European Journal of Human Genetics (2015) 23, 530–535

www.nature.com/ejhg

EJHG Open

ARTICLE

Evaluation of European coeliac disease risk variants in a north Indian population

Sabyasachi Senapati[1,6], Javier Gutierrez-Achury[2,6], Ajit Sood[3], Vandana Midha[4], Agata Szperl[2], Jihane Romanos[2], Alexandra Zhernakova[2], Lude Franke[2], Santos Alonso[5], B K Thelma[1,7], Cisca Wijmenga*[,2,7] and Gosia Trynka[2,7,8]

Studies in European populations have contributed to a better understanding of the genetics of complex diseases, for example, in coeliac disease (CeD), studies of over 23 000 European samples have reported association to the HLA locus and another 39 loci. However, these associations have not been evaluated in detail in other ethnicities. We sought to better understand how disease-associated loci that have been mapped in Europeans translate to a disease risk for a population with a different ethnic background. We therefore performed a validation of European risk loci for CeD in 497 cases and 736 controls of north Indian origin. Using a dense-genotyping platform (Immunochip), we confirmed the strong association to the HLA region (rs2854275, $P = 8.2 \times 10^{-49}$). Three loci showed suggestive association (rs4948256, $P = 9.3 \times 10^{-7}$, rs4758538, $P = 8.6 \times 10^{-5}$ and rs17080877, $P = 2.7 \times 10^{-5}$). We directly replicated five previously reported European variants ($P < 0.05$; mapping to loci harbouring *FASLG/TNFSF18*, *SCHIP1/IL12A*, *PFKFB3/PRKCQ*, *ZMIZ1* and *ICOSLG*). Using a transferability test, we further confirmed association at *PFKFB3/PRKCQ* (rs2387397, $P = 2.8 \times 10^{-4}$) and *PTPRK/THEMIS* (rs55743914, $P = 3.4 \times 10^{-4}$). The north Indian population has a higher degree of consanguinity than Europeans and we therefore explored the role of recessively acting variants, which replicated the HLA locus (rs9271850, $P = 3.7 \times 10^{-23}$) and suggested a role of additional four loci. To our knowledge, this is the first replication study of CeD variants in a non-European population.
European Journal of Human Genetics (2015) **23,** 530–535; doi:10.1038/ejhg.2014.137; published online 23 July 2014

INTRODUCTION

Hundreds of common variants have been established for immune-mediated diseases, mainly from genome-wide association studies (GWAS) in populations of European ancestry.[1] A large proportion of the European association signals replicate within non-European populations. Yet, additional risk variants can be found using multi-ethnic cohorts.[2–5]

Coeliac disease (CeD) is a common, autoimmune disorder, with a similar prevalence among Europeans and north Indians (1–3%).[6] The largest genetic risk to CeD is conferred by the HLA haplotypes, which account for approximately 35–40% of the risk.[7] Recent progress in understanding the genetic architecture of CeD has identified 39 non-HLA loci.[8–11] Although these analyses provided much insight into genetic risk factors for CeD, they were largely based on European populations, so that the risk of these associations in other non-European populations needed to be assessed.

We explored the contribution of reported risk variants and searched for new associated variants using a cohort of 1233 north Indian individuals. We acknowledge the fact that about half of this cohort contributed to the previously published Immunochip study. However, owing to the large size of the European cohorts, it is likely that any effect derived from the north Indian population would have been overwhelmed by the European effects (620 north Indian samples, which is only 2.5% of the total 24 269 samples). In this study, we have doubled the sample size of the north Indian cohort and analysed the transferability of known European variants to this population. We also took advantage of the higher degree of consanguinity in the north Indian population[12] and assessed the excess homozygosity to explore the possibility of recessively acting risk variants.

MATERIALS AND METHODS

Ethical statement

This study was approved by the respective institutional and university ethical committees. Informed written consent was received from all participants.

Study populations

We used two ethnically distinct cohorts from northern India and the Netherlands. The Indian cases ($n = 497$) and controls ($n = 736$) were recruited from Dayanand Medical College and Hospital, Ludhiana, in Punjab (northern India). The Indian controls included blood donors who tested negative for CeD serology. All the Indian subjects had self-reported northern Indian ethnicity. The Dutch cases ($n = 1150$) and controls ($n = 1173$) were recruited from three university medical centres (Utrecht, Leiden and VUmc Amsterdam) in the Netherlands. A small proportion of cases were recruited via the Dutch CeD patients' society. In both cohorts, Indian and Dutch CeD patients were diagnosed according to standard clinical, serological and histopathological

[1]Department of Genetics, University of Delhi South Campus, New Delhi, India; [2]Department of Genetics, University of Groningen, University Medical Hospital Groningen, Groningen, The Netherlands; [3]Department of Gastroenterology, Dayanand Medical College and Hospital, Ludhiana, India; [4]Department of Medicine, Dayanand Medical College and Hospital, Ludhiana, India; [5]Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, Leioa, Spain
*Correspondence: Professor C Wijmenga, Department of Genetics, University of Groningen, University Medical Centre Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands. Tel: +31 50 361710; Fax: +31 50 3617230; E-mail: c.wijmenga@umcg.nl
[6]These authors contributed equally to this study.
[7]These authors contributed equally to this study.
[8]Current address: Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
Received 14 October 2013; revised 10 June 2014; accepted 18 June 2014; published online 23 July 2014

criteria, following the ESPGHAN criteria.[13] Most of these samples have been described elsewhere.[11]

DNA extraction and genotyping

Most of the DNA samples were derived from blood, or from saliva for a small proportion of the Dutch cases and controls. Samples were hybridized on the Immunochip platform, a custom-made chip with 196 524 markers.[11] Genotyping was carried out according to Illumina's protocol at the genotyping facility of the University Medical Centre Groningen. The variant calls were exported from Genome Studio with human genome GRCh36/hg18 coordinates and further converted into GRCh37/hg19 coordinates using the UCSC liftOver tool.[14]

Genotype data quality control

We manually inspected cluster plots and adjusted them if required for about 150 000 markers. For about 20 000 SNPs, the assay performed poorly, and those variants were marked as uncalled and excluded from the final analysis. This left 178 111 high-quality SNPs that were exported from Illumina's Genome Studio. We required samples to have $\geq$98% call rate. Individuals showing a high degree of genetic relatedness (estimated kinship coefficient >0.2) or with a discordant sex were removed. Population outliers were identified by principal component analysis (PCA) implemented in EIGENSOFT 5.0.1.[15,16] We excluded markers with a call rate $<$99% or with significant deviation from Hardy–Weinberg equilibrium ($P>1\times10^{-3}$) and were left with 176 799 variants. A large number of Immunochip variants was of low frequency, but our study was underpowered to assess their association in the north Indian cohort; we therefore filtered out SNPs with a minor allele frequency (MAF) $<$10%. This excluded 88 083 and 87 442 variants in the north Indian and Dutch cohorts, respectively. We further only used the set of variants shared across the two cohorts, resulting in 88 716 SNPs in the final analysis. Genomic inflation for each cohort was estimated using 3016 'neutral' variants present on the Immunochip array for the replication of the GWAS for reading and writing skills and therefore unlikely to be confounded by the immune signals.

Statistical analysis

Chip-wide analysis. For both populations, we applied additive and recessive models using logistic regression, including the first four principal components and gender as covariates. We used an Immunochip-wide *P*-value cutoff of $P=3.5\times10^{-6}$ calculated as 0.05 divided by 14 136 linkage disequilibrium (LD)-independent SNPs ($r^2<0.1$ computed across 100 SNPs, with a 10-SNP sliding window) with MAF$>$10% and based on the north Indian cohort to determine a significant association. We used $P<1\times10^{-4}$ to report a suggestive association. To reassess the significance of the associations detected in this test, we also performed a minimum of 100 000 permutations to compute empirical *P*-values.

Replication. We used two approaches to test for replication of the 39 non-HLA CeD loci that have been reported in Europeans[11] in the north Indian cohort: the exact SNP replication, and the locus transferability.

- Exact SNP replication: the reported index SNPs (representing primary as well as secondary, independent signals) were tested for association in the north Indian cohort. We considered replication to be significant at $P<0.05$. During the quality control, we removed 13 of the total of 57 independent SNPs associated to the 39 loci because of MAF$<$10%. The exact SNP replication was applied to 44 reported SNPs, representing 38 distinct loci.
- Locus transferability test: we performed a transferability test to account for LD differences across different ethnic populations. If the causal SNP is poorly tagged by the associated SNP reported in the discovery population (European), the differences in LD between the north Indian and Dutch populations could result in either a lack of replication or a different localization of the association signal. Published LD boundaries were used to define each of the 39 non-HLA loci.[11] First, we identified all the LD-correlated markers for each locus, measured by $r^2\geq0.05$ between the index SNP and all the variants present in CeD loci using the Dutch controls. We then checked the association status of these variants in the north Indian Immunochip data set for transferability. SNP rs13132308 from the *IL2/IL21* locus was not present in the north Indian data set and was replaced by its best proxy SNP rs13151961 ($r^2>0.9$ based on the Dutch controls). The *ELMO1* locus and *SCHIP1/IL12A* locus were excluded because the index variants for these loci, or their proxies failed the 10% MAF filter. We required at least one significant variant per locus. The significance threshold was assessed in two ways. First, across all tested loci (37) we identified 142 LD-independent SNPs (pairwise $r^2<0.1$) and used $P=0.05/142=3.52\times10^{-4}$ as the significance threshold. Second, we used a permutation-based test as an alternative way to account for LD structure. We performed 1000 permutations for which we randomized phenotype labels. We identified a nominal *P*-value for each of the 37 loci in each of the permutation results. We then defined the 5th percentile of the nominal *P*-values as a significance threshold.

Runs of homozygosity (ROH). The north Indian population has a reported higher degree of consanguinity,[12,17–19] therefore we sought to identify if there were significantly associated runs of homozygous variants. We first pruned both data sets for LD using an $r^2\geq0.8$ as a threshold to avoid detecting false-positive regions because of homozygous stretches of LD-correlated SNPs.[20–22] This is particularly likely to occur for the Immunochip, as it includes regions with a high density of markers that are often in strong LD with each other. Then, we identified regions with ROH, which were defined as at least 50 contiguous homozygous SNPs with no interruption (without heterozygous positions in between) and with less than 500-kb distance between neighboring SNPs. Association of ROH regions was performed using PLINK v1.07.[23] We considered a locus significant after Bonferroni correction using the total number of ROH found per data set (8) in our analysis with a $P<0.006$. We assessed the false discovery rate (FDR) by running 10 000 permutations in which we randomly shuffled the phenotypes and reassessed the significance of association.

Data availability

Genotype data for the north Indian cohort are available through the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/) under the accession number EGAS00001000849. The summary statistics can be obtained by contacting the authors directly.

RESULTS

We obtained high-quality genotype data for 88 716 common SNPs that were shared between the north Indian and Dutch cohorts. PCA revealed different clustering patterns between the two cohorts, consistent with their different ethnicities. Furthermore, the case–control samples were homogeneous within each of the data sets (Supplementary Figure 1). We used the first four principal components to control for possible population substructure and did not observe significant genomic inflation in any of the data sets ($\lambda_{\text{India}}=1.036$, $\lambda_{\text{Netherlands}}=1.039$).

The strongest association in the north Indian cohort was present at the HLA locus. Comparable to the Dutch sample, the SNP rs2854275 was the top signal ($p_{\text{India}}=8.17\times10^{-49}$, $\text{OR}_{\text{India}}=6.97$ (5.38–9.04)); $p_{\text{Dutch}}=4.413\times10^{-121}$, $\text{OR}_{\text{Dutch}}=10.08$ (8.3–12.23)). This replicated the known, very strong effect of HLA in CeD. No other locus reached genome-wide significance (Figure 1 and Table 3). The rs4948256 at 10q21.2 in the *ANK3* gene reached our Immunochip-wide significance threshold ($P=9.28\times10^{-7}$, $p_{\text{permuted}}=1\times10^{-6}$). The rs4758538 at 11p15.4 in the *OSBPL5* gene ($P=8.57\times10^{-5}$, $p_{\text{permuted}}=7.93\times10^{-5}$) and rs17080877 at 18q22.2 near the *DOK6* gene ($P=2.68\times10^{-5}$, $p_{\text{permuted}}=1.7\times10^{-5}$) both were suggestively associated. None of these associations could be replicated in the Dutch cohort.

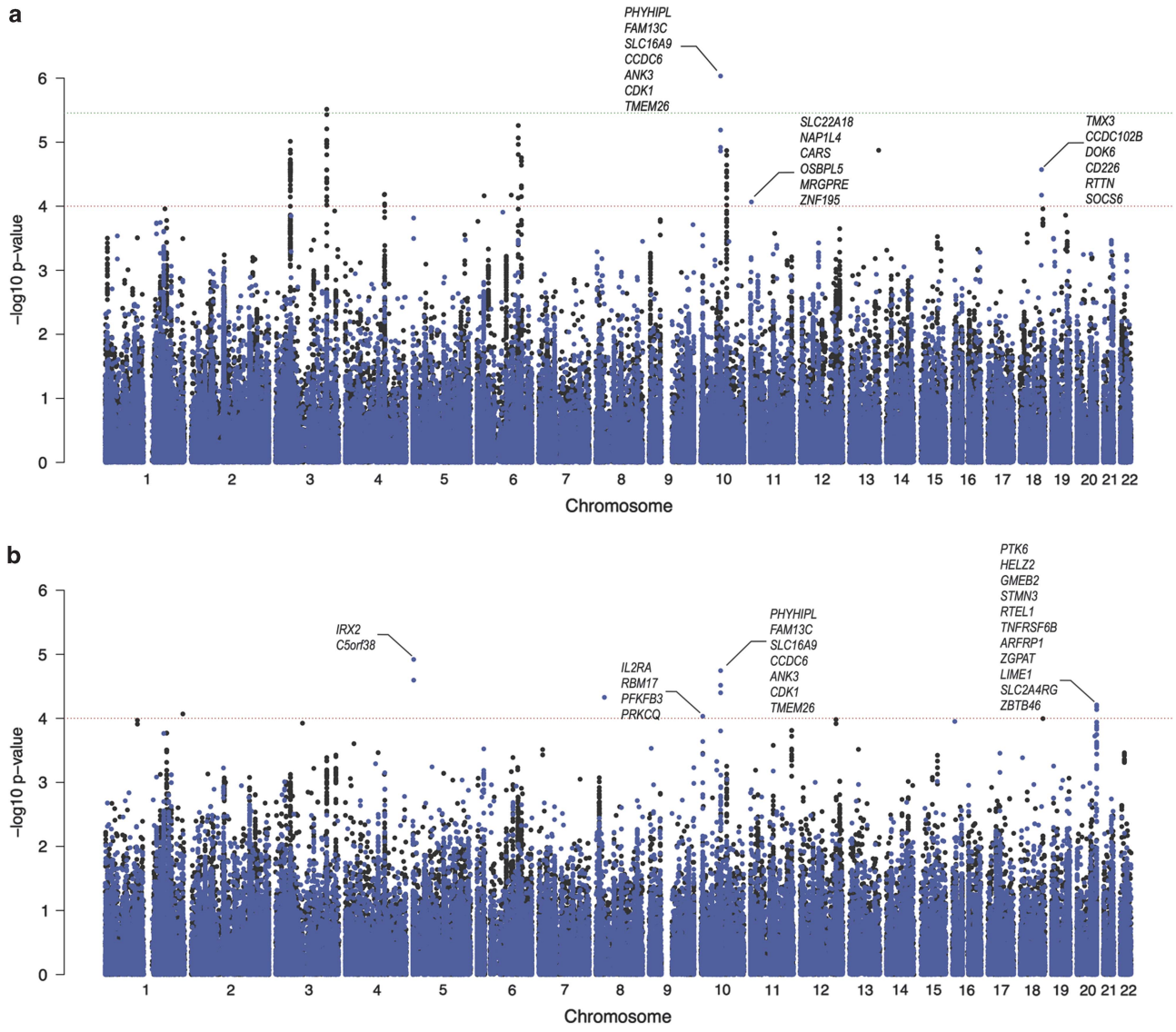

Figure 1 Additive and recessive association results in the north Indian population. Manhattan plot of association under (**a**) an additive and (**b**) a recessive model for the north Indian (blue dots) and Dutch populations (black dots). The Immunochip-wide significance cutoff ($P=3.5\times10^{-6}$) is depicted as a green dotted line and suggestive significance threshold as a red dotted line ($P=1.0\times10^{-4}$). We excluded the HLA region.

Replication of European associations in the north Indian cohort using exact and transferability tests

The exact association in the north Indian cohort was observed at five previously reported loci, corresponding to *FASLG/TNFRSF18* (rs12142280, $P=0.002$, OR = 0.69 (0.54–0.87)), *ICOSLG* (rs58911644, $P=0.004$, OR = 0.68 (0.51–0.88)), *PFKFB3/PRKCQ* (rs2387397, $P=0.02$, OR = 0.77 (0.61–0.96)), *SCHIP1/IL12A* (rs2561288, $P=0.03$, OR = 1.2 (1.01–1.44)) and *ZMIZ1* (rs1250552, $P=0.03$, OR = 1.2 (1.01–1.44)) (Table 1). At all these variants, the directions of association were the same as those previously reported (Supplementary Table 1). Three of these five loci, *FASLG/TNFRSF18*, *ZMIZ1* and *ICOSLG*, previously reached $P<0.05$ in the north Indian samples used in the initial study by Trynka *et al*[11] (Supplementary Table 1).

Using the transferability analysis of 39 loci, we observed two associations, which passed two independent significance thresholds (Table 2). One was at the *PFKFB3/PRKCQ* locus represented by the SNP rs744254 ($P=2.8\times10^{-4}$, OR = 0.70 (0.57–0.84)); the other was at the *PTPRK/THEMIS* locus, rs4142030 ($P=3.4\times10^{-4}$, OR = 1.39 (1.16–1.66)) (Supplementary Table 2).

Recessive analysis in the north Indian population

The ROH analysis identified 16 genomic regions with >50 consecutive homozygous SNPs each; two regions showed suggestive association with CeD in the north Indian cohort. They were located on chromosomes 2q11.2 ($P=4.1\times10^{-3}$, FDR = 0.11) and 6p21.33 ($P=1.4\times10^{-4}$, FDR = 0.0001) (Supplementary Table 3). None of those regions replicated in the Dutch cohort.

Complementary to the ROH analysis, we also applied a recessive model to test for association across all 88 716 SNPs. Apart from the SNP rs9271850 ($P=3.67\times10^{-23}$, OR = 0.12 (0.08–0.18)) located in the HLA locus, we observed four variants (MAF > 0.25) at four loci that reached the threshold of suggestive association ($P<1\times10^{-4}$; Figure 1 and Table 3). The rs744253 at 10p15.1 ($P=9.27\times10^{-5}$, OR = 0.3859; $p_{\text{permuted}}=6.7\times10^{-5}$) corresponds to *PFKFB3/PRKCQ* locus that was replicated in the direct association and

transferability tests described above. The remaining three loci were new suggestive associations, with the strongest observed at the intergenic SNP rs963872 ($P = 1.2 \times 10^{-5}$, OR = 2.01 (1.47–2.75); $p_{permuted} = 1.2 \times 10^{-5}$) at 5p15.33, near *IRX2* gene, followed by rs10994257 ($P = 1.8 \times 10^{-5}$, OR = 2.5 (1.67–3.99); $p_{permuted} = 5 \times 10^{-6}$) at 10q21.2 mapping to the *ANK3* gene and rs6010998 ($P = 6.17 \times 10^{-5}$, OR = 2.8 (1.67–4.66); $p_{permuted} = 2.5 \times 10^{-5}$) at 20q13.33 in the *RTEL1* gene.

DISCUSSION

Recent trans-ethnic studies have indicated that a large proportion of common disease- or trait-associated variants established in populations of European descent also contribute to the risk in other ethnic groups.[4,24,25] Yet this observation cannot be generalized because only a few of the GWAS were conducted in non-European populations.[26] Novel loci have been reported in these limited studies.[27,28] Thus, more multi-ethnic studies may help identify not only shared disease determinants, but also additional risk variants, which may provide leads toward a better understanding of the disease biology.[29]

To our knowledge, there have been very few genetic association studies for CeD performed in non-European populations, and those that have been published have focused only on the HLA effects based on very small sample size.[30–36] In this study, we attempted to validate known genetic determinants of CeD and identify potentially novel loci in the north Indian population. We used the Immunochip genotyping platform for this study because it offers the highest coverage of regions associated to immune-related disorders, including those conferring risk to CeD.

Given our modest sample size and the limited power of our study, it is likely that the real replication rate of the European associations in the north Indian population is much greater than we can report here. In this context, it should be mentioned that although half of the samples of the north Indian cohort were previously analysed in the study by Trynka *et al*,[11] the study focused strongly on the effects in samples of European ancestry (97% of the 24 269 samples were of European descent). In this study, we wanted to focus specifically on the role of the European variants in the north Indian population and to explore potential new associations that could have been diluted in the previous study because of the dominance of the European cohorts.

Table 1 Five loci that were significantly ($P < 0.05$) replicated in the north Indian cohort in the exact SNP test

Locus	*Chr*	*SNP ID*	*Genomic position (bp)*[a]	*Allele*	*Frequency cases*	*Frequency controls*	P-*value*	*OR (95% CI)*
FASLG/ TNFSF18	1	rs12142280	172 864 652	A	0.15	0.19	2.38E–03	0.69 (0.54–0.87)
SCHIP1/IL12A	3	rs2561288	159 674 928	T	0.48	0.48	0.038	1.2 (1.01–1.44)
PFKFB3/PRKCQ	10	rs2387397	6 390 192	G	0.19	0.23	0.020	0.77 (0.62–0.96)
ZMIZ1	10	rs1250552	81 058 027	A	0.45	0.41	0.039	0.83 (0.70–0.99)
ICOSLG	21	rs58911644	45 629 121	T	0.11	0.15	4.23E–03	0.68 (0.52–0.88)

[a]Coordinates are provided in human genome reference hg19.

Table 2 Two loci significantly replicated in the north Indian cohort in the transferability test

Locus	*Chr*	*Immunochip*	*Top transferable-SNP*	*Genomic position (bp)*[a]	*Allele*	*MAF*	P-*value*	*OR (95% CI)*	P-*value cutoff (LD-independent SNPs)*	*Permuted, 5%-ile* P-*value cutoff*
PTPRK/THEMIS	6	rs55743914	rs4142030	128 154 686	G	0.40	3.39E–04	1.39 (1.16–?1.67)	6.0E–04	1.32E–03
PFKFB3/PRKCQ	10	rs2387397	rs744254	6 392 848	G	0.36	2.80E–04	0.70 (0.58–0.84)	4.0E–04	8.25E–04

[a]Coordinates are provided in human genome reference hg19.

Table 3 Association results for Immunochip-wide additive and recessive models in the north Indian population compared with the same model applied to the Dutch cohort

								India		*The Netherlands*	
Locus	*Chr*[a]	*SNP*	*Genomic position (bp)*[a]	*Allele*	*Frequency Cases*	*Frequency Controls*	*Model*	P-*value*	*OR*	P-*value*	*OR*
HLA	6	rs2854275	32 736 406	T	0.50	0.15	Additive	6.97	8.17E–49	4.41E–121	10.08
ANK3	10	rs4948256	61 956 912	A	0.30	0.22	Additive	1.66	9.28E–07	0.01395	1.202
OSBPL5	11	rs4758538	3 162 394	T	0.23	0.28	Additive	0.65	8.57E–05	0.8804	0.9889
Intergenic (close to *DOK6*)	18	rs17080877	67 016 479	C	0.14	0.10	Additive	1.82	2.68E–05	0.4274	0.9276
HLA	6	rs9271850	32 595 060	A	0.33	0.60	Recessive	0.12	3.68E–23	2.18E–47	0.1374
Intergenic (close to *IRX2*)	5	rs963872	2 857 542	C	0.48	0.42	Recessive	2.01	1.20E–05	0.4407	1.168
PFKFB3/PRKCQ	10	rs744253	6 393 009	A	0.29	0.34	Recessive	0.39	9.27E–05	0.7184	0.9385
ANK3	10	rs10994257	61 982 048	T	0.34	0.26	Recessive	2.59	1.80E–05	0.7731	1.041
RTEL1/TNFRSF6B	20	rs6010998	62 294 402	A	0.28	0.25	Recessive	2.81	6.17E–05	0.1459	1.279

We used an Immunochip-wide *P*-value cutoff of $P = 3.5 \times 10^{-6}$ to report a significant association and $P = 1 \times 10^{-4}$ for suggestive association.
[a]Coordinates are provided in human genome reference hg19.

Our study suffers from a small sample size and lack of replication, for these reasons we acknowledge that the results presented here need to be validated further in future replication studies.

We successfully replicated HLA and five of the 38 tested European SNPs (13%); these included the loci harbouring the *FASLG/TNFSF18*, *SCHIP1/IL12A*, *PFKFB/PRKCQ*, *ZMIZ1* and *ICOSLG* genes. The transferability test further confirmed association to the *PFKFB/PRKCQ* locus and identified an additional association to a variant at the *PTPRK/THEMIS* locus with suggestive evidence of association. The top association at this locus, rs4142030, has a very low correlation with the European index variants ($r^2 = 0.083$, $D' = 0.53$) and is located in intron 2 of *THEMIS*, in contrast to the European-associated SNP, which is located in intron 30 of *PTPRK*. Given that our study was underpowered, we acknowledge that these suggestive associations must be further replicated in north Indian populations. However, it is tempting to speculate that this localization of the association signal could result from the independent effects of each of the genes in the different ethnicities. Both genes are of functional relevance to CeD aetiology and a recent study showed that the mRNA levels between these two genes have higher expression levels in CeD patients but not in controls.[37]

The analysis of the recessive effects implicated multiple genomic regions at a suggestive significance. With the run of homozygosity test, we found a suggestive association to a region at 2q11.2 that spans 1.14 Mb and includes the *LYG1*, *TXNDC9*, *EIF5B*, *REV1*, *AFF3* and *LONRF2* genes. This region has been associated with multiple immune-related phenotypes, including type 1 diabetes (T1D) and rheumatoid arthritis (RA).[38–40] The second region maps to the HLA locus at 6p21 and contains the *MICA* gene, which is also a locus with a pleiotropic effect in phenotypes such as Behcet's disease, RA or HCV-induced hepatocellular carcinoma.[41–47] The Immunochip-wide analysis under a recessive model suggested four loci that have also been linked to other phenotypes, for example, *ANK3*, which has been associated with psychiatric disorders,[48,49] and with other loci that include *PFKFB3/PRKCQ* and *RTEL1/TNFRSF6B* reported to be associated with immune-related diseases such as T1D, RA and irritable bowel disease.[39,50–63] Apart from HLA, the *ANK3* locus was the only one that passed Immunochip-wide significance threshold when we considered the additive model. None of these loci were replicated in the Dutch cohort, which could reflect a population-specific effect, but they require further replication in the north Indian population to robustly establish their association with a genome-wide significance. Although the genes mapping to the associated loci that we report here appear to be relevant candidates for CeD pathogenesis, our study was biased toward identifying biologically relevant genes because the Immunochip was designed to specifically target regions previously reported as associated to immune-related diseases.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Jackie Senior for carefully reading the manuscript. This study was funded by grants from the Celiac Disease Consortium and an Innovative Cluster approved by the Netherlands Genomics Initiative. Partial funding was provided by the Dutch Government (BSIK03009 to CW) and the Netherlands Organisation for Scientific Research (NWO, grant 918.66.620 to CW). CW received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC grant 2012-322698. GT was supported by a Rubicon grant from the Netherlands Organization for Scientific Research (NWO). BKT received funding from a JC Bose fellowship, Department of Science and Technology, Government of India. SS is supported by a Senior Research Fellowship from the Council for Scientific and Industrial Research (CSIR), New Delhi, India. AZ is supported by a grant from the Dutch Reumafonds (11-1-101) and a Rosalind Franklin Fellowship from the University of Groningen, The Netherlands.

1 Welter D, Macarthur J, Morales J *et al*: The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001–D1006.
2 Ng MC, Saxena R, Li J *et al*: Transferability and fine mapping of type 2 diabetes loci in African Americans: the Candidate Gene Association Resource Plus Study. *Diabetes* 2013; **62**: 965–976.
3 Liu CT, Ng MC, Rybin D *et al*: Transferability and fine-mapping of glucose and insulin quantitative trait loci across populations: CARe, the Candidate Gene Association Resource. *Diabetologia* 2012; **55**: 2970–2984.
4 Teslovich TM, Musunuru K, Smith AV *et al*: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; **466**: 707–713.
5 Bustamante CD, Burchard EG, De la Vega FM: Genomics for the world. *Nature* 2011; **475**: 163–165.
6 Makharia GK, Verma AK, Amarchand R *et al*: Prevalence of celiac disease in the northern part of India: a community based study. *J Gastroenterol Hepatol* 2011; **26**: 894–900.
7 Gutierrez-Achury J, Coutinho de Almeida R, Wijmenga C: Shared genetics in coeliac disease and other immune-mediated diseases. *J Intern Med* 2011; **269**: 591–603.
8 van Heel DA, Franke L, Hunt KA *et al*: A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007; **39**: 827–829.
9 Hunt KA, Zhernakova A, Turner G *et al*: Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008; **40**: 395–402.
10 Dubois PC, Trynka G, Franke L *et al*: Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010; **42**: 295–302.
11 Trynka G, Hunt KA, Bockett NA *et al*: Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* 2011; **43**: 1193–1201.
12 Reich D, Thangaraj K, Patterson N, Price AL, Singh L: Reconstructing Indian population history. *Nature* 2009; **461**: 489–494.
13 Husby S, Koletzko S, Korponay-Szabo IR *et al*: European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease. *J Pediatr Gastroenterol Nutr* 2012; **54**: 136–160.
14 Hinrichs AS, Karolchik D, Baertsch R *et al*: The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006; **34**: D590–D598.
15 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
16 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
17 Narang A, Jha P, Rawat V *et al*: Recent admixture in an Indian population of African ancestry. *Am J Hum Genet* 2011; **89**: 111–120.
18 Metspalu M, Romero IG, Yunusbayev B *et al*: Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet* 2011; **89**: 731–744.
19 Moorjani P, Thangaraj K, Patterson N *et al*: Genetic evidence for recent population mixture in India. *Am J Hum Genet* 2013; **93**: 422–438.
20 Lencz T, Lambert C, DeRosse P *et al*: Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci USA* 2007; **104**: 19942–19947.
21 McQuillan R, Leutenegger AL, Abdel-Rahman R *et al*: Runs of homozygosity in European populations. *Am J Hum Genet* 2008; **83**: 359–372.
22 Keller MC, Simonson MA, Ripke S *et al*: Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet* 2012; **8**: e1002656.
23 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
24 Waters KM, Stram DO, Hassanein MT *et al*: Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet* 2010; **6**: pii: e1001078.
25 Kurreeman F, Liao K, Chibnik L *et al*: Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet* 2011; **88**: 57–69.
26 Need AC, Goldstein DB: Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 2009; **25**: 489–494.
27 Negi S, Juyal G, Senapati S *et al*: A genome-wide association study reveals ARL15, a novel non-HLA susceptibility gene for rheumatoid arthritis in north Indians. *Arthritis Rheum* 2013; **65**: 3026–3035.
28 Tabassum R, Chauhan G, Dwivedi OP *et al*: Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* 2013; **62**: 977–986.
29 Okada Y, Wu D, Trynka G *et al*: Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; **506**: 376–381.

30 Brar P, Lee AR, Lewis SK, Bhagat G, Green PH: Celiac disease in African-Americans. *Digest Dis Sci* 2006; **51**: 1012–1015.
31 Almeida RC, Gandolfi L, De Nazare Klautau-Guimaraes M *et al*: Does celiac disease occur in Afro-derived Brazilian populations? *Am J Hum Biol* 2012; **24**: 710–712.
32 Remes-Troche JM, Ramirez-Iglesias MT, Rubio-Tapia A, Alonso-Ramos A, Velazquez A, Uscanga LF: Celiac disease could be a frequent disease in Mexico: prevalence of tissue transglutaminase antibody in healthy blood donors. *J Clin Gastroenterol* 2006; **40**: 697–700.
33 Perez-Bravo F, Araya M, Mondragon A *et al*: Genetic differences in HLA-DQA1* and DQB1* allelic distributions between celiac and control children in Santiago, Chile. *Hum Immunol* 1999; **60**: 262–267.
34 Akbari MR, Mohammadkhani A, Fakheri H *et al*: Screening of the adult population in Iran for coeliac disease: comparison of the tissue-transglutaminase antibody and anti-endomysial antibody tests. *Eur J Gastroenterol Hepatol* 2006; **18**: 1181–1186.
35 Srivastava A, Yachha SK, Mathias A, Parveen F, Poddar U, Agrawal S: Prevalence, human leukocyte antigen typing and strategy for screening among Asian first-degree relatives of children with celiac disease. *J Gastroenterol Hepatol* 2010; **25**: 319–324.
36 El-Akawi ZJ, Al-Hattab DM, Migdady MA: Frequency of HLA-DQA1*0501 and DQB1*0201 alleles in patients with coeliac disease, their first-degree relatives and controls in Jordan. *Ann Tropl Paediatr* 2010; **30**: 305–309.
37 Bondar C, Plaza-Izurieta L, Fernandez-Jimenez N *et al*: THEMIS and PTPRK in celiac intestinal mucosa: coexpression in disease and after *in vitro* gliadin challenge. *Eur J Hum Genet* 2013; **22**: 358–362.
38 Sandholm N, Salem RM, McKnight AJ *et al*: New susceptibility loci associated with kidney disease in type 1 diabetes. *PLoS Genet* 2012; **8**: e1002921.
39 Stahl EA, Raychaudhuri S, Remmers EF *et al*: Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 2010; **42**: 508–514.
40 Todd JA, Walker NM, Cooper JD *et al*: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007; **39**: 857–864.
41 Bratanic N, Smigoc Schweiger D, Mendez A, Bratina N, Battelino T, Vidan-Jeras B: An influence of HLA-A, B, DR, DQ, and MICA on the occurrence of celiac disease in patients with type 1 diabetes. *Tissue Antigens* 2010; **76**: 208–215.
42 Tinto N, Ciacci C, Calcagno G *et al*: Increased prevalence of celiac disease without gastrointestinal symptoms in adults MICA 5.1 homozygous subjects from the Campania area. *Digest Liver Dis* 2008; **40**: 248–252.
43 Martin-Pagola A, Ortiz L, Perez de Nanclares G, Vitoria JC, Castano L, Bilbao JR: Analysis of the expression of MICA in small intestinal mucosa of patients with celiac disease. *J Clin Immunol* 2003; **23**: 498–503.
44 Fernandez L, Fernandez-Arquero M, Gual L *et al*: Triplet repeat polymorphism in the transmembrane region of the MICA gene in celiac disease. *Tissue Antigens* 2002; **59**: 219–222.
45 Kumar V, Kato N, Urabe Y *et al*: Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat Genet* 2011; **43**: 455–458.
46 Hou S, Yang Z, Du L *et al*: Identification of a susceptibility locus in STAT4 for Behcet's disease in Han Chinese in a genome-wide association study. *Arthritis Rheum* 2012; **64**: 4104–4113.
47 Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ: Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Hum Mol Genet* 2011; **20**: 3494–3506.
48 Chen SJ, Chao YL, Chen CY *et al*: Prevalence of autoimmune diseases in in-patients with schizophrenia: nationwide population-based study. *Br J Psychiatry* 2012; **200**: 374–380.
49 Ludvigsson JF, Osby U, Ekbom A, Montgomery SM: Coeliac disease and risk of schizophrenia and other psychosis: a general population cohort study. *Scand J Gastroenterol* 2007; **42**: 179–185.
50 Anderson CA, Boucher G, Lees CW *et al*: Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011; **43**: 246–252.
51 Athanasiu L, Mattingsdal M, Kahler AK *et al*: Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *J Psychiatr Res* 2010; **44**: 748–753.
52 Barrett JC, Clayton DG, Concannon P *et al*: Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 2009; **41**: 703–707.
53 Chen DT, Jiang X, Akula N *et al*: Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol Psychiatry* 2013; **18**: 195–205.
54 Cooper JD, Smyth DJ, Smiles AM *et al*: Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 2008; **40**: 1399–1401.
55 Ferreira MA, O'Donovan MC, Meng YA *et al*: Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 2008; **40**: 1056–1058.
56 Franke A, McGovern DP, Barrett JC *et al*: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**: 1118–1125.
57 Hsu YH, Zillikens MC, Wilson SG *et al*: An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS Genet* 2010; **6**: e1000977.
58 Liu Y, Blackwood DH, Caesar S *et al*: Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol Psychiatry* 2011; **16**: 2–4.
59 Rajaraman P, Melin BS, Wang Z *et al*: Genome-wide association study of glioma and meta-analysis. *Hum Genet* 2012; **131**: 1877–1888.
60 Raychaudhuri S, Remmers EF, Lee AT *et al*: Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* 2008; **40**: 1216–1223.
61 Sanson M, Hosking FJ, Shete S *et al*: Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet* 2011; **20**: 2897–2904.
62 Shete S, Hosking FJ, Robertson LB *et al*: Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* 2009; **41**: 899–904.
63 Sklar P, Ripke S, Scott LJ *et al*: Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 2011; **43**: 977–983.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)