

GWAS for serum galactose-deficient IgA1 implicates critical genes of the *O*-glycosylation pathway

Krzysztof Kiryluk¹, Yifu Li¹, Zina Moldoveanu², Hitoshi Suzuki³, Colin Reily², Ping Hou⁴, Jingyuan Xie⁵, Nikol Mladkova¹, Sindhuri Prakash¹, Clara Fischman¹, Samantha Shapiro¹, Robert A. LeDesma¹, Drew Bradbury¹, Iuliana Ionita-Laza⁶, Frank Eitner^{7,8}, Thomas Rauen⁷, Nicolas Maillard⁹, Francois Berthoux⁹, Jürgen Floege⁷, Nan Chen⁵, Hong Zhang⁴, Francesco Scolari^{10,11}, Robert J. Wyatt^{12,13}, Bruce A. Julian¹⁴, Ali G. Gharavi¹ and Jan Novak²

Affiliations:

1. Dept. of Medicine, Div. of Nephrology, College of Physicians and Surgeons, Columbia University, New York, New York, USA
2. Dept. of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama, USA
3. Division of Nephrology, Dept. of Internal Medicine, Juntendo University Faculty of Medicine, Tokyo, Japan
4. Renal Div., Peking University First Hospital, Peking University Institute of Nephrology, Beijing, China
5. Dept. of Nephrology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
6. Dept. of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York, USA
7. Dept. of Nephrology, RWTH University of Aachen, Aachen, Germany
8. Kidney Diseases Research, Bayer Pharma AG, Wuppertal, Germany
9. Nephrology, Dialysis, and Renal Transplantation Dept., University North Hospital, Saint Etienne, France
10. Div. of Nephrology, Azienda Ospedaliera Spedali Civili of Brescia, Montichiari Hospital, Univ of Brescia, Brescia, Italy
11. Dept. of Medical and Surgical Specialties, Radiological Sciences, University of Brescia, Brescia, Italy
12. Div. of Pediatric Nephrology, University of Tennessee Health Sciences Center, Memphis, Tennessee, USA
13. Children's Foundation Research Institute, Le Bonheur Children's Hospital, Memphis, Tennessee, USA
14. Dept. of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA.

Corresponding Author:

Krzysztof Kiryluk M.D., M.S.
 Department of Medicine
 Division of Nephrology
 Columbia University
 1150 St Nicholas Ave
 Russ Berrie Pavilion #412
 New York, NY 10032
 Tel: 212-851-4926
kk473@columbia.edu

Abstract:

Aberrant *O*-glycosylation of serum immunoglobulin A1 (IgA1) represents a heritable pathogenic defect in IgA nephropathy, the most common form of glomerulonephritis worldwide, but specific genetic factors involved in its determination are not known. We performed a quantitative GWAS for serum levels of galactose-deficient IgA1 (Gd-IgA1) in 2,633 subjects of European and East Asian ancestry and discovered two genome-wide significant loci, in *CIGALT1* (rs13226913, $P = 3.2 \times 10^{-11}$) and *CIGALTIC1* (rs5910940, $P = 2.7 \times 10^{-8}$). These genes encode molecular partners essential for enzymatic *O*-glycosylation of IgA1. We demonstrated that these two loci explain approximately 7% of variability in circulating Gd-IgA1 in Europeans, but only 2% in East Asians. Notably, the Gd-IgA1-increasing allele of rs13226913 is common in Europeans, but rare in East Asians. Moreover, rs13226913 represents a strong cis-eQTL for *CIGALT1*, which encodes the key enzyme responsible for the transfer of galactose to *O*-linked glycans on IgA1. By *in vitro* siRNA knock-down studies, we confirmed that mRNA levels of both *CIGALT1* and *CIGALTIC1* determine the rate of secretion of Gd-IgA1 in IgA1-producing cells. Our findings provide novel insights into the genetic regulation of *O*-glycosylation and are relevant not only to IgA nephropathy, but also to other complex traits associated with *O*-glycosylation defects, including inflammatory bowel disease, hematologic disease, and cancer.

Author Summary:

O-glycosylation is a common type of post-translational modification of proteins; specific abnormalities in the mechanism of *O*-glycosylation have been implicated in cancer, inflammatory and blood diseases. However, the molecular basis of abnormal *O*-glycosylation in these complex disorders is not known. We studied the genetic basis of defective *O*-glycosylation of serum Immunoglobulin A1 (IgA1), which represents the key pathogenic defect in IgA nephropathy, the most common form of primary glomerulonephritis worldwide. We report our results of the first genome-wide association study for this trait using serum assays in 2,633 individuals of European and East Asian ancestry. In our genome scan, we observed two significant signals with large effects, on chromosomes 7p21.3 and Xq24, jointly explaining about 7% of trait variability. These signals implicate two genes that encode molecular partners essential for enzymatic *O*-glycosylation of IgA1 and mucins, and represent potential new targets for therapy.

Introduction:

N- and *O*-glycosylation is a fundamental post-translational modification of proteins in mammalian cells. Abnormalities in glycosylation have been linked to a broad range of human diseases, including neurologic disorders, immune-mediated and inflammatory diseases as well as cancer. Protein glycosylation is mediated by a large family of enzymes that have cell and tissue specific activity, and can generate highly diverse glycan structures that are important for signaling, cell-cell and cell-matrix interactions. The combinatorial possibilities of glycan structures imparted by the large number of glycosylation enzymes complicate a systematic analysis of protein glycosylation patterns and identification of critical steps involved in the activity, concentration, and regulation in any given cell or tissue. In such setting, genetic studies of congenital defects of glycosylation in humans have provided significant insight into non-redundant regulatory nodes in this pathway. The majority of these Mendelian disorders arise from loss of function mutations that severely perturb protein glycosylation across a range of tissues and produce a wide range of organ dysfunction in early life. However, less pronounced abnormalities in protein glycosylation have also been detected in complex disorders such as autoimmunity and cancer, suggesting that more subtle defects in this pathway can have important consequences for human health.

IgA nephropathy (IgAN), the most common cause of glomerulonephritis and a common cause of kidney failure worldwide, is a prototypical example of an immune-mediated disorder characterized by abnormal glycosylation¹. In humans, the hinge region heavy chains of immunoglobulin A1 (IgA1) are *O*-glycosylated with a variety of glycoforms in circulation. In healthy individuals, the prevailing glycoforms include the *N*-acetylgalactosamine (GalNAc)-galactose disaccharide and its sialylated forms. In IgAN, galactose-deficient IgA1 (Gd-IgA1) glycoforms are significantly more abundant compared to healthy controls². These under-galactosylated glycoforms are secreted by IgA1-producing cells while galactosylation of other circulating *O*-glycosylated proteins is preserved, suggesting a specific defect within IgA1-producing cells³. The pathogenetic mechanism of IgAN involves an autoimmune response resulting in production of IgA or IgG autoantibodies against circulating Gd-IgA1, and formation of immune complexes (Gd-IgA1 complexed with IgA/IgG autoantibody) that deposit in

the kidney and cause tissue injury^{1,4}. Consistent with this mechanism, Gd-IgA1 is the predominant glycoform in circulating immune complexes and in the glomerular immune-deposits in patients with IgAN⁵⁻⁸ and elevated serum levels of Gd-IgA1 (auto-antigen) and anti-glycan antibodies (auto-antibody) are associated with more aggressive disease and accelerated progression to end-stage kidney failure^{9,10}.

The design of a simple lectin-based ELISA assay, using a GalNAc-specific lectin from *Helix aspersa* (HAA), enables high-throughput screening of sera to quantify the levels of circulating Gd-IgA1². Using this assay, we have demonstrated that the serum levels of Gd-IgA1 represent a normally distributed quantitative trait in healthy populations, but up to two thirds of IgAN patients have levels above the 95th percentile for healthy controls. Examining family members of probands with familial and sporadic forms of IgAN, we also showed that elevated serum Gd-IgA1 levels segregate independently of total IgA levels and have high heritability (estimated at 50-70%)^{11,12}. Moreover, many healthy family members exhibited very high Gd-IgA1 levels, identifying elevated Gd-IgA1 as a heritable risk factor that precedes the development of IgAN.

To date, multiethnic genome-wide association studies involving over 20,000 individuals have identified 15 risk loci predisposing to IgAN, highlighting the importance of innate and adaptive immunity in this disorder. Power analyses indicated that discovery of additional risk loci using the case control design will require significant expansion in sample size. However, a systematic analysis of quantitative endophenotypes that are linked to disease pathogenesis, such as Gd-IgA1, has not been conducted to date and may provides the opportunity to discover additional pathogenic pathways using a smaller sample size. In this study, we performed the first GWAS for serum Gd-IgA1 levels, and successfully mapped new loci with surprisingly large contributions to the heritability of the circulating level of Gd-IgA1 independently of IgA levels.

Results

In order to test if serum levels of Gd-IgA1 remain stable over time, we first performed measurements of total serum immunoglobulin levels along with Gd-IgA1 levels at baseline and at four years of follow-up in 32 individuals of European ancestry followed longitudinally (**Figure 1**). While total IgG and IgA levels varied with time, Gd-IgA1 levels (normalized for total IgA) remained remarkably stable over a 4-year period ($r^2 = 0.92$, $P = 1.8 \times 10^{-13}$), demonstrating that *O*-glycosylation of IgA1 is minimally affected by random environmental factors.

We next used HAA lectin-based ELISA to analyze single time-point sera of 1,195 individuals in our discovery cohorts composed of 950 individuals of East Asian ancestry (483 biopsy-documented IgAN cases and 467 controls) and 245 individuals of European ancestry (141 biopsy-documented IgAN cases and 104 controls, **Table 1**). As previously reported, serum Gd-IgA1 levels were positively correlated with age (East Asians $r = 0.13$, $P = 8.9 \times 10^{-5}$; Europeans $r = 0.15$, $P = 1.7 \times 10^{-2}$) and total IgA levels (East Asians $r = 0.75$, $P < 2.2 \times 10^{-16}$; Europeans $r = 0.56$, $P < 2.2 \times 10^{-16}$), but were independent of gender ($P > 0.05$). In both cohorts, Gd-IgA1 levels were also significantly higher in IgAN cases compared to controls independently of age and total IgA levels (adjusted $P < 2.2 \times 10^{-16}$ in each individual cohort), providing a large-scale replication of prior findings.

We next performed a GWAS for serum levels of Gd-IgA1 in these cohorts with and without adjustment for total IgA levels. For genome-wide analysis, we used a linear model with individual SNPs coded as additive genetic predictors, and the outcome defined as standardized residuals of serum Gd-IgA1 after normalization and additional adjustment for case/control status, age, ancestry and cohort membership (see Methods). Each ethnicity-defined discovery cohort was analyzed separately and the results were meta-analyzed to prioritize top signals for follow-up. With this approach, we observed minimal genomic inflation in the combined genome-wide analyses ($\lambda=1.01$), indicating negligible effect of population stratification.

We first examined potential associations with known IgAN susceptibility loci, but found no statistically significant or suggestive signals between Gd-IgA1 levels and known IgAN risk alleles (**Supplemental Table 1**). In addition, we found no association between the global polygenetic risk score for IgAN, which captures the combined effect of all IgAN risk loci, and Gd-IgA1 levels. We also did not detect any associations of Gd-IgA1 levels with loci previously linked to variation in total IgA levels¹³⁻¹⁵, IgA deficiency¹⁶ or *N*-glycosylation of IgG¹⁷. At the same time, we replicated previously reported association of total IgA with *ELL2* (rs56219066, $P=8.5 \times 10^{-3}$)¹⁴, confirming that genetic regulation of IgA levels is distinct from Gd-IgA1 levels. These data thus indicated the presence of yet undiscovered loci controlling variation in Gd-IgA1 levels.

We next examined genome-wide distribution of P-values from the discovery stage to identify novel loci associated with Gd-IgA1 levels. Although no signals reached genome-wide significance in the discovery stage, we observed a number of suggestive ($P < 5 \times 10^{-4}$) loci that we followed up in 1,438 additional individuals of East Asian (N=653) and European (N=785) ancestry (**Supplementary Figure 1**). Subsequently, we analyzed all cohorts (N=2,633) jointly to identify genome-wide significant loci (**Table 2, Supplementary Table 2**). Our power calculations demonstrate that our design provides adequate power to detect variants explaining $\geq 1.5\%$ of overall trait variance at a genome-wide significant alpha 5×10^{-8} (**Supplementary Table 3**).

In the combined analysis, two distinct genomic loci, on chromosomes 7p21.3 and Xq24, reached genome-wide significance (**Figure 2a**). The strongest association was located within a 200-kb interval on chromosome 7p21.3 (**Figure 2b**), explaining 4% of trait variance in Europeans and ~1% in Asians (**Supplementary Table 4**). The only gene within this locus is *CIGALT1*, encoding core 1 synthase, glycoprotein-*N*-acetylgalactosamine 3-beta-galactosyltransferase 1. The top signal was represented by rs13226913 ($P=3.2 \times 10^{-11}$), an intronic SNP within *CIGALT1*. This locus is further supported by rs1008897 ($P=9.1 \times 10^{-10}$) in partial LD with rs13226913 ($r^2=0.33$, $D'=0.91$ in Europeans and $r^2=0.52$, $D'=0.73$ in Asians). After mutual conditioning, both SNPs continue to be associated with the phenotype, suggesting a complex pattern of association at this locus (**Supplementary Table 5**).

The protein encoded by *CIGALT1* generates the common core 1 *O*-glycan structure by transferring galactose (Gal) from UDP-Gal to GalNAc- α -1-Ser/Thr. Core 1 is a precursor for *O*-glycans in the hinge region of circulating IgA1, as well as many extended mucin-type *O*-glycans on cell surfaces. In humans, *CIGALT1* is abundantly expressed in IgA1-secreting cells¹⁸, as well as in EBV-transformed lymphocytes, gastrointestinal tract, lungs, and kidneys¹⁹. The top SNP, rs13226913, is not in LD with any coding variants, but it perfectly tags several SNPs intersecting the ENCODE and Roadmap enhancers and promoters in immune cells, including EBV-immortalized B cells and primary CD19+ cells (**Supplementary Table 6**). Interrogation of eQTL databases revealed that rs13226913 has a highly significant cis-eQTL effect on *CIGALT1* in peripheral blood ($P=3.9 \times 10^{-23}$) with the T allele associated with lower mRNA levels (**Supplementary Table 7**). Consistent with this finding, rs13226913 imparts an additive effect with each T (derived) increasing Gd-IgA1 levels by 0.22 standard deviation units (95%CI: 0.10-0.30).

The second genome-wide significant locus comprises a 100-kb interval on chromosome Xq24 (**Figure 2c**) and explains an additional 2.7% of the overall trait variance in Europeans and 1.2% in Asians (**Supplementary Table 4**). The top signal at this locus is represented by rs5910940 ($P=2.7 \times 10^{-8}$), a SNP 3' downstream from *CIGALTIC1*. The T (derived) allele increases serum Gd-IgA1 levels by 0.14 standard deviation units per allele (95%CI: 0.11-0.17). Our post-hoc examination of genotypic effects suggests a dominant effect of the rs5910940-T allele in females (dominant model $P=7.9 \times 10^{-9}$, **Supplementary Table 8**), although skewed inactivation of chromosome X in IgA1-producing cells could also potentially explain this effect.

CIGALTIC1 encodes a transmembrane protein that is similar to the core 1 beta1,3-galactosyltransferase 1 encoded by *CIGALT1*. However, its gene product (known as Cosmc) lacks the galactosyltransferase activity, but instead acts as a molecular chaperone required for the folding, stability, and full activity of C1GalT1²⁰. *CIGALTIC1* is also ubiquitously expressed in multiple tissues, including IgA1-secreting cells¹⁸, other blood cells, gastrointestinal tract, kidneys, and lungs¹⁹. Because sex chromosomes are not included in most eQTL analyses, we were not able to confirm if rs5910940 has an effect on the expression of *CIGALTIC1* based on available

datasets. However, rs5910940 tags a 2-bp insertion in the active promoter of *CIGALTIC1* in B-lymphocytes and leukemia cell lines (**Supplementary Table 9**). Considering the known functional dependency of *CIGALT1* and *CIGALTIC1*, we also tested for potential epistasis between these two loci, but did not detect any significant genetic interactions.

Taken together, these data predict an additive regulatory effect of rs13226913 and rs5910940, resulting in lower *CIGALT1* and *CIGALTIC1* expression, and leading to increased production of Gd-IgA1. We next performed siRNA knock-down studies in human cultured IgA1-secreting cell lines to confirm the effect of lower *CIGALT1* and *CIGALTIC1* transcript abundance on the production of Gd-IgA1 (**Figure 3**). Consistent with the observed genetic effect, *in vitro* knock-down of *CIGALT1* resulted in 30-50% increased production of Gd-IgA1 by the cells derived from IgAN patients ($P=0.025$) as well as healthy controls ($P=0.011$). Similar to *CIGALT1*, *in vitro* siRNA knock-down of *CIGALTIC1* in IgA1-producing cell lines significantly increased the production of Gd-IgA1 in healthy individuals ($P=0.032$) and a similar trend was observed in IgAN patients ($P=0.066$, **Figure 3**). Consistent with the genetic data, there were no multiplicative effects on Gd-IgA1 production with combined siRNA knock-down in IgA1-secreting lines.

Jointly, the newly discovered *CIGALT1* and *CIGALTIC1* loci explain up to **7%** of variance in Gd-IgA1 levels in Europeans and **2%** in Asians (**Supplementary Table 4**). Further examination of effect estimates by ethnicity confirms that the European cohorts predominantly drive these associations (**Supplementary Table 10**). Notably, the derived (T) allele of rs13226913 at *CIGALT1* locus is considerably more frequent in Europeans (freq. 47%) compared to Asians (freq. 10%), additionally contributing to the difference in variance explained between ethnicities. Subsequent examination of allelic frequencies in the Human Genome Diversity Panel (**Figure 4**) confirms that the derived allele of rs13226913 is rare or absent in some Asian populations, while being the predominant (major) allele in Europeans (freq. $\geq 50\%$). In contrast, the T (derived) allele of rs5910940 at *CIGALTIC1* locus is equally frequent in Asian and European populations (freq. $\sim 50\%$), but nearly fixed in

selected African populations. These findings suggest potential involvement of geographically confined selective pressures acting on the loci controlling the *O*-glycosylation process.

Lastly, we detected additional suggestive signals, including a locus on chromosome 7p13 that warrants further follow-up in larger cohorts (**Supplementary Figure 2a and 2b**). This locus is represented by rs978056 ($P=3.3 \times 10^{-5}$), an intronic SNP in *HECW1* (encoding E3 ubiquitin ligase) previously studied in the context of colon and breast cancer (**Supplementary Figure 3**). Based on the analysis of known protein-protein interactions, HECW1 is a second-degree neighbor of C1GalT1 and Cosmc, with ubiquitin C as a common interacting protein (**Supplementary Figure 2c**).

Discussion:

Genetic studies of immune endophenotypes have provided novel insights into the genetic architecture of complex traits and enhanced sub-classification of several autoimmune and inflammatory disorders. The power of immune endophenotypes is best exemplified by recent genetic studies of ANCA titers in vasculitis²¹, IgE levels in asthma^{22,23}, or studies of IgG *N*-glycosylation and autoimmunity¹⁷. Taking a similar approach, we performed the first GWAS for aberrant *O*-glycosylation of IgA1.

Abnormalities in the *O*-glycan synthesis have been linked to several human diseases, including IgA nephropathy, inflammatory bowel disease (IBD), hematologic diseases, and cancer. Mucin-type *O*-glycans produced by epithelial cells are critical for the formation of protective viscous barrier with anti-microbial properties at the mucosal surfaces of the gastrointestinal, urogenital and respiratory systems. Recent studies indicate that proper *O*-glycosylation of mucins is required for intestinal integrity in mice^{24,25} and may play a role in human susceptibility to IBD^{26,27}. In addition, *O*-glycosylation can affect the structure and immunogenicity of the modified proteins. For example, defective *O*-glycosylation represents the key pathogenic feature of Tn-syndrome²⁸, where acquired

enzymatic defect in the addition of galactose to *O*-glycans leads to exposed terminal N-acetylgalactosamine residue (Tn-antigen) on the surface of red blood cells, triggering polyagglutination by naturally occurring anti-Tn antibodies²⁸. Moreover, Tn and sialyl-Tn represent oncofetal antigens that are over-expressed in human cancers and may directly influence cancer growth, metastasis and survival, but the exact molecular perturbations that lead to *O*-glycosylation defects in tumor cells are presently not known²⁹.

Similar to Tn-syndrome, the pathogenesis of IgA nephropathy involves autoimmune response to Tn antigens. In this case, the Tn-antigen is exposed at the hinge region of IgA1 molecules as a result of aberrant *O*-glycosylation of IgA1 in the endoplasmic reticulum of IgA1-producing cells⁸. In patients with IgA nephropathy, the galactose-deficient IgA1 (Gd-IgA1) is recognized by circulating anti-Tn antibodies⁴, leading to the formation of nephritogenic immune complexes⁵⁻⁸. Several independent studies, including in healthy twins and in families with IgA nephropathy, have demonstrated that serum levels of Gd-IgA1 have high heritability, providing high level of support for a genetic determination of this trait and a strong rationale for this study^{11,12,30}.

In this study, we quantified the levels of Gd-IgA1 in sera of 2,633 subjects of European and East Asian ancestry using a simple lectin-based ELISA assay. Using GWAS approach, we discovered two genome-wide significant loci, on chromosomes 7p21.3 and Xq24, both with large effects on circulating levels of Gd-IgA1. The 7p21.3 locus contains *C1GALT1* gene, which encodes human core 1 β 1–3-galactosyltransferase (C1GalT1), the key enzyme responsible for the addition of galactose to the Tn antigen. Mice deficient in C1galt1 protein develop thrombocytopenia and kidney disease attributed to defective *O*-glycosylation of cell-surface proteins³¹. Moreover, C1galt1 deficiency in mice results in a defective mucus layer leading to spontaneous colonic inflammation that is dependent on the exposure to intestinal microbiota^{24,25}. C1GalT1 requires a molecular chaperone, Cosmc, which ensures that the enzyme is properly folded within the ER; loss of Cosmc activity results in C1GalT1 being degraded in the proteosome²⁸. Interestingly, Cosmc is encoded by *C1GALT1C1* residing within our second genome-wide significant locus on chromosome Xq24. We also localized a suggestive locus on chromosome 7p13 that encodes an E3 ubiquitin ligase, but it is presently not known if this protein participates in the proteosomal

degradation of C1GalT1. This signal will require further follow up. Importantly, our study demonstrates that there are several common genetic variants with relatively large effects on IgA1 *O*-glycosylation. These effects are conveyed by different genes, but converge on a single enzymatic step in the *O*-glycosylation pathway.

Our results contribute new insights into the genetic regulation of *O*-glycan synthesis, and demonstrate that a simple lectin-based assay can be used effectively to map genetic regulators of *O*-glycosylation of serum proteins. Given the high heritability of this trait, it is likely that additional loci contribute to variation in Gd-IgA1 levels. In particular, the inheritance pattern in IgAN kindreds suggested segregation of a major dominant gene, suggesting a potential role of additional rare alleles with large effects¹¹. A search in larger population-based studies that includes both common and rare variants is likely to uncover additional genetic determinants of *O*-glycosylation defects and elucidate mechanisms leading to IgA nephropathy and related disorders.

Materials and Methods:

Study Design and Power Analysis. The study was designed in two stages (**Supplementary Figure 1**). Stage 1 (the discovery phase) involved a genome-wide meta-analysis of two discovery cohorts: the Chinese cohort of 950 individuals (483 cases and 467 controls, all Han Chinese ancestry, genotyped with Illumina 660-quadruplet chip), and the US cohort of 245 individuals (141 cases and 104 controls, all European ancestry, genotyped with the Illumina 550v3 chip). Genome-wide scan was performed in both cohorts and fixed-effects meta-analysis was applied to prioritize signals for follow-up studies. Stage 2 (the replication phase) involved genotyping of the top signals from stage 1 in five additional cohorts of European and Asian ancestry (1,438 individuals in total, **Table 1**). We carried out power calculations for this design and for a range of effect sizes under the following assumptions: standard normal trait distribution, additive risk model, no heterogeneity in association, marker allelic frequency of 0.25 (average MAF for the microarrays used), perfect LD between a marker and a causal allele, a follow-up significance threshold of $P < 5 \times 10^{-4}$, and a joint significance level of $P < 5 \times 10^{-8}$. These calculations demonstrate

that we have adequate power to detect variants explaining $\geq 1.5\%$ of overall trait variance (**Supplementary Table 3**). Our study was conducted according to the principles expressed in the Declaration of Helsinki; all subjects provided informed consent to participate in genetic studies, and the Institutional Review Board of Columbia University as well as local ethics review committees for each of the individual cohorts approved our study protocol.

Phenotype Measurements and Quality Control. The level of serum total IgA was determined using standard ELISA³². The level of serum Gd-IgA1 was determined using custom HAA-based ELISA assay^{11,12,32}. This method relies on the detection of HAA-lectin binding to desialylated galactose-deficient glycans (Tn antigens) of serum IgA1 immunocaptured on ELISA plates. Because in humans, IgA1, but not IgA2, has *O*-glycans, this assay effectively quantifies the absolute level of Gd-IgA1 in units/ml serum. We have optimized this assay for high-throughput use. Briefly, 96-well plates were coated with F(ab')₂ fragment of goat IgG anti-human IgA at 3 μ g/ml. Plates were blocked with 1% BSA in PBS containing 0.05% Tween 20, and serial two-fold dilutions of samples and standards in blocking solution were incubated overnight at room temperature. To remove terminal sialic acid, the samples were treated with 100 μ L (1 mU) per well of neuraminidase (Roche) in 10 mM sodium acetate buffer (pH=5) for 3 h at 37°C. Next, the samples were incubated with GalNAc-specific biotinylated HAA lectin (Sigma-Aldrich) for 3 h at 37°C. The bound lectin was detected with avidin-horseradish peroxidase conjugate, followed by the peroxidase substrate, o-phenylenediamine-H₂O₂ (Sigma); the reaction was stopped with 1 M sulfuric acid. The concentration of Gd-IgA1 was calculated by interpolating the optical densities read at 490 nm on calibration curves constructed using a myeloma Gd-IgA1 standard. The intra-assay coefficients of variation (CVs) for calibration curves plotted by a 4-parameter model ranged from 2-10% for the extremes of the curves and 1-5% in the middle region. If higher values were noted, the samples were re-analyzed. The inter-assay CV was also consistently under 5% and our prior studies demonstrated excellent reproducibility of this assay³². In the final analysis, we applied a correction for potential plate effects using the same replicate samples across all plates. After corrections, serum Gd-IgA1 levels for each cohort were tested for normality by the Shapiro-Wilk test,

assisted by visual inspection of histograms and qq-plots. Non-normal trait distributions were transformed using logarithmic transformation. The log-transformed traits were regressed against age and case-control status to derive standardized residuals. Summary statistics (mean, SD, skewness, and kurtosis) were derived for the distribution of standardized residuals, which were then used as a quantitative trait in GWAS analysis. Summary statistics, normality testing, transformations, plots, and regression analyses were performed with R 3.0 software package (CRAN).

GWAS Discovery (Stage 1). The genotyping, genotype quality control, and ancestry analyses of the discovery cohorts have been previously described^{33,34}. Briefly, we implemented strict quality control filters for each cohort, eliminating samples with low call rates, duplicates, ancestry outliers, monomorphic and rare markers (MAF<1%), samples with cryptic relatedness, and samples with a detected sex mismatch. After all quality control steps, the Chinese Discovery Cohort was composed of 950 individuals typed with 508,112 SNPs, while the US Discovery Cohort was composed of 245 individuals typed with 531,778 SNPs. In total, 468,781 SNPs overlap between the cohorts. We used principal component-based ancestry matching algorithms to reduce any potential bias from population stratification (Spectral-GEM software)^{35,36}, as described in prior studies of these cohorts^{33,34}. Primary association testing for the Gd-IgA1 phenotype (expressed as standardized residuals) was performed within each cohort individually under an additive linear model in PLINK³⁷. Significant principal components of ancestry were included as covariates in the association analysis of each individual cohort. Additionally, we performed regression analyses with and without adjustment for serum total IgA levels. Adjusted effect estimates and standard errors were derived for each SNP and the results were combined across the genome using an inverse variance-weighted method (METAL software)³⁸. Genome-wide distributions of *P* values were examined visually using quantile-quantile plots for each individual cohort, as well as for the combined analysis. We estimated the genomic inflation factors³⁹, which were negligible for each individual discovery cohort ($\lambda = 1.011$ and 1.013 for the Chinese and US cohorts, respectively). The final meta-analysis QQ-plots showed no global departures from the expected distribution of *P* values and the overall genomic inflation factor was estimated at 1.010 (**Supplementary Figure 1**).

Follow-up of Suggestive Signals (Stage 2). We next prioritized the top 50 SNPs for replication among the top suggestive SNPs with $P < 5 \times 10^{-4}$ from the GWAS analyses. First, we clustered the top signals into distinct loci based on their genomic coordinates and metrics of LD. Conditional regression analysis was carried out to detect independent association between signals within the same genomic regions. For genotyping in replication cohorts, we prioritized the independent SNPs with the lowest P-value in each independent locus. We additionally required that each SNP be successfully genotyped in both discovery cohorts. We also excluded loci supported by only a single SNP ('singleton signals' defined by the absence of supporting signals with $P < 0.01$ within the same LD block). If genotyping of the top SNP failed, we selected a backup SNP on the basis of its strength of association, LD with the top SNP, quality of genotyping, and success in the design of working primers. Additionally, we added SNPs for which the signals became more significant ($P < 5 \times 10^{-4}$) after adjustment for serum total IgA levels. In all, we successfully acquired and analyzed genotype data for 50 carefully selected SNPs in 1,438 independent replication samples across 5 cohorts. The ethnic composition of the replication cohorts, genotyping methods, and genotype call rates are summarized in **Table 1**. Association analyses were first carried out individually within each of the cohorts using the same methods as in the discovery study. The results were next combined using a fixed-effects model (**Supplementary Table 2**). For each SNP, we derived pooled effect estimates, their standard errors, and 95% confidence intervals. To declare genome-wide significance, we used the generally accepted threshold of $P < 5 \times 10^{-8}$, initially proposed for Europeans genotyped with high-density platforms based on extrapolation to infinite marker density⁴⁰.

Chromosome X Analysis: We performed two types of association tests for X-linked markers. Our primary association test involved sex-stratified meta-analysis of chromosome X markers: each male and female sub-cohort was analyzed separately and the association statistics were combined across all sub-cohorts using fixed effects meta-analysis. This approach is not affected by the type of allele coding in males and allows for different effect size estimates between males and females²⁶. In secondary analyses, we assumed complete X-inactivation in females and a similar effect size between males and females. In this test, females are considered to have 0, 1, or 2

copies of an allele as in an autosomal analysis while males are considered to have 0 or 2 copies of the same allele (*i.e.*, male hemizygotes are equivalent to female homozygotes). The main limitation of this approach relates the assumption of complete X inactivation. Because approximately 15-25% of X-linked genes escape inactivation in female-derived fibroblasts⁴¹ and chromosome X inactivation has not been studied in IgA1-secreting cells, this analysis was performed only on an exploratory basis, but the results were consistent with sex-stratified analyses.

Tests of Alternative Inheritance Models and Epistasis: For the genome-wide significant loci, we explored two alternative genetic models (dominant and recessive) and compared these models using Bayesian Information Criterion (**Supplementary Table 8**). We also tested for all pairwise genetic interactions between the suggestive and significant loci using two different tests. First, we used a 1-degree-of-freedom likelihood ratio test (LRT) comparing two nested linear regression models: one with main effects only, and one with main effects and additive interaction terms. Second, we performed a more general 4-df genotypic interaction test. In this test, we compared a model with allelic effects, dominant effects, and their interaction terms with a reduced model without any of the interaction terms. All models were stratified by sex and cohort. The analyses were performed in R 3.0 software package (CRAN).

Functional Annotation of Significant and Suggestive Loci: To interrogate putative functional SNPs that were not typed in our dataset, we systematically identified all variants that were in high LD ($r^2 > 0.5$) with our top SNPs based on 1000 Genomes data. These variants were further annotated using ANNOVAR⁴², SeattleSeq⁴³, SNP Nexus⁴⁴, FunciSNP⁴⁵, HaploReg⁴⁶, and ChroMos⁴⁷. Additionally, we identified all genes whose expression was correlated with the top SNPs in cis- or trans- using the following eQTL datasets: (1) meta-analysis of transcriptional profiles from peripheral blood cells of 5,311 Europeans⁴⁸, (2) primary immune cells (B-cells and monocytes) from 288 healthy Europeans⁴⁹, and (3) the latest release of GTEx data across multiple tissue types^{19,50}. We utilized, automated MEDLINE text mining tools to assess network connectivity between genes residing in implicated GWAS loci, including GRAIL⁵¹, e-LiSe⁵², and FACTA+⁵³. We also interrogated all known protein-protein interaction networks for connectivity between candidate genes using the Disease Association Protein-

Protein Link Evaluator (DAPPLE)⁵⁴ and Protein Interaction Network Analysis platform (PINA2)⁵⁵. Network graphs were visualized in Cytoscape version 2.8.

siRNA Knock-down Studies in IgA1 Secreting Cell Lines: IgA1-secreting cell lines from five patients with IgAN and five healthy controls were transfected using ON-TARGETplus SMARTpool siRNAs (Thermo Fisher Scientific, Lafayette, CO, USA) specific for human *C1GALT1*, *COSMC*, or both. The ON-TARGETplus Non-targeting Pool siRNAs was used as a control. We followed our previously published protocol for Amaxa nucleofector II (Lonza, Allendale, NJ, USA)⁵⁶. Twenty-four hours after transfection, the knock-down efficiency was determined by qRT-PCR with previously described primers^{8,56}. The knockdown was expressed as cDNA level of the individual gene normalized to GAPDH after respective siRNA treatment, divided by the respective value obtained after treatment by non-targeting siRNA. The effect of siRNA knock-down on the phenotype (the degree of galactose-deficiency of IgA1) was based on the reactivity of secreted IgA1 with a lectin from *Helix aspersa* specific for terminal GalNAc, as described^{8,56}.

Acknowledgements:

We are grateful to all study participants for their contribution to this work. This study was supported by the following NIH grants from the National Institute for Diabetes and Digestive Kidney Diseases (NIDDK): K23DK090207 (K.K.), R03DK099564 (K.K.), R01DK105124 (K.K.), K01DK106341 (C.R.), R01DK078244 (J.N.), and R01DK082753 (A.G.G., J.N.), and by the Center for Glomerular Diseases at Columbia University.

Table and Figure Legend:

Table 1. Study Cohorts: the final numbers of cases and controls by cohort after implementation of all quality control filters.

Table 2. Combined results for new significant and suggestive GWAS signals: serum Gd-IgA1 levels were determined using HAA lectin-based ELISA, normalized and adjusted for age, case-control status and serum total IgA levels.

Figure 1. Longitudinal measurements of serum immunoglobulin levels and Gd-IgA1 levels over 4 years of follow-up. Initial and 4-year follow-up levels of (a) total serum IgG, (b) total serum IgA, and (c) serum Gd-IgA1 normalized for total serum IgA. Panels (d, e, f) represent scatter plots of initial (x-axis) versus follow-up (y-axis) values. P-values correspond to the Pearson's test of correlation; r^2 : squared correlation coefficient.

Figure 2. Results of the combined meta-analysis of serum Gd-IgA1 levels in 2,633 individuals of European and Asian ancestry: Manhattan plot (a) and regional plots for two distinct genome-wide significant loci, the *C1GALT1* locus (b) and the *C1GALT1C1* locus (c). The x-axis presents physical distance in kilobases (hg18 coordinates), and the y-axis presents $-\log P$ values for association statistics. The dotted horizontal line in a represents a genome-wide significance level ($P=5 \times 10^{-8}$). The regional plots contain all genotyped and imputed SNPs in the region meta-analyzed between the discovery and replication cohorts.

Figure 3. siRNA knock-down of C1GalT1, Cosmc and Cosmc+C1GalT1 in IgA1-secreting cell lines increases Gd-IgA1 production: (a) knock-down in IgA1-secreting cell lines from healthy controls; mock-control (n=5), non-targeting siRNA (n=7), C1GalT1 siRNA (n=5), Cosmc siRNA (n=7), and Cosmc+C1GalT1 siRNA (n=2); (b) knock-down in IgA1-secreting cell lines from IgAN patients; mock-control (n=5), non-targeting siRNA (n=7), C1GalT1 (n=5), Cosmc siRNA (n=7), and Cosmc+C1GalT1 siRNA (n=2); (c) relative change in mRNA in IgA1-secreting cell lines after siRNA knock-down of C1GalT1 (n=5), Cosmc (n=7), and Cosmc+C1GalT1 (n=2) compared to non-targeting siRNA control.

Figure 4. Genotypic effects and worldwide allelic frequency distribution for the two top genome-wide significant loci: (a) Mean trait values (+/- standard errors) by rs13226913 genotype at the *C1GALT1* locus. (b) The distribution of rs13226913 alleles across HGD populations. (c) Mean trait values (+/- standard errors) by rs5910940 genotype at the *C1GALT1C1* locus. (d) The distribution of rs5910940 alleles across HGD populations. The allelic distribution plots were modified from the HGD Selection Browser. The trait values were expressed as standard normal residuals of log-transformed serum Gd-IgA1 levels after adjustment for age, serum total IgA levels, case-control status and cohort membership.

Table 1. Study Cohorts: the final numbers of cases and controls by cohort after implementation of all quality control filters.

GWAS Cohorts*	Ancestry	N _{Cases}	N _{Controls}	N _{Total}	Genotyping Rate	Genotyping Platform**
Chinese Discovery Cohort	East Asian	483	467	950	99.9%	Illumina 660-quad
US Discovery Cohort	European	141	104	245	99.8%	Illumina 550v3
Total Discovery:		624	571	1,195		
Japanese Replication Cohort	East Asian	122	80	202	99.5%	KASP TM , LGC Genomics
Chinese Replication Cohort	East Asian	451	0	451	98.5%	KASP TM , LGC Genomics
German Replication Cohort	European	191	164	355	99.2%	KASP TM , LGC Genomics
French Replication Cohort	European	74	0	74	99.2%	KASP TM , LGC Genomics
US Replication Cohort	European	122	234	356	99.0%	KASP TM , LGC Genomics
Total Replication:		960	478	1,438		
Total All Cohorts:		1,584	1,049	2,633		

* Only individuals with the overall genotyping rate >95% (discovery) or >90% (replication) were included in the analysis.

** KASPTM: Kompetitive Allele Specific PCR (a proprietary SNP-typing technology by LGC Genomics; accuracy >99.8%).

Table 2. Combined results for significant and suggestive GWAS signals: serum Gd-IgA1 levels were adjusted for age, case-control status, and serum total IgA levels. Chromosome X analyses were stratified by sex.

Chr	Position (kb)	SNP	Allele*	Discovery Cohorts N=1,195			Replication Cohorts N=1,438			All Cohorts N=2,633			Hetero** P-value	Genes in Locus
				Effect	SE	P-value	Effect	SE	P-value	Effect	SE	P-value		
7	7213	rs13226913	T	0.20	0.05	2.2E-04	0.23	0.04	3.0E-08	0.22	0.03	3.2E-11	0.43 (NS)	C1GALT1
7	7239	rs1008897	G	0.19	0.06	4.6E-04	0.22	0.04	4.6E-07	0.21	0.03	9.1E-10	0.93 (NS)	C1GALT1
X	119642	rs5910940	T	0.13	0.03	2.3E-04	0.11	0.03	3.8E-05	0.14	0.03	2.7E-08	0.88 (NS)	C1GALT1C1
X	119698	rs2196262	A	0.11	0.03	1.2E-03	0.10	0.03	3.3E-04	0.12	0.02	1.4E-06	0.46 (NS)	C1GALT1C1
7	43345	rs978056	G	0.10	0.03	1.2E-03	0.07	0.03	7.5E-03	0.08	0.02	3.3E-05	0.15 (NS)	HECW1

* Gd-IgA1-increasing allele is provided as reference; ** P-value for the test of heterogeneity of effects

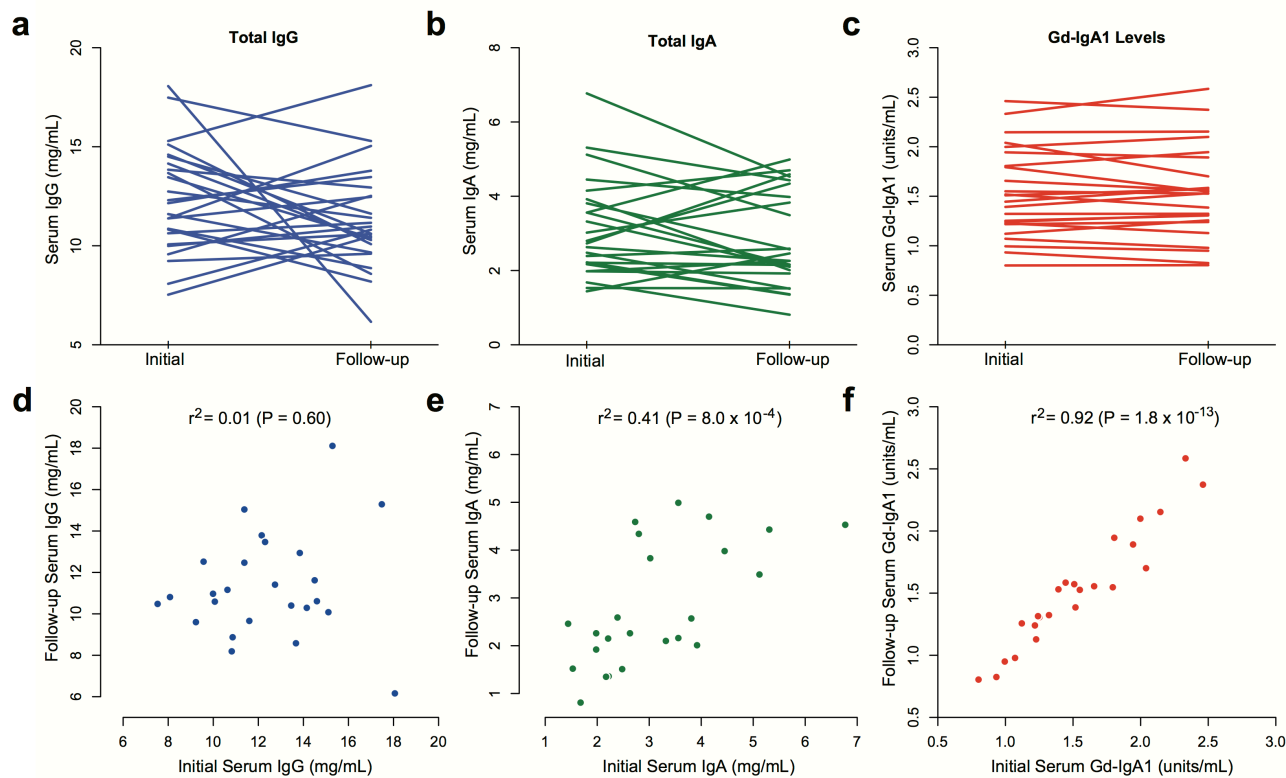


Figure 1. Longitudinal measurements of serum immunoglobulin levels and Gd-IgA1 levels over 4 years of follow-up.

Initial and follow-up levels of (a) total serum IgG, (b) total serum IgA, and (c) serum Gd-IgA1 normalized to total serum IgA. Panels (d, e, f) represent scatter plots of values at initial visit (x-axis) and 4-year follow-up visit (y-axis). P -values correspond to the Pearson's test of correlation; r^2 : squared correlation coefficient.

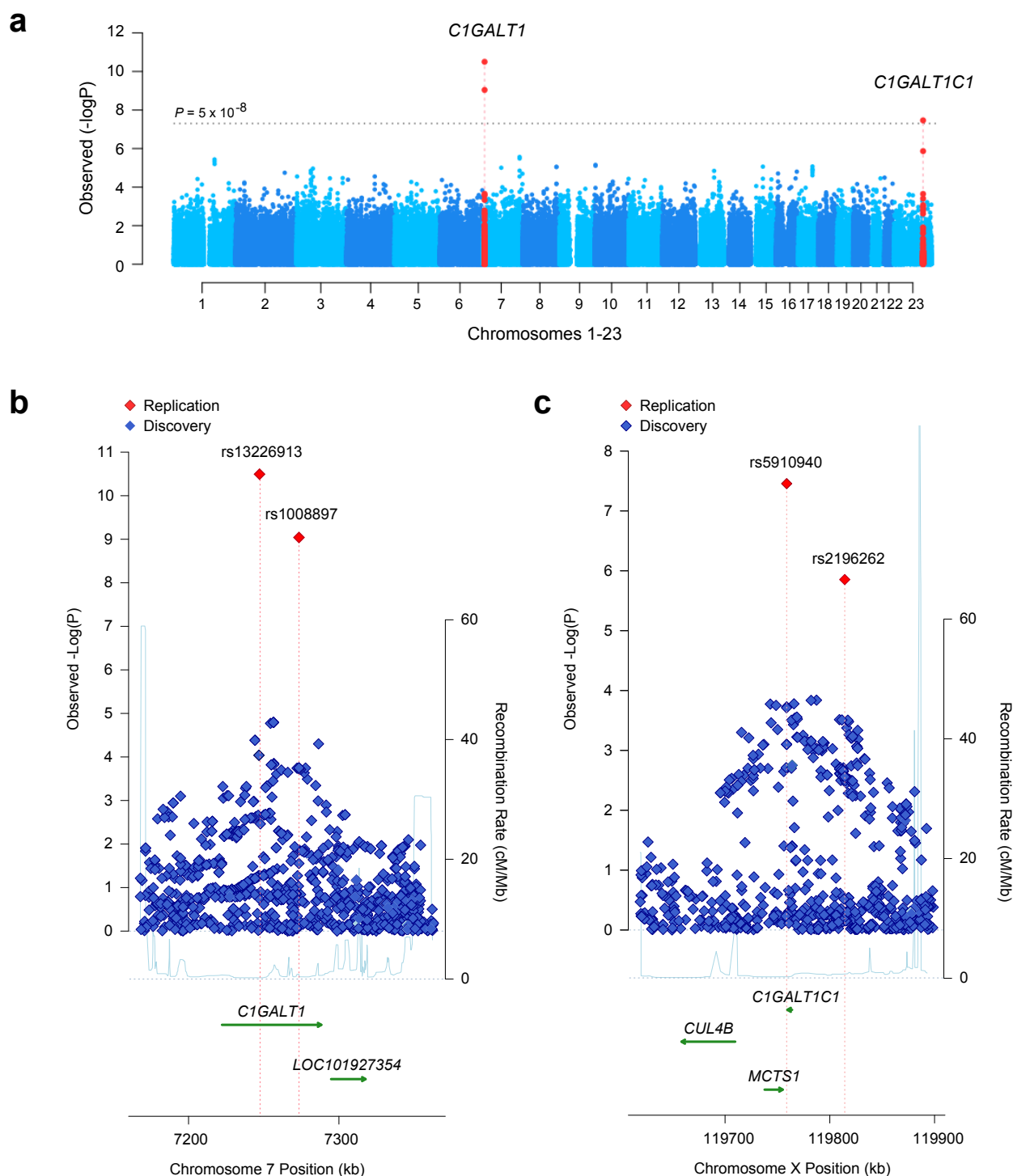


Figure 2. Results of the combined meta-analysis of serum Gd-IgA1 levels in 2,633 individuals of European and Asian ancestry

Manhattan plot (**a**) and regional plots for two distinct genome-wide significant loci, *C1GALT1* locus (**b**) and *C1GALT1C1* locus (**c**). The x-axis presents physical distance in kilobases (hg18 coordinates), and the y-axis presents $-\log P$ values for association statistics. The dotted horizontal line in **a** represents a genome-wide significance level ($P=5 \times 10^{-8}$). The regional plots contain all genotyped and imputed SNPs in the region meta-analyzed between the discovery and replication cohorts.

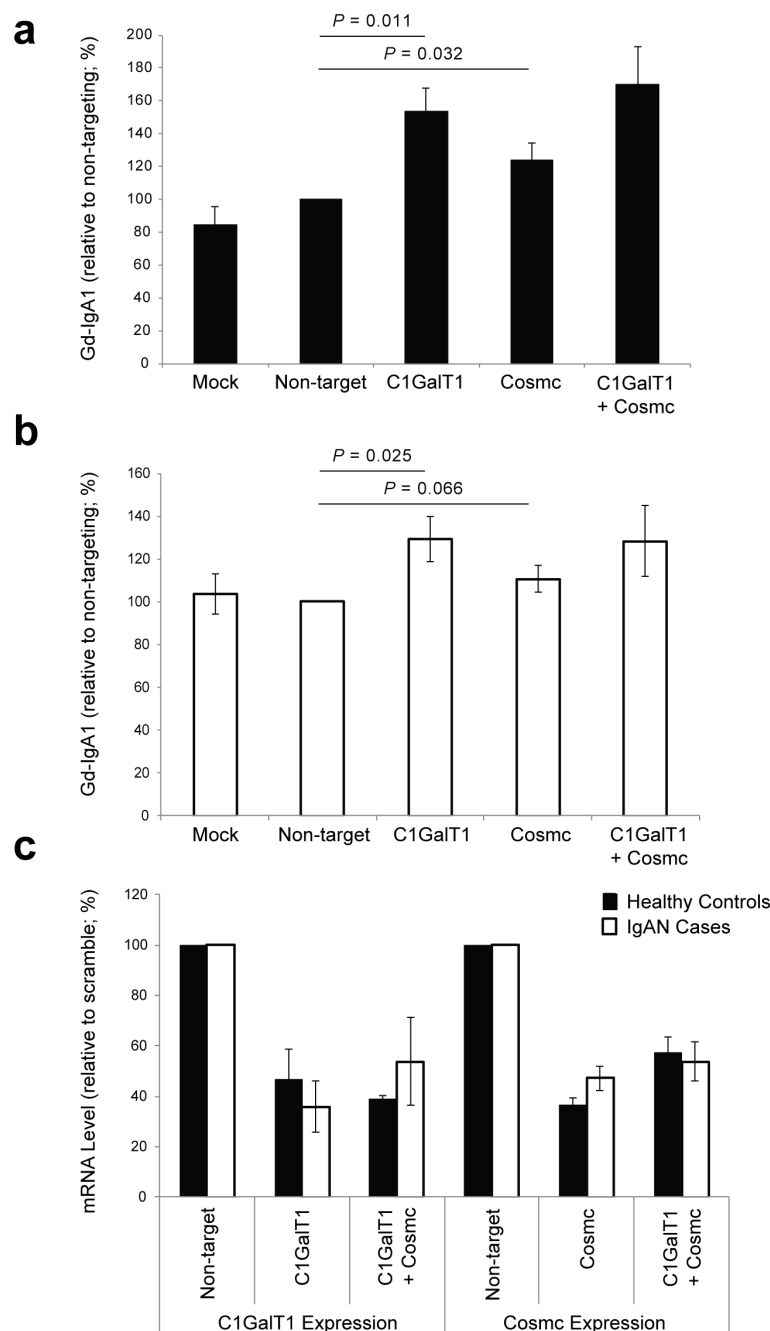


Figure 3. siRNA knock-down of C1GalT1, Cosmc and Cosmc+C1GalT1 in IgA1-secreting cell lines increases Gd-IgA1 production.

(a) Knock-down in IgA1-secreting cell lines from healthy controls; mock-control (n=5), non-targeting siRNA (n=7), C1GalT1 siRNA (n=5), Cosmc siRNA (n=7), and Cosmc+C1GalT1 siRNA (n=2); **(b)** Knock-down in IgA1-secreting cell lines from IgAN patients; mock-control (n=5), non-targeting siRNA (n=7), C1GalT1 siRNA (n=5), Cosmc siRNA (n=7), and Cosmc+C1GalT1 siRNA (n=2); **(c)** Relative change in mRNA in IgA1-secreting cell lines after siRNA knock-down of C1GalT1 (n=5), Cosmc (n=7), and Cosmc+C1GalT1 (n=2) compared to non-targeting siRNA control.

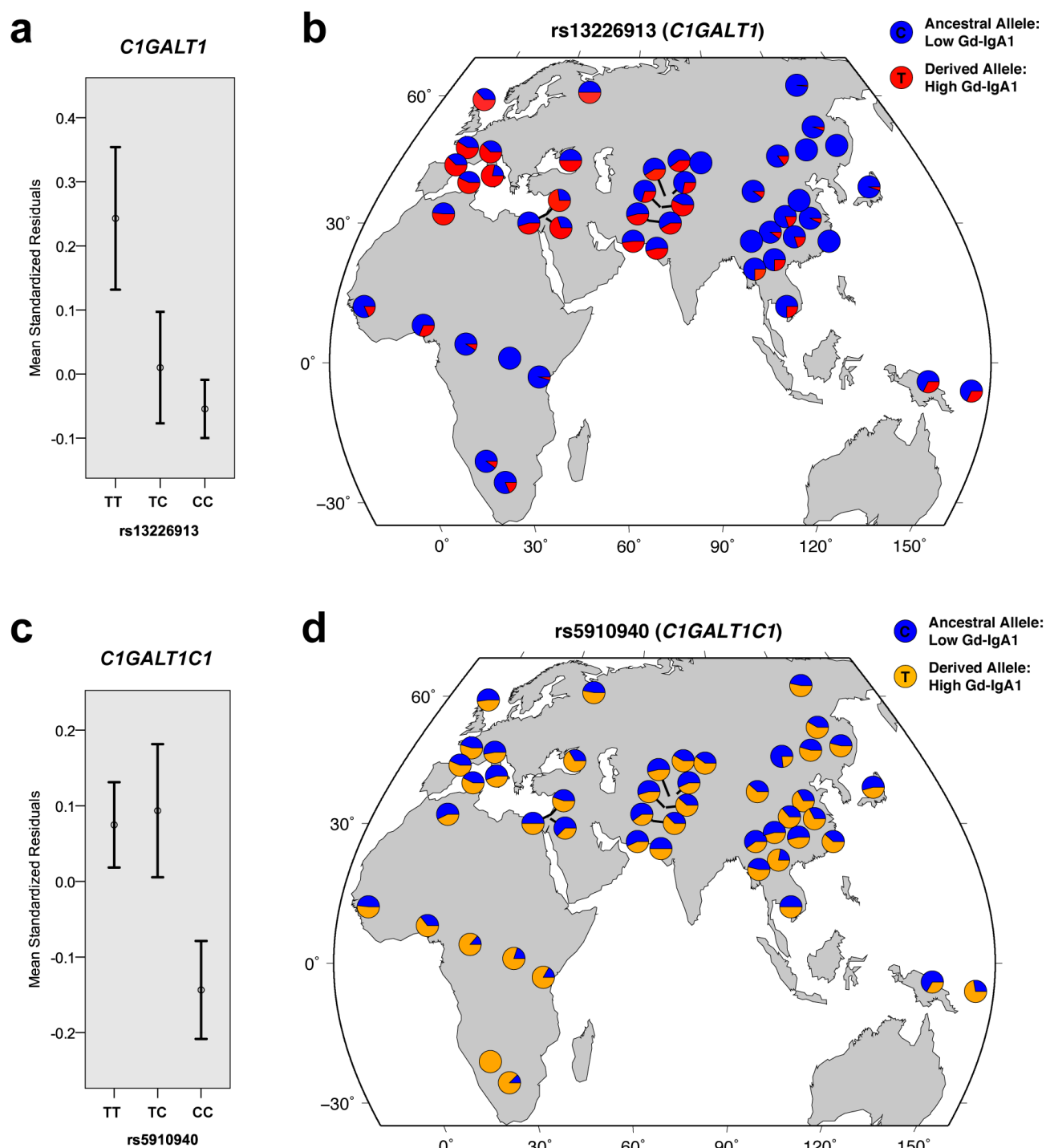


Figure 4. Genotypic effects and worldwide allelic frequency distribution for the two top genome-wide significant loci.

(a) Mean trait values (+/- standard errors) by rs13226913 genotype at the *C1GALT1* locus. (b) The distribution of rs13226913 alleles across HGDP populations. (c) Mean trait values (+/- standard errors) by rs5910940 genotype at the *C1GALT1C1* locus. (d) The distribution of rs5910940 alleles across HGDP populations. The allelic distribution plots were modified from the HGDP Selection Browser. The trait values were expressed as standard normal residuals of log-transformed serum Gd-IgA1 levels after adjustment for age, serum total IgA levels, case-control status and cohort membership.

References:

1. Magistroni, R., D'Agati, V.D., Appel, G.B. & Kiryluk, K. New developments in the genetics, pathogenesis, and therapy of IgA nephropathy. *Kidney Int* **88**, 974-89 (2015).
2. Moldoveanu, Z. *et al.* Patients with IgA nephropathy have increased serum galactose-deficient IgA1 levels. *Kidney Int.* **71**, 1148-1154 (2007).
3. Smith, A.C., de Wolff, J.F., Molyneux, K., Feehally, J. & Barratt, J. O-glycosylation of serum IgD in IgA nephropathy. *J Am Soc Nephrol* **17**, 1192-9 (2006).
4. Suzuki, H. *et al.* Aberrantly glycosylated IgA1 in IgA nephropathy patients is recognized by IgG antibodies with restricted heterogeneity. *J Clin Invest* **119**, 1668-77 (2009).
5. Allen, A.C. *et al.* Mesangial IgA1 in IgA nephropathy exhibits aberrant O-glycosylation: observations in three patients. *Kidney Int* **60**, 969-73 (2001).
6. Hiki, Y. *et al.* Mass spectrometry proves under-O-glycosylation of glomerular IgA1 in IgA nephropathy. *Kidney Int* **59**, 1077-85 (2001).
7. Tomana, M. *et al.* Circulating immune complexes in IgA nephropathy consist of IgA1 with galactose-deficient hinge region and antiglycan antibodies. *J Clin Invest* **104**, 73-81 (1999).
8. Suzuki, H. *et al.* IgA1-secreting cell lines from patients with IgA nephropathy produce aberrantly glycosylated IgA1. *J Clin Invest* **118**, 629-39 (2008).
9. Zhao, N. *et al.* The level of galactose-deficient IgA1 in the sera of patients with IgA nephropathy is associated with disease progression. *Kidney Int* **82**, 790-6 (2012).
10. Berthou, F. *et al.* Autoantibodies targeting galactose-deficient IgA1 associate with progression of IgA nephropathy. *J Am Soc Nephrol* **23**, 1579-87 (2012).
11. Gharavi, A.G. *et al.* Aberrant IgA1 glycosylation is inherited in familial and sporadic IgA nephropathy. *J Am Soc Nephrol* **19**, 1008-14 (2008).
12. Kiryluk, K. *et al.* Aberrant glycosylation of IgA1 is inherited in both pediatric IgA nephropathy and Henoch-Schonlein purpura nephritis. *Kidney Int* **80**, 79-87 (2011).
13. 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).
14. Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nat Commun* **6**, 7213 (2015).
15. 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).
16. Ferreira, R.C. *et al.* Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat Genet* **42**, 777-80 (2010).
17. Lauc, G. *et al.* Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet* **9**, e1003225 (2013).
18. Qin, W. *et al.* Peripheral B lymphocyte beta1,3-galactosyltransferase and chaperone expression in immunoglobulin A nephropathy. *J Intern Med* **258**, 467-77 (2005).
19. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
20. Ju, T. & Cummings, R.D. A unique molecular chaperone Cosmc required for activity of the mammalian core 1 beta 3-galactosyltransferase. *Proc Natl Acad Sci U S A* **99**, 16613-8 (2002).
21. Lyons, P.A. *et al.* Genetically distinct subsets within ANCA-associated vasculitis. *N Engl J Med* **367**, 214-23 (2012).

22. Liang, L. *et al.* An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature* **520**, 670-4 (2015).
23. Weidinger, S. *et al.* Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. *PLoS Genet* **4**, e1000166 (2008).
24. Fu, J. *et al.* Loss of intestinal core 1-derived O-glycans causes spontaneous colitis in mice. *J Clin Invest* **121**, 1657-66 (2011).
25. Perez-Munoz, M.E. *et al.* Discordance between changes in the gut microbiota and pathogenicity in a mouse model of spontaneous colitis. *Gut Microbes* **5**, 286-95 (2014).
26. Chang, D. *et al.* Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One* **9**, e113684 (2014).
27. Theodoratou, E. *et al.* The role of glycosylation in IBD. *Nat Rev Gastroenterol Hepatol* **11**, 588-600 (2014).
28. Ju, T. & Cummings, R.D. Protein glycosylation: chaperone mutation in Tn syndrome. *Nature* **437**, 1252 (2005).
29. Radhakrishnan, P. *et al.* Immature truncated O-glycophenotype of cancer directly induces oncogenic features. *Proc Natl Acad Sci U S A* **111**, E4066-75 (2014).
30. Lomax-Browne, H.J. *et al.* IgA1 Glycosylation Is Heritable in Healthy Twins. *J Am Soc Nephrol* (2016).
31. Alexander, W.S. *et al.* Thrombocytopenia and kidney disease in mice with a mutation in the C1galt1 gene. *Proc Natl Acad Sci U S A* **103**, 16442-7 (2006).
32. Moldoveanu, Z. *et al.* Patients with IgA nephropathy have increased serum galactose-deficient IgA1 levels. *Kidney Int* **71**, 1148-54 (2007).
33. Gharavi, A.G. *et al.* Genome-wide association study identifies susceptibility loci for IgA nephropathy. *Nat Genet* **43**, 321-7 (2011).
34. Kiryluk, K. *et al.* Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens. *Nat Genet* **46**, 1187-96 (2014).
35. Lee, A.B., Luca, D. & Roeder, K. A Spectral Graph Approach to Discovering Genetic Ancestry. *Ann Appl Stat* **4**, 179-202 (2010).
36. Lee, A.B., Luca, D., Klei, L., Devlin, B. & Roeder, K. Discovering genetic ancestry using spectral graph theory. *Genet Epidemiol* **34**, 51-9 (2010).
37. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
38. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
39. Devlin, B., Roeder, K. & Bacanu, S.A. Unbiased methods for population-based association studies. *Genet Epidemiol* **21**, 273-84 (2001).
40. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* **32**, 227-34 (2008).
41. Carrel, L. & Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400-4 (2005).
42. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
43. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).

44. Chelala, C., Khan, A. & Lemoine, N.R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655-61 (2009).
45. Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. & Noushmehr, H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res* **40**, e139 (2012).
46. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
47. Barenboim, M. & Manke, T. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. *Bioinformatics* **29**, 2197-8 (2013).
48. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
49. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**, 502-10 (2012).
50. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
51. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).
52. Gladki, A., Siedlecki, P., Kaczanowski, S. & Zielenkiewicz, P. e-LiSe--an online tool for finding needles in the '(Medline) haystack'. *Bioinformatics* **24**, 1115-7 (2008).
53. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. & Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **27**, i111-9 (2011).
54. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
55. Cowley, M.J. *et al.* PINA v2.0: mining interactome modules. *Nucleic Acids Res* **40**, D862-5 (2012).
56. Suzuki, H. *et al.* Cytokines alter IgA1 O-glycosylation by dysregulating C1GalT1 and ST6GalNAc-II enzymes. *J Biol Chem* **289**, 5330-9 (2014).