Identification of new susceptibility loci for IgA nephropathy in Han Chinese

Ming Li1,2,*, Jia-Nee Foo3,*, Jin-Quan Wang4,*, Hui-Qi Low3, Xue-Qing Tang1,2, Kai-Yee Toh3, Pei-Ran Yin1,2, Chiea-Chuen Khor5,6, Yu-Fen Goh3, Ishak D. Irwan3, Ri-Cong Xu1,2, Anand K. Andiappan7, Jin-Xin Bei3, Olaf Rotzschke7, Meng-Hua Chen8, Ching-Yu Cheng5,6, Liang-Dan Sun9,10, Geng-Ru Jiang11, Tien-Yin Wong5,6, Hong-Li Lin12, Tin Aung5,6, Yun-Hua Liao13, Seang-Mei Saw14,15, Kun Ye15, Richard P. Ebstein16, Qin-Kai Chen17, Wei Shi18, Soo-Hong Chew9,19, Wei Shi18, Nan Chen27, Xue-Jun Zhang8,10, Yi-Xin Zeng28,29, Hong Zhang30, Zhi-Hong Liu31, Xue-Qing Yu31,32, Jian-Jun Liu31,32,33,34,35,36,37,38,39,40,41,42,43 & J.-H. L. Li (email: liuj3@gis.a-star.edu.sg).

IgA nephropathy (IgAN) is one of the most common primary glomerulonephritis. Previously identified genome-wide association study (GWAS) loci explain only a fraction of disease risk. To identify novel susceptibility loci in Han Chinese, we conduct a four-stage GWAS comprising 8,313 cases and 19,680 controls. Here, we show novel associations at ST6GAL1 on 3q27.3 (rs7634389, odds ratio (OR) = 1.13, \( P = 7.27 \times 10^{-10} \)), ACCS on 11p11.2 (rs2074038, OR = 1.14, \( P = 3.93 \times 10^{-9} \)) and ODFT-KLF10 on 8q22.3 (rs2033562, OR = 1.13, \( P = 1.14 \times 10^{-9} \)), validate a recently reported association at ITGAX-ITGAM on 16p11.2 (rs7190997, OR = 1.22, \( P = 2.26 \times 10^{-19} \)), and identify three independent signals within the DEFA locus (rs2738058, \( P = 1.15 \times 10^{-19} \); rs12716641, \( P = 9.53 \times 10^{-9} \); rs9314614, \( P = 4.25 \times 10^{-9} \), multivariate association). The risk variants on 3q27.3 and 11p11.2 show strong association with mRNA expression levels in blood cells while allele frequencies of the risk variants within ST6GAL1, ACCS and DEFA correlate with geographical variation in IgAN prevalence. Our findings expand our understanding on IgAN genetic susceptibility and provide novel biological insights into molecular mechanisms underlying IgAN.

1 Department of Nephrology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510080, China. 2 Key Laboratory of Nephrology, Ministry of Health and Guangdong Province, Guangzhou, Guangdong 510080, China. 3 Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore. 4 National Clinical Research Center of Kidney Diseases, Jinling Hospital, Nanjing University School of Medicine, Nanjing, Jiangsu 210002, China. 5 Singapore Eye Research Institute, Singapore 169856, Singapore. 6 Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore 119228, Singapore. 7 Singapore Immunology Network, Agency for Science, Technology and Research, Singapore 138648, Singapore. 8 Department of Nephrology, General Hospital of Ningxia Medical University, Yinchuan, Ningxia 750004, China. 9 Institute of Dermatology and Department of Dermatology, No.1 Hospital, Anhui Medical University, Hefei, Anhui, 230032, China. 10 State Key Laboratory Incubation Base of Dermatology, Ministry of National Science and Technology, Hefei, Anhui 230032, China. 11 Department of Nephrology, XinHua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200092, China. 12 Department of Nephrology, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi 330006, China. 13 Department of Nephrology, The First Affiliated Hospital, Guangxi Medical University, Nanning 530021, China. 14 Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore 117549, Singapore. 15 Department of Nephrology, The People's Hospital of Guangxi Autonomous Region, Nanning, Guangxi 530021, China. 16 Department of Psychology, National University of Singapore, Singapore 117570, Singapore. 17 Department of Nephrology, The First Affiliated Hospital of Nanochang University, Nananch, Jiangxi 330006, China. 18 Department of Nephrology, Guangdong General Hospital, Guangzhou, Guangdong 510080, China. 19 Department of Economics, National University of Singapore, Singapore 117570, Singapore. 20 Department of Nephrology, Fuzhou General Hospital of Nanjing Military Command, Fuzhou, Fujian 350025, China. 21 Shandong Provincial Institute of Dermatology and Venerology, Shandong Academy of Medical Science, Jinan, Shandong 250000, China. 22 Department of Hepatobiliary Oncology, State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong 510080, China. 23 Department of Nephrology, Tongji Hospital, Tongji Medical College of Huazhong University of Science & Technology, Wuhan, Hubei 430030, China. 24 Duke-National University of Singapore Graduate Medical School, Singapore 169857, Singapore. 25 Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, National University Health System, Singapore 119228, Singapore. 26 Department of Nephrology, Sichuan Provincial People's Hospital, Chengdu, Sichuan 610072, China. 27 Department of Nephrology, RuiJin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200025, China. 28 State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou 510080, China. 29 Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 1000730, China. 30 Renal Division, Peking University First Hospital, Peking University, Institute of Nephrology, Beijing 100034, China. 31 School of Biological Sciences, Anhui Medical University, Hefei, Anhui 230032, China. 32 Institute of Dermatology and Department of Dermatology, No.1 Hospital, Anhui Medical University, Hefei, Anhui 230032, China. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to X-Q.Y. (email: yuxq@mail.sysu.edu.cn) or to J.-H.L. (email: liuj3@gis.a-star.edu.sg).

Received 12 Dec 2014 | Accepted 23 Apr 2015 | Published 1 Jun 2015 DOI: 10.1038/ncomms8270

© 2015 Macmillan Publishers Limited. All rights reserved.
IgA nephropathy (IgAN) is the most common primary glomerulonephritis and a major cause of end-stage renal disease in the Chinese population. It is characterized by the deposition of IgA in the mesangial area of glomeruli, and up to 40% of the cases progress to end-stage renal diseases within 20 years of disease onset.1-2 There is a marked regional difference in frequency of IgAN. It occurs with highest frequency in Asian populations, accounting for 45–58.2% of primary glomerular disease, with modest frequency in Caucasians (USA and Europe) and with lower frequency in the African population.2,3 These differences, together with evidence of familial clustering, strongly suggest the presence of a substantial genetic contribution to the disease. Previous genome-wide association studies (GWASs) have identified five loci significantly associated with IgAN, namely chromosome 1q32 (CFHR3-CFH1 genes), 6p21 (MHC), 8p23 (DEFA gene cluster), 17p13.1 (TNFSF13) and 22q12 (HORMAD2), providing the first valuable insights into genetic risk factors underlying disease mechanisms.4-5 These findings explain only a fraction of the disease risk and it is clear that more genetic loci remain to be discovered.

In this study, we conduct a four-stage GWAS comprising 8,313 cases and 19,680 controls of Han Chinese ancestry (Supplementary Fig. 1). We identify novel loci on 3q27.3 (ST6GAL1), 11p11.2 (ACCS) and 8q22.3 (ODF1-KLF10), validate a recently reported association on 16p11.2 (ITGAX-ITGAM), and identify three independent signals within the DEFA locus on 8p23. These findings significantly expand our understanding of the genetic susceptibility to IgAN.

Results

Genome-wide discovery analysis. We performed a new GWAS analysis (stage 1) by combining our published GWAS data set (consisting 1,434 cases and 4,270 controls) with an additional 6,511 control subjects of Chinese Han ethnicity, including 981 subjects from Guangdong, 523 subjects from Shandong and 5,007 subjects from Singapore (Supplementary Table 1). In each data set, we imputed untyped single-nucleotide polymorphisms (SNPs) using the 1,000-genome multi-ethnic reference panel (Feb 2012, IMPUTE v2) (refs 6-8). After merging shared SNPs across the data sets and stringent quality control filtering (see Methods), we analysed a total of 3,792,949 autosomal SNPs in 1,434 cases and 10,661 controls. With the expanded sample size and deep imputation of untyped variants, this new GWAS analysis has improved statistical power over the previous GWAS to detect risk variants in the range of odds ratio (OR) < 1.3 and minor allele frequency (MAF) < 10% (Supplementary Table 2).

We analysed genotype dosages taking into account imputation uncertainties,6-8 using a logistic regression model assuming an additive effect for allelic risk and adjusting for population stratification using the first five principal components (PCs) as covariates (Supplementary Figs 2–5). The genomic inflation factor after PC correction was low (λGC = 1.087; 1.073 for SNPs with MAF ≥5%, λ1,000 = 1.034), suggesting minimal effects of population stratification in our discovery GWAS. In addition, we examined genotypes of 3,731,832 SNPs imputed with high confidence (info score > 0.8, with call rates >95% for genotype with probabilities >0.9) using a multivariate linear mixed model implemented in GEMMA11,12 (λGC = 1.039; 1.015 for SNPs with MAF ≥5%, λ1,000 = 1.015) (Supplementary Figs 2 and 3), which gave results that were largely consistent with the logistic regression analysis with PC correction (Tables 1 and 2, Supplementary Data 1).

Our genome-wide discovery analysis provided strong supporting evidence for previously published associations4,5 at 17p13 and 22q12 (Supplementary Table 3, Supplementary Figs 6–8).

A similar trend of association was also observed at 1q32, even though the result was not statistically significant (Supplementary Table 3). We also observed strong and consistent evidence for the three independent association signals within the major histocompatibility complex (MHC) region 6p21 that we previously reported (Supplementary Table 3). In addition, we identified a novel independent signal at rs2295119 in the MHC region that remained genome-wide significant after conditioning on all previously reported human leukocyte antigen (HLA) SNPs (OR = 1.359, logistic regression P unconditioned = 3.24 × 10^-11, Pconditioned = 7.52 × 10^-10). This SNP is in linkage disequilibrium (LD) with rs9277554 (r^2 = 0.556, D’ = 0.993) which was found to be independently associated but did not reach genome-wide significance in our previous study.5

We performed HLA imputation on the expanded GWAS data set using genotyped SNPs within chr6: 20–40 Mb (build 37), the SNP2HLA tool13 and the Pan-Asian reference panel.14,15 In addition to the associations at the previously reported four-digit HLA alleles (Supplementary Table 4), we also observed strong associations at two-digit alleles DPB1*02 (OR = 1.32, logistic regression P = 1.77 × 10^-10) tagged by rs2295119 (r^2 = 0.94), and DRB1*04 (OR = 1.45, logistic regression P = 3.16 × 10^-11), tagged by previously reported SNPs rs6660895 (r^2 = 0.44) and rs1794275 (r^2 = 0.13) (Supplementary Table 4). Detailed HLA sequencing and typing analyses will be needed to further understand these associations.

Our analysis also suggested three independent association signals within the DEFA gene cluster on 8p23 (each P < 10^-4, r^2 < 0.1 between each pair of SNPs). After excluding the SNPs within the five previously identified regions, a notable excess of extremely small P values was observed on the quantile–quantile plot compared with the expected null distribution (Supplementary Fig. 9), which suggests the existence of additional associations beyond the ones already identified.

Validation analysis. We first selected the top 136 independent SNPs exceeding P < 1 × 10^-4 in either the PC-adjusted logistic regression or GEMMA analysis that are not within the known loci (hypothesis free). Finally, by including the three independent SNPs within the 8p23 locus, a total of 139 SNPs were selected for the first validation study, but only the assays for 122 were successfully designed for multiplex genotyping analysis by Sequonem.

As the initial validation, 115 SNPs were successfully genotyped in 2,651 IgAN cases and 2,907 controls of Han Chinese ethnicity recruited from the Southern region of China (stage 2). We performed logistic regression analysis of the validation samples and combined the association statistics across the discovery (PC-adjusted or GEMMA) and validation samples by a fixed-effects meta-analysis (Supplementary Data 1). We then took forward the top 22 SNPs with P < 1 × 10^-4 in the meta-analysis of the combined stages 1 and 2 samples for further validation analysis in an independent set of samples (stage 3) consisting of 2,428 IgAN cases and 4,202 controls recruited from the Northern (1,463 cases and 1,683 controls) and Southern (965 cases and 2,519 controls) regions of China (Supplementary Table 5). Finally, the top eight SNPs with P < 5 × 10^-7 in the combined samples of stages 1 + 2 + 3 and showing consistent associations across all four sample collections were analysed in an additional independent set of 1,800 IgAN cases and 1,910 controls (stage 4) recruited from Northern (704 cases and 805 controls) and Southern (1,096 cases and 1,105 controls) China (Table 1). We then conducted a full meta-analysis of all the stages 1–4 samples (Supplementary Fig. 1) by analysing the Northern and Southern samples of each of the four stages as independent sample
collections (six independent samples in total) to minimize bias resulting from population stratification.

From the combined analysis of a total of 8,313 IgaAN cases and 19,680 controls, we identified SNPs at four out of the five novel loci reaching genome-wide significance ($P < 5 \times 10^{-8}$), rs7190997 within the ITGAM-ITGAX locus on 16p11.2 (OR = 1.22, fixed-effects meta-analysis $P = 2.26 \times 10^{-19}$), rs2074038 at the ST6GAL1 locus on 3q27.3 (OR = 1.176, meta-analysis $P = 1.096 \times 10^{-16}$), chr16:31368178 on 16p11.2 (OR = 1.15, meta-analysis $P = 1.032 \times 10^{-15}$), on 8q22.3 (OR = 1.26%, meta-analysis $P = 1.059 \times 10^{-16}$), and rs7634389 at the ST6GAL1 locus on 3q27.3 (OR = 1.176, meta-analysis $P = 1.096 \times 10^{-16}$) (Table 1, Fig. 1). The fifth locus rs11264799 (FCRL3) on 1q23.1 remained suggestive (OR = 1.14, meta-analysis $P = 2.00 \times 10^{-7}$) (Supplementary Table 6). We also confirmed all three independent association signals within the DEFA locus at rs2738058 (OR = 1.23, meta-analysis $P = 1.15 \times 10^{-19}$), rs12716641 (OR = 1.15, meta-analysis $P = 9.53 \times 10^{-19}$) and rs9314614 (OR = 1.13, meta-analysis $P = 4.25 \times 10^{-9}$) through a multivariate association analysis (Table 2, Fig. 2). All the novel associations showed consistent effects across all the independent sample collections without evidence of heterogeneity and obtained statistically significant in the validation samples after correction for multiple testing (Bonferroni corrected $P < 0.05/154 = 3.25 \times 10^{-4}$; Tables 1 and 2).

To further ensure that none of the novel associations were influenced by population stratification and/or batch effects among the expanded control samples of stage 1 (genotyped in different arrays), we re-examined the imputation info scores and allele frequencies of all the validated SNPs across the different control data sets and found them to be high quality and very consistent across genotyping arrays and sample collections (Supplementary Table 7). We re-examined our reported SNPs in a subset of samples (1,434 cases and 4,270 controls) that were analysed in our previous GWAS, on which the imputation was performed as a batch from 444,882 overlapping autosomal SNPs that were genotyped on different Illumina chips and found that the strengths of these associations (ORs) were consistent with the results from the full data set, indicating that the new evidences of these associations were driven by improved statistical power and genetic variation coverage rather than systemic bias due to the
inclusion of additional control samples that were imputed in separate data sets (Supplementary Table 7). Good genotyping clusters were observed across all genotyping platforms for the reported SNPs (Supplementary Figs 10–12) with allele frequencies of imputed SNPs closely matching those genotyped in the validation samples (Tables 1 and 2).

Next, we re-ran the full meta-analysis by dividing the discovery data set into Northern (414 cases and 2,306 controls) and Southern (1,020 cases and 8,355 controls) clusters with adjustment of the top five principal components re-calculated within each cluster (see Methods) (Supplementary Fig. 5), and the full meta-analysis results were consistent at the top loci (Supplementary Table 8). The association effects were not significantly different between the combined (discovery and validation) Northern and Southern samples ($P_{\text{het}} > 0.05$) (Supplementary Table 9). In addition, we did another full meta-analysis with lambda GC correction of the PC-adjusted results from the discovery data set. All the novel loci remained genome-wide significant (Supplementary Table 9). Finally, the associations were not influenced by age and gender, and similar effects were observed in males and females (Supplementary Table 10).

While this manuscript was under preparation, an independent GWAS on IgAN conducted in Europeans and Asians was published and reported three novel loci (ITGAX, VAV3 and CARD9 (ref. 16)). Of the two previously reported independent associations within ITGAX locus, our top SNP (rs7190997) at ITGAX shows a high LD with the reported SNP rs11150612 ($r^2 = 0.71$) with the reported SNP rs2738048. For the GWAS samples, only the result from logistic regression analyses (GWAS logistic) was used in the fixed-effects meta-analysis with the validation samples.

| SNP         | Risk allele frequency (%) | Risk allele frequency (%) | $P$ value | OR (95% CI) | Multivariate analysis$^*$ |
|-------------|----------------------------|---------------------------|-----------|-------------|---------------------------|
| rs2738058   | chr8:6821617               |                           |           |             |                           |
| T/C         | 73.0                       | 66.6                      | 5.75E-10  | 1.310 (1.203-1.427) | 2.06E-08 (1.298 (1.85-1.422)) |
| rs9314614   | chr8:6697731               |                           |           |             |                           |
| C/G         | 39.7                       | 36.0                      | 3.54E-05  | 1.191 (1.097-1.295) | 2.85E-04 (1.161 (1.071-1.258)) |
| rs12716641  | chr8:6898998               |                           |           |             |                           |
| T/C         | 78.0                       | 74.3                      | 7.80E-07  | 1.260 (1.150-1.381) | 4.40E-03 (1.153 (1.045-1.272)) |

Only rs2738058 is in LD ($r^2 = 0.71$) with the reported SNP rs2738048. For the GWAS samples, only the result from logistic regression analyses (GWAS logistic) was used in the fixed-effects meta-analysis with the validation samples.

$^*$Pairwise LD ($r^2$) between rs2738058 and rs9314614 was 0.074. There is no LD between rs9314614 and rs12716641 ($r^2 = 0$).
heterogeneity across all the six sample collections (CI) = 1.07–1.31, meta-analysis P = 0.0011) without evidence of heterogeneity across all the six sample collections ($I^2 = 0\%$, Cochran’s Q test $P_{\text{heterogeneity}} = 0.70$). The effect size in our samples was, however, smaller than what was previously reported, and a slightly larger effect was observed in our Northern (OR = 1.26) than Southern (OR = 1.12) samples (Supplementary Table 11).

Independent associations within the DEFA locus. We discovered three independent signals at the DEFA locus (Table 2, Supplementary Table 12). Of these, only rs2738058 is in LD with the previously reported SNP rs2738048 ($r^2 = 0.71$) (ref. 5), and the other two signals at rs12716641 were located 76–77 kb and 124–125 kb away from rs2738058 and rs2738048, respectively (Fig. 2). While rs12716641 is located within the DEFA gene cluster, rs9314614 is located in the intron of the long-coding RNA GS1-24F4.2 and separated from the DEFA gene cluster by two recombination hotspots, likely representing an independent novel locus. Furthermore, all three are poorly correlated with rs10086568, an independent association within the DEFA locus reported in ref. 16 ($r^2 < 0.1$ in our samples and 1,000 genomes Asians, although $r^2 = 0.17$ with rs12716641 in Europeans). Multiple copy-number variants (CNVs) of DEFA1-A3 exist within this region$^{17}$ and were previously found to be associated with Crohn’s disease$^{18}$ and severe sepsis$^{19}$. Recently, rs4300027 has been reported to tag the CNVs in Europeans ($r^2 = 0.35$) (ref. 20). rs4300027 is in moderate LD with rs2738048 ($r^2 = 0.15$), rs2738058 ($r^2 = 0.12$) and rs12716641 ($r^2 = 0.28$), but not rs9314614 ($r^2 = 0.002$). This suggests that the associations at rs2738048/rs2738058 and rs12716641 may

Figure 1 | Recombination plots of the novel loci reaching genome-wide significance. (a) rs7190997 at 1p11.2, (b) rs7634389 at 3q27.3, (c) rs2074038 at 11p11.2 and (d) rs2033562 at 8q22.3, showing $P$ values obtained in the GWAS discovery (logistic regression) and in the combined analysis of GWAS and validation 1, 2 and 3 samples (fixed-effects meta-analysis).

Figure 2 | Recombination plots of the three independent loci at the defensin locus, showing $P$ values obtained in the GWAS discovery (logistic regression) and in the combined analysis of GWAS and validation 1, 2 and 3 samples (fixed-effects meta-analysis). The two novel signals are in low linkage disequilibrium with rs2738058, which tags the previously reported SNP rs2738048 and are separated from these SNPs by regions of high recombination rates.
implicate the role of DEFA1-A3 CNVs in IgAN. Further fine mapping and functional study will be needed to investigate the association of the DEFA1-A3 CNVs with IgAN and its relationship with the complex association patterns at these SNPs in Europeans and Asians.

Functional investigation through eQTL (expression quantitative trait loci) and GRAIL (genetic relationships across implicated loci) analyses. We then investigated potential biological effects of the novel associated SNPs and loci by looking for effects on mRNA expression levels (eQTLs)\textsuperscript{[21–22]}, ENCODE annotations of associated variants\textsuperscript{[23–24]}, documented associations with other diseases and known biological functions of nearby genes (Table 3, Supplementary Tables 13–15, Supplementary Data 2 and 3). Three out of the four novel loci showed strong associations with the expression levels of the genes ITGAX, ITGAM, ACCS, EXT2 and ST6GAL1 in peripheral blood cells and others, suggesting an important role of regulatory variants of these genes in IgAN risk. All five SNPs also tag variants that lie in predicted protein binding sites. Although no strong eQTL effects were observed for rs2033562 near ODF1-KLF10 in blood cells, ENCODE annotations suggest it may play a role in the expression of the nearby UBR5 gene\textsuperscript{[24]}. Furthermore, although there was no evidence for eQTL effects of rs2738048/rs2738058 and rs12716641, a moderate association with expression of the gene DEFB1 was observed at rs9314614 in monocytes (linear model $P = 4.22 \times 10^{-4}$; Supplementary Tables 13 and 15)\textsuperscript{[25]}, providing further support that this SNP could tag a novel independent locus and pathway despite its close proximity to the DEFA cluster.

These variants are also not in strong LD with those previously associated with other traits and diseases in Europeans or Asians (Table 3, Supplementary Table 13)\textsuperscript{[25]}, and a GRAIL (https://www.broadinstitute.org/mpg/grail)\textsuperscript{[26]} analysis of previously reported and novel loci did not identify any major pathways that can account for their associations with IgAN. We anticipate that these loci highlight multiple different pathways including mucosal immune response\textsuperscript{[16]} that may jointly influence IgAN pathogenesis through changes in gene expression levels. Since most of the existing

### Table 3 | Functional annotations of novel loci and associated SNPs (details on eQTLs and ENCODE annotations in Supplementary Tables 14 and 15, Supplementary Data 2 and 3)\textsuperscript{[16,21–25,37–43]}

| SNP/locus | eQTL/ENCODE | Other diseases | Known functions of genes |
|-----------|-------------|----------------|--------------------------|
| rs2074038 11p11.2 | Risk allele is strongly associated with increased ACCS gene expression levels in peripheral blood cells ($P < 9.81 \times 10^{-198}$), monocytes ($P = 1.07 \times 10^{-36}$), and B-cells ($P = 1.46 \times 10^{-29}$). It is also associated with increased EXT2 expression levels in peripheral blood cells ($P = 2.63 \times 10^{-23}$). This SNP and others in LD are also located at sites predicted with high likelihood to affect chromatin structure and protein binding. | Not previously associated with any other trait. | ACCS encodes a1-aminocyclopropane-1-carboxylate synthase homologue that belongs to the class-I pyridoxal-phosphate-dependent aminotransferase family. Shown to interact with the protein encoded by FBFI (Fas (TNFRSF6) binding factor 1), a keratin-binding protein that is required for epithelial cell polarization, apical junction complex assembly and angiogenesis. ST6GAL2 encodes ST6 beta galactosaminide alpha-2,6-sialyltransferase, a member of the sialyltransferase family involved in the generation of the cell-surface carbohydrate determinants and differentiation antigens. Regulates macrophage apoptosis via alpha2-6 sialylation of the TNFRI death receptor. May play a regulatory role in innate immune response. Up-regulated in human cancers. KLFI encodes a transcriptional repressor that acts as an effector of transforming growth factor beta signalling and activity of this protein may inhibit the growth of cancers. UBR5 encodes a E3 ubiquitin ligase which interacts with the deubiquitinating DUBA which in turns plays a role in IL-17 production in T-cells and inflammatory response in the small intestine. |
| rs7634389 3q27.3 | Risk allele is in strong LD ($r^2 > 0.9$ in our samples, HapMap Europeans and Asians) with variants associated with decreased expression levels of ST6GAL1 in peripheral blood cells (rs3821819; $P = 5.96 \times 10^{-20}$) and B-cells (rs17776120; $P = 1.61 \times 10^{-7}$). This SNP and others in LD may affect chromatin structure and protein binding. | IgG glycosylation, Type 2 diabetes, oesophageal cancer. Drug-induced liver injury. In moderate LD with rs11710456 in IgG glycosylation in Europeans ($r^2 = 0.528$) but not Asians ($r^2 = 0.12$). | |
| rs2033562 8q22.3 | No evidence for effects on gene expression levels in peripheral blood cells, monocytes, B-cells. In LD with SNPs with potential regulatory effects on UBR5 expression, chromatin structure and protein binding. | Chronic lymphocytic leukaemia (CLL) Not in LD with CLL-associated SNP. | |
expression data sets were generated in healthy individuals that were mostly of European descent\cite{21,22}, a more careful analysis will be needed to examine the association of these genotypes on the expression levels of these genes in Asian subjects, and to evaluate their biological and clinical relevance in IgAN patients and controls.

At the ST6GAL1 locus, we observed a trend of increasing allele frequencies of risk allele C at rs7634389 from Africans (16%) to Europeans (37.5%) to Asians (40%) in HapMap populations (Supplementary Table 16). Similarly, the ACCS SNP rs2074038 risk allele T is absent in Africans, present at moderate frequency in Europe (12%) and highest frequency in Asia (33%). Similar observations were made at two independent variants in the DEFA cluster (Supplementary Table 16). Consistent with previous reports on the geographical distribution of risk variants within 1q32, 22q12 (ref. 1), and more recently, 16p11.2 (ref. 16), these trends of increasing risk allele frequencies suggest that pooled differences in risk allele frequencies of these loci combined may contribute to differences in disease prevalence across different world populations\cite{1,4,16}.

**Discussion**

This study has several advantages over our previous study. Firstly, we have performed deep genome imputation in the discovery samples, leading to the discovery of loci tagged by SNPs that were not previously represented on SNP arrays. Secondly, the addition of 6,391 controls led to a moderate gain in statistical power to detect risk alleles with small effect sizes (OR<1.2). Thirdly, we have expanded our validation data sets to include both Northern and Southern Chinese samples that have provided firm independent replication of the association signals. We used geographic matching as a proxy for genetic matching to control the effects of population stratification in our validation data sets, as has been done in previous studies on the Chinese population\cite{27}, although ancestry informative markers may also be helpful. We also acknowledge a number of limitations. Despite deep imputation, we expect that the coverage of low frequency and rare variants (minor allele frequencies <5%) is limited given that most of the imputed rare variants in our study had a low information score that did not pass our strict quality control thresholds and were therefore excluded from further analysis. Also, with the current sample size, we had limited power to analyse lower-frequency variants with minor allele frequencies below 5% (Supplementary Table 2). Further studies will be needed to directly sequence or genotype rare variants on a larger number of cases and controls to evaluate their role in IgAN risk.

We have conducted the largest study on IgAN in the Han Chinese population to date by analysing a total of 8,313 IgAN cases and 19,680 controls. We have discovered three new loci at 11p11.2, 8q22.3 and 3q27.3, two novel independent associations within the DEFA-DEFB gene cluster and validated the recently reported locus at 16p11.2 (ref. 16). We estimate that these novel association signals explain about 1.7% of the disease variance and 5.5% of the variance in combination with the previously published loci\cite{42}. Some risk variants show significant difference in frequency and may contribute to IgAN prevalence differences across world populations. Our study has significantly expanded our understanding on the genetic basis of IgAN susceptibility and provided novel insight into the mechanisms underlying the development of IgAN.

**Methods**

**Study subjects.** The original genome-wide discovery analysis involved 1,523 cases (from Southern China) and 4,276 controls (972 controls from Southern China, 1,228 controls from Northern China and 2,076 Chinese controls from Singapore who share the same ancestral origin as the other Chinese controls)\cite{5}. To boost the statistical power of the current study, we included 6,511 control subjects of Chinese Han ethnicity from several of our previous GWAS including 981 subjects from Guangdong, 523 subjects from Shandong and 5,007 subjects from Singapore. For the validation study, three independent case–control samples were recruited from China as replication 1 (2,651 IgAN cases and 2,907 controls), replication 2 (2,428 cases and 4,202 controls) and replication 3 (1,800 IgAN cases and 1,910 controls) (Supplementary Note 1).

All the cases were histopathologically diagnosed by biopsy according to the following criteria: (i) immunofluorescence showing at least 2 + (scale 0 to 3 +) mesangial deposition of IgA, with IgA comprising the dominant immunoglobulin deposited in the glomeruli and (ii) excluding individuals with cirrhosis, Henoch–Schönlein purpura nephritis, hepatitis B-associated glomerulonephritis, HIV risk allele T is absent in Africans, present at moderate frequency in Europe (12%) and highest frequency in Asia (33%). Similar observations were made at two independent variants in the DEFA cluster (Supplementary Table 16). Consistent with previous results on the geographical distribution of risk variants within 1q32, 22q12 (ref. 1), and more recently, 16p11.2 (ref. 16), these trends of increasing risk allele frequencies suggest that pooled differences in risk allele frequencies of these loci combined may contribute to differences in disease prevalence across different world populations\cite{1,4,16}.

Sample genotyping and quality control. Genomic DNA was isolated from whole blood using a Qiaamp DNA extraction kit and quantified using a PicoGreen assay (Invitrogen). Genotyping analysis of the discovery samples was conducted using Human600-Quad (1,523 cases), Human610-Quad (1,953 southern and 1,228 northern Chinese) and the Human550K (1,022 Singaporean Chinese and 1,383 Shandong controls), Human 550K (1,022 Singaporean Chinese controls), Human 1M-Duo (930 Singaporean Chinese controls) and Human OmniExpress (1,133 Singaporean Chinese controls) BeadChips (Illumina). We excluded SNPs from the X, Y and mitochondrial chromosomes and focused all further analyses on autosomes. We performed identity by descent analysis using PLINK\cite{30} and 104 first degree relative pairs were identified; the relative with a lower sample call rate was excluded. Principal components analysis\cite{31} (Eigenvest v3.0: http://genepath.med.harvard.edu/~reich/Software.htm) was done on using a set on 47,462 common SNPs (MAF >1%) that were derived from 250,201 genotyped SNPs overlapping across all arrays. These SNPs were pruned to remove SNPs in LD (r^2 > 0.1, using PLINK --indep-pairwise 50 5 0.1) after exclusion of SNPs in the five conserved long-range LD regions in Chinese, namely the HLA region on chromosome 6, inversions on chromosomes 8 and 5 and two regions on chromosome 11 (refs. 5, 27). After principal components analysis, 16 outliers were detected based on principal component analysis (PCA) 1–5. We excluded 12,095 samples (1,434 cases and 10,661 controls) were left for the final analysis.

**Genotype imputation and quality control.** The software IMPUTE version 2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) was used for imputing genotype data of untyped SNPs in each data set following pre-phasing using SHAPEIT (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)\cite{6,8,22}. We imputed on the basis of only those genotyped SNPs that passed quality control thresholds, call rate >95%, Hardy–Weinberg equilibrium (HWE) P > 1 x 10^-4 in controls) in each of the data sets. The imputation was performed by using the multi-ethnic 1,000-genome reference panel (dated March 2012) consisting of 1,092 individuals from Africa, Asia, Europe and the Americas, which has been shown to outperform imputations with reference panels of matched ancestry\cite{5,20}. Quality control criteria after imputation are as follows: MAF >1%, Impute info >0.5 if MAF <0.5%, Impute info >0.8 if MAF <0.5%, HWE P in controls >1 x 10^-8 and average maximum posterior probabilities >0.99, leaving 3,792,915 SNPs for genotype dosage analyses in the final merged data set. For analysis of threshold-selected genotypes of 3,731,832 SNPs using GEMMA (http://www.xdlab.org/software.html)\cite{31,12}, we used the following quality control filters: call rate >0.95 after setting all genotypes with probabilities <0.9 to missing, MAF >0.01, HWE P in controls >0.05, test of differential missingness between cases and controls P >1E–8, imputation info >0.8.

**HLA imputation and analysis.** Imputation of two- and four-digit classical HLA alleles was performed using the SNP2HLA tool (https://www.broadinstitute.org/mpg/snp2hla/) and the Pan-Asian reference panel\cite{13} using the same set of SNPs overlapping across all arrays. These SNPs were pruned to remove SNPs in LD (r^2 > 0.1, using PLINK --indep-pairwise 50 5 0.1) after exclusion of SNPs in the five conserved long-range LD regions in Chinese, namely the HLA region on chromosome 6, inversions on chromosomes 8 and 5 and two regions on chromosome 11 (refs. 5, 27). After principal components analysis, 16 outliers were detected based on principal component analysis (PCA) 1–5. We excluded 12,095 samples (1,434 cases and 10,661 controls) were left for the final analysis.

**Genotyping and quality control in the validation study.** Genotyping analysis of the SNPs selected for validation was performed using the MassArray system from Sequenom. Loci-specific PCR and detection primers were designed using the MassArray Assay Design 3.0 software (Sequenom). SNPs that failed Sequenom design or genotyping in validation 2 and 3 samples were genotyped using Taqman assays (Life Technologies). TaqMan reactions were carried out in 5 µl volumes containing 10–20 ng DNA according to the manufacturer’s protocols. Fluorescence data were acquired on the ABI PRISM 7900 HT and SDS 2.4 software (Life Technologies) was used to call genotypes. For all SNPs, we examined the clustering patterns of genotypes and selected peak and slopes and confirmed that the genotype
calls were of good quality. All SNPs with call rates <95% and/or HWE \( P < 1 \times 10^{-6} \) in controls and all samples with call rates of <90% were removed from further analysis, with each batch of validation 1–3. After quality control, 124 SNPs from validation 1, 24 SNPs from validation 2 and 10 SNPs from validation 3 (including CFH SNP rs6677604) were left for further analysis.

**Association tests.** Imputed dosage data were analysed using SNPTEST ([https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html](https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html))\(^{39}\). Genome-wide case-control analysis was performed using frequentist tests under a missing data logistic regression model, as implemented in SNPTEST. We assumed an additive model for allelic risk, with the first five principal components as covariates to control for population stratification. To evaluate the effects of mixing Northern and Southern Chinese samples in the discovery analysis, we also split the discovery cohort into two clusters based on the first two principal components\(^{39}\), one of which is predominantly Northern Chinese (414 cases and 2,306 controls), and the other predominantly Southern Chinese (1,020 cases and 8,355 controls). In each cluster, we re-ran the principal components analysis and performed the association analysis adjusted for PCA1–5 in that cluster, then meta-analysed the PC-adjusted results.

The results were found to be similar and hence we kept the results from the first analysis in which we combined Northern and Southern samples. Finally, we ran Wald tests using a multivariate linear mixed model to correct for population stratification as implemented in the package GEMMA\(^{8}\). SNPs with \( P \) values differing by more than three \( \log_{10} \) between the two methods were excluded from further validation. Lambda 1,000 was calculated as a standardized estimate of the genomic inflation regardless of the sample size of the study\(^{39}\), using the following formula:

\[
\lambda = \frac{\text{mean}(1/\text{chr}.)}{\text{mean}(1/\text{all}.)} + 1/1,000
\]

For the validation studies, we performed the trend test in a logistic regression model, analysing samples from Northern and Southern regions of China in validation 2 as separate sample collections to control for potential confounding by population stratification. To combine the association statistics from the GWAS and the three validation cohorts, with the first five principal components as covariates we conducted a fixed-effects inverse variance meta-analysis using PLINK ([http://pngu.mgh.harvard.edu/~purcell/plink/](http://pngu.mgh.harvard.edu/~purcell/plink/))\(^{30}\). Tests for independent association signals were carried out using conditional logistic regression analyses implemented in PLINK\(^{30}\). Haplotypes analyses were conducted by phasing genotypes of interest using PHASE ([http://phase.isc.uji.es](http://phase.isc.uji.es)) software as implemented in PHASE. Analysed phased haplotypes using logistic regression analyses on PLINK\(^{30}\). Regional association plots, centred on the top SNP, were generated using LocusZoom ([http://csg.sph.umich.edu/locuszoom/](http://csg.sph.umich.edu/locuszoom/))\(^{36}\) to represent results from the logistic regression and fixed-effects meta-analysis at each stage of the study.

Furthermore, to evaluate the effects of age and gender on the association results, we entered age and gender as covariates into the logistic regression model and compared the results without adjustment within the same subset of samples with age and gender information available (6,523 cases and 17,349 controls; 84%). Each of the four stages were analysed separately and the results combined in a meta-analysis as before. Stratified analysis was performed by analysing males (3,067 cases and 9,993 controls) and females (3,324 cases and 7,371 controls) separately, among 6,391 cases and 17,364 controls with gender information available. Association results were also based on the combined meta-analysis across all four stages. Cochran’s Q test was used to test for heterogeneity in effect sizes between males and females, and also between Northern and Southern samples.

**Fraction of variance explained by loci.** The percentage of the total variance explained was estimated using the Nagelkerke’s pseudo \( R^2 \) using the R package ([http://cran.r-project.org/web/packages/fmsb/index.html](http://cran.r-project.org/web/packages/fmsb/index.html)), from the result of entering SNP genotypes and affection status into the glm function in R (v.2.15.1).

**References.**

1. Kiryluk, K. et al. Geographic differences in genetic susceptibility to IgA nephropathy: GWAS replication study and geospatial risk analysis. *PLoS Genet.* **8**, e1002765 (2012).
2. Wyatt, R. J. & Julian, B. A. IgA nephropathy. *N. Engl. J. Med.* **368**, 2402–2414 (2013).
3. Kiryluk, K., Novak, I. & Gharavi, A. G. Pathogenesis of immunoglobulin A nephropathy: recent insight from genetic studies. *Annu. Rev. Med.* **64**, 339–356 (2013).
4. Gharavi, A. G. et al. Genome-wide association study identifies susceptibility loci for IgA nephropathy. *Nat. Genet.* **43**, 321–327 (2011).
5. Yu, X., Chou, A. L., Zeng, Z. & Gharavi, A. G. Beta-defensin regions at 8p22-p23. *PLoS One* **8**, e1002765 (2012).
6. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
7. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
8. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
41. Liu, Z. et al. ST6Gal-I regulates macrophage apoptosis via alpha2-6 sialylation of the TNFR1 death receptor. *J. Biol. Chem.* **286**, 39654–39662 (2011).
42. Kudo, T. et al. Up-regulation of a set of glycosyltransferase genes in human colorectal cancer. *Lab. Invest.* **78**, 797–811 (1998).
43. Rutz, S. et al. Deubiquitinase DUBA is a post-translational brake on interleukin-17 production in T cells. *Nature* **518**, 417–421 (2015).

**Acknowledgements**

This work was funded by Guangdong Department of Science & Technology Translational Medicine Center grant (2011A080300002), the Specialized Research Fund for the Doctoral Program of Higher Education of China (2011017110088), the Science and Technology Planning Project of Guangdong Province, China (2013B051000019), Guangzhou Committee of Science and Technology, China (2012J5100031), Young Scientists Fund of National Natural Science Foundation of China (81200489), Yong Scholars Fund for the Doctoral Program of Higher Education of China (20120171120087), Natural Science Foundation of Guangdong Province, China (2014A030313136) and the Agency for Science & Technology and Research (A*STAR) of Singapore. We thank the staff in The First Affiliated Hospital of Sun Yat-sen University for help with sample collection, DNA extraction and clinical data collection, as well as the human genetics genotyping team at the Genome Institute of Singapore for help with the sample genotyping.

**Author contributions**

J.J.L. and X.Q.Y. organized and designed the study. M.L., J.N.F., K.Y.T., I.D.I., C.C.K. and Y.F.G. conducted and supervised the genotyping of samples. J.J.L., J.N.F. and H.Q.L. contributed to the design and execution of statistical analyses. P.R.Y., R.C.X. and X.Q.T. contributed to DNA extraction and clinical data collection. P.R.Y., R.C.X., X.Q.T., J.Q.W., A.K.A., J.X.B., O.R., M.H.C., C.Y.C., I.D.S., G.R.I., T.Y.W., H.L.L., T.A., Y.H.L., S.M.S., K.Y., R.P.E., Q.K.C., W.S., S.H.C., J.C., F.R.Z., S.P.L., G.X., E.S.T., L.W., N.C., X.I.Z., Y.X.Z., H.Z. and Z.H.L. conducted the recruitment of the samples. J.N.F., M.L., J.J.L. and X.Q.Y. drafted the manuscript. All authors contributed to the writing of the manuscript.

**Additional information**

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Li, M. et al. Identification of new susceptibility loci for IgA nephropathy in Han Chinese. *Nat. Commun.* **6**:7270 doi: 10.1038/ncomms8270 (2015).