Novel celiac disease genetic determinants related to the immune response

Karen A Hunt1, Alexandra Zhernakova2, Graham Turner3, Graham A.R. Heap1, Lude Franke2, Marcel Bruinenberg4, Jihane Romanos4, Lotte C. Dinesen5, Anthony W Ryan3, Davinder Panesar1, Rhian Gwilliam6, Fumihiko Takeuchi6, William M McLaren6, Geoffrey KT Holmes7, Peter D Howdle8, Julian RF Walters9, David S Sanders10, Raymond J Playford1, Gosia Trynka4, Chris JJ Mulder11, M Luisa Mean12,13, Wieke HM Verbeek11, Valerie Trimble3, Fiona M Stevens14, Colm O’Morain3, Nicholas P Kennedy3, Dermot Kelleher3, Daniel J Pennington1, David P Strachan15, Wendy L McArdle16, Charles A Mein17, Martin C Wapenaar4, Panos Deloukas6, Ralph McGinnis6, Ross McManus3,1, Cisca Wijmenga2,4,17, and David A van Heel1,4

1Institute of Cell and Molecular Science, Queen Mary University of London, 4 Newark Street, London E1 2AT, UK 2Complex Genetics Section, DBG-Dept Biomedical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands 3Departments of Clinical Medicine, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland 4Genetics Department, University Medical Center and Groningen University, Groningen, 9700 RB Groningen, The Netherlands 5Gastroenterology Unit, University of Oxford, Oxford OX3 7BN, UK 6Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK 7Department of Gastroenterology, Derbyshire Royal Infirmary, London Road, Derby DE1 2QY, UK 8Academic Medical Unit, Leeds Institute of Molecular Medicine, St James’s University Hospital, Leeds LS9 7TF 9Gastroenterology Section, Imperial College London, Hammersmith Hospital, London W12 0HS, UK 10Department of Gastroenterology and Liver Unit, Royal Hallamshire Hospital, Sheffield S10 2JF, UK 11Department of Gastroenterology, VU Medical Center, 1007 MB Amsterdam, The Netherlands 12Department of Pediatric Gastroenterology, VU Medical Center, 1007 MB Amsterdam, The Netherlands 13Department of Paediatric Gastroenterology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands 14Department of Medicine, National University of Ireland, Galway, Ireland 15Division of Community Health Sciences, St George’s University of London, Cranmer Terrace, London SW17 0RE, UK 16Avon Longitudinal Study of Parents and Children (ALSPAC), University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK 17The Genome Centre, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

Author Information Correspondence and requests for materials should be addressed to D.A.vH (d.vanheel@qmul.ac.uk). *These authors were joint principal investigators and contributed equally to this work

Author Contributions Statement LF and DAVH analysed UK GWAS data, selected SNPs and designed assays for golden gate genotyping. Substantial contributions to sample collections were made by DAVH, LD, GKTH, PH, JRFW, DSS (UK2 cases); DPS, WLMcA (1958 cohort controls); CJM, WV, MLM (DUTCH samples); VT, FMS, COM, NPK, DK (IRISH samples). UKGWAS genotyping was performed as described in PD lab2. KAH extracted UKGWAS and UK2 celiac DNA samples and performed UK2 sample golden gate genotyping. GrahamT and AWR prepared Irish DNA samples. GrahamT, AWR and KAH performed Irish sample golden gate genotyping. UK2 and IRISH genotyping was performed in CAM lab, DP performed quality control steps. AZ prepared DUTCH celiac and control DNA samples, and AZ and JR performed DUTCH sample golden gate genotyping in CW lab. DVH and KAH performed final golden gate genotype clustering on all samples, with assistance from RG. LD and DAVH collected Paxgene RNA celiac blood samples, GH extracted Paxgene RNA, GH and MB performed expression chips in CW lab. GH and LF analysed expression data. GosiaT performed IL18RAP re-sequencing. MCW processed intestinal biopsies, MB and MCW performed expression chips in CW lab, MCW and GH analysed expression data. DJP performed analysis of genes in intestinal T cell subsets. KAH and GH performed bioinformatics and annotation of celiac risk variant regions DAVH, RMM, CW were Principal Investigators and directed respectively the UK, IRISH and DUTCH sample collections and with RJP designed overall strategy and obtained funding for the study. DAVH directed the entire study, performed statistical analysis and generated the figures. DAVH and CW wrote the paper. RMcG, FT and WMMcL performed additional statistical analysis.

Published in final edited form as:
Abstract

Our celiac disease genome-wide association study identified IL2/IL21 region risk variants. We genotyped 1,020 of the most associated non-HLA markers in a further 1,643 cases and 3,406 controls. Joint analysis including the genome-wide association study data (767 cases, 1422 controls) identified seven new risk regions (P < 5 × 10^{-7}). Six regions harbor genes controlling immune responses, including: CCR3, IL12A, IL18RAP, RGS1, SH2B3 (nsSNP rs3184504), TAGAP. Whole blood IL18RAP mRNA expression correlated with IL18RAP genotype. Type 1 diabetes and celiac disease share HLA-DQ, IL2/IL21, CCR3 and SH2B3 risk regions. Extensive genome-wide association study follow-up has identified additional celiac disease risk variants in relevant biological pathways.

RESULTS

We initially selected 1,164 non-HLA SNPs from the UKGWAS for follow up, comprising 1,088 single SNPs with association results of P < 0.00275 and 76 additional non-synonymous SNPs (nsSNP) with association results between P ≥ 0.00275 and P < 0.01. After exclusion of SNPs unlikely to be successful based on Illumina GoldenGate assay design criteria, and inspection of actual genotype clusters, we analysed 1,025 SNPs in follow-up collections. These collections comprised celiac cases and controls of Northern European origin and of similar phenotypes to the UKGWAS samples. We did not genotype additional markers from candidate regions, as our primary aim was to identify new risk regions rather than precisely localise the actual causal genetic variants. This design represents an unbiased search of the genome for susceptibility variants. The markers included 8 SNPs from the IL2-IL21 region that were reported to be associated to celiac disease in our previous study, and we additionally selected 5 SNPs to tag celiac disease associated HLA-DQ2/8 haplotypes. We excluded samples failing quality control criteria (see Methods), and analysed 719 cases and 1,561 population controls from the UK (UK2 collection), 416 cases.
and 957 bloodbank controls from Ireland (IRISH), 508 cases and 888 bloodbank controls from the Netherlands (DUTCH).

Observed association statistics for the UKGWAS SNPs in the follow-up collections markedly deviate from expected findings (Fig. 1a). We report 21 non-HLA SNPs from 8 distinct chromosomal regions meeting a genome wide significance threshold in all 7,238 samples of \( P_{\text{overall}} < 5 \times 10^{-7} \) (Table 1). Results from the WTCCC7 and other recent GWA studies have shown that the majority of markers at a \( P < 5 \times 10^{-7} \) significance level will be true findings, although independent replication by other investigators is necessary for definitive validation. Only one of these eight regions, the \( IL2-IL21 \) region\(^2\), has been previously reported in celiac disease. Breslow-Day tests were non-significant for each of the eight regions implying consistent effect sizes and direction across the four collections, and accuracy of the reported Cochran-Mantel-Haenszel test odds ratios. The observation of generally weaker association evidence in the IRISH dataset (Supplementary Data 1) is therefore likely to be a reflection of the smaller sample-size of this collection, rather than ethnic heterogeneity.

We did not observe any evidence for gene-gene interactions (departure from log-additive effects) between these regions, nor between the \( HLA \) and these regions. Association statistics, linkage disequilibrium plots, and Ensembl genes for each of the eight regions are shown in Table 1 and Fig. 2, and more detailed statistics for all markers in each collection are shown in Supplementary Data 1. Celiac disease HLA-DQ associations reflect the ability of antigen presenting cells to present toxic cereal epitopes to T cells. Remarkably, seven of the eight identified non-HLA regions also contain biologically plausible candidate genes involved in the immune response.

We inspected the data for possible bias due to population differences between cases and controls, genotyping artefact, missing genotype data or other factors\(^8\). The overall genotype call rate across all samples and SNPs was high at 99.94% (details for each marker in Supplementary Data 1). Inspection of cluster plots for the most associated SNPs from each of the eight non-HLA celiac associated regions (Supplementary Fig. 1), consistent findings from two assay chemistries (Infinium and GoldenGate) in the UK populations, and multiple associated SNPs for some regions suggested that results were not due to laboratory generated false positive findings. We assessed the median distribution of test statistics using genomic control\(^9\) (\( \lambda_{\text{GC}} = 1.0 \) indicates a null distribution with no inflation of test statistics).

In the full UKGWAS genome wide association scan dataset (using 767 non-first degree related cases and 1422 controls genotyped for 307,411 non-HLA SNPs) there was minimal evidence for inflation of test statistics (\( \lambda_{\text{GC}} = 1.03 \)). In the UK2 IRISH DUTCH follow-up samples, genotyped for 1020 non-HLA SNPs there was also little evidence for inflation (\( \lambda_{\text{GC}} = 1.02 \)). When we also excluded 28 SNPs mapping to the eight non-HLA celiac associated regions there was no evidence for test statistic bias (\( \lambda_{\text{GC}} = 1.00 \)) in the follow-up samples, and the residual Q-Q plot showed little deviation from expected results (Fig. 1b). Furthermore when we used an alternative principal components approach, each of the eight non-HLA regions again met the genome wide significance threshold (\( P_{\text{overall}} < 5 \times 10^{-7} \)) when corrected by EIGENSTRAT analysis (Supplementary Data 1).

We correlated \( cis \) gene expression in whole blood RNA samples (n=109) with SNPs from the eight non-HLA celiac associated regions, to test for possible functional effects of these markers. Eight putative candidate genes from the regions were expressed above background levels in these samples. We analysed 14 pairs of SNPs and expressed gene probes (Supplementary Data 2). We also assessed tissue and cell type expression based on published literature and the GNF SymAtlas database\(^10\) for putative candidate genes from each region. We analysed gene expression in small intestinal tissue (where celiac disease...
manifests) from healthy controls, and treated and untreated celiac disease individuals. We summarise each region below.

### 4q27

A previously identified SNP rs68228442, located ∼24 kb 3’ of *IL21*, showed the strongest association with celiac disease in the current study (Poverall=2.82 × 10⁻¹³). Independent replication, with the same allele and direction, of our previous report is provided by the UK2 collection ([Supplementary Data 1](#)), rs6822844 P UK2=0.0017. This SNP is among a cluster of 8 associated SNPs that are in a block of strong linkage disequilibrium containing four genes (*KIAA1109-ADAD1-IL2-IL21*). Both *IL2* and *IL21* are strong candidate genes because of their role in T cell activation. A SNP allele from this region was also reported in a recent type 1 diabetes GWAS (conferring susceptibility), and Graves’ disease (conferring reduced risk). A further Dutch study suggested association of rs6822844 with type 1 diabetes and rheumatoid arthritis, in the same direction as the celiac disease data. These findings suggest the 4q27 region might represent a more general autoimmune locus, although whether effects are due to one or multiple causal variants and the exact nature of these effects is currently unclear.

### 1q31

The most significant SNP outside the HLA and IL2-IL21 regions is rs2816316 (P overall = 2.58 × 10⁻¹¹) located within a ∼70kb LD block containing the gene *RGS1* (regulator of G-protein signaling 1). rs2816316 maps 8kb distal to the 5’ end of *RGS1*. *RGS* family genes attenuate the signalling activity of G-proteins by acting as GTPase activating proteins. *RGS1* acts to regulate chemokine receptor signaling and is known to be involved in B-cell activation and proliferation. *RGS1*⁻/⁻ mice have enhanced B cell movement into and out of lymph nodes, and heightened dendritic cell migratory response to chemokines. We found *RGS1* to be expressed in human small intestinal biopsies ([Supplementary Fig. 2](#)). Interestingly, we previously observed *RGS1* to be strongly expressed in murine intestinal intra-epithelial lymphocytes, and now confirm that T cell *RGS1* expression appears to be specific to the intestinal intra-epithelial lymphocyte compartment and is not found in conventional splenic or thymic αβ T cells ([Supplementary Fig. 3](#)). Intestinal intra-epithelial lymphocytes play a key role in epithelial cell death and the development of villous atrophy in celiac disease.

### 2q11 - 2q12

Two associated SNPs, rs917997 (P overall=8.49 × 10⁻¹⁰) and rs13015714 (both in LD, r²=0.95 in UKGWAS controls) map to a ∼400kb linkage disequilibrium block. Interestingly, in the WTCCC GWAS, Crohn’s disease shows modest association (proxy SNP for rs917997, P~10⁻⁴, [Supplementary Data 3](#)). This LD block contains four genes, two of which are receptors for the IL-18 protein (*IL18RAP* and *IL18R1*). The IL-18 pathway is highly relevant as mature IL-18 induces T cell interferon-γ synthesis, a key cytokine involved in the mucosal inflammation of celiac disease. IL-18 binds to targeted cells through a receptor comprising an α chain (IL18Ra, IL18R1) and a β chain (IL18Rβ, AcPL, IL18RAP). Mature IL-18 is expressed in the intestinal mucosa of active, treated and latent celiac patients but not in healthy controls. *IL18RAP* is strongly expressed in unstimulated T cells and NK cells (GNF SymAtlas10), and is expressed in small intestinal biopsies ([Supplementary Fig. 2](#)). We observed a large cis effect (ANOVA $P=3.2 × 10^{-5}$) of rs917997 genotypes on the level of *IL18RAP* mRNA expression in whole blood from treated celiac patients (Fig. 3), accounting for 16.1% of the population variance in *IL18RAP* expression. We observed a significant allele dosage effect on expression ([post-hoc](#)) regression testing for linear trend $P<0.0001$). Individuals
homozygous for the minor rs917997 A allele (which is more common in celiac than control subjects) expressed the lowest levels of IL18RAP mRNA, heterozygotes intermediate and G allele homozygotes the highest levels.

We sequenced the coding regions of IL18RAP in 23 celiac disease patients and 8 control individuals, and found 19 variants, 17 of which were already in dbSNP. No variants (dbSNP or from resequencing) map to the region of the Illumina IL18RAP expression probe. None of the variants is predicted to have functional consequences. One new variant (c. 1210+17A>G) was observed in a single control and the other new variant (c. 1384+70_1384+71insT) was found in two celiac individuals.

3p21
SNP rs6441961 (P overall = 3.14 × 10^{-7}) maps within a large cluster of chemokine receptor genes on 3p21, including CCR1, CCR2, CCRL2, CCR3, CCR5 and CCXCR1. rs6441961 lies 44kb 3′ of the nearest gene (CCR3). LD block definition is hampered by poor HapMap coverage of this region due to structural variation22. Chemokines and their receptors are critical for the recruitment of effector immune cells to the site of inflammation. In the WTCCC GWAS, type 1 diabetes shows modest association in the same direction with the same allele of SNP rs6441961 (P∼10^{-5}, Supplementary Data 3) suggesting a possible common mechanism between both immune-mediated diseases.

3q25 - 3q26
Two SNPs (rs17810546, rs9811792) in a ~70kb linkage disequilibrium block show strong association to celiac disease (rs17810546 P overall = 1.07 × 10^{-9}). This region is immediately 5′ of IL12A (interleukin-12 A). Interestingly, the two associated SNPs (r^2=0.19 in UKGWAS controls) may represent independent association signals. SNP rs17810546 also shows modest correlation with SNPs in both IL12A and SCHIP1 (schwannomin interacting protein 1). IL12A encodes the IL12p35 subunit that together with IL12p40 form IL12p70, i.e. the heterodimeric IL-12 cytokine that has a broad range of biological activities on T and natural killer cells. IL-12 induces interferon-γ secreting Th1 cells, one of the immunological hallmarks of celiac disease.

3q28
Multiple correlated SNPs (rs1464510 P overall = 5.33 × 10^{-9}) within a ~70 kb linkage disequilibrium block show association with celiac disease. This block is either 5′ of the RefSeq gene LPP, or intronic for other possible isoforms of LPP. The LPP gene shows very high expression in the small intestine (Supplementary Fig. 2), and may play a structural role at sites of cell adhesion in maintaining cell shape and motility. Relatively little is known about LPP, and how genetic variation in this region might predispose to celiac disease is unclear.

6q25
SNP rs1738074 (P overall = 6.71 × 10^{-8}) on chromosome 6q25.3 maps to a ~200 kb linkage disequilibrium block containing TAGAP (T-cell activation GTPase activating protein), itself within a larger region of weaker LD containing RSPH3 (radial spokehead-like 3). TAGAP is expressed in activated T cells23, has three isoforms, and is a Rho GTPase-activating protein important for modulating cytoskeletal changes although little is known about its role in immune function.
12q24

Association signals arise from two correlated SNPs rs653178 ($P_{\text{overall}} = 8 \times 10^{-8}$) and rs3184504 ($r^2=0.99$ in UKGWAS controls). These markers map in the vicinity of \(SH2B3\) (also known as \(LNK\)) and \(ATXN2\), modest LD is seen over a broader region of \(~1\text{Mb}\) containing multiple other genes. Strong association with type 1 diabetes is reported in this region, with rs3184504 (same allele and direction as celiac disease) entirely accounting for the association signal\(^{11}\). \(SH2B3\) (also known as \(LNK\), lymphocyte adaptor protein) is strongly expressed in monocytes and dendritic cells, as well as to a lesser extent in resting B, T cells and NK cells (GNF SymAtlas\(^{10}\)). We found \(SH2B3\) to be strongly expressed in the small intestine, higher expression in inflamed celiac biopsies may reflect leukocyte recruitment and activation (Supplementary Fig. 2). The rs3184504 marker is a non-synonymous SNP in exon 3 of \(SH2B3\) leading to a R262W amino acid change in the pleckstrin homology domain. This domain may be important in plasma membrane targeting. \(SH2B3\) regulates T cell receptor, growth factor, and cytokine receptor-mediated signalling implicated in leukocyte and myeloid cell homeostasis\(^{24,25}\). \(SH2B3^{-/-}\) mice have increased responses to multiple cytokines\(^{26}\).

**DISCUSSION**

We previously performed the first genome wide association study in celiac disease and reported risk variants in the \(IL2-IL21\) region contributing to disease susceptibility\(^{2}\). Other than \(IL2-IL21\), only the long recognised \(HLA-DQ\) region has been convincingly associated with celiac disease. Our current study assessed 1,020 top celiac GWAS SNPs (including all markers with \(P < 0.00275\) in the original study) in further follow-up samples, and revealed seven new regions contributing significantly towards celiac disease risk. Our strategy of testing a further set of nsSNPs selected on the basis of a lower significance threshold did not yield any additional findings. The most significant SNPs for each of the seven new regions are common (minor allele frequency \(>0.1\)) albeit the observed effect sizes in the follow-up collections are modest (with increased or decreased odds ratios between 1.19 and 1.34). These are very similar magnitudes of effect sizes to other markers recently identified for other common diseases\(^{7,11,27}\). We note that the markers selected for the Illumina Hap300 are mainly tag SNPs, possibly mapping to relatively large regions of linkage disequilibrium, and are unlikely to be the causal variants. Since none of the variants are expected to be causal (although rs3184504 might be an exception), these odds ratios may be underestimates of the true effect sizes.

These seven new regions, together with \(IL2-IL21\) explain \(~3-4\%\) of the heritability of celiac disease. These estimates are based on the current markers and may be higher once the causal variants from each region are known. Consideration of the current findings, together with the strong risk attributed by the \(HLA-DQ2\) and \(-DQ8\) locus (in this study estimated to be \(~35\%\)), suggests many further susceptibility regions remain undetected. Additional regions might be identified by performing genome wide association on a larger sample size and in other collections (and possibly at higher marker density). Since not all disease susceptibility variants may be common SNPs, future attempts should also focus on rare variants and copy number polymorphisms - the pathways identified in this study may provide a good starting point.

Celiac disease is considered a Th1 mediated disorder with an important role for HLA-DQ2 and HLA-DQ8 in presenting wheat (and rye, barley) peptides to CD4\(^+\) T cells in the lamina propria. Tissue transglutaminase deamidation of immunodominant wheat peptides leads to higher DQ2 affinity, and tissue transglutaminase is the major target of the autoantibody response in celiac disease. Although we have not proved causality for any SNP variant, nor precisely identified the biologically important gene from each of the new regions, it is
remarkable that candidate genes mapping to seven of the eight non-HLA regions have clear functions in the immune response. Specific pathways include chemokine and cytokines, and B and T cell activation. The regions that we have identified cover important critical steps required for progression towards full-blown celiac disease. These new susceptibility genes may also partly explain why only a minority of HLA-DQ2/DQ8 positive individuals develop celiac disease, although (unknown) environmental factors are likely to play a role.

We provide a plausible biological mechanism for variants from two of the celiac disease associated regions: for SH2B3 a non-synonymous SNP in a functional domain, and for IL18RAP an effect of SNP genotype on mRNA expression. The IL18RAP findings, where the minor allele is more common in celiac subjects than controls and associated with lower IL18RAP mRNA expression, might suggest that leucocytes from celiac individuals signal less in response to IL-18, and generate less interferon-γ. This is somewhat counter-intuitive with the observed intestinal inflammation and high mucosal interferon-γ levels in celiac disease. However similarly counter-intuitive pictures are observed in (i) Crohn’s disease where disease associated NOD2 mutations are loss-of-function (i.e. less NF-κB activation in response to NOD2 agonist), yet there is excessive intestinal inflammation20, and (ii) autoimmune diseases where a gain of function PTPN22 mutation leads to reduced B and T cell receptor mediated signalling, yet there is an observed excessive systemic inflammatory response in a variety of autoimmune conditions21. Our understanding of the complex biology of celiac disease, including that of regulatory T cells, remains incomplete.

It has been suggested that common pathogenic effector pathways exist between HLA-associated autoimmune disorders, such as celiac disease, type I diabetes and rheumatoid arthritis. In these conditions the key elements regulating the attack or destruction of tissues or organ systems by an inappropriate or excessive immune response include genetic factors, innate and adaptive immune regulation, and triggering environmental factors such as infections or dietary components. T cell activation is a generic process playing a role in all autoimmune disorders. Published data suggests IL2-IL21 risk variants (not necessarily the same alleles) may predispose to celiac disease, type I diabetes, Graves disease and rheumatoid arthritis11,28. Consideration of the current celiac regions suggests and the WTCCC seven disease GWAS dataset7 suggests possible common mechanisms between celiac disease and type 1 diabetes at the SH2B3 region and the 3p21 CCR gene region, and celiac disease and Crohn’s disease at the IL18RAP region. It will be interesting to investigate the current celiac regions in other autoimmune disorders, and to dissect separate disease-specific genes from more generic genes. This information may be useful to unravel the pathological mechanisms of autoimmune diseases and may provide targets for intervention.

Our results have considerably enhanced our understanding of celiac disease pathogenesis. The identification of risk factors for celiac disease that point towards involvement of genes implicated in various aspects of immunity further highlights the importance of such responses for celiac disease development. Fine mapping and deep resequencing of the regions, and functional characterization of the potential candidate genes, will be required to firmly establish the true causal genes and their role in the disease process.

**METHODS**

**Subjects**

Detailed characteristics of UKGWAS, IRISH and DUTCH samples are provided in Supplementary Table 1, and our previously published study2. All subjects were of white Northern European origin. UK2 subjects were recruited similarly to the previously described UKGWAS subjects2 - cases were recruited from hospital clinics and controls from the...
British 1958 Birth Cohort, except for n=374 cases and n=176 unrelated controls recruited
direct through Coeliac UK advertisement. For DNA procedures see Supplementary
Methods. We note that \textit{HLA-DQ2.5cis} genotype frequencies (Supplementary Table 1,
inferrred from rs2187668) were similarly high across the celiac collections (carrier rate
range 87.4% to 89.1%), and similarly low across the control populations (carrier rate range
25.5% to 33.8%), suggesting broadly comparable phenotypic ascertainment across the four
collections. Informed consent was obtained from all subjects. Ethical approval was from
Oxfordshire REC B or East London and the City REC 1 (UKGWAS, UK2), the Medical
Ethical Committee of the University Medical Center Utrecht (DUTCH), and the Institutional
Ethics Committee of St James’s Hospital (IRISH).

\textbf{Marker selection and genotyping}

We selected from our published celiac disease genome wide association study\textsuperscript{2} of 310,605
post-quality control markers (Hardy-Weinberg equilibrium \(P>0.0001\) in controls) single
non-\textit{HLA} SNPs with two-tailed allele count \(\chi^2\) test \(P<0.00275\), or if nsSNPs \(P<0.01\).

Genotyping data and clustering of SNP genotypes was managed in BeadStudio. Samples
with \(<95\%\) call rate over 1025 SNPs, SNPs with \(<95\%\) call rate over remaining samples,
SNPs with poor amplification or poor genotype cloud clustering were excluded.

\textbf{Quality control steps}

Using the 1025 SNP dataset, pairwise comparisons of identity-by-descent were made for all
samples (UKGWAS, UK2, IRISH, DUTCH) using PLINK\textsuperscript{29}. We detected a higher
proportion of 1st degree relatives in the all sample dataset (98 pairs) than in the initial UK
celiac GWAS (11 pairs), and therefore in the current analyses excluded the lowest call rate
sample from each pair of 1° relatives from the entire study dataset. Minor but insignificant
changes are therefore present in the UKGWAS dataset results compared to our previous
publication\textsuperscript{2}.

Potential ethnic outlier samples were excluded (n=3, outliers had previously been excluded
from the UKGWAS dataset) using the nearest neighbour allele sharing method in PLINK
(samples with Z scores \(<-3\) with \(>1\) of \(5\) nearest neighbours excluded). We applied a filter
for SNP selection for follow up GoldenGate genotyping based on Hardy Weinberg (HWE)
equilibrium \(P>0.0001\) in Infinium genotyped controls from the UKGWAS. HWE \(P\) values in
each of the follow-up UK2, IRISH and Dutch collections are shown in \textit{Supplementary
Data 1}. Because of the prior filter step, we did not specifically exclude any follow-up SNPs
based on HWE analysis. All of the top association findings were in HWE in controls.

\textbf{Genotype statistical analysis}

Cochran-Mantel-Haenszel allele count chi-squared association tests were performed using
PLINK\textsuperscript{29} with 4 clusters: UKGWAS (Infinium assay), UK2 (GoldenGate assay), IRISH and
DUTCH collections. All \(P\) values are two-tailed. We observed minimal evidence for bias in
the Cochran-Mantel-Haenszel test statistics due to population stratification or other factors
(see \(\lambda_{GC}\) values \(\leq 0.03\) in \textit{RESULTS}), and present uncorrected statistics. To test for
differences in association across the datasets as manifested by heterogeneity of odds ratios at
disease-associated regions, we applied the Breslow-Day test to SNPs in Table 1. We also
evaluated these SNPs, and \textit{HLA-DQ2.5cis} (inferred by rs2187668 genotype) for possible
epistatic (gene-gene) interaction in predisposing to celiac disease. Breslow-Day and epistatic
interaction tests were carried out as implemented in PLINK. We also tested for epistatic
interactions using the model of Howard et al\textsuperscript{30}. For the PLINK analysis of interaction,
genotypes at the SNPs are evaluated as allele “doses” (e.g. 0, 1, or 2) whereas for the model
of Howard et al heterozygotes and homozygotes for the higher-risk allele are pooled
together and compared against the lower-risk homozygote genotype. The PLINK interaction
analysis and our analysis of the model of Howard et al (implemented in R statistical software, www.r-project.org) both tested for statistically significant departure from the model of log-additive odds ratios as evaluated by logistic regression.

We also investigated possible influence of stratification by an alternative principle components approach in which each dataset was separately evaluated by EIGENSTRAT software31. After eliminating SNPs in the 9 regions showing association with celiac disease, EIGENSTRAT was applied to the remaining SNPs (drawn from 310,605 SNPs for UKGWAS, or from 1025 SNPs for UK2, IRISH, DUTCH) to determine eigenvectors that correct for possible population stratification in cases and controls. For each case-control set, EIGENSTRAT calculates the Armitage Trend test chi-square statistic for each genotyped SNP and then applies a SNP-specific correction based on the eigenvectors that individually adjusts the value of Armitage chi-square. Following EIGENSTRAT correction, we also corrected with the residual genomic control factor for each collection and then combined the post-correction chi-squares across the 4 studies according to the meta-analysis method of Stouffer32. This enabled us to achieve a corrected, study-wide Z-score and P value for each SNP that can be compared to the Cochran-Mantel-Haenszel P value and shows that all 9 regions remain significant at $P < 5 \times 10^{-7}$ after correction by this alternative approach (Supplementary Data 1). The corrected P values for SNPs not in the 9 coeliac-associated regions also generate a QQ-plot that shows no evidence of overdispersion and is almost identical to the QQ-plot (Fig. 1b) of the Cochran-Mantel-Haenszel P values.

Calculation of familial clustering for each region is described in Supplementary Methods.

**P value threshold for genomewide statistical significance**

For statistical significance of disease association results in the combined genome scan (UKGWAS) and follow-up datasets (UK2, DUTCH, IRISH), we applied the same genomewide significance threshold ($P < 5 \times 10^{-7}$) adopted by the WTCCC7. It should be noted that this threshold is close to the conservative Bonferroni corrected significance threshold as suggested by Skol et al.3 for a two-stage study such as ours in which genome-scan and subsequent results from stage-2 followup samples are combined. Since 307,411 non-HLA SNPs were tested in our genome scan, the genomewide significance threshold based on Skol et al. criteria would be approximately $0.05/307,411 \approx 2 \times 10^{-7}$.

**Whole blood expression analysis**

PAXgene blood RNA sample was collected from unselected Illumina Hap300 genotyped UKGWAS celiac cases. All were selected on the basis of gluten free diet treatment for >6 months, to avoid possible bias in gene expression levels due to active inflammation. Notable features of this experimental design were: use of primary cells, analysis of expression from unstimulated leucocytes, and the possibility that the samples would be enriched for disease causal genetic variants compared to healthy controls. RNA was extracted using the PAXgene protocol, hybridised to Illumina HumanRef-8v2 expression arrays, scanned and processed in BeadStudio. Data were cubic spline normalised.

We obtained whole blood PAXgene expression data from 114 unique celiac individuals who had been genotyped in the UKGWAS. To assess the quality of the RNA for each sample, we assessed 6,298 genes with two different probes present on the Illumina HumanRef8v2 chip. We correlated intensities for all probes for each sample (Median Pearson correlation all samples = 0.22) and removed n=2 outlier samples that had low correlation (Pearson correlation < 0.10). Analysis of transcripts mapping to the non-pseudoautosomal region of chromosome Y, allowed for determining sex, resulting in the removal of n=2 samples with mismatched gender. We removed n=1 further sample where genotypes did not correlate as
expected with cis expression level across the genome. We analysed data from 109 individuals after applying these quality control criteria. ANOVA tests were performed to investigate possible cis effects of the strongest SNP associations on candidate gene expression from disease associated regions.

**Intestinal biopsy expression analysis**

Duodenal tissue biopsies were collected in RNAlater (Ambion), RNA was extracted using TRIzol (Invitrogen) and glass beads, hybridised to HumanRef-8v2 arrays and analysed as for the PAXgene samples. Expression profiling of T cell subsets is described in Supplementary Methods.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

We acknowledge funding from Coeliac UK (to DAvH); the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW); the European Union (STREP 036383); the Netherlands Organization for Scientific Research (VICI grant 918.66.620 to CW); the Science Foundation Ireland; the Irish Health Research Board; Hitachi Europe Ltd.; and the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to DAvH; New Blood Fellowship to RMcM; support for the work of RMcG and PD). We thank the Barts and The London Genome Centre for genotyping support; J Swift, P. Kumar, D.P. Jewell, S.P.L. Travis, K. Moriarty for collection of UKGWAS and additional samples. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.. We thank A. Monsuur for patient recruitment, G. Meijer and J. Meijer for histology review, K. Duran for DNA extraction, H. van Someren and F. Mulder for clinical database management, M. Plateel and the Genotyping facilities at UMC and UMC Utrecht for technical assistance (the Netherlands). We thank M. Abuazakouk, R. McLoughlin, K. Brophy, C. Feighery, J. McPartlin for sample collection. Irish Control DNA was supplied by Irish Blood Transfusion Service/Trinity College Dublin Biobank. We thank the Wellcome Trust Centre for Human Genetics, University of Oxford for provision of computing facilities and G. McVean for recombination rate data. We thank all celiac and control individuals for participating in this study.

**References**

Figure 1. Quantile-quantile plot for association results in follow-up samples

(a) Q-Q plot of association results (Cochran-Mantel-Haenszel test) for 1020 non-HLA SNPs in UK2, IRISH and DUTCH follow-up samples. Data points in red indicate SNPs shown in Table 1 with $P_{\text{overall}} < 5 \times 10^{-7}$ in all samples including UKGWAS. Straight line indicates expected results under null hypothesis.

(b) Q-Q plot of residual association results (Cochran-Mantel-Haenszel test) for 992 non-HLA SNPs, excluding 28 SNPs mapping to the eight celiac associated regions described in Table 1, in UK2, IRISH and DUTCH follow-up samples.
Figure 2. Linkage disequilibrium structure and association results for eight non-HLA celiac disease associated regions
Chromosomal positions based on NCBI build 36 coordinates, showing Ensembl (release 48) genes. Allele count $P$ values are shown for SNPs analysed in UKGWAS samples (black diamonds), and Cochran-Mantel-Haenszel association test $P$ values (red circles) for SNPs analysed in all UKGWAS, UK2, IRISH and DUTCH samples.
Figure 3. Correlation of rs917997 genotype with whole blood cis IL18RAP mRNA expression
Expression levels (bars show group means) were determined in samples from gluten-free diet treated celiac individuals.
<table>
<thead>
<tr>
<th>Cytogenetic location</th>
<th>Position (bp)</th>
<th>SNP</th>
<th>Rank UKGWAS</th>
<th>P UKGWAS</th>
<th>P follow-up</th>
<th>Odds ratio&lt;sup&gt;b&lt;/sup&gt; follow-up [95% CI]</th>
<th>P overall</th>
<th>Putative candidate gene(s)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=2189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=5049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=7238</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q31</td>
<td>190803436</td>
<td>rs2816316</td>
<td>382</td>
<td>0.0012</td>
<td>5.11 × 10&lt;sup&gt;-9&lt;/sup&gt;</td>
<td>0.71 [0.63 - 0.80]</td>
<td>2.58 × 10&lt;sup&gt;-11&lt;/sup&gt;</td>
<td>RGS1</td>
</tr>
<tr>
<td>2q11-2q12</td>
<td>102338297</td>
<td>rs13015714</td>
<td>149</td>
<td>0.00041</td>
<td>2.66 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>1.27 [1.15 - 1.40]</td>
<td>4.37 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>IL1RL1, IL18R1, IL18RAP, SLC9A4</td>
</tr>
<tr>
<td>2q13</td>
<td>102437000</td>
<td>rs917997</td>
<td>41</td>
<td>9.06 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.27 [1.15 - 1.40]</td>
<td>8.49 × 10&lt;sup&gt;-10&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3p21</td>
<td>46327388</td>
<td>rs6441961</td>
<td>1004</td>
<td>0.0028</td>
<td>3.21 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.21 [1.10 - 1.32]</td>
<td>3.14 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>CCR1, CCR3</td>
</tr>
<tr>
<td>3q25 - 3q26</td>
<td>161147744</td>
<td>rs17810546</td>
<td>111</td>
<td>0.00034</td>
<td>7.77 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>1.34 [1.19 - 1.51]</td>
<td>1.07 × 10&lt;sup&gt;-9&lt;/sup&gt;</td>
<td>IL12A, SCHIP1</td>
</tr>
<tr>
<td>3q28</td>
<td>189570322</td>
<td>rs9851967</td>
<td>453</td>
<td>0.00091</td>
<td>7.13 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>0.82 [0.76 - 0.90]</td>
<td>2.45 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>LPP</td>
</tr>
<tr>
<td>3q28</td>
<td>189571948</td>
<td>rs13076312</td>
<td>186</td>
<td>0.00034</td>
<td>1.21 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.21 [1.11 - 1.31]</td>
<td>1.78 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>3q28</td>
<td>189595248</td>
<td>rs1464510</td>
<td>52</td>
<td>7.65 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.21 [1.11 - 1.31]</td>
<td>5.33 × 10&lt;sup&gt;-10&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3q28</td>
<td>189599810</td>
<td>rs1559810</td>
<td>59</td>
<td>0.00010</td>
<td>3.28 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.19 [1.10 - 1.30]</td>
<td>1.99 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>123292459</td>
<td>rs11938795</td>
<td>88</td>
<td>0.00026</td>
<td>1.07 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>0.80 [0.73 - 0.89]</td>
<td>1.30 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td>IL2, IL21</td>
</tr>
<tr>
<td>4q27</td>
<td>123334952</td>
<td>rs13151961</td>
<td>10</td>
<td>5.48 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>4.32 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td>0.72 [0.65 - 0.81]</td>
<td>1.53 × 10&lt;sup&gt;-12&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>12347763</td>
<td>rs13119723</td>
<td>1</td>
<td>2.51 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>1.23 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>0.73 [0.65 - 0.82]</td>
<td>5.94 × 10&lt;sup&gt;-13&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>12347563</td>
<td>rs11734090</td>
<td>106</td>
<td>0.00034</td>
<td>2.95 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>0.81 [0.74 - 0.90]</td>
<td>4.76 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>123560609</td>
<td>rs7684187</td>
<td>245</td>
<td>0.00091</td>
<td>3.11 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>0.82 [0.75 - 0.90]</td>
<td>1.13 × 10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>123727951</td>
<td>rs12642902</td>
<td>25</td>
<td>4.59 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>3.75 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>0.82 [0.75 - 0.89]</td>
<td>2.12 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>4q27</td>
<td>123728871</td>
<td>rs6822844</td>
<td>7</td>
<td>4.80 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>9.84 × 10&lt;sup&gt;-9&lt;/sup&gt;</td>
<td>0.71 [0.63 - 0.80]</td>
<td>2.82 × 10&lt;sup&gt;-11&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>6q25</td>
<td>159385965</td>
<td>rs1738074</td>
<td>503</td>
<td>0.0016</td>
<td>1.19 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.21 [1.11 - 1.31]</td>
<td>6.71 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td>TAGAP</td>
</tr>
<tr>
<td>12q24</td>
<td>110368991</td>
<td>rs3184504</td>
<td>318</td>
<td>0.0011</td>
<td>3.22 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.19 [1.10 - 1.30]</td>
<td>1.33 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>SH2B3(LNK), ATXN2</td>
</tr>
<tr>
<td>12q24</td>
<td>110492139</td>
<td>rs653178</td>
<td>207</td>
<td>0.00065</td>
<td>2.98 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>1.19 [1.10 - 1.30]</td>
<td>8.00 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

Cochran-Mantel-Haenszel association analysis of 1020 SNPs (excluding HLA region). SNPs with P overall < 5 × 10<sup>-7</sup> (the WTCCC threshold for reporting possible associations<sup>7</sup>) are shown. Association statistics are reported for twenty one SNPs from eight regions, including the previously reported IL2/IL21 region. Chromosomal positions based on NCBI build-36 coordinates. Data from the UKGWAS was re-analysed with removal of a small number of first degree relatives compared to previous publication<sup>2</sup>.
Odds ratios for the UKGWAS and overall study are reported in Supplementary Data 1.

Named genes either map to the same strong LD block as associated SNPs, or the SNPs physically map within the gene - causality is not proven.

crs3184504 is a non-synonymous SNP in SH2B3 (R262W).