Chapter 6

Genetic Markers in Celiac Disease

Nora Fernández-Jiménez, Leticia Plaza-Izurieta, Jose Ramón Bilbao

Department of Genetics, Physical Anthropology and Animal Physiology, Basque Country University (UPV-EHU), BioCruces Research Institute, Bizkaia, Spain.

immunogenetics.let@gmail.com

Doi: http://dx.doi.org/10.3926/oms.232

How to cite this chapter

Fernandez-Jimenez N, Plaza-Izurieta L, Bilbao JR. *Genetic Markers in Celiac Disease.* In Rodrigo L and Peña AS, editors. *Celiac Disease and Non-Celiac Gluten Sensitivity.* Barcelona, Spain: OmniaScience; 2014. p. 103-121.

Abstract

Although the mode of inheritance of celiac disease (CD) is not completely understood, there is abundant evidence supporting the implication of genetic factors in susceptibility to CD and its heritability has been estimated to be of about 87%.

It has been known for a long time that certain HLA alleles are the major contributors to CD risk. However, despite playing a determinant role in the pathogenesis of the disease, their contribution to inheritance is modest (<50%) and it is believed that there must exist several non-HLA susceptibility loci, each one of them with a very small effect on the overall risk.

Consequently, during the last years, a great amount of effort has been made to locate and identify those additional susceptibility genes that might explain the genetics of the disease. Linkage studies in families, candidate gene association studies and (more recently) genome-wide association studies (GWAS) analyzing hundreds of thousands of Single-Nucleotide Polymorphisms (SNPs) have been performed. These approaches have identified several genes that are associated with CD, but not all of them have been confirmed in subsequent studies. Besides, the contribution of the identified genes remains modest, and a large part of the genetics of CD still remains to be clarified.

1. Introduction

Although the pattern of inheritance of celiac disease (CD) is still unknown, it has long been known that heredity is involved in the predisposition to the disease. Prevalence studies in affected families, especially those based on comparing twins, have been useful to estimate the proportions in which genetic and environmental factors contribute to the risk of disease development. According to these studies, Genetics plays an important role in both the initiation and subsequent development of CD. It is generally accepted that the proportion of pairs of monozygotic or identical twins in which both siblings suffer from the disease is of 75-86%, while among dizygotic or fraternal twins (who, like all siblings, share an average of 50% of the genome) this match is reduced to 16-20%. This difference between mono- and dizygotic twins has been used to calculate the magnitude of the genetic component in CD, which is higher than in other complex diseases of immunological origin, such as type 1 diabetes (about 30% concordance between identical twins and 6% for fraternal twins).[1] Furthermore, in CD, the correlation between sibling pairs and non-identical twins is almost the same, so that the environmental component would have a minimal effect on the risk of developing the disease. All this supports the idea that there is a strong genetic component in the development of celiac disease. At present, it is estimated that the heritability of CD (ratio of risk for a disease attributable to genetic factors versus environmental factors) is close to 87%.[2]

It has long been known that a large part of the genetic risk of developing CD is due to the presence of certain human leukocyte antigen (HLA) alleles. Despite their crucial role in the pathogenesis of the disease, the contribution of HLA to CD heritability is modest, so there has been a great deal of speculation about the existence of numerous susceptibility *loci* not linked to HLA, each of which would have a very small effect on the overall risk.

2. The HLA region and Celiac Disease

2.1. HLA Region

The Human Leukocyte Antigen or HLA is the name given to the Major Histocompatibility Complex (MHC) in humans. It is a *superlocus* located on the short arm of chromosome 6 that contains a large number of genes related to the immune system. HLA genes are responsible for encoding antigen-presenting proteins expressed on the surface of most human cells and constitute a major component in the ability to discriminate between self and non-self.

HLA genes influence the development of numerous inflammatory and autoimmune disorders, as well as susceptibility to infectious diseases, such as malaria and AIDS. However, due to the complexity of this region, the genetic components and specific pathogenic mechanisms for most of these diseases are unknown. The HLA region is one of the genome regions with the highest gene density. One explanation for this phenomenon is that, in this region, a high level of expression is favored.[3]

2.2. Contribution to genetic risk and susceptibility genes

As mentioned above, the HLA region is the most important CD susceptibility *locus* and accounts for about 50% of the heritability of the disease. The first evidence of association between HLA and CD was published in 1972 and was found using serological methods. Due to the high degree of linkage disequilibrium in the area, early studies identified HLA-A1, HLA-B8 and HLA-DR3 as the etiologic variants in the region, but molecular studies have shown that the factors directly involved are the HLA class II genes that code for HLA-DQ2 and HLA-DQ8 molecules. The association of HLA-DQ2 with the disease is the strongest; around 90% of celiac patients have at least one copy of the HLA-DQ2.5 heterodimer (formed by the combination of DQA1*05 and DQB1*02 alleles, responsible for encoding α and β heterodimer chains, respectively). On the other hand, 20-30% of the general population also carries this risk variant, demonstrating that although crucial for the disease, HLA-DQ2 alone is insufficient to develop it. The vast majority of CD patients lacking HLA-DQ2 carry the DQ8 variant present in the haplotype consisting of alleles DQA1*03:01 and DQB1*03:02.[4] A very small proportion of the patients are negative for both DQ2 and DQ8, but it has been observed that in most cases, these individuals have at least one of the two alleles encoding DQ2 molecule, i.e., DQA1*05 and DQB1*02.[4,5]

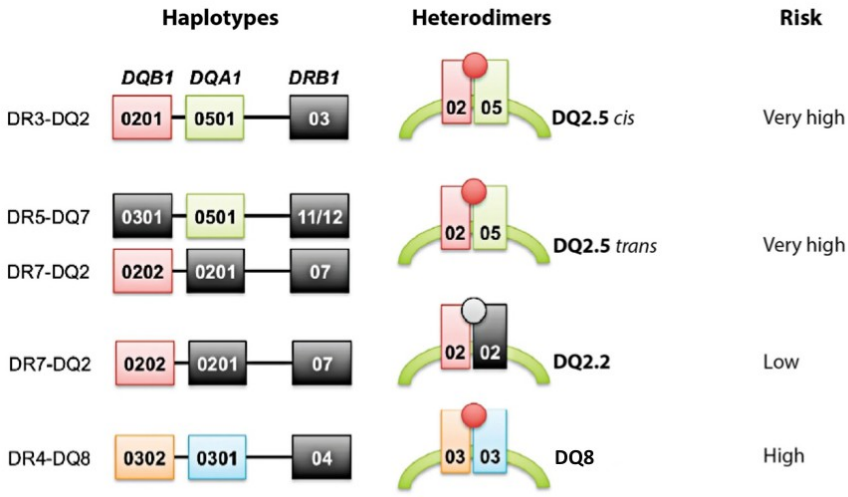

*Figure 1. Association of the HLA locus with CD. The HLA-DQ2 molecule is the major genetic risk factor for CD. Most celiac patients express the HLA-DQ2.5 heterodimer encoded by HLA-DQA1*05 (α chain) and HLA-DQB1*02 (β chain) alleles, which can be in cis in the haplotype DR3-DQ2 or in trans in DR5-DQ7 and DR7-DQ2.2 heterozygotes. The HLA-DQ2.2 dimer, an HLA-DQ2 variant (encoded by HLA-DQA1*02:01 and HLA-DQB1*02 alleles), confers a low risk of developing CD. Most DQ2-negative patients express HLA-DQ8 encoded by the haplotype DR4-DQ8.*

Risk variants DQ2 and DQ8 are in linkage disequilibrium (closely associated) with the HLA-DRB1 variants DR3 and DR4, respectively. Therefore, when referring to these risk variants, we might speak of haplotypes DR3-DQ2 and DR4-DQ8.[6] Haplotypes encoding the heterodimer HLA-DQ2.5

risk have been associated with CD in most populations (Figure 1). In certain haplotypes, such as DR3-DQ2, both alleles of the HLA-DQ2.5 heterodimer (DQA1*05:01 and DQB1*02:01) are located on the same chromosome and are encoded in cis. In heterozygous individuals with the DR5-DQ7 and DR7-DQ2 haplotypes, the two molecules are encoded in different chromosomes, or in *trans* (Figure 1). The differences between both HLA-DQ2.5 heterodimers affect an amino acid in the signal peptide of the DQα chains (DQA1*05:05 versus DQA1*05:01) and a residue in the membrane region of the DQβ chains (DQB1*02:01 versus DQB1*02:02) and seem not to have functional consequences, so they are attributed a similar risk. However, the risk conferred by another variant of the HLA DQ2 molecule, HLA-DQ2.2 dimer is very low (Figure 1).[7,8]

The degree of CD susceptibility is related to the number of DQ2.5 heterodimers. Individuals homozygous for DR3-DQ2 or DR3-DQ2/DR7-DQ2 heterozygotes express higher levels of DQ2.5 heterodimers and have maximum genetic risk of developing CD.[8-10] In this regard it is noteworthy that patients with refractory CD, unresponsive to the gluten-free diet, have a higher degree of DR3-DQ2 homozygosity (44-62%) compared to other celiac patients (20-24%). A similar allelic dose effect has been suggested for DQ8 molecules.

Together with the genes encoding DQ molecules, the HLA region contains other genes involved in the immune response that could also influence susceptibility to CD. Several studies have suggested that polymorphisms in genes such as *MICA*, *MICB* or *TNF* could contribute to the risk of developing the disease. However, most studies have not taken into account the high linkage disequilibrium between these genes and HLA-DQ and results are inconclusive. Sequencing and comprehensive mapping of the HLA region will help determine if it contains other susceptibility factors.

Despite the important contribution of HLA genes to the genetic risk, disease concordance for HLA identical siblings is only about 30%, so we can conclude that HLA genes are important but not sufficient to develop CD.[7]

2.3. Role in pathogenesis

The strong association of HLA class II genes with CD is explained by the fundamental role of CD4+ T lymphocytes in the pathogenesis of the disease. In fact, there are CD4+ T cells that recognize gluten peptides in the intestinal mucosa of celiac patients, but not in healthy individuals. These CD4+ cells present in the intestine of celiac patients are typically characterized by the HLA-DQ2 or-DQ8 molecules.[9]

When genetically susceptible individuals (expressing HLA-DQ2 or -DQ8 molecules) are exposed to certain gluten epitopes, these epitopes are presented by antigen presenting cells, stimulating the proliferation of gluten-specific CD4+ T cells.

An important milestone in the understanding of the molecular basis of the association of HLA with CD was the discovery that binding of HLA-DQ2 and -DQ8 molecules to gluten depends on enzymatic modifications of these peptides by the enzyme transglutaminase (TG2). This enzyme catalyzes a reaction that increases the negative charge of gluten epitopes, and enhances their binding to the HLA-DQ2 and -DQ8 molecules thus triggering the presentation of gluten peptides to T cells.

Given the importance of HLA molecules in the activation of autoreactive T cells against gluten, it makes sense that any distinct differences in their coding sequence may cause an alteration in any step of this process. Thus, polymorphisms in the sequence encoding the antigen-binding portion may cause changes in binding affinity, favoring the recognition of gluten peptides. Furthermore, certain polymorphisms located in regulatory regions may cause a sub-expression or over-expression of the HLA molecules, decreasing or increasing the immune response to gluten.

3. Search for genetic susceptibility genes in CD

In recent years, a great effort has been made to locate and identify susceptibility genes outside the HLA region and which may explain the Genetics of CD. For this purpose, two methods of analysis have been generally used: linkage studies in families and association studies. More recently, CD has been investigated by means of Genome-Wide Association Studies (GWAS) in which thousands of single nucleotide polymorphisms or SNPs have been analyzed. Through these studies, several genes associated with CD have been identified, but not all the observed associations have been subsequently confirmed.

3.1. Linkage regions and positional candidate genes

Linkage studies in families allow the identification of chromosomal regions repeatedly and consistently inherited by those family members affected by the disease through several generations. Through this type of analysis, the genome regions potentially involved in the pathogenesis of diseases can be further pinpointed. The genes located in these regions are positional candidate genes because their location confers upon them the suspicion of being involved in the pathogenesis of the disease.

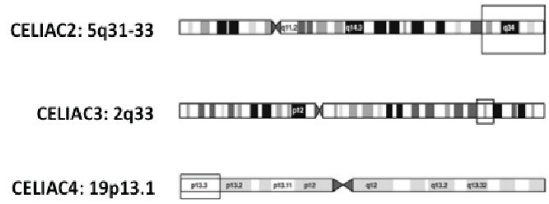

Figure 2. Linkage regions replicated in different family studies.

To date, four candidate regions linked to CD have been identified: the first is the HLA region or CELIAC1, which is the most important genetic component in CD and which has already been discussed in depth previously. The other three regions are called CELIAC2, CELIAC3 and CELIAC4 (Figure 2), but analyses of these loci have not always been conclusive and consistent.

3.1.1. CELIAC2

The CELIAC2 region is located on chromosome 5q31-33 and was first identified by Greco et al. in 1998.[11] The replication of this *locus* has not been universal, and no gene functionally implicated in the disease has been identified. This region contains a set of genes coding for several cytokines, which may play a role in regulating the immune system and inflammation.[12] Anyway, specific genes associated with CD have not been identified yet.

Several studies have focused on specific candidates such as the *IL12B* genes or the *SPINK* family of genes, but no consistent associations have been found for any of them.[13] Thirteen potentially functional variants of *IL4*, *IL5*, *IL9*, *IL13*, *IL17B* genes and *NR3C1*, all in the CELIAC2 *locus*, were genotyped in the Irish population, but none of these variants or haplotypes showed association with the disease.[14]

On the other hand, in an association study of genes selected because they are differentially expressed in the disease and are located in linkage regions, evidence of association with the *YIPF5* gene, also located in this region, was observed.[15] In a subsequent study in the Finnish and Hungarian populations linkage of this region with CD was confirmed, and evidence of association with *YIPF5* was observed again, although not in a consistent fashion.[16]

Despite being a major risk *locus* described in several linkage studies, no gene has been found which may explain its association with disease.

3.1.2. CELIAC3

CELIAC3 was first identified in 1999 by Holopainen et al.[17] This region is located in the 2q33 chromosomal region and it contains, among others, genes that regulate the *CD28*, *CTLA4* and *ICOS* lymphocyte responses, which will be discussed further on.

In this first study, seven different genetic markers were analyzed in 100 families. The *D2S116* microsatellite presented the highest nonparametric linkage score in this study, a furthermore, significant association between the marker and the disease was detected. The linkage between CD and this *locus* has been replicated in several subsequent studies using different genetic markers (microsatellites and SNPs), in addition to the abovementioned microsatellite.

The CELIAC3 *locus* contains the *CD28*, *CTLA4* and *ICOS* genes, which are located in a block of around 300kb that controls various aspects of the T cell response. The binding of *CD28* and *ICOS* to their respective ligands creates a positive signal for cytokine proliferation and activation, while the binding of *CTLA4* creates a signal that negatively regulates T cell activation. The association of the *CTLA4* gene with CD has been described in several populations, but results have not always been positive. A study in which all SNPs in this gene were analyzed suggests that haplotypes rather than SNPs are more strongly associated with the disease. However, data for these variables and/or haplotypes in the disease are needed to determine whether the association is with the *CTLA4* gene or with another neighboring gene.[18]

In the study that analyzed differentially expressed genes in the disease located at the linkage regions (mentioned above)[13], the gene that showed the strongest association with the disease was *SERPINE2*. This gene is important in the initial stages of extracellular matrix formation, a process that is altered in CD. A subsequent study was unable to replicate the association of the

disease with *SERPINE2* so, despite multiple attempts, the genetic factor that confers risk in the CELIAC3 region has not yet been identified.[19]

3.1.3. CELIAC4

The CELIAC4 *locus* is found on chromosome region 19p13.1 and was first identified by van Belzen et al in 2003.[20] This region contains more than 140 genes and some of them participate in the immune response and inflammation.

In the study, which identified CELIAC4, 82 families with affected members were analyzed. It was observed that the microsatellite D19S899 has a significant linkage peak with CD, with a LOD (*log of likelihood ratio*) score of 4.31. Besides, this genetic marker was significantly associated with the disease when 216 CD patients and 216 controls were analyzed to confirm the results obtained in the linkage study. However, not all subsequent replication studies have achieved positive results on the linkage of this region and CD.[21]

The best CELIAC4 region candidate is the myosin *IXB* gene (*MYO9B*), since it encodes a myosin molecule probably involved in enterocyte actin remodeling. The specific function of *MYO9B* is unknown, but it is known to contain a protein domain similar to that of the genes involved in tight junctions, so that it has been hypothesized that variations in this gene may result in the disruption of the intestinal barrier, thus allowing immunogenic peptides to cross.[22] However, not all association studies conducted have found a positive association with *MYO9B*. There are about 140 additional genes in the region, some of which are involved in immunity and inflammation (*CYP4F3*, *HSH2D*, *IL12RB1*, *IFI30* and *KIR*, for example) and which might be good candidates. A study that analyzed ten genes from this region in a Dutch population found evidence of association with the *CYP4F3* and *CYP4F2* genes, both involved in the inhibition of leukotriene, a potent inflammatory mediator.[23] *ICAM-1*, a gene found in this region that is important for intercellular adhesion, also showed association in a French population.[24] These weak associations must be replicated in independent populations in order to determine the contribution of these genes to the development of the disease.

3.2. Functional candidate genes

3.2.1. Innate immune response genes

The involvement of the innate immune system in the development of CD is increasingly evident, therefore several of the innate response genes have been studied in search of risk polymorphisms. One study analyzed functional polymorphisms located in regulatory regions of different proinflammatory mediators (*IL-1α*, *IL-1β*, *IL-1RN*, *IL-18*, *RANTES* and *MCP-1*). None of the genes analyzed in this study, except *RANTES*, which has a dubious association, was connected with the risk of developing CD.[25]

The KIR (Killer Immunoglobulin-like receptors) gene family has also been studied in CD because it contains innate immune response candidate genes. These receptors are located in the 19q13.4 region, which has presented evidence for linkage with the disease and encode receptors for NK (Natural Killer) and certain T cells that modulate cytolytic activity through interaction with HLA class I ligands, participating in the innate immune response. The gene content, genotypes and

haplotypes in the KIR genes from a Basque population were analyzed and it was observed that the frequency of the *KIR2DL5B*(+)/*KIR2DL5A*(-) combination was significantly higher in individuals with CD. This association was replicated in a Spanish population (odds ratio 3.63) suggesting the involvement of the KIR2DL5B gene with an increased risk of CD, probably due to the lack of an efficient inhibitory signal.[26] On the other hand, another study found that the *KIR3DL1* inhibitor gene was overexpressed in the intestinal mucosa in active disease, presumably due to increased subpopulations of lymphocytes with an NK phenotype.[27]

Toll-like receptors (TLR), which take part in pathogen recognition and immune response stimulation, have also been analyzed in search of association with CD. Although it has been shown that their expression is altered in patients, no association was found between polymorphisms in these genes and CD. Similarly, no association was found for copy number variation (CNV) of *TLR2* and *TLR4* with the disease.[28]

In turn, β-defensins form a cluster with variable number of copies in the population and are part of the innate immune response, acting as natural antibiotics. The genes that comprise this family have been previously associated with autoimmune and inflammatory diseases such as psoriasis or Crohn's disease. Although no association has been detected between SNPs in these genes and CD, there is an association between the copy number of the gene cluster and CD, since a lower presence of high copy numbers was observed (>4) among patients, suggesting a protective role of β-defensins in the disease.[28]

As mentioned previously, stress response genes *MICA* and *MICB* have also been studied in search of risk variants, but the location of these genes in the CELIAC1 *locus* has hindered conclusions about their independent contribution, due to high linkage disequilibrium in the HLA region.[29]

Although the innate immune system is activated in celiac patients, none of the candidate genes studied exhibited a strong association with the disease, so it may be assumed that many genes of the innate immune system, each with a weak effect, contribute to the development of disease activating the innate response.

3.2.2. Adaptive immune response genes

The Th1 response is one of the major inflammatory responses in CD and the characteristic cytokine of this type of response is IFN-γ. Production of this cytokine is significantly increased in active disease, reaching 240-fold higher levels in cases with total atrophy. The *IFNG* gene was studied in three Dutch and Finnish population cohorts, no differences between the allelic distributions of cases and controls were found. So far, there is no evidence that *IFNG* variants might predispose to the disease, despite its being highly overexpressed in the mucosa of celiac patients.[30]

Th17 cells have also been implicated in CD pathogenesis. Signaling by means of IL23 and its receptor (IL23R) is a key element in the differentiation of T cells towards Th17 cells. The *IL23R* gene has been associated with other autoimmune and/or inflammatory diseases such as psoriasis or ulcerative colitis. A coding variant in *IL23R* gene was analyzed in a Dutch population but was not associated with the disease.[31] However, the analysis of this same variant in a Spanish population showed an increase of the minor allele in patients, as opposed to what was observed in other diseases.[32] A later study found evidence of linkage in the *IL23R* gene region in Hungarian, Finnish and Italian populations, but no association was found with the studied polymorphisms.[33]

However, a recent study in Spanish populations in which the association of 101 SNPs in 16 genes related to the Th17 response (including *IL23R*) indicates that there is no association with the disease.[34]

On the other hand, the *CIITA* gene appears to be the major regulator of HLA class II genes. This gene has a complex expression pattern and two polymorphisms located in its promoter have been associated with other autoimmune diseases. These polymorphisms were analyzed in a Spanish CD cohort but no significant differences between patients and controls were detected.[9] On the contrary, the second GWAS does show an association between CD and the region containing the *CIITA* gene.

To date, no candidate gene from the adaptive immune response has been strongly associated with risk of developing CD.

3.2.3. Genes involved in intestinal epithelium remodeling

It has been reported that the permeability of the intestinal epithelium is increased in CD patients in response to gliadin. This alteration of the intestinal barrier is associated with structural changes in intercellular junctions. Due to its possible role in intestinal epithelium remodeling, the *MYO9B* gene in the CELIAC4 linkage region has been scanned for disease-associated polymorphisms.[23] A 2008 study analyzed 197 SNPs from 41 genes associated with intercellular communication in Dutch and British populations. Two of the genes, *PARD3* (2 SNPs) and *MAGI2* (2 SNPs) showed weak association with the disease in the Dutch population. Replication in a British population confirmed association with one *PARD3* SNP. The combined analysis of both populations confirmed the association for both genes with Odds ratio values of 1.23 for *PARD3* and 1.19 for *MAGI2*. These genes also showed positive association with ulcerative colitis, suggesting a common causal defect in the intestinal barrier for both diseases.[36]

3.2.4. Cell signaling pathways

Several signaling pathways are altered in CD, including the Jak-Stat signaling pathway, the kappa B (NFkB) transcription factor signaling pathway, the MAPK signaling pathway or the transforming growth factor beta (TGFB) signaling pathway.[15] Several genes of these pathways have been analyzed in search of an association with CD.

One of these genes is STAT1, whose expression is altered in the disease and is also a positional candidate since it is found in the CELIAC3 *locus*. An analysis was performed of five tag polymorphisms covering the entire gene in a Dutch population, but there was no evidence of association with CD.[37]

The *NFKB1* gene has also been studied in search of a genetic association with CD, but despite the fact that this transcription factor is constitutively active in the mucosa of celiac patients, no polymorphisms have been found to explain its increased activity in the disease. It has been suggested that the pathogenic effects attributed to this transcription factor may be caused by a regulatory defect instead of a polymorphism in the transcription factor itself. Genes located upstream in the biological cascade may be responsible for the increased genetic risk generating a higher NFkB-dependent transcriptional activity. Two of the genes identified in a GWAS follow-up

study (*REL* and *TNFAIP3*) are located in this cascade and may be responsible for its deregulation. Recently, a regulatory polymorphism in *UBD*, a gene involved in NFkB activation has been associated with the disease in a Spanish population study. This gene is overexpressed in patients with active disease and the associated allelic polymorphism has a significant correlation with gene expression levels.[38]

The modifications observed in these complex biological pathways can alter the expression of genes located further downstream in the route, so that the analysis of individual genes can give rise to error. An exhaustive analysis of these routes may be crucial for the selection of association study candidates.

3.2.5. Extracellular Matrix

The extracellular matrix appears degraded in the intestinal epithelium of celiac patients. Metalloproteinases are enzymes that degrade matrix components, and it has been recorded that their expression is increased in the active stages of the disease, contributing to the morphological alterations of the intestinal mucosa. Therefore, these genes have been studied on several occasions in search of susceptibility variants. In any case, functional polymorphisms of the *MMP-1* gene have not been associated with CD.[39]

4. Genome-wide association studies in celiac disease

GWAS allow for fast scanning of markers in complete sets of DNA or genomes of several individuals, with the purpose of finding genetic variations associated with a particular disease. Having identified these genetic associations, researchers can use this information to develop new and improved technologies to detect, treat and prevent diseases. These studies are especially useful in finding genetic variations that contribute to the development of common and complex diseases, such as asthma, cancer, diabetes, and (in this case) CD.

In order to conduct a GWAS, researchers use two groups of participants: individuals with the disease under study and individuals with characteristics similar to those above but who do not have the disease. This is an association study on a genome-wide scale.

The full DNA or genome of each individual is purified from a blood sample. This DNA is placed on a chip and is automatically scanned in the laboratory. These devices strategically inspect the samples looking for genetic variation markers, in this case, SNPs.

If it is discovered that certain genetic variations turn out to be significantly more frequent in patients than in healthy individuals, it is said that these variations are associated with the disease. These associated genetic variations may be important markers since they could point to the region in the human genome wherein lies the variation responsible for the disease. The associated variant itself need not necessarily be the direct cause of the disease; it could simply be pointing to the region where for the true causal variant should be sought. Due to this, in most cases it may be necessary to continue with the investigation, for example, by sequencing this particular region so as to identify the exact genetic variant implicated in the disease, or by

performing functional studies in order to find an association between specific variants and gene expression levels.

GWAS allow the definition of a new class of genetic variants associated with diseases. Association studies based on pedigrees use families in which clusters associated with the disease are useful to identify rare variants with great risk effect. On the other hand, GWAS depend on population-based samples and therefore require common variants with a more modest effect (since it will not be feasible to observe rare variants), which could not be observed using a traditional linkage-based approach.

4.1. Outcome of the first GWAS

In the first genome-wide study conducted on CD, 778 individuals with CD and 1,422 healthy controls were studied. Association analyses were performed on 310,605 SNPs with a minor allele frequency above 1%.[40]

As expected, the largest association was found around the HLA *locus*. The *rs2187668-A* allele was shown to be an efficient marker for HLA-DQ2.5cis, the most common HLA DQ2 haplotype associated with CD. In this first study, it was shown that 89.2 % of patients in the UK had one or two copies of HLA-DQ2.5cis, compared to 25.5 % in the control population.

Outside the HLA region a number of associated SNPs higher than what would be expected by chance alone was observed, 56 SNPs had an association with $p<10^{-4}$. Some of these SNPs are located close together, which suggests that the excess of SNPs with low p-values may be due to a true association of SNPs in linkage disequilibrium with the disease-causing variants.

The only SNP outside HLA that demonstrated significant association was *rs13119723*, in the 4q27 region, located in a linkage disequilibrium block containing the *IL2* and *IL21* genes. These results were repeated in collections of Dutch and Irish patients and controls.

It was estimated that the *IL2-IL21* alone could explain only 1% of the increased familial risk for CD, suggesting the existence of other susceptibility genes that had not yet been identified. For this reason, a study was undertaken to analyze the 1,164 most significant SNPs from the first study in a further 1,643 CD cases and 3,406 non-celiac controls from three independent European collections.[41] The associated regions identified in this new study were scrutinized for candidate genes that could play a role in the development of CD, especially those genes somehow implicated in the immune response (Figure 3).

It is important to be able to replicate the results of genetic discoveries in different populations when establishing a genetic effect in the predisposition to a disease. For this reason, efforts have been made to replicate, in several independent populations, the results obtained in the first GWAS, with different results, possibly due to population variations and the sample size of each one of these studies.

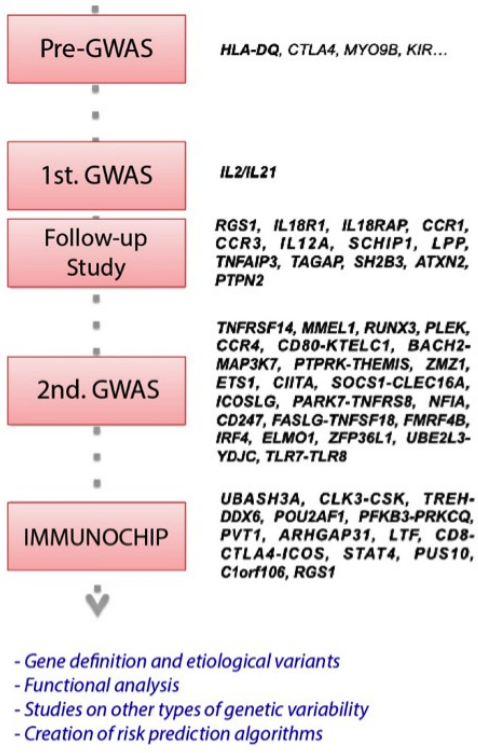

Figure 3. Advances in the Genetics of CD. After the Immunochip study, 40 loci that contribute to the risk of CD have been identified. Now is the time to perform functional studies in order to identify etiological variants and determine the practical applications of the association results (blue).

4.2. Outcome of the second GWAS

The second GWAS on CD was performed in 2009. To this end, an analysis was performed on 292,387 SNPs outside the HLA region in DNA samples from 4,533 individuals with celiac disease and 10,750 healthy controls of European origin. In addition, 231,362 non-HLA SNPs were also studied in 3,796 celiac patients and 8,154 controls.[42]

Thirteen new risk regions with significant evidence of association (Figure 3) were identified. There are several genes with immune functions in these regions: *BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLCD16A*, *ETS1*, *ICOSLG*, *RUNX3*, *THEMIS*, *TNFRSF14* and *ZMIZ1*. Another thirteen regions did not achieve significant association but did point to a trend and also contain genes with immune functions, including *CD247*, F*ASLG-TNFSF18-TNFSF4, IRF4*, *TLR7-TLR8, TNFRSF9* and *YDJC*.

4.3. Immunochip

The most recent large-scale project performed to identify variants associated with CD and other autoimmune diseases is the *Immunochip Project*. Regarding CD, more than 200,000 variants from approximately 12,000 celiac patients and 12,000 controls from 7 geographical regions were analyzed.[43]

The analysis was performed on 183 loci related to the immune system that are outside the HLA region; 39 showed significant association with CD; the 26 regions identified in the GWAS plus 13 new *loci*. All associated variants have a minor allele frequency above 5%, that is, they are all common variants. Low frequency variants associated with the disease have only been detected in 4 loci. The advantage of the Immunochip over GWAS lies in the possibility of fine mapping the regions to locate and identify causal signals, due to the fact that in the Immunochip genotyping is much denser. One example of this is that out of 54 independent signals outside HLA that are found in high density in the 36 genotyped *loci*, 29 are located around a single gene (Figure 3).

After functional annotation of associated regions, one of the main conclusions that have been reached is that there are very few markers in the coding regions of genes, although some markers are close to transcription start sites and others in the 3 'UTR regions.

Some potentially causative genes proposed due to the existence of signals near the 5' or 3' regulatory regions are *THEMIS/PTPRK*, *TAGAP*, *ETS1*, *RUNX3* and *RGS1*. Some of them had already been proposed after the previous GWAS.

4.4. Replication of association studies and functional analysis of candidate genes

In 2011, 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.[44] These 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.

A second work takes a step forward and identifies two genes (*PTPRK* and *THEMIS*), located in the same associated region, which are co-expressed both in active disease and in response to *in vitro* stimulation by gliadin from intestinal biopsies of celiac patients with inactive disease who have adhered to the diet for at least two years.[45] Therefore, it seems that the variants associated in this region affect the expression of different genes, not constitutively from the time of birth of the future celiac patient, but after a toxic stimulus triggers an immune response.

The implications of this finding are of great importance because they highlight the existence of common regulatory mechanisms for different genes in the DNA sequence that only have an effect in the presence of a disease-provoking immunogenic stimulus.

These and other studies emphasize the need for functional studies and to avoid the selection of hypothetical susceptibility genes using arbitrary criteria. Similarly, this reveals that much about the immense complexity of the regulatory genome remains to be discovered and it opens the door to comprehensive analysis of the noncoding genome variants, the study of nonmessenger

RNA molecules and to the levels of expression of their trans targets at outermost positions of the genome.

4.5. Conclusions

Despite the enormous efforts of the past decades, genetic and molecular mechanisms underlying this disease have not yet been fully explained. GWAS and subsequent studies have begun to unravel the genetic contribution to the pathogenesis of CD. Although from the genetic standpoint, diseases of immune etiology show wide differences in the number of loci involved, the effect of each of these and the environmental factors involved, it is true that there is a strong overlap between this family of disorders. This overlap must involve the participation of common biological pathways and suggests that strategies for treatment may also be shared. However, the interpretation of association studies must be done with caution since it is true that each of the identified loci contains more than one gene. Strategies to identify potential etiological variants are indicated in Figure 3, and could, in the future, identify functional alterations underlying autoimmune diseases. Over time, these pathogenic variants may be included in risk prediction algorithms and allow for the diagnosis of individuals with a high genetic predisposition before the onset of symptoms, which could result in an improved quality of life and decreased healthcare costs. In addition, they could open the door to new therapeutic targets for CD itself and for other diseases of autoimmune etiology.

References

1. Sollid LM. Thorsby E. *HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis*. Gastroenterol. 1993; 105: 910-22.
2. Greco L. Romino R. Coto I. et al. *The first large population based twin study of coeliac disease*. Gut. 2002; 50: 624-8. http://dx.doi.org/10.1136/gut.50.5.624
3. Horton R. Wilming L. Rand V. Lovering RC. Bruford EA. Khodiyar VK. et al. *Gene map of the extended human MHC*. Nat Rev Genet. 2004; 5: 889-99. http://dx.doi.org/10.1038/nrg1489
4. Karell K. Louka AS. Moodie SJ. Ascher H. Clot F. Greco L. et al. *HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease*. Hum Immunol. 2003; 64: 469-77. http://dx.doi.org/10.1016/S0198-8859(03)00027-2
5. 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: 188-92. http://dx.doi.org/10.1016/0198-8859(92)90104-U
6. Sollid LM. Thorsby E. *Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer*. J Exp Med. 1989; 169: 345–50. http://dx.doi.org/10.1084/jem.169.1.345
7. Sollid LM. *Coeliac disease: dissecting a complex inflammatory disorder.* Nat Rev Immunol. 2002; 2: 647–55. http://dx.doi.org/10.1038/nri885
8. 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: 215-20. http://dx.doi.org/10.1038/sj.gene.6364061
9. Ploski R. et al. *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: 173-7. http://dx.doi.org/10.1111/j.1399-0039.1993.tb01998.x
10. Lundin KE. Scott, H. Hansen T. Paulsen G. Halstensen TS. Fausa O. 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: 187-96. http://dx.doi.org/10.1084/jem.178.1.187
11. Greco L. Corazza GR. Babron MC. Clot F. Fulchignoni-Lataud MCV. Percopo S. et al. *Genome search in celiac disease.* Am J Hum Genet. 1998; 62: 35-41. http://dx.doi.org/10.1086/301754
12. Greco L. Babron MC. Corazza GR. Percopo S. Sica R. Clot F. et al. *Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families.* Ann Hum Genet. 2001; 65: 35-41. http://dx.doi.org/10.1046/j.1469-1809.2001.6510035.x
13. Seegers D. Borm ME. Van Belzen MJ. et al. *IL12B and IRF1 gene polymorphisms and susceptibility to celiac disease.* Eur J Immunogenet. 2003; 30: 421-5. http://dx.doi.org/10.1111/j.1365-2370.2003.00428.x
14. Ryan AW. Thornton JM. Brophy K. Daly JS. McLoughlin RM. O'Morain C. et al. *Chromosome 5q candidate genes in coeliac disease: genetic variation at IL4, IL5, IL9, IL13, IL17B and NR3C1.* Tissue Antigens. 2005; 65: 150-5. http://dx.doi.org/10.1111/j.1399-0039.2005.00354.x
15. Castellanos-Rubio A. *Combined functional and positional gene information for the identification of susceptibility variants in celiac disease.* Gastroenterol. 2008; 134: 738-46. http://dx.doi.org/10.1053/j.gastro.2007.11.041

16. Koskinen LL. Einarsdottir E. Korponay-Szabo IR. et al. *Fine mapping of the CELIAC2 locus on chromosome 5q31-q33 in the Finnish and Hungarian populations.* Tissue Antigens. 2009; 74: 408-16. http://dx.doi.org/10.1111/j.1399-0039.2009.01359.x
17. Holopainen P. Naluai AT. Moodie S. Percopo S. Coto I. Clot F. et al. *Candidate gene region 2q33 in European families with coeliac disease.* Tissue Antigens. 2004; 63: 212-22. http://dx.doi.org/10.1111/j.1399-0039.2004.00189.x
18. Brophy K. Ryan AW. Thornton JM. et al. *Haplotypes in the CTLA4 region are associated with coeliac disease in the Irish population.* Genes Immun. 2006; 7: 19-26. http://dx.doi.org/10.1038/sj.gene.6364265
19. Dema B. Martínez A. Fernández-Arquero M. Maluenda C. Polanco I. De la Concha EG. et al. *Lack of replication of celiac disease risk variants reported in a Spanish population using an independent Spanish sample.* Genes Immun. 2009; 10: 659-61. http://dx.doi.org/10.1038/gene.2009.54
20. Van Belzen MJ. Meijer JWR. Sandkuijl LA. Bardoel AFJ. Mulder CJJ. et al. *A major non-HLA locus in celiac disease maps to chromosome 19.* Gastroenterol. 2003; 125: 1032-41. http://dx.doi.org/10.1016/S0016-5085(03)01205-8
21. Capilla A. Donat E. Planelles D. Espinós C. Ribes-Koninckx C. Palau F. *Genetic analyses of celiac disease in a Spanish population confirm association with CELIAC3 but not with CELIAC4.* Tissue Antigens. 2007; 70: 324-9. http://dx.doi.org/10.1111/j.1399-0039.2007.00899.x
22. Monsuur AJ. De Bakker PIW. Alizadeh BZ. Xhernakova A. Bevova MR. Strengman E. et al. *Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect.* Nat Genet. 2005; 37: 1341-4. http://dx.doi.org/10.1038/ng1680
23. Curley CR. Monsuur AJ. Wapenaar MC. Rioux JD. Wijmenga C. *A functional candidate screen for coeliac disease genes.* Eur J Hum Genet. 2006; 14: 1215-22. http://dx.doi.org/10.1038/sj.ejhg.5201687
24. Abel M. Cellier C. Kumar N. Cerf-Bensussan N. Schmitz J. Caillat-Zucman S. *Adulthood-onset celiac disease is associated with intercellular adhesion molecule-1 (ICAM-1) gene polymorphism.* Hum Immunol. 2006; 67: 612-7. http://dx.doi.org/10.1016/j.humimm.2006.04.011
25. Rueda B. Zhernakova A. López-Nevot MA. Martín J. Koeleman BPC. *Association study of functional genetic variants of innate immunity related genes in celiac disease.* BMC Med Genet. 2005; 3: 6-29. http://dx.doi.org/10.1186/1471-2350-6-29
26. Santin I. Castellanos-Rubio A. Perez de Nanclares G. Vitoria JC. Castaño L. Bilbao JR. *Association of KIR2DL5B gene with celiac disease supports the susceptibility locus on 19q13.4.* Genes Immun. 2007; 8: 171-6. http://dx.doi.org/10.1038/sj.gene.6364367
27. Fernandez-Jimenez N. et al. *Upregulation of KIR3DL1 gene expression in intestinal mucosa in active celiac disease.* Hum Immunol. 2011; 72: 617-20. http://dx.doi.org/10.1016/j.humimm.2011.04.008
28. Fernandez-Jimenez N. Santín I. Irastorza I. Plaza-Izurieta L. Castellanos-Rubio A. Vitoria JC. Bilbao JR. *Analysis of beta-defensin and Toll-like receptor gene copy number variation in celiac disease.* Hum Immunol. 2010; 71: 833-6. http://dx.doi.org/10.1016/j.humimm.2010.05.012
29. Martín-Pagola A. Pérez-Nanclares G. Ortiz L. Vitoria JC. Hualde I. Zaballa R. et al. *MICA response to gliadin in intestinal mucosa from celiac patients.* Immunogenetics. 2004; 56: 549-54. http://dx.doi.org/10.1007/s00251-004-0724-8

30. Wapenaar MC. Van Belzen MJ. Fransen JH. Fariña Sarasqueta A. Houwen RHJ. Meijer JWR. et al. *The interferon gamma gene in celiac disease: augmented expression correlates with tissue damage but no evidence for genetic susceptibility.* J. Autoimmun. 2004; 23: 183-90. http://dx.doi.org/10.1016/j.jaut.2004.05.004
31. Weersma RK. Zhernakova A. Nolte IM. Lefebvre C. Rioux JD. Mulder F. et al. *ATG16L1 and IL23R are associated with inflammatory bowel diseases but not with celiac disease in the Netherlands.* Am J Gastroenterol. 2008; 103: 621-7. http://dx.doi.org/10.1111/j.1572-0241.2007.01660.x
32. Núñez C. Dema B. Cénit MC. Polanco I. Maluenda C. Arroyo R. et al. *IL23R: a susceptibility locus for celiac disease and multiple sclerosis?* Genes Immun. 2008; 9: 289-93. http://dx.doi.org/10.1038/gene.2008.16
33. Einarsdottir E. Koskinen LLE. Dukes E. Kainu K. Suomela S. Lappalainen M. 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; 28: 10-8. http://dx.doi.org/10.1186/1471-2350-10-8
34. Medrano LM. García-Magariños M. Dema G. Espino L. Polanco I. Figueredo MA. et al. *Th17-related genes and celiac disease susceptibility.* PLoS One. 2012; 7: e31244. http://dx.doi.org/10.1371/journal.pone.0031244
35. Dema B. Martínez A. Fernández-Arquero M. Maluenda C. Polanco I. Figueredo MA. et al. *Autoimmune disease association signals in CIITA and KIAA0350 are not involved in celiac disease susceptibility.* Tissue Antigens. 2009; 73: 326-9. http://dx.doi.org/10.1111/j.1399-0039.2009.01216.x
36. Wapenaar MC. Monsuur AJ. Van Bodegraven AA. Weersma RK. Bevova MR. Linskens RK. et al. *Associations with tight junction genes PARD3 and MAGI2 in Dutch patients point to a common barrier defect for coeliac disease and ulcerative colitis.* Gut. 2008; 57: 463-7. http://dx.doi.org/10.1136/gut.2007.133132
37. Diosdado B. Monsuur AJ. Mearin ML. Mulder C. Wijmenga C. *The downstream modulator of interferon-gamma, STAT1 is not genetically associated to the Dutch coeliac disease population.* Eur J Hum Genet. 2006; 14: 1120-4. http://dx.doi.org/10.1038/sj.ejhg.5201667
38. Castellanos-Rubio A. Santin I. Irastorza I. Sanchez-Valverde F. Castaño L. Vitoria JC. et al. *A regulatory single nucleotide polymorphism in the ubiquitin D gene associated with celiac disease.* Hum Immunol. 2010; 71: 96-9. http://dx.doi.org/10.1016/j.humimm.2009.09.359
39. Ciccocioppo R. Di Sabatino A. Bauer M. Della Riccia DN. Bizzini F. Biagi F. et al. *Matrix metalloproteinase pattern in celiac duodenal mucosa.* Lab. Invest. 2005; 85: 397-407. http://dx.doi.org/10.1038/labinvest.3700225
40. Van Heel DA. et al. *A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21.* Nat Genet. 2008; 39: 827-9. http://dx.doi.org/10.1038/ng2058
41. Hunt, KA. Zhernakova A. Turner G. Heap GAR. Franke L. Bruinenberg M. et al. *Newly identified genetic risk variants for celiac disease related to the immune response.* Nat Genet. 2008; 40: 395-402. http://dx.doi.org/10.1038/ng.102
42. Dubois PC. Trynka G. Franke L. Hunt KA. Romanos J. Curtotti A. et al. *Multiple common variants for celiac disease influencing immune gene expression.* Nat Genet. 2010; 42: 295-302. http://dx.doi.org/10.1038/ng.543

43. Trynka G. Hunt KA. Bockett NA. Romanos J. Mistry V. Szperl A. et al. *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nat Genet. 2011; 43: 1193-201. http://dx.doi.org/10.1038/ng.998
44. 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.* J Med Genet. 2011; 48: 493-6. http://dx.doi.org/10.1136/jmg.2011.089714
45. Bondar C. Plaza-Izurieta L. Fernandez-Jimenez N. Irastorza I. Withoff S. CEGEC. Wijmenga C. Chirdo F. Bilbao JR. *THEMIS and PTPRK in celiac intestinal mucosa: coexpression in disease and after in vitro gliadin challenge*. Eur J Hum Genet. 2013; 22: 358-62. http://dx.doi.org/10.1038/ejhg.2013.136