OPEN ACCESS Freely available online

PLoS GENETICS

A Meta-Analysis of Genome-Wide Association Scans Identifies IL18RAP, PTPN2, TAGAP, and PUS10 As Shared Risk Loci for Crohn's Disease and Celiac Disease

Eleonora A. M. Festen[1,2☯], Philippe Goyette[3☯], Todd Green[4☯], Gabrielle Boucher[3], Claudine Beauchamp[3], Gosia Trynka[2], Patrick C. Dubois[5], Caroline Lagacé[3], Pieter C. F. Stokkers[6], Daan W. Hommes[7], Donatella Barisani[8], Orazio Palmieri[9], Vito Annese[9], David A. van Heel[5], Rinse K. Weersma[1¶*], Mark J. Daly[4,10¶], Cisca Wijmenga[2¶], John D. Rioux[3¶*]

1 Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands, **2** Department of Genetics, University Medical Centre Groningen and University of Groningen, Groningen, The Netherlands, **3** Research Center, Université de Montréal and the Montreal Heart Institute, Montreal, Quebec, Canada, **4** The Broad Institute, Cambridge, Massachusetts, United States of America, **5** Centre for Digestive Diseases, Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, **6** Department of Gastroenterology and Hepatology, Academic Medical Centre, Amsterdam, The Netherlands, **7** Department of Gastroenterology and Hepatology, Leiden University Medical Centre, Leiden, The Netherlands, **8** Department of Experimental Medicine, Faculty of Medicine, University of Milano-Bicocca, Monza, Italy, **9** U. O. Gastroenterologia ed Endoscopia Digestiva, Ospedale "Casa Sollievo della Sofferenza", IRCCS, San Giovanni Rotondo, Italy, **10** Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Crohn's disease (CD) and celiac disease (CelD) are chronic intestinal inflammatory diseases, involving genetic and environmental factors in their pathogenesis. The two diseases can co-occur within families, and studies suggest that CelD patients have a higher risk to develop CD than the general population. These observations suggest that CD and CelD may share common genetic risk loci. Two such shared loci, *IL18RAP* and *PTPN2*, have already been identified independently in these two diseases. The aim of our study was to explicitly identify shared risk loci for these diseases by combining results from genome-wide association study (GWAS) datasets of CD and CelD. Specifically, GWAS results from CelD (768 cases, 1,422 controls) and CD (3,230 cases, 4,829 controls) were combined in a meta-analysis. Nine independent regions had nominal association p-value $<1.0\times10^{-5}$ in this meta-analysis and showed evidence of association to the individual diseases in the original scans (p-value $<1\times10^{-2}$ in CelD and $<1\times10^{-3}$ in CD). These include the two previously reported shared loci, *IL18RAP* and *PTPN2*, with p-values of 3.37×10^{-8} and 6.39×10^{-9}, respectively, in the meta-analysis. The other seven had not been reported as shared loci and thus were tested in additional CelD (3,149 cases and 4,714 controls) and CD (1,835 cases and 1,669 controls) cohorts. Two of these loci, *TAGAP* and *PUS10*, showed significant evidence of replication (Bonferroni corrected p-values <0.0071) in the combined CelD and CD replication cohorts and were firmly established as shared risk loci of genome-wide significance, with overall combined p-values of 1.55×10^{-10} and 1.38×10^{-11} respectively. Through a meta-analysis of GWAS data from CD and CelD, we have identified four shared risk loci: *PTPN2*, *IL18RAP*, *TAGAP*, and *PUS10*. The combined analysis of the two datasets provided the power, lacking in the individual GWAS for single diseases, to detect shared loci with a relatively small effect.

Citation: Festen EAM, Goyette P, Green T, Boucher G, Beauchamp C, et al. (2011) A Meta-Analysis of Genome-Wide Association Scans Identifies IL18RAP, PTPN2, TAGAP, and PUS10 As Shared Risk Loci for Crohn's Disease and Celiac Disease. PLoS Genet 7(1): e1001283. doi:10.1371/journal.pgen.1001283

Editor: Michel Georges, University of Liège, Belgium

Received March 10, 2010; **Accepted** December 20, 2010; **Published** January 27, 2011

Funding: The study was supported by the Celiac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009); by the Netherlands Organization for Scientific Research (NWO, VICI grant 918.66.620 to CW, AGIKO grant 92.003.533 to EAMF, KF grant 907.00.281 to RKW); by the Wellcome Trust (grant WT084743MA to DAvH); by the National Institutes of Allergy and Infectious Diseases (grant AI065687, AI067152 to JDR); by the National Institute of Diabetes and Digestive and Kidney Diseases (grant DK064869, DK062432 to JDR); and by the Crohn's and Colitis Foundation of America (grant SRA512 to JDR). EAMF is a MD-medical research trainee with financial support from The Netherlands Organisation for Health Research and Development. GT was awarded a Ter Meulen Fund travel grant by the Royal Netherlands Academy of Arts and Sciences (KNAW). PCD is a MRC Clinical Training Fellow (G0700545). The British 1958 Birth Cohort collection was funded by the UK Medical Research Council (grant G0000934) and the Wellcome Trust (grant 068545/Z/02). Funding for the project Wellcome Trust Case-Control Consortium 2 data was provided by the Wellcome Trust under award 085475. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: R.K.Weersma@int.umcg.nl (RKW); john.david.rioux@umontreal.ca (JDR)

☯ These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Author Summary

Celiac disease and Crohn's disease are both chronic inflammatory diseases of the digestive tract. Both of these diseases are complex genetic traits with multiple genetic and non-genetic risk factors. Recent genome-wide association (GWA) studies have identified some of the genetic risk factors for these diseases. Interestingly, in addition to some similarities in phenotype, these studies have shown that CelD and CD share some genetic risk factors. Specifically, by comparing the results of independent GWA studies of CD and CelD, two genetic risk loci were found in common: the *PTPN2* locus and the *IL18RAP* locus. Therefore, in order to directly test for additional shared genetic risk factors, we combined the GWA results from two large studies of CelD and CD, essentially creating a combined phenotype with anyone with CD or CelD being coded as affected. Association results were then replicated in additional cohorts of CelD and CD. It is expected that shared risk loci should show association in this analysis, whereas the signal of risk loci specific to either of the two diseases should be diluted. With this method of meta-analysis, we identified next to *PTPN2* and *IL18 RAP* two loci harbouring *TAGAP* and *PUS10* as shared risk loci for Crohn's disease and celiac disease at genome-wide significance.

Introduction

Crohn's disease (CD) and celiac disease (CelD) are both chronic intestinal inflammatory diseases. In CD inflammation can occur throughout the gastrointestinal tract but most commonly affects the ileal part of the small intestine. While the causative antigen(s) for this inflammation is unknown, it is thought that the disease arises as a reaction to the normal commensal flora of the bowel in a genetically susceptible individual [1,2]. In CelD inflammation is limited to the small intestine. CelD is caused by a reaction to gluten, a dietary peptide present in wheat, barley and rye [3,4]. In both CelD and CD contact between antigens and antigen-presenting cells (APCs) seems to be facilitated by an initial increase in intestinal permeability [5]. In both diseases the subsequent inflammatory response follows a T helper 1 pattern characterized by tumor necrosis factor beta (TNF-β) and interferon gamma (IFN-γ) production and a T helper 17 response marked by the production of interleukin 17 [5].

Although uncommon, it has been observed that CelD and CD can co-occur within families or even within individual patients; there appears to be a greater prevalence of CD among CelD patients than in the general population, although the relatively low prevalence of CD makes it difficult to establish this effect [6]. It is now well accepted that the risk for CD and CelD is partly determined by genetic factors, and recently many genetic risk factors for CelD and CD have been identified. Two genetic risk loci were previously shown to be shared between CelD and CD: a locus on 18p11 containing the *PTPN2* (protein tyrosine phosphatase, non-receptor type 2) gene and a locus on 2q12 containing the *IL18RAP* (interleukin 18 receptor accessory protein) gene [7–13]. While these observations confirm the existence of shared risk loci for CD and CelD, additional such shared risk loci are likely to exist.

There are two possible approaches for identifying shared risk loci. One approach is to test known risk loci from one disease in patient-control cohorts from the other disease. This approach has already been successfully applied in a cross study between CelD and type 1 diabetes (T1D), where four shared risk loci were identified some of which were previously unknown to be associated to CelD [11]. However, this approach relies on previously identified risk alleles, indicating that there probably are many more unknown common risk loci for T1D and CelD. In addition, some of the shared loci will not have a large enough effect in the individual diseases to have been identified by previous genetic studies. A second approach that tackles this problem is to analyze genetic data from two similar diseases as a single unified disease phenotype against healthy controls. Such an analysis would be expected to dilute disease-specific genetic associations, but increase the power for finding shared genetic risk loci of small effect in the individual diseases. The availability of genome-wide association studies (GWAS) performed in both CelD and CD provides large case-control genotyping datasets that enabled us to perform a cross-disease genome-wide meta-analysis in the aim of identifying novel shared risk loci.

To identify novel shared risk loci between CelD and CD, we performed a meta-analysis of two recently published GWAS: a large meta-analysis of three CD GWAS by the International IBD Genetics Consortium and a CelD GWAS in a British population. To confirm identified risk loci, we used a combination of Italian and Dutch CD cohorts and of British, Italian and Dutch CelD cohorts.

Results

Meta-analysis

We have performed a meta-analysis of 471,504 SNPs from genome-wide datasets of CD (3230 cases, 4829 controls) and CelD (768 cases, 1422 controls) in order to identify shared risk loci between these 2 diseases.

A quantile-quantile (Q-Q) plot of the association p-values for single-SNP Z scores from the meta-analysis was performed (Figure S1) and shows an excess of significant associations above what would be expected by chance. We observe a low inflation factor of 1.08, which is expected given the inflation observed in each of the original studies: 1.05 for CelD and 1.16 for CD. A Manhattan plot of the current study (Figure S2) highlights many strong association signals, several of which corresponding to previously reported CD and CelD loci; however, most of these show strong association in only one of the 2 diseases and have thus not been followed up due to the design of the current study. In addition, given the design of the original CelD GWAS, which included only individuals that were positive for the risk-associated allele *HLA-DQ2*, association to the major histocompatibility complex (MHC) region in CelD was of no relevance since it was artificially inflated. Therefore the MHC region (Chr6:22700000..35000000 from the NCBI B36 genome build) was removed from our analysis.

The meta-analysis of the CD and CelD datasets identified 25 SNPs, from 10 independent regions, that met our criteria for association (association with CelD at p-value $<1\times10^{-2}$ and with CD at p-value $<1\times10^{-3}$ in the original scans, as well as a nominal association p-value more significant than 1.0×10^{-5} in the meta-analysis) (Table 1 and Table S2). This is more than expected by chance, as we would expect no more than 3 independent regions to meet our criteria, which encouraged us to explore these specific loci further.

The strongest association signal identified in our scan was to the well accepted CD associated risk locus *CARD15* (p-values of 3.42×10^{-32}, 3.77×10^{-3} and 1.30×10^{-21} in the CD, CelD and scan datasets respectively). Given the strength and the width of the association signal peak at this locus in the CD dataset, our chances of detecting a false positive shared signal at this locus in the scan were artificially increased. Because of this, the *CARD15* locus was

Table 1. Results from the meta-analysis, replication, and combined analysis.

		Initial Analysis				Replication					Combined Analysis		
Locus	SNP	Risk Allele	CelD	CD	meta-analysis	CelD	CelD	CD	CD	meta-analysis	CelD	CD	meta-analysis
			p-value	p-value	p-value	OR**	p-value	OR**	p-value	p-value	p-value	p-value	p-value
IL18RAP*	rs6708413	G	6.50×10^{-5}	2.05×10^{-5}	8.37×10^{-8}	N.T.	N.T.	N.T.	N.T.	N.T.			
PTPN2*	rs16939895	C	2.06×10^{-3}	4.62×10^{-9}	6.39×10^{-9}	N.T.	N.T.	N.T.	N.T.	N.T.			
PUS10	rs10188217	C	8.30×10^{-3}	6.14×10^{-6}	3.33×10^{-6}	**1.14**	**4.52×10^{-5}**	**1.16**	**2.90×10^{-3}**	**6.03×10^{-7}**	**1.34×10^{-6}**	**6.16×10^{-8}**	**1.38×10^{-11}**
CCDC91	rs10771427	G	2.09×10^{-3}	5.72×10^{-5}	4.00×10^{-6}	1.00	0.94	1.02	0.79	0.80			
3q13	rs1517605	T	5.60×10^{-4}	1.98×10^{-4}	3.22×10^{-6}	1.02	0.67	0.98	0.73	0.95			
TAGAP	rs212388	C	1.22×10^{-4}	1.23×10^{-5}	7.86×10^{-7}	**1.14**	**8.88×10^{-5}**	**1.10**	**0.048**	**3.03×10^{-5}**	**7.00×10^{-7}**	**2.13×10^{-6}**	**1.55×10^{-10}**
KCNG4	rs40254	A	5.69×10^{-3}	2.97×10^{-5}	6.58×10^{-6}	0.96	0.19	1.00	0.98	0.34			
C10orf72	rs4317904#	A	2.09×10^{-4}	2.80×10^{-4}	1.87×10^{-6}	1.04	0.26	1.06	0.24	0.10			
FERMT1	rs6516104#	T	5.87×10^{-5}	6.19×10^{-4}	1.36×10^{-6}	1.04	0.47	0.97	0.73	0.76			

The meta-analysis was performed using a directional non-weighed Z-score method as explained in the methods section. Combined analyses were performed using a directional weighed Z-score method within diseases and a directional non-weighed Z-score method between diseases. Results for the combined analysis are given only for the SNPs that pass the replication thresholds (directionality in each disease and p-value <0.0071 for the combined replication data).

*, IL18RAP and PTPN2 were not followed up because they are known shared risk loci for CD and CelD;

**, OR for the replication is reported for the allele identified as the risk allele in the initial scan;

#, SNPs that were imputed in the CelD replication datasets.

N.T., not tested. CelD = Celiac disease, CD = Crohn's disease, SNP = single nucleotide polymorphism, IL18RAP = interleukin 18 receptor accessory protein, PTPN2 = protein tyrosine phosphatase, PUS10 = pseudouridylate synthase 10, CCDC91 = coiled-coil domain containing 91, TAGAP = T-cell activation RhoGTPase activating protein, KCNG4 = potassium voltage-gated channel, subfamily G, member 4, non-receptor type 2, C10orf72 = Chromosome 10 open reading frame 72, FERMT1 =-fermitin family homolog 1, 3q13, 9q13, 17p13 = intergenic regions.

doi:10.1371/journal.pgen.1001283.t001

not moved forward to replication. An evaluation of the association signal in the *in silico* CelD GWAS replication datasets confirmed that this locus did not show replication in CelD. Several of the SNPs meeting our criteria for association mapped to the known shared risk loci *IL18RAP* and *PTPN2*. Identifying these shared risk loci in the initial phase of our analysis provides proof of the effectiveness of our method. Interestingly, these two loci either reach or are very near genome-wide significance in the current meta-analysis (p-value of 8.37×10^{-8} for *IL18RAP* and of 6.39×10^{-9} for *PTPN2*), validating their previously identified role in both CD and CelD. The remaining 12 SNPs were located in seven independent regions and for each of these loci we selected the most associated SNP for testing for evidence of replication.

Replication phase

All SNPs selected for follow-up were genotyped in additional replication cohorts of CelD patients (n = 3149) and healthy controls (n = 4714) and of CD patients (n = 1941) and healthy controls (n = 1669). Given that these putative shared risk loci were selected through the combined analysis of our CD and CelD scan cohorts, a positive threshold for replication was therefore set at a corrected p-value of 0.0071 (Bonferroni corrected p-value of 0.05 for 7 independent tests) in the combined analysis of CD and CelD replication cohorts.

Only two of the 7 loci tested, *PUS10* (pseudouridylate synthase 10; RefSeq NM_144709.2) and *TAGAP* (T-cell activation GTPase activating protein; RefSeq NM_054114), showed significant replication (p-values of 6.03×10^{-7} and 3.03×10^{-6} respectively) with matching direction of association between the scan and replication datasets, as well as replication p-values more significant than 0.05 in both CD and CelD replication cohorts independently (Table 1). While neither *PUS10* nor *TAGAP* were identified as loci of genome-wide significance in the combined dataset from each disease (p-values $=1.34\times10^{-6}$ and 7.00×10^{-7} in CelD and p-values $=6.16\times10^{-8}$ and 2.13×10^{-6} in CD respectively), both reach genome-wide significance in a combined analysis of CD and CelD cohorts (p-values of 1.38×10^{-11} and 1.55×10^{-10} respectively). Based on the results calculated from the replication datasets, we also observe that the effects at these 2 loci are similar in size and direction for both CD and CelD (Table 1).

Discussion

By performing a meta-analysis of GWAS data from CD and CelD as a single disease phenotype, we have identified four risk loci shared by these 2 diseases: *PTPN2*, *IL18RAP*, *TAGAP* and *PUS10*. This meta-analysis approach provided the power, lacking in individual disease-specific GWAS datasets, to identify shared risk loci with small effects in each single disease. This approach is a powerful and versatile way of identifying shared risk loci. In fact, two of the shared loci described here, *TAGAP* and *PUS10*, would not have reached genome-wide significance without the power gained from the combined samples (scan and replication) of these 2 diseases. As the GWAS for the individual diseases increase in power, we can expect the power of the current approach to also increase enabling us to identify further shared loci.

The *TAGAP* locus identified in the current study as a shared risk factor for CD and CelD is located on chromosome 6q25.3, within a 200-kb block of linkage disequilibrium (LD). This *TAGAP* locus was previously identified as a CelD risk locus [9] but not found in previous studies of CD. *TAGAP* is the best candidate of four genes in this region of strong LD [9]. *TAGAP* is a member of the Rho-GTPase protein family, which release GTP from GTP-bound Rho, thereby acting as a molecular switch. The gene is expressed in activated T cells and appears to be important for modulating cytoskeletal changes [14]. Little is known about the

exact role of *TAGAP* in immune function, but it has been found to be co-regulated with *IL2* and is expected to play a role in T-cell activation [14].

The current study also identifies a shared risk locus between CD and CelD in the *PUS10* gene region, a locus previously described as a risk locus for CD [15]. This locus was recently identified as a risk locus in both ulcerative colitis (UC) and CelD, indicating that this locus may be a shared risk locus for these three diseases [7,16]. This latter finding further validates the approach use in this study to identify risk factors that are shared across diseases. Interestingly, the UC study identified three independent signals in this region which seem to be shared differently across these three diseases: one signal seems to be shared only between CD and UC, a second only between CelD and UC, while the third, identified in the current study, seems shared between all three diseases. Further analysis of this locus will be necessary in order to clarify the role of these different alleles in disease risk.

In this study we aimed to find shared genetic risk factors for CelD and CD by meta-analysis of GWAS data of both diseases, defining a single phenotype for these analyses. Using readily available data, we were able to reliably establish four shared loci: *PTPN2*, *IL18RAP*, *TAGAP* and *PUS10*. For many diseases with overlapping phenotypic characteristics, GWAS data is available and joint analysis of GWAS datasets of these related diseases could lead to the identification of many new shared susceptibility loci.

Materials and Methods

Subjects

For the CD aspect of the meta-analysis, we used the previously published data (available at http://www.broadinstitute.org/~jcbarret/ibd-meta/) from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) meta-analysis of 3230 CD cases and 4829 healthy controls taken from three independent CD GWAS (Table S1) [15]. A more in depth description of these cohorts and their origin can be obtained from the original publication of this meta-analysis.

Two independent cohorts were used for the CD replication phase (Table S1). The first consisted of 1217 Dutch CD cases from three Dutch university medical centers: the Academic Medical Centre Amsterdam (n = 661), the University Medical Centre Groningen (n = 322) and the University Medical Centre Leiden (n = 234); the 804 Dutch controls used for this replication cohort were obtained from cohorts of healthy partners of IBD patients from the UMC Leiden (n = 151) and the UMC Groningen (n = 120) and from healthy blood donors recruited through the Sanquin Blood bank by the UMC Utrecht and the VUMC Amsterdam (n = 533) [17,18]. The second replication cohort consisted of an Italian IBD case – control cohort (724 CD patients and 892 controls) collected at the S. Giovanni Rotondo "CSS" (SGRC) Hospital in Italy. This cohort has previously been used and characterized in several association reports [19,20].

All patients and controls were of European Caucasian descent. The diagnosis of CD required objective evidence of inflammation from radiologic, endoscopic, and/or histopathologic evaluation. All affected subjects fulfilled clinical criteria for CD. Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants.

For the celiac disease aspect of this meta-analysis, data from a previously published genome-wide scan (768 British cases, 1422 British controls) for CelD was used (Table S1). A more in depth description of this cohorts and its origin can be obtained from the original publications [9,12].

For the replication phase in CelD we used the genotyping results from a second celiac GWAS in three independent CelD cohorts (Table S1). From this study we received data from 3149 cases and 4714 controls from three European populations (UK, the Netherlands and Italy) genotyped on Custom Illumina Human 670-Quad, Hap550 and 1.2 M slides [13]. UK CelD cases were recruited from hospital outpatient clinics (n = 434) and directly through Coeliac UK advertisement (n = 1415) [9]; UK controls were recruited from the 1958 birth cohort and UK National Blood Service for the Welcome Trust Case Control Consortium (WTCCC) (n = 3786). Dutch CelD cases were collected by the UMC Utrecht, Leiden UMC and VUMC Amsterdam from outpatient clinics (n = 803); Dutch controls were recruited through the Sanquin Blood bank by the UMC Utrecht and the VUMC Amsterdam (n = 385). Italian CelD cases (n = 497) and controls (n = 543) were collected by a CelD referral centre (Centro per la prevenzione e diagnosi della malattia celiaca, Fondazione IRCCS Ospedale Maggiore Policlinico) in northern Italy.

All affected individuals were unrelated and were diagnosed according to the revised ESPGAN criteria (1990). The cohorts encompassed individuals that showed a Marsh II or Marsh III lesion in the initial diagnostic small-bowel biopsy specimens, or presented with dermatitis herpetiformis and were *HLA-DQ2* positive. Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants. Some controls were shared between the WTCCC component of the CD meta-analysis and the two UK CelD cohorts (one used in the original scan and one used as replication). This was taken into account as explained in the meta-analysis description.

Imputation of GWAS data

Imputation of the CelD datasets used in the initial and the replication phases of this study were performed with BEAGLE using HapMap phase II and HapMap phase III as reference datasets [21]. A minimum quality score for statistical certainty of the imputation of 0.98 was adhered to.

Imputation of the CD dataset for the initial CD-CelD meta-analysis was performed with the programs MACH and IMPUTE, using HapMap phase II as a reference dataset, as previously described [15,22,23].

Meta-analysis

For the original CD dataset, association tests were described previously [15]. Briefly, results for each SNP from three independent GWAS were summarized as Z-scores and combined in a weighted fashion into a single test statistic; imputation uncertainty was taken into account into Z-score and weight calculation using empirical variance calculated from allele dosage. For the original CelD GWAS scan, best guess imputed genotype frequency data was obtained, and association P-values were calculated using chi-square tests (1 df) of SNP allele counts.

The initial meta-analysis was performed using the statistical program R (http://www.r-project.org/). For both the CD and the CelD dataset the p-values signifying the evidence for association were converted to directional Z-scores, and an overall Z-score and two-tailed p-value for the average of the individuals was subsequently calculated. Given the fact that some controls were shared between one component of the CD meta-analysis and the CelD scan, we expect a correlation of 0.187 between CD and CelD Z-scores. We took this correlation into account in the variance term of the overall Z-score [24].

Unweighed Z-scores were used when combining the data from CD and CelD in the initial meta-analysis, since the CD cohort was

substantially larger than the CelD cohort and weighing would lead to an overrepresentation of the CD signal in the meta-analysis.

A locus was selected for replication when SNPs met the following criteria: a p-value for the locus of $<1\times10^{-5}$ in the meta-analysis, in combination with a p-value of $<1\times10^{-2}$ in the CelD dataset and a p-value of $<1\times10^{-3}$ in the CD dataset. Different thresholds for CD and CelD for inclusion in the replication phase were used in order to reflect the difference in power between the scans for the two phenotypes. For each of the loci that met these criteria, the most strongly associated SNP was analyzed in the CD and CelD replication cohorts.

In order to evaluate the expected number of SNPs that would pass our thresholds (p-value $<1\times10^{-2}$, $<1\times10^{-3}$ and $<1\times10^{-5}$ for CelD, CD and meta-analysis, respectively) by chance, we first evaluated the probability for a particular SNP to reach those thresholds under the null hypothesis of no association and the expected correlation between the two datasets. This probability can be evaluated from the distribution of two correlated normal variables (correlation of 0.186), combined as described for the meta-analysis. We evaluated this probability to be approximately 6.0×10^{-6}. If the 468,378 SNPs tested in the scan were independent, we would then expect less than 3 ($468{,}378*6.0\times10^{-6}$) independent SNPs to be selected by chance. Under a binomial model, we evaluated the probability that 9 or more independent SNPs passes our thresholds to be lower than 0.0025. Those are obviously upper bounds, as we know correlation exists among the SNPs tested.

Replication phase

For replication in CD, SNPs selected for testing were designed into multiplex assays, and genotyped using primer extension chemistry and mass spectrometric analysis (iPlex assay, Sequenom, San Diego, California, USA) on the Sequenom MassArray. This was performed at the Laboratory for Genetics and Genomic Medicine of Inflammation (www.inflammgen.org) of the Université de Montreal. Quality control was performed, excluding samples showing >10% missing data, as well as SNPs with >10% missing data or significantly out of Hardy-Weinberg equilibrium (p-value <0.001). The overall genotyping call rate in the CD replication dataset following quality control analyses was >99%. The CD replication datasets from the two groups (Dutch and Italian) were combined and analyzed using a weighted and directional Z-score approach.

For replication in CelD, genotype frequencies and association data for five replication SNPs were obtained from genome-wide genotyping datasets on Illumina Human 670Quad or 610Quad Genotyping BeadChips (Illumina, Inc., San Diego, CA,). Each GWAS dataset was analyzed using PLINK 1.05 and association p-values were calculated using chi-square tests (1 df) of SNP allele counts [25]. Two of the replication SNPs were not included on the Illumina Human 670Quad or 610Quad Genotyping BeadChips. For these SNPs best guess genotype frequency data was obtained by imputation as described above, and association p-values were calculated using chi-square tests (1 df) of SNP allele counts The CelD replication datasets from the three groups (UK, the Netherlands and Italy) were combined and analyzed using a weighted and directional Z-score approach.

Since selection of specific SNPs for replication was based on their association p-values in the combined CD-CelD dataset, a significant threshold for replication was set at a Bonferroni corrected p-value of 0.05 for 7 independent tests (p-value more significant than 0.0071) in the combined CD-CelD replication dataset. As for the initial meta-analysis, the data from the replication in the CD and CelD cohorts were combined through an unweighed Z-score approach. In addition, for a SNP to be replicated, both effect and direction of association trend needed to match between scan and replication within each disease.

For each SNP showing positive replication, an overall disease-specific association p-value, combining the scan and replication data, was also calculated using a weighted meta-analysis approach. Finally, an overall CD-Celiac meta-analysis of the scan and replication phases of this study was obtained, by combining these within-CD and within-Celiac datasets in an unweighed meta-analysis. Given the fact that some controls were shared between the within-CD and the within-CelD, we expect a correlation of 0.149 between CD and CelD Z-scores. As for the initial scan, we took this correlation into account in the variance term of the overall Z-score as per Lin and colleagues [24].

Supporting Information

Figure S1 Q-Q plot of meta-analysis scan. A meta-analysis scan of 471,504 SNPs from genomewide datasets of CD and CelD was performed using a directional non-weighed Z-score method (as explained in the methods section). The Q-Q plot was generated from the p-values for single-SNP Z scores. Given the strength of association signal for the MHC region (chr6:22,700,000.35,000,000) in CelD, this region overwhelmed the Q-Q plot and was therefore removed from the dataset for purpose of clarity; removing the MHC region had little impact on the observed inflation.
Found at: doi:10.1371/journal.pgen.1001283.s001 (0.06 MB DOC)

Figure S2 Manhattan plot of meta-analysis scan. A meta-analysis scan of 471,504 SNPs from genomewide datasets of CD and CelD was performed using a directional non-weighed Z-score method (as explained in the methods section). The Manhattan plot was generated from the p-values for single-SNP Z scores. Given the strength of association signal for the MHC region (chr6:22,700,000.35,000,000) in CelD, this region overwhelmed the Manhattan plot and was therefore removed from the dataset for purpose of clarity.
Found at: doi:10.1371/journal.pgen.1001283.s002 (0.10 MB DOC)

Table S1 Subjects included in study.
Found at: doi:10.1371/journal.pgen.1001283.s003 (0.14 MB DOC)

Table S2 Complete list of associated markers from meta-analysis.
Found at: doi:10.1371/journal.pgen.1001283.s004 (0.01 MB XLSX)

Acknowledgments

The authors would like to acknowledge the use of the published results from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) meta-analysis of 3,230 CD cases and 4,829 healthy controls taken from three independent CD GWA (results available at http://www.broadinstitute.org/~jcbarret/ibd-meta/). We acknowledge use of DNA from the British 1958 Birth Cohort collection. We thank all clinicians and Coeliac UK for assistance with direct recruitment of celiac disease individuals. We thank C. J. J. Mulder, M. L. Mearin, W. H. M. Verbeek, G. J. Tack, V. M. Wolters, R. H. J. Houwen for collection of Dutch CelD samples. We thank A. Andriulli, F. Bossa, D. Scimeca, and A. Latiano for collection of Italian CD samples. We thank P. Deloukas, R. Gwilliam, and the Wellcome Trust Sanger Institute for UK control genotyping. We thank K. Hunt, V. Mistry, and G. Heap for sample preparation; A. Curtotti and C. Mein for assistance with UK case genotyping. This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We thank all the individuals who participated in the study and J. Senior for critically reading the manuscript.

Author Contributions

Conceived and designed the experiments: EAMF PG TG MJD CW JDR. Performed the experiments: CB PCD CL DAvH. Analyzed the data: EAMF PG TG GB CB GT PCD CL DAvH MJD CW JDR. Contributed reagents/materials/analysis tools: PCFS DWH DB OP VA DAvH RKW CW JDR. Wrote the paper: EAMF PG TG GB CB RKW MJD CW JDR.

References

1. Baumgart DC, Carding SR (2007) Inflammatory bowel disease: cause and immunobiology. Lancet 369: 1627–1640.
2. Baumgart DC, Sandborn WJ (2007) Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet 369: 1641–1657.
3. Hunt KA, van Heel DA (2009) Recent advances in coeliac disease genetics. Gut 58: 473–476.
4. van Heel DA, West J (2006) Recent advances in coeliac disease. Gut 55: 1037–1046.
5. Festen EA, Szperl AM, Weersma RK, Wijmenga C, Wapenaar MC (2009) Inflammatory bowel disease and celiac disease: overlaps in the pathology and genetics, and their potential drug targets. Endocr Metab Immune Disord Drug Targets 9: 199–218.
6. Tursi A, Giorgetti GM, Brandimarte G, Elisei W (2005) High prevalence of celiac disease among patients affected by Crohn's disease. Inflamm Bowel Dis 11: 662–666.
7. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295–302.
8. Festen EA, Goyette P, Scott R, Annese V, Zhernakova A, et al. (2009) Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut 58: 799–804.
9. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40: 395–402.
10. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39: 830–832.
11. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med 359: 2767–2777.
12. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39: 827–829.
13. Zhernakova A, Festen EM, Franke L, Trynka G, van Diemen CC, et al. (2008) Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. Am J Hum Genet 82: 1202–1210.
14. Chang IF, Hsiao HY (2005) Induction of RhoGAP and pathological changes characteristic of Alzheimer's disease by UAHFEMF discharge in rat brain. Curr Alzheimer Res 2: 559–569.
15. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
16. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, et al. (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet 42: 332–337.
17. Festen EA, Stokkers PC, van Diemen CC, van Bodegraven AA, Boezen HM, et al. (2010) Genetic analysis in a Dutch study sample identifies more ulcerative colitis susceptibility loci and shows their additive role in disease risk. Am J Gastroenterol 105: 395–402.
18. Weersma RK, Stokkers PC, van Bodegraven AA, van Hogezand RA, Verspaget HW, et al. (2009) Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. Gut 58: 388–395.
19. De Jager PL, Franchimont D, Waliszewska A, Bitton A, Cohen A, et al. (2007) The role of the Toll receptor pathway in susceptibility to inflammatory bowel diseases. Genes Immun 8: 387–397.
20. Tello-Ruiz MK, Curley C, DelMonte T, Giallourakis C, Kirby A, et al. (2006) Haplotype-based association analysis of 56 functional candidate genes in the IBD6 locus on chromosome 19. Eur J Hum Genet 14: 780–790.
21. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210–223.
22. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913.
23. Li Y, Abecasis GR (2006) MACH 1.0: Rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet S79: 2290.
24. Lin DY, Sullivan PF (2009) Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet 85: 862–872.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.