Multiple common variants for celiac disease influencing immune gene expression.

https://escholarship.org/uc/item/54d9w7gw

Nature genetics, 42(4)

1546-1718

Dubois, Patrick CA
Trynka, Gosia
Franke, Lude
et al.

2010-04-01

10.1038/ng.543

Peer reviewed
Multiple common variants for celiac disease influencing immune gene expression

Patrick CA Dubois1,37, Gosia Trynka2,37, Lude Franke1,2, Karen A Hunt1, Jihane Romanos2, Alessandra Curtotti3, Alexandra Zhernakova4, Graham AR Heap3, Róza Ádány5, Arpo Aromaa6, Maria Teresa Bardella7,8, Leonard H van den Berg9, Nicholas A Bockett1, Emilio G. de la Concha10, Bárbara Dema10, Rudolf SN Fehrmann2, Miguel Fernández-Arquero10, Szilvia Fiatal5,11, Elvira Grandone12, Peter M Green13, Harry JM Groen14, Rhian Gwilliam15, Roderick HJ Houwen16, Sarah E Hunt15, Katri Kaukinen17, Dermot Kelleher18, Ilma Korponay-Szabó19,20, Kalle Kurppa17, Padraic MacMathuna21, Markku Mäki17, Maria Cristina Mazzilli22, Owen T McCann15, M Luisa Mearin23, Charles A Mein3, Muddassar M Mirza13, Vanisha Mistry1, Barbara Mora22, Katherine I Morley15, Chris J Mulder24, Joseph A Murray25, Concepción Núñez10, Elvira Oosterom2, Roel A Ophoff26,27, Isabel Polanco28, Leena Peltonen15,29, Mathieu Platteeuw2, Anna Rybak30, Veikko Salomaa6, Joachim J Schweizer23, Maria Pia Sperandeo31, Greetje J Tack24, Graham Turner18, Jan H Veldink9, Wieke HM Verbeek24, Rinse K Weersma2,32, Victoriem M Wolters16, Elena Urcelay10, Bozena Cukrowska33, Luigi Greco31, Susan L. Neuhausen34, Ross McManus18, Donatella Barisani35, Panos Deloukas15, Jeffrey C Barrett15, Paivi Saavalainen36, Cisca Wijmenga2, and David A van Heel1

1Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK 2Genetics Department, University Medical Center and Groningen University, 9700 RB Groningen, The Netherlands 3The Genome Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK 4Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands 5Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary 6National Institute for Health and Welfare, Helsinki 00271, Finland 7Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy 8Department of Medical Sciences, University of Milan, Italy 9Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Heidelberglaan 100, Utrecht, The Netherlands 10Clinical Immunology Department, Hospital Clínico San Carlos, Madrid, Spain 11Public Health Research Group of Hungarian Academy of Sciences, Medical & Health Science Center, University of Debrecen, Debrecen, Hungary 12Unita’ di Aterosclerosi e Trombosi, I.R.C.C.S Casa Sollievo della Sofferenza, S. Giovanni Rotondo, Foggia, Italy 13NIHR GSTT/KCL Comprehensive Biomedical Research Centre King’s College London School of Medicine 7th Floor Tower Wing Guy’s Hospital London SE19RT United Kingdom 14Dept. of

Correspondence to DAvH (d.vanheel@qmul.ac.uk) or regarding expression analyses to LF (lude@ludesign.nl).

These authors contributed equally.

AUTHOR CONTRIBUTION STATEMENT
DAvH and CW designed, co-ordinated and led the study. Experiments were performed in the labs of CW, DAvH, CAM, PD and PMG. Major contributions were: 1) DNA sample preparation - PCAD, GT, KA, JR, AZ, PS 2) genotyping - PCAD, GT, KA, AC, JR, RG; 3) expression data generation - HIMG, LhtvDB, RAO, Rkw, LF; 4) case/control association analyses - PCAD, GT, LF, JCB, DAvH; 5) expression analyses - LF, GAH, RSNF; 6) manuscript preparation - PCAD, GT, LF, RSNF, GAH, JCB, CW, DAvH. Other authors contributed variously to sample collection and all other aspects of the study. All authors reviewed the final manuscript.

The authors declare no competing financial interests.

DATABASE ACCESSION NUMBERS
Expression data is available in GEO (http://www.ncbi.nlm.nih.gov/geo/) as GSE11501 and GSE20142.
Abstract

We performed a second-generation genome wide association study of 4,533 celiac disease cases and 10,750 controls. We genotyped 113 selected SNPs with $P_{\text{GWAS}} < 10^{-4}$, and 18 SNPs from 14 known loci, in a further 4,918 cases and 5,684 controls. Variants from 13 new regions reached genome wide significance ($P_{\text{combined}} < 5 \times 10^{-8}$), most contain immune function genes (BACH2, CCR4, CD80, CIITA/SOCS1/CLEC16A, ICOSLG, ZMIZ1) with ETS1, RUNX3, THEMIS and TNFRSF14 playing key roles in thymic T cell selection. A further 13 regions had suggestive association evidence. In an expression quantitative trait meta-analysis of 1,469 whole blood samples, 20 of 38 (52.6%) tested loci had celiac risk variants correlated ($P<0.0028$, FDR 5%) with cis gene expression.

Celiac disease is a common heritable chronic inflammatory condition of the small intestine induced by dietary wheat, rye and barley, as well as other unidentified environmental factors, in susceptible individuals. Specific HLA-DQA1 and HLA-DQB1 risk alleles are necessary, but not sufficient, for disease development. The well defined role of HLA-DQ heterodimers, encoded by these alleles, is to present cereal peptides to CD4+ T cells, activating an inflammatory immune response in the intestine. A single genome wide association study (GWAS) has been performed in celiac disease, which identified the IL2/IL21 risk locus. Subsequent studies probing the GWAS information in greater depth have identified a further 12 risk regions. Most of these regions contain a candidate gene functional in the immune system, although only in the case of HLA-DQA1 and HLA-DQB1 have the causal variants been established. Many of the known celiac loci overlap with other immune-related diseases. In order to identify additional risk variants, particularly of smaller effect size, we performed a second-generation GWAS using over six times as many
samples as the previous GWAS and a denser genome-wide SNP set. We followed up promising findings in a large collection of independent samples.

The GWAS included five European celiac disease case and control sample collections including the previously reported celiac disease dataset1. We performed stringent data quality control (Online Methods), including calling genotypes using a custom algorithm on both large sample sets, and where possible cases and controls together (Online Methods). We tested 292,387 non-HLA SNPs from the Illumina Hap300 marker set for association in 4,533 celiac disease cases and 10,750 controls of European descent (Table 1). A further 231,362 additional non-HLA markers from the Illumina Hap550 marker set were tested for association in a subset of 3,796 celiac disease cases and 8,154 controls. All markers were from autosomes or the X chromosome. Genotype call rates were >99.9% in both datasets. The overdispersion factor of association test statistics, $\lambda_{GC}=1.12$, was similar to that observed in other GWAS of this sample size7,8. Findings were not substantially altered by imputation of missing genotypes for 737 celiac disease cases genotyped on the Hap300 BeadChip and corresponding controls (Table 1, collection 1). Here we present results for directly genotyped samples, as around half the additional Hap550 markers cannot be accurately imputed from Hap300 data9 (including the new ETS1 locus finding in this study). Results for the top 1000 markers are available in Supplementary Data 1, but because of concerns regarding identity detection of individuals10, results for all markers are available only on request to the corresponding author.

For follow-up, we first inspected genotype clouds for the 417 non-HLA SNPs meeting $P_{GWAS}<10^{-4}$, being aware that top GWAS association signals may be enriched for genotyping artefact, and excluded 22 SNPs from further analysis using a low threshold for possible bias. We selected SNPs from 113 loci for replication. Markers that passed design and genotyping quality control included: a) 18 SNPs from all 14 previously identified celiac disease risk loci (including a tag SNP for the major celiac disease associated HLA-DQ2.5cis haplotype1); b) 13 SNPs from all 7 novel regions with $P_{GWAS}<5\times10^{-7}$; c) 86 SNPs from 59 of 68 novel regions with $5\times10^{-7}>P_{GWAS}>5\times10^{-5}$ in stage 1; d) 14 SNPs from 13 of 30 novel regions with $5\times10^{-5}>P_{GWAS}>10^{-4}$ in stage 1 (for this last category, we mostly chose regions with immune system genes). Two SNPs were selected per region for: regions with stronger association; regions with possible multiple independent associations; and/or containing genes of obvious biological interest. We successfully genotyped 131 SNPs in 7 independent follow-up cohorts comprising 4,918 celiac disease cases and 5,684 controls of European descent. Genotype call rates were >99.9% in each collection. Primary association analyses of the combined GWAS and follow-up data were performed with a two-sided 2×2×12 Cochran-Mantel-Haenszel test.

**RESULTS**

**Celiac disease risk variants**

The HLA locus and all 13 other previously reported celiac disease risk loci showed evidence for association at a genome wide significance threshold ($P_{combined}<5\times10^{-8}$) (Table 2). We note that some loci were previously reported using less stringent criteria (e.g. the $P<5\times10^{-7}$ recommended by the 2007 WTCCC study11), but that in the current, much larger sample set, all known loci meet recently proposed $P<5\times10^{-8}$ thresholds12,13.

We identified 13 novel risk regions with genome-wide significant evidence ($P_{combined}<5\times10^{-8}$) of association, including regions containing the BACH2, CCR4, CD80, CIITA/ SOCS1/CLEC16A, ETS1, ICOSLG, RUNX3, THEMIS, TNFRSF14, and ZMIZ1 genes which are of obvious immunological function (Table 2). A further 13 regions met ‘suggestive’ criteria for association (either $10^{-6}>P_{combined}>5\times10^{-8}$ and/or $P_{GWAS}<10^{-4}$).
and \( P_{\text{followup}} < 0.01 \). These regions also contain multiple genes of obvious immunological function, including \( CD247, FASLG/TNFSF18/TNFSF4, IRF4, TLR7/TLR8, TNFRSF9 \) and \( YDIC \). Six of the 39 non-HLA regions show evidence for the presence of multiple independently associated variants in a conditional logistic regression analysis (Supplementary Table 2).

We tested the 40 SNPs with the strongest association (Table 2) from each of the known genome-wide significant, new genome-wide significant, and new suggestive loci for evidence of heterogeneity across the 12 collections studied. Only the \( HLA \) region was significant (Breslow-Day test \( P < 0.05 / 40 \) tests, \( rs2187668 \) \( P = 4.8 \times 10^{-8} \)) which is consistent with the well described North-South gradient in HLA allele frequency in European populations, and more specifically for \( HLA-DQ \) in celiac disease.\(^{14}\)

We observed no evidence for interaction between each of the 26 genome-wide significant non-\( HLA \) loci, which is consistent with what has been reported for complex diseases so far. However, we did observe weak evidence for lower effect sizes at non-\( HLA \) loci in high risk \( HLA-DQ2.5\')s homozygotes, similar to what has been previously observed in type 1 diabetes.\(^{7}\)

To obtain more insight into the functional relatedness of the celiac loci, we applied GRAIL, a statistical tool that utilizes text mining of PubMed abstracts to annotate candidate genes from loci associated with common disease risk.\(^{15,16}\) In order to assess the sensitivity of this tool (using known loci as a positive control), we first performed a ‘leave-one-out’ analysis of the 27 genome-wide significant celiac disease loci (including \( HLA-DQ \)). GRAIL scores of \( P_{\text{ext}} < 0.01 \) were obtained for 12 of the 27 loci (44% sensitivity, Table 2). Factors that limit the sensitivity of GRAIL include biological pathways being both known (a 2006 dataset is used to avoid GWAS era studies), and published in the literature. We then applied GRAIL analysis, using the 27 known regions as a seed, to all 49 regions (49 SNPs) with \( 10^{-3} > P_{\text{combined}} > 5 \times 10^{-8} \) and obtained GRAIL \( P_{\text{ext}} < 0.01 \) for 9 regions (18.4%). As a control, only 5.5% (279 of 5033) of randomly selected Hap550 SNPs reached this threshold. In addition to the five ‘suggestive’ loci shown in Table 2, GRAIL annotated four further interesting gene regions of lower significance in the combined association results: \( rs944141/PDCD1LG2 \) \( (P_{\text{combined}} = 4.4 \times 10^{-6}) \), \( rs976881/TNFRSF8 \) \( (P_{\text{combined}} = 2.1 \times 10^{-4}) \), \( rs4682103/CD200/BTLA \) \( (P_{\text{combined}} = 6.8 \times 10^{-6}) \) and \( rs4919611/NFKB2 \) \( (P_{\text{combined}} = 6.1 \times 10^{-5}) \). These appeared to further enrich for genes of immunological interest which are not GRAIL annotated in the \( 10^{-3} > P_{\text{combined}} > 5 \times 10^{-8} \) significance window, including \( rs3828599/TNIP1 \) \( (P_{\text{combined}} = 1.55 \times 10^{-4}) \), \( rs8027604/PTPN9 \) \( (P_{\text{combined}} = 1.4 \times 10^{-6}) \), \( rs944141/CD274 \) \( (P_{\text{combined}} = 4.4 \times 10^{-6}) \). Some of these findings, for which neither genome-wide significant nor suggestive association is achieved, are likely to comprise part of a longer tail of disease predisposing common variants, of weaker effect sizes. Definitive assessment of these biologically plausible regions would require genotyping and association studies using much larger sample collections than the present study.

We previously showed considerable overlap between celiac disease and type 1 diabetes risk loci,\(^{17}\) as well as celiac disease and rheumatoid arthritis,\(^{18}\) and more generally, there is now substantial evidence for shared risk loci between the common chronic immune mediated diseases.\(^{6}\) To update these observations, we searched ‘A Catalog of Published Genome Wide Association Studies’ (18 Nov 2009)\(^{19}\) and the HuGE database.\(^{20}\) We found some evidence (requiring a published association report of \( P < 1 \times 10^{-5} \)) of shared loci with at least one other inflammatory or immune mediated disease for 18 of the current 27 genome-wide significant celiac risk regions. We defined shared regions as the broad LD block, however different SNPs are often reported in different diseases, and at only three of the 18 shared regions are associations across all diseases with the same SNP or a proxy SNP in \( r^2 > 0.8 \)
HapMap CEU. Currently 9 regions appear celiac disease specific and may reflect distinctive disease biology including the regions containing rs296547 and rs9792269, and the regions around CCR4, CD80, ITGA4, LPP, PLEK, RUNX3 and THEMIS. In fact, locus sharing between diseases is probably greater, due to both stochastic variation in results from sample size limitations, and regions with a genuinely stronger effect size in one disease and weaker effect size in another.

Genetic variation in ETS1 has recently been reported to be associated with systemic lupus erythematosus (SLE) in the Chinese population, although is not associated with SLE in European populations21. The most strongly associated celiac (European population) SNP rs11221332 and the most strongly associated SLE (Chinese population) SNP rs6590330 map 70kb apart. Inspection of the HapMap phase II data shows broadly similar linkage disequilibrium patterns between Chinese (CHB) and European (CEU) populations in this region, with the two associated SNPs in separate non-adjacent linkage disequilibrium blocks. Thus distinct common variants within the same gene are predisposing to different autoimmune diseases across different ethnic groups.

**Function of celiac risk variants**

Celiac risk variants in the HLA alter protein structure and function4. However we identified only four non-synonymous SNPs with evidence for celiac disease association \( (P_{GWAS} < 10^{-4}) \) from the other 26 genome-wide significant associated regions (rs3748816/MMEL1, rs3816281/PLEK, rs196432/RUNX3, rs3184504/SHP2B3). Although comprehensive regional resequencing is required to test the possibility that coding variants contribute to the observed association signals, more subtle effects of genetic variation on gene expression are the more likely major functional mechanism for complex disease genes. With this in mind, we performed a meta-analysis of new and published genome-wide expression quantitative trait loci (eQTL) datasets comprising 1,469 human whole blood (PAXgene) samples reflecting primary leucocyte gene expression. We applied a new method, *Transcriptional Components*, to remove a substantial proportion of inter-individual non-genetic expression variation and performed eQTL meta-analysis on the residual expression variation (Online Methods).

We assessed 38 of the 39 genome-wide significant and suggestive celiac disease associated non-HLA loci (Table 2) for cis expression - genotype correlations. We tested the SNP with the strongest association from each region. However for five regions the most associated SNP was not genotyped in the eQTL samples (Hap300 data), instead for four of these we tested a proxy SNP \( (r^2 > 0.5 \text{ in } \text{HapMap CEU}) \). In addition, for six loci showing evidence of multiple independent associations in conditional regression analyses, we tested a second SNP showing independent celiac disease association for eQTL analysis. In total we assessed 44 independent non-HLA SNP associations in peripheral whole blood samples genotyped on the Illumina Hap300 BeadChip and either Illumina Ref8 or HT12 expression arrays, correlating each SNP with data from gene-probes mapping within a 1Mb window.

We identified significant (Spearman \( P<0.0028 \), corresponding to 5% false discovery rate) eQTLs at 20 of 38 (52.6%) non-HLA celiac loci tested (Table 3, Supplementary Figures 2 & 3). Some loci had evidence of eQTLs with multiple probes, genes or SNPs (Table 3). We assessed whether the number of SNPs with cis-eQTL effects out of the 44 SNPs that we tested, was significantly higher than expected. We observed that eQTL SNPs on average have a substantially higher MAF than non-eQTL SNPs in the 294,767 SNPs tested. In order to correct for this we selected 44 random SNPs that had an equal MAF distribution, and determined for how many of these MAF-matched SNPs eQTLs were observed. We observed a significantly higher number of eQTL SNPs \( \left( P=9.3 \times 10^{-5}, 10^6 \text{ permutations} \right) \) amongst the celiac associated SNPs than expected by chance (22 observed eQTL SNPs, vs. 7.8 expected
eQTL SNPs). Therefore the celiac disease associated regions are greatly enriched for eQTLs. These data suggest some risk variants may influence celiac disease susceptibility through a mechanism of altered gene expression. Candidate genes with a significant eQTL, where the peak eQTL signal and peak case/control association signal are similar (Supplementary Figure 3), include MMEL1, NSF, PARK7, PLEK, TAGAP, RRP1, UBE2L3 and ZMIZ1.

We also assessed co-expression of genes mapping within 500kb of SNPs showing strongest case/control association from the 40 genome-wide significant and suggestive celiac disease loci in an analysis of the 33,109 human Affymetrix Gene Expression Omnibus dataset. This analysis loses power to detect tissue specific correlations by use of numerous tissue types, but greatly gains power by the large sample size. We detected several distinct co-expression clusters (Pearson correlation coefficient between genes >0.5), including four clusters of immune-related genes which contain at least one gene from 37 of the 40 genome-wide significant and suggestive loci (Fig. 1). These data further demonstrate that genes from celiac disease risk loci map to multiple distinct immunological pathways involved in disease pathogenesis.

**DISCUSSION**

We previously reported that most celiac genetic risk variants mapped near genes that are functional in the immune system22, and this remains true for the 13 new genome-wide significant, and 13 new suggestive, risk variants from the current study. We can now refine these observations and highlight specific immunological pathways relevant to celiac disease pathogenesis:

1) **T cell development in the thymus**

The rs802734 LD block contains the recently identified gene THEMIS ‘THymus-Expressed Molecule Involved in Selection’. THEMIS plays a key regulatory role in both positive and negative T-cell selection during late thymocyte development23. Furthermore, the rs10903122 LD block contains RUNX3, a master regulator of CD8+ T lymphocyte development in the thymus24,25. TNFRSF14 (LIGHTR, rs3748816 LD block) has widespread peripheral leucocyte functions as well as a critical role in promoting thymocyte apoptosis26. The ETS1 transcription factor (rs11221332 LD block) is also active in peripheral leucocytes, however it is also a key player in thymic CD8+ lineage differentiation, acting in part by promoting RUNX3 expression27.

The importance of the thymus in autoimmune disease pathogenesis has been previously emphasised by the established role of thymectomy in the treatment of myasthenia gravis. In type 1 diabetes, it was shown that disease associated genetic variation in the insulin gene INS causes altered thymic insulin expression and subsequent T cell tolerance for insulin as a self-protein28. However, the importance of thymic T cell regulation has not been previously recognised in the aetiology of celiac disease. It is conceivable that the associated variants may alter biological processes prior to thymic MHC-ligand interactions. Alternatively it is now clear that exogenous antigen presentation and selection occurs in the thymus via migratory dendritic cells - this has been demonstrated for skin, and has been hypothesised for food antigens29,30. These findings suggest research into immuno-/pharmaco-logical modifiers of T cell tolerance more generally in autoimmune diseases.

2) **Innate immune detection of viral RNA**

Although the association signal at rs5979785 ($P_{combined}=6.36\times10^{-8}$) in the TLR7/TLR8 region is just outside our genome wide significance threshold, we observe a strong effect of
rs5979785 on TLR8 expression in whole blood. Both TLRs recognise viral RNA. Taken together with the recent observation of rare loss of function mutations in the enteroviral response gene IFIH1 protective against type 1 diabetes31, these findings suggest viral infection (and the nature of the host response to infection) as a putative environmental trigger common to these autoimmune diseases.

3) T and B cell co-stimulation (or co-inhibition)
This class of molecules controls the strength and nature of the response to T or B (immunoglobulin) cell receptor activation by antigens. We observe multiple regions with genes (CTLA4/ICOS/CD28, TNFRSF14, CD80, ICOSLG, TNFRSF9, TNFSF4) from this class of ligand-receptor pairs suggesting fine control of the adaptive immune response might be altered in at-risk individuals.

4) Cytokines, chemokines and their receptors
Our previous report discussed the function of the 2q11-12 interleukin receptor cluster (IL18RAP, etc), the 3p21 chemokine receptor cluster (CCR5, etc.) and the loci containing IL2/IL21 and IL12A22. We now report additional loci containing TNFSF18 and CCR4.

We estimate that the current celiac disease variants, including the major celiac disease associated HLA variant, HLA-DQ2.5cis, less common celiac disease associated haplotypes in the HLA (HLA-DQ8, HLA-DQ2.5trans; HLADQ2.2), and the additional 26 definitively implicated loci explain about 20% of total celiac disease variance, which would represent 40% of genetic variance, assuming a heritability of 0.5. A long tail of low effect size common variants, along with highly penetrant rare variants (both at the established loci and elsewhere in the genome), may substantially contribute to the remaining heritability.

We observed different haplotypes within the ETS1 region associated with coeliac disease in Europeans, and SLE in the Chinese population. We further note for some autoimmune diseases studied in European origin populations, that although the same LD block has been associated, the association is with a different haplotype. In some cases, the same variants are associated, but the direction of association is opposite (e.g. rs917997/IL18RAP in celiac disease versus type 1 diabetes). We believe further exploration of these signals may reveal critical differences in the nature of the immune system perturbation between these diseases.

Previous investigators have observed that only a small proportion of GWAS associations are coding variants, and have suggested that these may instead influence regulation of gene expression. Here, we show that over half the celiac disease associated variants are correlated with expression changes in nearby genes. This mechanism is likely to explain the function of some risk variants for other common, complex diseases. Further research, however, is needed to definitively determine at each locus both the celiac disease causal variants and their functional mechanisms.

Supplementary Material
Refer to Web version on PubMed Central for supplementary material.

Acknowledgments
We thank Celiac UK for assistance with direct recruitment of celiac disease individuals, and UK clinicians (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis, K. Moriarty) who recruited celiac disease blood samples described in our previous studies1,22. We thank the genotyping facility of the UMCG (J. Smolonska, P. van der Vlies) for generating part of the GWAS and replication data and the gene expression data; R. Booij and M. Weenstra are thanked for preparation of Italian samples. H. Ahola, A. Heimonen, L. Koskinen, E. Einarsdottir and K. Löytynoja are thanked for their work on
Appendix

METHODS

Subjects

Written informed consent was obtained from all subjects, with Ethics Committee / Institutional Review Board approval. All individuals are of European ancestry. Affected celiac individuals were diagnosed according to standard clinical, serological and histopathological criteria including small intestinal biopsy. DNA samples were from blood, lymphoblastoid cell lines or saliva. A more detailed description of subjects is provided in Supplementary Information.

GWAS genotyping

See Table 1. UK(1) case and control genotyping was previously described1,7. Illumina 670-Quad and 1.2M-Duo (custom chips designed for the WTCCC2 and comprising Hap550 / 1M and common CNV content) and 610-Quad genotyping was performed in London, Hinxton
and Groningen. Bead intensity data was normalized for each sample in BeadStudio, R and theta values exported and genotype calling performed using a custom algorithm1,35. A detailed description of genotype calling steps is provided in Supplementary Information.

Quality control steps were performed in the following order: Very low call rate samples and SNPs were first excluded. SNPs were excluded from all sample collections if any collection showed call rates were <95% or deviation from Hardy-Weinberg equilibrium (P<0.0001) in controls. Samples were excluded for call rate <98%, incompatible recorded gender and genotype inferred gender, ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap Phase II data), duplicates and first degree relatives. 22 of 417 SNPs showing apparent association (P_{GWAS}<10^{-4}) were excluded after visual inspection of R theta plots suggested possible bias.

The over-dispersion factor of association test statistics (genomic control inflation factor), \( \lambda_{GC} \), was calculated using observed versus expected values for all SNPs in PLINK.

**Follow-up genotyping**

See Table 1. Finnish controls (12) were genotyped on the 610-Quad BeadChip, other samples were genotyped using Illumina GoldenGate BeadXpress assays in London, Hinxton and Groningen. Genotyping calling was performed in BeadStudio for combined cases and controls in each separate collection, with the exception of the Finnish collection, and whole genome amplified samples (89 Irish cases and 106 Spanish controls). Quality control steps were performed as for the GWAS. 131 of 144 SNPs passed quality control and visual inspection of genotype clouds.

**SNP association analysis**

Analyses were performed using PLINK v1.0736, mostly using the Cochran-Mantel-Haenzel test. Logistic regression analyses were used to define the independence of association signals within the same linkage disequilibrium block, with group membership included as a factorized covariate.

Genotype imputation was performed for samples genotyped on the Hap300 using BEAGLE and CEU, TSI, MEX and GIH reference samples from HapMap3. Association analysis was performed using logistic regression on posterior genotype probabilities, with group membership included as a factorized covariate.

Structured association tests were performed using PLINK, as described using genetically matched cases and controls within collections identified by identity by state similarity across autosomal non-HLA SNPs as described34 (settings --ppc 0.001 --cc, clusters constrained by the 5 collections). Principal components analysis was performed using EIGENSTRAT and a set of 12,810 autosomal non-HLA SNPs chosen for low LD and ancestry information37,38, association tests were corrected for the top 10 principal components and combined using weighted Z scores.

The fraction of additive variance was calculated using a liability threshold model39 assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the combined replication panel. Genetic variance was calculated assuming 50% heritability.

**GRAIL analysis**

We performed GRAIL analysis (http://www.broadinstitute.org/mpg/grail/grail.php) using HG18 and Dec2006 PubMed datasets, default settings for SNP rs number submission, and
the 27 genome-wide significant celiac disease risk loci (most associated SNP) as seeds. As a query we used either associated SNPs, or 101 × 50 randomly chosen Hap550 SNP datasets (5050 SNPs, of which 5033 mapped to the GRAIL database).

Identification of Transcriptional Components

We noted that the power of eQTL studies in humans is limited by substantial observed inter-individual variation in expression measurements due to non-genetic factors, and therefore developed a method, Transcriptional Components, to remove a large component of this variation (manuscript in preparation). Expression data from 42,349 heterogeneous human samples hybridized to Affymetrix HG-U133A (GEO accession number: GPL96) or HG-U133 Plus 2.0 (GEO accession number: GPL570) Genechips were downloaded 40 (Fig. 1, step 1). Samples missing data for >150 probes were excluded, and only probes available on both platforms were analysed, resulting in expression data for 22,106 probes and 41,408 samples. We performed quantile normalization using the median rank distribution and log₂ transformed the data - ensuring an identical distribution of expression signals for every sample, discarding previous normalization and transformation steps.

Initial quality control (QC) was performed by applying principal component analysis (PCA) on the sample correlation matrix (pair-wise Pearson correlation coefficients between all samples). The first principal component (PC), explaining ~80-90% of the total variance, describes probe-specific variance. 6,375 samples with correlation R < 0.75 of the sample array with this PC were considered outliers of lesser quality and excluded from analysis. We excluded entire GEO datasets where >25% of the samples were outliers (probably expression ratios versus a reference, not absolute data). The final dataset comprised 33,109 samples (17,568 GPL96 and 15,541 GPL570 samples), and we repeated the normalization and transformation on the originally deposited expression values of these post-quality control samples.

We next applied PCA on the pairwise 22,106 × 22,106 probe Pearson correlation coefficient matrix assayed on the 33,109 sample dataset (our fast C++ tool, MATool, is available upon request), attempting to simplify the structure of the data. Here, PCA represents a transformation of a set of correlated probes into sets of uncorrelated linear additions of probe expression signals (eigenvectors) that we name Transcriptional Components (TCs). Each TC is a weighted sum of probe expression signals and eigenvector probe coefficients. These TC-scores can be calculated for each observed expression array sample (reflecting the TC activity per sample).

Subjects for expression - genotype correlation

We obtained peripheral blood DNA and RNA (PAXgene) from Dutch and UK individuals who were disease cases or controls for GWAS studies (Supplementary Table 1). All samples had been genotyped for a common SNP set on Illumina platforms. Analysis was confined to 294,767 SNPs that had a MAF ≥ 5%, call-rate ≥ 95% and exact HWE P > 0.001. RNA from the samples was either hybridized to Illumina HumanRef-8 v2 arrays (229 samples, Ref-8v2) or Illumina HumanHT-12 arrays (1,240 samples, HT-12), and raw probe intensity extracted using BeadStudio. The Ref-8v2 samples were jointly quantile normalized and log₂ transformed, and similarly for the HT-12 samples. Subsequent analyses were also conducted separately for these datasets, up to the eventual eQTL mapping, that uses a meta-analysis framework, combining eQTL results from both arrays. HT-12 and Ref-8v2 arrays are different, but share many probes with identical probe sequences. Illumina sometimes use different probe identifiers for the same probe sequences - in meta-analysis and Table 3, the label HT-12 was used if both HT-12 and Ref-8v2 had the same sequence.
Re-mapping of probes

If probes mapped incorrectly, or cross-hybridized to multiple genomic loci, it might be that an eQTL would be detected that would be deemed a trans-eQTLs. To prevent this, we used a mapping approach versus a known reference that we developed for high-throughput short sequence RNAseq data. We took the DNA sequence as synthesised for each cDNA probe, and aligned it versus a transcript masked gDNA genome combined with cDNA sequences. A more detailed description of probe re-mapping is provided in Supplementary Information. Probes that did not map, or mapped to multiple different locations were removed.

Affymetrix transcriptional components applied to Illumina expression data

TC-scores can be inferred in new (non-Affymetrix) datasets for every new individual sample. For the Illumina samples (used for the cis-eQTL mapping), only Illumina probes that could be mapped to any of our 22,106 Affymetrix probes were used (www.switchtoi.com/probemapping.ilmn). The TC-score of sample for the th TC is defined as,

\[ \text{TC-score}_{ij} = \sum_{t=1}^{n} a_{tj} \times v_{ij} \]

where \( a_{tj} \) is defined as the th Affymetrix probe coefficient for the th TC; \( a_{ti} \) is the Illumina expression measurement for the th mapped probe for sample . We inferred the Illumina TC-scores for the top 1,000 TCs.

Removal of transcriptional component effects from Illumina expression data

Because our Illumina eQTL dataset (n = 1,469) is much less heterogeneous than the Affymetrix dataset (n = 33,109), we expect that some TCs will hardly vary. We therefore performed a PCA on the covariance matrix of the top 1,000 inferred TC-scores for the Illumina dataset to effectively compress the TC data into a small set of ‘aggregate TCs’ (aTCs). As aTCs are orthogonal, we used linear regression to eliminate the effect of the top 50 aTCs. We correlated the TC-scores for each peripheral blood sample with probe expression levels. We then used the resulting residual gene expression data for subsequent cis-eQTL mapping.

cis-eQTL mapping

We used the residual gene expression data (Fig. 1) in a meta-analysis framework, as described. In brief, analyses were confined to those probe-SNP pairs for which the distance from probe transcript midpoint to SNP genomic location was less than 500 kb. To prevent spurious associations due to outliers, a non-parametric Spearman’s rank correlation analysis was performed. When a particular probe-SNP pair was present in both the HT12 and H8v2 datasets, an overall, joint p-value was calculated using a weighted Z-method (square root of the dataset sample number). To correct for multiple testing we controlled the false discovery rate (FDR). The distribution of observed \( P \) values was used to calculate the FDR, by permuting expression phenotypes relative to genotypes times within the HT12 and H8v2 dataset. Finally, we removed any probes from analysis which contain a known SNP (1000Genomes CEU SNP data, April 2009 release).

References

1. 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]
2. van Heel DA, West J. Recent advances in coeliac disease. Gut. 2006; 55:1037–46. [PubMed: 16766754]
3. Sollid LM, et al. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/ beta heterodimer. J Exp Med. 1989; 169:345–50. [PubMed: 2909659]
4. Kim CY, Quarsten H, Bergseng E, Khosla C, Sollid LM. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci U S A. 2004; 101:4175–9. [PubMed: 15020763]
5. Henderson KN, et al. A Structural and Immunological Basis for the Role of Human Leukocyte Antigen DQ8 in Celiac Disease. Immunity. 2007; 27:23–34. [PubMed: 17629515]
6. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet. 2009; 10:43–55. [PubMed: 19092835]
7. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009
8. Barrett JC, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn’s disease. Nat Genet. 2008; 40:955–62. [PubMed: 18587394]
9. Anderson CA, et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet. 2008; 83:112–9. [PubMed: 18589396]
10. Jacobs KB, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet. 2009
11. 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]
12. Pe’er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008; 32:381–5. [PubMed: 18348202]
13. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008; 32:227–34. [PubMed: 1830295]
14. Karell K, et al. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum Immunol. 2003; 64:469–77. [PubMed: 12651074]
15. Raychaudhuri S, et al. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat Genet. 2009
16. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5:e1000534. [PubMed: 19557189]
17. Smyth DJ, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med. 2008; 359:2767–77. [PubMed: 19073967]
18. Coenen MJ, et al. Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum Mol Genet. 2009
19. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–7. [PubMed: 19474294]
20. Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: Disease-centered and Gene-centered Views of the Evolving Knowledge of Human Genetic Associations. Bioinformatics. 2009
21. Han JW, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet. 2009; 41:1234–7. [PubMed: 19838193]
22. Hunt KA, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet. 2008; 40:395–402. [PubMed: 18311140]
23. Allen PM. Themis imposes new law and order on positive selection. Nat Immunol. 2009; 10:805–6. [PubMed: 19621038]
24. Sato T, et al. Dual functions of Runx proteins for reactivating CD8 and silencing CD4 at the commitment process into CD8 thymocytes. Immunity. 2005; 22:317–28. [PubMed: 15780989]
25. Woolf E, et al. Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. Proc Natl Acad Sci U S A. 2003; 100:7731–6. [PubMed: 12796513]
26. Wang J, Fu YX. LIGHT (a cellular ligand for herpes virus entry mediator and lymphotixin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide. J Immunol. 2003; 170:3986–93. [PubMed: 12682226]
27. Zamisch M, et al. The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus. J Exp Med. 2009
28. Vafiadis P, et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. Nat Genet. 1997; 15:289–92. [PubMed: 9054944]
29. Bonasio R, et al. Clonal deletion of thymocytes by circulating dendritic cells homing to the thymus. Nat Immunol. 2006; 7:1092–100. [PubMed: 16951687]
30. Klein L, Hinterberger M, Wirmserberger G, Kyewski B. Antigen presentation in the thymus for positive selection and central tolerance induction. Nat Rev Immunol. 2009; 9:833–844. [PubMed: 19935803]
31. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. Science. 2009
32. Trynka G, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut. 2009; 58:1078–83. [PubMed: 19240061]
33. Garner CP, et al. Replication of celiac disease UK genome-wide association study results in a US population. Hum Mol Genet. 2009; 18:4219–25. [PubMed: 19648293]
34. Plenge RM, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet. 2007; 39:1477–82. [PubMed: 17982456]
35. Franke L, et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet. 2008; 82:1316–33. [PubMed: 18519066]
36. 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]
37. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]
38. Yu K, et al. Population substructure and control selection in genome-wide association studies. PLoS One. 2008; 3:e2551. [PubMed: 18596976]
39. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000; 405:847–56. [PubMed: 10866211]
40. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207–10. [PubMed: 11752295]
41. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19:185–93. [PubMed: 12538238]
42. Sherlock G. Analysis of large-scale gene expression data. Brief Bioinform. 2001; 2:350–62. [PubMed: 11808747]
43. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000; 97:10101–6. [PubMed: 10963673]
44. Heap GA, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. Hum Mol Genet. 2009
45. Heap GA, et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med Genomics. 2009; 2:1. [PubMed: 19128478]
46. Franke L, Jansen RC. eQTL analysis in humans. Methods Mol Biol. 2009; 573:311–28. [PubMed: 19763935]
Figure 1. Co-expression analysis of genes mapping to 40 genome-wide significant and suggestive celiac disease regions in 33,109 heterogenous human samples from the Gene Expression Omnibus

Genes mapping within a 1Mb window of associated SNPs (Table 2) were tested for interaction with genes from other loci. Interactions with Pearson correlation $>0.5$ shown ($P<10^{-100}$). Only the genes known to contain causal mutations ($HLA-$DQA1, $HLA-$DQB1) were analysed from the HLA region. “HLA-DQB2 / HLA-DQB1” is a single expression probeset mapping to both genes. No probe for THEMIS was present on the earlier version of the U133 array, however in a subset analysis of U133 Plus2.0 data, THEMIS is co-expressed in the major immune gene cluster.
### Table 1

Sample collections and genotyping platforms

| Collection | Country | Celiac disease cases | Controls |
|------------|---------|----------------------|----------|
|            |         | Sample size (pre-QC) | Sample size (post-QC) | Platform<sup>c</sup> | Sample size (pre-QC) | Sample size (post-QC) | Platform<sup>c</sup> |
| Stage 1: Genome wide association |         |                      |                      |                       |                      |                      |                       |
| 1<sup>e</sup> | UK      | 778                  | 737                  | Illumina Hap300v1-1   | 2,596                | 2,596                | Illumina Hap550-2v3   |
| 2<sup>e</sup> | UK      | 1,922                | 1,849                | Illumina 670-QuadCustom_v1 | 5,069                | 4,936                | Illumina 1.2M-DuoCustom_v1 |
| 3<sup>e</sup> | Finland | 674                  | 647                  | Illumina 670-QuadCustom_v1 | 1,839                | 1,829                | Illumina 600-Quad     |
| 4<sup>b</sup> | Netherlands | 876              | 803                  | Illumina 670-QuadCustom_v1 | 980                  | 846                  | Illumina 670-QuadCustom_v1 |
| 5<sup>e</sup> | Italy   | 541                  | 497                  | Illumina 670-QuadCustom_v1 | 580                  | 543                  | Illumina 670-QuadCustom_v1 |
| Analysis of Hap300 markers |         |                      |                      |                       |                      |                      |                       |
| Analysis of additional Hap550 markers |         |                      |                      |                       |                      |                      |                       |
| Stage 2: Follow-up |         |                      |                      |                       |                      |                      |                       |
| 6          | USA     | 987                  | 973                  | Illumina GoldenGate   | 615                  | 555                  | Illumina GoldenGate   |
| 7          | Hungary | 979                  | 965                  | Illumina GoldenGate   | 1,026                | 1,067                | Illumina GoldenGate   |
| 8<sup>i</sup> | Ireland | 653                  | 597                  | Illumina GoldenGate   | 1,489                | 1,456                | Illumina GoldenGate   |
| 9          | Poland  | 599                  | 564                  | Illumina GoldenGate   | 745                  | 716                  | Illumina GoldenGate   |
| 10         | Spain   | 558                  | 550                  | Illumina GoldenGate   | 465                  | 433                  | Illumina GoldenGate   |
| 11<sup>e</sup> | Italy   | 1,056                | 1,010                | Illumina GoldenGate   | 864                  | 804                  | Illumina GoldenGate   |
| 12<sup>e</sup> | Finland | 270                  | 259                  | Illumina GoldenGate   | 653<sup>j</sup>      | 653                  | Illumina 600-Quad     |
| Subtotal   |         |                      |                      |                       |                      |                      |                       |
| Analysis of Hap300 markers, and follow-up (91 SNPs) |         |                      |                      |                       |                      |                      |                       |
| Analysis of additional Hap550 markers, and follow-up (60 SNPs) |         |                      |                      |                       |                      |                      |                       |

<sup>a</sup>Sample numbers attempted for genotyping, before any quality control (QC) steps were applied.

<sup>b</sup>Sample numbers after all quality control (QC) steps (used in the association analysis).

<sup>c</sup>All platforms contain a common set of Hap300 markers; the Hap550, 610-Quad, 670-Quad and 1.2M contain a common set of Hap550 markers.
dFinnish stage 2 controls were individuals within the Finrisk collection for whom Illumina 610-Quad genotype data became available after the completion of stage 1.

eAs an additional quality control step, we performed case-case and control-control comparisons for collection 1 versus 2, and collection 3 versus 12, for the 40 SNPs in Table 2 and observed no markers with P<0.01. We did observe (as expected) differences for collection 5 versus 11, from Northern and Southern Italy, respectively.

fAll 737 post-QC cases reported in a previous GWAS.

g690 of the post-QC cases and 1150 of the post-QC controls were included in previous GWAS follow-up studies.

h498 of the post-QC cases and 767 of the post-QC controls were included in previous GWAS follow-up studies.

i352 of the post-QC cases and 921 of the post-QC controls were included in previous GWAS follow-up studies.

jSome of these data were generated elsewhere, and some prior quality control steps (information not available) had been applied.
Table 2

Genomic regions with the strongest association signals for celiac disease

| Chr | Position (bp) | SNP | LD block (Mb) | Minor allele | Minor allele freq | P_{GWAS} 4533 cases, 80750 controls | P_{indep} 4953 cases, 5684 controls | P_{meta} 9451 cases, 16434 controls | Odds ratioc [95% CI] | Multiple independent association signalsd | Ref | ReSeq Genes in LD block | Genes of Interest |
|-----|---------------|-----|--------------|--------------|------------------|-------------------|-----------------|--------------------------|-------------------|----------------------------------|-----|----------------|-----------------|
| 1   | 119030416     | rs2161316 | 119.73-119.81 | C            | 0.160            | 1.45×10^{-12}    | 5.95×10^{-6}   | 8.95×10^{-6}             | 2.20×10^{-17}    | 0.80 [0.76-0.84]                 | 22  | 1               | RGS1             |
| 2   | 6040333       | rs13304364 | 60.78-61.74   | G            | 0.401            | 4.92×10^{-8}    | 5.17×10^{-8}   | 3.31×10^{-8}             | 3.46×10^{-13}    | 1.15 [1.11-1.20]                 | 32  | 8               | REL, AIMS2        |
| 2   | 102347000     | rs917999   | 102.22-102.37 | A            | 0.236            | 5.97×10^{-5}    | 7.83×10^{-4}   | 1.11×10^{-5}             | 1.34 [1.19-1.53] | yes                             |
| 2   | 181704220     | rs13050763 | 181.50-181.97 | G            | 0.448            | 2.02×10^{-6}    | 3.21×10^{-5}   | 4.74×10^{-5}             | 1.13 [1.09-1.18] | 22  | 5               | ITGA4, UBE2E3     |
| 2   | 204510825     | rs4653774  | 204.40-204.64 | A            | 0.223            | 8.80×10^{-8}    | 9.49×10^{-4}   | 5.80×10^{-9}             | 1.14 [1.09-1.19] | 17  | 2               | CTLA4, KIF15, CD8 |
| 3   | 46210205      | rs13089911 | 45.00-46.17   | A            | 0.097            | 2.53×10^{-11}   | 1.96×10^{-7}   | 3.26×10^{-17}             | 1.30 [1.23-1.39] | yes  |
| 3   | 16114744      | rs1780546  | 161.07-161.23 | G            | 0.125            | 4.56×10^{-8}    | 9.57×10^{-12} | 9.39×10^{-28}             | 1.36 [1.29-1.44] | yes  |
| 3   | 18959328      | rs1404510  | 189.55-189.62 | A            | 0.485            | 9.49×10^{-24}   | 3.63×10^{-18} | 5.98×10^{-30}             | 1.29 [1.25-1.34] | 22  | 1               | LPP              |
| 4   | 123334052     | rs13151961 | 123.19-123.78 | G            | 0.142            | 6.31×10^{-8}    | 4.45×10^{-11} | 2.18×10^{-7}             | 0.74 [0.70-0.78] | 1   |
| 6   | 32713862      | rs2187668  | gene identified | A          | 0.258            | <10^{-5}        | <10^{-9}        | 6.23 [5.95-6.52]         | yes  |
| 6   | 133814761     | rs2327832  | 137.92-138.37 | G            | 0.216            | 1.41×10^{-14}   | 1.97×10^{-6}   | 4.46×10^{-19}             | 1.23 [1.17-1.28] | 32  | 0               | TNFAP3           |
| 6   | 15938965      | rs178074   | 159.24-159.45 | A            | 0.434            | 3.14×10^{-8}    | 1.56×10^{-9}   | 2.94×10^{-35}             | 1.16 [1.12-1.21] | 22  | 2               | TACAP            |
| 12  | 110492139     | rs653178   | 110.19-111.31 | G            | 0.495            | 6.03×10^{-4}    | 1.47×10^{-6}   | 7.15×10^{-31}             | 1.20 [1.15-1.24] | 22  | 13              | SH2R3            |
| 18  | 12799340      | rs1803217  | 12.73-12.91   | G            | 0.168            | 5.52×10^{-7}    | 1.04×10^{-4}   | 2.52×10^{-10}             | 1.17 [1.12-1.23] | 17  | 1               | TNFRSF14, MMEL3 |

New loci, genome-wide significant evidence (P_{meta} < 5 × 10^{-8})

| 1   | 2516606       | rs3488816  | 2.40-2.78     | G            | 0.339            | 4.93×10^{-7}    | 1.17×10^{-5}   | 3.28×10^{-9}             | 0.89 [0.85-0.92] | 4   |
| 2   | 2517616       | rs10903122 | 25.11-25.18   | A            | 0.480            | 3.21×10^{-5}    | 8.44×10^{-7}   | 1.73×10^{-10}            | 0.89 [0.85-0.92] | 22  | 4               | RUNX3            |
| 3   | 199153800     | rs265467   | 199.12-199.31 | A            | 0.357            | 6.46×10^{-5}    | 1.34×10^{-5}   | 4.19×10^{-9}             | 0.89 [0.86-0.92] | 2   |
| 2   | 6845249       | rs17035578f| 68.39-68.54   | G            | 0.278            | 1.34×10^{-5}    | 1.41×10^{-4}   | 7.79×10^{-9}             | 0.88 [0.84-0.92] | 2   |
| 3   | 32990473      | rs13314983f| 32.90-33.06   | C            | 0.464            | 6.87×10^{-6}    | 1.09×10^{-4}   | 3.27×10^{-9}             | 1.13 [0.98-1.17] | 2   |
| 12  | 20260846      | rs11712665f| 130.59-130.78 | C            | 0.394            | 5.40×10^{-7}    | 1.72×10^{-5}   | 8.03×10^{-9}             | 1.13 [0.86-1.17] | 5   |
| 6   | 9098333       | rs10806425 | 90.96-91.10   | A            | 0.597            | 9.46×10^{-6}    | 9.29×10^{-6}   | 3.89×10^{-10}            | 1.13 [0.86-1.17] | 1   |
| 6   | 128320491     | rs802734   | 127.99-128.38 | G            | 0.311            | 1.36×10^{-6}    | 1.70×10^{-9}   | 2.62×10^{-14}            | 1.17 [1.12-1.22] | 2   |

Note: Author manuscript; available in PMC 2010 October 01.
| Chr | Position (bp) | SNP     | LD block (Mb) | Minor allele | Minor allele freq | $P_{\text{GWAS}}$ 4533 cases, 18789 controls | $P_{\text{follow-up}}$ 498 cases, 5844 controls | $P_{\text{conditional}}$ 9451 cases, 5844 controls | Odds ratio $^c$ [95% CI] | Multiple independent association $^d$ | Ref | RefSeq | Genes of Interest and GRAIL annotation $^e$ |
|-----|---------------|---------|---------------|-------------|------------------|---------------------------------------------|-----------------------------------------------|-------------------------------------------------|--------------------------|---------------------------|-----|--------|--------------------------------------|
| 8   | 129333771     | rs9722699 | 129.21-129.37 | G           | 0.238           | 8.14 x 10^-6                                 | 1.00 x 10^-4                                  | 3.28 x 10^-6                                   | 0.88 [0.84-0.91]                           | yes | 0    | ZMIZI |
| 11  | 127866184     | rs11221312 | 127.84-127.99 | A           | 0.237           | 4.74 x 10^-11                                 | 9.98 x 10^-7                                  | 5.28 x 10^-16                                  | 1.21 [1.16-1.27]                           | yes | 1    | EFS1 |
| 16  | 11311394      | rs1292822 | 11.22-11.39   | A           | 0.161           | 1.07 x 10^-5                                  | 7.59 x 10^-5                                  | 3.12 x 10^-6                                   | 0.86 [0.82-0.91]                           | yes | 4    | CRTC3, SOCS1, CLEC16A |
| 21  | 44471849      | rs4819388  | 44.42-44.47   | A           | 0.280           | 3.24 x 10^-5                                  | 1.66 x 10^-5                                  | 2.46 x 10^-9                                   | 0.88 [0.84-0.92]                           | yes | 2    | KCOSLG |

New loci, suggestive evidence (either A. $10^{-6}$ $P_{\text{GWAS}}$ or $5 x 10^{-8}$ and/or B. $P_{\text{GWAS}}$ $< 10^{-4}$ and $P_{\text{follow-up}}$ $< 0.01$)

| Chr | Position (bp) | SNP     | LD block (Mb) | Minor allele | Minor allele freq | $P_{\text{GWAS}}$ 4533 cases, 18789 controls | $P_{\text{follow-up}}$ 498 cases, 5844 controls | $P_{\text{conditional}}$ 9451 cases, 5844 controls | Odds ratio $^c$ [95% CI] | Multiple independent association $^d$ | Ref | RefSeq | Genes of Interest and GRAIL annotation $^e$ |
|-----|---------------|---------|---------------|-------------|------------------|---------------------------------------------|-----------------------------------------------|-------------------------------------------------|--------------------------|---------------------------|-----|--------|--------------------------------------|
| 1   | 7969259       | rs1272742 | 7.84-8.13     | A           | 0.185           | 3.06 x 10^-5                                  | 8.21 x 10^-6                                  | 9.11 x 10^-6                                   | 1.14 [1.09-1.20]                           | yes | 4    | PARK7, TINF5 |
| 6   | 61564451      | rs6691768 | 61.32-61.62   | G           | 0.378           | 2.63 x 10^-5                                  | 1.16 x 10^-5                                  | 1.19 x 10^-7                                   | 0.90 [0.87-0.94]                           | yes | 1    | NFA |
| 1   | 165678008     | rs864537  | 165.43-165.71 | A           | 0.391           | 1.01 x 10^-7                                  | 9.25 x 10^-5                                  | 3.80 x 10^-7                                   | 0.91 [0.87-0.94]                           | yes | 1    | CD247 |
| 1   | 170977623     | rs859607  | 170.87-171.20 | A           | 0.486           | 8.15 x 10^-5                                  | 5.68 x 10^-5                                  | 1.75 x 10^-6                                   | 1.10 [0.06-1.14]                           | yes | 1    | FASLG, TINF5, TINF54 |
| 3   | 69335889      | rs6005529 | 69.27-69.37   | A           | 0.097           | 4.84 x 10^-4                                  | 7.66 x 10^-4                                  | 1.46 x 10^-7                                   | 1.19 [1.12-1.27]                           | yes | 1    | FRMD4B |
| 3   | 170974795     | rs1036599 | 170.84-171.09 | A           | 0.252           | 2.99 x 10^-7                                  | 6.63 x 10^-4                                  | 4.57 x 10^-7                                   | 1.12 [0.07-1.16]                           | yes | 3    | ? |
| 6   | 328566        | rs1031380  | 0.32-0.40     | A           | 0.080           | 9.14 x 10^-4                                  | 1.48 x 10^-5                                  | 5.58 x 10^-6                                   | 1.21 [1.03-1.29]                           | yes | 1    | RHAG |
| 7   | 37341035      | rs6744941 | 37.32-37.41   | A           | 0.170           | 1.37 x 10^-5                                  | 2.63 x 10^-5                                  | 1.56 x 10^-7                                   | 1.14 [0.09-1.20]                           | yes | 1    | ELMO1 |
| 11  | 49737316      | rs276201   | 49.63-49.96   | A           | 0.184           | 3.35 x 10^-5                                  | 5.06 x 10^-5                                  | 6.64 x 10^-7                                   | 1.13 [0.08-1.18]                           | yes | 1    | ? |
| 16  | 68347957      | rs4899260 | 68.24-68.39   | A           | 0.263           | 4.55 x 10^-5                                  | 2.21 x 10^-4                                  | 3.92 x 10^-7                                   | 1.12 [1.07-1.16]                           | yes | 2    | ZFP56,1 |
| 17  | 42230599      | rs2074404 | 41.40-42.25   | C           | 0.250           | 5.03 x 10^-5                                  | 5.96 x 10^-5                                  | 1.23 x 10^-6                                   | 0.90 [0.86-0.94]                           | yes | 10   | ? |
| 22  | 20312892      | rs229428   | 20.14-20.35   | A           | 0.201           | 2.49 x 10^-4                                  | 4.13 x 10^-4                                  | 1.84 x 10^-7                                   | 1.13 [0.08-1.19]                           | yes | 6    | UBE2L3, YDC |
| X   | 12881445      | rs5979785  | 12.82-12.93   | G           | 0.263           | 6.32 x 10^-6                                  | 2.18 x 10^-5                                  | 6.36 x 10^-6                                   | 0.84 [0.84-0.92]                           | yes | 1    | IL87, TL18 |

$^a$The most significantly associated SNP from each region is shown.

$^b$LD regions were defined by extending 0.1 cM to the left and right of the focal SNP as defined by the HapMap3 recombination map. All chromosomal positions are based on NCBI build 36 coordinates.

$^c$Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls.

$^d$Evidence from logistic regression at a genome-wide significant or suggestive level of significance after conditioning on other associated SNPs (see Supplementary Table 2).

$^e$Selected named genes within or adjacent to the same LD block as the associated SNPs, causality is not proven. In particular, other genes and other causal mechanisms may exist. Gene names underlined are identified from GRAIL15,16 analysis (Online Methods) with $P_{\text{ext}}<0.001$.
These markers were present on the Hap550 but not Hap300 SNP sets, and are not genotyped for 737 cases and 2596 controls in the stage I GWAS, and combined dataset analyses. Only minor changes in P values were observed when these genotypes were imputed and included in analysis.

The IRF4 region (specifically rs9738805, $r^2 = 0.08$ with rs1033180 in HapMap CEU) was previously identified as showing strong geographical differentiation. Association with celiac disease was still observed after correction for population stratification using either a structured association approach (corrected $GWAS P = 5.16 \times 10^{-6}$, 478×2×2 CMH test) or principal components correction (uncorrected $GWAS P = 7.05 \times 10^{-5}$, corrected $GWAS P = 2.28 \times 10^{-5}$, Cochran-Armitage trend test combined using weighted Z score) (Online Methods). However, definitive exclusion of population stratification would require family-based association studies.
| SNP | Chr | SNP position | Probe | Centre | Position | Illumina Array AddressID | Expression dataset | Gene name | eQTL P-value |
|-----|-----|-------------|-------|--------|----------|--------------------------|-------------------|-----------|--------------|
| rs3748816 | 1 | 2516606 | 2515115 | 2482221 | 650452 | HT-12 | 1.66 × 10^{-5} |
| rs3748816 | 1 | 2516606 | 2515115 | 2300328 | 712214 | Ref-8v2 | 3.50 × 10^{-10} |
| rs3748816 | 1 | 2516606 | 2513115 | 2001279 | 2.05 × 10^{-12} | MIREJ | 3.00 × 10^{-10} |
| rs3748816 | 1 | 19885380 | 19885380 | 24569180 | 2.45 × 10^{-12} | MIREJ | 3.00 × 10^{-10} |
| rs3816281 | 2 | 61263810 | 61263810 | 1170200 | 1.50 × 10^{-10} |
| rs917997 | 2 | 102418571 | 102418571 | 601200 | 7.97 × 10^{-6} |
| rs13098911 | 3 | 46255176 | 46255176 | 3796171 | 5.50 × 10^{-10} |
| rs6441794 | 3 | 46255176 | 46255176 | 2190671 | 1.02 × 10^{-10} |
| rs11922594 | 3 | 120683364 | 120683364 | 6550288 | 5.09 × 10^{-17} |
| rs10806425 | 6 | 80622540 | 80622540 | 3520349 | 1.92 × 10^{-3} |
| rs653178 | 12 | 110970417 | 110970417 | 400323 | 1.44 × 10^{-3} |
| SNP  | Chr | SNP position | Probe Centre Position | Illumina Array AddressID | Expression dataset | Gene name   | eQTL P value |
|------|-----|--------------|-----------------------|--------------------------|--------------------|------------|-------------|
| rs653178 | 12 | 110492139    | 110849406             | 2070376                  | HT-12              | TMEM116    | 3.60 × 10⁻⁶ |
| rs653178 | 12 | 110492139    | 110849406             | 3190129                  | Ref-8v2            | TMEM116    | 1.51 × 10⁻³ |
| rs12928822 | 16 | 11311394     | 11335627              | 4540072                  | Ref-8v2 + HT-12    | C16s175    | 1.02 × 10⁻⁴ |
| rs4819388 | 21 | 44473849     | 44049657              | 7200373                  | Ref-8v2            | RRP1       | 2.62 × 10⁻³ |

Loci with suggestive evidence (either A. $10^{-6} < P_{combined} < 5 \times 10^{-8}$ and/or B. $P_{GWAS} < 10^{-4}$ and $P_{follow-up} < 0.01$)

- rs12727642 | 1 | 796259       | 7956338               | 61093                    | Ref-8v2 + HT-12    | PARK7      | 9.76 × 10⁻⁶ |
- rs685537 | 1 | 165678008    | 165710482             | 6209400                  | Ref-8v2 + HT-12    | CD247      | 1.77 × 10⁻³ |
- rs685537 | 1 | 165678008    | 165710482             | 3890689                  | HT-12              | CD247      | 2.93 × 10⁻⁷ |
- rs674495 | 7 | 3734035     | 37157561              | 2730154                  | Ref-8v2 + HT-12    | ELM101     | 5.40 × 10⁻⁶ |
- rs2074404 | 17 | 42220599    | 41824045              | 3532672                  | Ref-8v2 + HT-12    | LRRC37A    | 1.17 × 10⁻⁴ |
- rs2074404 | 17 | 42220599    | 42106695             | 5260138                  | Ref-8v2 + HT-12    | NSF        | 1.20 × 10⁻⁴ |
- rs2074404 | 17 | 42220599    | 42106695             | 1440484                  | Ref-8v2 + HT-12    | NSF        | 4.26 × 10⁻⁴ |
- rs2074404 | 17 | 42220599    | 42223012             | 4003615                  | HT-12              | WNT3       | 2.77 × 10⁻³ |
- rs2074404 | 17 | 42220599    | 42485154             | 4880307                  | HT-12              | LOC389397  | 1.78 × 10⁻⁹ |
- rs228428 | 22 | 2031292     | 20308188             | 1230242                  | Ref-8v2 + HT-12    | UBE2L3     | 1.96 × 10⁻⁹ |
- rs5979785 | X  | 12881445    | 12842944             | 6480360                  | Ref-8v2 + HT-12    | TLR8       | 3.88 × 10⁻³ |
- rs5979785 | X  | 12881445    | 12842944             | 3390612                  | Ref-8v2 + HT-12    | TLR8       | 1.07 × 10⁻⁹ |

See Supplementary Figures 2 & 3 for detailed results, and Supplementary Table 3 for more detail of Illumina expression probes.

a We tested the SNP with the strongest association from 34 of 39 non-HLA loci ($P_{combined} < 10^{-6}$, Table 2), Hap300 proxy SNPs for 4 further loci, and a second independently associated SNP from 6 loci, for correlation with gene expression in PAXgene blood RNA in up to 1,349 individuals. 1 locus (containing ETS1) where an adequate proxy SNP was not available was not included for the eQTL analysis. SNP-gene expression correlations were tested for probes within a 1 Mb window. Results are presented for SNPs showing significant correlations with cis gene expression after controlling false discovery rate at 5% (corresponding to $P<0.0028$).
b All chromosomal positions are based on NCBI build-36 coordinates. Probe centre position was determined by re-mapping probe sequences to the human transcriptome and calculated from the mid-point of the transcript start and transcript end positions in genomic co-ordinates.
c 'HT-12' comprise 1240 individuals with blood gene expression assayed using Illumina Human HT-12v3 arrays, 'Ref-8v2' comprise 229 individuals with blood gene expression assayed using Illumina Human-Ref-8v2 arrays (Online Methods).
d Spearman rank correlation of genotype and residual variance in transcript expression. Meta-analysis eQTL P value shown if both datasets had identical probes.
e Second, independently associated SNP from this locus.
f Proxy SNP, $r^2=0.61$ in HapMap CEU with most associated SNP rs1712165.
Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.

Different Illumina probe sequences with the same Probe Centre Position.