

Published in final edited form as:

Nat Genet. 2008 April ; 40(4): 395–402. doi:10.1038/ng.102.

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 Mearin^{12,13}, 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 McManus^{3,*}, Cisca Wijmenga^{2,4,*}, and David A van Heel^{1,*}

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 (E-mail: 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.

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 ($P < 5 \times 10^{-7}$). 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.

Celiac disease is common in Western populations with a prevalence of approximately 1%¹. *HLA-DQ2* or *-DQ8* restricted T cell responses to dietary proteins in wheat, rye, and barley lead to small intestinal inflammation and intestinal villous atrophy with consequent clinical features. Several recent immunological advances have identified dominant epitopes, the role of tissue transglutaminase, and the crystal structures of *HLA-DQ* molecules binding wheat peptides¹. However the primary factors (both environmental and genetic) predisposing to disease in the ~30% of the population carrying susceptible *HLA-DQ* types are mostly unknown.

To identify additional celiac disease susceptibility genes, we recently tested 310,605 SNPs in a genome wide association study of 778 celiac cases and 1,422 population controls from the United Kingdom (UKGWAS), using the Illumina HumanHap300 BeadChip². The only SNP outside the *HLA* region demonstrating genome-wide significance was rs13119723 on 4q27, located in a ~500 kb block of linkage disequilibrium (LD) containing the *IL2* and *IL21* genes². Independent replication of SNPs from the *IL2-IL21* region was established in both Dutch and Irish collections of celiac patients and controls. We estimate, using the current markers, that the *IL2-IL21* region explains less than 1% of the increased familial risk to celiac disease, suggesting that there are additional unidentified susceptibility genes. Since we observed a greater number of significantly associated SNPs in the UKGWAS than would be expected by chance, we proceeded to study >1,000 of the most significant UKGWAS association results in a further 1,643 celiac cases and 3,406 controls from three independent European celiac disease collections. This two-stage strategy, involving a joint analysis of all data, substantially reduces the genotyping requirements versus performing whole genome genotyping on all samples and has been shown to maintain sufficient statistical power³.

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 $P \geq 0.00275$ 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^{4,5}. 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 study², and we additionally selected 5 SNPs to tag celiac disease associated *HLA-DQ2/8* haplotypes⁶. 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 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 $P < 5 \times 10^{-7}$ "genome-wide" 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* region², 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 factors⁸. 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 control⁹ ($\lambda_{GC} = 1.0$ 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 ($\lambda_{GC} = 1.03$). In the UK2 IRISH DUTCH follow-up samples, genotyped for 1020 non-*HLA* SNPs there was also little evidence for inflation ($\lambda_{GC} = 1.02$). When we also excluded 28 SNPs mapping to the eight non-*HLA* celiac associated regions there was no evidence for test statistic bias ($\lambda_{GC} = 1.00$) 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 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 database¹⁰ 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 (P overall = 2.82×10^{-13}). Independent replication, with the same allele and direction, of our previous report² 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^{2,11,12}. A SNP allele from this region was also reported in a recent type 1 diabetes GWAS (conferring susceptibility), and Graves' disease (conferring reduced risk)¹¹. A further Dutch study suggested association of rs6822844 with type 1 diabetes and rheumatoid arthritis, in the same direction as the celiac disease data¹³. 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 nodes¹⁴, and heightened dendritic cell migratory response to chemokines¹⁵. 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 lymphocytes¹⁶, 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 $\alpha\beta$ 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 disease¹⁷.

2q11 - 2q12

Two associated SNPs, rs917997 (P overall = 8.49×10^{-10}) and rs13015714 (both in LD, $r^2=0.95$ in UKGWAS controls) map to a ~400kb linkage disequilibrium block. Interestingly, in the WTCCC GWAS⁷, Crohn's disease shows modest association (proxy SNP for rs917997, $P \sim 10^{-4}$, **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 (*IL18R α* , *IL18R1*) and a β chain (*IL18R β* , *AcPL*, *IL18RAP*). Mature IL-18 is expressed in the intestinal mucosa of active, treated and latent celiac patients but not in healthy controls¹⁹.

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 $P = 3.2 \times 10^{-5}$) 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 GWAS⁷, type 1 diabetes shows modest association in the same direction with the same allele of SNP rs6441961 ($P \sim 10^{-5}$, **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 overall = 8×10^{-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 ~1Mb 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 signal¹¹. *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 SymAtlas¹⁰). 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 cytokines²⁶.

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 susceptibility². 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 SNP associations (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- κ B 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 1 diabetes, Graves disease and rheumatoid arthritis^{11,28}. 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 *HLA-DQ2.5cis* genotype frequencies (Supplementary Table 1, inferred from rs2187668) were similarly high across the celiac collections (carriage rate range 87.4% to 89.1%), and similarly low across the control populations (carriage 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).

Marker selection and genotyping

We selected from our published celiac disease genome wide association study² of 310,605 post-quality control markers (Hardy-Weinberg equilibrium $P > 0.0001$ in controls) single non-*HLA* SNPs with two-tailed allele count χ^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.

Quality control steps

Using the 1025 SNP dataset, pairwise comparisons of identity-by-descent were made for all samples (UKGWAS, UK2, IRISH, DUTCH) using PLINK²⁹. 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 1st relatives from the entire study dataset. Minor but insignificant changes are therefore present in the UKGWAS dataset results compared to our previous publication².

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 **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.

Genotype statistical analysis

Cochran-Mantel-Haenszel allele count chi-squared association tests were performed using PLINK²⁹ 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 λ_{GC} values ≤ 1.03 in 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 *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 al³⁰. 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 genome-wide 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 RNeasy (Qiagen), 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 UMCG and UMC Utrecht for technical assistance (the Netherlands). We thank M. Abuzakouk, 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

1. van Heel DA, West J. Recent advances in coeliac disease. *Gut* 2006;55:1037–46. [PubMed: 16766754]
2. van Heel DA, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007;39:827–9. [PubMed: 17558408]
3. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–13. [PubMed: 16415888]
4. Chanock SJ, et al. Replicating genotype-phenotype associations. *Nature* 2007;447:655–60. [PubMed: 17554299]
5. Clarke GM, Carter KW, Palmer LJ, Morris AP, Cardon LR. Fine Mapping versus Replication in Whole-Genome Association Studies. *Am J Hum Genet* 2007;81:995–1005. [PubMed: 17924341]
6. de Bakker PI, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006;38:1166–72. [PubMed: 16998491]
7. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–78. [PubMed: 17554300]
8. Plenge RM, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* 2007;39:1477–82. [PubMed: 17982456]
9. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004. [PubMed: 11315092]
10. Su AI, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004;101:6062–7. [PubMed: 15075390]
11. Todd JA, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007;39:857–64. [PubMed: 17554260]
12. Yamanouchi J, et al. Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. *Nat Genet*. 2007
13. Zhernakova A, et al. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *Am J Hum Genet* 2007;81:1284–8. [PubMed: 17999365]
14. Han SB, et al. Rgs1 and Gna12 regulate the entrance of B lymphocytes into lymph nodes and B cell motility within lymph node follicles. *Immunity* 2005;22:343–54. [PubMed: 15780991]
15. Shi GX, Harrison K, Han SB, Moratz C, Kehrl JH. Toll-like receptor signaling alters the expression of regulator of G protein signaling proteins in dendritic cells: implications for G protein-coupled receptor signaling. *J Immunol* 2004;172:5175–84. [PubMed: 15100254]

16. Pennington DJ, et al. The inter-relatedness and interdependence of mouse T cell receptor gammadelta + and alphabeta+ cells. *Nat Immunol* 2003;4:991–8. [PubMed: 14502287]
17. Hue S, et al. A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. *Immunity* 2004;21:367–77. [PubMed: 15357948]
18. Wu C, et al. IL-18 receptor beta-induced changes in the presentation of IL-18 binding sites affect ligand binding and signal transduction. *J Immunol* 2003;170:5571–7. [PubMed: 12759435]
19. Salvati VM, et al. Interleukin 18 and associated markers of T helper cell type 1 activity in coeliac disease. *Gut* 2002;50:186–90. [PubMed: 11788557]
20. Abraham C, Cho JH. Functional Consequences of NOD2 (CARD15) Mutations. *Inflamm Bowel Dis* 2006;12:641–50. [PubMed: 16804402]
21. Rieck M, et al. Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. *J Immunol* 2007;179:4704–10. [PubMed: 17878369]
22. Iafrate AJ, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–51. [PubMed: 15286789]
23. Mao M, et al. T lymphocyte activation gene identification by coregulated expression on DNA microarrays. *Genomics* 2004;83:989–99. [PubMed: 15177553]
24. Fitau J, Bouliday G, Coulon F, Quillard T, Charreau B. The adaptor molecule Lnk negatively regulates tumor necrosis factor-alpha-dependent VCAM-1 expression in endothelial cells through inhibition of the ERK1 and -2 pathways. *J Biol Chem* 2006;281:20148–59. [PubMed: 16644735]
25. Li Y, He X, Schembri-King J, Jakes S, Hayashi J. Cloning and characterization of human Lnk, an adaptor protein with pleckstrin homology and Src homology 2 domains that can inhibit T cell activation. *J Immunol* 2000;164:5199–206. [PubMed: 10799879]
26. Velazquez L, et al. Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice. *J Exp Med* 2002;195:1599–611. [PubMed: 12070287]
27. Zeggini E, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;316:1336–41. [PubMed: 17463249]
28. Zhernakova A, et al. Novel association in chromosome 4q27 region to rheumatoid arthritis and confirmation to type 1 diabetes points to a general risk locus for autoimmune diseases. *Am J Hum Genet*. 2007online preprint 1 Aug 2007
29. Purcell S, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 2007;81:559–75. [PubMed: 17701901]
30. Howard TD, et al. Gene-gene interaction in asthma: IL4RA and IL13 in a Dutch population with asthma. *Am J Hum Genet* 2002;70:230–6. [PubMed: 11709756]
31. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–9. [PubMed: 16862161]
32. Rosenthal, R. *Meta-Analytic Procedures for Social Research*. Sage Publications Inc.; 1991.

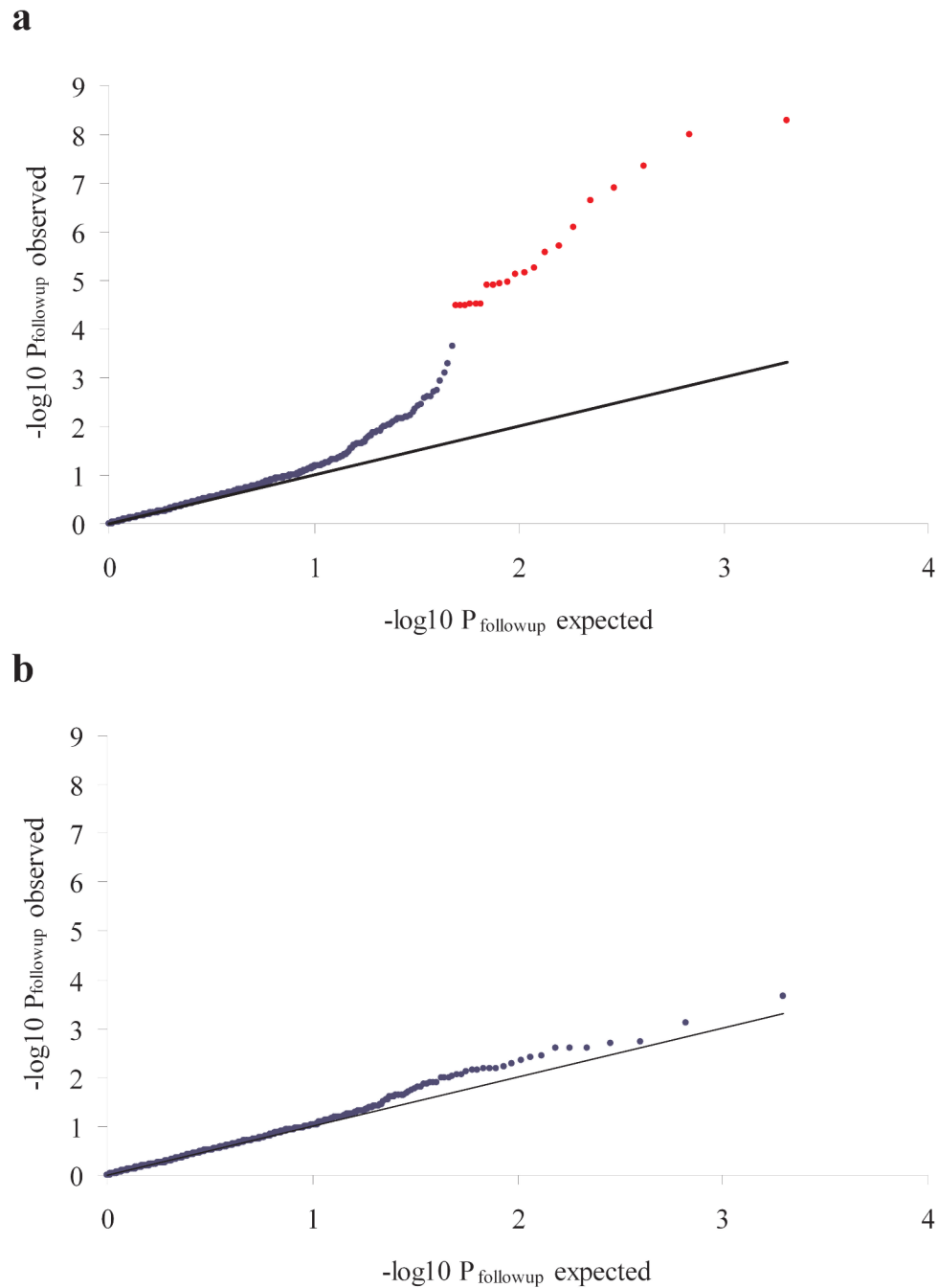


Figure 1. Quantile-quantile plot for association results in follow-up samples

(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.

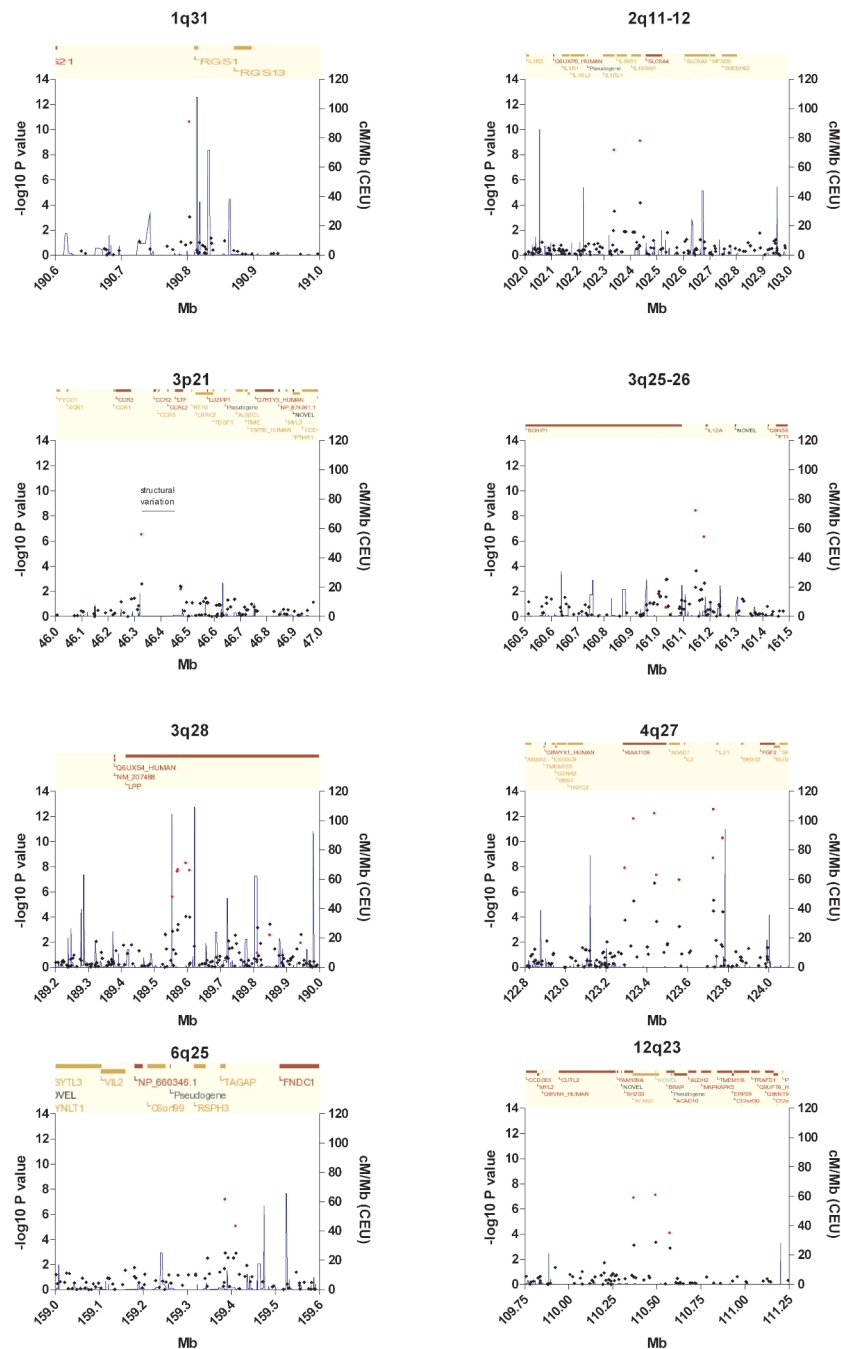


Figure 2. Linkage disequilibrium structure and association results for eight non-*HLA* celiac disease associated regions

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.

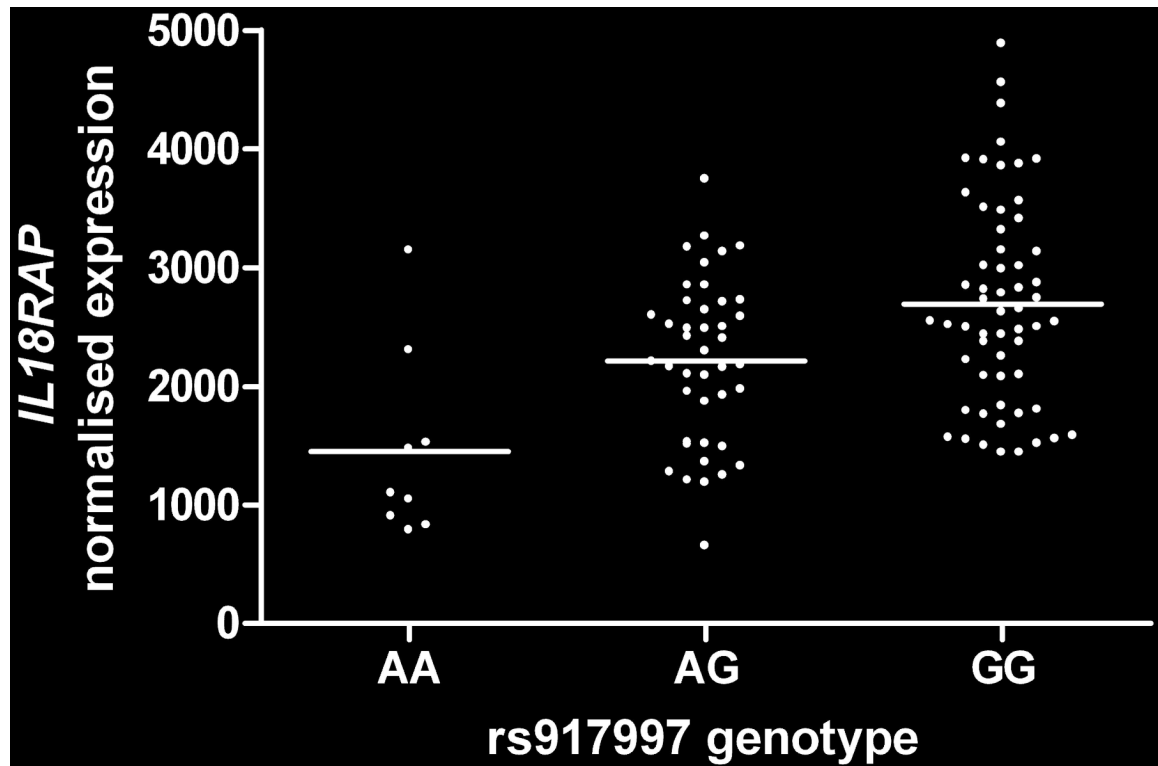


Figure 3. Correlation of rs917997 genotype with whole blood *cis* *IL18RAP* mRNA expression
 Expression levels (bars show group means) were determined in samples from gluten-free diet treated celiac individuals.

Table 1
Genomic regions with the strongest association signals for celiac disease

Cytogenetic location	Position (bp)	SNP	Rank UKGWAS	P UKGWAS	P follow-up	Odds ratio ^a follow-up [95% CI]	P overall	Putative candidate genes ^b
Sample size				n=2189	n=5049		n=7238	
Assay chemistry				Infinitum	GoldenGate			
1q31	190803436	rs2816316	382	0.0012	5.11×10^{-9}	0.71 [0.63 - 0.80]	2.58×10^{-11}	RGS1
2q11-2q12	102338297	rs13015714	149	0.00041	2.66×10^{-6}	1.27 [1.15 - 1.40]	4.37×10^{-9}	IL1RL1, IL18R1, IL18RAP, SLC9A4
	102437000	rs917997	41	9.06×10^{-5}	1.97×10^{-6}	1.27 [1.15 - 1.40]	8.49×10^{-10}	
3p21	46327388	rs6441961	1004	0.0028	3.21×10^{-5}	1.21 [1.10 - 1.32]	3.14×10^{-7}	CCR1, CCR3
3q25 - 3q26	161147744	rs17810546	111	0.00034	7.77×10^{-7}	1.34 [1.19 - 1.51]	1.07×10^{-9}	IL12A, SCHIP1
	161179692	rs9811792	898	0.0028	5.42×10^{-6}	1.21 [1.12 - 1.32]	5.24×10^{-8}	
3q28	189570322	rs9851967	453	0.00091	7.13×10^{-6}	0.82 [0.76 - 0.90]	2.45×10^{-8}	LPP
	189571948	rs13076312	186	0.00034	1.21×10^{-5}	1.21 [1.11 - 1.31]	1.78×10^{-8}	
	189595248	rs1464510	52	7.65×10^{-5}	1.21×10^{-5}	1.21 [1.11 - 1.31]	5.33×10^{-9}	
	189607048	rs1559810	59	0.00010	3.28×10^{-5}	1.19 [1.10 - 1.30]	1.99×10^{-8}	
4q27	123292459	rs11938795	88	0.00026	1.07×10^{-5}	0.80 [0.73 - 0.89]	1.30×10^{-8}	IL2, IL21
	123334952	rs13151961	10	5.48×10^{-6}	4.32×10^{-8}	0.72 [0.65 - 0.81]	1.53×10^{-12}	
	123437763	rs13119723	1	2.51×10^{-7}	1.23×10^{-7}	0.73 [0.65 - 0.82]	5.94×10^{-13}	
	123447563	rs11734090	106	0.00034	2.95×10^{-5}	0.81 [0.74 - 0.90]	4.76×10^{-8}	
	123560609	rs7684187	245	0.00091	3.11×10^{-5}	0.82 [0.75 - 0.90]	1.13×10^{-7}	
	123727951	rs12642902	25	4.59×10^{-5}	7.35×10^{-6}	0.82 [0.75 - 0.89]	2.12×10^{-9}	
	123728871	rs6822844	7	4.80×10^{-6}	9.84×10^{-9}	0.71 [0.63 - 0.80]	2.82×10^{-13}	
	123774157	rs6840978	31	4.79×10^{-5}	2.32×10^{-7}	0.75 [0.68 - 0.84]	5.53×10^{-11}	
6q25	159385965	rs1738074	503	0.0016	1.19×10^{-5}	1.21 [1.11 - 1.31]	6.71×10^{-8}	TAGAP
12q24	110368991	rs3184504 ^c	318	0.0011	3.22×10^{-5}	1.19 [1.10 - 1.30]	1.33×10^{-7}	SH2B3(LNK), ATXN2
	110492139	rs653178	207	0.00065	2.98×10^{-5}	1.19 [1.10 - 1.30]	8.00×10^{-8}	

Cochran-Mantel-Haenszel association analysis of 1020 SNPs (excluding *HLA* region). SNPs with P overall $< 5 \times 10^{-7}$ (the WTCCC threshold for reporting possible associations⁷) 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 publication².

^aOdds ratios for the UKGWAS, and overall study are reported in Supplementary Data 1.

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

^C rs3184504 is a non-synonymous SNP in SH2B3 (R262W).