 5**).

OVOL1/RELA locus. The locus on chr.11q13.1 (rs10896045, Beta=0.07, $P=2.57 \times 10^{-22}$) encodes multiple candidate transcripts including *OVOL1*, *RELA*, and several others. The co-localization analysis suggested that this locus was shared with IgA nephropathy, a kidney disease due to IgA deposition in the glomeruli, with the IgA-increasing allele associated with increased risk of IgA nephropathy. Previous GWAS also pointed to the protective role of this locus from atopic dermatitis (**Figure 2c**). While the index SNP localizes to the intronic portion of *OVOL1*, a putative transcription factor of poorly defined function, the nearby *RELA* gene encodes a subunit of NF- κ B complex⁵¹. We found no significant eQTL effects for either *RELA* or *OVOL1* in blood or primary immune cells. However, the lead SNP had significant eQTL effects on *OVOL1* transcript in thyroid

($P=3.1 \times 10^{-12}$), spleen ($P=1.5 \times 10^{-6}$) and EBV-transformed lymphocytes (4.7×10^{-9}) with the IgA-increasing allele (rs10896045-A) associated with high expression of *OVOL1* gene in all three GTEx tissues.

ANKRD55/IL6ST locus. The locus on chr.5q11.2 (rs6859219, Beta=0.07, $P=1.41 \times 10^{-20}$) encodes *ANKRD55* (closest gene) and *IL6ST*. This locus has previously been associated with IgG glycosylation⁵² and increased risk of several immune-mediated disorders, including multiple sclerosis, rheumatoid arthritis, and Crohn's disease (Figure 1c, Supplementary Table 5). In all cases, the IgA-increasing allele represented a risk factor for these immune-mediated traits. We co-localized this locus with cis-eQTL for *ANKRD55* in whole blood (PP4=0.98) and T regulatory cells (PP4=0.51) with concordant effect of *ANKRD55* expression and IgA levels. *ANKRD55* gene expression is known to be specific to CD4⁺ T cells, and mouse studies suggest that this transcript is induced by inflammation and displays important T-cell regulatory functions⁵³.

IL1R1 locus. The locus on chr.2q11.2 (rs13427957, Beta=0.04, $P=6.19 \times 10^{-10}$) contains the *IL1R1* gene encoding Interleukin 1 Receptor Type 1, an important T-cell receptor involved in many cytokine-induced immune and inflammatory responses⁵⁴. This locus co-localizes with eQTL for *MIR4772* in Th1 cells (PP4=0.69) with concordant effect on IgA levels, and *IL1R1* is predicted to be a target gene of *MIR4772* by miRDB⁵⁵ and TargetScan⁵⁶. A mouse knock-out of *IL1R1* in Th2 cells had decreased IgA levels⁵⁷. These observations support the role of *IL1R1*, but the specific genetic mechanism at this locus will require dedicated studies in human T cells.

POU2AF1 locus. The novel locus on chr.11q23.1 (rs4938518, Beta=0.056, $P=7.01 \times 10^{-16}$) encodes *POU2AF1* (POU class 2 homeobox associating factor 1), the gene involved in B-cell antigen responses required for the formation of germinal centers⁵⁸. Mouse knock-out of *POU2AF1* leads to increased B-cell apoptosis and decreased IgA production⁵⁹. The IgA-decreasing allele at this locus has been associated with increased risk of primary biliary cholangitis (Figure 2c and Supplementary Table 5).

Additional loci. The *CTF1* locus on 16p11.2 (rs1458201, Beta=0.052, $P=2.02 \times 10^{-11}$) encodes cardiotrophin 1 involved in multiple immune-related pathways including cytokine signaling in immune system and interleukins,

IL-6-type cytokine receptor ligand interactions and signaling (**Figure 4a**). The *CITED2* locus on chr.6q24.1 (rs17069163, Beta=0.05, $P=5.94 \times 10^{-11}$) encodes a CBP/P300 interacting trans-activator, which functions as a molecular switch of TGF- α and TGF- β induced signaling⁶⁰ and has previously been implicated in immune homeostasis and tolerance⁶¹. The locus on chr.1q41 (rs73100295, Beta=0.36, $P=3.91 \times 10^{-08}$) encodes the *GPATCH2* gene, but its role in the immune system regulation is unknown. The co-localization analysis of the locus on chr.7q11.23 (rs55722505, Beta=0.048, $P=8.61 \times 10^{-14}$) suggested several candidate effector genes including *SRCRB4D*, *DTX2* and *YWHAG* genes in whole blood, *ZP3* and *SSC4D* in monocytes, and *POMZP3* in naïve CD8 and B cells (**Figure 3a**). Lastly, we confirmed the previously described effect of the HLA locus, and our stepwise conditional analysis identified at least eight independent SNPs contributing to the HLA signal, indicating highly complex association patterns at this locus.

Convergence with relevant mouse phenotypes. To further prioritize causal genes at our GWAS loci, we tested for significant gene set overlaps with human orthologs that cause immune-related phenotypes when disrupted in mice. This included gene sets for abnormal IgA levels (120 genes), abnormal immune tolerance (408 genes) and abnormal response to infection (562 genes) from the Mouse Genome Informatics (MGI) database⁵⁹. We identified significant overlap for 13 genes linked to abnormal IgA levels in mice (enrichment $P=5.8 \times 10^{-5}$), 29 linked to abnormal immune tolerance (enrichment $P=6.3 \times 10^{-6}$) and 42 linked to abnormal response to infection (enrichment $P=1.1 \times 10^{-5}$) (**Figure 4b and Supplementary Tables 8-10**). Among the 120 genes with abnormal IgA phenotypes in mice, six additional genes (*TRIM13*, *NCR1*, *IRF5*, *FGR*, *ARHGAP15* and *REL*) surpassed a Bonferroni-corrected threshold ($P=0.05/22169=2.26 \times 10^{-6}$) when tested against our GWAS results; these represent plausible new candidate genes for regulation of IgA production in humans.

Relationship to the risk of IgA nephropathy and tonsillectomy. Given the known role of the IgA system in the pathogenesis of IgA nephropathy, a kidney disease caused by glomerular deposition of IgA in the setting of pharyngitis, tonsillitis, and other mucosal infections, we specifically tested for shared genetic architecture between IgA levels, IgA nephropathy (see accompanying manuscript), and tonsillectomy⁶² by systematic lookups and co-localization analyses of genome-wide significant loci. Among all 31 independent alleles associated with higher IgA levels, 12 had nominal associations with increased risk of IgA nephropathy

($P < 0.05$), all with concordant effects (**Supplementary Table 11**). Among these, there were five genome-wide significant loci with high probability of shared causal variants ($PP_4 > 0.7$), *TNFSF4/TNFSF18*, *ANKRD55/IL6ST*, *OVOL1/RELA*, *SH2B3*, and *HORMAD2/LIF* (**Supplementary Table 12**). There were also four loci with an overlapping genomic position, but high probability of different causal variants between the two traits ($PP_3 > 0.7$), *HLA*, *CTF1*, *TRAF3* and *TNFSF8/TNFSF15*. Between IgA levels and tonsillectomy, we observed two co-localized loci (*SH2B3* and *HORMAD2/LIF*), both with concordant effects, and another two loci (*HLA* and *CTF1*) with high probability of different causal variants (**Supplementary Table 13**). Remarkably, the *HORMAD2/LIF* locus was genome-wide significant in all three GWAS and co-localized across all three traits, suggesting a common genetic mechanism (**Figure 5a and b**).

To further explore the genetic relationships between these traits, we performed two sample Mendelian Randomization (MR) between the co-localized IgA level-associated loci as an exposure instrument and IgA nephropathy and tonsillectomy as disease outcomes. Using this strategy, we estimated significant causal effects between serum IgA levels and IgA nephropathy (inverse variance-weighted OR 9.70 per SD of exposure, 95%CI: 6.80-13.8, $P < 0.001$; **Figure 5c**), supporting IgA level as a causal mediator of disease risk for these loci. Sensitivity analysis confirmed that all co-localized loci contributed with concordant effects, there were no outlier effects, and there was no evidence of directional horizontal pleiotropy (Egger intercept test $P = 0.58$). Moreover, this effect remained highly significant when instrumental variables were expanded to encompass all genome-wide significant non-HLA loci for serum IgA levels (inverse variance-weighted OR=2.49 per SD of IgA level, 95%CI: 1.56-3.99, $P < 0.001$). In contrast, we detected no significant causal effects between IgA levels and tonsillectomy ($P = 0.26$), or tonsillectomy and IgA nephropathy ($P = 0.96$).

Genetic correlations with other complex traits studied by GWAS. We assessed genome-wide genetic correlations (r_g) of IgA levels with 52 complex traits and diseases, including 13 immune-mediated disorders, 23 infectious diseases, and 16 cardio-metabolic traits using stratified LD score regression^{21,63} (**Figure 6a and Supplementary Table 14**). In the analysis that excluded the HLA region (to remove potential bias from large effects and extended LD at this locus), we confirmed positive genetic correlation with IgA nephropathy ($r_g = 0.35$, $P = 0.002$), tonsillectomy ($r_g = 0.20$, $P = 0.01$), type 2 diabetes ($r_g = 0.18$, $P = 0.01$), and BMI ($r_g = 0.13$, $P = 0.03$). We observed a negative genetic correlation with Crohn's disease ($r_g = -0.19$, $P = 0.005$), celiac disease ($r_g = -0.21$,

$P=0.007$), and inflammatory bowel disease ($r_g=-0.13$, $P=0.04$). The observed negative correlation with traits that involve gut inflammation could be potentially explained by the protective anti-inflammatory effects of mucosal IgA. Among infectious disease GWAS, we observed mainly negative genetic correlations, including with bacterial meningitis ($r_g=-0.47$, $P=0.005$) and shingles ($r_g=-0.46$, $P=0.009$), despite the fact that most existing GWAS for infections are either underpowered or based only on self-report⁶². For most, but not all phenotypes, genetic correlations with and without HLA were comparable, as summarized in **Supplementary Figure 5 and Supplementary Table 14**.

Phenome-wide association studies (PheWAS). To detect additional genetic associations, we derived a genome-wide polygenic score (GPS) for serum IgA levels and tested for its phenotypic associations using meta-PheWAS approach across the UK Biobank and eMERGE-III datasets (but removing any participants included in the discovery GWAS, see **Online Methods**). In the combined analysis of 556,656 participants, we detected 31 significant phenotypic associations of the GPS (**Figure 6b and Supplementary Table 15**). This included several protective associations with immune and inflammatory disorders, such as celiac disease (OR per SD=0.54, $P=4.6 \times 10^{-227}$), hypothyroidism ($OR_{SD}=0.94$, $P=1.8 \times 10^{-19}$), type 1 diabetes ($OR_{SD}=0.91$, $P=8.2 \times 10^{-12}$), and psoriasis ($OR_{SD}=0.91$, $P=1.7 \times 10^{-10}$). Among significant risk associations were disorders of iron metabolism ($OR_{SD}=1.31$, $P=4.0 \times 10^{-19}$) and hematuria, a common manifestation of IgAN ($OR_{SD}=1.04$, $P=1.0 \times 10^{-8}$). To assess which of these associations were driven by the HLA region, we next repeated meta-PheWAS after excluding all HLA variants from the GPS (**Supplementary Figure 6 and Supplementary Table 15**). In this analysis, we detected only two phenome-wide-significant associations: a protective association with Celiac disease (OR per SD=0.86, $P=4.6 \times 10^{-12}$) and a risk association with morbid obesity ($OR_{SD}=1.09$, $P=3.0 \times 10^{-5}$). These associations were direction-consistent with our genome-wide genetic correlation analyses.

Given these results, we next applied Mendelian randomization approach to resolve potential causal relationships between serum IgA levels, Celiac disease, and BMI. For instrumental variables, we used independent genome-wide significant alleles (excluding HLA region) from this study, and from the largest studies for Celiac disease⁶⁴ and BMI⁶⁵. Interestingly, we detected no significant causal effects between serum IgA levels and Celiac disease or BMI. In reverse causality analyses, however, we observed a highly significant causal effect of BMI on serum IgA levels (inverse variance weighted effect = 0.12, 95%CI: 0.05-0.19, $P<0.001$).

In sensitivity analyses, there was evidence of directional pleiotropy (Egger intercept test $P=0.004$), but the causal effect of BMI became stronger and more significant when the balanced pleiotropy assumption was relaxed (Egger regression effect = 0.38, 95%CI: 0.19-0.58, $P<0.001$). These results suggest that elevated serum IgA levels in individuals with metabolic syndrome may represent a consequence rather than a cause of obesity.

Genetic correlations with microbiome diversity. Because many of the detected loci for IgA levels implicate dysregulation of the mucosal immune system and intestinal inflammation, we investigated genome-wide genetic relationships between IgA levels and intestinal microbiome diversity and abundance based on recent GWAS studies^{49,66}. Notably, we observed a negative genome-wide genetic correlation between IgA levels and microbiome diversity ($r_g=-0.22$, $P=0.034$). We then interrogated the significance of each individual genome-wide significant IgA level locus against GWAS summary statistics for intestinal microbiome diversity⁴⁹. This analysis demonstrated an excess of genetic associations compared to null expectation, with the strongest associations at the *TNFSF8/TNFSF15* and *HORMAD2/LIF* loci (**Figure 7a**). Next, we tested the significant IgA loci against GWAS summary statistics for abundance of 123 intestinal microbiome species and 52 genera⁶⁶. *Dorea longicatena* was the most strongly associated species as ranked by the sum of the Z-scores for the 20 independently significant SNPs from GWAS for IgA levels, with the strongest association for *TNFSF4/TNFSF18* and *HDAC7/VDR* loci (**Figure 7b**). The opposed effects of these loci between IgA levels and *Dorea longicatena* are consistent with the notion that IgA-increasing alleles may reduce microbial diversity by altering the abundance of selected species. In the genus-level analyses, *Dorea* and *Subdoligranulum* were among the top associations based on the ranking of z-scores (**Figure 7c**).

Discussion

Our study provides multiple insights into the genetic regulation of serum IgA levels and dissects shared genetic effects between IgA levels and several human diseases. Our large multi-ethnic GWAS identified 20 genome-wide significant loci, 11 of which were novel. These loci encode genes enriched in immune-related pathways, with 13 candidate genes demonstrating IgA abnormalities when genetically manipulated in mice. The complementary enrichment analyses based on gene expression across multiple tissues/cell types highlighted primary immune cells, mainly T cells, B cells and monocytes/macrophages, as the most likely effector cell types regulating IgA production.

Previous smaller GWAS for serum IgA levels have been limited to European or East Asian ancestry, while African and other ancestries have not been included in these studies. Our multi-ethnic cohort included four diverse ancestral groups and allowed us to demonstrate that African ancestry was associated with higher serum IgA levels compared to other ancestries. Consistent with this observation, IgA-increasing alleles at 15 out of the 20 genome-wide significant loci were more frequent in populations of African ancestry compared to non-African populations. These observations could be explained by multi-locus adaptation potentially favoring IgA-increasing alleles in this ancestral group, but environmental exposures correlated with African ancestry may also be contributing to the observed differences. Although we observed no clear signature of polygenic adaptation, our study was not designed to test these hypotheses and the existing methods for detecting multi-locus selection lack sensitivity²².

Our systematic co-localization analysis of GWAS associations and eQTLs in whole blood and 13 different immune cell types prioritized new biologically plausible candidate genes and target cell types for 14 of the 20 significant loci, providing an extensive resource for follow up studies. We also observed that many eQTL co-localizations were present only in specific immune cell subtypes. As an example, variants within the *LITAF* locus co-localize specifically with *LITAF* eQTL but only in monocytes, pointing to monocyte lineage as the most likely causal cell type for this locus. In contrast, *TRAF3*, *HDAC7*, *IL1R1* and *ANKRD55* loci co-localize with eQTLs specifically in human T cell lineages, prioritizing T cells for functional studies of these loci.

In the analysis of genetic relationships with other traits, we observed shared polygenic determination between IgA levels and several immune, infections, renal, and cardio-metabolic traits. Our findings support the protective role of IgA in susceptibility to various infectious pathogens as well as inflammatory bowel disease.

At the same time, we observed a positive genetic correlation between IgA levels and IgA nephropathy, a common form of kidney disease due to IgA deposition. In particular, Mendelian randomization analysis suggested a causal role for elevated IgA levels in the pathogenesis of IgA nephropathy. Ongoing IgA nephropathy clinical trials targeting pathways that reduce IgA levels will shed light on this hypothesis⁶⁷.

We also observed an inverse genetic correlation of serum IgA levels with intestinal microbiome diversity, which may be related to the associations of IgA levels with obesity and inflammatory bowel disease. The species abundance of *Dorea longicatena* in particular was most strongly correlated with GWAS loci for IgA levels in our study, and associated with obesity and markers for insulin resistance in prior studies^{68,69}. At the genus level, the intestinal abundance of *Subdoligranulum* was negatively associated with IgA-increasing alleles, and this genus has previously been reported to be depleted in individuals with immune-mediated diseases⁷⁰. IgA-microbiome interactions can regulate the composition of microbiota at mucosal surfaces but may also shape the host IgA response⁷¹. We note that current GWAS for intestinal microbiome are underpowered, thus larger studies are needed to dissect the precise case-effect relationships between these traits using Mendelian randomization approaches.

In summary, we identified 20 genome-wide significant loci associated with serum IgA levels, and we prioritized potential effector genes and cell types for 70% of the loci. We demonstrated that IgA levels are positively correlated with African ancestry. We further characterized shared genetic architecture between serum IgA levels and other complex traits, demonstrating that while IgA-increasing alleles appear to have protective effects against infections, they may represent risk factors for selected auto-immune, kidney, and cardiometabolic diseases. Importantly, the risk of IgA nephropathy appears causally related to elevated IgA levels, and several genome-wide significant loci colocalize between these traits. Lastly, the inverse genetic correlation between IgA levels and intestinal microbiome diversity suggested a potential link with a metabolic syndrome that warrants follow-up investigation.

Methods

Measurement of serum IgA levels

Serum total IgA levels were measured via a previously optimized and validated 'sandwich' ELISA protocol. 96-well plates were coated with an IgA capture antibody (AffiniPure F(ab')₂ Fragment of Goat Anti-Human Serum IgA, α Chain Specific; Jackson Immuno Research, #09-006-011), covered, and let to sit on a shaker for 60 minutes. The plate was subsequently washed three times with a TBS Wash buffer. Next, 250-uL Blocking Buffer was added to the wells, and the plate was again covered, and set on the shaker for 90 minutes. After aspirating the wells of the blocking buffer, 100 uL/well of samples and standards were added, all in duplicate. Standards were diluted 1:3 in series beginning with 500 ng/mL. The plate was left to incubate for 120 minutes, after which samples were aspirated and the plate was washed thoroughly five times. Next, 100uL of secondary/detection antibody (Mouse monoclonal Ab to Human IgA horseradish peroxidase conjugate, Abcam Bio #ab7383), was added at 62.5ng/mL to the wells and left to incubate for 60 minutes. The plate was washed an additional five times. Next, 100 uL of TMB 2 component ELISA substrate (KPL, 50-76-00) was added to the wells, and left to incubate for 12 minutes. 100 uL of 2M H₂SO₄ (Sigma Aldrich, #339741) was added to each well at the end of the 12 minutes to stop the reaction, per substrate manufacturer recommendation. The samples were then measured using the BIO-TEK PowerWaveTM XS and KC-Jr. plate reading software. The plate was read at 450 nm, with a 630 nm reference wavelength, per substrate manufacturer recommendation. Samples with an internal CV greater than 10% were re-run. Samples that fall outside of the standard range (500 ng/mL - 0.2 ng/mL) were re-run at appropriate dilutions.

MESA cohorts

The Multi-Ethnic Study of Atherosclerosis (MESA) is a diverse population-based cohort of participants recruited prospectively for studies of cardiovascular disease⁷². The recruitment took place across six clinical centers in the United States. The participants were genotyped using the Affymetrix Human SNP array 6.0⁷³. We measured serum IgA levels using standardized ELISA protocol (see above) in N=5420 participants, and these individuals were included in the GWAS analysis. The standard genotype quality control (QC) filters included per-SNP genotyping rate >95%, per-individual genotyping rate >90%, MAF >0.01, and HWE test p-

value $>1 \times 10^{-5}$ within each ancestry group. We assessed for cryptic relatedness and duplicates, and we excluded one individual from any pairs with estimated pairwise kinship coefficients >0.05 . Gender of each individual was imputed based on the analysis of sex chromosome markers and individuals with gender mismatch against records were excluded. The imputation analysis was carried out using Minimac3⁷⁴ after pre-phasing in Eagle V2.3⁷⁵ and using 1000 Genomes (Phase 3) as reference⁷⁶. A total of 11,102,943 common high-quality markers ($R^2 > 0.8$ and $MAF > 0.01$) were imputed and used in the downstream analyses. The principal component analysis (PCA)⁷⁷ was used to assign genetic ancestry to the participants based on co-clustering with the major continental reference populations of the 1000 Genomes Phase 3⁷⁶. After exclusions of ancestry outliers, we identified 5 distinct ancestral clusters (European: N=2,280, African: N=1,275, Admixed 1: N=474 and Admixed 2: N=750, and East Asian: N=641). Subsequent GWAS analyses were then performed within each ancestry cluster. For ancestry adjustment, we re-run PCA within each cluster, defined significant PCs by Tracy-Widom test, and included the significant PCs for each cluster as covariates in the association testing.

Electronic Medical Records and Genomics (eMERGE-III) cohorts

The eMERGE consortium consists of 12 medical centers with electronic health records (EHRs) linked to genome-wide genotype data for 102,138 individuals. The serum IgA level was extracted by a lab value query performed by eight active sites in the eMERGE phase III. The genotyping and imputation of the eMERGE cohort have been previously described in detail⁷⁸⁻⁸⁰. Briefly, we implemented the minimac3 imputation model with HRC1.1 as references for each genotyping platform in a separate batch. After imputation, we merged all the 81 imputed batches based on genomic position. After QC filters including $R^2 \geq 0.8$ and $MAF \geq 0.01$ in $\geq 75\%$ of 81 imputation batches, around 7 million high-quality markers were used in our downstream analysis. Further, we excluded duplicates and cryptic relatedness in the given cohort determined by the estimated pairwise kinship coefficients >0.05 . The genetic ancestry for each individual was assigned based on PCA with reference populations of the 1000 Genomes Phase 3⁷⁶. After exclusions of ancestry outliers, the final dataset with matched serum IgA phenotypes consisted of 6,047 individuals across five ancestral groups (European: N=4,261, African: N=476, East Asian: N=73, Admixed 1: N=235 and Admixed 2: N=1002). Like

the MESA cohort analysis, we repeated the PCA within each ancestral cluster and defined significant PCs to be used as covariates in association testing.

German, French, Chinese, Japanese, and U.S. cohorts

The description of these cohorts including recruitment and measurements of serum IgA levels have been published previously⁸¹. For this study, we included only individuals who were ascertained as 'healthy population controls' for GWAS studies, thus these cohorts were not enriched in any specific disease type. The genotyping was performed using Illumina 550v3 (US cohort), Illumina 370-Duo (French cohort), Illumina MEGA v1.0 (Chinese and Japanese cohorts) and Illumina MEGA v1.1 (German cohort). The standard genotype QC filters included per-SNP genotyping rate >95%, per-individual genotyping rate >90%, MAF >0.01, and HWE test p-value >1x10⁻⁵ within each cohort. We assessed for cryptic relatedness and duplicates, and we excluded one individual from any pairs with estimated pairwise kinship coefficients >0.05. Gender of each individual was imputed based on the analysis of sex chromosome markers and individuals with gender mismatch against records were excluded. The imputation was performed using MACH 1.0 for pre-phasing and then Minimac3 for imputation based on ancestry-matched reference panels of 1000 Genome Project (Phase 3). We performed PCA of each dataset to exclude outliers and define the number of significant PCs by Tracy-Widom test. The final numbers of individuals and high-quality markers (R²>0.8, MAF>0.01) used in downstream analyses were as follows: the German cohort included N=156 healthy individuals and 7,612,078 markers; the French cohort included N=103 healthy population controls and 7,096,980 markers, the Chinese cohort N=467 healthy individuals and 5,113,877 markers, the Japanese cohort N=776 healthy individuals and 6,673,613 markers; and the U.S. cohort N=93 healthy individuals and 7,439,363 markers.

Swedish cohort

We obtained summary statistics for GWAS for total IgA levels for 9,617 participants; the ascertainment, genotyping, and analysis of this cohorts has been published previously¹⁶. In order to improve marker density, we re-imputed this cohort based on association estimates of the genotyped markers from the summary statistics using ImPG V1.0⁸² software and the 1000 Genomes (Phase 3) European reference⁷⁶. Using ImPG

software, we derived the posterior mean of z-scores at untyped SNPs given the z-scores at typed SNPs and the correlation matrix among all pairs of SNPs induced by their linkage disequilibrium (LD) that were estimated using the reference panel. The effect size and standard error for each imputed SNP were then estimated based on its imputed z-score and reference allelic frequency as described previously⁸³. A total of 6,907,390 variants were imputed with high quality ($R^2 > 0.8$ and $MAF > 0.01$) and included in our downstream analysis.

Decode cohorts

The analysis of this cohort composed of 16,883 participants from Iceland and 2,151 individuals from southern Sweden has been published previously²⁰. The association summary statistics for IgA levels including effect size, P-value and minor allele frequencies for all variants with $P < 1 \times 10^{-6}$ were provided by the authors in their publication²⁰. In order to incorporate this dataset using fixed-effects meta-analysis, the unbiased standard error for each variant was derived using the following equation⁸⁴:

$$SE = \sqrt{\frac{1 - 2p(1-p)b^2}{2p(1-p)n}}$$

where b is the standardized allelic effect on IgA levels, p is the minor allele frequency, and n is the sample size.

Genome-wide association studies and multiethnic meta-analysis

We conducted a multiethnic meta-analysis of 17 discovery cohorts including a total of 41,263 individuals. Multiethnic cohorts were classified into ancestry-specific strata based on global PCA analysis. In each sub-cohort, serum IgA levels were log-transformed and expressed as standard-normalized residuals from regression of log-transformed IgA levels against age and sex. We performed genome-wide association testing in each cohort for the markers that were imputed at high quality ($r^2 > 0.8$) using a linear regression model under additive coding of the dosage genotypes, and with adjustment for cohort-specific significant principal components (PCs) of ancestry⁸⁵. To quantify potential inflation of type I error due to stratification or technical artifacts, we estimated the genomic inflation factor for each cohort but detected no substantial inflation with $\lambda < 1.05$ in each individual study. We performed a fixed-effects meta-analysis to combine the results of

all 17 individual cohort summary statistics using METAL⁸⁶. The quantile-quantile plot of the final meta-analysis showed no global departures from the expected null distribution, with the genomic inflation factor estimated at 1.016 (**Supplementary Figure 2**). The genome-wide significant signals were defined by the generally accepted $P < 5.0 \times 10^{-8}$ and signals with $P < 1.0 \times 10^{-6}$ were considered as suggestive.

Conditional analyses

To detect independent variant associations at each genome-wide significant locus, we performed stepwise conditional analysis using meta-analysis summary statistics and a multi-SNP-based conditional & joint association analysis (COJO) in GCTA software⁸⁷. This method approximates the variance-covariance matrix between association statistics with LD information from an external reference panel. In our analysis, we used 1000 Genome Project Phase 3⁷⁶ reference populations of the same ancestries as our discovery cohorts.

Tissue/cell type and pathway enrichment analyses

The region of each GWAS locus was defined by first selecting all proxy SNPs in LD ($r^2 > 0.5$) with the lead SNP, then extending the genomic region 250 kb upstream and downstream of the first and last proxy SNP based on genomic position. Each region was then annotated using Ensembl human gene annotations. Gene sets were created for all genome-wide significant regions but excluding the HLA region. For tissue/cell type enrichment, we used DEPICT (Data-driven Expression-Prioritized Integration for Complex Traits) to test for tissues and cell-types in which genes from the associated regions were highly expressed as previously described²⁴. Next, for each tissue, empirical enrichment p-values were computed by repeatedly sampling random sets of loci from the entire genome to estimate the null distribution for the enrichment statistic. For pathway enrichment analysis, we used established pathways from the databases including Molecular Signatures Database (MSigDB C2), KEGG, BioCyC, REACTOME, Pathway Interaction Database. Statistical significance for enrichment was set at FDR q-value < 0.05 .

Functional annotations

We first defined each GWAS locus by +/- 400kb of the genome-wide significant index SNP. We annotated all transcripts within these intervals using the latest assembly of the human genome (hg19) to create sets of

positional candidate genes for each locus. Using ANNOVAR software⁸⁸, we annotated all variants within the region that were in LD ($r^2 > 0.5$) with the top SNP, including all known coding, splicing, and 3'UTR and 5'UTR variants. To prioritize the candidate genes at each of the GWAS loci, we next performed colocalization analyses based on our meta-analysis statistics and gene expression QTLs in whole blood quantified from 31,684 individuals²⁸, as well as 13 human immune cell types from the Database of Immune Cell eQTLs (DICE) project²⁹. After harmonization of effect alleles, we identified all co-localized eQTLs mapping to the region of the index SNP ± 400 kb using Coloc Package in R⁸⁹. Co-localization with PP4 greater than 0.7 was considered as strong evidence in support of shared causal SNPs, while PP4 > 0.5 was considered as suggestive. To test for histone QTL effects in immune cells, we interrogated all GWAS index SNPs and their proxies ($r^2 > 0.5$) against histone QTL of three major human immune cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T-cells) based on the analysis of ~200 individuals³⁸.

Protein-Protein Interactions (PPI) network analyses

Protein-protein interactions among the positional candidate genes at the GWAS loci were predicted using InWeb_InBioMap⁹⁰, a curated and computationally derived regulatory network of 420,000 interactions (Supplementary Figure 3). We used high confidence interactions defined by the recommended cut off confidence score < 0.1 . The candidate gene network contained a total of 48 genes and 52 interactions. The network components were grouped into ten modules based on their connectivity. Functional and pathway enrichments within each module were identified based on Gene Ontology, KEGG, and Reactome databases. In addition, we used ToppGene Suite⁹¹ to calculate interaction enrichment p-values for each gene. A Bonferroni-corrected $P < 0.05$ was used as enrichment significance cut-off.

Intersection with related mouse phenotypes

We evaluated genes that when knocked out cause abnormal immune phenotypes in mice based on the comprehensive MGI phenotype ontology database. The following mouse phenotypes were evaluated: abnormal IgA levels (MP:0020171); abnormal immune tolerance (MP:0005000); and abnormal response to infection (MP:0005025). The human orthologs of these genes were obtained with the Human–Mouse Disease Connection web tool (<http://www.informatics.jax.org/humanDisease.html>). The significance of intersections

between these gene sets and the list of positional candidate genes from GWAS was determined using a hypergeometric test.

Heritability and genetic correlations with other phenotypes

SNP-based heritability of IgA level GWAS was estimated by LD score regression model using the LDSC software²¹. The LD score for each SNP was estimated based on LD matrices derived from ancestry-matched 1000 Genome Project Phase 3 populations. To investigate the shared genetic architecture between IgA levels and other phenotypes, we first collected the summary statistics of autoimmune and inflammatory disorders and cardio-metabolic traits from the LD-hub⁹² or GWAS catalog²⁵, and summary statistics for infection-related phenotypes were provided by 23andMe⁶². For each phenotype, we used GWAS summary statistics from the largest GWAS available with a minimum coverage of 2 million SNPs. We excluded traits with estimated SNP-based heritability <1%. The genetic correlations were next estimated using bivariate LD score regression²¹ with ancestry-matched 1000 Genome Project Phase 3 as the LD reference panel. To additionally assess the effects of our GWAS loci on the intestinal microbiome, we interrogated independently significant lead SNPs from our GWAS in summary statistics for the largest available GWAS for microbiome abundance and diversity⁴⁹ by examining quantile-quantile plots and testing for an excess of significant associations with microbiome traits as compared to null expectation.

Pleiotropic annotation of GWAS loci

To systematically cross-annotate our loci against all previously published GWAS findings, we downloaded the GWAS catalogue (9/19 download)²⁵. We filtered all published SNPs that were associated with any disease phenotype or trait at a genome-wide significance ($P < 5 \times 10^{-8}$) and resided within the genomic regions of association with IgA level. For each SNP associations, we manually verified the direction of effect for a reference allele based on original GWAS publications. Next, each selected SNP from the catalogue was queried against our GWAS results to extract the odds ratios and *P*-values for associations with IgA levels. The directionality of allelic effects was assessed to identify pleiotropic alleles with concordant or opposed effects. We calculated a maximum r^2 between SNPs associated with each catalogued trait and the independent SNPs from our study based on 1000 Genomes Project Phase 3 data⁹³. We defined shared susceptibility alleles if

r^2 between the top SNPs exceeded 0.5. We constructed a susceptibility overlap map that connected each of the IgA GWAS loci to the previously associated GWAS traits and highlights associations with SNPs in high LD with the top signals. The map was visualized with Cytoscape v.3.6 software⁹⁴. Moreover, given a large number of overlapping GWAS loci between IgA levels, IgAN, and tonsillectomy, we performed systematic locus colocalization analyses based on regional summary statistics using Coloc software⁸⁹.

Meta-phenome-wide association study (Meta-PheWAS)

For the purpose of testing cumulative effects of genetic determinants of serum IgA levels on other traits, we first derived a genome-wide polygenic score (GPS) for IgA levels using LDpred algorithm⁹⁵ with GWAS summary statistics for IgA levels (assuming 1% fraction of causal variants) and using 1000 Genomes Phase 3 (all populations) for LD reference⁷⁶. We then performed a meta-PheWAS analysis for the GPS across the UK Biobank (UKBB, N=460,364 participants) and Electronic Medical Records and Genomics-III (eMERGE-III, N=96,292, after excluding those participants analyzed in the GWAS for IgA levels). The eMERGE-III genotype data was processed in the same way as for GWAS described above, but for ancestry adjustments we performed principal component analysis of the entire eMERGE-III cohort using FlashPCA⁹⁶ on a set of 48,509 common (MAF>0.01) and independent variants (pruned in PLINK with --indep-pairwise 500 50 0.05 command). The first 3 PCs were included as a covariate in PheWAS. The UKBB is a large prospective population-based cohort that enrolled individuals ages 40-69 for genetic studies⁹⁷. All 488,377 UKBB participants underwent genotyping with Affymetrix's UK Biobank Axiom and UK BiLEVE Axiom arrays with genotype imputation using a 1000 Genomes reference panel with IMPUTE4 software⁹⁸⁻¹⁰⁰. We applied similar QC filters to eMERGE-III, retaining 9,233,643 common (MAF $\geq$ 0.01) variants imputed with high confidence ($R^2 \geq$ 0.8). For principal component analysis using FlashPCA⁹⁶, we used a set of 35,226 variants with MAF>0.01 and pruned using --indep-pairwise 500 50 0.05 command in PLINK. The first 3 PCs were used as covariates in PheWAS. To harmonize coded diagnoses between UKBB and eMERGE-III we converted all available ICD-10 codes to ICD-9-CM system given that the great majority eMERGE-III diagnoses were coded using ICD-9-CM. After the conversion, eMERGE participants had a total of 20,783 ICD codes that were then mapped to 1,817 distinct phecodes. The 488,377 UKBB participants had a total of 10,221 ICD codes mapped to 1,523 phecodes. Phenome-wide associations were then performed using the PheWAS R package¹⁰¹. The

case definition required a minimum of two ICD-9 codes from the “case” grouping of each phecode, while “control” group had no ICD-9 codes relevant to the tested phecode. In total, 1,523 overlapping phecodes were tested in both UKBB and eMERGE-III using logistic regression after adjusting each analysis for age, sex, study site, genotyping batch, and 3 PCs of ancestry. The meta-PheWAS across both datasets was performed using metagen under fixed effects model in PheWAS R library¹⁰¹. To establish significant disease associations in PheWAS, we set the Bonferroni-corrected statistical significance threshold at 3.28×10^{-5} ($0.05/1,523$) correcting for 1,523 independent phecodes tested.

Mendelian randomization analyses

Two sample mendelian randomization (MR) analyses were performed using genetic variants as instruments to test the causal effects between an exposure and an outcome. Only SNPs independently associated with the exposure at a genome-wide significance ($P < 5 \times 10^{-8}$) were used as instruments in MR studies. We excluded HLA alleles from all instruments, since these alleles are likely to exhibit strong pleiotropic associations with a wide range of immune-related outcomes. For primary hypothesis testing, we used inverse variance weighted MR model under the assumption of balanced pleiotropy by meta-analyzing SNP specific Wald estimates using multiplicative random effects. The random effects model was chosen to account for any potential heterogeneity. Given a total of eight bi-directional MR tests performed between IgA levels and IgAN, tonsillectomy, Celiac disease, and BMI, we used a Bonferroni-corrected significance threshold $\alpha = 0.05/8 = 0.00625$. We additionally tested for the presence of directed horizontal pleiotropy using Egger test for non-zero intercept. Additional sensitivity analyses were performed by testing for outlier effects, and relaxing the assumption of balanced horizontal pleiotropy by using median-based estimator, mode-based estimator, and Egger regression methods. The testing was conducted using the TwoSampleMR package¹⁰².

Acknowledgements

We are grateful to all study participants for contributing DNA and serum samples for the purpose of our genetic studies. We would also like to acknowledge Dr. Andre Franke, Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany and Dr. Alexandra Zhernakova, Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands for

providing summary statistics for the intestinal microbiome GWAS. This work was funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant numbers R01-DK105124 (KK), RC2-DK116690 (KK), R01-DK078244 (JN, BAJ), and R01-DK082753 (AGG, JN, KK, BAJ) with additional support by R01-LM013061 (KK, CW), R01-LM006910 (GH), U01-HG008680 (KK), U01-AI152960 (KK), K25-DK128563 (AK), UL1-TR001873 (AK, KK), K01-DK106341 (CR), and R03-DK122194 (CR). JF and FE were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRU 5011 – Project-ID 445703531. TR is funded by R01-MD012467, R01-NS029993, R01-NS040807, U24-NS107267, U19-AG065169, UL1-TR002736, KL2-TR002737, and the Florida Department of Health. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, 75N92020D00005, 75N92020D00002, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, CA, USA) and the Broad Institute of Harvard and MIT (Boston, MA, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. Parts of this study have been conducted using the UKBB Resource under UKBB project ID number 41849. The eMERGE network is funded by the National Human Genome Research Institute (NHGRI) through the following grants: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008680 (Columbia University Health Sciences); U01HG008685 (Mass General Brigham); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children’s Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008684 (Children’s Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); U01HG008664 (Baylor College of Medicine); and U54MD007593 (Meharry Medical College). The NIDDK, NHLBI, and NHGRI had no role in the design of this study, data analysis or interpretation, or writing the manuscript.

Author Contributions

K. Kiryluk and A.G. Gharavi conceived the study, provided overall supervision of the project, and made the decision to publish the findings; L. Liu performed GWAS meta-analyses, conditional analyses, fine-mapping studies, functional annotations, and genetic correlation analyses. She also performed genotype quality control and primary GWAS analyses of the MESA, European, and U.S.-based discovery sub-cohorts. A. Khan performed genotype quality control and GWAS analyses in the eMERGE-III sub-cohorts, and PheWAS analyses in eMERGE-III and UKBB datasets. N. Shang analyzed IgA levels derived from electronic health records. E. Sanchez-Rodriguez and F. Zanoni contributed to the analyses of genetic correlations and overlapping susceptibility with IgA nephropathy. X. Zhou, H. Zhang, J. Xie and N. Chan coordinated recruitment and serum IgA measurements in the Chinese GWAS cohort. H. Suzuki coordinated recruitment and serum IgA measurements in the Japanese GWAS cohort. J. Floege, T. Rauen, and F. Eitner recruited and characterized the German GWAS cohort. B. A. Julian, R. J. Wyatt, J. Novak, and S. Hall recruited and characterized the U.S. GWAS cohort. J. Zhang and S. Hall managed the study database and DNA samples. P. Krithivasan performed genotype calling for microarray data. O. Balderes, R. A. LeDesma, C. Fischman, Y. Li, Z. Moldoveanu, C. Reily, J. T. Sheff, B. Knoppova, and J. Novak performed laboratory measurements of serum IgA levels. E. C. Oelsner and R. G. Barr contributed the MESA SHARe phenotype and genotype data. G. Hripacsak, C. Weng, M. Elkind, T. Rundek, J. Harley, H. Moncrieffe, L.C. Kottyan, B. Namjou-Khales, T.L. Walunas, R. Knevel, S. Raychaudhuri, E.W. Karlson, S. J. Hebring, J. C. Denny, I. B. Stanaway, and D. Crosslin contributed eMERGE-III phenotype and genotype data. L. Hammarström, A. Viktorin, P. Magnusson contributed summary statistics for the Swedish discovery cohort. I. Ionita-Laza consulted on the statistical analysis of genetic data. K. Kiryluk, A.G. Gharavi, N. Steers, and J. Novak provided biological interpretation of the GWAS loci. L. Liu and K. Kiryluk wrote the manuscript. All authors have read and approved the final version of manuscript.

Declaration of Interests

Dr. Kiryluk has served on an advisory board for Goldfinch Bio and Gilead Sciences. Dr. Gharavi has served on an advisory board for Novartis, Travere and Natera and receives research grant funding from the Renal Research Institute and Natera. Dr. Moncrieffe is presently employed by Janssen Pharmaceutical Companies

of Johnson & Johnson. Dr. Eitner is currently employed by Bayer AG. Drs. Julian and Novak are co-founders, co-owners of, and consultants for Reliant Glycosciences, LLC and are co-inventors on US patent application 14/318,082 (assigned to UAB Research Foundation). The other authors report no relevant conflicts of interest.

Ethics Statement

All subjects provided informed consent to participate in genetic studies, and the Institutional Review Board of Columbia University approved our studies under the following protocol numbers: IRB-AAAC7385 (primary analysis), IRB-AAQ9205 (eMERGE-III analysis), IRB-AAAC9458 (MESA SHARE analysis), and IRB-AAAS3500 (UK Biobank analysis).

Data Availability

The MESA SHARE genotype and phenotype data (including serum IgA levels measured in this study) are available through dbGAP, accession number phs000209.v13.p3. The Electronic Medical Records and Genomics-III (eMERGE-III) imputed genotype and phenotype data are available through dbGAP, accession number: phs001584.v2.p2. The UK Biobank genotype and phenotype data are available through the UK Biobank web portal. Genotype and phenotype data for all other cohorts is in the process of dbGAP submission (accession number pending). GWAS summary statistics will be available for download from http://www.columbiamedicine.org/divisions/kiryluk/study_gwas_stat.php at the time of publication.

Web Resources

COLOC: <https://cran.r-project.org/web/packages/coloc/>
 DEPICT: <https://data.broadinstitute.org/mpg/depict/index.html>
 DICE: <https://dice-database.org/>
 EAGLE: <https://data.broadinstitute.org/alkesgroup/Eagle/>
 eQTLGen: <https://www.eqtlgen.org/>
 FUN-LDA: <http://www.columbia.edu/~ii2135/funlda.html>
 GCTA-COJO: <https://cnsgenomics.com/software/gcta/#COJO>
 GeNets: <http://apps.broadinstitute.org/genets>
 GSEA: <http://software.broadinstitute.org/gsea/msigdb/>
 GWAS catalog: <https://www.ebi.ac.uk/gwas>

784 InWeb: <http://apps.broadinstitute.org/genets#InWeb> InBiomap
785 KING: <http://people.virginia.edu/~wc9c/KING/>
786 LD hub: <http://ldsc.broadinstitute.org/ldhub>
787 LDpred: <https://github.com/bvilhjal/ldpred>
788 LDSC: <https://github.com/bulik/ldsc>
789 Metabolomics GWAS Server: <http://metabolomics.helmholtz-muenchen.de/gwas/>
790 METAL: <http://csg.sph.umich.edu/abecasis/Metal/>
791 MGI: <http://www.informatics.jax.org>
792 MINIMAC3: <http://genome.sph.umich.edu/wiki/Minimac3>
793 MIP: <https://imputationserver.sph.umich.edu>
794 PheWAS: <https://github.com/PheWAS/PheWAS>
795 PheWeb: <http://pheweb.sph.umich.edu/SAIGE-UKB/>
796 PLINK: <https://www.cog-genomics.org/plink/1.9/>
797 RELI: <https://github.com/WeirauchLab/RELI>
798 STRING: <https://string-db.org>
799 ToppGene: <https://toppgene.cchmc.org>
800

References

1. Woof, J.M. & Kerr, M.A. The function of immunoglobulin A in immunity. *J Pathol* **208**, 270-82 (2006).
2. Fasching, C.E. *et al.* Impact of the molecular form of immunoglobulin A on functional activity in defense against *Streptococcus pneumoniae*. *Infect Immun* **75**, 1801-10 (2007).
3. Woof, J.M. & Russell, M.W. Structure and function relationships in IgA. *Mucosal Immunol* **4**, 590-7 (2011).
4. Corthesy, B. Role of secretory IgA in infection and maintenance of homeostasis. *Autoimmun Rev* **12**, 661-5 (2013).
5. Yu, H.Q. *et al.* Distinct features of SARS-CoV-2-specific IgA response in COVID-19 patients. *Eur Respir J* **56**(2020).
6. Ma, H. *et al.* Serum IgA, IgM, and IgG responses in COVID-19. *Cell Mol Immunol* **17**, 773-775 (2020).
7. Sterlin, D. *et al.* IgA dominates the early neutralizing antibody response to SARS-CoV-2. *Sci Transl Med* **13**(2021).
8. Maeda, A. *et al.* Significance of serum IgA levels and serum IgA/C3 ratio in diagnostic analysis of patients with IgA nephropathy. *J Clin Lab Anal* **17**, 73-6 (2003).
9. Papista, C., Berthelot, L. & Monteiro, R.C. Dysfunctions of the IgA system: a common link between intestinal and renal diseases. *Cell Mol Immunol* **8**, 126-34 (2011).
10. Rodriguez-Segade, S. *et al.* High serum IgA concentrations in patients with diabetes mellitus: age-wise distribution and relation to chronic complications. *Clin Chem* **42**, 1064-7 (1996).
11. Gonzalez-Quintela, A. *et al.* Serum levels of immunoglobulins (IgG, IgA, IgM) in a general adult population and their relationship with alcohol consumption, smoking and common metabolic abnormalities. *Clin Exp Immunol* **151**, 42-50 (2008).
12. Di Franco, P. *et al.* Genetic and environmental influences on serum levels of immunoglobulins and complement components in monozygotic and dizygotic twins. *J Clin Lab Immunol* **27**, 5-10 (1988).
13. Stoica, G., Macarie, E., Michiu, V. & Stoica, R.C. Biologic variation of human immunoglobulin concentration. I. Sex-age specific effects on serum levels of IgG, IgA, IgM and IgD. *Med Interne* **18**, 323-32 (1980).
14. Grundbacher, F.J. & Shreffler, D.C. Changes in human serum immunoglobulin levels with age and sex. *Z Immunitätsforsch Allerg Klin Immunol* **141**, 20-6 (1970).
15. Lomax-Browne, H.J. *et al.* IgA1 Glycosylation Is Heritable in Healthy Twins. *J Am Soc Nephrol* **28**, 64-68 (2017).
16. Viktorin, A. *et al.* IgA measurements in over 12 000 Swedish twins reveal sex differential heritability and regulatory locus near CD30L. *Hum Mol Genet* **23**, 4177-84 (2014).
17. Hatagima, A., Cabello, P.H. & Krieger, H. Causal analysis of the variability of IgA, IgG, and IgM immunoglobulin levels. *Hum Biol* **71**, 219-29 (1999).
18. Grundbacher, F.J. Heritability estimates and genetic and environmental correlations for the human immunoglobulins G, M, and A. *Am J Hum Genet* **26**, 1-12 (1974).
19. Yang, C. *et al.* Genome-wide association study identifies TNFSF13 as a susceptibility gene for IgA in a South Chinese population in smokers. *Immunogenetics* **64**, 747-53 (2012).

- 844 20. Jonsson, S. *et al.* Identification of sequence variants influencing immunoglobulin levels.
845 *Nat Genet* **49**, 1182-1191 (2017).
- 846 21. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from
847 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
- 848 22. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**,
849 760-764 (2016).
- 850 23. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list
851 enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-11
852 (2009).
- 853 24. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using
854 predicted gene functions. *Nat Commun* **6**, 5890 (2015).
- 855 25. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
856 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-
857 D1012 (2019).
- 858 26. Ingram, D.G. & Friedman, N.R. Toward Adenotonsillectomy in Children: A Review for the
859 General Pediatrician. *JAMA Pediatr* **169**, 1155-61 (2015).
- 860 27. Shilatifard, A. *et al.* ELL2, a new member of an ELL family of RNA polymerase II
861 elongation factors. *Proc Natl Acad Sci U S A* **94**, 3639-43 (1997).
- 862 28. Vosa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci
863 and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300-1310
864 (2021).
- 865 29. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene
866 Expression. *Cell* **175**, 1701-1715 e16 (2018).
- 867 30. Wahl, A. *et al.* Genome-Wide Association Study on Immunoglobulin G Glycosylation
868 Patterns. *Front Immunol* **9**, 277 (2018).
- 869 31. Durst, K.L. & Hiebert, S.W. Role of RUNX family members in transcriptional repression
870 and gene silencing. *Oncogene* **23**, 4220-4 (2004).
- 871 32. Ito, Y., Bae, S.C. & Chuang, L.S. The RUNX family: developmental regulators in cancer.
872 *Nat Rev Cancer* **15**, 81-95 (2015).
- 873 33. Watanabe, K. *et al.* Requirement for Runx proteins in IgA class switching acting
874 downstream of TGF-beta 1 and retinoic acid signaling. *J Immunol* **184**, 2785-92 (2010).
- 875 34. Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with
876 multiple myeloma. *Nat Commun* **6**, 7213 (2015).
- 877 35. Martinelli, N. *et al.* FADS genotypes and desaturase activity estimated by the ratio of
878 arachidonic acid to linoleic acid are associated with inflammation and coronary artery
879 disease. *Am J Clin Nutr* **88**, 941-9 (2008).
- 880 36. Ameer, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific
881 haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids.
882 *Am J Hum Genet* **90**, 809-20 (2012).
- 883 37. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed
884 Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
- 885 38. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human
886 Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).

- 887 39. Graham, D.B. *et al.* TMEM258 Is a Component of the Oligosaccharyltransferase Complex
888 Controlling ER Stress and Intestinal Inflammation. *Cell Rep* **17**, 2955-2965 (2016).
- 889 40. Bushell, K.N. *et al.* LITAF mediation of increased TNF-alpha secretion from inflamed
890 colonic lamina propria macrophages. *PLoS One* **6**, e25849 (2011).
- 891 41. Stucchi, A. *et al.* A new transcription factor that regulates TNF-alpha gene expression,
892 LITAF, is increased in intestinal tissues from patients with CD and UC. *Inflamm Bowel Dis*
893 **12**, 581-7 (2006).
- 894 42. Tang, X., Metzger, D., Leeman, S. & Amar, S. LPS-induced TNF-alpha factor (LITAF)-
895 deficient mice express reduced LPS-induced cytokine: Evidence for LITAF-dependent LPS
896 signaling pathways. *Proc Natl Acad Sci U S A* **103**, 13777-82 (2006).
- 897 43. He, J.Q., Saha, S.K., Kang, J.R., Zarnegar, B. & Cheng, G. Specificity of TRAF3 in its
898 negative regulation of the noncanonical NF-kappa B pathway. *J Biol Chem* **282**, 3688-94
899 (2007).
- 900 44. Bista, P. *et al.* TRAF3 controls activation of the canonical and alternative NFkappaB by
901 the lymphotoxin beta receptor. *J Biol Chem* **285**, 12971-8 (2010).
- 902 45. Chen, Z. *et al.* TRAF3 Acts as a Checkpoint of B Cell Receptor Signaling to Control
903 Antibody Class Switch Recombination and Anergy. *J Immunol* **205**, 830-841 (2020).
- 904 46. Xie, P., Kraus, Z.J., Stunz, L.L., Liu, Y. & Bishop, G.A. TNF receptor-associated factor 3 is
905 required for T cell-mediated immunity and TCR/CD28 signaling. *J Immunol* **186**, 143-55
906 (2011).
- 907 47. Arkee, T., Hostager, B.S., Houtman, J.C.D. & Bishop, G.A. TRAF3 in T Cells Restrains
908 Negative Regulators of LAT to Promote TCR/CD28 Signaling. *J Immunol* (2021).
- 909 48. Kasler, H.G., Lee, I.S., Lim, H.W. & Verdin, E. Histone Deacetylase 7 mediates tissue-
910 specific autoimmunity via control of innate effector function in invariant Natural Killer T
911 Cells. *Elife* **7**(2018).
- 912 49. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D
913 receptor and other host factors influencing the gut microbiota. *Nat Genet* **48**, 1396-1406
914 (2016).
- 915 50. Sonar, S. & Lal, G. Role of Tumor Necrosis Factor Superfamily in Neuroinflammation and
916 Autoimmunity. *Front Immunol* **6**, 364 (2015).
- 917 51. Vallabhapurapu, S. & Karin, M. Regulation and function of NF-kappaB transcription
918 factors in the immune system. *Annu Rev Immunol* **27**, 693-733 (2009).
- 919 52. Lauc, G. *et al.* Loci associated with N-glycosylation of human immunoglobulin G show
920 pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet* **9**,
921 e1003225 (2013).
- 922 53. Lopez de Lapuente, A. *et al.* Novel Insights into the Multiple Sclerosis Risk Gene
923 ANKRD55. *J Immunol* **196**, 4553-65 (2016).
- 924 54. Dinarello, C.A. Overview of the IL-1 family in innate inflammation and acquired
925 immunity. *Immunol Rev* **281**, 8-27 (2018).
- 926 55. Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA
927 targets. *Nucleic Acids Res* **48**, D127-D131 (2020).
- 928 56. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target
929 sites in mammalian mRNAs. *Elife* **4**(2015).

- 930 57. Schmitz, N., Kurrer, M. & Kopf, M. The IL-1 receptor 1 is critical for Th2 cell type airway
931 immune responses in a mild but not in a more severe asthma model. *Eur J Immunol* **33**,
932 991-1000 (2003).
- 933 58. Hodson, D.J. *et al.* Regulation of normal B-cell differentiation and malignant B-cell
934 survival by OCT2. *Proc Natl Acad Sci U S A* **113**, E2039-46 (2016).
- 935 59. Bult, C.J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res* **47**, D801-D806
936 (2019).
- 937 60. Chou, Y.T. *et al.* CITED2 functions as a molecular switch of cytokine-induced
938 proliferation and quiescence. *Cell Death Differ* **19**, 2015-28 (2012).
- 939 61. Batlle, E. & Massague, J. Transforming Growth Factor-beta Signaling in Immunity and
940 Cancer. *Immunity* **50**, 924-940 (2019).
- 941 62. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify
942 susceptibility loci for multiple common infections. *Nat Commun* **8**, 599 (2017).
- 943 63. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits.
944 *Nat Genet* **47**, 1236-41 (2015).
- 945 64. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare
946 variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
- 947 65. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity
948 biology. *Nature* **518**, 197-206 (2015).
- 949 66. Bonder, M.J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**,
950 1407-1412 (2016).
- 951 67. Cheung, C.K., Rajasekaran, A., Barratt, J. & Rizk, D.V. An Update on the Current State of
952 Management and Clinical Trials for IgA Nephropathy. *J Clin Med* **10**(2021).
- 953 68. Brahe, L.K. *et al.* Specific gut microbiota features and metabolic markers in
954 postmenopausal women with obesity. *Nutr Diabetes* **5**, e159 (2015).
- 955 69. Companys, J. *et al.* Gut Microbiota Profile and Its Association with Clinical Variables and
956 Dietary Intake in Overweight/Obese and Lean Subjects: A Cross-Sectional Study.
957 *Nutrients* **13**(2021).
- 958 70. Forbes, J.D. *et al.* A comparative study of the gut microbiota in immune-mediated
959 inflammatory diseases-does a common dysbiosis exist? *Microbiome* **6**, 221 (2018).
- 960 71. Huus, K.E., Petersen, C. & Finlay, B.B. Diversity and dynamism of IgA-microbiota
961 interactions. *Nat Rev Immunol* **21**, 514-525 (2021).
- 962 72. Bild, D.E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J*
963 *Epidemiol* **156**, 871-81 (2002).
- 964 73. Zhao, X. *et al.* Whole genome sequence analysis of pulmonary function and COPD in
965 19,996 multi-ethnic participants. *Nat Commun* **11**, 5182 (2020).
- 966 74. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**,
967 1284-1287 (2016).
- 968 75. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium
969 panel. *Nat Genet* **48**, 1443-1448 (2016).
- 970 76. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**,
971 68-74 (2015).
- 972 77. Galinsky, K.J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of
973 ADH1B in Europe and East Asia. *Am J Hum Genet* **98**, 456-472 (2016).

974 78. Stanaway, I.B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to ~40
975 million variants genome wide and association with the herpes zoster medical record
976 phenotype. *Genet Epidemiol* **43**, 63-81 (2019).

977 79. Khan, A. *et al.* Medical Records-Based Genetic Studies of the Complement System. *J Am*
978 *Soc Nephrol* (2021).

979 80. Shang, N. *et al.* Medical records-based chronic kidney disease phenotype for clinical care
980 and "big data" observational and genetic studies. *Npj Digital Medicine* **4**(2021).

981 81. Kiryluk, K. *et al.* GWAS for serum galactose-deficient IgA1 implicates critical genes of the
982 O-glycosylation pathway. *PLoS Genet* **13**, e1006609 (2017).

983 82. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances
984 evidence of functional enrichment. *Bioinformatics* **30**, 2906-14 (2014).

985 83. Hoffmann, T.J. *et al.* Genome-wide association analyses using electronic health records
986 identify new loci influencing blood pressure variation. *Nat Genet* **49**, 54-64 (2017).

987 84. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts
988 complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).

989 85. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide
990 association studies. *Nat Genet* **38**, 904-9 (2006).

991 86. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of
992 genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

993 87. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics
994 identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3
995 (2012).

996 88. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants
997 from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

998 89. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic
999 association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

1000 90. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic
1001 interpretation. *Nat Methods* **14**, 61-64 (2017).

1002 91. Kaimal, V., Bardes, E.E., Tabar, S.C., Jegga, A.G. & Aronow, B.J. ToppCluster: a multiple
1003 gene list feature analyzer for comparative enrichment clustering and network-based
1004 dissection of biological systems. *Nucleic Acids Res* **38**, W96-102 (2010).

1005 92. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score
1006 regression that maximizes the potential of summary level GWAS data for SNP
1007 heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).

1008 93. Kiryluk, K. *et al.* Discovery of new risk loci for IgA nephropathy implicates genes involved
1009 in immunity against intestinal pathogens. *Nat Genet* **46**, 1187-96 (2014).

1010 94. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
1011 biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

1012 95. Vilhjalmsen, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic
1013 Risk Scores. *American Journal of Human Genetics* **97**, 576-592 (2015).

1014 96. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide
1015 data. *PLoS One* **9**, e93766 (2014).

1016 97. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a
1017 wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).

1018 98. Altshuler, D.M. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-+
1019 (2015).
1020 99. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate
1021 genotype imputation in genome-wide association studies through pre-phasing. *Nature*
1022 *Genetics* **44**, 955-+ (2012).
1023 100. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
1024 *Nature* **562**, 203-+ (2018).
1025 101. Denny, J.C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to
1026 discover gene-disease associations. *Bioinformatics* **26**, 1205-10 (2010).
1027 102. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the
1028 human phenome. *Elife* **7**(2018).
1029

1030 **Table 1 Baseline characteristics of participants in the GWAS cohorts.**

Cohorts	Ancestry	No. of Individuals	N _{Male}	N _{Female}	Mean Age
MESA (European)	European	2,280	1132	1148	64.17
MESA (African)	African	1,275	599	676	64.05
MESA (Admixed 1)	Admixed	474	209	265	63.38
MESA (Admixed 2)	Admixed	750	369	381	63.17
MESA (East Asian)	East Asian	641	319	322	62.48
eMERGE (European)	European	4,261	1622	2640	56.45
eMERGE (African)	African	476	326	150	50.06
eMERGE (East Asian)	East Asian	73	29	44	45.38
eMERGE (Admixed 1)	Admixed	235	84	151	45.63
eMERGE (Admixed 2)	Admixed	1,002	427	575	54.97
German (European)	European	156	104	52	44.88
French (European)	European	103	30	73	--
Chinese (European)	East Asian	467	318	149	32.39
Japanese (European)	East Asian	776	523	252	33.79
U.S. (European)	European	93	53	40	35.66
Swedish (PMID: 24676358)	European	9,167	4361	4806	64.50
Decode (PMID: 28628107)	European	19,034	--	--	--
All Discovery		41,263	6144	6918	51.18

1031

Table 2 Effect estimates for independently associated variants with serum IgA levels by trans-ethnic meta-analysis.

Locus	CHR	BP (hg19)	SNP	Effect allele	BETA	P value	Note
<i>RUNX3</i>	1	25291697	rs188468174	C	0.876	3.42E-92	Known
<i>TNFSF4/TNFSF18</i>	1	173163568	rs7518129	G	0.056	1.06E-16	Known
	1	173172158	rs4916314	A	0.061	3.46E-14	
<i>GPATCH2</i>	1	217563106	rs73100295	T	0.356	3.91E-08	New
<i>IL1R1</i>	2	102689031	rs13427957	C	0.040	6.19E-10	New
<i>ELL2</i>	5	95277555	rs3777175	G	0.084	7.81E-30	Known
	5	95231115	rs3777207	T	0.080	1.16E-25	
<i>ANKRD55/IL6ST</i>	5	95286609	rs200990904	C	0.065	2.02E-14	Known
	5	55438580	rs6859219	C	0.073	1.41E-20	
	6	31106893	rs1265094	A	0.076	1.86E-31	
	6	32710119	rs5018343	A	0.070	2.94E-24	
<i>HLA</i>	6	32449523	chr6:32449523	T	0.078	4.32E-25	Known
	6	29943715	rs3823363	C	0.062	3.13E-22	
	6	31438243	rs2596450	G	0.058	1.33E-17	
	6	32804299	rs3819717	C	0.047	1.01E-13	
	6	31324152	rs1050628	G	0.126	3.56E-16	
	6	32343369	rs3117106	T	0.073	1.44E-12	
<i>RUNX2</i>	6	45526470	rs1200427	T	0.059	6.85E-14	New
<i>CITED2</i>	6	139975943	rs17069163	T	0.050	5.94E-11	New
<i>ZP3/SSC4D</i>	7	76034150	rs55722505	C	0.048	8.61E-14	New
<i>TNFSF8/TNFSF15</i>	9	117692882	rs3181356	T	0.069	1.13E-18	Known
<i>FADS2/FADS1</i>	11	61595564	rs968567	T	0.116	2.42E-41	Known
	11	61557803	rs102275	C	0.073	3.60E-29	
<i>OVOL1/RELA</i>	11	65555524	rs10896045	A	0.066	2.57E-22	New
<i>POU2AF1</i>	11	111267394	rs4938518	T	0.056	7.01E-16	New
<i>HDAC7, VDR</i>	12	48214825	rs7487637	G	0.054	9.97E-15	New
<i>SH2B3</i>	12	111833788	rs10774624	G	0.050	1.37E-13	Known
<i>RCOR1/TRAFF3</i>	14	103239630	rs12147883	C	0.048	5.42E-14	New
<i>LITAF</i>	16	11717832	rs113962704	T	0.054	1.91E-12	New
<i>CTF1</i>	16	30916129	rs1458201	A	0.052	2.02E-11	New
<i>HORMAD2/LIF</i>	22	30448399	rs193473	A	0.067	1.58E-20	Known
	22	30269907	rs9614090	A	0.052	5.56E-16

Figure 1 | Ancestral differences in serum IgA levels: **a.** discovery analysis across the four major ancestral groups in MESA demonstrates that African ancestry is associated with higher adjusted IgA levels, **b.** replication analysis in all non-MESA study participants confirms higher mean IgA levels in individuals of African ancestry, **c.** mean adjusted IgA levels ($\pm$ 95% confidence intervals) as a function of African ancestry fraction, demonstrating that individuals in the upper quartile ($>75\%$) of African ancestry have the highest serum IgA levels (N=5420 MESA participants); standardized residuals generated by regression of log-transformed serum IgA levels against age and sex were significantly correlated with the African ancestry fraction ($P=4.6 \times 10^{-33}$), and this relationship remained highly significant after additional adjustment for BMI and diabetes ($P=3.7 \times 10^{-23}$). The boxplots depict the median (horizontal line), upper/lower quartiles (boxes), and range (whiskers); the red diamond point denotes the mean value per ancestry group; * $P<0.05$; ** $P<0.01$; *** $P<0.001$.

Figure 3 | Colocalization and enrichment analyses across different tissue/cell types. (a) Colocalization analysis with gene expression QTLs (eQTLs) in whole blood and primary immune cells; each row indicates one gene and each column denotes one tissue/cell type; the color from red to white shows the probability of sharing the same causal variant between GWAS loci and eQTLs (PP4). **(b)** Tissue/cell type enrichment in DEPICT; the y-axis represents the $-\log_{10}$ of the p-value and x-axis shows the first level MeSH tissue and cell type annotations. **(c)** Integrative analysis of eQTL and hQTL for *LITAF* locus. The upper and lower panels show the regional plots of the *LITAF* locus for IgA GWAS and eQTLs in monocyte, respectively. The y-axis represents the $-\log_{10}$ of the p-value and x-axis shows the chromosome positions. The lowest panel denotes the positions of *LITAF* gene and a hQTL peak (H3K27ac) in monocyte.

Figure 4 | Pathway and gene set enrichment analyses. (a) Pathway enrichment analysis for genes at the significant GWAS loci; (b) Gene-set enrichment for genes that cause abnormal IgA level; (c) abnormal immune tolerance; and (d) abnormal response to infection when genetically manipulated in mice. The y axis shows $-\log_{10}(P \text{ value})$ for association with serum IgA in the trans-ancestry meta-analysis for the variant with the lowest P value in each candidate gene. The dashed line corresponds to genome-wide significance ($P=5 \times 10^{-8}$). Enrichment P values correspond to the observed number of genes with association signals below the genome-wide threshold against the number expected under binomial distribution.

Figure 5 | Colocalization and Mendelian randomization analyses based on GWAS for serum IgA levels, IgA nephropathy and tonsillectomy. (a) Regional plot of the *HORMAD2/LIF* locus for IgA levels (top panel), IgA nephropathy (middle panel), and tonsillectomy (lower panel). (b) Co-localization analysis of *HORMAD2/LIF* locus across the three traits; PP4 is the posterior probability of co-localization. (c) Mendelian randomization analysis using IgA level as an exposure, IgA nephropathy as an outcome, and co-localizing loci as instruments. The x and y axis represent effect sizes of the genetic variants associated with the exposure and outcome, respectively. IVW: inverse variance weighted.

Figure 6 | Genetic relationships between IgA levels and human disease traits. (a) genome-wide genetic correlation analyses between IgA levels and autoimmune, infectious, and cardio-metabolic traits after exclusion of the HLA region ($*P < 0.05$). Supplementary Table 14 provides a comparison of genetic correlations with and without HLA. **(b)** Meta-PheWAS of genome-wide polygenic score (GPS) for IgA levels across eMERGE-III and UKBB biobanks (total $N = 556,656$). The y axis shows $-\log_{10}(P\text{-value})$ and is truncated to accommodate the top association with celiac disease. The red line corresponds to the Bonferroni-corrected significance threshold for 1,523 phecodes tested ($\alpha = 0.05/1,523 = 3.28 \times 10^{-5}$). Each triangle represents an individual phenotype (phecode) tested as an outcome against the GPS for IgA levels as a predictor; an upward triangle indicates a positive (risk) association, while a downward triangle indicates a negative (protective) association. All associations are adjusted for age, sex, site, genotyping batch, and principal components of ancestry. The phenotypes are grouped by organ system (or relevant disease category) and sorted based on their statistical significance within each group. Supplementary Table 15 provides a comparison of significant PheWAS associations with and without HLA.

Figure 7 | Genetic relationship between IgA levels and human gut microbiome. Quantile-quantile plots for associations of 20 significant IgA SNPs with (a) overall microbiome diversity; (b) abundance of each of the 123 species; (c) abundance of each of the 52 genera. The x axis shows the expected $-\log_{10}(P \text{ value})$, and the y axis represents the observed $-\log_{10}(P \text{ value})$.