eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

FUNCTIONAL IMPLICATION OF GWAS CANDIDATE GENES IN CELIAC DISEASE PATHOGENESIS

TESIS DOCTORAL

Leticia Plaza Izurieta

2016

TESIS DOCTORAL

FUNCTIONAL IMPLICATION OF GWAS CANDIDATE GENES IN CELIAC DISEASE PATHOGENESIS

Leticia Plaza Izurieta

2016

Directores:

Jose Ramón Bilbao Catalá

Juan Carlos Vitoria Cormenzana

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

This work was funded by a Predoctoral Fellowship from the Basque Department of Education, University and Research to Leticia Plaza (RBFI-2012-450) and Research Project Grants from the Spanish Ministry of Science and Innovation (07/0796 and 10/0310) and Basque Departments of Health (2006/111030 and 2011/111034) and Industry (SAIO-2008/00231).

Abbreviations

AIDS	acquired immune deficiency syndrome
APC	antigen presenting cells
CD	Celiac disease
CEGEC	Spanish Consortium for Genetics of Celiac Disease
EGFR	epidermal growth factor receptor
EMA	anti-endomysium autoantibodies
EpCAM	Epithelial cell adhesion molecule
ESPGHAN	European Society for Pediatric Gastroenterology, Hepatology and Nutrition
FBS	Fetal bovine serum
GFD	gluten free diet
GWAS	Genome Wide Association Studies
HLA	Human Leucocyte Antigen
IELs	Intraepithelial lymphocytes
KIR	Killer Immunoglobulin-like receptor
LD	Linkage disequilibrium
MHC	Mayor Histocompatibility Complex
NFkB	Nuclear kappa B transcription factor
NK	Natural killer
OR	Odds ratio
PCR	Polymerase chain reaction
RA	rheumatoid arthritis
RPLPO	large ribosomal protein
RT-PCR	Real-time quantitative reverse transcription PCR

SD	standard deviation
SNP	Single Nucleotide Polymorphism
T1D	type 1 diabetes
TF	transcription factor
TG2	transglutaminase
TGA	anti-tissue transglutaminase autoantibodies
TLR	Toll-like receptor
WGA	Whole genome amplification

List of original publications

Chapter I

Plaza-Izurieta L, Castellanos-Rubio A, Irastorza I, Fernandez-Jimenez N, Gutierrez G, CEGEC, Bilbao JR. Revisiting genome wide associaton studies in celiac disease: replication study in Spanish population and expression analysis of candidate genes. Journal of medical genetics. 2011, 48:493-496.

Chapter II

Plaza-Izurieta L, Fernandez-Jimenez N, Irastorza I, Jauregui-Miguel A, Romero-Garmendia I, Vitoria JC, Bilbao JR. Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease. European Journal of Human Genetics. 2015, 23(8):1100-5.

Chapter III

Plaza-Izurieta L, Jauregui-Miguel A., Romero-Garmendia I, Garcia-Etxebarria K., Legarda M., Irastorza I., Bilbao JR. Immunochip candidate genes study in CD intestinal cell populations. (in preparation).

Project justification and scope

Celiac disease (CD) is a chronic immune mediated disorder with a high prevalence. It is believed that prevention will be crucial for the eradication of this disorder, and for that purpose, efficient mechanisms of prediction and early diagnosis need to be developed. In a temporal scale, the presence of clinical symptoms can be considered an advance stage of the disease-progression process. This active disease stage would be preceded by the presence of immunological markers, such as circulating autoantibodies against tTG (tissue transglutaminase), reflecting an ongoing immune mediated tissue-destruction process that initiates only among genetically predisposed individuals.

Therefore, it becomes essential to define which genes are involved in disease susceptibility, in order to understand the pathogenic mechanisms underlying CD development and also to provide genetic markers capable of discriminating individuals at risk of this disease, which would allow predictive diagnosis to be performed prior to the activation of the autoimmune response and would improve the selection of candidates for putative immune prevention trials.

In order to dissect the genetics of this complex autoimmune disease, the current project has focused on the search of functional genetic determinants in celiac disease using the Genome wide association studies performed in celiac disease and the candidate genes proposed form those results as starting point.

Introduction

1. Celiac disease

Gluten sensitive enteropathy (MIM 212750) or celiac disease (CD) is a chronic, immune-mediated inflammatory disorder characterized by flattened villi on the small bowel mucosa, caused by intolerance to ingested gluten and related proteins in wheat, rye and barley that develops in genetically susceptible individuals.

1.1. Clinical features and diagnosis

The adverse effects of ingested gluten were not recognized until 1950 [1], although the clinical picture of CD had been first described by Samuel Gee more than 60 years before [2]. If untreated, classical CD presents with a range of symptoms and signs that can be divided into intestinal features, such as diarrhea, abdominal distension or vomiting, and those caused by malabsorption, like failure to thrive (low weight, lack of fat, hair thinning) or psychomotor impairment (muscle wasting) [3]. Other atypical symptoms are also associated with CD, and include neurological events, dental enamel defects, infertility, osteoporosis, joint symptoms and elevated liver-enzyme concentrations [4]. From a histological point of view, when a susceptible person is on a gluten-containing diet, there are gradual changes in the small intestinal mucosa that result in a lesion with villous atrophy and crypt hyperplasia (Figure 1).

The degree and severity of gluten-induced mucosal alterations are described in the Marsh-Oberhuber classification [5]. At first, there is an infiltration by intraepithelial lymphocytes (IELs) of the villous epithelium (Marsh I), which is followed by hypertrophic crypts (Marsh II), while the villi are not shortened. In the more advanced stage (Marsh III), crypts are hypertrophic, the *lamina propria* is swollen, and there is either severe partial, subtotal or total villous atrophy (Figure 1). Together with the damage of the small intestinal mucosa, CD is characterized by the presence of different gluten-dependent serum autoantibodies, such as anti-endomysium (EMA) or anti-tissue transglutaminase (TGA) antibodies among others [6].

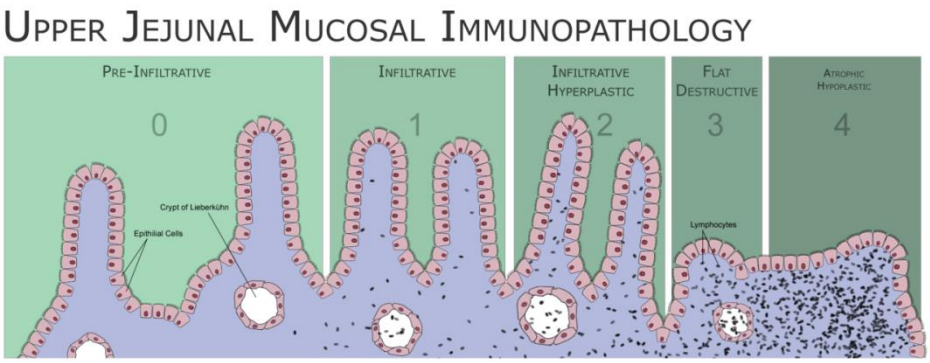


Figure 1: Gluten-induced mucosal changes in different stages according to Marsh classification. Image from www.theglutensyndrome.net.

Taking into account these pathological features, the diagnostic criteria for CD have been established by the European Society for Pediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) and until recently, have been based on the presence of characteristic histological injuries in a biopsy of small intestine and by positive serologic results, although the latter were not essential [7]. In contrast, in the diagnostic guidelines published in 2012, duodenal biopsy can be excluded in symptomatic children with IgA class TGA titers above 10 times the upper limit of normal levels. Apart from that, Human Leucocyte Antigen (HLA) genotyping is helpful since CD is very unlikely if risk haplotypes are absent [8].

1.2. **Epidemiology**

Until the end of the last century, CD was considered a comparatively uncommon disorder, with prevalence rates of 1/1000 in Europe [3]. However, more recent population studies have shown that the prevalence of CD is around 1% in Western Europe, although there are differences among populations [9]. Different investigations have suggested that the incidence of childhood CD may have been rising during 1980s and 1990s, and this has been related to infant feeding practices [10]. On the other hand, the diagnosis of adult CD has also risen dramatically in most areas of the world where there are data available [11-13]. Environmental risk factors with seasonal patterns, including certain viral infections have been proposed as risk factors for CD [14].

Typically, CD has been regarded as a disorder affecting almost exclusively people of European origin, but the increased reliability of serological test has improved the accuracy of estimates of CD prevalence demonstrating a frequency ranging from 1:100 to 1:200 in unselected populations of North America and Australia [15,16]. CD was also believed to be rare in Latin America [17-19], North Africa [20,21] and the Middle East [22,23] where there were only limited cases and occasional observations of CD. Additionally, CD has been historically considered absent in the Far East (China, Japan, Korea…) [24]. However, recent screening studies performed in these areas have demonstrated that the prevalence of CD has been underestimated and that it is, instead, similar to that of the so-called Western countries. With the spread of the modern Western diet, including gluten-containing cereals (specially wheat) to all parts of the world, CD has become a global Public Health problem, and also affects the populations of developing countries [16].

In terms of gender, females are more commonly affected than males, and among patients presenting the disease during their fertile years, a female to male ratio of almost 3 to 1 has been observed [25].

Until now, the only proven treatment for CD is a strict and life-long removal of gluten from the diet, which is achieved by the elimination of wheat, barley, and rye cereal products [3,26]. However, complying with gluten free diet (GFD) is difficult and it is thought to decrease quality of life. Moreover, inadequately treated and untreated patients are predisposed to complications such as short stature, nutritional deficiencies, osteoporosis, secondary autoimmune disorders, malignancies, infertility and poor outcome of pregnancies [27].

2. **Pathogenesis of celiac disease**

The recent advances in our knowledge on the mechanisms that take part in the development of celiac disease have made it one of the best-understood HLA-linked disorders. However, several pathogenic processes still remain to be described.

It has been known for some time that CD is a T cell mediated disease: gluten peptides cross the epithelium into the *lamina propria* and are deamidated by tissue transglutaminase to be presented by DQ2+ and/or DQ8+ antigen presenting cells (APCs) to pathogenic CD4+ T cells. This triggers a Th1-mediated response that leads to the

infiltration of the epithelial *lamina propria* by inflammatory cells, together with crypt hyperplasia, and villous atrophy [28]. However, studies in the last decade have also stressed the role of the innate immune response in the pathogenesis of the disease, and it has been shown that gliadin can also activate a non-T cell mediated response [29,30] (Figure 2).

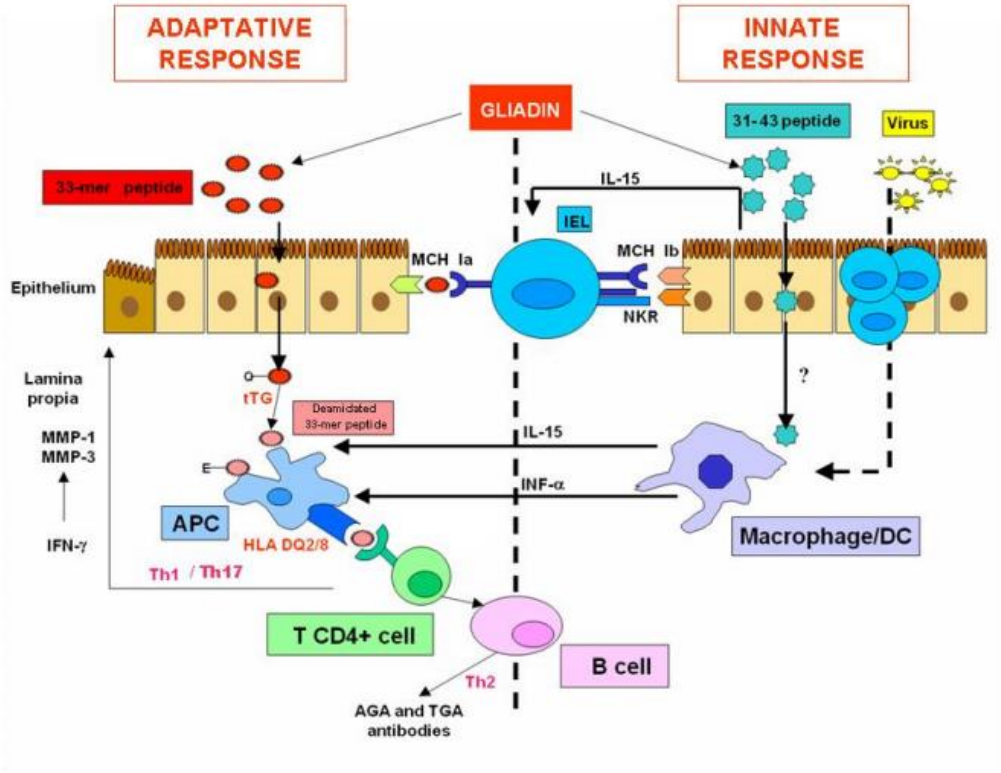


Figure 2: Pathogenic mechanisms of celiac disease: adaptive and innate immune branches. Adapted from Castellanos-Rubio et al., 2010 [31].

2.1. **Gluten**

Gluten is a mixture of monomeric, alcohol soluble glutenins, and polymeric, prolamin rich gliadins that is found in the endosperm of cereals like wheat, barley and rye. CD is triggered by the exposure to gluten proteins in the diet. Gluten proteins are very resistant to degradation by intestinal proteases so that long fragments (10-50 residues) are present

in the gut lumen. These fragments are good substrates for the enzyme tissue transglutaminase type 2 (TG2), which can deamidate gluten peptides increasing their ability to bind to HLA-DQ2 or HLA-DQ8 molecules leading to a gluten-specific CD4+ Th1 T-cell response. In addition, gluten can trigger CD8+ T cell responses in the *lamina propria* and may expand the intraepithelial lymphocyte population independently of Mayor Histocompatibility Complex (MHC) presentation [28,32].

Different in vitro studies of intestinal organ culture, primary antigen presenting cells, and epithelial and monocytic cell lines support this idea. A peptide spanning aminoacids 31-43 of the alpha gliadin molecule (31-43 peptide) has been thoroughly studied for its innate immune stimulatory properties.

This peptide is not bound by HLA-DQ2 and does not induce T cell-specific responses in the gut. Within a few hours of challenge, it is able to induce apoptosis or inhibit the epidermal growth factor receptor (EGFR) endocytic pathway, far before the mechanism of gluten peptide presentation to CD4+T lymphocytes [32,33]. In support of this idea, it has been observed that intestinal biopsies of CD patients incubated with gliadin show an upregulation of *MICA*, a stress-induced molecule that interacts with the *NKG2D* receptor expressed on Tγδ and NK cells, and is capable of activating innate cytotoxic and cytokine production responses in the initial stages of the disease, linking innate with adaptive immunity [34].

2.2. **Transglutaminase**

In CD patients, gluten induces the secretion of IgA-class autoantibodies against TG2. This enzyme is a ubiquitously expressed multifunctional protein which is usually active in the extracellular space and catalyzes the covalent and irreversible cross-linking of a protein with a glutamine residue to a second protein with a lysine residue [35,36]. Gluten is rich in prolines and glutamines, and has very few negative residues (necessary to bind to the groove of HLA-DQ2 or -DQ8) so that gluten derived peptides must be first deamidated by TG2 [37]. TG2 deamidation of specific gliadin peptides transforms gliadin peptides from a non-stimulatory molecule into an efficient T-cell antigen capable of evoking a massive secretion of local cytokines, and lead to alterations in enterocyte differentiation and proliferation. Additionally, TG2 mediated crosslinking between gliadin peptides and the enzyme leads to the formation of TG2-gliadin complexes that trigger the production of autoantibodies [38]. A unique 33-mer peptide harboring six partly overlapping copies of

three T-cell epitopes is the most potent T cell stimulator after its deamidation by tissue transglutaminase.

Whether anti-transglutaminase antibodies participate in the pathogenesis of the typical mucosal lesion of the disease, or only represent a bystander event in CD is still unclear. Biological effects of CD autoantibodies on cell cycle, apoptosis, angiogenesis and intestinal permeability have been reported, suggesting that TG2 antibodies could be pathologically relevant [38,39].

2.3. **Adaptive immunity**

Adaptive immunity includes T cell-mediated and humoral immunity, and both of them are activated in the small intestinal mucosa of CD patients, with gliadin as the recognized antigen. CD4+ T lymphocytes from the small intestinal mucosa recognize deamidated gliadin peptides bound to HLA-DQ2 and HLA-DQ8 heterodimers on APCs [40,41]. Gliadin-specific T lymphocytes from celiac mucosa are mainly of the Th1 phenotype and release prevalently proinflammatory cytokines, dominated by IFN-γ [42,43]. In addition to IFN-γ, other Th1-inducing cytokines such as interleukin 18 and IFN-α are also increased [44-46]. A different linage of CD4+ T-helper cells that differentiate in the presence of IL6 and TGFβ and produce interleukin 17 cytokine-family members (Th17 lymphocytes) has been identified and seem to be responsible for pathogenic effects previously attributed to the IL12/INFγ network [47]. Both Th1 and Th17 responses are present in the active CD lesion, a phenomenon that has also been described in other immune-mediated conditions [48-50].

2.4. **Innate immunity**

The innate immune response represents the first line of defense against pathogens, and is activated during the first stages of exposure to an infectious agent. In CD, the innate immune responds to gliadin in a Tαβ-lymphocyte independent manner and contributes to the creation of the proinflammatory environment necessary for subsequent T cell activation in patients carrying HLA-DQ2 or DQ8. Several in vivo challenge studies have demonstrated that peptide 31-43 from α-gliadin is capable of inducing disease symptoms, and several CD characteristic changes have been observed in biopsy cultures [51-53]. This peptide does not appear to stimulate a T cell-mediated response [54,55], so it is

likely that the toxicity of peptide 31-43 is based on its capacity of activating the innate immune response.

Several studies have implicated *MyD88*, the major signal transducer of Toll-like receptor 4 (*TLR4*) on monocytes, macrophages and dendritic cells, and TLR4 itself as the primary receptor for innate responses to cereal proteins [28]. Innate immune activation of IELs by gluten induces expression of *MICA* on the intestinal epithelium, which serves as ligand for the NKG2D receptor on natural killer, γδT cells and on subsets of CD4+ and CD8+ T cells. Epithelial MICA together with upregulated IL15 leads to the activation of *NKG2D* on IELs triggering antigen-specific lymphocyte-mediated cytotoxicity. Finally, IL21 has emerged as an additional driving force of innate immunity that often acts in concert with IL15 [30,56].

3. **Genetics of celiac disease**

Even though the inheritance model of CD is still unknown, it has been known for a long time that Genetics participates in the susceptibility to the disease. Studies on the prevalence of CD in affected families, and especially those comparing twin pairs, have been very useful to estimate the proportion in which both environmental and genetic factors contribute to the development of this disorder. According to these studies, genetics is a fundamental player both in the triggering and in the latter development of CD.

In general, it is well accepted that the proportion of monozygotic or identical twins concordant for CD is around 75-86%, while in the case of dizygotic twins, this proportion is reduced to 16-20%. This difference between mono- and di-zygotic twins has allowed scientists to estimate the genetic component of CD, which is higher than what has been described for other immunological complex diseases, such as type 1 diabetes (T1D) (around 30% concordance in monozygotic and 6% in dizygotic twins) [57]. Additionally, concordance rates between sibling pairs and dizygotic twins are almost the same, indicating that the environmental component has a minimum contribution to the risk of developing CD.

In summary, accumulated evidence suggests that CD has a very strong genetic component and it has been calculated that the heritability of this disease (proportion of the risk of suffering from CD attributable to genetic factors, compared to environmental determinants) is around 87% [58]. The largest portion of the genetic risk to develop CD

comes from the presence of certain HLA alleles. However, even if the role of these HLA molecules is essential in the pathogenesis of the disease, their contribution to the heredity is modest. In a recent publication from Gutierrez-Achury *et al*. it has been calculated that the classical *HLA-DQA1* and *HLA-DQB1 loci* alone explain 23% of the CD heritability risk, whereas a newly discovered 5 HLA novel variants reported an additional 18% of genetic variance [59], which in total explain approximately 40% of CD risk, and thus, it has been hypothesized on the existence of many small effect, non-HLA susceptibility *loci*.

3.1. HLA region and celiac disease

3.1.1. HLA region

HLA is the name for the MHC in humans; it is a super *locus* located on the chromosomal region 6p21 and contains a large number of genes related to the immune response. HLA genes encode antigen presenting proteins that are expressed in most human cells and are essential for the ability of the organism in distinguishing between self and foreign molecules.

HLA genes are involved in many inflammatory and autoimmune disorders and also contribute to the susceptibility to develop infectious diseases such as acquired immune deficiency syndrome (AIDS) or malaria. However, due to the high genetic complexity of the region, most of the particular genetic factors and pathogenic mechanisms underlying the susceptibility to each of these disorders remain unknown. In fact, the HLA region presents the highest genic density of the entire genome and a very strong gene expression seems to be favored [60].

3.1.2. Contribution to the genetic risk and susceptibility genes

As previously pointed out, the HLA region is the most important susceptibility *locus* in CD and explains around 40% (23% classic and 18% novel variants) of the genetic component of the disease. The first evidences supporting the association between HLA and CD where published in 1973 and were detected using serological methods [61]. Due to the strong linkage disequilibrium present in the area, initial studies identified HLA-A1, HLA-B8 and HLA-DR3 as the etiological variants in the region, but subsequent molecular studies have revealed that the factors directly implicated are the HLA class II genes encoding both HLA-DQ2 and -DQ8 molecules (Figure 3).

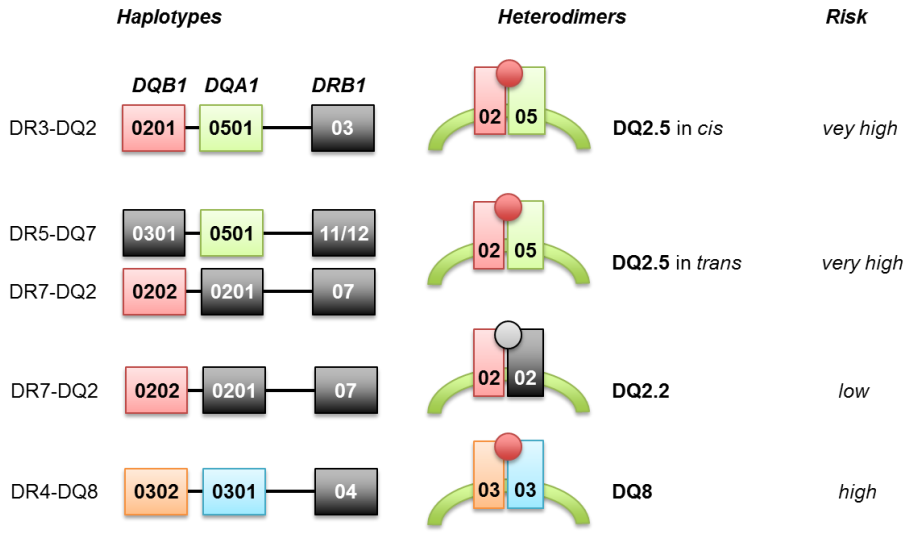


Figure 3: Association of the HLA locus with CD. HLA-DQ2 molecule is the major factor conferring risk to CD. Most celiac patients express the heterodimer HLA-DQ2.5, encoded by the alleles HLA-DQA1*05 (α chain) and HLA-DQB1*02 (β chain), that can be present in cis in the DR3-DQ2 haplotype or in trans, in the heterozygotes DR5-DQ7 and DR7-DQ2.2. The HLA-DQ2.2 dimer, a variant of HLA-DQ2 encoded by the alleles HLA-DQA1*02:01 and HLA-DQB1*02:02, confer a low risk to develop the disease. Most of the patients that are negative for DQ2 express HLA-DQ8, encoded by the DR4-DQ8 haplotype. (Adapted from Abadie et al., Annu Rev Immunol (2011) [62])

The strongest association has been found with HLA-DQ2, and 90% of celiac patients present at least one copy of the HLA-DQ2.5 heterodimer (formed by the combination of the products of DQA1*05 and DQB1*02 alleles, that encode the α and β chains of the heterodimer, respectively). On the other hand, 20-30% of the non-celiac population also presents this HLA-DQ2 variant, making it clear that, even though it is very important, it is not sufficient to develop the disease. Most of the patients who do not carry the HLA-DQ2 genotype are HLA-DQ8 carriers and so have at least one copy of the haplotype containing DQA1*03:01 and DQB1*03:02 alleles [4]. A very small portion of the patients are negative for both DQ2 and DQ8, but it has been observed that in these few cases, individuals present at least one of the two alleles encoding the DQ2 molecule (DQA1*05 or DQB1*02) [63,64].

HLA-DQ2 and -DQ8 variants are in linkage disequilibrium with DR3 and DR4, respectively. Thus, we often refer to these risk variants as DR3-DQ2 and DR4-DQ8 haplotypes [65]. In

several haplotypes, as is the case of DR3-DQ2, the two alleles of the HLA-DQ2.5 heterodimer (DQA1*05:01 and DQB1*02:01) are located in the same chromosome and therefore, encoded in *cis*. In the heterozygous individuals carrying DR5-DQ7 and DR7-DQ2 haplotypes, the two molecules taking part in the risk heterodimer are encoded in *trans* because they are located in different chromosomes. The differences between these two types of HLA-DQ2.5 rely on a single amino acid of the DQ α chain (DQA1*05:01 vs. DQA1*05:05) and another residue of the membrane region of the DQ β chain (DQB1*02:01 versus DQB1*02:02), but they seem not to have any functional consequences and they are associated with a similar risk effect. However, the risk conferred by another HLA-DQ2 variant, the HLA-DQ2.2 dimer, is very low [57,66].

There is also a relationship between the degree of susceptibility to CD and the number of DQ2.5 heterodimers. Homozygous individuals with two DR3-DQ2 haplotypes as well as the heterozygous patients presenting DR3-DQ2/DR7-DQ2 express the highest levels of DQ2.5 heterodimers and thus, confer the maximum genetic risk to develop CD [67-69]. In this sense, it has to be mentioned that patients with refractory CD (those not responding to GFD) present a higher degree of homozygosity for DR3-DQ2 (44-62%) than other celiac patients (20-24%). A similar dose-dependent effect has also been suggested for DQ8 molecules.

Apart from the genes encoding DQ molecules, the HLA region also contains many other genes that participate to the immune response and that could contribute to the susceptibility to CD. Several studies have postulated that polymorphisms in genes such as *MICA*, *MICB* or *TNF* could contribute to the genetic risk to develop this disorder. Nonetheless, most of these works have not paid enough attention to the strong linkage disequilibrium among genes and results are not conclusive. Although HLA genes importantly contribute to the genetic susceptibility, the concordance of the disease in siblings identical for HLA genotype approaches only 30%, so that we can conclude that HLA genes are important but not sufficient to develop CD [66].

3.1.3. **Role of HLA in the pathogenesis of CD**

The strong association of the HLA class II genes with CD is directly linked to the fundamental role of CD4+ T lymphocytes in the pathogenesis of the disease. In fact, CD4+ T cells that are able to recognize gluten-derived peptides are present in the intestinal mucosa of celiac patients, but not in the case of healthy, non-celiac individuals.

When genetically susceptible individuals are exposed to certain gluten-derived epitopes, they are presented by the HLA-DQ2/HLA-DQ8 molecules on the surface of APC, stimulating the proliferation of gluten-specific CD4+ T cells [28].

An important landmark in the molecular basis underlying the association between HLA and CD was the discovery that the binding capacity between the HLA-DQ2 and/or -DQ8 and the gliadin peptides increases substantially when the latter have been enzymatically modified by the enzyme TG2. As mentioned before, the enzyme catalyzes a reaction that provokes the increase of negative charges in the gluten-derived peptides, favoring their binding to certain HLA molecules (DQ2 and DQ8) and thus, triggering the presentation of these gluten peptides to CD4+ T cells.

Given the importance of HLA molecules in the activation of auto-reactive gluten-specific T cells, it is expected that any modification in their coding sequence will provoke alterations in different steps of this process. In this way, polymorphisms in the sequence encoding the antigen binding sites could affect affinity, favoring or hampering the recognition of the gluten-derived peptides [70]. On the other hand, several polymorphisms located in regulatory sites can repress or enhance the expression of the HLA molecules, reducing or augmenting the immune response to gluten.

3.2. **Genome-wide association studies in CD**

From 2006, when first GWA study was published [71], the way to approach genetic studies of complex traits and diseases has changed. GWA studies have evolved over the last ten years into a powerful tool that enable researchers to scan a great number of genetic markers in large genomic DNA sample sets, with the aim of finding genetic variants associated with a particular disease. These studies are especially useful when we try to find susceptibility variants contributing to the genetic background of complex diseases that are common in the population, as is the case of CD.

The unit of genetic variation in Genome Wide Association studies (GWAS) is the Single Nucleotide Polymorphism (SNP). The large majority of them have a minimal impact on biological systems, but SNPs can also have functional consequences such as amino acid changes, changes to mRNA transcript stability, and changes to transcription factor binding affinity [72].

To perform a GWAS, scientists use two types of participants: individuals affected by the studied disease (cases) and individuals with similar characteristics of those belonging to the first group (sex, age…) but not suffering from the disorder (controls). The optimal selection of both case and control samples is crucial in the GWAS design. For case selection it is important to minimize phenotypic heterogeneity and there is a growing focus on the application of GWA methodologies to population-based cohorts, although most published GWA studies have featured case-control designs [73]. Moreover, the presence of individuals with different ancestral and demographic backgrounds could cause population stratification. If cases and controls differ to this respect, markers that are informative for the ethnic origin of the sample might be confounded with disease status, leading to spurious association. Selection bias is also a common mistake particularly in controls, when the sample set might not be representative of the wider population that is purported to represent [73]. Sample size is also one of the most important issue in GWA studies, and the conclusion is clear: the more samples the better [74].

Associated variants, SNPs that have higher frequency in cases than in controls, indicate genomic regions in which disease causing variants could reside. The associated SNPs are not always the causal variants and it is thought that they are merely pointing to associated regions that should be more deeply scanned. This is the reason why it is imperative to keep on with the investigation, deep-sequencing the associated region with the aim of identifying the exact genetic change involved or performing functional analyses and trying to correlate specific variants or alleles with expression levels.

Thus, GWAS constitute a method that allows capturing a new type of genetic variation. Family-based association studies that use pedigrees usually identify rare variants with large phenotypic effects, while these genome-wide approaches rely on population-based sample sets and therefore, are used to find more common variation with modest effects (Figure 4).

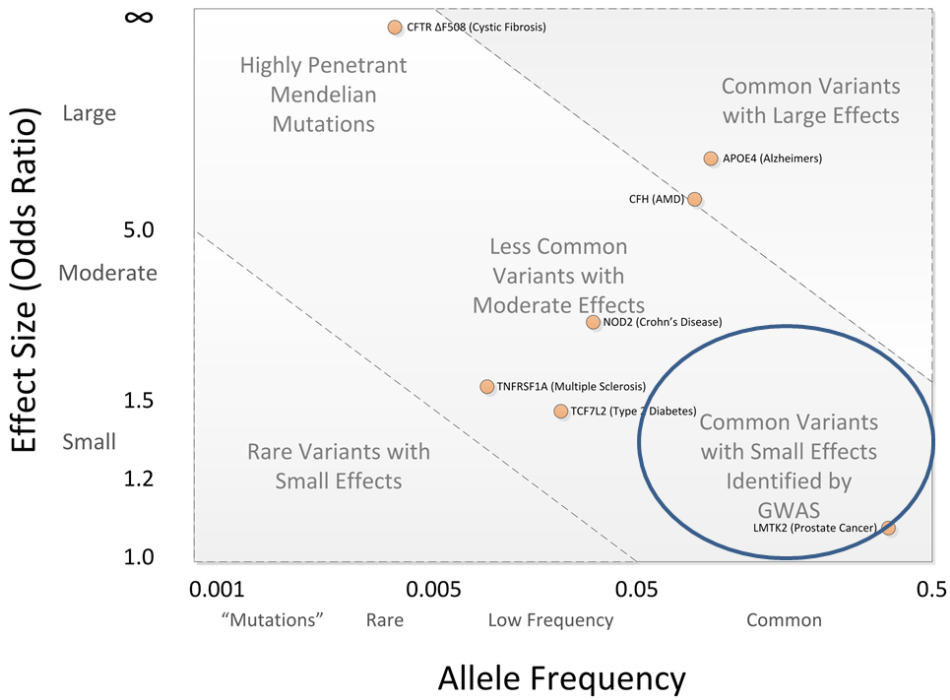


Figure 4: Spectrum of Disease Allele Effects. Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines. (Adapted from Bush et al. (2012) doi:10.1371/journal.pcbi.1002822.g001 [75]).

3.2.1. **Results of the first GWAS**

In the first GWAS performed in CD, 778 affected and 1,422 healthy control individuals were studied. Association analyses were performed for up to 310,605 SNPs that showed population-scale frequencies higher than 1% for their minor alleles [71]. The strongest association was obviously found in the HLA *locus* and rs2187668-A allele was identified as an efficient marker for the HLA-DQ2.5 *cis* encoding HLA-DR3-DQ2 haplotype. As already mentioned, this is the most common HLA-DQ2 haplotype associated with CD. In this first GWAS it was shown that in the case of the patients coming from the United Kingdom, at least one copy of HLA-DQ2.5 *cis* haplotype was present in 89.2% of patients, while it was found only in the 25.5% of the control group.

Outside the HLA region, 56 associated SNPs were found, with $p<10^{-4}$. Some of these SNPs were close to each other, suggesting that they could be localized in *loci* showing real associations with the disease and that these polymorphisms could be in linkage disequilibrium (LD) with the causal variants contributing to the complex genetic component of CD.

However, the only SNP significantly associated with the disease ($p<10^{-6}$) was rs13119723, in the 4q27 region, in which an LD block containing *IL2* and *IL21* is present. These results were replicated in Dutch and Irish patients and controls. Besides, it was estimated that this region might only explain around 1% of the genetic risk to develop CD, suggesting the existence of other susceptibility genes which had not yet been identified. An additional set of 1.643 cases and 3.406 controls from three different independent European cohorts was analyzed for the 1.164 most associated SNPs[76]. The associated regions following this replication study were investigated in order to find candidate genes that could be functionally implicated in the development of CD, paying especial attention to those genes taking part in the immune response (Figure 5).

3.2.2. **Results of the second GWAS**

The second GWAS in CD was performed in 2009. Up to 292.387 non-HLA SNPs were analyzed in 4.533 celiac individuals and in 10.750 healthy controls, all from European origin. Moreover, 231.362 additional SNPs outside the HLA region were studied in an independent cohort of 3.796 affected patients and 8.154 controls [77].

Thirteen previously unknown risk regions were found with significant evidences of association (Figure 5). In these regions there are several genes with immune functions: *BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLEC16A*, *ETS1*, *ICOSLG*, *RUNX3*, *THEMIS*, *TNFRSF14* and *ZMIZ1*. Other 13 *loci* did not reach the genome-wide statistical significance but seemed to be somehow related to disease and contained genes with implications in the immune system, including *CD247*, *FASLG-TNFSF18-TNFSF4*, *IRF4*, *TLR7-TLR8*, *TNFRSF9* and *YDJC*.

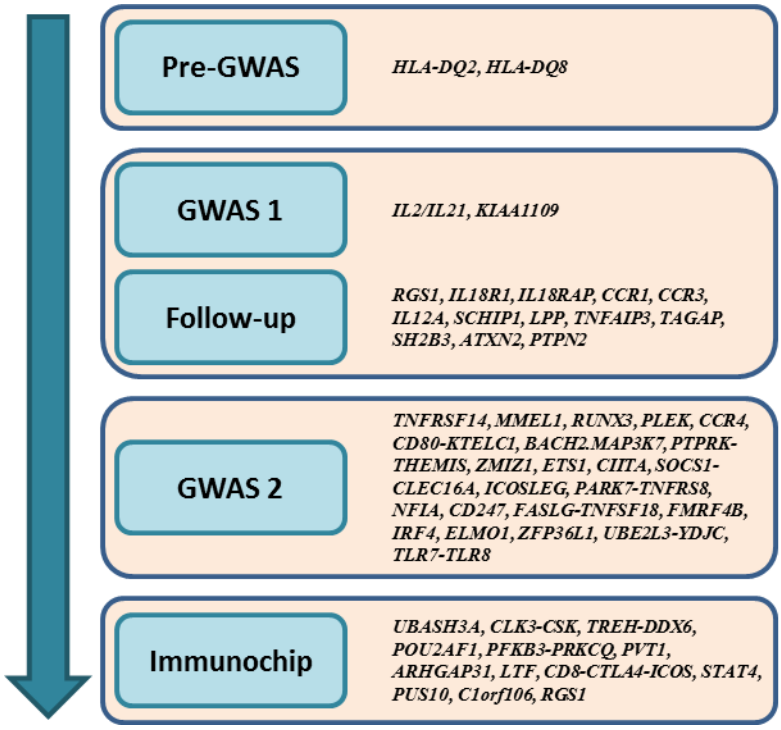


Figure 5: Advances in the Genetics of CD. After the Immunochip study, 40 loci have been found to contribute to the genetic risk to develop the disease. It remains elusive which are the causal variants underlying these associations, so that functional studies will be needed in order to determine which the possible applications of these large association studies will be (adapted from Kumar V. et al, DOI 10.1007/s00281-012-0312-1 [78]).

3.2.3. **The Immunochip**

The last large-scale project aiming to identify a larger portion of associated variants was the so called Immunochip, in which, among other thousands of samples coming from patients suffering from different immune-mediated disorders, more than 12,000 celiac individuals and 12,000 controls were analyzed for more than 200,000 variants [79].

In this huge study, 183 immune-related non-*HLA loci* were interrogated. Thirty nine regions showed a significant association with CD (the 26 *loci* identified with the two GWA studies plus 13 more) (Figure 5). All the associated variants were common and had a minor allele frequency higher than 5%. Association of low frequency variants was only

found for 4 regions. In this work, fine-mapping of the regions of interest was performed, allowing scientists to better recognize the etiological variants.

After the functional annotation of the associated SNPs, one of the main conclusions was that the genetic variants lying in coding regions are very few, although some of them are located close to the transcription start sites of several genes or on their 3'UTR regions. Some of the proposed candidate genes, given the fact that they have significant signals close to their regulatory regions, were the following: *THEMIS/PTPRK*, *TAGAP*, *ETS1*, *RUNX3* and *RGS1*. Several of these had been already proposed after the GWA studies.

3.2.4. **GWAS replication studies**

Association results obtained in genome wide studies must be replicated to determine whether the genetic makeup of CD is robust and comparable across different populations. CD associated regions have been analyzed in many independent populations, with some contradictory results.

The first replication study of the CD GWAS results was carried out in 2009 in an Italian population, composed by 538 CD patients and 593 healthy controls. From the eight *loci* identified in the first GWAS and it follow-up, in the studied Italian population 4 *loci* showed susceptibility to CD (*IL2/IL21, RGS1, IL12A/SCHIP1, SH2B3*), 2 regions showed moderate association (*LPP, TAGAP*), whereas there were 2 more regions with no association (*CCR3, IL18RAP*)[80].

Additionally, in the same year a Spanish population was tested for association in the 2 *loci* that could not be replicated in the Italian cohort, rs917997 (2q12) and rs6441961 (3p21). In a case-control study with 722 CD patients and 794 non-CD controls, only the association of the 3p21 genetic region with CD susceptibility was confirmed[81].

With a similar approach, the top 1020 non-HLA SNPs from the GWAS were genotyped and analyzed for association in 906 CD patients and 3819 controls from a US population. Five of eight regions identified in the GWAS follow-up study were strongly associated with CD, including regions on 1q31, 3q25, 3q28, 4q27 and 12q24. The strongest association was found in region 4q27 (*IL2/IL21*), which was also the most associated region in the GWAS and the follow up study. Second most associated signal in this study was located at 3q28, harboring *LPP*. In addition, this study provided new evidence for

association not previously reported, located on 2q31 harboring an already candidate gene *ITGA4* [82].

In 2010, a Swedish/Norwegian population in whom association with *IL2/IL21* region was earlier replicated, was analyzed aiming to replicate the remaining regions in a family cohort using transmission disequilibrium test, which is not prone to population stratification as a source of false-positive results[83]. 325 Swedish/Norwegian CD families were genotyped for 9 associated SNPs, from which 5 SNPs (rs2816316-1q31, rs6441961-3p21, rs17810564/rs9811792-3q25-26, rs1464510-3q128) showed significant association, SNP (rs917997-2q11-12) showed borderline but not significant association, and no evidence of association was found in the remaining 2 SNPs (rs13015714-2q11-12, rs1738074-6q25). rs3184504 in 12q24 region (*SH2B3*) was not analyzed because of assay failure.

All those contradictory studies, stresses the importance of analyzing large samples to obtain robust results that can be replicated. However, as is the case of the majority of studies, it is also evident that the strength of these signals is relatively small, and the ability to detect significant association often depends on minute allele frequency differences across populations, which could account, at least in part, for some of the negative results.

3.2.5. **Functional follow up of the association studies**

Another major objective of the genetic association studies is to identify the functional culprits that are implicated in the development of the disease, in order to increase our understanding of its pathogenesis. To investigate their putative role in the disease process, the expression of candidate genes proposed following GWAS is a good approach.

In a work performed by our group, we observed that two genes (*PTPRK* and *THEMIS*), situated in the same associated *locus*, were coexpressed both in the active disease and after in vitro stimulation with gliadin of celiac GFD-treated biopsies. The work concluded that the associated SNPs in this region could affect the expression levels of the surrounding genes, but instead of influencing them in a constitutive way, we speculated that the associated genetic variants might underlie the regulation of both *PTPRK* and *THEMIS* in response to the immunogenic insult, that is, the exposure to dietary gluten.

These findings point to common regulatory mechanisms encoded in the DNA sequence that might control the expression patterns of different genes, which can be switched on only under the immunogenic stimulus [84].

On the other hand, in order to elucidate the substantial fraction of heritability that remains unexplained in most complex diseases, a novel hypothesis has recently been postulated. It has been called the "rare-variant synthetic genome-wide-association hypothesis" and it is based on the assumption that unobserved rare causal variants lead to association detected at common tag variants. However, a recent work in which sequencing and genotyping for coding exons of 25 GWAS risk genes were performed in 41,911 UK residents of white European origin (24,892 subjects with six autoimmune disease phenotypes and 17,019 controls) has revealed that rare coding-region variants at known *loci* have a negligible role in common autoimmune disease susceptibility, including CD [85].

In conclusion, these works and other similar studies stress the need of developing functional studies and the importance of avoiding arbitrary selection of susceptibility candidate genes. Additionally, they reveal the huge work that remains to be done in order to identify the elements underlying the complex regulatory system of the genome, while opening the door to future studies, in which the scientific community will need to exhaustively analyze both different classes of variation (such as structural variants of the genome or epigenetic features) and the vast noncoding genome, in order to shed light on the complex genetics of common disorders and to be able to understand the effect of the disease-associated variants found by the numerous GWA studies.

Aims

The present work has two main objectives that aim to contribute to decipher celiac disease pathogenesis:

I. To analyze the CD associated genetic variants proposed in Genome Wide Association Studies in the Spanish population.

   a. To replicate the association results from GWAS in the Spanish population.
   b. To look for novel associated signals and regions in the Spanish population not previously described.

II. To question the implication of the proposed candidate genes in disease development.

   a. To analyze the expression of the CD candidate genes in the disease tissue of celiac patients and controls.
   b. To analyze the expression of the CD candidate genes in intestinal cell populations of celiac patients and controls and to compare the results to whole biopsies.
   c. To determine whether disease-associated variants have any influence on candidate gene expression.
   d. To perform coexpression analyses in order to reveal possible common regulatory elements altered in CD.

Material and Methods

1. **Subjects**

CD was diagnosed according to the European Society for Pediatric Gastroenterology, Hepatology and Nutrition (ESPHGAN) criteria in force at the time of recruitment, including determination of antibodies against gliadin and endomysium (EMA) or tissue trans-glutaminase (TGA) as well as a confirmatory small bowel biopsy.

For the analysis of the effects of chronic or long-term exposure to gliadin in celiac patients, 3 subject groups were analyzed:

- Active CD patients: newly diagnosed CD patients with clinically active disease (positive for CD-associated antibodies and presenting atrophy of intestinal villi with crypt hyperplasia) who were on a non-restricted (gluten-containing) diet at that time.

- Treated CD patients: normalized CD patients (asymptomatic, antibody-negative and with a recovered intestinal epithelium) who had been on a strict GFD for more than two years.

- Control group: non-celiac individuals not suffering from inflammation at the time of endoscopy where used as a control sample set.

2. **Ethical approval**

All the studies performed are part of the research projects 03-11032, 04-1170 and 06-11030, which have been approved by the local Institutional Ethics Boards (Hospital Universitario de Cruces code CEIC-E09/10 and Basque Clinical Trials and Ethics Committee code PI2013072). All samples were collected between 2003 and 2015 during routine diagnosis endoscopy and after informed consent from patients or their parents were obtained.

3. SNP Genotyping

3.1. DNA samples

3.1.1. GWAS replication study

This study included DNA from 1094 CD patients and 540 healthy controls. Samples were obtained from CEGEC (Spanish Consortium for Genetics of Celiac Disease). 950 DNA samples, 475 CD and 475 controls, came from Hospital Universitario de Cruces and Hospital Universitario Araba (Basque Country); 95 CD samples from Hospital San Juan de Dios (Catalunya), 190 from Hospital Universitario de Valladolid (Castilla Leon); 190 samples, 95 CD and 95 controls, from Hospital Virgen del Camino (Navarra); 190 CD samples from Hospital General de Asturias (Asturias) and 60 CD samples from CATLAB (Catalunya) (Table 1)

Table 1: DNA sample set for the GWAS replication study

CEGEC centers	Community	CD patients	Controls
Hospital Universitario de Cruces Hospital Universitario Araba	Euskal Herria	475	475
Hospital San Juan de Dios	Catalunya	95	-
Hospital Universitario de Valladolid	Castilla León	190	-
Hospital Virgen del Camino	Navarra	95	95
Hospital General de Asturias	Asturias	190	-
CATLAB	Catalunya	60	-

3.1.2. Immunochip

From the 12,041 CD cases and 12,228 controls genotyped in the immunochip project. 545 CD patients and 308 healthy adult blood donors came from CEGEC`s collection (Table 2). Results on this thesis will include only result from the CEGEC sample set.

Table 2: Immunochip sample collection

Population sample	CD patients	Controls
UK	7,728	8,274
The Netherlands	1,123	1,147
Poland	505	533
Spain-CEGEC	**545**	**308**
Spain-Madrid	537	320
Italy	1,374	1,255
India	229	391
TOTAL	12,041	12,228

3.2. **DNA extraction**

3.2.1. **GWAS replication study**

Genomic DNA from Hospital Universitario de Cruces and Hospital Universitario Araba samples was extracted from 150 µl of frozen whole blood using ABI PRISM™ 6100 Nucleic Acid Prep Station (Applied Biosystems, Foster City, CA, USA), and resuspended in elution solution 2, included in the extraction kit. Blood digestion, cell lysis, purification and washing were performed according to the manufacturer´s protocol. Samples from other centers were extracted with their own standard procedures. All samples were checked for quality and quantity by measuring the absorbance at 260nm.

3.2.2. **Immunochip**

DNA from Hospital Universitario de Cruces samples was extracted from 200 µl of frozen blood using NucleoSpin Genomic DNA Blood kit (Macherey-Nagel, Düren, Germany) following manufacturer´s instructions, and resuspended in ddH_2O. DNA was quantified using Quanti-it PicoGreen dsDNA reagent (Invitrogen, Carlsbad, CA, USA) and concentrations were adjusted to 50 ng/µl with a Biomek NXP Laboratory Automation Workstation (Beckman Coulter, Fullerton, CA, USA).

3.3. Whole genome amplification

Whole genome amplification (WGA) was performed by isothermal strand displacement using the GenomiPhi V2 DNA Amplification Kit (QIAGEN Gmbh, Hilden, Germany) in 480 patient and 384 healthy control DNA samples from Hospital Universitario de Cruces due to the small amount of DNA available. DNA was briefly heat-denatured and cooled in sample buffer, containing random hexamers that non-specifically bind to the DNA. A master-mix containing DNA polymerase, additional random hexamers, nucleotides, salts and buffers was added and isothermal amplification proceed at 30°C for 1.5 hours. After amplification the enzyme is heat inactivated during 10 minute incubation at 65°C.

3.4. Single Nucleotide Polymorphism selection

3.4.1. GWAS replication study

Ten SNPs reported to tag the seven regions identified by Hunt *et al.* [86] were selected and subjected to genotyping: rs2816316 (1q31, *RGS1*), rs917997 and rs13015714 (2q11-12, *IL18RAP/IL18R1*), rs6441961 (3p21, *CCR1/CCR3/CCR2*), rs17810546 and rs9811792 (3q25-26, *IL12A/SCHIP1*), rs1464510 (3q28, *LPP*), rs6822844 and rs13119723 (4q27 , *IL2/IL21* and *KIAA1109*), rs1738074 (6q25, *TAGAP*) and rs3184504 (12q24, *SH2B3*) (Table 3).

3.4.2. Immunochip

The marker selection for the Immunochip project is extensively described by Trynka *et al.* [79]. In total, the consortium selected 186 distinct *loci* containing markers that were genome wide significant ($p<5x10^{-8}$) from 12 autoimmune diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). Briefly, all 1000 Genomes Project pilot phase CEU population variants within 0.1 cM of the lead SNP for each disease and region were selected. Additional genomic region resequencing content was submitted for Immunochip analysis at specific *loci* from cases with CD, Crohn's disease and type 1 diabetes and controls.

3.5. Single Nucleotide Polymorphism genotyping

3.5.1. GWAS replication study

SNPs were genotyped using commercially available TaqMan allelic discrimination assays developed by Applied Biosystems, which include two allele-specific MGB probes containing distinct fluorescent dyes and a PCR primer pair to detect specific SNP targets (Table 3). Genotyping was performed following the manufacturer's specifications. Amplification was performed in a standard PCR thermal cycler, and pre- and post-amplification fluorescent was measured on an ABIPrism 7900HT sequence detection system (Applied Biosystems, Foster City, CA, USA). DNA samples were processed in 384 well plates, each of then containing four negative controls, prepared with a Biomek NXp automated liquid handler (Beckman Coulter, Fullerton, CA, USA).

Table 3: TaqMan Genotyping assay list

Candidate Gene	Region	SNP	Assay ID
RGS1	1q31	rs2816316	C_15810686_10
IL18R1	2q11-12	rs13015714	C_31439507_10
IL18RAP	2q11-12	rs917997	C_345197_1
CCR3	3p21	rs6441961	C_26450554_10
SCHIP	3q25-26	rs17810546	C_32594555_10
IL12A	3q25-26	rs9811792	C_2936004_10
LPP	3q28	rs1464510	C_8239299_10
KIAA1109	4q27	rs13119723	C_26404981_10
IL21	4q27	rs6822844	C_28983601_10
TAGAP	6q25	rs1738074	C_2966098_10
SH2B3	12q24	rs3184504	C_2981072_10

3.5.2. Immunochip

Samples were genotyped using the Immunochip according to Illumina's protocols (at labs in London, UK; Hinxton, UK; Groningen, The Netherlands; and Charlottesville, Virginia, USA). NCBI build 36 (hg18) mapping was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm).

Additionally, genotyping of 44 top-associated SNPs from the Immunochip project was performed with a Fluidigm Biomark dynamic array (48.48) and SNPtype assays (Fluidigm Corp.) in 26 samples with expression results in which DNA was available. Eight samples had been previously genotyped in the Immunochip sample set and were used as quality control for the new genotyping. Three samples had to be removed from the study due to failed genotyping, resulting in a total number of 23 samples, 14 controls and 9 celiac patients. The assay design was performed by the Fluidigm Assay Design Group. Seven of the target SNPs did not fulfill the established assay design requirements due to adjacent SNPs within 20–30 bases on each side of the target SNP, GC content >65% or triallelic SNPs. After an in-deep analysis of those seven SNPs, taking into account the allelic frequencies of the target SNP and the adjacent SNPs and the frequency of each allele in the case of the unique triallelic SNP (rs61907765) in Ensembl, we decided to omit this obstacle in the design of six SNPs and to remove the SNP rs60215663 from the analysis due to smaller minor-allele frequency than adjacent SNPs (Supplementary table 1).

3.6. **Data analysis**

3.6.1. **Single Nucleotide genotyping**

SDS version 2.3 software was used for genotype calling and minor allele frequencies were compared in 2x2 contingency tables. Allelic and genotypic frequencies in cases and controls were compared using Fisher's exact test and X^2 tests (genotypic test for trend and dominant and recessive models), respectively, using EPI-INFO v.6.0 (Centers for Disease Control, Atlanta, GA).

3.6.2. **Immunochip statistical analyses**

Case-control association analyses for the whole Immunochip sample set were performed with PLINK v1.07. Graphs were plotted using LocusZoom39 [87].

Additionally, CEGEC sample set genotype data was analysed independently in a case-control association study, trying to identify population specific association signals.

4. Functional analysis of candidate genes

4.1. Biopsy samples

Biopsy specimens from the distal duodenum of patients were obtained using standard clinical procedures by Pediatric Gastroenterologist; a portion of the sample was used for diagnostic pathology examination and another in the present investigation. Intestinal biopsies from CD children at the time of diagnosis were compared with tissue samples from the same patients in remission after treatment with a GFD for more than 2 years, and to non-celiac controls with no inflammation at the time of endoscopy. A detailed description of each studied group is found previously in subjects section.

4.1.1. GWAS replication study

Intestinal biopsies from 29 CD children at diagnosis (18 females/11 males) and from the same patients after > 2 years on GFD were compared. Eight tissue samples from non-celiac individuals were used as controls.

4.1.2. Immunochip

Intestinal biopsies pairs (diagnosis and after > 2 years on GFD) from 15 CD children, and 15 tissue samples from non-celiac controls were analyzed.

4.2. Cell populations from biopsies

The principal cell populations of intestinal biopsies are enterocytes and immune cells. Enterocytes are characterized for being CD326 positive cells, that means that enterocytes expressed the Epithelial cell adhesion molecule (EpCAM) in their surface, a molecule that is involved in cell signaling, migration, proliferation and differentiation. On the other hand, immune cells express the CD45 antigen on their surface, a protein encoded by the *PTPRC* gene that is a member of the protein tyrosine phosphatase family.

These characteristics make it possible to separate both cell populations from a biopsy sample, and allow the independent study of both cell types in subsequent analysis.

For that purpose, MACS magnetic cell separation technology was used (Miltenyi Biotec), following manufacturer´s protocol for CD45 MicroBeads and MS separation columns. Briefly, biopsy samples were collected and processed freshly. Cells were mechanically separated by agitation in complete medium (RPMI + antibiotics + FBS + DTT + EDTA) for 1 hour. Cells released to the media were collected by centrifugation after filtering the media though pre-separation filters (20 µm) and the remaining *lamina propria* was stored in RLT buffer at -80ºC.

Dead cell removal kit was used in a first purification step to prepare a viable single-cell homogenous suspension. In a second purification step, cells were labeled with CD45 magnetic microbeads to separate the cell mixture in two fractions, the CD45 positive immune cells (attached to the column) and CD45 negative enterocytes (flow through). Both fractions were stored in RLT lysis buffer at -80ºC for a posterior nucleic acid extraction.

In order to confirm the purity of both cell fractions we performed cytometer analyses (Figure 6).

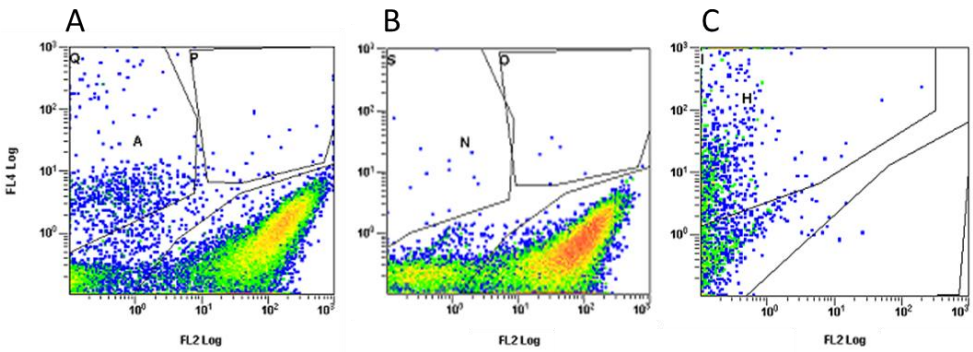


Figure 6: Cytometer results of cell fractions separated from whole biopsies. FL2 in X axis correspond to CD326 anti-antibody and FL4 in Y axis to CD45 anti-antibody. A: Cell mixture before separation. B: CD45- and CD326+ cell fraction. C: CD45+ and CD326- cell fraction.

4.3. **RNA extraction**

4.3.1. **GWAS replication study**

Biopsies were stored in liquid nitrogen until use. Frozen tissue samples were disrupted with disposable plastic pellet pestles (Kontes, Vineland, NJ, USA) in 1.5-ml microcentrifuge tubes. At the beginning of this thesis, biopsies were homogenized using a QIAshredder column (QIAGEN Gmbh, Hilden, Germany) and RNA was isolated using RNeasy Micro Kit (QIAGEN) with DNase I treatment and subsequently stored at -80ºC until use. RNA was quantified using a Nanodrop spectrophotometer (Thermo Scientific, Waltham, MA) and the concentration of all RNA samples was adjusted to 8 ng/µl.

4.3.2. **Immunochip genes in biopsies and cell populations**

Due to the design of new experiments were microRNA was also isolated, RNA for immunochip candidate genes study was isolated using NucleoSpin microRNA kit (Macherey-Nagel, Düren, Germany) following manufacturer´s instructions. Small and large RNAs were collected in a single fraction (total RNA), and stored at -80ºC until use. RNA was quantified using a Nanodrop spectrophotometer (Thermo Scientific, Waltham, MA) and the concentration of all RNA samples was adjusted to 8 ng/µl.

4.4. **Gene expression: RT-PCR**

4.4.1. **GWAS replication study**

GWAS candidate gene expression was quantified by real-time quantitative reverse transcription PCR (RT-PCR). Expression assays for each gene were purchased as commercial Assay-on-Demand (Applied Biosystems, Foster City, CA, USA), each assay consisting of a pair of unlabeled primers and a FAM labeled MGB probe (Table 4). The expression of the housekeeping gene *RPLPO* (large ribosomal protein) was simultaneously quantified in each experiment (VIC labeled MGB probe) and used as an endogenous control of input RNA. Experiments were carried out in triplicate in a 7900 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) using 8 ng of RNA and a single-reaction enzyme mixture (QuantiTect Probe RT-PCR kit, QIAGEN). Relative expression of each gene was calculated using the accurate Ct method [34] and normalized to one of the control samples.

Table 4: TaqMan Gene Expression assay list for GWAS candidate gene analysis

Gene Symbol	Assay ID
RGS1	Hs00175260
IL18R1	Hs00175381
IL18RAP	Hs00187256
SCHIP	Hs00205829
IL12A	Hs00168405
LPP	Hs00194400
KIAA1109	Hs00361070
IL21	Hs00222327
TAGAP	Hs00611823
SH2B3	Hs00193878

4.4.2. **Immunochip**

RNA was normalized to 8 ng/µl and converted to cDNA using the AffinityScript cDNA Synthesis kit (Agilent Technologies, Santa Clara, CA, USA) following the manufacturer's protocol. Gene expression analyses were performed using Fluidigm Biomark 48.48 dynamic arrays (Fluidigm Corp., South San Francisco, CA, USA) and commercially available TaqMan Gene Expression assays (Table 5). Housekeeping gene RPLPO was simultaneously quantified and used as an endogenous control of input RNA (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA). Relative expression in each sample was calculated using the accurate Ct method [34] and normalized to the average expression value of the 15 control samples. Gene expression results are publicly available at the Gene Expression Omnibus data repository (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE61849.

Table 5: TaqMan Gene Expression assay list for Immunochip candidate gene analysis

Gene Symbol	Assay ID	Gene Symbol	Assay ID
ADAD1	Hs00330122_m1	*PVT1*	Hs01069044_m1
ARHGAP31	Hs00393361_m1	*TMEM187*	Hs01920894_s1
ATXN2	Hs00268077_m1	*TNFSF18*	Hs00183225_m1
BACH2	Hs00222364_m1	*TREH*	Hs00389383_m1
CCR1	Hs00928897_s1	*TTC34*	Hs01128292_m1
CCR2	Hs00704702_s1	*UBASH3A*	Hs00957643_m1
CD28	Hs01007422_m1	*UBE2E2*	Hs00994287_m1
CIITA	Hs00172094_m1	*UBE2L3*	Hs00748530_s1
CLK3	Hs00999875_m1	*YDJC*	Hs00419214_g1
CSK	Hs01062585_m1	*ZFP36L1*	Hs00245183_m1
DDX6	Hs00898915_g1	*ZMIZ1*	Hs00393480_m1
ELMO1	Hs00404994_m1	*SOCS1*	Hs00705164_s1
GLB1	Hs01035168_m1	*FASLG*	Hs00181225_m1
HCFC1	Hs00232039_m1	*ITGA4*	Hs00168433_m1
MMEL1	Hs00364353_m1	*ICOS*	Hs00359999_m1
OLIG3	Hs00703087_s1	*CTLA4*	Hs03044418_m1
PFKFB3	Hs00998700_m1	*IRF4*	Hs01056533_m1
PLEK	Hs00950975_m1	*IRAK1*	Hs01018347_m1
POU2AF1	Hs01573371_m1	*STAT4*	Hs01028017_m1
PRM1	Hs00358158_g1	*PRKCQ*	Hs00989970_m1
PRM2	Hs04187294_g1	*LTF*	Hs00914334_m1
PTPN2	Hs00959886_g1	*CCR4*	Hs00747615_s1
PUS10	Hs00328708_m1	*ICOSLG*	Hs00323621_m1

Candidate genes identified in the CEGEC sample-set association study were also analyzed with the same procedure and in the same samples. The commercially available TaqMan gene expression assays are summarized in Table 6.

Table 6: TaqMan Gene Expression assay list for CEGEC candidate gene analysis

Gene Symbol	Assay ID
BLK	Hs01017452_m1
CAPSL	Hs00921468_m1
FAM167A	Hs00697562_m1
GALC	Hs01012300_m1
GPR65	Hs00269247_s1
IL12RB2	Hs00155486_m1
IL23R	Hs00332759_m1
IL7R	Hs00902334_m1
PSMG2	Hs00220315_m1
SERBP1	Hs00967385_g1
UGT3A1	Hs01014799_m1
UGT3A2	Hs04177793_m1

4.5. **Data analysis and statistics**

4.5.1. **GWAS replication study**

Differences in gene expression levels were analyzed with the nonparametric Wilcoxon matched-pairs rank test (diagnosis *vs* treated) and Mann Whitney test (non-celiac *vs* both disease groups) using Instat v.3.05 (GraphPad Software, Inc. La Jolla, CA, USA). Two-tailed p-values, below 0.05 were considered significant.

4.5.2. **Immunochip**

Differences in gene expression levels were analyzed with nonparametric Wilcoxon matched pairs rank test (diagnosis *vs.* treated) and Mann–Whitney U-test (non-celiac *vs* both disease groups). All statistic calculations were performed in GraphPad Prism 5 (GraphPad Software, La Jolla, CA, USA). Extreme outliers exceeding >3 standard deviation SD from the mean of each group were considered methodological errors and were removed from statistical comparisons.

4.6. **Coexpression analysis**

Coexpression between immunochip candidate gene pairs was calculated using Pearson´s correlation. Merlin 1.1.2 software was used to test association between SNP genotype and candidate gene expression [88]. The association was tested independently in each of the studied groups in order to avoid false associations due to duplicated genotypes in CD sample pairs.

4.7. **Genotype-phenotype correlation**

The relationship between the expression of the gene in active CD and the genotype of the SNP was calculated using the R Correlation Coefficient (http://www.fon.hum.uva.nl/Service/Statistics/Correlation_coefficient.html) in those samples for which both expression levels and SNP genotypes were available.

Chapter 1:

Revisiting genome wide association studies in celiac disease: replication study in Spanish population and expression analysis of candidate genes.

1. **Introduction**

CD is a chronic, immune mediated disorder caused by intolerance to ingested gluten that develops in genetically susceptible individuals and affects approximately 1% of Caucasians [32,89]. The major susceptibility locus is located in the MHC region on chromosome 6p21. More than 90% of patients with CD express the HLA-DQ2 heterodimer, and those lacking HLA-DQ2 present the HLA-DQ8 molecule. However, HLA-DQ2 is common in the general population, being present in approximately 30% of Caucasians. Overall, the HLA locus is thought to explain approximately 40% of the heritability of CD, and a large effort has been put into the search of other loci that may contribute to the genetic predisposition to the disease [37,90,91].

As in other complex disorders, GWAS have also been performed in CD, and the first such study in 778 CD patients and 1422 controls from the UK identified strong association in a 500 kb LD block in 4q27 [71]. The region contains the *IL2* and *IL21* genes that encode for cytokines involved in T cell maturation and proliferation, and are interesting functional candidates for CD pathogenesis. Moreover, this region has also been associated with type 1 diabetes (T1D) and rheumatoid arthritis (RA), suggesting that it could be a general autoimmune susceptibility locus [92].

A subsequent follow-up of the top SNPs of this first study identified seven new risk regions in Europeans from the Netherlands, UK and Ireland [76]. Several of these loci also harbor immune related genes, like *RGS1* (1q31), *IL18R1/ IL18RAP* (2q11e2q12), *CCR3* (3p21), *IL12A/SCHIP1* (3q25e3q26), *TAGAP* (6q25), and *SH2B3* (12q24). Again, some have also been associated with T1D and Crohn's disease, adding further support to their implication in immune mediated disorders [93,94].

Very recently, a second GWAS and replication study in more than 9,000 cases and 15,000 controls has extended the genome wide significant association to a further 13 loci, most of which contain genes involved in the immune response (*BACH2, CCR4, CD80, CIITA-SOCS1-CLEC16A, ICOSLG*, and *ZMIZ1*) or in T cell maturation in the thymus (*ETS1, RUNX3, THEMIS*, and *TNFRSF14*) [77]. Analyses of the SNP effects on gene expression have been performed in whole blood, showing some correlation with cis gene expression.

GWAS have broadened our view of the genetic makeup of CD, and have shown that a considerable number of low penetrant variants are responsible for the inherited risk of

developing the disease. However, the individual contribution of each polymorphism is low (OR <1.5) because the differences in allele frequencies between cases and controls seldom go beyond 10%. Additionally, population related differences are often diluted because GWAS sample sets are a combination of non-uniform cohorts of different sizes with differential contributions to the overall association signal. Moreover, it must be remembered that association studies can only pinpoint the location of signals; subsequent selection of candidate genes has been aprioristic and frequently biased by our current view of disease pathogenesis, with no experimental results to support any functional involvement of these genes in the target tissue of patients with CD.

To investigate these issues, we performed a replication study in eight loci identified in the first CD GWAS in a Spanish population, and analyzed the expression of 10 proposed candidate genes to question their implication in CD development, as well as the influence of the associated SNPs in their expression.

2. **Methods**

Celiac disease was diagnosed according to the ESPHGAN criteria, including determination of antibodies against gliadin and endomysium (EMA) or tissue transglutaminase (TGA) as well as a confirmatory small bowel biopsy. This study was approved by the Institutional Board of all the CEGEC centers.

DNA samples were extracted from whole blood using conventional methods; 11 SNPs reported to tag the eight regions identified in the first CD GWAS were genotyped in 1,094 CD patients and 540 adult blood donors from the CEGEC collection, using commercially available TaqMan allelic discrimination assays (Applied Biosystems, Foster City, California, USA) on an ABI7900HT sequence detection system (Applied Biosystems).

SDS version 2.3 software was used for genotype calling and minor allele frequencies were compared in 232 contingency tables using c2 tests using EPI-INFO v.6.0 (Centers for Disease Control and Prevention, Atlanta, Georgia, USA). Results were combined with those from a previous Spanish sample of 558 patients and 465 controls genotyped in the replication of a second CD GWAS for joint association analysis [77].

The expression of *RGS1, IL18RAP, IL18R1, IL12A, SCHIP1, LPP, IL21, KIAA1109, TAGAP*, and *SH2B3* was quantified by RT-PCR using commercial Assay-on-Demand sets (Applied Biosystems) and a single reaction enzyme mixture (QuantiTect Probe RT-PCR kit, Qiagen, Hilden, Germany) with 8 ng of total RNA from intestinal biopsies from 29 CD children at diagnosis (18 females/11 males, on a gluten-containing diet, with CD associated antibodies and atrophy of intestinal villi and crypt hyperplasia), and from the same patients after >2 years on GFD (asymptomatic, antibody negative, and normalized intestinal epithelium). Eight tissue samples from non-celiac individuals were used as controls. Experiments were done in triplicate in an ABI7900HT system and housekeeping gene *RPLPO* was simultaneously quantified and used as an endogenous control of input RNA. Relative expression in each sample was calculated using the accurate Ct method and normalized to one of the control samples, as previously described [34].

Differences in gene expression levels were analyzed with the non-parametric Wilcoxon matched pairs rank test (diagnosis vs treated groups) and Mann-Whitney U test (non-celiac vs both disease groups) using Instat v.3.05 (GraphPad Inc, La Jolla, California, USA). Extreme outliers exceeding >3SD from the mean of each group were considered

methodological errors and were removed from statistical comparisons (one GFD treated CD sample in *SCHIP1* and an active disease patient in *KIAA1109*). Due to the limited amount of RNA from each biopsy, only 15 biopsy pairs were analyzed for each gene, to uniformly represent, when possible, the three genotypes in each SNP.

3. **Results and discussion**

The recent GWAS performed in CD offer a catalogue of associated genomic regions that may underlie the genetic risk to the disease [76,77,91]. However, associations must be replicated to determine whether the genetic makeup of CD is robust and comparable across different populations; most importantly, the functional implication of the proposed genes and variants must be experimentally addressed by analysis in disease tissue or cell models. In the present work, we performed genetic association of eight genomic regions identified after the first GWAS in a sample from the Spanish population and performed functional studies in proposed candidate genes [76,91].

Out of the 11 SNPs genotyped in the eight associated loci, significant association with CD in our cohort from the Spanish population was detected in three regions, 1q31, 2q11e2q12, and 3q25, which harbor candidate genes *RGS1, IL18RAP/IL18R1*, and *SCHIP1*, respectively, involved mainly in the immune response (Table 7). Combined analysis of our genotyping results with those of a previously studied Spanish sample set confirmed these associations, and two other regions reached statistical significance: 3q28 (*LPP*), which had been identified in that study but not in the present report; and 4q27 (*KIAA1109*), which did not show association in any of the Spanish collections. On the other hand, the signal in 3p21 (*CCR3*) detected in the previous GWAS replication disappeared after combining both samples. Taken together, our results follow other similar studies, and are a confirmation of our current knowledge on the genetics of CD, stressing the importance of analyzing large samples to obtain robust results that can be replicated. However, as is the case of the majority of studies, it is also evident that the strength of these signals is relatively small, and the ability to detect significant association often depends on minute allele frequency differences across populations, which could account, at least in part, for some of the negative results [71,76,77,80].

On the other hand, another major objective of genetic association studies is to identify the functional culprits that are implicated in the development of the disease, in order to increase our understanding of its pathogenesis. To investigate their putative role in the disease process, the expression of candidate genes proposed following GWAS was tested in the disease tissue at diagnosis and after GFD, and compared to non-celiac tissue. Four of the 10 genes analyzed were significantly overexpressed in active CD samples when compared with non-CD tissue: *IL18RAP, IL21, SH2B3*, and *IL12A* are all involved in T cell signaling and participate in the activation of Th1 and/or Th17 responses, which are

responsible for mucosal inflammation in active CD [47]. Alteration of gene expression in active CD tissue is, however, at least partly the consequence of disease mediated phenomena (e.g., the proinflammatory environment) and might not reflect a primary event of genetic origin. In fact, we could not detect any effect of the associated SNP genotype on the expression of *IL18RAP* or *IL21*, which were only overexpressed in active mucosa. In turn, *SH2B3* is constitutively upregulated in patient mucosa, independent of disease status, suggesting a defect that precedes disease development and could be due to a genetic variant. Indeed, the presence of rs3184504*T is associated with higher expression of *SH2B3* in intestinal mucosa of active CD patients.

Table 7: Association study of GWAS identified SNPs in our Spanish cohort and that published in the second GWAS follow-up [77] . Significant associations ($p<0.05$) are shown in bold type. *NCBI 36.3 refseq coordinates. **Minor allele in bold type. NA, not available.

					Spanish CEGEC				Spanish GWAs				Meta-analysis				
Gene	**SNP**	**CHR**	**BP***	**Alleles****	**MAF-A**	**MAF-U**	**p value**	**OR**	**MAF-A**	**MAF-U**	**p value**	**OR**	**MAF-A**	**MAF-U**	**p value**	**OR**	**CI**
RGS1	rs2816316	1	190803436	A:**C**	0.1518	0.1994	**0.0008**	0.72	0.1536	0.1882	**0.0421**	0.78	0.1525	0.1944	**0.0001**	**0.75**	0,64-0,87
IL18RAP	rs917997	2	102437000	C:**T**	0.2806	0.2407	**0.0159**	1.23	0.2791	0.2691	0.6206	1.05	0.2801	0.2534	**0.0357**	**1.15**	1,01-1,31
IL18R1	rs13015714	2	102338297	**G**:T	0.2797	0.2273	**0.0032**	1.32	NA	NA	NA	NA	0.2797	0.2273	**0.0032**	**1.32**	1,09-1,60
CCR3	rs6441961	3	46327388	C:**T**	0.3307	0.3439	0.4538	0.94	0.3582	0.2875	**0.0009**	1.38	0.3400	0.3187	0.1152	1.1	0,97-1,24
SCHIP1	rs17810546	3	161147744	A:**G**	0.1197	0.0939	**0.0288**	1.31	0.1066	0.09469	0.3868	1.14	0.1153	0.0942	**0.0358**	**1.25**	1,03-1,52
IL12A	rs9811792	3	161179692	**C**:T	0.4936	0.4971	0.8511	0.99	NA	NA	NA	NA	0.4936	0.4971	0.8511	0.99	0,85-1,15
LPP	rs1464510	3	189595248	**A**:C	0.4495	0.4190	0.1108	1.13	0.4618	0.403	**0.0090**	1.27	0.4539	0.4116	**0.0036**	**1.19**	1,06-1,34
IL2/IL21	rs6822844	4	123728871	G:**T**	0.0950	0.1101	0.1786	0.85	NA	NA	NA	NA	0.0950	0.1101	0.1786	0.85	0,66-1,09
KIAA1109	rs13119723	4	123437763	A:**G**	0.0941	0.1122	0.1175	0.82	0.1082	0.1339	0.0804	0.78	0.0991	0.1222	**0.0108**	**0.79**	0,66-0,95
TAGAP	rs1738074	6	159385965	C:**T**	0.4229	0.4065	0.3812	1.07	0.4427	0.3995	0.0544	1.19	0.4297	0.4033	0.0646	1.11	0,99-1,25
SH2B3	rs3184504	12	110368991	C:**T**	0.4551	0.4341	0.2674	1.09	NA	NA	NA	NA	0.4551	0.4341	0.2674	1.09	0,93-1,27

Interestingly, this polymorphism is a non-synonymous SNP that provokes amino acid change R262W in an important domain of the protein and has been under positive selection because it provides a more efficient antibacterial response [95]. We were not able to detect genetic association of this variant with the disease in our Spanish samples, although it has indeed been observed in several other European groups[76,91]. A similar situation is observed in *IL12A*, which is also upregulated in the GFD group, and where a single copy of rs9811792*C is capable of increasing mRNA production, thus suggesting a regulatory role for this variant, although no significant association of this SNP was observed in our study.

Two other genes, *LPP* and *SCHIP1*, both with unclear function in the disease process, also showed some alteration in biopsies from patients compared to non-celiac controls. *LPP* is highly expressed in the small intestine and has been implicated in cell adhesion and could have a structural role in epithelium maintenance [96]. This gene appears to be activated in patients compared to controls, and interestingly, significant differences are seen only in the GFD treated group, supporting a genetically driven constitutive alteration. The presence of rs1464510*A, which is associated with higher risk of the disease in the combined Spanish sample, seems however to have an effect opposite to what is expected, since rs1464510*AA biopsies from active CD patients have lower *LPP* mRNA levels compared to tissue from rs1464510*AC and rs1464510*CC patients. *SCHIP1* is underexpressed in patient tissue samples, and although the decrease is more evident in biopsies from individuals with active CD, a similar trend is observed when controls are compared to GFD treated patients. As in the case of *LPP*, the effect of allele rs17810546*G on *SCHIP1* gene expression is in the opposite direction; one would anticipate it to be associated with a lower amount of mRNA, but biopsies from active patients with the risk allele showed higher expression level of *SCHIP1*. In the case of *KIAA1109*, there are decreased expression levels in active patients compared to, but there is no evidence that the associated allele rs13119723*A participates in gene regulation. No differences in the expression levels of *RGS1*, *IL18R1*, and *TAGAP* were observed among the different groups of biopsies.

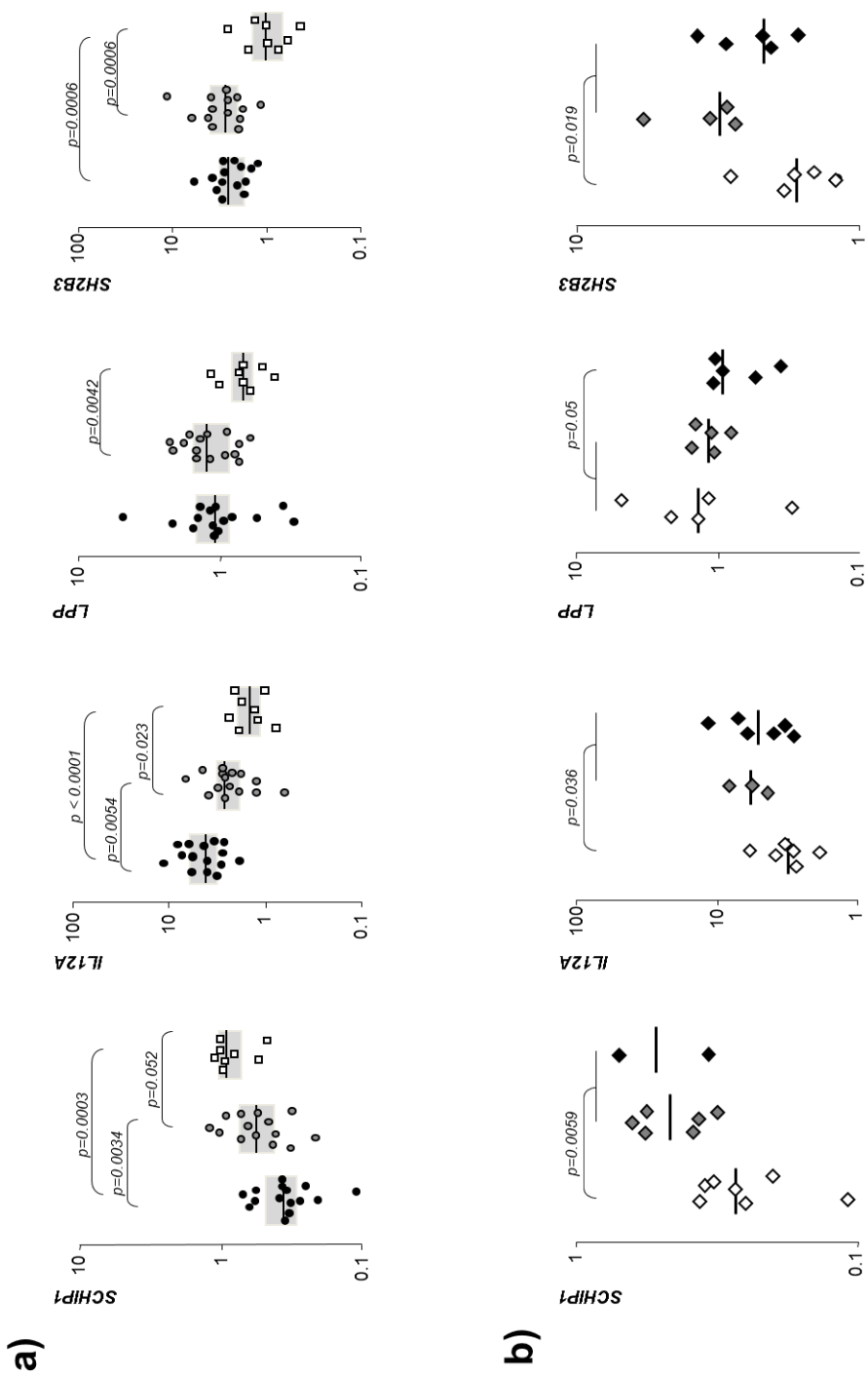

a)
SCHIP1
p=0.0003
p=0.0034
p=0.052
IL12A
p < 0.0001
p=0.0054
p=0.023
LPP
p=0.0042
SH2B3
p=0.0006
p=0.0006
b)
SCHIP1
p=0.0059
IL12A
p=0.036
LPP
p=0.05
SH2B3
p=0.019

Figure 7: Expression analyses of proposed candidate genes from genome wide association studies (GWAS) regions in intestinal mucosa. Only genes showing significant differences between gluten-free diet (GFD) patients and controls are shown. (A) Comparison of active (black circles) and treated (grey circles) celiac disease (CD) patients and non-celiac controls (white squares). (B) Effect of single nucleotide polymorphism (SNP) genotypes on relative gene expression levels in active CD patients homozygous for the protective allele (white diamonds), heterozygous patients (grey diamonds), and patients homozygous for the risk variant (black diamonds). Note that the risk allele in each SNP is the one that is more frequent among CD patients, not necessarily the minor allele.

Overall, there are different functional relationships between the expression of candidate genes in associated regions and SNP genotypes, and the link with disease pathogenesis is not straightforward. Taking into account that genetic variation could influence basal expression levels in non-inflamed CD patients, it is interesting to see that in the four genes where differences in expression levels prevailed between GFD-CD and control mucosa (*SCHIP1, IL12A, LPP*, and *SH2B3*), an effect of the associated GWAS SNP could be observed. These findings could indicate that these genes and their associated variants are indeed aetiological players in the pathogenesis of the disease. However, the biology of complex diseases is certainly much more complex than a direct SNP altered gene function-disease relationship, and we must be very cautious when proposing aetiological genes and pathogenic mechanisms based only on association peaks. Larger functional studies (both in sample size and in the number of genes that are analyzed in each associated genomic region), as well as experimental studies in cell or animal models, will clarify whether GWAS hits can be useful for the identification of relevant aetiological variants.

Chapter 2:

Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease.

1. **Introduction**

CD is a common (prevalence 1:100) chronic immune mediated enteropathy caused by intolerance to ingested gluten that develops in genetically predisposed individuals. The typical histological findings in active CD comprise villous atrophy, crypt hyperplasia and lymphocytic infiltration of the small intestinal mucosa, and the only effective treatment is strict lifelong GFD [62].The major CD susceptibility locus maps to the MHC region on chromosome 6p21 and has been estimated to be responsible for 40% of the genetic contribution to CD; in fact, virtually all patients are HLA–DQ2- or HLA–DQ8-positive [65]. However, risk HLA variants are necessary but not sufficient for CD development, as those alleles are also common in general population, pointing to the contribution of other loci to the genetic predisposition to develop the disease.

To date, two GWAS have been performed in CD, revealing 26 regions of genetic susceptibility to the disease [71,76,77]. More recently, 13 additional susceptibility loci have been discovered with the Immunochip genotyping array, where immune mediated disease loci containing markers that had achieved genome wide significance ($p<5 \times 10^{-8}$) in 12 diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, CD, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis) were densely genotyped [79]. Many of the loci identified are also associated with other autoimmune or chronic immune-mediated diseases, with particular overlapping between CD, type 1 diabetes [97] and rheumatoid arthritis [98].

Several genes within those regions have been proposed as etiological candidates, most of them previously related to the immune response or to T-cell maturation, and it has been suggested that they might participate in the different stages of the pathogenesis of CD. However, association studies are only able to pinpoint the location of susceptibility loci and the subsequent selection of candidate genes is often aprioristic and biased by the current paradigm of CD pathogenesis, with no robust experimental results to support any functional involvement of those candidate genes in the target tissue of CD patients. So far, the large-scale studies performed in CD have discovered a total of 57 independent CD association signals from 39 non-HLA loci [79]. Twenty-nine of those regions map to a single protein coding gene, whereas the majority seem to localize to intergenic regions, suggesting more than one possible causal gene or some yet unidentified functional

elements of the genome. Overall, 66 candidate genes have been proposed based on their localization under the association peaks, but there is a need to perform functional studies in the disease target tissue to prove the causative mechanism suggested for each association signal.

In a previous work, our group analyzed the expression of the 10 candidate genes proposed in the first GWAS in intestinal biopsies from patients and controls [76], to determine the influence of associated SNP genotypes in their expression and their possible implication on CD development. We observed that several genes were differentially expressed depending on disease status, and found different functional relationships between the expression of candidate genes and SNP genotypes [99].

In the present work, we wanted to question the implication of the additional proposed candidate genes in disease development. To investigate their putative role in the disease process we selected an additional set of 45 candidate genes with known function (table 8) and analyzed their expression in the disease tissue of celiac patients at diagnosis and after more than 2 years on GFD, and compared it with non-celiac controls. We also aimed to determine whether disease-associated variants have any influence on gene expression, considering the genotypes of the top-associated SNPs in the Immunochip project for each candidate gene. Moreover, we performed coexpression analyses in order to reveal possible common regulatory elements, which could be altered in celiac patients on account of inflammation or owing to predisposing genetic determinants.

Table 8: Studied candidate genes and associated SNPs.

dbSNP ID (release 138)	**HGVS name (GRCh38)**	**Candidate Protein-coding Genes**
rs4445406	1:g.2607961T>C	*MMEL1* *TTC34*
rs12068671	1:g.172711891T>C	*FASLG*
rs859637	1:g.172741860T>C	*TNSF18*
rs13003464	2:g.60959694A>G	*PUS10*
rs10167650	2:g.68418428T>G	*PLEK*
rs1018326	2:g.181143073T>C	*ITGA4*
rs6715106	2:g.191048308A>G	*STAT4*
rs6752770	2:g.191108837A>G	
rs12998748	2:g.191083911G>T	
rs1980422	2:g.203745673C>T	*CD28*
rs34037980	2:g.203905331A>G	*CTLA4*
rs10207814	2:g.203595238C>T	*ICOS*
rs4678523	3:g.32996229T>C	*CCR4* *GLB1*
rs2097282	3:g.46336534C>T	*CCR1*
rs7616215	3:g.46164194C>T	*CCR2* *LTF*
rs61579022	3:g.119404431G>A	*ARHGAP31*
rs1050976	6:g.408079C>T	*IRF4*
rs12203592	6:g.396321C>T	
rs7753008	6:g.90099920T>C	*BACH2*
rs17264332	6:g.137684378A>G	*OLIG3*
rs77027760	6:g.137680924G>A	
rs79758729	7:g.37378851A>G	*ELMO1*
rs10808568	8:g.128251814A>C	*PVT1*

rs2387397	10:g.6348230G>C	*PFKFB3* *PRKCQ*
rs1250552	10:g.79298270A>G	*ZMIZ1*
rs7104791	11:g.111326133T>C	*POU2AF1*
rs10892258	11:g.118709156G>A	*TREH* *DDX6*
rs61907765	11:g.128522042C>T	*ETS1*
rs3184504	12:g.111446804T>C	*ATXN2*
rs11851414	14:g.68792785T>C	*ZFP36L1*
rs1378938	15:g.74804102T>C	*CLK3* *CSK*
rs6498114	16:g.10870261G>T	*CIITA*
rs243323	16:g.11267345A>G	*SOCS1*
rs80073729	16:g.11279940G>A	*PRM1*
rs9673543	16:g.11291099A>G	*PRM2*
rs11875687 rs62097857	18:g.12843138T>C 18:g.12857759G>A	*PTPN2*
rs1893592	21:g.42434957A>C	*UBASH3A*
rs58911644	21:g.44209238A>T	*ICOSLG*
rs4821124	22:g.21625000T>C	*UBE2L3* *YDJC*
rs13397	X:g.153982797G>A	*HCFC1* *TMEM187* *IRAK1*

2. Material and methods

2.1. Patients and biopsies

CD was diagnosed according to the European Society of Pediatric Gastroenterology Hepatology and Nutrition criteria in force at the time of recruitment, including anti-gliadin, anti-endomysium and anti-transglutaminase antibody determinations as well as a confirmatory small bowel biopsy. The study was approved by the Institutional Boards (Cruces University Hospital code CEIC-E09/10 and Basque Clinical Trials and Ethics Committee code PI2013072) and analyses were performed after informed consent was obtained from all subjects or their parents. Biopsy specimens from the distal duodenum of each patient were obtained during routine diagnosis endoscopy.

The sample set consisted of 15 CD children at diagnosis (on a gluten containing diet, with CD-associated antibodies, atrophy of intestinal villi and crypt hyperplasia), and the same patients in remission after being treated with GFD for > 2 years (asymptomatic, antibody negative and normalized intestinal epithelium at that time), plus 15 tissue samples from non-celiac individuals not suffering from inflammation at the time of endoscopy used as controls. Total RNA was extracted from small bowel biopsies using the NucleoSpin microRNA kit (Macherey-Nagel, Düren, Germany) following manufacturer's instructions.

2.2. RNA samples and gene expression

RNA was normalized to 8 ng/ µl and converted to cDNA using the AffinityScript cDNA Synthesis kit (Agilent Technologies, Santa Clara, CA, USA) following manufacturer's protocol. Gene expression analyses were performed using Fluidigm Biomark 48.48 dynamic arrays (Fluidigm Corp., South San Francisco, CA, USA) and commercially available TaqMan Gene Expression assays. Housekeeping gene *RPLPO* was simultaneously quantified and used as an endogenous control of input RNA (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA). Relative expression in each sample was calculated using the accurate Ct method [34] and normalized to the average expression value of the 15 control samples as previously described. Gene expression results are publicly available at the Gene Expression Omnibus data repository (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE61849.

Differences in gene expression levels were analyzed with nonparametric Wilcoxon matched pairs rank test (diagnosis vs treated) and Mann–Whitney U-test (non-celiac vs

both disease groups). Coexpression was calculated using Pearson correlation. All statistic calculations were performed in GraphPad Prism 5 (GraphPad Software, La Jolla, CA, USA). Extreme outliers exceeding >3 SD from the mean of each group were considered methodological errors and were removed from statistical comparisons.

2.3. **SNP genotyping**

Genotyping of 44 top-associated SNPs from the Immunochip project was performed with a Fluidigm Biomark dynamic array (48.48) and SNPtype assays (Fluidigm Corp.) in 26 samples with expression results in which DNA was available. Eight samples were already genotyped in the Immunochip sample set and were used as quality control for the new genotyping. Three samples had to be removed from the study due to failed genotyping, resulting in a total number of 23 samples, 14 controls and 9 celiac patients. The assay design was performed by the Fluidigm Assay Design Group. Seven of the target SNPs did not fulfill the established assay design requirements due to adjacent SNPs within 20–30 bases on each side of the target SNP, GC content > 65% or triallelic SNPs. After an in-deep analysis of those seven SNPs, taking into account the allelic frequencies of the target SNP and the adjacent SNPs and the frequency of each allele in the case of the unique triallelic SNP (rs61907765) in Ensembl, we decided to omit this obstacle in the design of six SNPs and to remove the SNP rs60215663 from the analysis due to smaller minor-allele frequency than adjacent SNPs.

2.4. **Coexpression analysis**

Merlin 1.1.2 software was used to test association between SNP genotype and candidate gene expression [88].The association was tested independently in each of the studied groups in order to avoid false associations due to duplicated genotypes in CD sample pairs.

3. Results

3.1. Differentially expressed genes in CD

Fifteen out of the forty-five genes analyzed were differentially expressed when comparing the fold change between active disease samples and non-celiac controls. Nine of the genes were significantly overexpressed in active CD (*CTLA4, ICOS, CIITA, FASLG, PLEK, PVT1, CD28, UBASH3A* and *SOCS1*), whereas the other six genes (*ATXN2, ICOSLG, ARHGAP31, ZFP36L1, CCR2* and *TREH*) were downregulated (Figure 8-A). As could be expected due to the aprioristic selection of the candidate genes, GO-term analysis of the altered genes showed enrichment of immune response related processes such as regulation of T cells, lymphocyte and leukocyte activation and proliferation, lymphocyte costimulation and so on. The most relevant genes behind this enrichment are *ICOSLG* (inducible T-cell costimulator ligand); *CCR2* (chemokine (C–C motif) receptor 2), a receptor for a chemokine which specifically mediates monocyte chemotaxis and is involved in monocyte infiltration in inflammatory diseases; *PLEK* (pleckstrin); *CTLA4* (cytotoxic T-lymphocyte-associated protein 4), a member of the immunoglobulin superfamily that encodes a protein which transmits an inhibitory signal to T cells; CD28, an essential protein for T-cell proliferation and survival, cytokine production and T-helper type-2 development and *ICOS* (inducible T-cell co-stimulator), which also belongs to the CD28 and CTLA4 cell-surface receptor family and has an important role in cell-cell signaling, immune response and regulation of cell proliferation. *ICOS*, *CD28* and *CTLA4* are located on the CELIAC3 locus, a well-known region that has been linked to several autoimmune disorders, including CD, originally identified by Holopainen et al [100] and that has been replicated several times in posterior studies. When treated patients and non-celiac controls were compared, only three genes showed significant expression differences (*ATXN2*, *CCR2* and *CCR4*), being constitutively downregulated in the disease group (Figure 8-B).

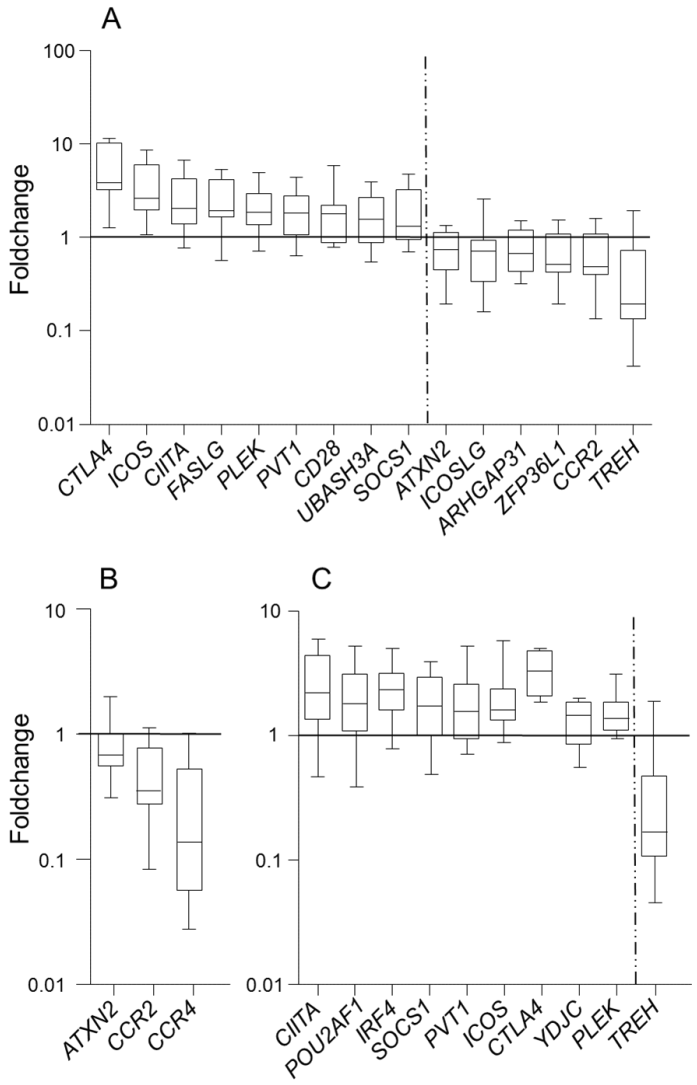


Figure 8: Expression fold change of differentially expressed genes. (a) Active CD vs controls, (b) treated CD vs controls and (c) active vs treated CD.

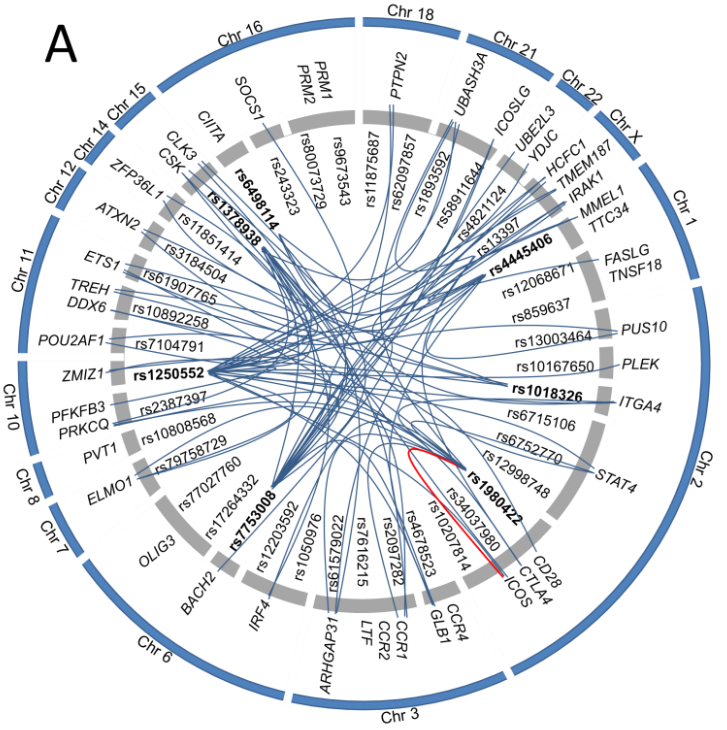
The comparison between active and treated disease mucosa identified differential expression in ten genes, nine of which were upregulated in the active disease (*CIITA, POU2AF1, IRF4, SOCS1, PVT1, ICOS, CTLA4, YDJC* and *PLEK*) and one was downregulated (*TREH*) (Figure 8-C). As in the case of active disease vs controls, the enriched GO terms are

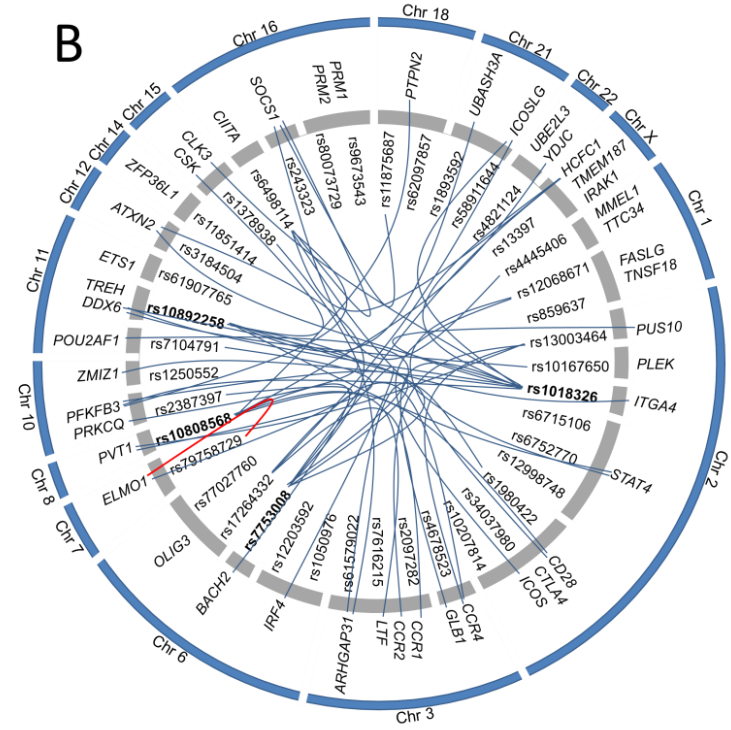
related to the regulation of immune cell activation, due to the altered expression of *CTLA4*, *ICOS* and *PLEK* as previously, plus *IRF4* (interferon regulatory factor 4), an important transcription factor in the regulation of interferon in response to infection by viruses, which is lymphocyte specific and negatively regulates TLR signaling, a pathway that is central to the activation of innate immune system. Apart from that, GO terms related to interferon-gamma response are also enriched in this case, due to three genes that are upregulated in the active disease attributable to the inflammatory process, *CIITA* (class II MHC transactivator), *IRF4* and *SOCS1* (suppressor of cytokine signaling 1).

3.2. **Genotype effect in gene expression**

Despite the limited number of biological samples in our study, we also searched for relationships between SNP genotypes and gene expression levels. We were able to include 14 individuals from the control group and 9 sample pairs from the disease group, for whom both genotypes and expression results were available. For this reason, it was often impossible to have all three genotypes present in every group; heterozygous and minor-allele homozygous samples were combined in order to increase statistical power.

We detected genotype effects of a number of SNPs on the expression of several genes, but surprisingly, the effect seemed to be stimulus dependent, as it was different among the groups. Moreover, most eQTLs were in *trans* and only four candidate genes located under the association peak were influenced by its putative regulatory SNP; rs1980422-*ICOS* in debuts, rs79758729-*ELMO1* in treated patients, rs12068671-*TNSF18* and rs13397-*TMEM187* in controls (Figure 9). In an attempt to explain this result, we scrutinized the genomic region around each associated SNP in search for putative regulatory elements that could be altering the expression of genes in *trans*. We conducted searches in different databases available online, such as Haploreg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) [101], Ensembl (http://www.ensembl.org) [102] and the UCSC Genome browser (http://genome.ucsc.edu) [103]. As expected, elements affected by the potentially regulatory SNPs included open chromatin regions, novel protein-coding sequences, processed antisense transcripts, pseudogenes, microRNAs, novel lincRNAs and altered protein-binding motifs. This finding opens the door for further studies in order to determine whether any of those sequences could have a real functional role in gene regulation and development of CD.


A
Chr 18
Chr 21
Chr 22
Chr X
Chr 1
Chr 2
Chr 3
Chr 6
Chr 7
Chr 8
Chr 10
Chr 11
Chr 12
Chr 14
Chr 15
Chr 16
PRM1
PRM2
SOCS1
PTPN2
UBASH3A
ICOSLG
UBE2L3
YDJC
HCFC1
TMEM187
IRAK1
MMEL1
TTC34
FASLG
TNSF18
PUS10
PLEK
ITGA4
STAT4
CD28
CTLA4
ICOS
CCR4
GLB1
CCR1
CCR2
LTF
ARHGAP31
IRF4
BACH2
OLIG3
ELMO1
PVT1
PRKCQ
PFKFB3
ZMIZ1
POU2AF1
TREH
DDX6
ETS1
ATXN2
ZFP36L1
CSK
CLK3
CIITA
rs6498114
rs1378938
rs11851414
rs3184504
rs61907765
rs10892258
rs7104791
rs1250552
rs2387397
rs10808568
rs79758729
rs77027760
rs17264332
rs7753008
rs12203592
rs1050976
rs61579022
rs7616215
rs2097282
rs4678523
rs11875687
rs62097857
rs1893592
rs58911644
rs4821124
rs13397
rs4445406
rs12068671
rs859637
rs13003464
rs10167650
rs1018326
rs6715106
rs6752770
rs12998748
rs1980422
rs34037980
rs10207814


B
Chr 18
Chr 21
Chr 22
Chr X
Chr 1
Chr 2
Chr 3
Chr 6
Chr 7
Chr 8
Chr 10
Chr 11
Chr 12
Chr 14
Chr 15
Chr 16
PRM1
PRM2
SOCS1
PTPN2
UBASH3A
ICOSLG
UBE2L3
YDJC
HCFC1
TMEM187
IRAK1
MMEL1
TTC34
FASLG
TNSF18
PUS10
PLEK
ITGA4
STAT4
CD28
CTLA4
ICOS
CCR4
GLB1
CCR1
CCR2
LTF
ARHGAP31
IRF4
BACH2
OLIG3
ELMO1
PVT1
PRKCQ
PFKFB3
ZMIZ1
POU2AF1
TREH
DDX6
ETS1
ATXN2
ZFP36L1
CSK
CLK3
CIITA
rs6498114
rs1378938
rs11851414
rs3184504
rs61907765
rs10892258
rs7104791
rs1250552
rs2387397
rs10808568
rs79758729
rs77027760
rs17264332
rs7753008
rs12203592
rs1050976
rs61579022
rs7616215
rs2097282
rs4678523
rs11875687
rs62097857
rs1893592
rs58911644
rs4821124
rs13397
rs4445406
rs12068671
rs859637
rs13003464
rs10167650
rs1018326
rs6715106
rs6752770
rs12998748
rs1980422
rs34037980
rs10207814

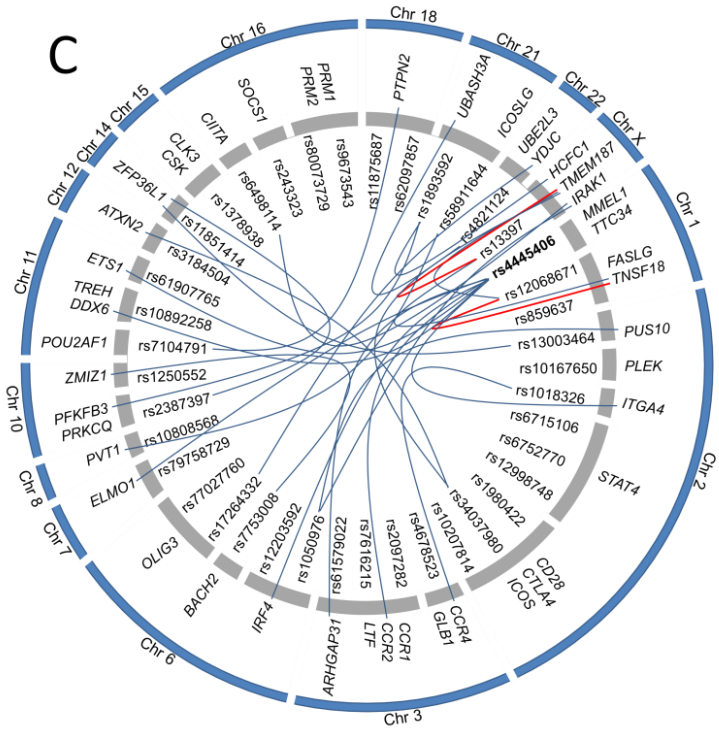


Figure 9: SNP genotype effect on candidate gene expression for the different disease statuses. (a) At diagnosis, (b) > 2 years GFD and (c) controls. Statistical analyses performed with Merlin 1.1.2 software; p value < 0.005 was fixed as significant SNP effect. SNPs with an effect on multiple genes are shown in bold. Blue lines indicate trans-eQTLs and red lines cis-eQTLs.

3.3. **Coexpressed gene patterns in CD**

Coexpression analyses were performed to identify possible common regulation signatures that could be altered in celiac patients on account of inflammation or owing to genetic determinants. Interestingly, we observed different correlation patterns among genes in the three study groups, from higher to lower coexpression levels in gluten-consuming celiac patients at diagnosis, treated patients and non-celiac controls, respectively (Figure 10). The selection of those genes that were coexpressed in both groups of patients, but not in non-celiac controls, identified a subset of 18 genes that were tightly correlated in patients that seemed to be putatively under the control of three SNPs (Figure 11).

Figure 10: Gene pair coexpression matrixes for the different disease statuses. Each small square represents the p value for the correlation of the expression level in a specific gene pair. Red, dark pink, light pink and white indicate Pearson's correlation p values of $p<0.0001$, $p<0.001$, $p<0.01$ and $p>0.01$ respectively.

One of those SNPs, rs1018326, is located on chromosome two, in an intergenic region between *UBE2E3* and *ITGA4*, on top of a known lincRNA (AC104820.2) whose function has not been described yet. This RNA gene has five transcripts (spliced variants), ranging from 342 to 1,771 base pairs length. The expression of AC104820.2 was significantly altered between biopsy pairs from the same patients in different stages of the disease, being upregulated in active biopsies (Figure 12). We did not observe these differences when comparing unpaired biopsies from independent active and treated CD patients, stressing the enormous variability among CD patients and the need for strict sample pairing for efficient comparisons.

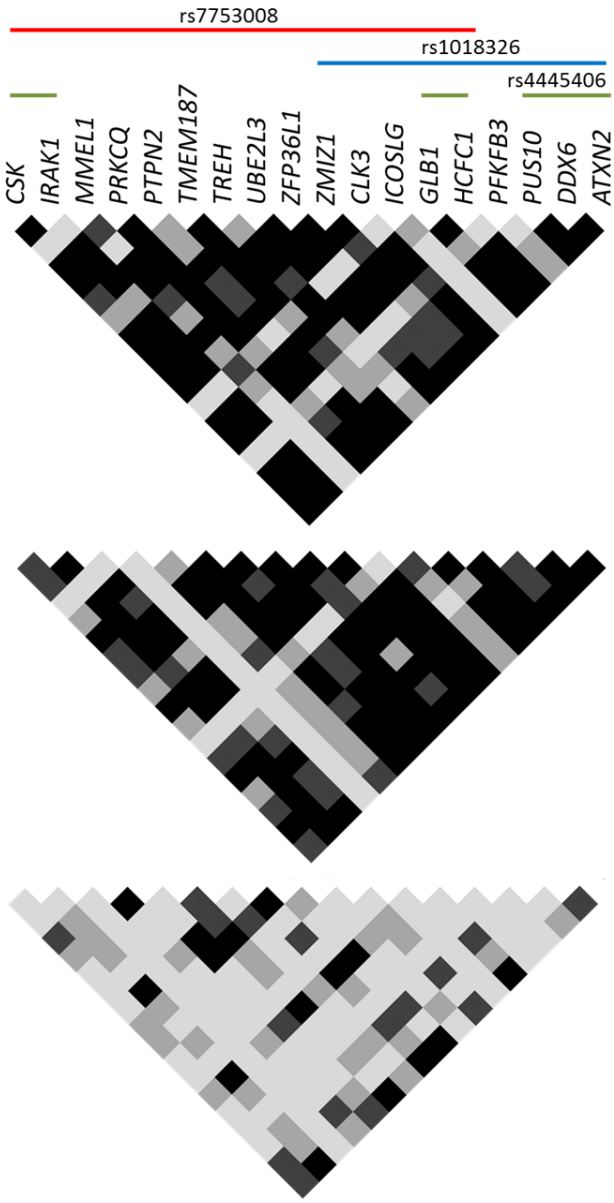


Figure 11: Gene pair coexpression matrixes for the different disease statuses on a subset of genes correlated in patients but not in controls. (a) At diagnosis, (b) 4 2 years GFD and (c) controls. Each small square represents the P-value for the correlation of the expression level in a specific gene pair. Black, dark gray, light gray and white indicate Pearson's correlation P-value of $p < 0.0001$, $p < 0.001$, $p < 0.01$ and $p > 0.01$, respectively. SNPs with trans-eQTLs for those genes are shown.

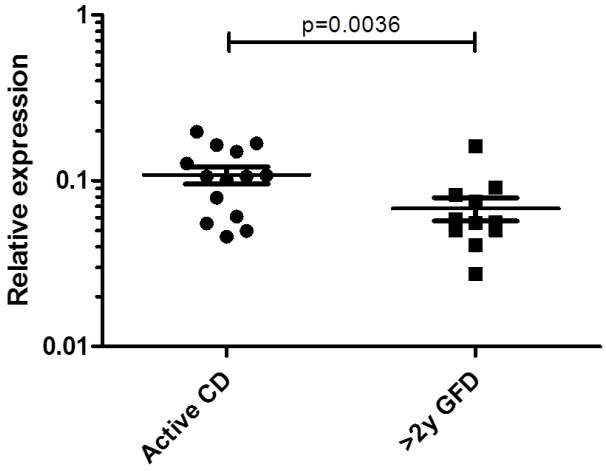


Figure 12: Relative expression of AC104820.2 lincRNA in a set of 11 biopsy pairs. Paired t-test was applied for statistical analysis.

4. **Discussion**

Candidate gene selection following large-scale SNP association studies is often aprioristic and greatly influenced by the current knowledge of the pathogenic mechanisms that are thought to be involved in the disease, but functional studies are the only unbiased approach to identify real functional players. Until now, only a small number of studies have performed deep analyses of associated regions prior to proposing candidate-susceptibility genes: a genetic and functional analysis of *THEMIS* and *PTPRK*, the two candidate genes located on the CD association peak chr6: 127.99-128.38 Mb found a significant correlation between the expression levels of both genes in CD patients that was absent in the control group [84]. Although this finding could suggest a possible role for both of the genes, it shows the existence of a common regulatory relationship that could reside in the noncoding albeit functional intergenic region. Using a different approach, fine mapping of the *LPP* locus to identify possible functional variants revealed six SNPs that overlap regulatory sites, with rs4686484 having a possible effect on *LPP* gene expression in CD patients [104]. Finally, Östensson M. *et al* [105] recently performed pathway analyses and two-locus interaction studies to further investigate association signals. They found some differentially expressed genes in the small intestine mucosa from CD patients, and identified susceptibility genes from top-scoring regions that could be gathered into several categories. They suggested that those genes and pathways together could reveal a new potential biological mechanism that could influence the genesis of CD and other chronic inflammatory disorders.

Although the effects of associated SNPs on gene expression has been previously studied in CD, and several *cis*- and *trans*-eQTLs have been found, expression data have always been obtained from peripheral blood samples [105]. A recently published work analyzed the effect of regulatory variants upon monocyte activation and concludes that a significant proportion of variants may show activity only in a context specific manner, proposing that only considering the genetic, cellular and environmental context relevant to the disease will make it possible to resolve functional genetic variants more extensively [106]. This is the case of the results obtained in the present study, where we are able to distinguish different eQTLs determined by prolonged gliadin insult and inflammation that are present in CD patients at diagnosis. Furthermore into the complexity of the mechanisms involved in the functional translation of associated genetic variants, we observe that SNP rs1018326 seems to exert its effect through a lincRNA, AC104820.2, for whom disease-related expression changes are only evident when biopsies from the same patients taken at different stages of the disease are compared.

Concerning the coexpression patterns identified, the higher correlation observed in patients could suggest a coordinated alteration that again points towards complex regulatory mechanisms. In the case of active disease, a higher degree of correlation could be explained as a consequence of the inflammatory milieu provoked by the ingestion of gliadin, taking into account that all the candidate genes proposed are related to the immune response. However, this correlation is maintained in the group of treated patients, even after gliadin withdrawal from the diet for >2 years, suggesting at least two possible explanations: an intrinsic, constitutive alteration of those genes in celiac individuals that is independent from gliadin ingestion or a response pattern caused by the gliadin insult, that is not reversible after 2 years on GFD and probably requires a longer time to reach basal expression levels.

An opposite coexpression scenario has recently been described by our group in the case of the NFκB pathway in CD [107]. In that case, the strongest correlation was found in the control group, suggesting a very tight regulatory control of the pathway in a healthy gut, and an alteration of this pathway in the disease. These opposed results make sense if we take into account that NFκB coexpression is indeed expected to be the normal situation because genes that are part of the same pathway are expected to be under the same regulatory mechanisms. In the case of the disease-associated loci, even though enriched in immune-related genes, they would not be expected to react in a coordinated manner upon an environmental challenge unless they are related to the same regulatory variation. In this work we are analyzing the expression of many candidate genes that have in common the implication in the immune response, which is altered in CD, so the coordinated alteration of those genes could be understood.

The idea put forward in the present study needs robust experimental confirmation to be proven, and there are still many pieces to be put together in the puzzle of the common disease genetic susceptibility. However, it is clear that the effects of associated variants go far beyond the over simplistic idea of transcriptional control at a nearby locus. The complex interactions that maintain a coordinated, healthy response to an environmental challenge are written on our genome and the disruption of those subtle fine-tuning mechanisms emerge as the initial cause of a series of events that eventually lead to disease.

Chapter 3:

Immunochip candidate genes study in CD intestinal cell populations

1. **Introduction**

Celiac disease (CD) is a chronic immune-mediated disorder caused by intolerance to ingested gluten that develops in genetically susceptible individuals. Typical histological findings in active CD comprise villous atrophy, crypt hyperplasia and lymphocytic infiltration of the small intestinal mucosa, and the only effective treatment is strict lifelong gluten-free diet (GFD)[62.] The main genetic contributor to CD is the MHC region on chromosome 6p21 and has been estimated to be responsible of 40% of the genetic susceptibility to CD. However, risk HLA variants (HLA-DQ2/-DQ8) are also common in general population making necessary the contribution of other loci to genetic predisposition of developing CD.

GWAS have been the strategy to discover the associated genomic regions behind the risk of developing CD (among other diseases). Results from GWAS and the more recent Immunochip project revealed 39 CD associated regions outside HLA [71,79,86].

Population specific differences became evident every time an association study tries to be replicated in an independent population [80-83,99]. The main reason behind this matter could be the genetic background of the samples used in those big association studies such as the Immunochip project, where in the case of CD, more than half of the samples (7.728/12.041 cases and 8.274/12.228 controls) where from the UK collection [79]. Therefore, new approaches in the analysis of the dense genotyping data available that correct for the population specific differences could generate distinct results with new association signals and candidate genes or regulatory elements.

Overall, 66 candidate genes have been proposed based on their location under association peaks, and several functional studies have been carried out, mainly in peripheral blood mononuclear cells (PBMC) and a few in the disease target tissue, the intestinal epithelia [84,99,108].

Intestinal biopsy samples are a heterogeneous mixture of cells that could give false positive or negative results when analyzing candidate genes that could have differential expression in different cell types. The principal cell populations of intestinal biopsies are enterocytes and cells from the immune system, mainly intraepithelial lymphocytes (IELs)[109]. Enterocytes are characterized for being CD326 positive cells that express EpCAM in their surface, a molecule that is involved in cell signaling, migration, proliferation and differentiation. Otherwise, immune cells express the CD45 antigen on their surface, a protein encoded by the *PTPRC* gene that is a member of the protein tyrosine

phosphatase family. These characteristics make it possible to separate both cell populations from a biopsy sample, and allow the independent study of both cell types in subsequent analysis and compare those results with previously published biopsy functional studies [108].

In the present work, taking into account that association results are not homogenous among populations, we first analyzed the genotype data available from the immunochip project in the Spanish origin sample set from the CEGEC, aiming to identify new associated regions that did not show association after the analysis of the whole sample. More importantly, we also wanted to question the implication of the proposed candidate genes in disease development taking into account the cell populations in the intestinal epithelia, a novel approach never used before. In this way, we were able to identify discordant expression levels in both cell populations.

2. Material and methods

2.1. Patients and samples

CD was diagnosed according to the ESPGHAN (European Society of Pediatric Gastroenterology Hepatology and Nutrition) criteria in force at the time of recruitment, including anti-gliadin (AGA), anti-endomysium (EMA) and anti-transglutaminase antibody (TGA) determinations as well as a confirmatory small bowel biopsy. The study was approved by the Institutional Boards (Cruces University Hospital code CEIC-E09/10 and Basque Clinical Trials and Ethics Committee code PI2013072) and analyses were performed after informed consent was obtained from all subjects or their parents. Biopsy specimens from the distal duodenum of each patient were obtained during routine diagnosis endoscopy, one sample was used for diagnostic pathology examination and the other in the present investigation. Eight Intestinal biopsies from CD children at the time of diagnosis were compared with 8 tissue samples from non-celiac controls with no inflammation at the time of endoscopy. Blood samples from 9 celiac patients were also collected for PBMC isolation for functional studies.

2.2. Cell population separation from intestinal biopsies

To separate epithelial and immune cells from biopsy samples, MACS magnetic cell separation technology was used (Miltenyi Biotec), following manufacturer´s protocol for CD45 MicroBeads and MS separation columns. Briefly, biopsy samples were collected and processed freshly. Cells were mechanically separated by agitation in complete medium (RPMI + antibiotics + FBS + DTT + EDTA) during 1 hour. Cells released to the media were collected by centrifugation after filtering the media with pre-separation filters (20 µm) and the remaining *lamina propria* was stored in RLT buffer at -80ºC.

Dead cell removal kit was used in a first purification step to prepare a viable single-cell homogenous suspension. In a second purification step, cells were labeled with CD45 magnetic microbeads to separate the cell mixture in two fractions, the CD45 positive immune cells (attached to the column) and CD45 negative enterocytes (flow through). Both fractions were stored in RLT buffer at -80ºC for a posterior nucleic acid extraction.

A Cytometry analysis of a small portion of both cell populations was carried out to probe the good results of the separation process.

2.3. **Peripheral Blood mononuclear cell isolation**

Blood samples from CD patients were freshly collected and separation of PBMCs was accomplished through density gradient centrifugation using Ficoll. After the centrifugation step, Ficoll separates layers of blood, with lymphocytes and monocytes under a layer of plasma. PBMC layer was carefully taken with a Pasteur pipette and after 2 washing steps with RPMI medium, PBMCs were cryopreserved at -80 in freezing medium until nucleic acid extraction.

2.4. **RNA samples and gene expression**

RNA extracted with the NucleoSpin microRNA kit (Macherey-Nagel, Düren, Germany) following manufacturer´s instructions was normalized to 8 ng/µl and converted to cDNA using the AffinityScript cDNA Synthesis kit (Agilent Technologies, Santa Clara, CA, USA) following manufacturer's protocol. Gene expression analyses were performed using Fluidigm Biomark 48.48 dynamic arrays (Fluidigm Corp., South San Francisco, CA, USA) and commercially available TaqMan Gene Expression assays. Housekeeping gene *RPLPO* was simultaneously quantified and used as an endogenous control of input RNA (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA). Relative expression in each sample was calculated using the accurate Ct method [34] and normalized to the average expression value of all the analyzed samples.

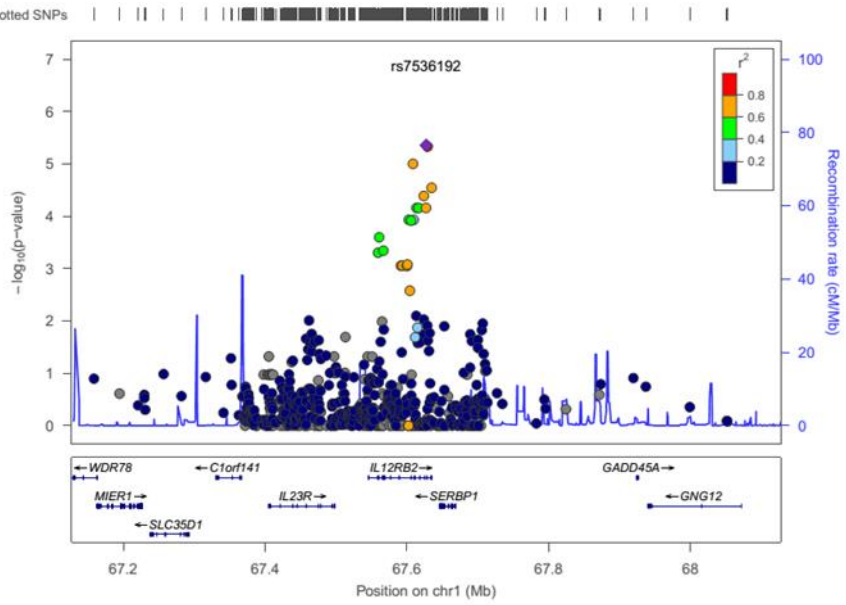
2.5. **Association study**

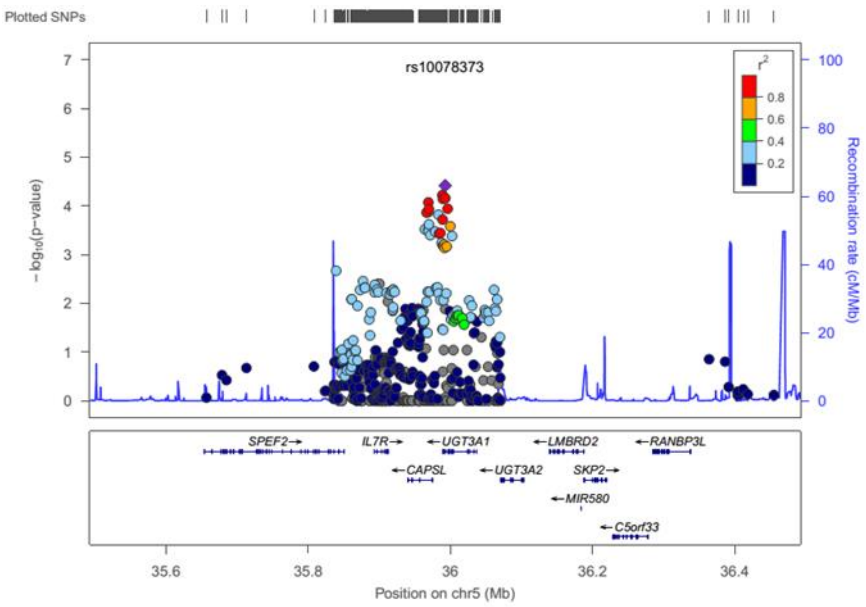
The CEGEC sample set genotyped in the Immunochip project consists of 545 celiac patients and 308 non-CD controls [79]. Association analysis of this sample set alone was performed using PLINK v1.07 [110], and the top 50 SNPs were selected. LD blocks with at least 4 associated SNPs were mapped using LocusZoom [87] for new candidate gene selection.

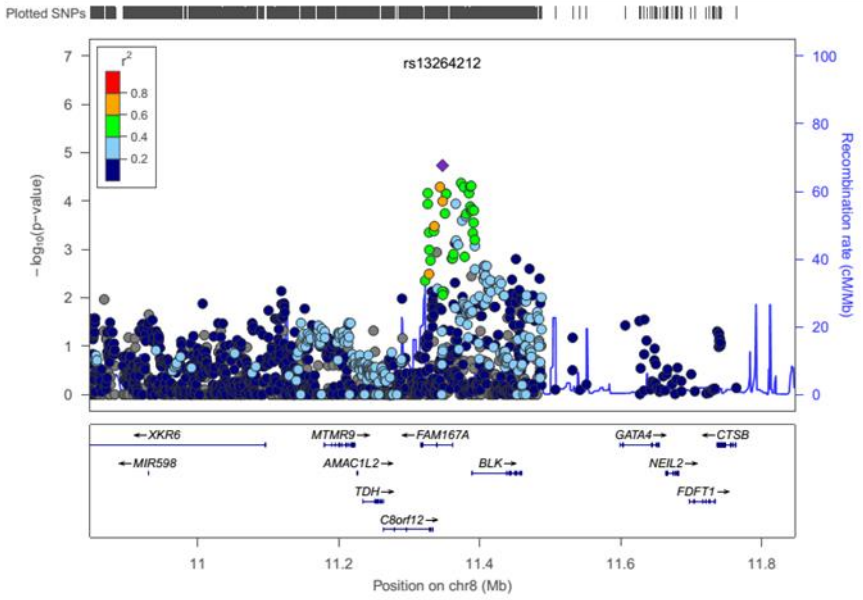
3. **Results**

The CEGEC sample set genotype data from the immunochip project was analyzed on its own, aiming to identify candidate genes hidden because of possible population differences. Top 50 most associated SNPs in CEGEC Spanish population were taken, with p values ranging from 7.90 x 10^{-5} to 1.53 x 10^{-12} (Supplementary table 2), Chromosome 6 was excluded from this analysis.

Interestingly, we were able to identify 5 regions with more than one associated SNP, which reinforce the possibility of being true real signals and not an artefact. The signals localized to dense genotyped regions are shown in Figure 13.




Plotted SNPs
rs10078373
r²
0.8
0.6
0.4
0.2
− log10(p−value)
Recombination rate (cM/Mb)
SPEF2
IL7R
UGT3A1
LMBRD2
RANBP3L
CAPSL
UGT3A2
SKP2
MIR580
C5orf33
Position on chr5 (Mb)


Plotted SNPs
rs13264212
r²
0.8
0.6
0.4
0.2
− log10(p−value)
Recombination rate (cM/Mb)
XKR6
MIR598
MTMR9
FAM167A
AMAC1L2
BLK
TDH
C8orf12
GATA4
CTSB
NEIL2
FDFT1
Position on chr8 (Mb)

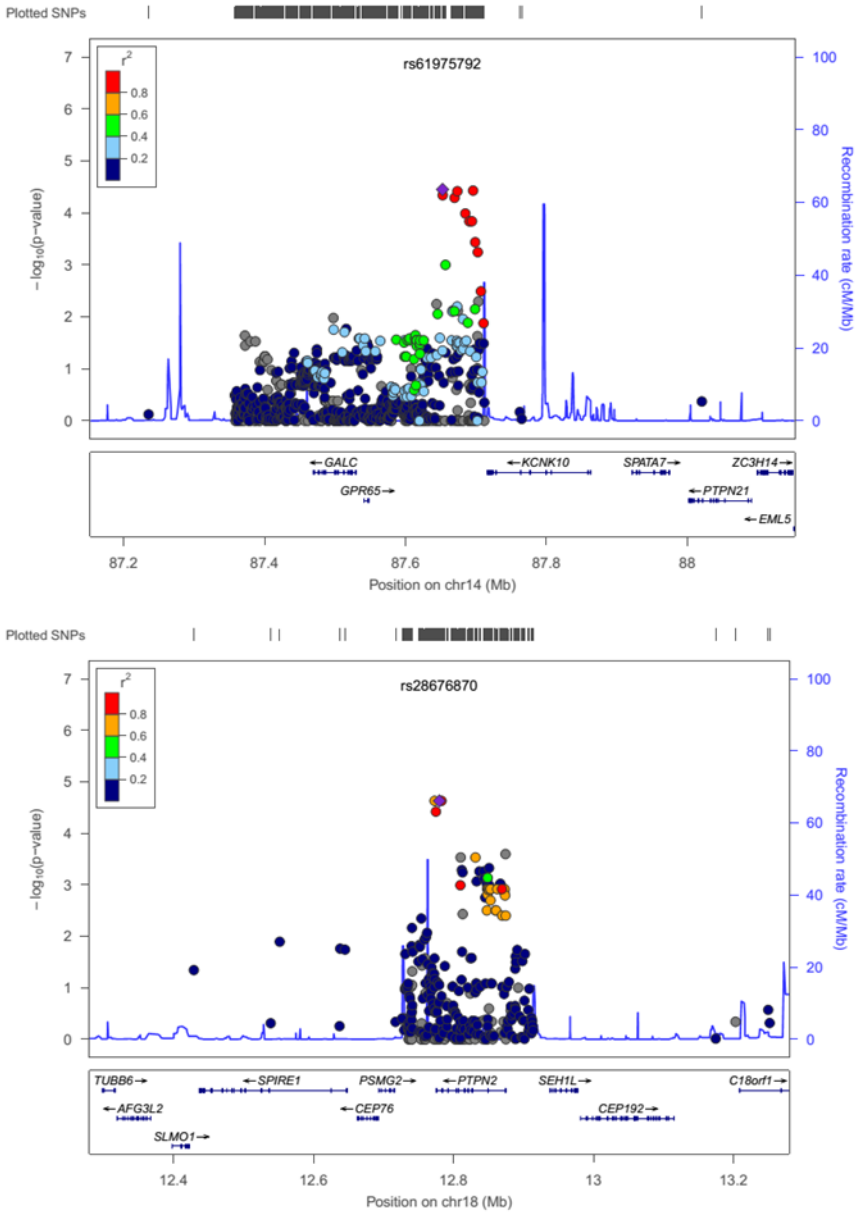


Figure 13: Locus Zoom plots for the 5 associated regions in CEGEC sample set. At each locus, the first signal is shown as a purple diamond and SNPs in LD are shown according to the r^2 value

These association signals suggest new candidate genes that localize to these regions. *IL23R*, *IL12RB2* and *SERBP1* in chromosome 1 (67353404 to 67390948, rs7536192 lowest p value= 4.38 x 10^{-06}), *IL7R*, *CAPSL*, *UGT3A1* and *UGT3A2* in chromosome 5 (35930601 to 35962030, rs10078373 lowest p value= 3.78 x10^{-05}), *FAM167A*, *C8orf12* and *BLK* in chromosome 8 (11433457 to 11457539, rs13264212 lowest p value= 1.82 x 10^{-05}), *GALC* and *GPR65* in chromosome 14 (88099567 to 88154762, rs61975792 lowest p value= 3.53 x 10^{-05}), and *PTPN2* and *PSMG2* in chromosome 18 (12777265 to 12804934, rs28676870 lowest p value= 2.33 x 10^{-05}).

From those genes, *PTPN2* is the only one that has been already proposed as a candidate gene from the immunochip data, with 2 SNPs showing independent signal, rs11875687 and rs62097858. Surprisingly, the p values for those SNPs in CEGEC sample set are 0.000856 and 0.999 respectively, and none of them is in LD with the most associated SNP (rs28676870) in our population (Figure 14), suggesting different haplotypes that could be population-specific. *IL23R* have also shown some association with CD in previous studies but with contradictory results in different populations [111-113], but no association is found in the immunochip.

For the subsequent functional analyses in this study, we added the new candidate genes associated in the CEGEC sample set to those previously identify by the immunochip project. Candidate genes proposed after Immunochip results in CD were studied by our group in a previous publication [108]. Our functional study revealed a great number of genes being differentially expressed in intestinal biopsy samples from celiac patients (active and treated disease) and controls.

From the 68 genes studied 4 were excluded from further analysis due to a failure of the assay in the expression study (*TTC34*, *BACH2*, *OLIG3*, *PRM2*). The remaining genes showed differences among sample type that will be described.

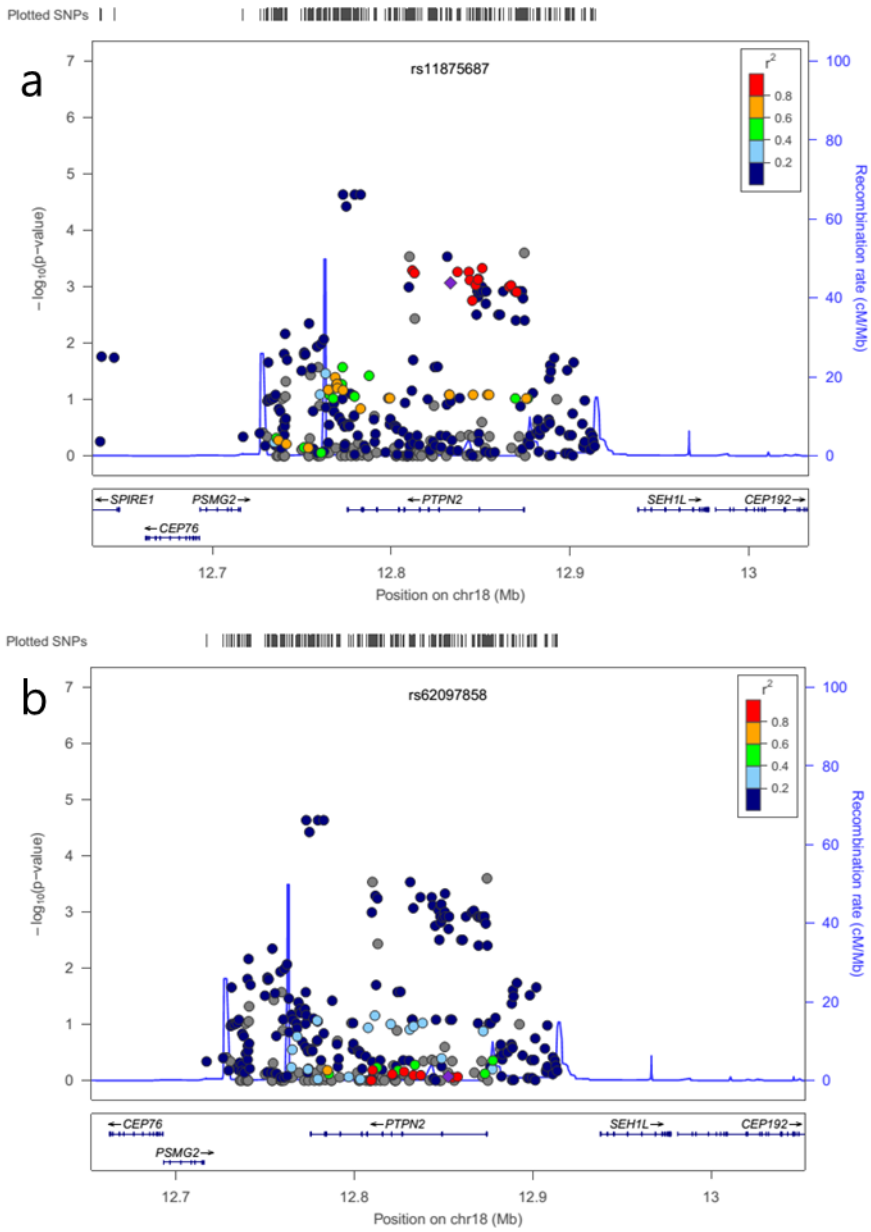


Figure 14: *PTPN2* region in detail. a) Immunochip most associated SNPs in b) Spanish CEGEC population.

In the epithelial cells (CD326+) 16 genes (*IL12RB2, FASLG, RGS1, PLEK, UBE2E2, LTF, IL12A, IRF4, TAGAP, PRKCQ, POU2AF1, ETS1, SH2B3, GPR65, CIITA* and *SOCS1*) showed differential expression ($p<0.01$) between active celiac patients and controls. All of them were upregulated in the disease samples (Figure 15). GO-term analysis of the altered genes showed enrichment of alpha-beta T cell activation, with *PRKCQ*, *IL12A* and *IRF4* as pivotal genes in this process. A subset of 7 genes (*IL12RB2, FASLG, LTF, IL12A, PRKCQ, ETS1* and *SOCS1*) showed pathway interaction in GeneMANIA prediction server [114]. KEGG pathway analysis of those genes situated them in Jak-STAT signaling pathway, cytokine-cytokine receptor interaction and inflammatory bowel disease (IBD), an autoimmune disease that shares many association loci with CD [115,116].

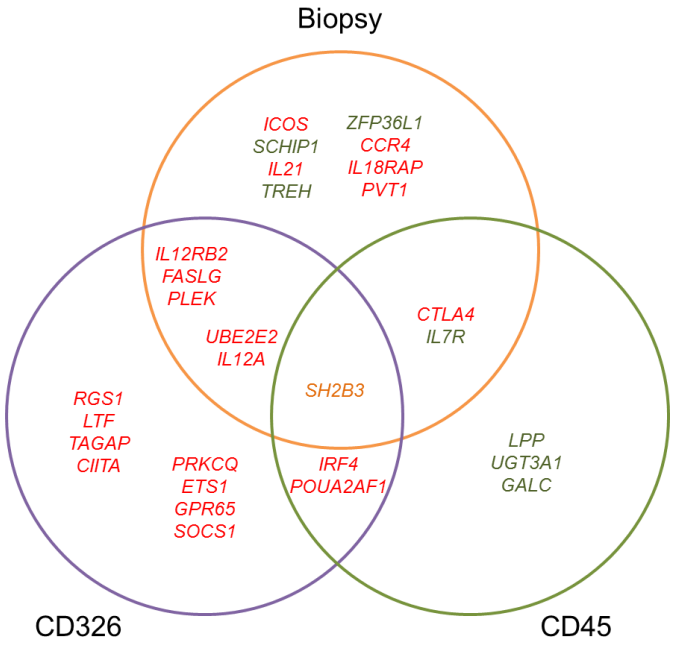


Figure 15: Differentially expressed genes among groups. Genes upregulated in celiac patients are shown in red, genes downregulated in celiac patients are shown in green, genes with different expression tendency in associated sample type are shown in orange. On the other hand, the immune cell samples (CD45+) showed altered expression in 8 of the studied genes. Five of those genes (*LPP, IL7R, UGT3A1, SH2B3* and *GALC*) were downregulated in active disease compared to controls, and the remaining 3 genes (*CTLA4, IRF4* and *POU2AF1*) were upregulated. GO term analysis did not return any significant process.

Interestingly, 3 genes had shared alteration in both cell types (*IRF4, POU2AF1* and *SH2B3*), but surprisingly *SH2B3* showed the opposite effect in CD326+ and CD45+ cells, being upregulated in epithelial cells and downregulated in immune cells in the disease samples.

After comparing the new results in this study with the results of the previous work in whole biopsies [99,108], differences among the 3 sample types arises, with only 1 gene (*SH2B3*) being differentially expressed in the 3 sample types, 5 genes that shared differences between epithelial cells and biopsies *(IL12RB2, FASLG, PLEK, UBE2E2* and *IL12A*), and 2 genes (*CTLA4* and *IL7R*) shared between IELs and biopsies. There were 8 more genes that only showed expression differences in biopsy samples *(IL18RAP, ICOS, CCR4, SCHIP1, IL21, PVT1, TREH* and *ZFP36L1*).

Comparing the results from each sample type, we were able to identify some altered genes that are specific from a unique cell type, genes in which we won't be able to notice differential expression when analyzing whole biopsies. Eight genes were differentially expressed ($p<0.01$) between cases and control exclusively in CD326 cell population (RGS1, *LTF, IRF4, TAGAP, PRKCQ, POU2AF1, ETS1* and *GPR65*), from which *GPR65* is a candidate gene identified in the CEGEC sample set. All the genes altered in CD326 cell population were upregulated in active disease compared to controls (Figure 16).

On the other hand, 5 genes showed differential expression ($p<0.01$) exclusive from CD45 cell population (*GALC, IRF4, LPP, POU2AF1* and *UGT3A1*). From those genes, *GALC* and *UGT3A1* were identified in CEGEG sample set. *GALC, LPP* and *UGT3A1* were downregulated in active disease compared to controls, while *IRF4* and *POU2AF1* were overexpressed in active celiac patients (Figure 16).

Interestingly, two genes in which differential expression was identified thanks to the new approach of separating cell populations from whole biopsies were differentially expressed in both cell types but not in whole biopsies. Both genes showed the same alteration tendency in CD326 and CD45 cells, being upregulated in active disease samples.

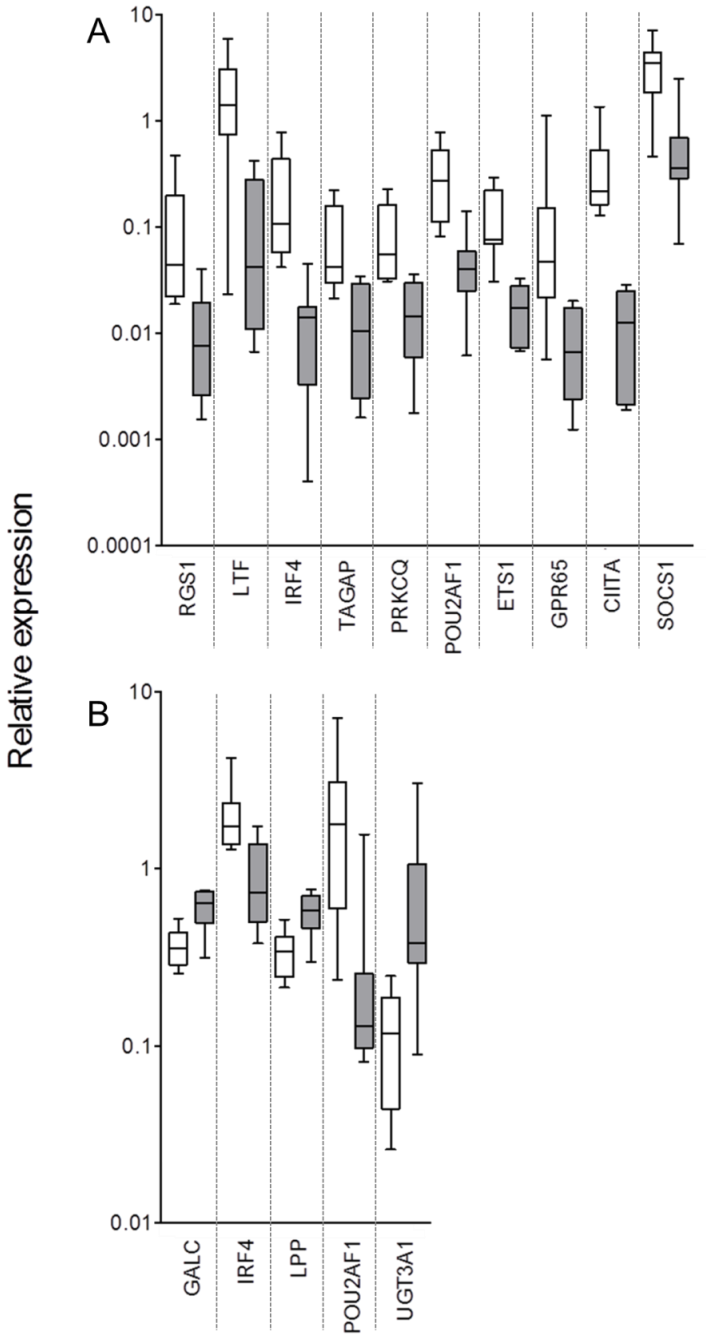


Figure 16: Relative expression of genes altered exclusively in cell populations. A: CD326+ cell population. B: CD45+ cell population. Graph shows median and 10-90 percentiles.

In addition, we analyzed gene expression in PBMCs from 9 active celiac patients and compared the results to gene expression levels in debut biopsies. Eighteen out of 52 analyzed immunochip candidate genes showed differential expression between PBMCs and whole biopsies. From those genes, 8 of them (*FASLG, IL18RAP, CTLA4, CCR4, IL12A, SH2B3, SOCS1, ICOSLG*) have been identified previously by our group as disease altered genes [99,108].

Coexpression analyses were performed to identify possible cell specific regulation signatures that could be altered due to disease status or owing to genetic determinants. CD45+ cell population seems to have a higher correlation of some of the studied genes in celiac patients when compared to expression patterns in controls. We identified a subset of 16 genes that were tightly correlated in CD45+ cells in active CD patients but not in controls (Figure 17).

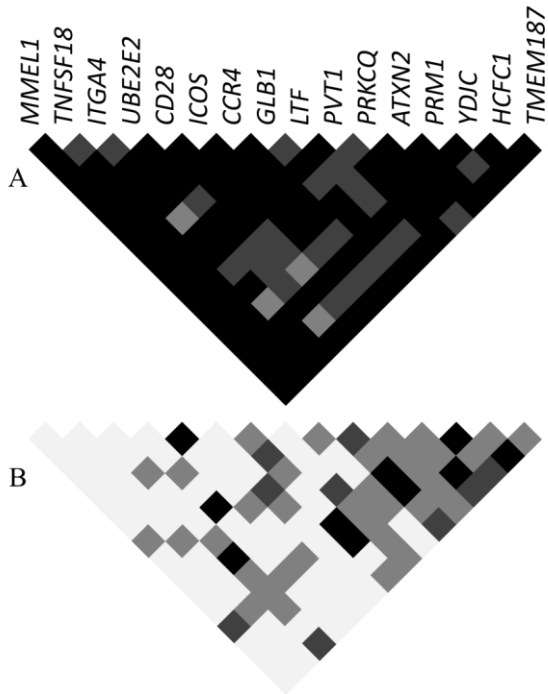


Figure 17: Gene pair coexpression matrixes on a subset of genes correlated in CD45+ cells in celiac patients but not in controls. A: CD45+ cells in celiac patients B: CD45+ cells in controls. Each small square represents the p-value for the correlation of the expression level in a specific gene pair. Black, dark gray, light gray and white indicate Pearson`s correlation p-value of $p<0.0001$, $p<0.001$, $p<0.01$ and $p>0.01$ respectively.

Go term analysis of the genes coexpressed in CD45+ cells during active CD, in which an inflammatory autoimmune process is going on, uncovers Go terms related to regulation and activation of effector immune cells. The most relevant genes behind this result are *PRKCQ*, *CD28*, *ICOS* and *TNFSF18*.

We were not able to identify any significant coexpression differences between celiac patients and controls in CD326+ epithelial cell samples.

4. **Discussion**

It is already known that sample size is one of the most important issues in GWA studies, and the conclusion is clear: the more samples the better [74]. That´s why the immunochip project tried to collect samples from many countries aiming to analyze the bigger sample set possible. But differences among populations have been found every time an association result obtained in CD GWAS have been tried to replicate [80-83,99].

Once again, we notices this population differences and were able to discover new associated regions analyzing the CEGEC sample set on its own, which indicates that the genetic background of the samples is important and should be taken into account when analyzing association results. Recently, our group performed a novel and very promising approach to analyze immunochip data, by stratifying individuals into groups taking into account their immunogenetic ancestry to scrutinize each of the group aiming to find novel genetic elements related to CD (article pending revision).

To probe the functional implication of the proposed candidate genes is compulsory to perform gene expression analysis in the damaged tissue, especially when the studied disease is known to be tissue specific, as is the case of CD. Till now, most of the functional studies have been performed in PBMCs and a few in intestinal biopsies of celiac patients and controls. But this could not be enough if we take into account the composition of the intestinal epithelia, in which epithelial cells are the most abundant cell type followed by the infiltrated immune cells, mostly lymphocytes [109].

Therefore, we proposed and carried out a new approach for functional studies by separating epithelial and immune cells from intestinal biopsies to analyze the contribution of each cell type in the development of CD.

We observed that most cell expression changes (16 out of 63 studied genes) are affecting epithelial cells, with all the deregulated genes being upregulated in active celiac patients. Moreover, 8 of the 16 genes are exclusively deregulated in CD326+ cells, which make it impossible to identify those gene expression alterations when analyzing whole biopsies. The overexpression of the genes related to the immune response in epithelial cells indicates the importance of the innate immune system in CD, based on the fact that gliadin peptides are able to trigger a non-T-cell-dependent response that could establish the proinflammatory environment necessary for subsequent T-cell activation and tissue destruction [29]. In the case of CD45+ cells, 3 of the 8 altered genes are exclusive of this cell population, and there were two more genes altered both in CD326+ and CD45+ cells.

Additionally, results in biopsies seem to be more similar to epithelial cells than to immune cells, with 6 altered genes shared between biopsies and CD326+ against only 2 genes shared with CD45+ cells. This result could be expected taking into account that the immune cell proportion is by far smaller in biopsies compared to epithelial cells. Moreover, the only gene that share differential expression in three studied sample types showed contrary tendencies, indicating that cell specific responses should be taking into account. Therefore, our new approach of separating both cell types for functional studies could help us to expose functional players in CD that keep hidden in gene expression analysis of whole biopsies.

On the other hand, our results in PBMCs from celiac patients in which we found many genes differentially expressed between PBMCs and biopsies, once again indicate the importance of performing the functional studies of candidate genes in the affected tissue. From our point of view, the expression differences found between PBMCs and biopsies make it difficult to use PBMC studies to analyze markers of CD.

The coexpression pattern identified in immune cells, with a subset of genes tightly correlated in celiac patients but not in controls, is an indicative of the immune response that is going on in the inflamed intestine of active CD patients, a coordinated expression of many genes in response to gliadin insult. Not finding any correlation pattern in epithelial cells, could indicate that the innate immune response of those non immune cells is more chaotic.

Final remarks

Millions of SNPs have been identified thanks to the Human Genome sequencing projects. Some of those SNPs, called tag SNPs, have been used as genetic markers in GWAS and allow the identification of thousands of susceptibility variants for many complex diseases. The two GWAS performed in CD, together with several follow-up studies, revealed a total of 26 non-HLA associated regions [71,76,77]. The most recent large-scale project performed to identify variants associated with CD and other autoimmune diseases is the Immunochip Project, in which a denser genotyping of 186 GWAS loci associated with 12 immune-related diseases identified 13 additional regions associated with CD [79].

Hence, there are a total of 39 non-HLA regions associated with CD, containing 57 independent association signals. Nineteen of those regions pinpoint to a single candidate gene, but only 3 associated SNPs are linked to protein-altering variants located in exonic regions, although some potentially causative genes have been proposed due to the existence of signals close to the 5' or 3' regulatory regions of nearby genes.

Even though most SNPs localize to noncoding intergenic and intronic regions, CD associated variants seem to be located in expression quantitative trait loci or eQTLs, genomic loci that regulate expression levels of mRNAs or proteins. After a meta-analysis of a genome-wide eQTL dataset of 1,469 human whole blood samples, supposed to reflect primary leukocyte gene expression, 38 genome-wide CD associated non-HLA loci were assessed for cis expression-genotype correlation [77]. Twenty significant eQTLs were identified, more than expected by chance, indicating that CD associated regions are greatly enriched for eQTLs. These data may indicate that some risk variants could have an influence in CD susceptibility by altering gene expression, however, there are many evidences indicating that cis-eQTLs differ between different tissues and can even have completely opposite effects.

Hence, it is imperative to perform functional analyses of the proposed candidate genes in the disease tissue, and this was one of the most important goals of the work presented in this thesis. For that purpose, the eight association peaks from the first CD GWAS were replicated in a Spanish population, identifying four genes (*IL12A, LPP, SCHIP1* and *SH2B3*) whose expression in the intestinal mucosa varied according to disease status and the genotype of the associated variant. Our results suggest that these genes may be constitutively altered in celiac patients, probably before the onset of observable symptoms of the disease, and therefore could have a primary role in its pathogenesis.

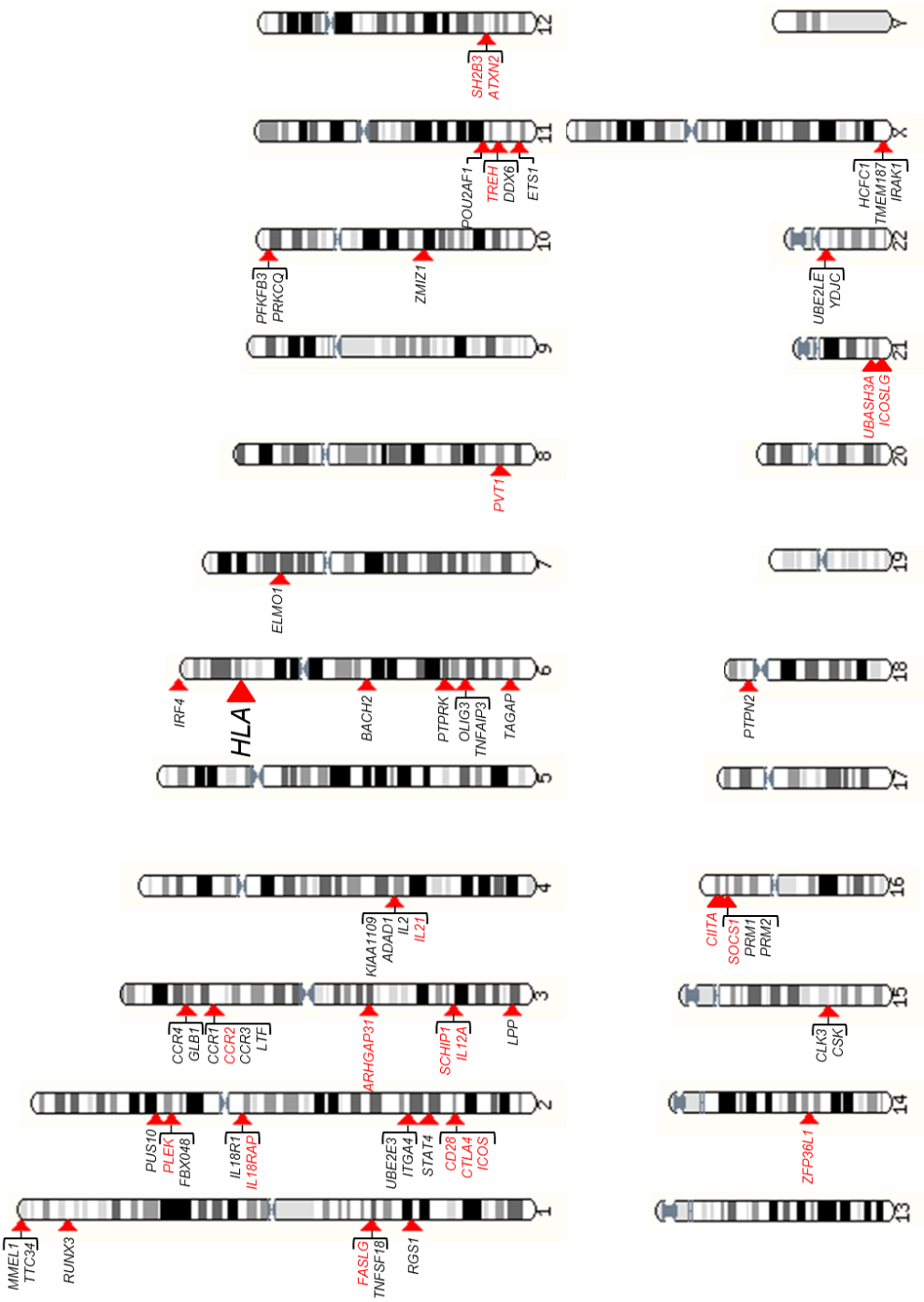


Figure 18. Celiac disease associated regions and proposed candidate genes. Genes highlighted in red showed differential expression in functional analysis.

Following the immunochip project, the list of candidate genes from associated regions increased considerably. The expression of those new candidate genes was analyzed in the disease tissue of celiac patients at diagnosis and after treatment, and compared to non-celiac controls. Moreover, the SNP genotype effect in gene expression was also investigated and coexpression analyses were performed. Several genes showed differential expression among disease groups, most of them related to immune response. Multiple trans- but only 4 cis-eQTLs were found, and surprisingly the genotype effect seems to be stimulus dependent as it differs among groups. Coexpression levels vary from higher to lower levels in active patients at diagnosis, treated patients and non-celiac controls respectively. A subset of 18 genes tightly correlated in both groups of patients but not in controls was also identified. Interestingly, this subset of genes was influenced by the genotype of 3 SNPs. These results strongly suggest that the effects of disease-associated SNPs go far beyond the over simplistic idea of transcriptional control at a nearby locus.

Gene expression patterns that are cell specific are the next step in the analysis of the candidate genes in CD. Due to the structured complexity of the intestinal tissue, where different cell types are combined in different proportions, gene expression analysis in whole biopsies could yield deceptive results, with truly altered genes being masked by the complex tissue. By isolating the 2 main cell populations in intestinal biopsies, CD45+ immune cells and CD326+ epithelial cells, and by analyzing gene expression separately in each of them, it is possible to get a more reliable idea about what is going on in celiac intestine.

This new methodological approach in the study of CD candidate genes makes it possible to identify gene expression alterations that are cell population specific, as is deeply described in chapter 3 of this thesis. To sum up, 8 genes showed alteration exclusive of CD326+ cells and 3 genes were exclusively deregulated in CD45+ cells, while there were 2 genes that shared alteration in both cell types but not in whole biopsies and only one gene shared among the 3 analyzed samples.

In conclusion, studies so far stress the need for developing functional studies as well as the importance of avoiding arbitrary selection of susceptibility candidate genes. Additionally, they reveal the huge work that remains to be done in order to identify the elements underlying the complex regulatory system of the genome, while opening the door to future studies, in which the scientific community will need to exhaustively analyze both different classes of variation (such as structural variants of the genome or epigenetic

features) and the vast noncoding genome, in order to shed light on the complex genetics of common disorders and to be able to understand the effect of the disease-associated variants found by the numerous GWA studies.

Conclusions

1. Genetic associations from GWAS must be replicated to determine whether the genetic makeup of CD is robust and comparable across different populations.

    a. Association results on this thesis follow other similar studies, and are a confirmation of our current knowledge on the genetics of CD, stressing the importance of analyzing large samples to obtain robust results that can be replicated.

    b. The strength of the associated signals is relatively small, and the ability to detect significant association often depends on minute allele frequency differences across populations, which could account, at least in part, for some of the negative results.

2. The functional implication of the proposed genes and variants must be experimentally addressed by analysis in disease tissue or cell models in order to identify real functional players of CD.

    a. Candidate gene selection following large-scale SNP association studies is often aprioristic and greatly influenced by the current knowledge of the pathogenic mechanisms that are thought to be involved in the disease, but functional studies are the only unbiased approach to identify real functional players.

3. Alteration of gene expression in active CD tissue is at least partly the consequence of disease mediated phenomena (e.g. the proinflammatory environment) and might not reflect a primary event of genetic origin.

4. There are different functional relationships between the expression of candidate genes in associated regions and SNP genotypes, and the link with disease pathogenesis is not straightforward.

    a. The biology of complex diseases is certainly much more complex than a direct SNP altered gene function-disease relationship, and we must be very cautious when proposing aetiological genes and pathogenic mechanisms based only on association peaks.

    b. The effects of associated variants go far beyond the oversimplistic idea of transcriptional control at a nearby locus. The complex interactions that maintain a coordinated, healthy response to an environmental challenge are written on our genome and the disruption of those subtle fine-tuning mechanisms emerge as the initial cause of a series of events that eventually lead to disease.

5. To question the implication of the proposed candidate genes in disease development taking into account the cell populations in the intestinal epithelia is a novel approach never used before. In this way, we are able to identify discordant expression levels in both cell populations, CD326+ epithelial cells and CD45+ immune cells.

    a. Overexpression of genes related to the immune response in epithelial cells indicates the importance of the innate immune system in CD, based on the fact that gliadin peptides are able to trigger a non-T-cell-dependent response that could establish the proinflammatory environment necessary for subsequent T-cell activation and tissue destruction.

    b. This new approach of separating both cell types for functional studies could help us to expose functional players in CD that keep hidden in gene expression analysis of whole biopsies.

Supplementary material

Supplementary table 1: Genotyping assays design for 44 top-associated SNPs from the Immunochip project, performed with a Fluidigm Biomark dynamic array (48.48) and SNPtype assays.

SNP-ID	Sequence
rs4445406	GGCCTCCTCTGGGCTTGGCTAGGGTCCTATGCTGGAGGAAGGGACAGACATTCCAGGGGA GGCTGCCTGTGATGAGGCAGGCCCCGGCTGTTAGGGGCCGGATGCCGTGGGGAGTGCCGT CTGACCTCCTGGGCTGCCCACTGGGGATAAAGGGCTGGGT[T/C]CCCGGCTGGGCTGYCGGTC CTGAGTTGCGTTCCAGTTTTCCACATGGTCTGGCCCCTGCCGTTTGTATATTCAGCCTCT CAGGCCAAAAATTAAGTAAAGGCGAGGACACAGGAAGTCATCTCCAGCTGGGRGTATTGC
rs12068671	CTGTTCATGAATATTTGCATAACTGCTAATTAATGCTTGGGAGAACTGAATCACTGGAAT GTCATGGGGGCAGATTATTTCTACTAGGCCTGGTGGAATCTTCTTGTGGTGGGCATGGCC TGCCTGGGCTTCTACAACTAACATCCAGGAGCAGAAGCCC[T/C]TCCATTTCTAGCAGTTGAT GAGGGTTGTAGTGCTGATGAACACTGATGACTTGGAAAGCTTTTAGAACTTAGATAGGGG TCGGGRGGAAGTGCAGACCCTTTGATTTTCCTTTCCCAGGATGGGAAAGAGCTCATACGA
rs859637	CCTTTCTGGATGGTGATTCCAATGTTAGTTACATTTTCAAGGCCTTTGCTGTGCTGCTCT GGGTCTGYCTGTGCATGTGCTACTCAGAGATAACTCCAGTGTGTCTTTAGYTACAGAATT AGAAGATTTCCTTCTCTAGCTCTTTCCCCTCAAGYATTTC[T/-/C]TTCCCCCAAATTC TCTGACTTCCAAGGGCTTGTTTTCCTAGTTATGTGGTCACTAGAAAGATAAGATTTAACT KGAAGCTATAGTGACCATAATACCCATAATTCTCCATGATCDGGGCTGCCTTTAGAGCAA
rs10800746	GGGCTTCGGACGAGCCCTGGGACCCCGGGCACAGGTACGGCTGCTCTGGAGGGACCCCGG GGCCAGGCAGGAGGGCTGTTTTGTTGGGGAGGGGCTGCAATTCTAGTATGTCCAAGGGTA CAGAGGCATCARAGCAGCAACCTAGTGGCCGGGTAATCCC[C/T]GTCTAGGAAGAACAGGCCA AAARATCATTTCCTYTTSGCTTTGGAAGGCACTCTTGGCTTCCCTGGGGCACCTCCTAAT TTTGCCTGCCATGTCTAGCTCAAGAGCATATACACTGGTGGAGATGATCCCCGAAGATAG
rs13003464	GCGCACCACCATGCCCAGCTGATTTTTGTATTTTTAGTAGACACAGGGTCTCACCAYGTT GCCCAGGCTGGTCTTGAATTCCTGGCTGCAAGTGATTYGCCCGCCTCAGCCWCCCAGAGT GCTGGGATTGCAGGTGCGCACCACTGYACTCGGCAGAGCA[A/G]TCATCTTGAGAGAGGACAG GCAGATATTCTCGAGTGCCTTGGCTGAAATCCARCCCAGGCCARCGGCACCCAKAACATG ATGCCATAGGCACAGGAAGCCTATGGCCAAGAGGCCCAGACTTGGTAGTGATTAACAAAA
rs10167650	ATTCTGAATTTTYCTACCACTTCTACAGAAGGGTATAATATCTCATAATCTGTAGACAGG TTAAATACCAGGCCTGCTTCATTAGTTCCTAACTAACCTTTCTGAGCCTTGTATWTGCAT GTGTGATATGAGCATATTGATACTTAGCTTGCACTGTTGA[T/G]GTGAGGATTAAATAAGATA ATGTAAATAAAATGACACTAATGTGGATATTCAATATATGGTAGTTTAACAAGATGTTAA AATGAGATTCTTTTTGTCTGTATTCAGCTGAAGCTCTCAAAAGAACCAGAACCAAATAGT
rs1018326	AGTAGAAGTATGAGAGGAACACTTATCTTTAAGAATAAGTGCAGACTTTTTGCTCTCCTC AAAATATCATCTTCATTATAATTCCAAAAATATAATGAAAGAAAAATTAAAATGATAACC TATCTTATTCTTACAATTAACTTAAGATTTTTTTGAGTTA[T/C]TTCCATCTATATACTGGAA TCTTAATACTGACCTTGGCATTAAAGTCCAGGTAGTATTAWTAATATGAATGTTACATTT GTGTTATTATTCTACAGTTTATTAAGTGCTTTTACTTATATAGTGGATTTAATTCTTCTG
rs6715106	ACTTTACGCTTTATATAATACTTTACTGAACTGAAATGYGGGTGGTGAGAGGGTGATTAT TTTCTCAGTGATTCAAAATAATGTCTAATGTTTACTTTAAATCTCTGATGCAATTGTGAG GCACATTTAGAACAAAATCCAGCAAGGAAGCTGACATACT[A/G]TACACAAGACTGGAGGAAA GACGCTCTGACCAAATAAAGCATTATWAATTTTCACAGCTTGCATGACTGTCATAGTCTT CATTATTTCATCAGTGTAGTCTAAGCTAKAACACTATTCAGTCTCACATTCKCAGGTTCC
rs6752770	AAAACRGTGATGAATAGGTTAAAAGGAGAAAACAAGTGCATGCTCTGGGACCRTGATTCT GTGGTTGGGCCTGAYGGATTTAAACCAGCTAAAAGGTCAGAGCCAAAATAAGTGCTATTC CCTAATGGAAACCTCCCTAAATTTCTAAAACACYGTGTAC[A/G]TGCCCTTCAAAATTTGATA AAGRTTCACTAACAAAACAGTTTTTCTTAATTGAATTTTCCATGTCTACTACACTTTCTG AAAAATAAAACYGTTTGTGTGACAAACCTTTTTTCCACTTGGCAAAAACAGAAACCTCTT

rs12998748	TTCCTCGCCAATGTCAAACACACTGGCCTGTGTGGGTATCAAACTGACTTACTCATAAAT TAAGTAAATAAGCCCAAATGCCTTTCAAGTTCGCATGAKTTTAGTAATCTTTGATGAAAA AAAGACAGTTTTAAAATTATTGCTAAAATAAAACAGCAAT[G/T]TCTTCAGAATTTAGACATT TGGCCTAAATTAGTCCRTCTTTGATAGGTGTTTTAAGATATAGAACTTTGTGGCTGGGTG TGGTGGCTCACGCCTGTAATTCYAGCACTTTGGGAGGCTGAGACAAGCAGATCATGAGGT
rs1980422	CTTCTTATACATTATTGAAGTCTTCATAGTGAACTACATCTTGCATGTGATGGGTGCTCA GAATATCTTAATGAGTGAATCAATGGATGGATGGWTAAATGAATGACTATCTTTCATTTG ATAAATATCCGCAAGCTATTTTGGTTTTTGACAAATTAGA[C/T]GAAACAGGTATTATGAAAA GACTTGGGAAAATTGAGGACAAATTAGTTAACTAGATACTATGAAAAGACTTGGGAAATT CATATTTTTAAAATTATTGAAATGTATTAATTAATATGCATTTTAAGACATTGTTTAGGC
rs34037980	TTTTTCTTTTAGCAGGAGCAAAGATAATCTAWCAAATCAAAGATGGTCTTTTCAATAAAT AATGCTGGAAAGACTGGACATCCACATGCAAAAACCAAGAACCAAAAACCAAAAACTGAA GCCAGAYACAGGTCTCCCACATCACAAAAATTAACTTCAA[A/G]TGGATAATAGACTTAAATG CTAAACACAAAACTATAAAACTCTTAGAAGATAACATTGGAGAAAATCTAGACGACCTTG GGTATGACTATGGCTTTTTAGATATAACACCAAAGRCACAATCTATGAAAGAAATAATTG
rs10207814	GTACCTTTCAGTCAATTTTGCTGTGAACCTGAAGCTGTTCTAAAATGTAGTCTATTAAAA AAGGAAAAAAAAGGAGAGATATTAATTGACCTTAACTTAATATATAGCAACAGTCTGTAAA GGCTGTTAGATGCACCAATGCAATGTAATGGAAGTGTTAT[C/T]CTGTGCTCAGCCTGGGGGC CAGGGAAAAGAATCACTCTCTTCTCCTTTCTAGTTTTACCATTTCTGGGGTGTTATTCTG AGAACCAAGCTTAAAAGAGACAAAGACAAGTGAGAGGCTTGTAGATTAGGACAATTAGCA
rs4678523	CCACATACATCTGTGAGATCTGGTAAACYGGCAGGATGACAGTATATTCCTCAGTGTTCY TGTGAGGGGCAAGTGAGTTAAACATAAACTCAGAACTGYGCATGGTGYAGAACAAGTCCT CAACAAATGAGAGCTACTGCTCCATGTCAGGACAGGAATA[T/C]CAGAAACACCACATCCTTT CCTATCTGTGCMCATACACAGCACACACACACGTGCATGCCTGACTACTGACATGTGTAT AGGCCTGTACTGTGTGAATATGTGAAGTTCTTATGTTGAAACACACAAAACGGGATCCAA
rs2097282	TTTACAAAGCTATAATAATCAAGACAATGTGTTACTGACATAATGATAGACATATTGATC AATATAGACAATGGATGAATGGAGTAGAATTGACAGTCTAGAAATAAAACCTTATATATA TGGTCAATTTATTTCTTTATGTATTCTGTCATCCAGAATA[C/T]ATAAAGAACACCTTCAACT CAATAACAACAATACCAACAAGGGTGCCAAGACCAATCAATGSAGGAAATAATAGTCTTT TCAAGAAATTTTGCTGAGACAACTAARTATTTATATGCAAAATAATGAGTTTGAATGGCT
rs7616215	GGGATCTCCTGATCTGAGGGTKGCAAAAMTCTGTGGGAGAAGCATGGTTTCCYTGGGTCA CACAWTCACTCACGGCTTCTCTTGGCTGGGAGTGGGGGTTCCCTTSGGTCTGTGTTGCTC CTGGGTGGGCTGACATCCTGCCCTGCTTTTCTACATTCTC[C/T]GTGGGTTGAACAGTTTCTC YGATCAGTCCTAATGTGAGCACCTGGATGTTTTAGTTGAAGGTGCTGAATTTACTCACCC CTTTTGTCCATCTCCGTGAGTGCCACACCCTGTAGCTGCTTCTAATCGGCCATCGTGGCC
rs61579022	TTTTCCCCCTCTCTCCTCTCTATAAATAAGAATACACATTTTGTATTAAGTTGGTCTGGG AAAGACACTAATGGAGACTCCAGCCTGCCTTCCTACCTGTGGCCAATGGGCAGTCAGGGC AGAAAGAGGGGCCTTGTGGAAGGCAGCCATCCAGATCTGA[G/A]ACTGAGGGTGGAAGGGGGC TGCTGAGAGGCTCAGGGGCAACAGGTCCCCTCACCCTGCTCTCTSAYGGGGTGAGGTYAA CTTCAGTGTCTTATCTTCTTCCATGATTATTCCATGTGGAGAACTTCACCTTCATTAATT
rs1050976	ATTGCGGCGAGACAAGCATGGAAAATCAGTGACATCTGATTGGCAGATGASCTTATKTCA AARGGAAGGGKGGCTTTGCATTTCTTGTGTTCTRTAGACTGCCATCATTGATGATCACTG TGAAAATTGACCAAGTGATGTGTTTACATTTACTGAAATG[C/T]GCTCTTTAATTTGTTGTAG ATTAGGTCTTGCTGGAAGACAGAGAAAACTTGCCTTTCAGTATTGACACTGACTAGAGTG ATGACTGCTTGTAGGTATGTCTGTGCYATTTCTCAGGGAAGTAAGATGTAAATTGAAGAA
rs12203592	AATATGCTTCTCARGTCTTCTGGGAAACAGATGTTTTGTGGAAGTGGAAGATTTTGGAAG TASTGCCTTATCATGTGAAACCACAGGGCAGCTGATCTCTTCAGGCTTTCYTGATGTGAA TGACAGCTTTGTTTCATCCACTTTGGTGGGTAAAAGAAGG[C/T]AAATTYCCCTGTGGTACTT TTGGTGCCAGGTTTAGCCATATGACGAAGSTTTACATAAAACAGTACAAGTATCTCCATT GTCCTTTATGRTCCTCCATGAGTGTTTTCACTTAGTCTGATGAAGGGTTCACTCCAGTCT

rs7753008	TTAACAGGTTTRTTGAGATATAATCAGCATATCATACAAYTCACACCTTTAAAATATACA ATTCAATGTTTTTTAGTATATTCACAGAGTTGTGYGACCATCACCACTATCTAATATCAG AACATTTCCATCATCCAAAAAAGAAAACCTACACCCATAG[T/C]RGTTACTCTTCATTTCCTT TCCCCTTCAACCCTTGGCAACCACCAATCTACTTTCTGTCTCTGTGGATTTTCCTATTCT ACACATTTTGTATAAATGGAATMATCTATTATGAGGCCTTTTGTGTCTAGCTTCTTTCAC
rs55743914	TGCCTTGGTTTTTCTGGTTTACTTTTCTGTGAGCATACACTACACAGCTGCTCCCAACCT GATCCCACCATGCACATACATATGTGTAACCCACAGATATCTCTYGTTAGAAGTAAAATG TGTTAGATAAGTGTAACTTTCTTTCTTACTCTGCATACAG[C/T]TTCACCTAGATGAACTGTC ACACACACACATACCTATCAGCTCTTATTTTGTAAGCTATGTTCTGACTTTAYSAGCCTT GACAAGCTCAAGTAGCATGATATGAACTGCCCTCTACAGTTGGTAAAAGCTGAAAACTTA
rs72975916	GCAAATATTTTTTGGCTATAACAGAAAAAAAATCCTAMAAATTTCCCAAATTAAATGAAA ATAGTTTCATGATTAGCAGTTTATGCAATACRTGATTAAGTGTCTGTGGGTAGAGGGGGA TGCTCAGCAATGTATTAATTAGGCAGGAAAGTCCCCACAA[C/T]GTCTCATTTTCACATCTGA TTCCCTATGTGTSTATGAAGTCAAATCAATAAGTAGGTATATGACTAACTAATSTCTAAG ATGCCTTCCAAATCTAAGATTCTGTGACTCACCGGGGCTTCYACCATGTTTGGCTTGCTG
rs17264332	TATTTGTTAAGCTGGTAAAACATTCTGACTTCTTGAGKTGTCTAATCTAGGCTTGCMTAC TCTTTTAGGGGATSTTTTAAAACTGTTTCTATGCTGGTATATAATCTCATATTCCTCCTA CTGTTATTTTATTAAGTACCTYATTTTATTTTTATTTTCT[A/G]CATGGTTCAGCCTAGTTGT TTCTATTAAAGCCAAGATAACTTCAATTGCTCAACAACAAGGTGAACTTGAACTGCTGAG CATGATCAATCTCATCAAATTGGCAAATGCTTATTGAGGGTCTATCTATCACATGYTGTG
rs10808568	GAACTTAGCACAGTGACTGCCAAATAGTGTTGACTTCATAAATAGTGAGCAATCATGAAA AAGCAGATTTTCCTTTTGTAGAACAATAAAACTCACACAAAGAAAAGCTACAGTTTCAAA TGGCTTTCTGGGAAACATAGTACAAATSTGAAATTCTGAA[A/C]TGAATCAGGCTACTTGGTA AGTAGATACTGACTTAAATCTAAACSAAGTAAGATGCCATCCTATTTCAAAAGGTGAGAA TAAGAGYGCAAAGGACCAGTGCAGCAGATCCYTCCTGATGGTGAGTCAGTGGTGGTGAGT
rs2387397	AAATGAATGACAGAGAAAAARGAAAAAGCCACCCAGSCTTGAAAAGTGGTAGCTTCTTGC YGCCGACTTCCTGCCRGTGCCAAGTCAGAKGGAGCTCTGCGTGTCCTCAGTTCACCCTCG CCAGGCCACACCGCATGCAAATAAGAAGCTGTTTCAGTKT[G/C]TGCCCATCTGAGAYGCTGA CATAAAAAAGGAAGAAAGAAAGGRAAAAAACCACCGTCACCACCAACAAAAAAACCCAAA ACTGATTAATTCTMGTAGATTRCAGCAGAGCAGGGGGCCTGGTGGTGACACATCCTGTTT
rs1250552	TCYCCAACCCTGAGCCTCTGTGTCCCCATTTATCACAGGGATGGGTAAGGGTCTCCACAG GCACCTGACACAGGAGCTCAGTGGTGAGACCCAGRTGAAGGGTGGGGGAGGCCCTTCCCAG GGCTGTCTYCTGAGCAGGACAGAGATCTGCGAGAGAGAAG[A/G]GTTTCCYTGGAGGAGGCTC TCCTCCCGAGGGGCATGGCTCCAGGCCCCTGGGATGAGTGCRGTRTATGTCCATAGCCRT GGCCCCTTCTGTCTCYGTARTCCCATAACTYRTGCRTGTCTTTTCTTTCCCTCCAGCCTC
rs7104791	ATCCTCCTCTGCYTCCTTGATTTTCCTCTTCATCTTTCTTCTCAACCTTTCTCCTTTCTT TTTATACCCTAGTTTTATTCCTTACCTCTGACAATCTTGCTTCTTTGACAAAACATTAAA CTGAGATCTCAGACTGGCATGCAACCTGATTAGTAGAAAC[T/C]TGCTCTTCCTGGCCTAGTG CAGTTCATCATGCAGCCTCTCTTCCTCYTCCCCTGGCACCTCACAGCAGTCAGCATCCAY GTGCTCCTGCTAACAAGTGCAAGGCCAAATGAAGAAAACACTAAKAATGCATTCACAAAA
rs10892258	ACCATATATTRGCCACTGGTTTAATGTACAAGATGTGTCTCCRGTCCCAATGGTAGACAT CCTATAGCCAAGACTCCATCCAATTGCCTGACCTCCACATGTCCCTACTCTGACTGGAAC CYTGTGGGTAGTTTAGGTCTCWGAGGATCAGTCAAGCTCA[G/A]AGCAAGATTGAATAACCCC TGGAAGGTGTGGATATTCYTATTTCCCCAGGATCAGAGGCATATTTACCATAAAGCTAAT GATGCTTATGCTTCAAGACTGCTCACTTACTAAAGCCCTGGTGTTAACATGATCATATGC
rs61907765	TTGGGTGGGGGGCCTCRGCCGTCGCCACTCACCCGGGGAGGGGAAAAGCTCCAGATCGACT TTTTYYGTCTTGATGATGGTGAGAGTCGGYTTGAGATCGACGGCCGCCTTCATRGTGCCA GGAGTGSGGGACGTACGGGATGGTAGCARGTTTRCAGTTA[C/G/T]YGTTGTTTTTCTTTTTAAT GAGGATTAGTAACAGGGGGAAGGGGACGGGGGAAATCCGACTTTCTTCCCAAAAATCTCA AATTCCCGCTGCCTTTCTTTCCCCCGCGCCCGGACGGTGCGCGCCCGGCACTCCAGGGGA

rs3184504	ATACTCTCTCTAAAAGGGGGACTCCTGGGGAGACTATAGACAAACTCAGGCCTGGCTGGA AGAAAGAGCMYCAGGAACAAGCCTTGAGTACCCCAACYTSGTCTCGTAGAGYTCAAGGCC CAAGCTACAAGCARCTTGCTCCAGCAYCCAGGAGGTCCGG[T/C]GGTGCACACRRCTTGAGAT GCCTGACAACCTTTACACCTTTGTGCTGAAGGTGAGTGACAAGGCTTTTCAMACCCTGGG GCAATACAAATACMTACACATACAGCAGACCCAACSCTGTTCCCTTCCTCCCTGCCAGGT
rs11851414	CCCTGCCAGGCAAACTTCGCCYCTCAAAACCCTGGCCTCCAGATKCACATGTAATCMCCG CCAGSAACTGTTGAAACTCAAAGGGTGGGAAGGACGGGGCCAAATTCCTTCAAACTTGGS AGAAATGCCGGAGGAGAAAAGAATCATCYCGCTGCACCAC[T/C]TTSCCCATTGCCTTCCAAG ACCCAAACTTTTGGGGGTTCTTTCTTAAGGCAAAAGAAAAAGACTTTTTGAAAAGCAAAT GCTCCGCCCCCCTTTACCTTGCATAAARCTTCGCTCAAGTCGAAGATGGTGGCAGACACG
rs1378938	TGGGGTGTTCCCCTGTCCTGTTGTGGCCCCGACCCTTCCTCTACCAGCCAACAGGACCTG AAATCCAGGRAGATCTGACTTSRAAGTCTACYTTGTGTCTGTTCCTCAGCCCCACTSTAG TTCCATCYACCAGTGCAACAGCCTCYTCGCCGCCTCCACT[T/C]CATCTTCTCTGCATACTGC CAGACTTAATTTTTYTAAAACTACACCATTCATTCAACAACAAATATTTGAATACCTGCT GTTTGCTAGGACTGAGGTTCATCAATAAGCAAAACACACCCTGTGGAGTCCACAGTTGGG
rs6498114	CTCTGGTCATGACCTGAGGAGGGAGACTATGGGGTYTCATGAGTCCTGGCCAATTGCAGT ACTTCCTGCMAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGTACTGCTTTAGCTTCCTCTTCT RAACAACCCAAGGGGCTACCGTAAATCTACCCCAAATAGG[G/T]TGGCAGTCTTCAAACTGAA ACCAGCCTGATGATTCAGAAGTCCCCTTCAGATATTGCAGTGCCTTAGGTGCAATTGTGC AGAAAGCCCAGTTCTGGWGATGGGAGCTTAGGGGTTTCGCTCTGTGCTACACTGAGCAAC
rs243323	GGTGCCCCCAGCTCATGTCCCAGCACCTTCAGATTCTATGCCAGCTCCTCTGGAACACAG GTTCTAACCCAGGCTCATTCCCTCACGGGGACATGACATSACCCTGCTCCCTTGATACAG CTCATTCTGACAAGTCTTGTGTGTTTGGACTGTAAGTTCC[A/G]TGAGGGCAGGGCCTTGTGT GTTCTCTAGTTCTGTTCCCAGAGCCMAGCACAGTCCCTGAACAGACTTTGGTCCCTTAGT AAAAGTTTGTTGAAATGGCTGCTTTGAGGAACGTTGAGGTTGGTACTATGTTACAGAATC
rs9673543	GGTCCAGAATTCAGACATCTCAGTAGAGGTGTCYAGTAASTAGGCAGCTGAGTATCTGAG TSTGACATTTAGCCTGGAAGTGATATATTTGGKCATTGTCAGTGTTTACAATCAGGAGAC TGGCCAGGATCACTGATTTAGTTTTTGCTGCTGCTGTAAC[A/G]AAGTACCATAAACTTTGTA ACTTAAAACAGCAYGAGTTCATTATGGTTTTTCGCYCCCAACTCTGTCASCCAGGMTGGA ATGCAGTGGCATGWTCACGGCTYATTGCACCCTTTACTTCCTAGGCTCAAGTGATCCTAC
rs11875687	CGTAAACTAATTCRGCAGAAAGGAAAAGCGCAGAAACCTTCAGGACTTCTTAGTACAGGT ACCTGGAATTTTGCGAYGCAGATGAAAAACATGGGTGAGGGGAGTTTGAGGGATCTAAAG ATGTGGATTCTGATCTCTCTTTTGGCTCTACCAACCTACA[T/C]GATTCCAGAYGTAGTCATG ATTTCCCAGAAAAGGGTGACTACTACTTGAGAAACACTATCAGCATCCCAGTCGCCGACA GTCCTATGATAAGGGGCACTCAGGGGTCTGTTTTCTTGTCTGTCAAWATCAGTGGCTCRA
rs62097857	TAGAGAATTATTCTCTGTAGCTCAGTGATTCTCAATYAGAGATTTGGAGGCAGGGGAAGG GCACCTAAGTATCAGAATTACTGCCCACACCCRCTATTCCCTGAACGCACACCCACACTG GAACTGCCACCGTGATGAGCCACCTGTTTCAGAGGAAGTG[G/A]GTYAAATCTTCAAGTGTGC TGCAACTGAAGAAAATGGAAAACCACTCTTCTAGCTACTTGTCTTTAAGAACACTTCTAA CTGGTAGTAAAGAAAAAAGCATTAAGAGAAACATAGTATTTAATAGTAATTTATAACAAA
rs1893592	TGAAACTGAACGTCTGATGCTTATTTATACTTCAGGGAAWTACTACAGGCCAGASCTGAA TTTCCCCTGCAGTCTGCCMASAYGGAGTCGTGGGATCAAAGACTTTGAAAAYKATCCCCC ATTATCATCGTGTGGCATTTTCCAGTCCAGAWTTGCAGGT[A/C]TGTTTGAGGACTGTCTAGT AGGAAAGGTAACAATAACARCAACACTGATTATGGCTAGCAGGCATCCAGCCTGAGCCCT AGCTACACACCCTTTGATACAGTGTTAGCTTTGTCTCCAGTCTGCTGAGGAGGCTGGGGC

rs58911644	TACAAAGGCCTACCTGCGTGAATGTCCACCTGCMTGGATGCCCACCCRCACGGATGCCCA CCTGCACAGTTATCCATCTGAACAGCCCTCTGGTGCCTCCTTTTGCTTGCACATGTTCAA ATCCTCCCAGTCCTACAGTCGATTGTTCCTAGTTGGGGTC[A/T]TTCTGGGCATGTTATTCTA AAACTTGTTTTTTCCCRTCTCCACATCAAAGGAGAAAGGTCTAGCTTGCTTCTTTGCTAT AGAGCAGGGCCTGAGTGAGGTCAGGACCACAGAGCAGACCCTATGAGTGTTGGTCAGGAC
rs4821124	GCCCAGAAACAGCCAGTTACCACCCTCCCCAGAGCCAGAGACCTTTAAGCCCCTCTGTCA ACTTCCTCTTCCCTACAGCCACCAGCCACCATTTTTGGCCTTGTTCACTATTAATTGTTC AGCATGAAATAAGCATCCATTTCATCACTYCTGTACCYKC[T/C]GCAGCCTGACTTCTGTCAC CTGCTGCCCCTTGCTGCTKCTCCTTTCATTTGGTGCCATCTAACCCAGCCCTGAAAGAAG CCAGAGCTGCTCTCAGCATTTCCCACCCTGTCCTGCCTGGCTCTTCCTTCTCTCTCTGCT
rs13397	CCTACTTAGCTTTGGGGGTGCTCTCTTGCCTGGGYTTTGTGGTCCTCAAGCTGTGTGACC ATCAGCTCRCACGGTGGYGTCTCTTCCAGTGCCTCACAGGCCACTKCTGGTCCAAGGTCT GTGAYGTGCTCCAGTTCCACTTTGCGTTTTTGTTTCTGAC[G/A]CATTTCAACACTCACCCAA GATWCCATCCCTCTGGCGGGAAGACGCGTTGAACCCAGGGAAGAACCTGCTGAAAACYRA TGACCCCCAGCATTGAAATGGACTCTGAGATGGCAGCGTGGTGCCAGTGTCAGACATCCT

Supplementary table 2: Top 50 associated SNPs from CEGEC sample set. SNPs with strong LD grouped in the same association peak are shown in bold.

SNP	CHR	BP	p value
rs1105297	1	43184183	7.49E-05
rs3790568	1	67608648	9.94E-06
rs58438451	1	67613706	6.93E-05
rs17129913	1	67616667	6.93E-05
rs7515827	1	67624052	4.09E-05
1kg_1_67626426	1	67626426	6.63E-05
rs7513724	1	67627123	6.93E-05
rs7536192	1	67627190	4.38E-06
rs12409092	1	67629219	4.66E-06
rs1874396	1	67635054	2.84E-05
rs13001423	2	100291772	4.21E-06
seq-NOVEL-10879	2	162954368	7.90E-05
rs11883509	2	181655408	2.86E-08
imm_3_49278660	3	49278660	7.53E-05
rs62259859	3	58471021	6.18E-05
rs63332460	3	161215534	1.08E-06
rs6451245	5	35988854	5.93E-05
rs10045685	5	35989454	7.12E-05
rs10078373	5	35992572	3.78E-05
rs1478449	5	35992959	6.87E-05
rs7785711	7	26828165	7.24E-08
rs4841530	8	11326556	6.82E-05
rs4840561	8	11343870	5.09E-05
rs13264212	8	11347259	1.82E-05
rs13253092	8	11352401	7.06E-05
rs13251015	8	11352458	7.06E-05
rs2244234	8	11373877	4.20E-05
rs7812879	8	11377591	5.11E-05
rs1478900	8	11385069	6.85E-05
rs2736343	8	11386659	4.99E-05
rs9694294	8	11388130	4.83E-05
rs2553878	8	54218936	1.09E-05
rs13377037	10	1058639	6.08E-05
rs11250242	10	1090064	4.28E-05
rs7984030	13	43528420	7.26E-06
rs11846679	14	21490247	1.34E-06
rs61975792	14	87652580	3.53E-05
rs7145673	14	87652974	4.55E-05
rs17124095	14	87669880	5.11E-05
rs1940550	14	87673873	3.82E-05
rs1570194	14	87696186	3.72E-05
imm_14_97450261	14	97450261	7.00E-05
rs16940147	15	56463411	4.97E-05
rs12593974	15	89952719	1.53E-12
rs33973997	18	12773118	2.33E-05
rs45551338	18	12774908	3.79E-05
rs28676870	18	12779681	2.33E-05
rs2032174	18	12782954	2.33E-05
rs2268278	21	35106809	4.54E-05
rs17648208	22	28828978	1.43E-09

Bibliography

1. WK D: Coeliakie. MD Thesis, 1950.

2. S. G: On the celiac disease. St Bart Hosp Rep., 1888, Vol 24, pp 17-20.

3. Feighery C: Fortnightly review: coeliac disease. *BMJ* 1999; **319:** 236-239.

4. Mäki M, Collin P: Coeliac disease. *Lancet* 1997; **349:** 1755-1759.

5. Marsh MN: Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). *Gastroenterology* 1992; **102:** 330-354.

6. Stenman SM, Lindfors K, Korponay-Szabo IR *et al*: Secretion of celiac disease autoantibodies after in vitro gliadin challenge is dependent on small-bowel mucosal transglutaminase 2-specific IgA deposits. *BMC Immunol* 2008; **9:** 6.

7. Mäki M: The humoral immune system in coeliac disease. *Baillieres Clin Gastroenterol* 1995; **9:** 231-249.

8. Husby S, Koletzko S, Korponay-Szabó IR *et al*: European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease. *J Pediatr Gastroenterol Nutr* 2012; **54:** 136-160.

9. Dubé C, Rostom A, Sy R *et al*: The prevalence of celiac disease in average-risk and at-risk Western European populations: a systematic review. *Gastroenterology* 2005; **128:** S57-67.

10. Ivarsson A, Persson LA, Nyström L *et al*: Epidemic of coeliac disease in Swedish children. *Acta Paediatr* 2000; **89:** 165-171.

11. Collin P, Reunala T, Rasmussen M *et al*: High incidence and prevalence of adult coeliac disease. Augmented diagnostic approach. *Scand J Gastroenterol* 1997; **32:** 1129-1133.

12. Bodé S, Gudmand-Høyer E: Incidence and prevalence of adult coeliac disease within a defined geographic area in Denmark. *Scand J Gastroenterol* 1996; **31:** 694-699.

13. Murray JA, Van Dyke C, Plevak MF, Dierkhising RA, Zinsmeister AR, Melton LJ: Trends in the identification and clinical features of celiac disease in a North American community, 1950-2001. *Clin Gastroenterol Hepatol* 2003; **1:** 19-27.

14. Plot L, Amital H: Infectious associations of Celiac disease. *Autoimmun Rev* 2009; **8:** 316-319.

15. Cataldo F, Montalto G: Celiac disease in the developing countries: a new and challenging public health problem. *World J Gastroenterol* 2007; **13:** 2153-2159.

16. de Kauwe AL, Chen Z, Anderson RP *et al*: Resistance to celiac disease in humanized HLA-DR3-DQ2-transgenic mice expressing specific anti-gliadin CD4+ T cells. *J Immunol* 2009; **182:** 7440-7450.

17. Rabassa EB, Sagaró E, Fragoso T, Castañeda C, Gra B: Coeliac disease in Cuban children. *Arch Dis Child* 1981; **56:** 128-131.

18. Sagaró E, Jimenez N: Family studies of coeliac disease in Cuba. *Arch Dis Child* 1981; **56:** 132-133.

19. Galvão LC, Gomes RC, Ramos AM: [Celiac disease: report of 20 cases in Rio Grande do Norte, Brazil]. *Arq Gastroenterol* 1992; **29:** 28-33.

20. al-Tawaty AI, Elbargathy SM: Coeliac disease in north-eastern Libya. *Ann Trop Paediatr* 1998; **18:** 27-30.

21. Suliman GI: Coeliac disease in Sudanese children. *Gut* 1978; **19:** 121-125.

22. Khuffash FA, Barakat MH, Shaltout AA, Farwana SS, Adnani MS, Tungekar MF: Coeliac disease among children in Kuwait: difficulties in diagnosis and management. *Gut* 1987; **28:** 1595-1599.

23. al-Hassany M: Coeliac disease in Iraqi children. *J Trop Pediatr Environ Child Health* 1975; **21:** 178-179.

24. Fasano A, Catassi C: Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum. *Gastroenterology* 2001; **120:** 636-651.

25. Feighery C, Weir DG, Whelan A *et al*: Diagnosis of gluten-sensitive enteropathy: is exclusive reliance on histology appropriate? *Eur J Gastroenterol Hepatol* 1998; **10:** 919-925.

26. Di Sabatino A, Corazza GR: Coeliac disease. *Lancet* 2009; **373:** 1480-1493.

27. Lerner A: New therapeutic strategies for celiac disease. *Autoimmun Rev* 2010; **9:** 144-147.

28. Schuppan D, Junker Y, Barisani D: Celiac disease: from pathogenesis to novel therapies. *Gastroenterology* 2009; **137:** 1912-1933.

29. Maiuri L, Ciacci C, Ricciardelli I *et al*: Association between innate response to gliadin and activation of pathogenic T cells in coeliac disease. *Lancet* 2003; **362:** 30-37.

30. Hüe S, Mention JJ, Monteiro RC *et al*: A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. *Immunity* 2004; **21:** 367-377.

31. Castellanos-Rubio A, Santin I, Martin-Pagola A *et al*: Long-term and acute effects of gliadin on small intestine of patients on potentially pathogenic networks in celiac disease. *Autoimmunity* 2010; **43:** 131-139.

32. Jabri B, Sollid LM: Tissue-mediated control of immunopathology in coeliac disease. *NatRevImmunol* 2009; **9:** 858-870.

33. Maiuri L, Picarelli A, Boirivant M *et al*: Definition of the initial immunologic modifications upon in vitro gliadin challenge in the small intestine of celiac patients. *Gastroenterology* 1996; **110:** 1368-1378.

34. Martin-Pagola A, Perez-Nanclares G, Ortiz L *et al*: MICA response to gliadin in intestinal mucosa from celiac patients. *Immunogenetics* 2004; **56:** 549-554.

35. Folk JE, Cole PW: Transglutaminase: mechanistic features of the active site as determined by kinetic and inhibitor studies. *Biochim Biophys Acta* 1966; **122:** 244-264.

36. Folk JE, Chung SI: Transglutaminases. *Methods Enzymol* 1985; **113:** 358-375.

37. Heap GA, van Heel DA: Genetics and pathogenesis of coeliac disease. *SeminImmunol* 2009; **21:** 346-354.

38. Caputo I, Barone MV, Martucciello S, Lepretti M, Esposito C: Tissue transglutaminase in celiac disease: role of autoantibodies. *Amino Acids* 2009; **36:** 693-699.

39. Lindfors K, Kaukinen K, Mäki M: A role for anti-transglutaminase 2 autoantibodies in the pathogenesis of coeliac disease? *Amino Acids* 2009; **36:** 685-691.

40. Mazzarella G, Maglio M, Paparo F *et al*: An immunodominant DQ8 restricted gliadin peptide activates small intestinal immune response in in vitro cultured mucosa from HLA-DQ8 positive but not HLA-DQ8 negative coeliac patients. *Gut* 2003; **52:** 57-62.

41. Lundin KE, Sollid LM, Qvigstad E *et al*: T lymphocyte recognition of a celiac disease-associated cis- or trans-encoded HLA-DQ alpha/beta-heterodimer. *J Immunol* 1990; **145:** 136-139.

42. Nilsen EM, Lundin KE, Krajci P, Scott H, Sollid LM, Brandtzaeg P: Gluten specific, HLA-DQ restricted T cells from coeliac mucosa produce cytokines with Th1 or Th0 profile dominated by interferon gamma. *Gut* 1995; **37:** 766-776.

43. Troncone R, Gianfrani C, Mazzarella G *et al*: Majority of gliadin-specific T-cell clones from celiac small intestinal mucosa produce interferon-gamma and interleukin-4. *Dig Dis Sci* 1998; **43:** 156-161.

44. Monteleone G, Pender SL, Alstead E *et al*: Role of interferon alpha in promoting T helper cell type 1 responses in the small intestine in coeliac disease. *Gut* 2001; **48:** 425-429.

45. León AJ, Garrote JA, Blanco-Quirós A *et al*: Interleukin 18 maintains a long-standing inflammation in coeliac disease patients. *Clin Exp Immunol* 2006; **146:** 479-485.

46. Steinman L: A brief history of T(H)17, the first major revision in the T(H)1/T(H)2 hypothesis of T cell-mediated tissue damage. *Nat Med* 2007; **13:** 139-145.

47. Castellanos-Rubio A, Santin I, Irastorza I, Castaño L, Carlos Vitoria J, Ramon Bilbao J: TH17 (and TH1) signatures of intestinal biopsies of CD patients in response to gliadin. *Autoimmunity* 2009; **42:** 69-73.

48. Harris KM, Fasano A, Mann DL: Monocytes differentiated with IL-15 support Th17 and Th1 responses to wheat gliadin: implications for celiac disease. *Clin Immunol* 2010; **135:** 430-439.

49. Monteleone I, Sarra M, Del Vecchio Blanco G *et al*: Characterization of IL-17A-producing cells in celiac disease mucosa. *J Immunol* 2010; **184:** 2211-2218.

50. Sjöström H, Lundin KE, Molberg O *et al*: Identification of a gliadin T-cell epitope in coeliac disease: general importance of gliadin deamidation for intestinal T-cell recognition. *Scand J Immunol* 1998; **48:** 111-115.

51. Sturgess R, Day P, Ellis HJ *et al*: Wheat peptide challenge in coeliac disease. *Lancet* 1994; **343:** 758-761.

52. Maiuri L, Troncone R, Mayer M *et al*: In vitro activities of A-gliadin-related synthetic peptides: damaging effect on the atrophic coeliac mucosa and activation of mucosal immune response in the treated coeliac mucosa. *Scand J Gastroenterol* 1996; **31:** 247-253.

53. Picarelli A, Di Tola M, Sabbatella L *et al*: 31-43 amino acid sequence of the alpha-gliadin induces anti-endomysial antibody production during in vitro challenge. *Scand J Gastroenterol* 1999; **34:** 1099-1102.

54. Arentz-Hansen H, Körner R, Molberg O *et al*: The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. *J Exp Med* 2000; **191:** 603-612.

55. Anderson RP, Degano P, Godkin AJ, Jewell DP, Hill AV: In vivo antigen challenge in celiac disease identifies a single transglutaminase-modified peptide as the dominant A-gliadin T-cell epitope. *Nat Med* 2000; **6:** 337-342.

56. Fina D, Sarra M, Caruso R *et al.*: Interleukin 21 contributes to the mucosal T helper cell type 1 response in coeliac disease. *Gut* 2008; **57:** 887-892.

57. Sollid LM, Thorsby E: HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology* 1993; **105:** 910-922.

58. Greco L, Romino R, Coto I *et al.*: The first large population based twin study of coeliac disease. *Gut* 2002; **50:** 624-628.

59. Gutierrez-Achury J, Zhernakova A, Pulit SL *et al.*: Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. *Nat Genet* 2015; **47:** 577-578.

60. Horton R, Wilming L, Rand V *et al.*: Gene map of the extended human MHC. *Nat Rev Genet* 2004; **5:** 889-899.

61. Ludwig H, Polymenidis Z, Granditsch G, Wick G: [Association of HL-A1 and HL-A8 with childhood celiac disease]. *Z Immunitatsforsch Exp Klin Immunol* 1973; **146:** 158-167.

62. Abadie V, Sollid LM, Barreiro LB, Jabri B: Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annu Rev Immunol* 2011; **29:** 493-525.

63. Karell K, Louka AS, Moodie SJ *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:** 469-477.

64. Spurkland A, Sollid LM, Polanco I, Vartdal F, Thorsby E: HLA-DR and -DQ genotypes of celiac disease patients serologically typed to be non-DR3 or non-DR5/7. *Hum Immunol* 1992; **35:** 188-192.

65. Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E: Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J Exp Med* 1989; **169:** 345-350.

66. Sollid LM: Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol* 2002; **2:** 647-655.

67. van Belzen MJ, Koeleman BP, Crusius JB *et al.*: Defining the contribution of the HLA region to cis DQ2-positive coeliac disease patients. *Genes Immun* 2004; **5:** 215-220.

68. Ploski R, Ek J, Thorsby E, Sollid LM: On the HLA-DQ(alpha 1*0501, beta 1*0201)-associated susceptibility in celiac disease: a possible gene dosage effect of DQB1*0201. *Tissue Antigens* 1993; **41:** 173-177.

69. Lundin KE, Scott H, Hansen T *et al*: Gliadin-specific, HLA-DQ(alpha 1*0501,beta 1*0201) restricted T cells isolated from the small intestinal mucosa of celiac disease patients. *J Exp Med* 1993; **178:** 187-196.

70. Hovhannisyan Z, Weiss A, Martin A *et al*: The role of HLA-DQ8 beta57 polymorphism in the anti-gluten T-cell response in coeliac disease. *Nature* 2008; **456:** 534-538.

71. van Heel DA, Franke L, Hunt KA *et al*: A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *NatGenet* 2007; **39:** 827-829.

72. Griffith OL, Montgomery SB, Bernier B *et al*: ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* 2008; **36:** D107-113.

73. McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9:** 356-369.

74. Consortium WTCC: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447:** 661-678.

75. Bush WS, Moore JH: Chapter 11: Genome-wide association studies. *PLoS Comput Biol* 2012; **8:** e1002822.

76. Hunt KA, Zhernakova A, Turner G *et al*: Newly identified genetic risk variants for celiac disease related to the immune response. *NatGenet* 2008; **40:** 395-402.

77. Dubois PC, Trynka G, Franke L *et al*: Multiple common variants for celiac disease influencing immune gene expression. *NatGenet* 2010; **42:** 295-302.

78. Kumar V, Wijmenga C, Withoff S: From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. *Semin Immunopathol* 2012; **34:** 567-580.

79. Trynka G, Hunt KA, Bockett NA *et al*: Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *NatGenet* 2011; **43:** 1193-1201.

80. Romanos J, Barisani D, Trynka G, Zhernakova A, Bardella MT, Wijmenga C: Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. *JMedGenet* 2009; **46:** 60-63.

81. Dema B, Martínez A, Fernández-Arquero M *et al*: Association of IL18RAP and CCR3 with coeliac disease in the Spanish population. *J Med Genet* 2009; **46:** 617-619.

82. Garner CP, Murray JA, Ding YC, Tien Z, van Heel DA, Neuhausen SL: Replication of celiac disease UK genome-wide association study results in a US population. *Hum Mol Genet* 2009; **18:** 4219-4225.

83. Amundsen SS, Rundberg J, Adamovic S *et al*: Four novel coeliac disease regions replicated in an association study of a Swedish-Norwegian family cohort. *Genes Immun* 2010; **11:** 79-86.

84. Bondar C, Plaza-Izurieta L, Fernandez-Jimenez N *et al*: THEMIS and PTPRK in celiac intestinal mucosa: coexpression in disease and after in vitro gliadin challenge. *Eur J Hum Genet* 2013.

85. Hunt KA, Mistry V, Bockett NA *et al*: Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 2013; **498:** 232-235.

86. Hunt KA, Zhernakova A, Turner G *et al*: Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008; **40:** 395-402.

87. Pruim RJ, Welch RP, Sanna S *et al*: LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26:** 2336-2337.

88. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30:** 97-101.

89. Green PH, Cellier C: Celiac disease. *NEnglJMed* 2007; **357:** 1731-1743.

90. Nistico L, Fagnani C, Coto I *et al*: Concordance, disease progression, and heritability of coeliac disease in Italian twins. *Gut* 2006; **55:** 803-808.

91. van Heel DA, Hunt K, Greco L, Wijmenga C: Genetics in coeliac disease. *BestPractResClinGastroenterol* 2005; **19:** 323-339.

92. Zhernakova A, Alizadeh BZ, Bevova M *et al*: Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *AmJHumGenet* 2007; **81:** 1284-1288.

93. Todd JA, Walker NM, Cooper JD *et al*: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *NatGenet* 2007; **39:** 857-864.

94. Zhernakova A, Festen EM, Franke L *et al*: Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. *AmJHumGenet* 2008; **82:** 1202-1210.

95. Zhernakova A, Elbers CC, Ferwerda B *et al*: Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *AmJHumGenet* 2010; **86:** 970-977.

96. Trynka G, Wijmenga C, van Heel DA: A genetic perspective on coeliac disease. *Trends Mol Med* 2010; **16:** 537-550.

97. Smyth DJ, Plagnol V, Walker NM *et al*: Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N Engl J Med* 2008; **359:** 2767-2777.

98. Zhernakova A, Stahl EA, Trynka G *et al*: Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet* 2011; **7:** e1002004.

99. Plaza-Izurieta L, Castellanos-Rubio A, Irastorza I, Fernandez-Jimenez N, Gutierrez G, Bilbao JR: Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of candidate genes. *JMedGenet* 2011; **48:** 493-496.

100. Holopainen P, Naluai AT, Moodie S *et al*: Candidate gene region 2q33 in European families with coeliac disease. *Tissue Antigens* 2004; **63:** 212-222.

101. Ward LD, Kellis M: HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012; **40:** D930-934.

102. Flicek P, Ahmed I, Amode MR *et al*: Ensembl 2013. *Nucleic Acids Res* 2013; **41:** D48-55.

103. Kent WJ, Sugnet CW, Furey TS *et al*: The human genome browser at UCSC. *Genome Res* 2002; **12:** 996-1006.

104. Almeida R, Ricaño-Ponce I, Kumar V *et al*: Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. *Hum Mol Genet* 2013.

105. Östensson M, Montén C, Bacelis J *et al*: A possible mechanism behind autoimmune disorders discovered by genome-wide linkage and association analysis in celiac disease. *PLoS One* 2013; **8:** e70174.

106. Fairfax BP, Humburg P, Makino S *et al*: Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 2014; **343:** 1246949.

107. Fernandez-Jimenez N, Castellanos-Rubio A, Plaza-Izurieta L *et al*: Coregulation and modulation of NFκB-related genes in celiac disease: uncovered aspects of gut mucosal inflammation. *Hum Mol Genet* 2013.

108. Plaza-Izurieta L, Fernandez-Jimenez N, Irastorza I *et al*: Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease. *Eur J Hum Genet* 2015; **23:** 1100-1105.

109. Mayhew TM, Myklebust R, Whybrow A, Jenkins R: Epithelial integrity, cell death and cell loss in mammalian small intestine. *Histol Histopathol* 1999; **14:** 257-267.

110. Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.

111. Núñez C, Dema B, Cénit MC *et al*: IL23R: a susceptibility locus for celiac disease and multiple sclerosis? *Genes Immun* 2008; **9:** 289-293.

112. Weersma RK, Zhernakova A, Nolte IM *et al*: ATG16L1 and IL23R are associated with inflammatory bowel diseases but not with celiac disease in the Netherlands. *Am J Gastroenterol* 2008; **103:** 621-627.

113. Einarsdottir E, Koskinen LL, Dukes E *et al*: IL23R in the Swedish, Finnish, Hungarian and Italian populations: association with IBD and psoriasis, and linkage to celiac disease. *BMC Med Genet* 2009; **10:** 8.

114. Warde-Farley D, Donaldson SL, Comes O *et al*: The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010; **38:** W214-220.

115. Parmar AS, Lappalainen M, Paavola-Sakki P *et al*: Association of celiac disease genes with inflammatory bowel disease in Finnish and Swedish patients. *Genes Immun* 2012; **13:** 474-480.

116. Lawlor G, Peppercorn MA: New genetic data support an association between celiac disease and inflammatory bowel disease. *Inflamm Bowel Dis* 2011; **17:** E80-81.