Jihane Romanos

Genetics of celiac disease and its diagnostic value.

Thesis, University of Groningen, with summary in English, Dutch, Portuguese, French, and Arabic.

The research presented in this thesis was mainly performed at the Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands, and was financially supported by KP6 EU grant 036383 (PREVENTCD).

Printing of this thesis was financially supported by: Rijksuniversiteit Groningen, University Medical Center Groningen, Groningen University Institute for Drug Exploration (GUIDE), and Celiac Disease Consortium.

Cover design and layout by Claudia Marcela Gonzaleza Arevalo (e-mail: argo1983@gmail.com).
Printed by EIKON PLUS, Krakow, Poland.



ISBN: 83-60391-70-X

GUIDE

cdc
celiac disease consortium

PREVENTCD
prevent coeliac disease

umcg

rijksuniversiteit
groningen

RIJKSUNIVERSITEIT GRONINGEN

# Genetics of celiac disease and its diagnostic value

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
maandag 12 december 2011
om 14.30 uur

door

**Jihane Romanos**

geboren op 23 januari 1982
te Haret-Sakhr, Libanon

| | |
|---|---|
| Promotor: | Prof. dr. C. Wijmenga |
| Beoordelingscommissie: | Prof. dr. G.H. de Bock |
| | Prof. dr. A.J. Oldehinkel |
| | Prof. dr. H.J. Verkade |

*"If you cannot work with love but only with distaste,*
*it is better that you should leave your work"*

Khalil Gibran

Paranimfen: Fany Messanvi
M. Elena Merlo

To Teta Vola, Teita Jeanette and Jido Georges

# Table of contents

# Preface and outline of the thesis

Before the Human Genome Project was completed in 2003, scientific and medical attention for genetic disorders was mainly focused on understanding rare single-gene disorders, such as Huntington's disease, Duchene muscular dystrophy, and cystic fibrosis, as well as on chromosomal abnormalities. In recent years, with the further completion of the international HapMap Project and the development of new methods for genotyping individual DNA samples for more than 500,000 markers, the attention of genomics and genetic researchers has shifted toward understanding the basis of common multifactorial disorders, such as celiac disease, type 2 diabetes, coronary heart disease and cancer.

Celiac disease (CD) is one of the most common inflammatory disorders of the small intestine caused by permanent gluten intolerance in susceptible individuals. It affects around 1% of Western populations but remains largely unrecognized. The only treatment is a lifelong gluten-free diet. The age at onset ranges from infancy to late adulthood, and the clinical presentation of the disease is highly variable ranging from common gastrointestinal symptoms like diarrhea and abdominal pain to a more systemic non-reversible symptoms like anemia, osteoporosis, and infertility.[1] Untreated CD are at increased risk to develop other autoimmune diseases like type 1 diabetes and autoimmune thyroiditis in addition to having higher rate of mortality compared to the general population.[2, 3]

CD is a multifactorial disorder: several genetic factors combined with an environmental trigger are necessary for the disease to develop. The major, well-studied genetic risk factor for CD is the human leukocyte antigen (HLA), more specifically the HLA-DQ locus, coding for the HLA-DQ2 and HLA-DQ8 molecules. These heterodimers contribute to 35-40% of CD etiology. Other non-HLA genes have been reported to be associated to CD but they have only a modest effect.

An important benefit from the study of the genetics of human disease is to be able to predict the risk that individuals may have of succumbing to a particular disease. Genetic testing of monogenic diseases – where there is a strong correlation between risk genotype and disease – has been employed successfully in a diverse range of applications from prenatal and newborn screening, to carrier testing and medical diagnosis. With the success of genome-wide association studies and the promises of whole-genome sequencing, attention has now shifted to translating this new wave of basic genetic knowledge into personalized medicine.

In this thesis, my overall aim is to discuss the identification of genetic risk variants for CD, replicate them in new populations, and develop a risk model which can improve diagnosis of CD. In the introduction, **chapter 1**, I describe the history of CD, its wide spectrum of clinical features, and its current diagnosis, pathogenesis and genetic background. More than 95% of patients carry HLA-DQ2 and/or DQ8 molecules, however, 30-40% of the general population also carry these molecules. Thus HLA is necessary but not

sufficient to develop the disease. It has a sensitivity of over 96% in most populations implying that individuals without HLA-DQ2 and/or DQ8 are unlikely to develop the disease.[4] A majority of members of the European Society of Pediatric Gastroenterology and Nutrition (ESPGHAN) have demanded modification of the current CD diagnostic criteria in order to include HLA testing as an additional screening parameter.[5] In **chapter 2**, I validate a novel HLA-genotyping method in three European populations. This method was developed in a Dutch population and used six HLA-tagging single nucleotide polymorphisms (SNPs); it is suitable for high-throughput approaches.[6]

The first genome-wide association study (GWAS) on CD and its follow-up identified 8 non-HLA loci that contribute significantly to CD risk.[7-9] In **chapter 3**, I replicate these findings in 538 cases and 593 controls from Italy and show that CD risk loci are differently associated in different populations. For example, CCR3 and IL18RAP are associated in UK and Dutch populations, but not in Italian and Irish cohorts. Different genes may be implicated in different populations due to human migration and genetic drift. A second GWAS and a fine-mapping project identified a total of 57 non-HLA SNPs to contribute to CD development.[10, 11]

Advances in technology and increased knowledge on biology and genetic risk factors for common complex diseases have led to the creation of genetic risk models that can be used to target diagnostic, preventive, and therapeutic interventions based on a person's genetic risk, or to complement existing risk models based on non-genetic factors, like family history or the presence of other diseases. As CD is a major socio-economic burden on patients, their families and society, improved diagnosis and early prevention would ease these negative effects. **Chapter 4** shows how testing multiple genetic loci simultaneously, which collectively result in superior prediction of CD, might be used as a diagnostic or screening tool to prevent long-term and irreversible complications. **Chapter 5** describes the genetic risk profile which I developed for CD based on HLA and non-HLA risk alleles, using cases and controls from our first GWAS. The study showed that using 10 non-HLA risk alleles can improve identification of high-risk individuals. To improve and validate the genetic risk model, I have increased the number of SNPs in the model to 26 and 57 variants and tested the model with 26 variants on two different cohorts (a nested case-control and a prospective cohorts) in **chapter 6**.

Genomic profiling for CD might not yet be applicable in clinical practice, but with some improvement, it might become part of the future diagnosis and treatment of CD. In **chapter 7**, I discuss who can benefit from this genetic profiling and how to improve the accuracy of the risk prediction. As a first step, this genetic profiling can mainly improve the diagnosis of CD in individuals at high-risk, such as first-degree relatives and individuals with other immune-mediated diseases. Maybe one day it will also be a good screening test for selecting newborns who could benefit from early intervention to prevent CD. This model can be improved by including more susceptibility variants, which can be rare, population-specific, have a parental origin effect, causative for an endo-phenotype of CD, or pathway-specific.

In conclusion, my thesis shows that risk profiling for CD may well have an application for identifying individuals at high risk for CD. Genomic profiling might lead to a future where there is personalized medicine for CD patients, where individuals could be categorized as having a low, intermediate or high risk of developing CD and could then benefit from early intervention to prevent CD or receive different treatments specific to their genetic background.

## References


1. *Tack, G. J., Verbeek, W. H. M., Schreurs, M. W. J. & Mulder, C. J. J. The spectrum of celiac disease: epidemiology, clinical aspects and treatment. Nat. Rev. Gastroenterol. Hepatol. 7, 204-13 (2010).*
2. *Biagi, F. & Corazza, G. R. Mortality in celiac disease. Nat. Rev. Gastroenterol. & Hepatol 7, 158-62 (2010).*
3. *Ludvigsson, J. F., Montgomery, S. M., Ekbom, A., Brandt, L. & Granath, F. Small-intestinal histopathology and mortality risk in celiac disease. JAMA 302, 1171-8 (2009).*
4. *Margaritte-Jeannin, P. et al. HLA-DQ relative risks for coeliac disease in European populations: a study of the European genetics cluster on coeliac disease. Tissue Antigens 63, 562-7 (2004).*
5. *Ribes-Koninckx, C. et al. Coeliac disease diagnosis: ESPGHAN 1990 Criteria or need for a change? Results of a questionnaire. J. Pediatr. Gastroenterol. Nutr. Epub June 28 (2011).*
6. *Monsuur, A. J. et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PloS one 3, e2270-e2270 (2008).*
7. *van Heel, D. A. et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827-9 (2007).*
8. *Hunt, K. A. et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395-402 (2008).*
9. *Trynka, G. et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 58, 1078-83 (2009).*
10. *Dubois, P. C. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295-302 (2010).*
11. *Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. in press (2011).*

# Molecular diagnosis of celiac disease: are we there yet?

Jihane Romanos [1,*], Anna Rybak [2,*], Cisca Wijmenga [1], and Martin C. Wapenaar [1]

[1] Department of Genetics, University Medical Center Groningen, University of Groningen, the Netherlands.
[2] Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland.

* These authors contributed equally to this work.

*Expert Opin. Med. Diagn. 2008; 2(4):399-416.*

# CHAPTER 1

CHAPTER 1

# Abstract

**Background:** Celiac disease (CD) is a complex genetic disorder of the small intestine resulting from aberrant cellular responses to gluten peptides. It may affect as much as 1% of the Western population and the only treatment is a lifelong gluten-free diet. Allelic variants of the HLA-DQ locus, coding for the HLA-DQ2 and HLA-DQ8 molecules, contribute to ~ 40% of CD etiology, whereas other genes, such as *MYO9B, CTLA4, IL2, IL21, PARD3* and *MAGI2*, have only a modest effect. Most of these genes have shown varied association among different populations and an overlap with other autoimmune or inflammatory disorders, indicating that such disorders may share common pathways. **Objectives:** In this review, a molecular approach into diagnostics of celiac disease is shown. **Conclusions:** Genome-wide association studies will allow more genes to be identified, and knowing how risk variants combine will help to predict better the risk for the individual. HLA typing can already be used to identify high-risk individuals.

**Keywords:** celiac disease, *CTLA4*, HLA-DQ2, HLA-DQ8, *IL2/IL21, MAGI2, MYO9B, PARD3*

# 1. Introduction

## 1.1 History of celiac disease

In 1888, Samuel Gee first described celiac disease (CD) as a disorder with onset usually between 1 and 5 years of age, with diarrhea, abdominal distension and failure to thrive as the most important symptoms [1]. The cause of the disease was unknown, but it was noticed that patients recovered when they were put on a restricted diet. Different diets were used, including a banana diet, until, in 1941, the Dutch pediatrician Willem Karel Dicke discovered that children with CD benefited when treated with a wheat-free diet [2]. In his thesis, he described the clinical improvement of five children when wheat, rye and oats were omitted from their diet and their relapse when these items were included in their diets again. He concluded that components from these flours caused CD. Paulley was the first to demonstrate changes in biopsies from the small intestine, although the classification that is now used was introduced and described in detail by Marsh in 1992 (Figure 1) [3]. In the mean time, CD was found to be associated with HLA-DQ2 [4], which was shown to present gluten to T cells [5,6]. The next milestone was the discovery of tissue transglutaminase as the autoantigen for the endomysial antibodies [7].

It is now well known that the storage protein gluten present in wheat, and the homologs secalin in rye and hordein in barely, are the triggering cause of CD. The introduction of serological tests for CD made population-based studies possible and these have shown that CD is the most common food intolerance, affecting around 1% of the general population (Table 1).

## 1.2 Clinical features and treatment of celiac disease

There is a wide spectrum of gastrointestinal and other symptoms that may appear in the course of CD. Classic CD patients present with steatorrhea, abdominal distension, edema, malabsorption and failure to thrive [8]. Other gastrointestinal symptoms, such as persistent diarrhea, persistent abdominal pain, constipation and vomiting, are also common. In the last two decades, the availability of simple serological tests has allowed us to recognize and diagnose also atypical celiac disease with non-gastrointestinal symptoms being presented in the first instance, such as iron-deficient anemia resistant to oral iron supplementation, osteopoenia or osteoporosis, dental enamel hypoplasia, delayed puberty, short stature, persistent modest elevation of serum aminotransferase levels (Table 2) [9], as well as a delay in the age of disease onset [10,11]. It seems that diagnosed CD patients represent only the tip of the iceberg among many more undiagnosed patients with atypical or asymptomatic CD [12,13]. A classification has therefore been established based on the clinical symptoms and pathomorphological features of CD. Classic or symptomatic CD refers to a presentation with the typical triad of symptoms: malabsorption, failure to thrive and persistent diarrhea. In asymptomatic or

'silent' CD, gastrointestinal symptoms are absent or the disease overlaps with other autoimmune or genetic diseases (Table 2) and is usually detected accidentally or by mass screening of high-risk groups with positive antiendomysium antibodies and/or antihuman tissue transglutaminase antibodies, or with typical lesions of the small intestine. A potential or latent form of CD is diagnosed in patients with positive autoantibodies and a typical HLA-predisposing genotype (-DQ2 or -DQ8; discussed later) but a normal or minimally abnormal mucosal architecture with an increased number of intraepithelial γδ lymphocytes. However, latent CD can also be applied to remission patients who presented with flat duodenal mucosa in the past and recovered on a gluten-free diet [14].

Patients who have been on a gluten-free diet for over 12 months and present with persisting villous atrophy with crypt hyperplasia and increased intraepithelial lymphocytes (IELs) in the small intestinal mucosa are referred to as having refractory CD (RCD) [15]. At present, two categories of RCD can be recognized: type I without aberrant T cells, which responds well to immunosuppressive therapy; and type II with a clonal outgrowth of aberrant T cells in the intestinal mucosa characterized by loss of antigen on IELs. Patients with the latter type have a high risk of developing enteropathy-associated T-cell lymphoma (EATL). RCD type II patients often suffer from severe malabsorption with weight loss, abdominal pain and diarrhea [15]. A diagnosis of RCD must include the reassessment of the initial diagnosis of CD in order to exclude other gastrointestinal diseases (Table 3), as well as an assessment of the patient's strict adherence to the gluten-free diet.

**Table 1**. Prevalence of celiac disease in different populations.

| Population | Prevalence (%) | References |
|---|---|---|
| Australia | 0.23 | [113] |
| Sweden | 0.49 | [114] |
| Spain | 0.26 | [115] |
| Italy | 0.33 | [116] |
| The Netherlands | 0.5 | [117] |
| North Ireland | 0.8 | [118] |
| Western populations | 1 | [25] |
| England | 1.15 | [119] |
| New Zealand | 1.2 | [120] |
| Finland | 1.99 | [121] |
| Romania | 2.22 | [122] |
| Northern Americans | 0.75 | [26] |
| Sahara | 5.6 | [123] |

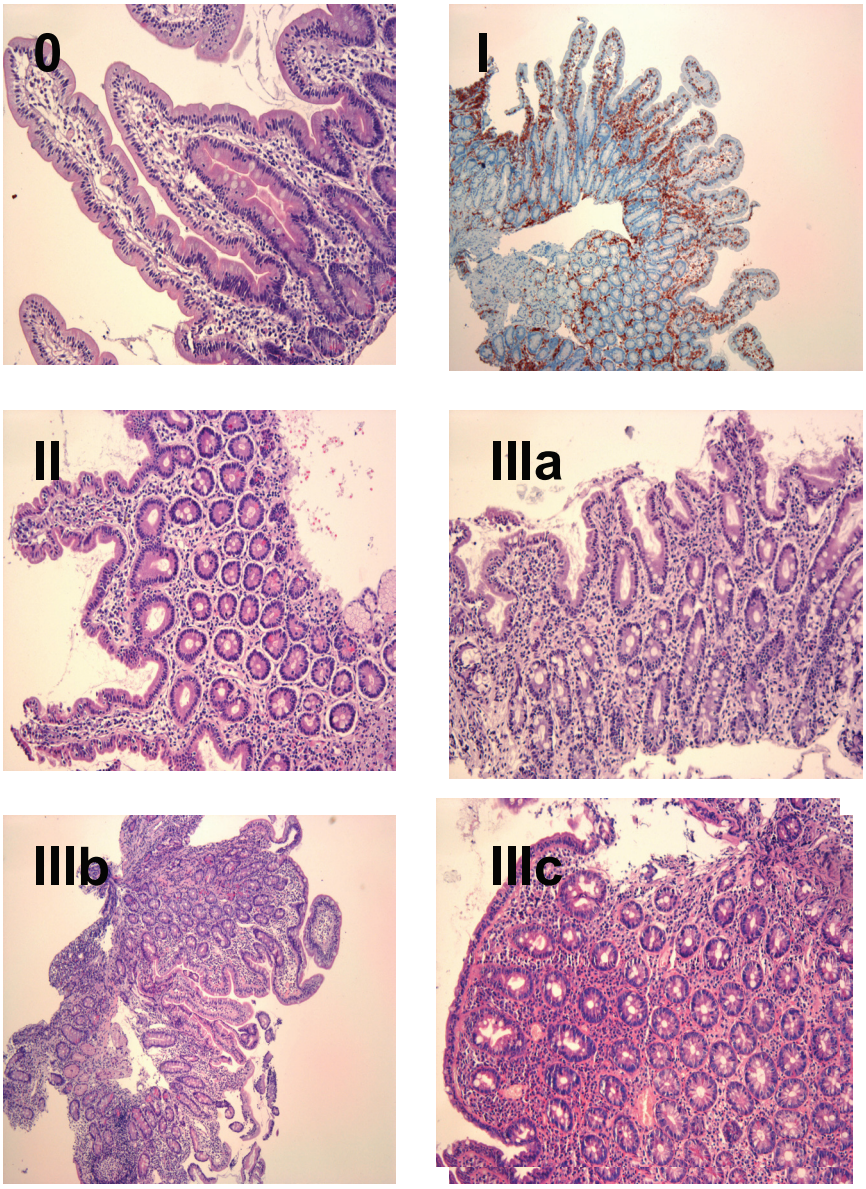



**Figure 1. Marsh classification. 0.** Pre-infiltrative mucosa. Normal stage of the intestinal mucosa. **I.** Increased number of the non-mitotic intraepithelial lymphocytes (IELs), more than 30 per 100 enterocytes. Normal mucosal architecture. **II.** Increased IEL infiltration, crypt hyperplasia (increase in crypt depth without a reduction in villous height). Villous:crypt (v:c) ratio = 2:1. **III.** Classical CD lesion with villous atrophy. **IIIa**. Partial atrophy, v:c ratio = 1:1. **IIIb**. Subtotal atrophy, v:c ≤ 1:2. **IIIc**. Total villous atrophy.

**Table 2.** The different presentations of celiac disease.

| | Symptoms | References |
|---|---|---|
| Classic celiac disease | Malabsorption | [8] |
| | Abdominal distension | |
| | Failure to thrive | |
| | Steatorrhea | |
| | Weight loss | |
| | Edema | |
| | Fatigue, lethargy | |
| | Irritability | |
| Other symptoms | Persistent diarrhea | [124] |
| | Vomiting | |
| | Constipation | |
| | Persistent abdominal pain | |
| | Dyspepsia | |
| | Poor appetite | |
| Atypical CD symptoms | Dermatitis herpetiformis | [9,11,12,14,125] |
| | Growth retardation | |
| | Iron-deficiency anemia | |
| | Dental enamel hypoplasia | |
| | Delayed puberty | |
| | Osteopenia, osteoporosis | |
| | Isolated hypertransaminasaemia | |
| | Infertility | |
| | Bone pains and fractures | |
| | Arthritis | |
| | Recurrent oral aphthae | |
| | Irritable bowel disease | |
| | Secondary hypoparathyroidism | |
| | Erythema nodosum | |
| | Esophageal reflux | |

**Table 2 (continued).** The different presentations of celiac disease.

| | Symptoms | References |
|---|---|---|
| Associated diseases | Autoimmune diseases | [9,11,12,14, 124-126] |
| | - Diabetes mellitus type I | |
| | - Autoimmune thyroiditis | |
| | - Autoimmune hepatitis | |
| | - Sjögren syndrome | |
| | - Idiopathic dilative cardiomyopathy | |
| | - Chronic autoimmune urticaria | |
| | - Alopecia | |
| | Neuropsychiatric conditions: | |
| | - Gluten ataxia | |
| | - Depression | |
| | - Anxiety | |
| | - Peripheral neuropathy | |
| | - Epilepsy (with or without cerebral calcifications) | |
| | - Migraine | |
| | Myocarditis | |
| | Macroamylasemia | |
| | Hyposplenism | |
| | Primary biliary cirrhosis | |
| | Idiopathic pulmonary hemosiderosis | |
| Associated genetic disorders | Down's, Turner's, William's syndromes | [14] |
| | Selective IgA deficiency | |

The current treatment for CD is a strict gluten-free diet for life, in which wheat, rye and barley must be avoided. There is a constant debate concerning including oats in the diet of celiac patients. Oats contain less prolamine in comparison with wheat, rye or barley: prolamine of wheat constitutes 40% of the cereal, whereas of oats only 15% [16,17]. It seems that oats can be symptomatically tolerated by most patients, but as the number of silent or atypical forms of CD increases, those patients wishing to consume a diet containing oats should have a regular follow-up. So far, the long-term effects of oats in the diet for celiac patients remain unknown [17]. Depending on the severity of the disease and the coexistence of any atypical symptoms, some patients may require special treatment, for example, vitamins or iron supplementation, repletion of fluids and electrolytes,

paraenteral nutrition, or occasionally steroids. Patients suffering from RCD type II require immunosuppressive therapy or even chemotherapy with autologous stem cell transplantation [15].

In most cases, following a gluten-free diet leads to subsidence of the symptoms, a decrease of the autoantibodies' titer, and a subsequent reversion of the intestinal lesions. Interestingly, children on a gluten-free diet report a quality of life comparable with that of a reference population, whereas patients who acquire the disease as an adult find it difficult to comply with the diet [18].

## 1.3 Diagnosis

The first criteria for the diagnosis of CD were established in 1970 by the European Society for Pediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN). The criteria were based on the characteristic histological changes in multiple small bowel biopsies, recovery on a gluten-free diet, and relapse after reintroduction of gluten into the diet [19]. In 1990 these criteria were revised and a gluten-challenge was prescribed only for patients younger than 2 years, owing to the many other diseases and conditions that may lead to villous atrophy in that period (Table 3) [20]. Nowadays, it is recommended to obtain multiple biopsies from the distal parts of the duodenum, with evaluation to include assessment of the pathomorphological changes of the mucosa, as described by Marsh (see Box 1) [3,9].

Serological tests used to identify CD are common, sensitive and have a high specificity. The most useful are IgA-class antiendomysial (EmA) and human tissue transglutaminase (tTG) antibodies. Both are targeted at tissue transglutaminase autoantigen and present a similar high sensitivity (86 – 100% for EmA, 77 – 100% for tTG) and specificity (90 – 100% for EmA, 91 – 100% for tTG) [21]. However, there are some pitfalls in the serological screening test for CD seen in patients with selective IgA deficiency (~ 2.6% of CD patients), patients with only IgG-class autoantibodies and normal serum IgA, and finally those with positive EmA antibodies, while typical tTG antibodies are lacking [22]. Fraser et al. [23] recommended that a combination of the antibodies should be used to confirm the diagnosis of CD, with anti-tTG and EmA antibodies, where at least one test should be based on IgG-class antibodies. New, quick commercial antibody tests have been developed and some of them can even detect total IgA serum level; however, they should not be used for the final diagnosis of CD, which should be based on the accepted standards.

There is a strong association between CD and HLA genotype. Over 95% of CD patients are HLA-DQ2-positive (discussed later) and the rest are DQ8-positive, but there are still no prospective studies evaluating the benefits of testing HLA genotype in clinical practice.

As medicine in the twenty-first century should be based on prevention and prophylaxis, mass screening for CD is being considered and this issue is now being discussed worldwide [24]. It is also recommended to test the first-degree relatives of celiac patients, as well as patients with other autoimmune diseases because they

show a higher prevalence of CD than in the general population [24]. Serology tests and HLA typing are both used for screening and provisional diagnosis, but ultimately only endoscopic biopsy sampling can confirm or reject the diagnosis.

## 1.4 Epidemiology

Owing to the numerous cases of undiagnosed patients suffering from CD, the prevalence of this disease has long been underestimated. Ratios of undiagnosed-to-diagnosed cases of CD were reported between ~ 5:1 and 13:1 [18]. Many studies have shown that CD affects 0.3 – 1.0% of the general population (Table 1) [18,24-26], with a female-to-male ratio around 2 – 3:1 [18]. Across epidemiological studies such as the multicenter study done by ESPGHAN in 1992, the variable incidence and different patients' age at the time of the first diagnosis were observed. Mass screening studies as well as twin and sib-pair studies have revealed that there is ~ 10% prevalence among first-degree relatives and up to 30% genetic predisposition risk among first-degree relatives with confirmed HLA-DQ2 or HLA-DQ8 genotype [18,27], whereas the risk for monozygotic twins increases to 75% [28]. This is a much higher rate than in any other condition with a multifactorial basis and it gives further

**Table 3.** Diseases and conditions with mucosal changes similar to those in celiac disease

| Diseases and conditions |
|---|
| Tropical sprue |
| HIV enteropathy |
| Combined immunodeficiency states |
| Radiation damage |
| Recent chemotherapy |
| Graft-vs-host disease |
| Chronic ischemia |
| Giardiasis |
| Whipple's disease |
| Crohn's disease |
| Eosinophilic gastroenteritis |
| Zollinger-Ellison syndrome |
| Autoimmune enteropathy |
| Cow's milk protein allergy |
| Protein energy malnutrition |

**Box 1 |** Marsh classification of stages in celiac disease

| | |
|---|---|
| Type 0 | Pre-infiltrative mucosa. Normal stage of the intestinal mucosa. |
| Type I | Increased number of the non-mitotic intraepithelial lymphocytes (IELs), more than 30 per 100 enterocytes. Normal mucosal architecture. |
| Type II | Increased IEL infiltration, crypt hyperplasia (increase in crypt depth without a reduction in villous height). |
| Type III | Classical CD lesion with villous atrophy: IIIa – partial, IIIb – subtotal, IIIc – total villous atrophy. |
| Type IV | Total villous atrophy and crypt hypoplasia. This is an irreversible lesion that is present in refractory CD patients. |

support for a genetic background underlying the disease. Greco et al. [28] have shown that the dizygotic twins-to-siblings ratio for CD appears to be close to one (11.1 versus 11%), which suggests that a shared environment (gluten antigen aside) has little or no effect on the concordance of dizygotic twins reared together.

# 2. Pathogenesis

The pathogenesis of CD is complex. It is the interaction between three major factors that drive the pathological processes that take place in the affected intestine: genetic make-up, innate and adaptive immunity, and gluten as the principal environmental trigger.

Gluten is present in wheat, barley and rye and consists of glutenins and gliadins. The latter are proteins rich in glutamine and proline residues and play an important role as glutamine-donor substrate for tissue transglutaminase (tTG). Deamidation of the gliadins with tTG results in negative charging of the gluten-peptides. Studies on HLA-DQ2 and HLA-DQ8 molecules, which are necessary for developing CD, revealed that they can bind negatively charged amino acids with a very high affinity. These complexes activate B cells to produce anti-tTG antibodies, as well as stimulate CD4+ T-cell clones' response, which results in production of inflammatory cytokines in which IFN-γ is dominant. Activation of the T cells starts inflammatory reactions that lead to mucosal damage in the small intestine [29,30]. Aside from this mechanism, intraepithelial T lymphocytes (CD8+), those seen to be under the control of the CD4+ T cells, express NK-cell receptors and kill the epithelial cells [31]. Furthermore, several studies have reported that gluten can directly induce a stress response in epithelial cells, owing to activation of the innate immune cells such as macrophages or dendritic cells (Figure 2) [31].

There is still discussion about whether the amount of dietary gluten ingested plays an important role in the initiation event of CD. Ivarsson has shown that children < 2 years of age presented with a significantly

higher incidence of CD when the amount of daily gluten ingestion increased twofold. The incidence fell rapidly when the consumption of gluten by young children was reduced [32].

Recently, an infectious etiology was proposed resulting from the molecular mimicry between gluten proteins and proteins produced during adenovirus or rotavirus infections [33-35]. Furthermore, Ivarsson et al. have shown that children who experienced 3 or more infectious episodes before 6 months of age had an increased risk of developing CD before 2 years of age [36]. However, further studies are needed to prove this association.

# 3. Genetics

In the past decade, tremendous progress has been achieved in unraveling the genetic etiology of CD. Twin- and family-based studies clearly show a strong genetic component to CD development, with inherited risk attributable to HLA and non-HLA factors [37]. Most of the CD-susceptible genes have been identified by genome linkage analysis, candidate gene studies and/or genome-wide association studies (GWA) (see Box 2). So far, four loci, termed CELIAC1 – CELIAC4, and several genes have been reported to be associated with CD in different populations, but these findings could not always be confirmed (Table 4). These identified genes outside HLA explain only a very small percentage of familial disease risk. Based on their function, the genes

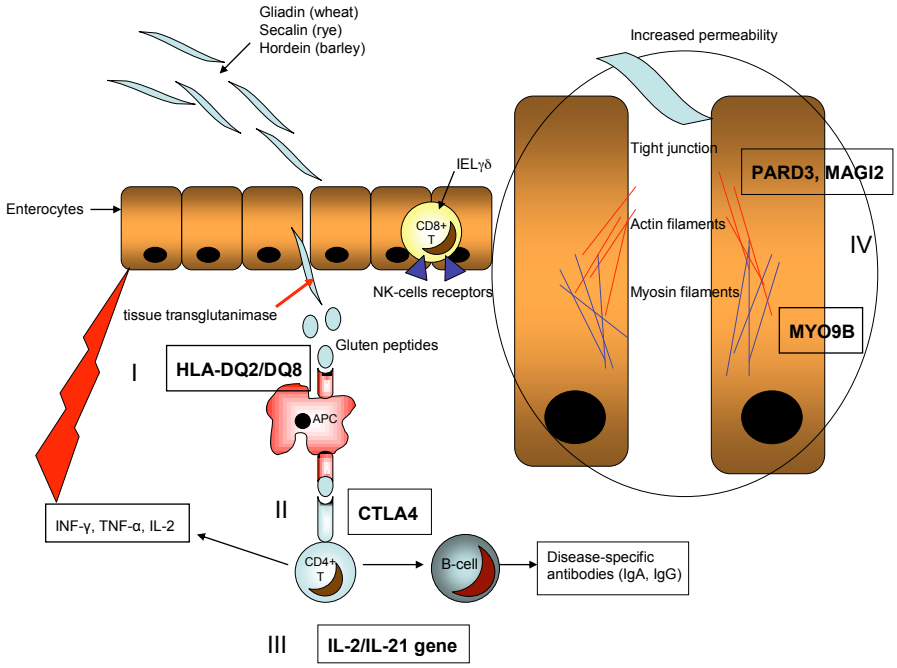



**Figure 2.** Pathogenesis of celiac disease.

can be divided into two categories: inflammation-related genes and potential mucosal barrier-related genes. In this section, the two categories of predisposing genes and their associations in different populations are discussed. Identifying CD susceptibility genes may improve the diagnostic and prognostic markers, provide a better understanding of the disease etiology, permit development of new therapeutics and clarify the clinical overlap of CD with other autoimmune disorders.

## 3.1 Inflammation-related genes

*3.1.1 The HLA complex (CELIAC1 locus)*

The HLA complex is the most gene-dense region of the mammalian genome. It is located on chromosome 6p21.3 and plays an important role in multiple autoimmune disorders, such as CD, type 1 diabetes, rheumatoid arthritis, multiple sclerosis, psoriasis and others. CD was first found to be associated with the HLA complex in 1972 using serological methods [38]. Later studies found that the strong CD association was actually with HLA-

**Box 2** | Types of studies performed to detect genetic susceptibility to a disease

Candidate gene approach: is the genetic study of genes suspected of influencing the development of a disease or trait of interest. It is based on generating a priori hypotheses about their etiological role in disease. The genes may be candidates because of their protein function, their expression in specific tissues, their location in a genomic region which is linked to the disease of interest, and/or their location in a chromosomal region influencing the trait of interest in animal models.

Genome-wide linkage analysis: is a genome-wide screen based on markers that co-segregate with the disease, thereby indicating the regions in the genome that contain disease susceptibility genes. Families with multiple patients or small families with affected sib-pairs are usually collected for this type of study. In a genome-wide linkage analysis, the whole genome can be screened simultaneously testing for linkage. In complex diseases, the analysis for linkage is based on counting the number of parental alleles at a locus shared identical-by-descent between the siblings. Deviation from Mendelian expected frequencies of allele-sharing indicates linkage.

Genome-wide association studies (GWA): are genome-wide scans based on >300,000 SNPs selected as representative of a region in which they are in linkage disequilibrium with other variants. These studies aim to pinpoint the genetic differences that correlate with and perhaps play a causative role in a particular disease, by comparing DNA samples from a group of patients who share the disease (or other biological trait) to those who do not. Large numbers of patients and controls are needed for this type of analysis in order to detect genes with low effects.

DQ2 and in a minority of cases with HLA-DQ8 [4,5,39]. The HLA-DQ molecules function as cell surface receptors for exogenous peptide antigens (gluten in the case of CD) on APC (antigen-presenting cells), presenting them to T helper cells (Figure 2). Studies have shown that T cells from the intestinal mucosa of CD patients preferentially recognize gliadin when presented by DQ2 or DQ8 [40,41]. The latter molecules are made up of α and β subunits to form a heterodimer and each subunit is encoded by HLA-DQA1 or HLA-DQB1 genes, respectively. These genes encode numerous polymorphic molecules, which differ in amino acids affecting the peptide binding and presentation repertoire. DQ2 (denoted serologically) is a group of DQ2.5 and DQ2.2 heterodimers encoded by DQA1*0501- DQB1*0201 and DQA1*0201-DQB1*0202, respectively. The HLA-DQB1*0201 allele and HLA-DQB1*0202 allele differ only by one or a few base pairs and are thought to have the same functional properties. This also holds for the HLA-DQA1*0505 allele of DQ7 and HLA-DQA1*0501 alleles of DQ2.5. It was found that 90 – 95% of celiac patients had the DQ2 molecule compared with 20 – 30% of healthy controls[5,39]. In this group, the risk of CD was shown to be increased in individuals homozygous for DQ2.5 [42]. A recent study showed that the HLA-DQ2.5 homozygote frequency is more than doubled (from 20.7 to 44.1%) in RCD type II patients [43], and increases further to 53.3% in patients with EATL. This HLA-DQ2.5 heterodimer can be encoded in cis (i.e., on the same chromosome) or in trans (i.e., on different chromosomes by DQ7/DQ2.2). In southern Europe, individuals carrying DQ2.5 in trans are more prevalent than in the north, where individuals carry DQ2.5 in cis more frequently [44]. The other major risk allele for CD is DQ8 (HLA-DQA1*03, HLA-DQB1*0302), which accounts for 6 – 10% of celiac patients, mainly in southern Europe [45].

The HLA-DQ has been implicated as a risk factor for CD in several populations, but it can only explain at most 40% of the familial aggregation observed [46]. Thus, various studies have suggested that immune-related genes, such as *TNF*, *MICA* and *MICB*, in the HLA region are also involved in CD susceptibility [47-49]. Unfortunately, no significant evidence for independent HLA risk factors was found because of the strong linkage disequilibrium across the HLA region [48].

Since HLA-DQ2 and HLA-DQ8 alleles are also found in the general population, carrying these two alleles is necessary but not sufficient for CD development, indicating the presence of non-HLA-susceptible gene(s). In the next sections, the non-HLA genes that have been found to contribute to CD are described.

*3.1.2 CD28-CTLA4-ICOS variants (CELIAC3 locus)*

Much less is known about the involvement of non-HLA genes in celiac disease. Several studies in different populations have reported and replicated linkage to the chromosome region 2q33 [50-58]. This region contains a cluster of T lymphocyte immune-regulating genes: *CD28*, cytotoxic T lymphocyte-associated antigen 4 (*CTLA4*) and inducible T-cell co-stimulator (*ICOS*). Being co-stimulatory molecules, CD28 and CTLA4 bind to receptors (B7 family) on the surface of antigen-presenting cells. Together with the antigen-specific T-cell receptor, they

**Table 4.** Identified susceptibility genes and their associations with diseases other than CD

| Gene | Locus | Chromosome | Function | Associated with disease other than CD | Way of identification |
|---|---|---|---|---|---|
| HLA-DQ | CELIAC 1 | 6p21.3 | Presenting peptides derived from extracellular proteins | Graves' disease, multiple sclerosis, systemic lupus erythematosus, type 1 diabetes [127], autoimmune thyroid disease [127], Abacavir hypersensitivity | O [4] |
| CTLA4 | CELIAC 3 | 2q33 | Regulating T-cell response, activation and proliferation | Type I diabetes [90], Grave's disease [128], rheumatoid arthritis [129], systemic lupus erythematosus [130], autoimmune thyroid disease and multiple sclerosis [131] | CG [53] |
| IL2/IL21 | | 4q26-q27 | Regulating the proliferation of T, B and NK lymphocytes | type I diabetes [60], rheumatoid arthritis [60] | GWA [59] |
| MYO9B | CELIAC 4 | 19p13.1 | Remodeling of the actin of cytoskeleton | Systemic lupus erythematosus [72], rheumatoid arthritis [72], inflammatory bowel disease [73-75], schizophrenia [76] | LS with GWA [69] |
| | CELIAC 2 | 5q31-33 | | Asthma [132], Crohn's disease [133] | LS [91] |
| | | 6q21-22 | | Type I diabetes, rheumatoid arthritis, multiple sclerosis [100] | LS [100] |

GC: candidate gene; GWA: genome-wide association; LS: linkage study; O: other

are necessary for T-cell activation (Figure 2). The CTLA4 molecule provides a negative signal to reduce T-cell activation, whereas CD28 functions as a positive regulator of T cells. The ICOS molecule, expressed on activated T cells, provides a positive proliferation and cytokine secretion signal. All three genes were selected as candidate genes for CD because of their involvement in different aspects of the T-cell response and their implication in a variety of chronic inflammatory and autoimmune diseases (Table 4). Using markers within or near *CTLA4*, several studies from different populations showed association with CD, whereas other studies failed to replicate the association (Table 5).

*3.1.3 IL2/IL21 region*

Recently, a genome-wide association study in the UK, with replication in Dutch and Irish populations, identified a new region on chromosome 4q27 as associated to CD [59]. The same region was also found to be associated to type 1 diabetes [60-62] and to rheumatoid arthritis[60]. The strongest association to CD was found with SNP marker rs6822844, 24 kb 5' of *IL21* gene (Table 5), suggesting a predisposition of genetic variation of this region to CD. This region harbors three known protein-coding genes (*TENR*, *IL2* and *IL21*) and a predicted gene of unknown function (*KIAA1109*). Based on the expression profile, *TENR* is unlikely to be involved in CD as it is exclusively expressed in testis. The function of the *KIAA1109* gene is unknown but it is widely expressed as multiple splice variants in multiple tissues. IL2 is secreted by antigen-stimulated T cells and is a key cytokine for T-cell activation and proliferation. It was found to be overexpressed in IEL clones isolated from CD patients compared with controls [63]. IL21 is also a T-cell-derived cytokine and it enhances B, T and NK cell proliferation. Both cytokines (IL2 and IL21) are implicated in the mechanisms of other intestinal inflammatory diseases, which makes them good candidate genes for CD.

## 3.2 Potential mucosal barrier-related genes

Impairment of the intestinal barrier may play an important role in the pathogenesis of CD and this would explain the unwanted passage of gluten peptides. Bjarnason et al. showed that untreated celiac patients and patients suffering from dermatitis herpetiformis (DH), as well as those on a gluten-free diet, have significantly increased intestinal permeability (measured by urinary excretion of 51Cr-EDTA) compared with that of controls. He therefore suggested there was a primary (inherited) defect in the intestinal mucosa of patients with CD and DH [64]. It has also been shown that impaired intestinal permeability is present long before the onset of the disease in animal models for gut inflammation as well as in patient cohorts [64,65]. Furthermore, studies on the ultrastructural morphology of the small intestinal mucosa revealed significant differences in the structure of the tight junctions among patients with active disease, patients on a gluten-free diet and normal controls [27,66]. Drago et al. also noticed zonulin (an intestinal protein involved in tight junction regulation) release

**Table 5.** Level of involvement of genes in different populations

| Gene | Population | Ass or not ass. | Most associated SNP | Allele, genotype or haplotype associated | p-value | Odds ratio | Ref. |
|---|---|---|---|---|---|---|---|
| CTLA4/CD28 | Basque (Spain) | Not ass. | CTLA4 +49 A/G | - | - | - | [134] |
| | British | Ass. | CTLA4 Haplotype | ACCGTG | 0.00067 | 1.41 (95% CI = 1.16-1.73) | [50] |
| | Dutch | Ass. | CT_60 A/G | G | 0.048 | 1.31 (95% CI = 0.99-1.73) | [48] |
| | Finnish (pooled data) | Ass. | CTLA4 +49 A/G | A | 0.002 | 1.29 (95% CI = 1.09-1.52) | [52] |
| | Finnish | Ass. | D2S116 | *136 | 0.018 | na | [135] |
| | French | Ass. | CTLA4 +49 A/G | AA | 0.002 | 2.36 (95% CI = 1.37-4.06) | [53] |
| | Irish | Ass. | CTLA4 - 658 C/T | T | 0.0263 | 1.62 (95% CI = 1.03-2.54) | [55] |
| | Italian | Ass. | CTLA4 +49 A/G | A | 0.03 | na | [54] |
| | Italian | Not ass. | CTLA4 +49 A/G | - | - | - | [86] |
| | Swedish | Ass. | CTLA4 +49 A/G | A | 0.02 | na | [56] |
| | Swedish/Norwegian | Ass. | CTLA4 +49 A/G | A | 0.007 | na | [57] |
| | Tunisian | Not ass. | CTLA4 +49 A/G | - | - | - | [86] |
| IL2/ IL21 | British | Ass. | rs6822844 | T | $4.6 \times 10^{-6}$ | na | [59] |
| | Dutch | Ass. | rs6822844 | T | $2.1 \times 10^{-5}$ | na | [59] |
| | Irish | Ass. | rs6822844 | T | 0.0013 | na | [59] |
| MYO9B | British | Not ass. | rs2305764 | - | - | - | [136] |
| | Dutch | Ass. | rs2305764 | A | $2.1 \times 10^{-6}$ | 2.27 (95% CI = 1.56-3.30) | [69] |
| | Finnish/Hungarian | Ass. | linkage | - | 0.00002 | - | [70] |
| | Italian (south) | Not ass. | rs2305764 | - | - | - | [137] |
| | Spanish (north: Madrid) | Not ass. | rs2305764 | - | - | - | [85] |
| | Spanish (south: Granada & Malaga) | Ass. | rs2305764 | AA | 0.01 | 2.3 (95% CI = 1.3-4.2) | [72] |
| | Swedish/Norwegian | Not ass. | - | - | - | - | [138] |

Ass. associated region; Not ass. not associated region; na data not available; CI confidence interval

and increased intestinal permeability when exposed to luminal gliadin [67].

The fact that genes potentially associated to CD may be involved in the proper functioning of the tight junctions and in the intestinal barrier suggests that intestinal barrier impairment may play an important, if not causative, role in developing CD.

*3.2.1 MYO9B gene (CELIAC4 locus)*

The Myosin IXB gene on chromosome 19 is so far the only CD gene that has been identified based on positional information from a linkage peak [68]. Some studies have shown significant association to the *MYO9B* gene in Dutch, Spanish, Hungarian and Finnish cohorts; however, it could not always be replicated in other populations (Table 5) [68-70]. Interestingly, this gene has been shown to be associated with RCD type II [71], with inflammatory bowel disease (IBD) in multiple populations [72-75] and with schizophrenia [76], suggesting that *MYO9B* is not a celiac-exclusive gene. IBD is also known to be associated with increased barrier permeability [65].

The *MYO9B* gene may play a role in actin cytoskeletal remodeling. Recent evidence suggests that tight regulation of cytoskeletal dynamics is key to the intestinal defense system, as it affects cellular shape, migration, adhesion, activation and phagocytosis of multiple cell types (such as epithelial cells, phagocytes, T cells) [77,78]. This is fundamental to the three consecutive immune defense lines: i) maintenance of a physical barrier through tight junction-mediated epithelial cell contacts; ii) mobilization of innate immunity through chemo-attraction and activation of phagocytes (particularly neutrophils); and iii) induction of adaptive immunity and tolerance through the formation of the T-cell synapse. Based on genetic evidence from tight junction genes (discussed later), it is tempting to speculate that *MYO9B* results in barrier impairment, although it cannot be excluded that it also influences neutrophil recruitment or is involved in T-cell immune synapse formation.

*3.2.2 PARD3 and MAGI2 genes*

A large-scale candidate gene study focused on tight junction pathway-related genes recently identified two genes as being associated to both CD and IBD: *PARD3* and *MAGI2*. Those genes encode tight junction adaptor proteins that act as membrane-associated scaffolds. This observation further supports a causal role for an impaired barrier function in CD [79]. In the same study, a Dutch ulcerative colitis group showed significant association to *MAGI2* and suggestive association to *PARD3*. No association was observed for Crohn's disease. Together with *MYO9B*, *MAGI2* and *PARD3* (partly) have been shown to be genetically associated with both CD and ulcerative colitis, suggesting they share a common etiology through tight junction-mediated intestinal barrier impairment.

### 3.3 Other genes

Many other candidate genes have been tested for association with CD. Apart from HLA and the genes presented above, the *PGPEPI, PREP, STAT1, IFN-G, SPINK1/2/4/5, FAS, MMP 1/3, IL12B, IRFI, DPPIV, TGM2, NOS2, KIR* and *LILE* gene clusters and *ELN* genes have been analyzed. No convincing disease association was found in most studies, possibly because of their low power to detect small effect sizes. Some genes, such as *IL10* [80], *DLG5*[81], *FcgRIIa* and *FcgRIIIa* [82], showed some association in one population, but replication in a second independent sample and other populations is still needed to confirm the association. Recently, a follow-up of a GWA study identified seven new regions [83]. Six of these loci harbor genes controlling immune response, including *CCR3, IL12A, IL18RAP, RGS1, SH2B3* and *TAGAP* genes. Although all markers had a 'genome-wide' significance, independent replication by other investigators is necessary for definitive validation.

## 4. Variable level of involvement of some genes in different populations

A priori, one would expect that within the same ethnic population, a disease would always demonstrate association to the same alleles. This, however, is not always the case. As Table 5 shows, CD is associated with *CTLA4* and *MYO9B* genes in some populations but not in others. Moreover, it was noticed that different single nucleotide polymorphisms (SNPs) or different alleles are associated in different populations. The A allele at position +49 in exon 1 of the *CTLA4* gene, for example,was first reported to be associated to CD by a French group, whereas a Dutch group found slight association with the G allele of the same SNP [51,53]. In addition, association was found with the CT60_G variant [84] and with CTLA4-658T [55]. A stronger association at the haplotype level rather than for a single variant was also suggested [50]. A large European study observed nominal linkage and association between CD and several marker alleles in the CTLA4 region, but it failed to identify the primary functional gene variant [58], whereas a study in an Irish population showed a stronger association to an extended haplotype (CTLA-4/CD28/ICOS) with CD compared with association with the haplotypes of individual genes [55]. The group suggested that a causal variant is probably in linkage disequilibrium with the extended haplotype around *CTLA4*.

Moreover, a polymorphism that is not associated with the disease in a certain population can still be linked to that disease in an apparently similar population: for example, the two Spanish studies on *MYO9B* in which the southern Spanish population showed association [72] whereas the northern population did not [85], and similarly the two Italian studies on *CTLA4* in which one found association with CD [54] whereas the other one did not [86].

These discrepancies among populations probably result from the complex interaction between

marker allele frequencies, number of founder mutations, disease heterogeneity and recombinations over time. Differences within European populations due to regional founder effects were also suggested for the *NOD2* and *DLG5* genes [87,88], and allele frequencies for *NOD2* risk alleles were also reported to vary significantly between European populations [89]. Other reasons for discrepancies in results could be that the sample size was too small to detect linkage disequilibrium, even if the factors mentioned above were minimal, or that the sample size was too small to detect a disease allele that has only a minor effect in the disease population. In cases with no or weak association, samples stratified for age of onset, disease subgroup, or HLA-DQ2 showed strengthened association between *CTLA4* and CD [54,90].

# 5. Finding the causal gene and variant

Much research effort has been invested in genome-wide linkage scans in CD, yet the true susceptibility gene or variant is proving hard to identify. Two loci on chromosomes 5 and 6 are of major interest and seem likely to contain disease predisposing genes.

The chromosome 5q31-33 region was first pointed out as a candidate region for CD in Italian cohorts (CELIAC2 locus) [91-93] and thereafter in other linkage studies [91,94,95]. Moreover, a meta- and pooled analysis of European CD data confirmed linkage to this region [96]. This same region coincides with linkage regions for IBD, giving more support to the hypothesis of common disease susceptibility. An extensive SNP association screen and studies on Crohn's disease [97] and on CD [98,99] have all failed to identify any susceptible gene(s) located in this region but have shown that a specific common haplotype confers susceptibility to each of the diseases. The main difficulty in identifying the causative gene(s) is because of the strong linkage disequilibrium across the region, resulting in multiple SNPs having equivalent genetic evidence [98].

Another interesting region is on chromosome 6q21-22, distinct from the HLA region, with suggested linkage in a genome-wide screen in a Dutch population [100]. The same group also found linkage to a region at 6q25.3 in a large CD family [101]. Furthermore, linkage of type 1 diabetes mellitus to the 6q21 region has been reported [102]. Although this 6q21-25 region has not shown any association in other genome-wide screens and no gene has been identified, these results are highly indicative of a susceptibility region for autoimmune diseases. The inability to detect the causative gene could be because the region contains several susceptibility genes that collectively contribute a moderate risk to CD susceptibility. Another explanation could be related to DNA copy number variations, which are in regions of complex genomic structure that are poorly genotyped by tagging SNPs. This same region in addition to six others was also identified in a recent GWA study [83]. Although these regions are in general much smaller than the linkage region, they still contain on average two to three genes. Thus, further studies are needed to identify the biologically important gene from each of these loci and

determine its causative variant.

To pinpoint the true causal gene is difficult, but still much less so than identifying the actual causative gene variant(s). Contrary to monogenic disorders in which one variant or mutation in a gene leads to the disease, complex genetic disorders are influenced by subtle changes in multiple genes. Thus, it is difficult to find the disease-causing variants because they have to occur frequently enough in the general population to coexist in one individual and cause CD. With the exception of HLA-DQ2 and HLA-DQ8, the causative variants of the genes listed above have not yet been identified. The associated variants are likely to be in strong linkage disequilibrium with the disease-causing variants, so we can expect multiple rare variants together to cause the association, as seen for the *CARD15* gene in IBD [97]. Thus, the challenge for the future is to identify the causal variants, for example, by sequence analysis, in animal models and/or by functional studies. The difficulty is to find which variant is pathogenic as not all variants will have a direct effect on proteins (non-synonymous, nonsense and frameshift mutations), but most are non-deleterious SNPs or will have an indirect effect on the proteins (synonymous, intronic and untranslated region [UTR] variants affecting splice sites and regulatory elements).

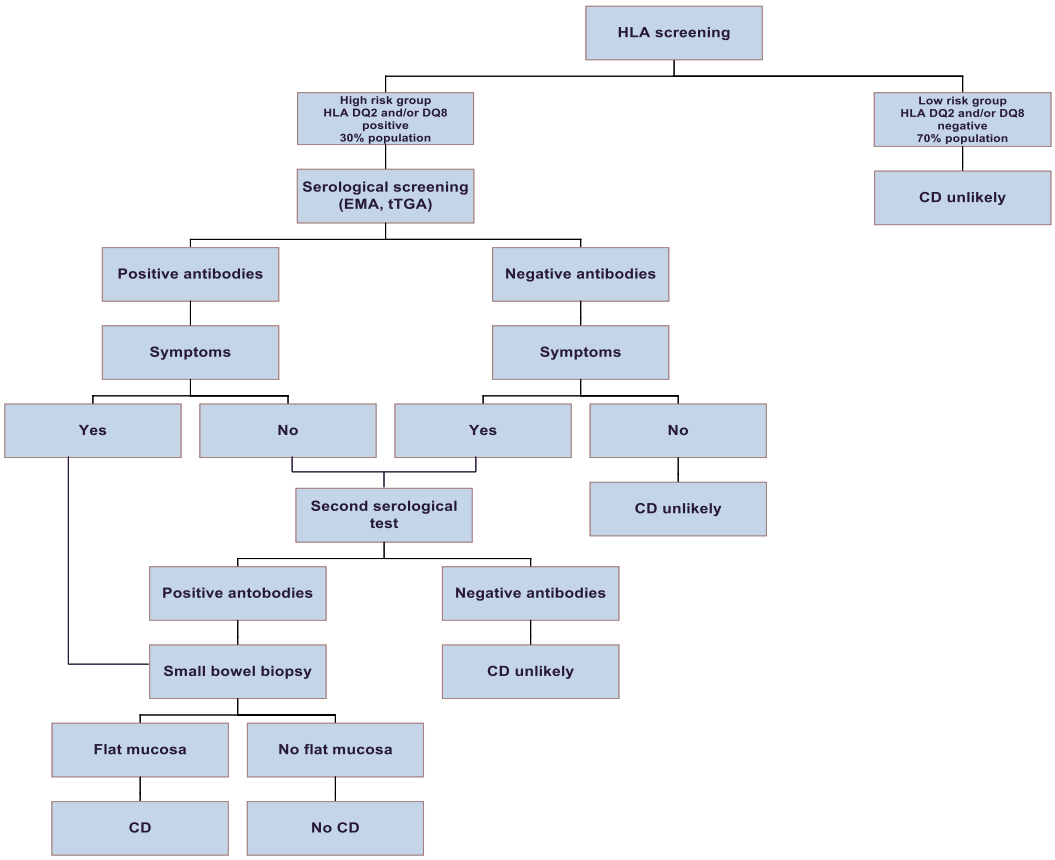



**Figure 3.** Screening strategy for celiac disease.

# 6. Future prospects for CD genetics

There is still much to discover and learn when it comes to CD. Linkage and candidate gene association studies have been successful to a limited extent, although they are time-consuming and have often failed to deliver definitive results. GWA studies are fast and provide a powerful and comprehensive approach to identify many new variants, genes and molecular pathways involved in CD. Owing to three recent advances, this type of study now has sufficient power to detect plausible effect sizes. First, the International HapMap resource [103], which documents patterns of genome-wide variation and linkage disequilibrium in four population samples, and greatly facilitates both the design and analysis of association studies. Second, the availability of dense genotype chips, containing sets of hundreds of thousands of SNPs that provide good coverage of much of the human genome, allows genotyping of hundreds or thousands of cases and controls simultaneously. Third, appropriately large and well-characterized clinical samples are now available, permitting replication studies and further stratification of the samples.

The estimated contribution from the HLA region to CD susceptibility is ~ 40% [46], whereas other genes account for a much lower percentage of the genetic risk. GWA studies will reveal several new susceptibility genes, which may offer a potential route to new therapies, improved diagnosis, and preventing disease complications, although rare genes will be missed. This can already be seen in multiple genome-wide studies. In the past few decades, many searches for CD genes have been performed and numerous putative loci have been identified [56,84,91,92,95], but no causative genes have been found. Most probably, the causative variants in these loci will have only a modest effect on the function of the underlying genes, making it difficult to pinpoint the culprit gene(s). Another explanation could be that the presence of several susceptibility genes together contributes moderately to CD etiology.

**Table 6.** Difference among traditional HLA typing and Tag SNP method.

| Traditional HLA - typing | HLA -typing using TaqSNP |
|---|---|
| - Indicate presence or abscence of DQ2 and DQ8. | - Indicate if patient is homozygous/ heterozygous for DQ2.2, DQ2.5, DQ7, DQ8. |
| - Difficult: several reactions. | - Easy: PCR reaction and end point measurments. |
| - Long: multiple step. | - Quick |
| - Expensive. | - Cheap. |
| - Sensitive to the quality of DNA. | - Insensitive to the quality of DNA. |
| - DNA: ~150ng/reaction | - DNA: 8ng/reaction |

# 7. Using genetic information for disease diagnostics

The study of the genetics of human disease helps to assess the risk that individuals may develop a certain disease. Knowledge of this risk can then be used by clinicians in prevention, diagnosis, prognosis and treatment. At present, clinicians use the patient's family history to help assess their risk of a disease. For monogenic diseases, modern molecular tools have improved the use of family histories to determine the genetic risk, whereas for complex diseases this is still too difficult owing to the many genes and mutations involved. When tested separately, each gene variant contributes mildly to the risk of developing a disease. Recently, studies have shown that combining multiple-risk alleles could improve the disease prediction for an individual [104,105].

Celiac disease is an important health problem because of its high prevalence, associated specific and non-specific morbidity, and long-term complications (Table 2). Thus, early diagnosis is very important, especially in high-risk groups such as first-degree relatives and those individuals with type 1 diabetes, iron-deficiency anemia, low bone mineral density, Down's syndrome, short stature or infertility. Possibly, individuals identified as high-risk may be helped by inducing oral tolerance during infancy through introduction of low amounts of gluten into their diet during breastfeeding [106]. However, there is still a need for long-term prospective studies to investigate this relationship further.

At present, serology testing is used as a first diagnosis tool but it needs to be repeated during an individual's lifetime because of fluctuations in serum antibody levels. HLA-DQ2 and HLA-DQ8 are strongly associated with CD (as shown above), thus researchers have tested this association as a basis for diagnosis [5]. Comparing these two tests showed that the specificity was 99% for tissue antitransglutaminase antibodies (TGA) or EMA, and 57% for HLA-DQ typing (both with 95% CIs), and the sensitivity was 81 versus 100%, respectively [107]. Thus, a possible screening strategy for high-risk groups or a general population study would involve two steps based first on selecting those at high risk of developing CD by HLA-DQ typing and, second, using repeated serological screening of those with more symptoms (Figure 3). Patients with HLA-DQ2 and/or HLA-DQ8, positive serology and other CD symptoms should be subjected to a small bowel endoscopy to confirm the diagnosis. The advantage of using HLA typing as a first screen is that 70% of the general population could then be dismissed from further testing. Current evidence suggests that non-HLA-DQ2 and/or non-HLA-DQ8 cases of CD are rare or non-existent.

Testing for HLA-risk molecules is routinely performed using methods such as PCR-single-strand conformation polymorphism (SSCP), sequence-specific oligonucleotide probing (SSOP) and/or PCR-sequence-specific primer kits (PCR-SSP) [108-111]. All these methods require several reactions, multiple steps such as amplification and hybridization to a membrane, special software or expertise in analyzing the results, and most of them are expensive. To make HLA typing more automated and relatively cheap, Monsuur et al. [112]

established a new approach using six tagging SNPs to predict whether an individual is heterozygous or homozygous for the DQ2.5, DQ2.2, DQ7 and DQ8 risk types. Genotyping the six SNPs is simple, fast and cost-effective compared with more classical techniques (Table 6). This method is characterized by a high sensitivity (> 96.8%), a high specificity (> 99.4%) and a high predictive value (> 94.0%). Although, the presence of HLA-DQ2 or HLA-DQ8 is not sufficient to diagnose the disease, it is an indication to do further serology tests and, later, a biopsy sampling. Thus, this method can be used to exclude the diagnosis of CD in the absence of HLA-DQ2 or HLA-DQ8 when screening high-risk groups or even whole populations.

# 8. Conclusion

There is still a lot to discover and learn about CD and the genes involved. We need to understand better the role of the susceptibility genes in CD by identifying the causal variants. Disease gene pathways, however, are already emerging, and point to a prominent role for inflammation. In addition, the complex interactions of these identified genes and how they affect the pathogenesis of this disorder need to be understood. It is expected that more susceptibility genes will be discovered in the years to come: some of them will be shared with other immune-related diseases, whereas others will be more exclusive to a certain group of disorders. The authors also expect to find gene variants specific to populations and to ethnic groups. Newly identified CD susceptibility genes will lead to better diagnostic tools, new targets for therapeutic intervention and an improved understanding of autoimmune diseases in general. With HLA typing, the 30% of the population carrying DQ2 and/or DQ8 who are at risk of developing CD and need to have further tests can already be identified, while the remainder (no DQ2 or DQ8) can be excluded from any further testing.

# 9. Expert opinion

CD is a severe food intolerance that occurs as a result of a complex interaction of environmental factors and multiple gene variants. It is estimated that at least 50 genes are involved in CD and that a person needs to carry a combination of several causal variants of these genes to develop the disorder. Thus, not all patients are expected to have the same combination of genetic variants and healthy people may also carry some of these variants, but not enough, or not in the right combination, to develop CD. Differences in ethnic background can lead to such genetic differences among populations. Moreover, certain variants may have been introduced into a specific population by a common ancestor. This may explain why populations may have various levels of associations to different genes (Table 5). HLA-DQ2 and HLA-DQ8 are, however, common risk factors in all populations.

On the one hand, the genes identified so far are not exclusive for CD but are also associated with other gastrointestinal or inflammatory disorders, showing an overlap in the genetic background of biologically related disorders. Examples are *CTLA4, IL2/IL21* and *MYO9B*. *CTLA4* and *IL2/IL21* are associated to CD but also involved in type 1 diabetes and rheumatoid arthritis, whereas *MYO9B* is involved in rheumatoid arthritis and ulcerative colitis. On the other hand, the prevalence of CD in patients with type 1 diabetes varies between 1 and 12%, which is higher than the prevalence in the general not-at-risk population [25]. This indicates that causative genes are present in common pathways of other immune-related disorders. It seems that they are mainly involved in one of two pathways: the immune response or the mucosal barrier.

With the advance of GWA, it is expected that many more CD susceptibility genes will be identified. Combining all the variants will improve CD risk assessment. As HLA-DQ2/8 contributes to 40% of the CD risk etiology, genetic screening should include HLA typing, thereby discharging 70% of the population (-DQ2- and -DQ8-negative people) from repeated autoantibody testing [113]. The ability to define high-risk and no-risk groups with simple HLA typing raises the important question of whether all newborns should be screened, particularly as intervention with the right amount of gluten during breastfeeding, and at the right time during infancy, may induce oral tolerance to gluten in high-risk newborns and thereby prevent the future development of CD.

**Declaration of interest**

J Romanos was supported by KP6 EU grant 036383 (PreventCD). A Rybak received a grant from Nutricia for a six-month stay at the Groningen Genetics Department.

## Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. *Vogten AJ, Pena AS. Coeliac disease: one century after Samuel Gee (1888). Neth J Med 1987;31(56):253-5*
2. *van Berge-Henegouwen GP, Mulder CJ. Pioneer in the gluten free diet: Willem Karel Dicke 1905 – 1962, over 50 years of gluten free diet. Gut 1993;34(11):1473-5*
3. *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(1):330-54*

   *•• First classification of the duodenal mucosa damage during celiac disease development.*
4. *Tosi R, Vismara D, Tanigaki N, et al. Evidence that celiac disease is primarily associated with a DC locus allelic specificity. Clin Immunol Immunopathol 1983;28(3):395-404*
5. *Sollid LM, Thorsby E. HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. Gastroenterology 1993;105(3):910-22*
6. *Koning F, Schuppan D, Cerf-Bensussan N, Sollid LM. Pathomechanisms in celiac disease. Best Pract Res Clin Gastroenterol 2005;19(3):373-87*
7. *Dieterich W, Ehnis T, Bauer M, et al. Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat Med 1997;3(7):797-801*

   *•• First identification of tissue transglutaminase as the endomysial autoantigen in celiac disease and the discussion of its possible role in the pathogenesis of the disease.*
8. *Dewar DH, Ciclitira PJ. Clinical features and diagnosis of celiac disease. Gastroenterology 2005;128 (4 Suppl 1):S19-24*
9. *Hill ID, Dirks MH, Liptak GS, et al. Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. J Pediatr Gastroenterol Nutr 2005;40(1):1-19*

   *• Practical indications for the diagnostics of celiac disease, showing value of the serological tests, HLA-typing and histology.*
10. *Ravikumara M, Tuthill DP, Jenkins HR. The changing clinical presentation of coeliac disease. Arch Dis Child 2006;91(12):969-71*
11. *Green PH, Jabri B. Coeliac disease. Lancet 2003;362(9381):383-91.*

    *• Review on the pathogenesis, diagnosis, treatment and complications of celiac disease.*
12. *Mearin ML. Celiac disease among children and adolescents. Curr Probl Pediatr Adolesc Health Care 2007;37(3):86-105*

13. *Fasano A, Catassi C. Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum. Gastroenterology 2001;120(3):636-51*

*•• Review describing the growing awareness of the diverse clinical manifestationsof celiac disease. Special emphasis is put on the diagnostic procedures and treatment protocol.*

14. *Fasano A. Clinical presentation of celiac disease in the pediatric population. Gastroenterology 2005;128 (4 Suppl 1):S68-73*
15. *Al-Toma A, Verbeek WH, Mulder CJ. Update on the management of refractory coeliac disease. J Gastrointestin Liver Dis 2007;16(1):57-63*

*•Overview of the available diagnostic and therapeutic methods for the refractory form of celiac disease.*

16. *Garsed K, Scott BB. Can oats be taken in a gluten-free diet? A systematic review. Scand J Gastroenterol 2007;42(2):171-8*
17. *Haboubi NY, Taylor S, Jones S. Coeliac disease and oats: a systematic review. Postgrad Med J 2006;82(972):672-8*
18. *Bai J, Zeballos E, Fried M, et al. WGO-OMGE practice guideline celiac disease. World Gastroenterol News 2005;10(2 Suppl):1-8*
19. *Meeuwisse GW. Diagnostic criteria in coeliac disease. Acta Pediatr Scand 1970;(59):461-3*
20. *Walker-Smith JA, Guandalini S,Schmitz J, et al. Revised criteria for diagnosis of coeliac disease. Report of working group of European Society of Paediatric Gastroenterology and Nutrition. Arch Dis Child 1990;65:909-11*
21. *Hill ID. What are the sensitivity and specificity of serologic tests for celiac disease? Do sensitivity and specificity vary in different populations? Gastroenterology 2005;128(4 Suppl 1):S25-32*
22. *Green PH, Barry M, Matsutani M. Serologic tests for celiac disease. Gastroenterology 2003;124(2):585-6*
23. *Fraser JS, Ellis HJ, Moodie S, Ciclitira PJ. Letters to the editor. Gastroenterology 2003;124(2):585-6*
24. *Mearin ML, Ivarsson A, Dickey W. Coeliac disease: is it time for mass screening? Best Pract Res Clin Gastroenterol 2005;19(3):441-52*

*• Discusses the advantages and disadvantages of population screening for celiac disease. By including HLA typing, up to two-thirds of the population could be excluded from further unnecessary (repeated) serological and histological tests.*

25. *Dube 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(4 Suppl 1):S57-67*
26. *Fasano A, Berti I, Gerarduzzi T, et al. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Arch Intern Med 2003;163(3):286-92*

*• Until recently, celiac disease was considered to be rare in the US. This study demonstrated that the prevalence of celiac disease in the US and European populations is comparable.*

27. *Book L, Zone JJ, Neuhausen SL. Prevalence of celiac disease among relatives of sib pairs with celiac disease in US families. Am J Gastroenterol 2003;98(2):377-81*

28. *Greco L, Romino R, Coto I, et al. The first large population based twin study of coeliac disease. Gut 2002;50(5):624-8*

    *• Substantial evidence for a very strong genetic component in celiac disease, which stems only partly from the HLA locus.*

29. *Dieterich W, Esslinger B, Schuppan D. Pathomechanisms in celiac disease. Int Arch Allergy Immunol 2003;132(2):98-108*

    *• Review of the pathogenesis of celiac disease with special attention given to the tTG enzyme. It also discusses possible alternative therapies based on bacterial prolyl endopeptidases, the inhibition of tTG activity with highly specific enzyme inhibitors, or HLA-DQ2/DQ8 blocking peptide analogues.*

30. *Dewar D, Pereira SP, Ciclitira PJ.The pathogenesis of coeliac disease. Int J Biochem Cell Biol 2004;36(1):17-24*

31. *Meresse B, Curran SA, Ciszewski C, et al. Reprogramming of CTLs into natural killer-like cells in celiac disease. J Exp Med 2006;203(5):1343-55*

    *•• A report on the oligoclonal expansions of intraepithelial cytotoxic T-lymphocytes (CTLs) that show genetic reprogramming of natural killer (NK) functions. This NK transformation of CTLs may underlie both the self-perpetuating, gluten- independent tissue damage and the uncontrolled CTL expansion that leads to malignant lymphomas.*

32. *Ivarsson A. The Swedish epidemic of coeliac disease explored using an epidemiological approach – some lessons to be learnt. Best Pract Res Clin Gastroenterol 2005;19(3):425-40*

33. *Kagnoff MF, Paterson YJ, Kumar PJ, et al. Evidence for the role of a human intestinal adenovirus in the pathogenesis of coeliac disease. Gut 1987;28(8):995-1001*

34. *Stene LC, Honeyman MC, Hoffenberg EJ, et al. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am J Gastroenterol 2006;101(10):2333-40*

35. *Zanoni G, Navone R, Lunardi C, et al. In celiac disease, a subset of autoantibodies against transglutaminase binds toll-like receptor 4 and induces activation of monocytes. PLoS Med 2006;3(9):e358*

    *• Identification of a rotavirus-like peptide using serum from untreated celiac disease patients. This peptide also showed homology to tissue transglutaminase, human heat shock protein 60, desmoglein 1 and Toll-like receptor 4. This suggests that rotavirus infection may be involved in celiac disease pathogenesis.*

36. *Ivarsson A, Hernell O, Nystrom L, Persson LA. Children born in the summer have increased risk for coeliac disease. J Epidemiol Community Health 2003;57(1):36-9*

37. *Louka AS, Sollid LM. HLA in coeliac disease: unravelling the complex genetics of a complex disorder. Tissue Antigens 2003;61(2):105-17*

38. *Falchuk ZM, Rogentine GN, Strober W. Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy. J Clin Invest 1972;51(6):1602-5*

39. *Sollid LM, Markussen G, Ek J, et al. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J Exp Med 1989;169(1):345-50*

40. *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(1):187-96*

    *•• Demonstrates that the preferential mucosal presentation of gluten-derived peptides in celiac disease is mediated by HLA-DQ(alpha 1*0501,beta 1*0201), explaining the strong HLA genetic association.*

41. *Lundin KE, Scott H, Fausa O, et al. T-cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8. Hum Immunol 1994;41(4):285-91*

42. *Congia M, Cucca F, Frau F, et al. A gene dosage effect of the DQA1*0501/ DQB1*0201 allelic combination influences the clinical heterogeneity of celiac disease. Hum Immunol 1994;40(2):138-42*

43. *Al-Toma A, Goerres MS, Meijer JW, et al. Human leukocyte antigen-DQ2 homozygosity and the development of refractory celiac disease and enteropathy-associated T-cell lymphoma. Clin Gastroenterol Hepatol 2006;4(3):315-9*

44. *Margaritte-Jeannin P, Babron MC, Bourgey M, et al. HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease. Tissue Antigens 2004;63(6):562-7*

    *• Assessment of a European population gradient in the HLA-DQ genotype group frequencies in celiac disease patients and parents. The relative risks associated with each DQ genotype group differ between northern and southern European countries.*

45. *Torres MI, Lopez Casado MA, Rios A. New aspects in celiac disease. World J Gastroenterol 2007;13(8):1156-61*

46. *Bevan S, Popat S, Braegger CP, et al. Contribution of the MHC region to the familial risk of coeliac disease. J Med Genet 1999;36(9):687-90*

47. *Bolognesi E, Karell K, Percopo S, et al. Additional factor in some HLA DR3/DQ2 haplotypes confers a fourfold increased genetic risk of celiac disease. Tissue Antigens 2003;61(4):308-16*

48. *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(3):215-20*

49. *Louka AS, Moodie SJ, Karell K, et al. A collaborative European search for non-DQA1*05-DQB1*02 celiac disease loci on HLA-DR3 haplotypes: analysis of transmission from homozygous parents. Hum Immunol 2003;64(3):350-8*

50. *Hunt KA, McGovern DP, Kumar PJ, et al. A common CTLA4 haplotype associated with coeliac disease. Eur J Hum Genet 2005;13(4):440-4*

51. *van Belzen MJ, Mulder CJ, Zhernakova A, et al. CTLA4 +49 A/G and CT60 polymorphisms in Dutch coeliac*

*disease patients. Eur J Hum Genet 2004;12(9):782-5*

52. *Rioux JD, Karinen H, Kocher K, et al. Genomewide search and association studies in a Finnish celiac disease population: identification of a novel locus and replication of the HLA and CTLA4 loci. Am J Med Genet A 2004;130(4):345-50*
53. *Djilali-Saiah I, Schmitz J, Harfouch-Hammoud E, et al. CTLA-4 gene polymorphism is associated with predisposition to coeliac disease. Gut 1998;43(2):187-9*

    *• First study that showed association of CTLA4 to celiac disease.*
54. *Mora B, Bonamico M, Indovina P, et al. CTLA-4 +49 A/G dimorphism in Italian patients with celiac disease. Hum Immunol 2003;64(2):297-301*
55. *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(1):19-26*
56. *Popat S, Hearle N, Wixey J, et al. Analysis of the CTLA4 gene in Swedish coeliac disease patients. Scand J Gastroenterol 2002;37(1):28-31*
57. *Naluai AT, Nilsson S, Samuelsson L, et al. The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders. Tissue Antigens 2000;56(4):350-5*
58. *Holopainen P, Naluai AT, Moodie S, et al. Candidate gene region 2q33 in European families with coeliac disease. Tissue Antigens 2004;63(3):212-22*
59. *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. Nat Genet 2007;39(7):827-9*

    *•• First genome-wide genetic association study in celiac disease revealing a new locus containing the IL2 and IL21 pro-inflammatory cytokine genes.*
60. *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. Am J Hum Genet 2007;81(6):1284-8*
61. *Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661-78*

    *•• Milestone paper describing a large, genome-wide genetic association study using a common control cohort; it identified genes involved in seven common, complex genetic, human diseases.*
62. *Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39(7):857-64*
63. *Kolkowski EC, Fernandez MA, Pujol-Borrell R, Jaraquemada D. Human intestinal alphabeta IEL clones in celiac disease show reduced IL-10 synthesis and enhanced IL-2 production. Cell Immunol 2006;244(1):1-9*

64. *Bjarnason I, Marsh MN, Price A, et al. Intestinal permeability in patients with coeliac disease and dermatitis herpetiformis. Gut 1985;26(11):1214-9*

65. *Buhner S, Buning C, Genschel J, et al. Genetic basis for increased intestinal permeability in families with Crohn's disease: role of CARD15 3020insC mutation? Gut 2006;55(3):342-7*

66. *Marsh MN, Swift JA, Williams ED. Studies of small-intestinal mucosa with the scanning electron microscope. BMJ 1968;4(5623):95-6*

67. *Drago S, El AR, Di PM, et al. Gliadin, zonulin and gut permeability: effects on celiac and non-celiac intestinal mucosa and intestinal cell lines. Scand J Gastroenterol 2006;41(4):408-19*

68. *Troncone R, Ivarsson A, Szajewska H, Mearin ML; on behalf of the members of the European multi-stakeholder platform on CD (CDEUSSA). Future research on Celiac disease. A position report from the European multi-stakeholder platform on Celiac disease (CDEUSSA). Aliment Pharmacol Ther 2008;27(11):1030-43.*

69. *Monsuur AJ, de Bakker PI,Alizadeh BZ, et al. Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nat Genet 2005;37(12):1341-4*

    *•• First celiac disease gene identified by genome-wide linkage and a subsequent association study. The structure of the gene suggests it could play a role in maintaining the intestinal barrier, the impairment of which is part of the etiology of celiac disease.*

70. *Koskinen LLE, Korponay-Szabo IR, Viiri K, et al. Myosin ixb gene region and gluten intolerance: linkage to coeliac disease and a putative dermatitis herpetiformis association. J Med Genet 2007 Dec; doi:10.1136/jmg.2007.053991*

71. *Wolters VM, Verbeek WH, Zhernakova A, et al. The MYO9B gene is a strong risk factor for developing refractory celiac disease. Clin Gastroenterol Hepatol 2007;5(12):1399-405*

    *•• First non-HLA gene associated to refractory celiac disease.*

72. *Sanchez E, Alizadeh BZ, Valdigem G, et al. MYO9B gene polymorphisms are associated with autoimmune diseases in Spanish population. Hum Immunol 2007;68(7):610-5*

73. *van Bodegraven AA, Curley CR,Hunt KA, et al. Genetic variation in myosin IXB is associated with ulcerative colitis. Gastroenterology 2006;131(6):1768-74*

    *• This paper links the etiology of ulcerative colitis with that of celiac disease through the common genetic association with the MYO9B gene.*

74. *Latiano A, Palmieri O, Valvano MR, et al. The association of MYO9B gene in Italian patients with inflammatory bowel diseases. Aliment Pharmacol Ther 2008;27(3):241-8*

75. *Nunez C, Oliver J, Mendoza JL, et al. MYO9B polymorphisms in patients with inflammatory bowel disease. Gut 2007;56(9):1321-2*

76. *Jungerius BJ, Bakker SC, Monsuur AJ, et al. Is MYO9B the missing link between schizophrenia and celiac*

*disease? Am J Med Genet B Neuropsychiatr Genet 2007;147B(3):351-5*

77. *Matter K, Balda MS. Signalling to and from tight junctions. Nat Rev Mol Cell Biol 2003;4(3):225-36*
78. *Revenu C, Athman R, Robine S, Louvard D. The co-workers of actin filaments: from cell structures to signals. Nat Rev Mol Cell Biol 2004;5(8):635-46*
79. *Wapenaar MC, Monsuur AJ, van Bodegraven AA, 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(4):463-7 • Genetic associations with PARD3 and MAGI2 further strengthen the involvement of the intestinal barrier in the etiology of celiac disease and ulcerative colitis.*
80. *Barisani D, Ceroni S, Meneveri R, et al. IL-10 polymorphisms are associated with early-onset celiac disease and severe mucosal damage in patients of Caucasian origin. Genet Med 2006;8(3):169-74*
81. *Festen EA, Zhernakova A, Wijmenga C, Weersma RK. Association of DLG5 variants with Gluten Sensitive Enteropathy. Gut 2008;57(7):1027-8.*
82. *Alizadeh BZ, Valdigem G, Coenen MJ, et al. Association analsis of functional variants of the FcgRIIa and FcgRIIIa genes with type 1 diabetes, celiac disease and rheumatoid arthritis. Hum Mol Genet 2007;16(21):2552-9*
83. *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(4):395-402*
84. *King AL, Yiannakou JY, Brett PM, et al. A genome-wide family-based linkage study of coeliac disease. Ann Hum Genet 2000;64(Pt 6):479-90*
85. *Nunez C, Marquez A, Varade J, et al. No evidence of association of the MYO9B polymorphisms with celiac disease in the Spanish population. Tissue Antigens 2006;68(6):489-92*
86. *Clot F, Fulchignoni-Lataud MC, Renoux C, et al. Linkage and association study of the CTLA-4 region in coeliac disease for Italian and Tunisian populations. Tissue Antigens 1999;54(5):527-30*
87. *Arnott ID, Nimmo ER, Drummond HE, et al. NOD2/CARD15, TLR4 and CD14 mutations in Scottish and Irish Crohn's disease patients: evidence for genetic heterogeneity within Europe? Genes Immun 2004;5(5):417-25*
88. *Tenesa A, Noble C, Satsangi J, Dunlop M. Association of DLG5 and inflammatory bowel disease across populations. Eur J Hum Genet 2006;14(3):259-60*
89. *Cavanaugh J. NOD2: ethnic and geographic differences. World J Gastroenterol 2006;12(23):3673-7*
90. *Zhernakova A, Eerligh P, Barrera P, et al. CTLA4 is differentially associated with autoimmune diseases in the Dutch population. Hum Genet 2005;118(1):58-66*
91. *Greco L, Corazza G, Babron MC, et al. Genome search in celiac disease. Am J Hum Genet 1998;62(3):669-75*
92. *Greco L, Babron MC, Corazza GR, et al. Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families. Ann Hum Genet 2001;65(Pt 1):35-41*
93. *Percopo S, Babron MC, Whalen M, et al. Saturation of the 5q31-q33 candidate region for coeliac disease. Ann*

*Hum Genet 2003;67(Pt 3):265-8*

94. *Naluai AT, Nilsson S, Gudjonsdottir AH, et al. Genome-wide linkage analysis of Scandinavian affected sib-pairs supports presence of susceptibility loci for celiac disease on chromosomes 5 and 11. Eur J Hum Genet 2001;9(12):938-44*
95. *Liu J, Juo SH, Holopainen P, et al. Genomewide linkage analysis of celiac disease in Finnish families. Am J Hum Genet 2002;70(1):51-9*
96. *Babron MC, Nilsson S, Adamovic S, et al. Meta and pooled analysis of European coeliac disease data. Eur J Hum Genet 2003;11(11):828-34*

    *• European meta and pooled data analysis showing significant linkage to a genetic risk factor for celiac disease in the 5q31-33 region. Individual linkage studies, however, were each below the level of significance for this region.*
97. *Rioux JD, Daly MJ, Silverberg MS, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 2001;29(2):223-8*
98. *Amundsen SS, Adamovic S, Hellqvist A, et al. A comprehensive screen for SNP associations on chromosome region 5q31-33 in Swedish/Norwegian celiac disease families. Eur J Hum Genet 2007;15(9):980-7*
99. *Ryan AW, Thornton JM, Brophy K, et al. Chromosome 5q candidate genes in coeliac disease: genetic variation at IL4, IL5, IL9, IL13, IL17B and NR3C1. Tissue Antigens 2005;65(2):150-5*
100. *van Belzen MJ, M jer JW, Sandkuijl LA, et al. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology 2003;125(4):1032-41*

    *• The first genome-wide linkage study in celiac disease that yielded significant linkage, pointing to a locus on 19 p13.1, which was later identified as the MYO9B gene.*
101. *van Belzen MJ, Vrolijk MM, Meijer JW, et al. A genomewide screen in a four-generation Dutch family with celiac disease: evidence for linkage to chromosomes 6 and 9. Am J Gastroenterol 2004;99(3):466-71*
102. *Pociot F, McDermott MF. Genetics of type 1 diabetes mellitus. Genes Immun 2002;3(5):235-49*
103. *International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437(7063):1299-320*

    *• Project documenting patterns of genome-wide variation and linkage disequilibrium in four population samples. This information is crucial for the efficient design of markers to be used in comprehensive genetic association studies.*
104. *Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007;17(10):1520-8*
105. *Weedon MN, McCarthy MI, Hitman G, et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 2006;3(10):e374*

    *•• A clear example taken from research on type 2 diabetes of how combining information from several known, common risk, polymorphisms allows the identification of population subgroups with markedly differing risks of*

*developing a complex disease compared with the results obtained using single polymorphisms.*

106. *Akobeng AK, Ramanan AV, Buchan I, Heller RF. Effect of breast feeding on risk of coeliac disease: a systematic review and meta-analysis of observational studies. Arch Dis Child 2006;91(1):39-43*
107. *Hadithi M, von Blomberg BM, Crusius JB, et al. Accuracy of serologic tests and HLA-DQ typing for diagnosing celiac disease. Ann Intern Med 2007;147(5):294-302*
108. *Carrington M, Miller T, White M, et al. Typing of HLA-DQA1 and DQB1 using DNA single-strand conformation polymorphism. Hum Immunol 1992;33(3):208-12*
109. *Ronningen KS, Spurkland A, Iwe T, et al. Distribution of HLA-DRB1, -DQA1 and -DQB1 alleles and DQA1-DQB1 genotypes among Norwegian patients with insulin-dependent diabetes mellitus. Tissue Antigens 1991;37(3):105-11*
110. *Olerup O, Zetterquist H. HLA-DR typing by PCR amplification with sequence- specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens 1992;39(5):225-35*
111. *Buyse I, Decorte R, Baens M, et al. Rapid DNA typing of class II HLA antigens using the polymerase chain reaction and reverse dot blot hybridization. Tissue Antigens 1993;41(1):1-14*
112. *Monsuur AJ, de Bakker PIW, Zhernakova A, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PloS ONE 2008;3(5):e2270.*
113. *Liu E, Rewers M, Eisenbarth GS. Genetic testing: who should do the testing and what is the role of genetic testing in the setting of celiac disease? Gastroenterology 2005;128(4 Suppl 1):S33-7*
114. *Hovell CJ, Collett JA, Vautier G, et al. High prevalence of coeliac disease in a population-based study from Western Australia: a case for screening? Med J Aust 2001;175(5):247-50*
115. *Lagerqvist C, Ivarsson A, Juto P, et al. Screening for adult coeliac disease – which serological marker(s) to use? J Intern Med 2001;250(3):241-8*
116. *Riestra S, Fernandez E, Rodrigo L, et al. Prevalence of Coeliac disease in the general population of northern Spain. Strategies of serologic screening. Scand J Gastroenterol 2000;35(4):398-402*
117. *Catassi C, Ratsch IM, Fabiani E, et al. Coeliac disease in the year 2000: exploring the iceberg. Lancet 1994;343(8891):200-3*
118. *George EK, Jansen TL, Mearin ML, Mulder CJ. Epidemiology of celiac disease in The Netherlands. J Pediatr Gastroenterol Nutr 1997;24(5):S7-9*
119. *Johnston SD, Watson RG, McMillan SA, et al. Prevalence of coeliac disease in Northern Ireland. Lancet 1997;350(9088):1370*
120. *West J, Logan RF, Hill PG, et al. Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. Gut 2003;52(7):960-5*

121. *Cook HB, Burt MJ, Collett JA, et al. Adult coeliac disease: prevalence and clinical significance. J Gastroenterol Hepatol 2000;15(9):1032-6*

122. *Lohi S, Mustalahti K, Kaukinen K, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther 2007;26(9):1217-25*

123. *Dobru D, Pascu O, Tanta M, et al. The prevalence of coeliac disease at endoscopy units in Romania: routine biopsies during gastroscopy are mandatory (a multicentre study). Rom J Gastroenterol 2003;12(2):97-100*

124. *Catassi C, Ratsch IM, Gandolfi L, et al. Why is coeliac disease endemic in the people of the Sahara? Lancet 1999;354(9179):647-*

125. *James MW, Scott BB. Coeliac disease: the cause of the various associated disorders? Eur J Gastroenterol Hepatol 2001;13(9):1119-21*

126. *Troncone R, Auricchio R, Paparo F, et al. Coeliac disease and extraintestinal autoimmunity. J Pediatr Gastroenterol Nutr 2004;39(Suppl 3):S740 1*

127. *Villalta D, Girolami D, Bidoli E, et al. High prevalence of celiac disease in autoimmune hepatitis detected by anti-tissue tranglutaminase autoantibodies. J Clin Lab Anal 2005;19(1):6-10*

128. *van Heel DA, Hunt K, Greco L, Wijmenga C. Genetics in coeliac disease. Best Pract Res Clin Gastroenterol 2005;19(3):323-39*

*•• Review on the involvement of susceptible genes in celiac disease.*

129. *Heward JM, Allahabadia A, Armitage M, et al. The development of Graves' disease and the CTLA-4 gene on chromosome 2q33. J Clin Endocrinol Metab 1999;84(7):2398-401*

130. *Plenge RM, Padyukov L, Remmers EF, et al. Replication of putative candidate- gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet 2005;77(6):1044-60*

131. *Hudson LL, Rocca K, Song YW, Pandey JP. CTLA-4 gene polymorphisms in systemic lupus erythematosus: a highly significant association with a determinant in the promoter region. Hum Genet 2002;111(4-5):452-5*

132. *Vaidya B, Pearce S. The emerging role of the CTLA-4 gene in autoimmune endocrinopathies. Eur J Endocrinol 2004;150(5):619-26*

133. *Postma DS, Bleecker ER, Amelung PJ,et al. Genetic susceptibility to asthma – bronchial hyperresponsiveness coinherited with a major gene for atopy. N Engl J Med 1995;333(14):894-900*

134. *Ma Y, Ohmen JD, Li Z, et al. A genome- wide search identifies potential new susceptibility loci for Crohn's disease. Inflamm Bowel Dis 1999;5(4):271-8*

135. *Martin-Pagola A, Perez de NG, Vitoria JC, et al. No association of CTLA4 gene with celiac disease in the Basque population. J Pediatr Gastroenterol Nutr 2003;37(2):142-5*

136. *Holopainen P, Arvas M, Sistonen P, et al. CD28/CTLA4 gene region on chromosome 2q33 confers genetic susceptibility to celiac disease. A linkage and family-based association study. Tissue Antigens 1999;53(5):470-5*

137. *Hunt KA, Monsuur AJ, McArdle WL, et al. Lack of association of MYO9B genetic variants with coeliac disease in a British cohort. Gut 2006;55(7):969-72*

138. *Cirillo G, Di Domenico MR, Corsi I, et al. Do MYO9B genetic variants predispose to coeliac disease? An association study in a cohort of South Italian children. Dig Liver Dis 2007;39(3):228-31*

139. *Amundsen SS, Monsuur AJ, Wapenaar MC, et al. Association analysis of MYO9B gene polymorphisms with celiac disease in a Swedish/Norwegian cohort. Hum Immunol 2006;67(4-5):341-5*

# Cost-effective HLA typing with tagging SNPs predicts celiac disease risk haplotypes in the Finnish, Hungarian, and Italian populations

Lotta Koskinen [1,*], Jihane Romanos [2,*], Katri Kaukinen [3], Kirsi Mustalahti [4], Ilma Korponay-Szabo [5], Donatella Barisani [6], Maria Teresa Bardella [7,8], Fabiana Ziberna [9], Serena Vatta [9], György Széles [10], Zsuzsa Pocsai [11], Kati Karell [12], Katri Haimila [12], Róza Ádány [11], Tarcisio Not [9], Alessandro Ventura [9], Markku Mäki [4], Jukka Partanen [12], Cisca Wijmenga [2], Päivi Saavalainen [1]

[1] Department of Medical Genetics, and Research Program for Molecular Medicine, Biomedicum Helsinki, Finland University of Helsinki, Helsinki Finland; [2] Genetics Department, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands; [3] Department of Gastroenterology and Alimentary Tract Surgery, Tampere University Hospital and Medical School, University of Tampere, Tampere, Finland; [4] Paediatric Research Centre, University of Tampere Medical School and Tampere University Hospital, University of Tampere, Finland; [5] Heim Pal Children's Hospital, Budapest and University of Debrecen, Hungary; [6] Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca, Monza,Italy; [7] Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Italy; [8] Department of Medical Sciences, University of Milan, Milan, Italy; [9] Department of Reproductive and Development Sciences, University of Trieste and IRCCS "Burlo Garofolo" Children Hospital, Trieste, Italy; [10] Faculty of Public Health, Department of Epidemiology and Biostatistics, University of Debrecen, Debrecen, Hungary; [11] Faculty of Public Health, Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary; [12] Research and Development, Finnish Red Cross Blood Service, Helsinki, Finland.

* These authors contributed equally to this work.

*Immunogenetics 2009; 61(4):247-56.*

# CHAPTER 2

# Abstract



Human leukocyte antigen (HLA) genes, located on chromosome 6p21.3, have a crucial role in susceptibility to various autoimmune and inflammatory diseases, such as celiac disease and type 1 diabetes. Certain HLA heterodimers, namely DQ2 (encoded by the DQA1*05 and DQB1*02 alleles) and DQ8 (DQA1*03 and DQB1*0302), are necessary for the development of celiac disease. Traditional genotyping of HLA genes is laborious, time-consuming, and expensive. A novel HLA-genotyping method, using six HLA-tagging single-nucleotide polymorphisms (SNPs) and suitable for high-throughput approaches, was described recently. Our aim was to validate this method in the Finnish, Hungarian, and Italian populations. The six previously reported HLA-tagging SNPs were genotyped in patients with celiac disease and in healthy individuals from Finland, Hungary, and two distinct regions of Italy. The potential of this method was evaluated in analyzing how well the tag SNP results correlate with the HLA genotypes previously determined using traditional HLA-typing methods. Using the tagging SNP method, it is possible to determine the celiac disease risk haplotypes accurately in Finnish, Hungarian, and Italian populations, with specificity and sensitivity ranging from 95% to 100%. In addition, it predicts homozygosity and heterozygosity for a risk haplotype, allowing studies on genotypic risk effects. The method is transferable between populations and therefore suited for large-scale research studies and screening of celiac disease among high-risk individuals or at the population level.

**Keywords:** HLA, Human leukocyte antigen, Celiac disease, Tagging SNP.

# Introduction



Human leukocyte antigen (HLA) genes are known to confer risk to several autoimmune and inflammatory disorders, such as celiac disease and type 1 diabetes. The HLA genes are located in the major histocompatibility complex (MHC) region on human chromosome 6p21.3. This region is highly polymorphic and it includes several genes that have an essential role in immune responses. MHC genes can be divided into three classes: class I genes (in particular, *HLA-A, HLA-B, HLA-C*), class II genes (*HLA-DR, HLA-DQ, HLA-DP*), and class III genes, which are a more heterogenic group of various, mostly immune-related, genes.

Celiac disease, or gluten intolerance, is a chronic inflammatory disease of the small intestine, with autoimmune features. It is triggered, in genetically susceptible individuals, by dietary gluten exposure from wheat, barley, or rye. Up to 1% of the Caucasian population has celiac disease, although the disease is often silent (symptom free) or presents with atypical or mild symptoms. To date, the only confirmed and functionally characterized genetic risk factors of celiac disease are HLA-DQ2 and HLA-DQ8 haplotypes, coded by MHC class II genes. Approximately 90% of the European Caucasian patients with celiac disease carry the HLA-DQ2 heterodimer coded by alleles DQA1*05 and DQB1*02 (Sollid et al. 1989; Karell et al. 2003). A majority of the patients carry the DQ2.5 (or DR3-DQ2) haplotype, where the alpha and beta chains of the DQ2 heterodimer are encoded together in cis on a DRB1*03 haplotype (including alleles DRB1*03, DQA1*0501, and DQB1*0201). The DQ heterodimer can also be encoded in trans configuration by the DQ2.2 (DR7-DQ2) and DQ7 (DR5/6-DQ7) haplotypes, with the DQA1*05 allele deriving from DRB1*11, *12, or *13 haplotypes (DRB1*11/12/ 13, DQA1*0505, DQB1*0301) and the DQB1*02 allele deriving from DRB1*07 haplotype (DRB1*07, DQA1*02, DQB1*0202; Sollid and Thorsby 1993; Mazzilli et al. 1992). The DQ8 heterodimer encoded by the DR4-DQ8 haplotype (DRB1*04, DQA1*03, DQB1*0302) is common in celiac patients who do not carry the DQ2 heterodimer (Spurkland et al. 1992). Both DQ2 and DQ8 heterodimers are known to have a central role in the pathogenesis of celiac disease (Molberg et al. 1998). Only a small number of patients (6%) carry neither DQ2 nor DQ8, and the vast majority of these carry just one chain of the DQ2 heterodimer, i.e., either the DQ2.2 or DQ7 haplotype (Karell et al. 2003; Polvi et al. 1998). Therefore, practically only individuals who carry either the DQ2 or DQ8 haplotype can become gluten intolerant, although other genetic and environmental risk factors are also required, as these haplotypes are common also in the healthy population (Sollid et al. 1989; Polvi et al. 1996).

HLA testing is traditionally performed using serology, by DNA-based methods with sequence-specific primer or sequence-specific oligonucleotide approaches or by hybridization and fluorescence detection. These methods are relatively laborious and costly for research or high- throughput screening purposes. Recently, Monsuur et al. (2008) described a method for detecting the HLA risk alleles for celiac disease using HLA-tagging

single-nucleotide polymorphisms (SNPs). In this method, six SNPs were reported to tag the risk haplotypes for celiac disease, and the genotyping could be performed in a high-throughput mode. The results showed that the sensitivity and specificity of this test to recognize the DQ2.2, DQ2.5, DQ7, and DQ8 haplotypes were above 99% in the Dutch population and very high also in the UK, Spanish, and Italian populations. The tag SNP selection was based on genotype data collected in the classical HLA genes and more than 7,500 common SNPs and insertion–deletion polymorphisms across the human MHC region (de Bakker et al. 2006).

In this study, we genotyped and evaluated the usefulness of these HLA-tagging SNPs in the Finnish, Hungarian, and two independent Italian sample sets and investigated further the distribution of celiac-disease-related HLA risk factors in these populations.

# Materials and methods

## Samples

*Finnish population*

The Finnish sample set consisted of three cohorts: 85 families with celiac disease including altogether 278 family members, 210 unrelated single cases with celiac disease, and 176 control individuals. Altogether, this material consisted of 430 celiac disease patients and 664 samples from Finnish individuals. The collection of the Finnish celiac disease families has been described earlier (Mustalahti et al. 2002b). The Finnish single cases with celiac disease were collected at the Department of Gastroenterology and Alimentary Tract Surgery in Tampere University Hospital. A majority of the patients were diagnosed according to the ESPGAN (1990) criteria, the rest being positive for the endomysial antibody (EmA) specific for celiac disease. Eighty-six EmA-negative adults collected at the same clinic were used as non-celiac controls in addition to 90 other controls representing the population density of Finland excluding Lapland and Northern Karelia.

*Hungarian population*

The Hungarian sample set consisted of 177 patients with celiac disease and 179 population controls. The collection of the Hungarian celiac patients and population controls has been described previously (Koskinen et al. 2008; Szeles et al. 2005; Central Statistical Office 2001, Hungary).

*Italian population*

The Italian samples were collected from two distinct regions of Italy. The first set was from Trieste in northeastern Italy and consisted of 134 celiac disease patients and 202 healthy individuals while the second set was from Milan, north Italy, and was composed of 543 cases and 592 controls. We analyzed the two sets separately.

All celiac patients were diagnosed in accordance with ESPGAN (1990) criteria and the intestinal biopsies were analyzed using the classification of Oberhuber et al. (1999). In addition, patients' serum samples tested positive for both anti-transglutaminase and anti-endomysium antibodies.

## Ethics

The collection of the patient and control materials were approved by the ethical committees of the Tampere University Hospital, Heim Pal Children's Hospital, Budapest, Hungary, the University of Debrecen, the Independent Local Ethical Committee of the Burlo Garofolo Children's Hospital in Trieste and the hospital Fondazione Istituto Di Ricovero e Cura a Carattere Scientifico (IRCCS) Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milano, Italy. All enrolled participants were informed about the study according to the study protocol and gave written informed consent.

## HLA typing

From the Finnish family cohort, 212 individuals had previously been genotyped for the *DQB1* and *DRB1* genes. The genotyping of the DQB1 polymorphisms was performed using the Olerup SSP DQB1 low-resolution kit (Olerup SSP AB, Saltsjöbaden, Sweden). DRB1 genotypes were determined using HLA-linked microsatellite markers. This method has been described earlier by Karell et al. (2000). For 136 Finnish unrelated patients and 52 controls without celiac disease, the HLA typing was performed using the DELFIA® Celiac Disease Hybridization Assay Kit (PerkinElmer Life and Analytical Sciences, Wallac Oy, Turku, Finland). This method detects the positivity or negativity for alleles DQA1*05, DQB1*02, and DQB1*0302. The HLA genotyping of the Finnish cohorts was performed in the European Foundation of Immunogenetics accredited tissue-typing laboratory of the Finnish Red Cross Blood Service, Helsinki, Finland.

From the Hungarian population, 78 patients had been genotyped for the DQB1 and DRB1 polymorphisms using the Olerup SSP DQ low-resolution and Olerup SSP DR low-resolution kits (Olerup SSP AB, Salt- sjöbaden, Sweden).

Among the Italian sets, 97 of the celiac cases from the Trieste region had been previously genotyped for the *DQB1* and *DQA1* genes in the Genetic Unit of IRCCS "Burlo Garofolo" Children's Hospital, Trieste, Italy, using the low- and high-resolution Dynal Classic SSP DQ Kits (Dynal A. S., Oslo, Norway) based on polymerase chain reaction (PCR) with allele-specific primers.

## SNP genotyping

Six HLA-tagging SNPs reported by Monsuur et al. (2008) were genotyped in the Finnish, Hungarian, and Italian materials using TaqMan chemistry and the On Demand assays by Applied Biosystems (Applied Biosystems,

**Table 1.** The six single nucleotide markers reported to tag the celiac disease HLA risk haplotypes by Monsuur et al.(2008).

| rs-number | Allele call | | Applied Biosystems assay number | Basepair position | Call rates | Tags DQ type | Positive predicting allele | Negative predicting allele |
|---|---|---|---|---|---|---|---|---|
| | VIC | FAM | | | | | | |
| rs2187668 | C | T | C__58662585_10 | 32713862 | 0.97 | DQ2.5 | T | |
| rs2395182 | G | T | C__11409965_10 | 32521295 | 0.96 | DQ2.2 | T | |
| rs4713586 | A | G | C__27950246_10 | 32767560 | 0.92 | DQ2.2 | | G |
| rs7775228 | C | T | C__29315313_10 | 32766057 | 0.98 | DQ2.2 | C | |
| rs4639334 | A | G | C__42975350_10 | 32710192 | 0.97 | DQ7 | A | |
| rs7454108 | C | T | C__29817179_10 | 32789461 | 0.96 | DQ8 | C | |

Foster City, CA, USA, www.appliedbiosystems.com; Table 1). The samples were genotyped by applying the standard protocol provided by Applied Biosystems. The PCR assays and allelic discrimination were run using an ABI PRISM 7900HT Sequence Detection System instrument (Applied Biosystems, Foster City, CA, USA). The control and case individuals were always run on same 384-well plates to prevent biased genotyping results due to technical issues. The genotype call rates shown in Supplementary Table 1 were higher than 95% except for SNP rs4713586 for which the call rate was only 92% in some populations even though 20 ng of DNA in a 4-µl reaction volume and 45 amplification cycles were used. The Hardy–Weinberg equilibrium (HWE) for the six SNPs was calculated in the cases and controls from Finnish, Hungarian, and two Italian populations (Supplementary Table 1). The controls for each of the four populations were in HWE for all six SNPs ($p > 0.05$). In the patient groups, some SNPs were out of HWE due to the strong HLA association in celiac disease rather than caused by genotyping errors. In the family dataset, the genotyping results were tested for Mendelian errors by the PedCheck program (O'Connell and Weeks 1998). When a Mendelian inconsistency was detected, the related genotypes were discarded from the analysis. There were no Mendelian errors for rs2395182, rs4713586, rs7775228, and rs4639334. The Mendelian error rate for both rs2187668 and rs7454108 was 0.6% in the Finnish families.

Interpreting the DQ7-tagging rs4639334 genotyping results required special caution since we identified four genotype clusters instead of the expected three clusters (denoting two different homozygotes and the heterozygotes) in the genotyping results of this SNP.

The extra cluster was located between the heterozygote cluster and the cluster of homozygotes for the minor allele. When we looked at publicly available databases for the DNA sequence spanning the SNP rs4639334 and genomic variation in that region, we found four SNPs that were located close to rs4639334 and thus in the sequence complementary to the probes. These SNPs are most probably causing this effect of more than three clusters appearing at the allele detection (Franke et al. 2008). According to current knowledge, there is no proxy to this SNP. The individuals showing an extra cluster in the rs4639334 were determined as DQ7 heterozygotes using traditional HLA typing. If samples in this cluster are mistakenly called as homozygotes, the existence of DQ7 is still correctly predicted, although, being aware of this problem, the genotypes can be called correctly. The fourth cluster was seen in controls of all populations but not in cases, which might result from different distribution of the various DQ7-positive haplotypes between the groups, although to prove this would require typing of DRB1 and other HLA alleles from the material. We were unable to determine the linkage disequilibrium between the DQ7- tagging SNP and the four SNPs under the probe since they are not genotyped by the HapMap project. The failure of Monsuur et al. (2008) to identify this problem can be due to the low frequency of the causative polymorphism in the Dutch population. In addition, the knowledge of SNPs under probes disrupting the binding of the probe was not known at the time of the experiment and those samples were considered to be dropouts. The future genome-wide association studies on celiac disease will hopefully give more light to the diversity of DQ7 and other haplotypes in population and will hopefully also reveal new alternative tagging SNPs to overcome the clustering problem with rs4639334.

**Table 2.** Determination of DQ types from the DELFIA results

| DELFIA result | | | |
|---|---|---|---|
| DQB1*02 | DQB1*0302 | DQA1*05 | DQ-type |
| + | + | + | DQ2.5 / DQ8 |
| + | + | - | DQ2.2 / DQ8 |
| + | - | - | DQ2.2 / DQX |
| - | + | + | DQ8 / DQ7 |
| - | - | + | DQ7 /DQX |
| + | - | + | DQ2.5 or DQ2.2 / DQ7 |
| - | + | - | DQ8 / DQX |
| - | - | - | DQX / DQX |

DQX = not DQ2.2, DQ2.5, DQ7, DQ8.

### Validation

The HLA alleles previously typed using the commercial kits, Olerup SSP and DELFIA, in addition to the microsatellite method for a portion of the samples, were used to determine the DQ types. These methods differ in their resolution. Using the Olerup SSP DQB1 low-resolution kit together with the DRB1, DQA1 low-resolution kit, or the microsatellite method, it is possible to detect the risk haplotypes known to be associated with celiac disease (DQ2.2, DQ2.5, DQ7, DQ8) in both homozygous and heterozygous forms in addition to other common HLA haplotypes.

With the DELFIA kit, it is possible to identify individuals who are carriers of HLA-DQA1*05, HLA-DQB1*02, and HLA-DQB1*0302 alleles. It is not possible to distinguish homozygotes from the heterozygotes in all combinations or DQ2.5-positive patients from those who have both DQ2.2 and DQ7. The DQ types were determined from the DELFIA results as shown in Table 2.

The results from the tagging SNP genotypes were compared to the previously established genotypes. The DQ types were determined from the tagging SNP results as in the study by Monsuur et al. (2008; see also Table 1). If a mismatch between the two typing methods was found, the DQ type was verified in our laboratory using the Olerup SSP DQB1 and DRB1 low-resolution kits (Olerup SSP AB, Saltsjöbaden, Sweden). The DQ-typing results from the traditional typing methods and the tagging SNP approach were compared to study the sensitivity, specificity, positive predictive value, and correlation (r2) of the test. From family materials, all available family members with sufficient DR-DQ data available were counted, but only one patient per family was selected for the haplotype frequency estimations.

### Statistical analysis

To assess the celiac diseases risk conferred by different HLA genotypes, we conducted logistic regression analysis using SPSS software version 14.0

## Results

### Validation

To predict the DQ2.2, DQ2.5, DQ7, and DQ8 risk haplotypes for celiac disease, six tagging SNPs were genotyped in 400 Finnish individuals (212 from family cohort, 136 unrelated patients, and 52 controls), 79 Hungarian patients, and 97 Trieste-Italian cases from whom traditional HLA genotypes were available. The genotype call rates, HWE, and Mendelian errors are described in “Materials and methods” and shown in Supplementary Table 1.

**Table 3**. Validation results in the Finnish, Hungarian and Italian (Trieste) samples.

| | Finland | Hungary | Italy (Trieste) | ALL |
|---|---|---|---|---|
| DQ2.2 | | | | |
| Number of chromosomes tested | 354 | 152 | 166 | 672 |
| Sensitivity | 1 | 1 | 1 | 1 |
| Specificity | 1 | 1 | 1 | 1 |
| PPV | 1 | 1 | 1 | 1 |
| r-squared | 1 | 1 | 1 | 1 |
| false results | 0 | 0 | 0 | 0 |
| DQ2.5 | | | | |
| Number of chromosomes tested | 388 | 158 | 186 | 732 |
| Sensitivity | 0.98 | 1 | 1 | 0.99 |
| Specificity | 0.99 | 1 | 1 | 0.99 |
| PPV | 0.99 | 1 | 1 | 0.99 |
| r-squared | 0.94 | 1 | 1 | 0.97 |
| false results | 0.015 | 0 | 0 | 0.008 |
| DQ7 | | | | |
| Number of chromosomes tested | 422 | 158 | 186 | 766 |
| Sensitivity | 1 | 1 | 1 | 1 |
| Specificity | 1 | 0.99 | 0.99 | 0.997 |
| PPV | 1 | 0.95 | 0.97 | 0.97 |
| r-squared | 1 | 0.94 | 0.96 | 0.97 |
| false results | 0 | 0.006 | 0.0005 | 0.0026 |
| DQ8 | | | | |
| Number of chromosomes tested | 736 | 158 | 192 | 1086 |
| Sensitivity | 0.95 | 1 | 1 | 0.97 |
| Specificity | 0.999 | 0.99 | 0.99 | 0.996 |
| PPV | 0.98 | 0.88 | 0.83 | 0.93 |
| r-squared | 0.93 | 0.87 | 0.82 | 0.81 |
| false results | 0.0041 | 0.006 | 0.01 | 0.0055 |

ALL = Finland, Hungary and Italy combined, PPV = positive predictive value.

In the Finnish sample sets, the validation for the DQ2.5, DQ2.2, and DQ7 haplotypes was done only for the family cohort. The cases and controls had been typed using the DELFIA assay, by which the presence of these haplotypes cannot be determined in all situations when the parental haplotypes are unknown. The results of the validation presented in Table 3 show sensitivity and specificity ranging between 0.95 and 1 (for more detailed information, see Supplementary Table 2). In the samples tested for HLA using the DELFIA kit, seven individuals were DQA1*05 positive and DQB1*02 negative, implying that these were DQ7 positive. Five of the samples were DQB1*02 positive and DQA1*05 negative and could thus be determined as DQ2.2 positives. The tagging SNP results to determine DQ7 and DQ2.2 matched these results. One hundred eight individuals were both DQA1*05 and DQB1*02 positive, so it could not be determined from the DELFIA results whether they had DQ2.5 or both DQ2.2 and DQ7. The tagging SNP method predicted 105 of these to carry the DQ2.5 haplotype and three of them to have the DQ2.2/DQ7 genotype. In the Hungarian and Trieste-Italian populations, the sensitivities and specificities of the test ranged between 0.99 and 1 (Table 3, for more detailed information, see Supplementary Table 2).

The Finnish, Hungarian, and Trieste-Italian materials were also analyzed together for the correlation. The overall correlation was high; the sensitivities for DQ2.2, DQ2.5, DQ7, and DQ8 were between 0.97 and 1, and the specificities were between 0.996 and 1 (Table 3, for more detailed information, see Supplementary Table 2). Out of all the 576 tested individuals, 12 (2.1%) showed different results in the tagging SNP assay when compared to the traditional HLA-typing results (Table 4). Nine of them were celiac-disease-affected individuals from the Finnish family cohort; one was a Hungarian celiac patient and two were Italian (Trieste) celiac patients. In addition, four patients from the Italian (Milan) sample set were predicted to have three HLA alleles by the tag SNP method but, when genotyped for the *DQB1* and *DRB1* genes using the Olerup SSP low-resolution kits, two of the samples showed a rare DR11-DQ2 haplotype (Table 4).

## Population samples

The six HLA risk haplotype tagging SNPs were also genotyped in an extended set of Finnish, Hungarian, and Italian (Milan and Trieste) cases and controls. The HLA haplotype and genotype frequencies determined from the SNP allele frequencies are presented in Tables 5 and 6. Excess of DQ2.5 among patients was seen both in haplotype (Table 5) and carrier (Table 6) frequencies in all three populations. According to the SNP genotype results, 90.2% of the Finnish patients with celiac disease carried the DQ2 heterodimer (88.3% DQ2.5 and 1.9% DQ2.2/DQ7), and 6.4% carried the DQ8 haplotype without DQ2. Of the Hungarian patients, 97.2% were carriers of the DQ2 heterodimer (87.5% DQ2.5 and 9.7% DQ2.2/DQ7 haplotype) and 2.3% of the patients were carriers of the DQ8 haplotype without DQ2. Among the Italian (Trieste) celiac patients, 74.7% were carriers for the DQ2.5 haplotype and 14.9% of them carried the DQ2.2/DQ7 haplotypes; 7.5% of them had the

DQ8 haplotype without DQ2. In the Italian (Milan) samples, 64.1% of the patients were carriers for the DQ2.5 haplotype, 24.3% carriers for the DQ2.2/DQ7 haplotype, and 6.2% carriers for the DQ8 haplotype without DQ2.

Dose effect of the DQ2.5 homozygosity on celiac disease risk was observed in all three populations when comparing the genotypes of cases and population controls (Table 6). Higher DQ2.2 haplotype and DQ2.5/DQ2.2 genotype frequencies were seen among the patients in Hungary and Italy but to a lesser extent in Finland (Tables 5 and 6). In addition, the tag SNP method allowed us to determine whether an individual is homozygous or heterozygous for the risk haplotypes, and thus we calculated the odds ratios of different risk HLA genotypes in different populations (Supplementary Table 3). Due to small sample size, the Hungarian sample set was uninformative for this analysis in addition to genotypes DQ2.2/DQ2.5, DQ2.5/DQ2.2, DQ2.2/DQ2.2, and DQ7/DQX in the Finnish sample set. Using logistic regression in the rest of the groups, we were able to see very high risk effects in genotypes with at least one DQ2.5 haplotype compared to no-risk genotype (DQX/DQX). A more moderate risk was identified in genotypes with DQ8 and DQ2.2 and no increase in risk in the presence of only DQ7. In addition, we studied the risk effect of the second haplotype in the presence of one DQ2.5 haplotype in all four population groups (data not shown). Only Italian-Milan population group had a sufficiently big sample size for this analysis, and thus the results are more reliable for this group than for the other groups with smaller sample size. The results suggest that homozygosity for DQ2.5 increases the risk of celiac disease 5.5-fold when compared to individuals with DQ2.5/DQX, while DQ2.2 increases the risk of celiac disease 3.1-fold. DQ7 or DQ8 in the presence of DQ2.5 did not confer additional risk to the disease.

# Discussion

Celiac disease is an important health problem because of its high prevalence, associated specific and non-specific morbidity, and long-term complications (Mearin et al. 2005; Romanos et al. 2008). Early diagnosis is very important, especially in high-risk groups such as first-degree relatives and individuals with type 1 diabetes, iron deficiency anemia, or Down syndrome. Although the presence of HLA-DQ2 or HLA-DQ8 is not sufficient alone for diagnosis, it indicates a need for further serology tests and later a biopsy sampling in the risk groups. Also, their absence reduces the risk for the disease to very low.

Testing for HLA risk molecules is routinely performed using methods which require several reactions, multiple steps such as amplification and hybridization to a membrane, and special software or expertise in analyzing the results, and most of them are expensive. To make HLA typing more automated and relatively cheap, Monsuur et al. (2008) established a new approach using six tagging SNPs to predict whether an individual is heterozygous or homozygous for the DQ2.5, DQ2.2, DQ7, and DQ8 risk genotypes. Genotyping the six

**Table 4.** The DQ types of the samples giving false results.

| Cohort | Sample | HLA-type: Chromosome 1 | HLA-type: Chromosome 2 | DQ type (SNPs) | Comments |
|---|---|---|---|---|---|
| **Finland** | 1. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*1402 DQB1*0301* | *DQ2.5*<br>*DQ2.5* | 2 *DQ2.5* predicted, only one present. Rare *DR14;DQ7* haplotype instead. |
| | 2. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*1501 DQB1*0601* | *DQX*<br>*DQX* | *DQ2.5* present but not predicted. |
| | 3. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*0401 DQB1*0302* | *DQ2.5*<br>*DQX* | *DQ8* present but not predicted. |
| | 4. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*0401 DQB1*0302* | *DQ2.5*<br>*DQ8 ; DQ8* | Two *DQ8* predicted, only one present. |
| | 5. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*0701 DQB1*0202* | *DQX*<br>*DQX* | *DQ2.5* present but not predicted. *DQ2.2* could not be predicted due to failed genotyping for this sample. |
| | 6. affected familymember | *DRB1*0401 DQB1*0302* | *DRB1*1601 DQB1*0501* | *DQX*<br>*DQX* | *DQ8* present but not predicted. |
| | 7. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*1301 DQB1*06* | *DQ2.5*<br>*DQ2.5* | 2 *DQ2.5* predicted, only one present. |
| | 8. affected familymember | *DRB1*0301 DQB1*0201* | *DRB1*0701 DQB1*0202* | *DQX*<br>*DQX* | *DQ2.5* present but not predicted. *DQ2.2* could not be predicted due to failed genotyping for this sample. |
| | 9. affected familymember | *DRB1*0301 DQB1*0201* | *DR2*<br>*DQ6* | *DQX*<br>*DQX* | *DQ2.5* present but not predicted. |
| **Hungary** | 10. affected case | *DRB1*0701 DQB1*0202* | *DRB1*1101 DQB1*0302* | *DQ2.2*<br>*DQ7 ; DQ8* | *DQ2.2*, *DQ7* and *DQ8* predicted, *DQ2.2* and a rare *DR11-DQ8haplotype* present (*DQA1*05* not necessarily included). |
| **Italy (Trieste)** | 11. affected case | *DRB1*0301 DQA*0501 DQB1*0201* | *DRB1*0801 DQA1*0301 DQB1*0302* | *DQ2.5*<br>*DQ8 ; DQ8* | Two *DQ8* predicted, one present (rare *DR8;DQ8*). |
| | 12. affected case | *DRB1*0701 DQA*0201 DQB1*02* | *DRB1*1102 DQA1*0505 DQB1*03* | *DQ2.2*<br>*DQ7 ; DQ7* | Two *DQ7* predicted, only one present. |
| **Italy (Milan)** | 13. affected case | *DRB1*0701-DQB1*0201/0202* | *DRB1*11-DQB1*0201/0202/0203* | *DQ2.2*<br>*DQ2.2 ;DQ7* | 2 *DQ2.2* predicted, only one present. *DQ7* predicted, but not present. Instead a rare *DR11-DQ2* haplotype present. |
| | 14. affected case | *DBBI*0301-DQB1*0201* | *DRB1*11-DQB1*0301* | *DQ2.5*<br>*DQ8 ; DQ7* | *DQ8* predicted but not present. *DQ2.5* and *DQ7* correctly predicted. |
| | 15. affected case | *DBBI*0301-DQB1*0201* | *DRB1*11-DQB1*0301* | *DQ2.5*<br>*DQ2.2 ;DQ7* | *DQ2.2* predicted but not present. *DQ2.5* and *DQ7* correctly predicted |
| | 16. affected case | *DRB1*0701-DQB1*0201/0202* | *DRB1*11-DQB1*0201/0202/0203* | *DQ2.2*<br>*DQ2.2 ;DQ7* | 2 *DQ2.2* predicted, only one present. *DQ7* predicted, but not present. Instead a rare *DR11-DQ2* haplotype present. |

DQX = not DQ2.2, DQ2.5, DQ7, DQ8.

**Table 5.** HLA haplotype frequencies (%) in the Finnish, Hungarian, and Italian controls and cases with celiac disease determined from the tagging SNP results.

| | Finland | | Hungary | | Italy (Trieste) | | Italy(Milan) | | CEU | |
|---|---|---|---|---|---|---|---|---|---|---|
| **DQ type** | **Controls (352)** | **Patients (530)** | **Controls (358)** | **Patients (352)** | **Controls (404)** | **Patients (268)** | **Controls (1164)** | **Patients (1070)** | | |
| ***DQ2.2*** | 6 | 5.8 | 8.9 | 20.1 | 11.9 | 17.2 | 10.8 | 27.1 | 13.3 (*DQA1*0201*) | |
| ***DQ2.5*** | 8.8 | 50.9 | 10.1 | 49.6 | 15.8 | 44.8 | 8.4 | 37.1 | 15.6 (*DQB1*0201*) | |
| ***DQ7*** | 6.8 | 4 | 21.5 | 13 | 29.5 | 16 | 27.8 | 17.9 | 16.7 (*DQB1*0301*) | |
| ***DQ8*** | 11.1 | 6.2 | 9.8 | 2.8 | 5.4 | 6.7 | 5.7 | 4.6 | 14.4 (*DQB1*0302*) | |
| **DQX** | 67.3 | 33 | 49.7 | 14.4 | 37.4 | 15.3 | 47.3 | 13.3 | | |
| **Total** | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | | |

The total number of chromosomes is shown in brackets. DQX = not DQ2.2, DQ2.5, DQ7, DQ8. *CEU = HLA allele frequencies in the CEU population from article by de Bakker et al.(2006)

SNPs is simple, fast, and cost-effective compared to more classical techniques. In the Dutch, Spanish, and Italian (Naples) analyzed populations, the sensitivity of this test was reported to be >0.991 and the specificity >0.996 (Monsuur et al. 2008). Since linkage disequilibrium in the MHC region can be different in different populations, our aim was to validate this method in four new populations.

We genotyped the six HLA-tagging SNPs in the Finnish, Hungarian, and northern Italian (Trieste) populations. We then compared the deduced genotype from the tagging SNP with the traditional HLA typing. Our results showed that the sensitivity and specificity to detect the celiac disease risk alleles in the Finnish celiac disease patients and controls were >0.95 and >0.99, respectively. In the Hungarian patients with celiac disease, the sensitivity was 1 for each risk allele and the specificity was >0.99. In the Italian (Trieste) patients with celiac disease, the sensitivity was 1 for each tested allele and the specificity was >0.99. These results imply that with this method it is possible to detect the DQ2 and DQ8 alleles with high accuracy. In addition, the method is transferable to other populations since it is proven to be a good test in five different geographic groups, Dutch, UK, Finnish, Hungarian, and Italian.

Moreover, additional sets of controls from each population were also genotyped as well as a case control material from Milan. Although no previous HLA-typing results were available from them, the HLA allele frequencies determined by the SNP method followed the known HLA

**Table 6.** HLA-genotype frequencies (%) in the Finnish, Hungarian and Italian controls and cases with celiac disease, determined from the tagging SNP results.

| Genotype | Finland | | Hungary | | Italy (Trieste) | | Italy (Milan) | |
|---|---|---|---|---|---|---|---|---|
| | Controls (176) | Patients (265) | Controls (179) | Patients (176) | Controls (202) | Patients (133) | Controls (582) | Patients (535) |
| **ALL DQ2+** | **17.6** | **90.2** | **21.8** | **97.2** | **36.1** | **89.6** | **22.5** | **88.4** |
| DQ2.5/DQX | 13.6 | 58.5 | 10.1 | 28.4 | 13.4 | 23.1 | 6.9 | 18.3 |
| DQ2.5/DQ2.5 | 0.6 | 13.6 | 2.2 | 11.9 | 3 | 14.9 | 0.7 | 10.1 |
| DQ2.5/DQ2.2 | 0 | 6.4 | 1.7 | 29.5 | 3 | 15.7 | 2.9 | 23.9 |
| DQ2.5/DQ7 | 1.1 | 4.9 | 3.4 | 14.2 | 7.9 | 15.7 | 4.1 | 9.3 |
| DQ2.5/DQ8 | 1.7 | 4.9 | 0.6 | 3.4 | 1.5 | 5.2 | 1.5 | 2.4 |
| DQ2.2/DQ7 | 0.6 | 1.9 | 3.9 | 9.7 | 7.4 | 14.9 | 6.4 | 24.3 |
| **ALL DQ2-, DQ8+** | **19.3** | **6.4** | **17.3** | **2.3** | **8.9** | **7.5** | **9.5** | **6.2** |
| DQ8/DQX | 15.9 | 4.2 | 10.1 | 0 | 5 | 3.7 | 5 | 2.6 |
| DQ8/DQ8 | 1.1 | 0.8 | 1.7 | 0 | 0 | 0.7 | 0.3 | 0.6 |
| DQ8/DQ2.2 | 1.7 | 1.1 | 2.2 | 1.1 | 1 | 1.5 | 0.7 | 2.1 |
| DQ8/DQ7 | 0.6 | 0.4 | 3.4 | 1.1 | 3 | 1.5 | 3.4 | 0.9 |
| **ALL DQ2-, DQ8-** | **63.1** | **3.4** | **60.9** | **0.6** | **55** | **2.9** | **68** | **5.4** |
| DQ2.2/DQX | 9.7 | 1.5 | 10.1 | 0 | 9.4 | 2.2 | 10.3 | 1.7 |
| DQ2.2/DQ2.2 | 0 | 0.4 | 0 | 0 | 1.5 | 0 | 0.7 | 1.1 |
| DQ7/DQX | 9.1 | 0.8 | 22.3 | 0.6 | 16.8 | 0 | 28.4 | 1.3 |
| DQ7/DQ7 | 1.1 | 0 | 5 | 0 | 11.9 | 0 | 6.7 | 0 |
| DQX/DQX | 43.2 | 0.8 | 23.5 | 0 | 15.3 | 0.7 | 22 | 1.3 |

The total number of chromosomes is shown in brackets. DQX= not DQ2.2, DQ2.5, DQ7, DQ8

allele frequencies in these populations. For instance, the haplotype frequencies of the Italian (Milan) set are comparable to the haplotype frequencies reported by Margaritte-Jeannin et al. (2004). In addition, being able to predict if an individual is homozygous or heterozygous for a risk haplotype allowed us to study the risk of celiac disease conferred by different HLA genotypes. Our data agree with several previous findings of dose effects of functional DQ2 heterodimers in risk of celiac disease (Margaritte-Jeannin et al. 2004; Ploski et al. 1993; Louka et al. 2002; Vader et al. 2003), most likely due to enhanced presentation of antigenic gliadin peptides to the immune system. In the presence of one DQ2.5, the second haplotype being DQ2.5 or DQ2.2 increased the risk of developing celiac disease. The sample size of our materials is, however, limited for making conclusions about the risk effects conferred by different DQ haplotypes. As the method described in this study is suitable for large-scale typing of celiac disease risk conferring HLA haplotypes, such analysis should be done in the future.

Recently, the tagging SNP method was also shown to be useful for HLA testing in type 1 diabetes and it can be useful in other disorders like systemic lupus erythematosus, rheumatoid arthritis, and other HLA DR3-DQ2 or DR4- DQ8 associated diseases (Barker et al. 2008). This method can also be used to exclude the diagnosis of celiac disease in the absence of HLA-DQ2 or HLA-DQ8 when screening high-risk groups (e.g., relatives of patients with celiac disease and patients with Down syndrome or type 1 diabetes), which has been reported to be cost reducing when compared to follow-up by antibody screening (Mearin et al. 2005; Romanos et al. 2008; Csizmadia et al. 2000; Kaukinen et al. 2002; Mustalahti et al. 2002a).

The main benefits of the SNP-typing method are significant cost and time savings and a true potential

as a high-throughput method. The TaqMan method used in this study is most cost reducing when applied to the 96 or 384 sample format, making this method most suitable for large sample cohorts, typically in research and population screening studies.

In conclusion, we confirm that the recently described HLA-tagging SNP genotyping method shows high specificity and sensitivity with celiac-disease-associated HLA risk haplotypes also in the Finnish, Hungarian, and Italian populations and is proven to be a novel cost-effective high-throughput method for HLA typing in celiac disease and other HLA-DQ2- and HLA-DQ8-associated diseases.

**Acknowledgements**

All the study subjects are warmly thanked for their participation in the study. We thank Hanne Ahola for excellent technical assistance and Cleo van Diemen for her statistical expertise. We thank Erzsébet Szathmári, Judit B. Kovács, Margit Lörincz, and Anikó Nagy for their work with the Hungarian families. Anna-Elina Lehesjoki and Albert de la Chapelle are acknowledged for providing us with the Finnish population samples.

This work and the study groups have been funded from the EU Commission by a Marie Curie Excellence Grant (FP6 contract MEXT- CT-2005-025270), the Academy of Finland, the Hungarian Scientific Research Fund (contract OTKA 61868), the University of Helsinki Funds, Biocentrum Helsinki, the Research Fund of Tampere University Hospital, the Competitive Research Funding of the Pirkanmaa Hospital District, the Yrjö Jahnsson Foundation, the Foundation of Pediatric Research, the Sigrid Juselius Foundation, the Finnish Cultural Foundation, the Maud Kuistila Memorial Foundation, the Finnish Society for Gastroenterological Research, the Finnish Celiac Disease Society, the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW), and KP6 EU grant 036383 (PREVENTCD).

The HLA-DQ haplotyping was invented at the University Medical Center Utrecht (UMC Utrecht) and will be developed and marketed by Genome Diagnostics BV. The UMC Utrecht may receive royalties from the worldwide sale of the technology. UMC Utrecht may distribute part of the royalty revenues to the inventors (Wijmenga C and Monsuur A). None of the authors report a financial or other links with Genome Diagnostics BV. Genome Diagnostics had no role in study design, data collection and analysis, decisions to publish, or preparation of the manuscript.

## References



*Barker JM, Triolo TM, Aly TA, Baschal EE, Babu SR, Kretowski A, Rewers MJ, Eisenbarth GS (2008) Two single nucleotide polymorphisms identify highest-risk diabetes human leukocyte antigen genotype: potential for rapid screening. Diabetes 57:3152–3155*

*Central Statistical Office (2001) Demographic yearbook of Hungary. Central Statistical Office, Budapest*

*Csizmadia CG, Mearin ML, Oren A, Kromhout A, Crusius JB, von Blomberg BM, Pena AS, Wiggers MN, Vandenbroucke JP (2000) Accuracy and cost-effectiveness of a new strategy to screen for celiac disease in children with Down syndrome. J Pediatr 137:756–761. doi:10.1067/mpd.2000.110421*

*de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ et al (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38:1166–1172. doi:10.1038/ng1885*

*ESPGAN (1990) Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Pediatric Gastroenterology and Nutrition. Arch Dis Child 65:909–911. doi:10.1136/adc.65.8.909*

*Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zhernakova A, Hunt KA, Blauw HM, van den Berg LH, Ophoff R, Deloukas P, van Heel DA, Wijmenga C (2008) Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet 82:1316–1333. doi:10.1016/j.ajhg.2008.05.008*

*Karell K, Klinger N, Holopainen P, Levo A, Partanen J (2000) Major histocompatibility complex (MHC)-linked microsatellite markers in a founder population. Tissue Antigens 56:45–51. doi:10.1034/ j.1399-0039.2000.560106.x*

*Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, Partanen J, European Genetics Cluster on Celiac Disease (2003) 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 64:469–477. doi:10.1016/ S0198-8859(03) 00027-2*

*Kaukinen K, Partanen J, Maki M, Collin P (2002) HLA-DQ typing in the diagnosis of celiac disease. Am J Gastroenterol 97:695–699. doi:10.1111/j.1572-0241.2002.05471.x*

*Koskinen LLE, Korponay-Szabo IR, Viiri K, Juuti-Uusitalo K, Kaukinen K, Lindfors K, Mustalahti K, Kurppa K, Adany R, Pocsai Z, Szeles G, Einarsdottir E, Wijmenga C, Maki M, Partanen J, Kere J, Saavalainen P (2008) Myosin IXB gene region and gluten intolerance: linkage to coeliac disease and a putative dermatitis herpetiformis association. J Med Genet 45:222–227. doi:10.1136/jmg.2007.053991*

*Louka AS, Nilsson S, Olsson M, Talseth B, Lie BA, Ek J, Gudjonsdottir AH, Ascher H, Sollid LM (2002) HLA in coeliac*

*disease families: a novel test of risk modification by the 'other' haplotype when at least one DQA1*05-DQB1*02 haplotype is carried. Tissue Antigens 60:147–154. doi:10.1034/j.1399- 0039.2002.600205.x*

*Margaritte-Jeannin P, Babron MC, Bourgey M, Louka AS, Clot F, Percopo S, Coto I, Hugot JP, Ascher H, Sollid LM, Greco L, Clerget-Darpoux F (2004) HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease. Tissue Antigens 63:562– 567. doi:10.1111/j.0001-2815.2004.00237.x*

*Mazzilli MC, Ferrante P, Mariani P, Martone E, Petronzelli F, Triglione P, Bonamico M (1992) A study of Italian pediatric celiac disease patients confirms that the primary HLA association is to the DQ(alpha 1*0501, beta 1*0201) heterodimer. Hum Immunol 33:133–139. doi:10.1016/0198-8859(92)90064-T*

*Mearin ML, Ivarsson A, Dickey W (2005) Coeliac disease: is it time for mass screening? Best Pract Res Clin Gastroenterol 19:441– 452. doi:10.1016/j.bpg.2005.02.004*

*Molberg O, Mcadam SN, Korner R, Quarsten H, Kristiansen C, Madsen L, Fugger L, Scott H, Noren O, Roepstorff P, Lundin KE, Sjostrom H, Sollid LM (1998) Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut- derived T cells in celiac disease. Nat Med 4:713–717. doi:10.1038/nm0698-713*

*Monsuur AJ, de Bakker PI, Zhernakova A, Pinto D, Verduijn W, Romanos J, Auricchio R, Lopez A, van Heel DA, Crusius JB, Wijmenga C (2008) Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PLoS One 3:e2270. doi:10.1371/journal. pone.0002270*

*Mustalahti K, Lohiniemi S, Collin P, Vuolteenaho N, Laippala P, Maki M (2002a) Gluten-free diet and quality of life in patients with screen-detected celiac disease. Eff Clin Pract 5:105–113*

*Mustalahti K, Sulkanen S, Holopainen P, Laurila K, Collin P, Partanen J, Maki M (2002b) Coeliac disease among healthy members of multiple case coeliac disease families. Scand J Gastroenterol 37:161–165. doi:10.1080/003655202753416812*

*Oberhuber G, Granditsch G, Vogelsang H (1999) The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur J Gastroenterol Hepatol 11:1185–1194. doi:10.1097/00042737-199910000-00019*

*O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266. doi:10.1086/301904*

*Ploski R, Ek J, Thorsby E, Sollid LM (1993) 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 41:173–177. doi:10.1111/j.1399-0039.1993.tb01998.x*

*Polvi A, Eland C, Koskimies S, Maki M, Partanen J (1996) HLA DQ and DP in Finnish families with celiac disease. Eur J Immunogenet 23:221–234. doi:10.1111/j.1744-313X.1996.tb00117.x*

*Polvi A, Arranz E, Fernandez-Arquero M, Collin P, Maki M, Sanz A, Calvo C, Maluenda C, Westman P, de la Concha EG,*

*Partanen J (1998) HLA-DQ2-negative celiac disease in Finland and Spain. Hum Immunol 59:169–175. doi:10.1016/ S0198-8859(98) 00008-1*

*Romanos J, Rybak A, Wijmenga C, Wapenaar MC (2008) Molecualr Diagnosis of celiac disease: are we there yet? Expert Opin Med Diagn 2:399–416. doi:10.1517/17530059.2.4.399*

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

*Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J Exp Med 169:345–350. doi:10.1084/jem.169.1.345*

*Spurkland A, Sollid LM, Polanco I, Vartdal F, Thorsby E (1992) HLA-DR and -DQ genotypes of celiac disease patients serolog- ically typed to be non-DR3 or non-DR5/7. Hum Immunol 35:188–192. doi:10.1016/0198-8859(92)90104-U*

*Szeles G, Voko Z, Jenei T, Kardos L, Pocsai Z, Bajtay A, Papp E, Pasti G, Kosa Z, Molnar I, Lun K, Adany R (2005) A preliminary evaluation of a health monitoring programme in Hungary. Eur J Public Health 15:26–32. doi:10.1093/ eurpub/ cki107*

*Vader W, Stepniak D, Kooy Y, Mearin L, Thompson A, van Rood JJ, Spaenij L, Koning F (2003) TheHLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten specific T cell responses. Proc Natl Acad Sci U S A 100:12390–12395. doi:10.1073/pnas.2135229100*

## Supplementary Data

**Supplementary table 1**. Results of the quality check in the Hungarian (HUN), Finnish (FIN), and Italian (ITA), case-control sample sets.

| HUN | | CASES | | | CONTROLS | | |
|---|---|---|---|---|---|---|---|
| | MARKER | MAF | CALLRATE | HWE | MAF | CALLRATE | HWE |
| DQ2.5 | RS2187668 | 0.491573 | 0.988889 | 0.000000 | 0.107955 | 0.977778 | 0.127148 |
| DQ2.2 | RS2395182 | 0.084000 | 0.988889 | 0.526800 | 0.186000 | 0.972222 | 0.473900 |
| DQ7 | RS4639334 | 0.136364 | 0.977778 | 0.145763 | 0.218750 | 0.977778 | 0.798729 |
| DQ2.2 | RS4713586 | 0.005952 | 0.933333 | 0.938135 | 0.024691 | 0.900000 | 0.747282 |
| DQ8 | RS7454108 | 0.028249 | 0.983333 | 0.698943 | 0.099432 | 0.977778 | 0.288871 |
| DQ2.2 | RS7775228 | 0.213483 | 0.988889 | 0.000293 | 0.122159 | 0.977778 | 0.334290 |
| **FIN** | | **CASES** | | | **CONTROLS** | | |
| | MARKER | MAF | CALLRATE | HWE | MAF | CALLRATE | HWE |
| DQ2.5 | RS2187668 | 0.286111 | 0.997230 | 0.000510 | 0.088068 | 0.994350 | 0.731878 |
| DQ2.2 | RS2395182 | 0.201950 | 0.994460 | 0.906508 | 0.278409 | 0.994350 | 0.610393 |
| DQ7 | RS4639334 | 0.053161 | 0.963989 | 0.278996 | 0.072254 | 0.977401 | 0.213500 |
| DQ2.2 | RS4713586 | 0.058917 | 0.869806 | 0.927040 | 0.094937 | 0.892655 | 0.694598 |
| DQ8 | RS7454108 | 0.095238 | 0.988920 | 0.446981 | 0.110795 | 0.994350 | 0.902259 |
| DQ2.2 | RS7775228 | 0.112500 | 1.000000 | 0.814796 | 0.153409 | 1.000000 | 0.934296 |
| **ITA-Trieste** | | **CASES** | | | **CONTROLS** | | |
| | MARKER | MAF | CALLRATE | HWE | MAF | CALLRATE | HWE |
| DQ2.5 | RS2187668 | 0.444444 | 0.971223 | 0.020137 | 0.164103 | 0.989848 | 0.695893 |
| DQ2.2 | RS2395182 | 0.068841 | 0.992806 | 0.073860 | 0.131980 | 1.000000 | 0.723657 |
| DQ7 | RS4639334 | 0.170370 | 0.971223 | 0.242722 | 0.286802 | 0.994949 | 0.094836 |
| DQ2.2 | RS4713586 | 0.003968 | 0.906475 | 0.964330 | 0.020408 | 0.994924 | 0.770542 |
| DQ8 | RS7454108 | 0.068841 | 0.992806 | 0.073860 | 0.053571 | 0.994924 | 0.537662 |
| DQ2.2 | RS7775228 | 0.172662 | 1.000000 | 0.013875 | 0.139594 | 1.000000 | 0.923835 |
| **ITA-Milan** | | **CASES** | | | **CONTROLS** | | |
| | MARKER | MAF | CALLRATE | HWE | MAF | CALLRATE | HWE |
| DQ2.5 | RS2187668 | 0.372007 | 1.000000 | 0.000215 | 0.082770 | 1.000000 | 0.975926 |
| DQ2.2 | RS2395182 | 0.065498 | 0.065498 | 0.102736 | 0.142736 | 1.000000 | 0.983591 |
| DQ7 | RS4639334 | 0.181400 | 0.998158 | 0.000000 | 0.277015 | 0.984797 | 0.235398 |
| DQ2.2 | RS4713586 | 0.009208 | 1.000000 | 0.828548 | 0.031732 | 0.984797 | 0.577782 |
| DQ8 | RS7454108 | 0.046041 | 1.000000 | 0.070823 | 0.055743 | 1.000000 | 0.900284 |
| DQ2.2 | RS7775228 | 0.285582 | 0.996317 | 0.000000 | 0.148649 | 1.000000 | 0.499128 |

MAF minor allele frequency; HWE Hardy-Weinberg equilibrium p-value

**Supplementary table 2.** Detailed sensitivities, specificities, positive predictive values, correlations ($r^2$) and false results in the Finnish, Hungarian and Italian materials.

FIN = Finnish population, HUN = Hungarian population, ITA = Italian population from Trieste region, ALL = all three populations combined; PPV = positive predictive value

| | | DQ2.2 FIN DR;DQ typed | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 26 | 0 | 26 |
| | - | 0 | 328 | 328 |
| Σ | | 26 | 328 | 354 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ2.5 FIN DR;DQ typed | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 198 | 2 | 200 |
| | - | 4 | 184 | 188 |
| Σ | | 202 | 186 | 388 |

Sensitivity 0.98
Specificity 0.99
PPV 0.99
r-squared 0.9391
false results 0.015

| | | DQ7 FIN DR;DQ typed | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 15 | 0 | 12 |
| | - | 0 | 407 | 407 |
| Σ | | 15 | 407 | 422 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ8 FIN DR;DQ & DELFIA | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 39 | 1 | 40 |
| | - | 2 | 694 | 696 |
| Σ | | 41 | 695 | 736 |

Sensitivity 0.95
Speci ficity 0.99856
PPV 0.975
r-squared 0.925
false results 0.0041

| | | DQ2.2 HUN | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 27 | 0 | 27 |
| | - | 0 | 125 | 125 |
| Σ | | 27 | 125 | 152 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ2.5 HUN | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 78 | 0 | 78 |
| | - | 0 | 80 | 80 |
| Σ | | 78 | 80 | 158 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ7 HUN | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 19 | 1 | 20 |
| | - | 0 | 138 | 138 |
| Σ | | 19 | 139 | 158 |

Sensitivity 1
Specificity 0.99
PPV 0.95
r-squared 0.9432
false results 0.006

| | | DQ8 HUN | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 7 | 1 | 8 |
| | - | 0 | 150 | 50 |
| Σ | | 7 | 151 | 158 |

Sensitivity 1
Specificity 0.99
PPV 0.875
r-squared 0.8692
false results 0.006

| | | DQ2.2 ITA | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 21 | 0 | 21 |
| | - | 0 | 145 | 145 |
| Σ | | 22 | 145 | 166 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ2.2 ALL | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 74 | 0 | 74 |
| | - | 0 | 598 | 598 |
| Σ | | 74 | 598 | 672 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ2.5 ITA | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 97 | 0 | 97 |
| | - | 0 | 89 | 89 |
| Σ | | 97 | 89 | 186 |

Sensitivity 1
Specificity 1
PPV 1
r-squared 1
false results 0

| | | DQ2.5 A LL | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 373 | 2 | 375 |
| | - | 4 | 353 | 357 |
| Σ | | 377 | 355 | 732 |

Sensitivity 0.98939
Specificity 0.994366
PPV 0.994667
r-squared 0.9675
false results 0.008197

| | | DQ7 ITA | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 29 | 1 | 30 |
| | - | 0 | 156 | 156 |
| Σ | | 29 | 157 | 186 |

Sensitivity 1
Specificity 0.99
PPV 0.97
r-squared 0.9615
false results 0.0005

| | | DQ7 ALL | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 63 | 2 | 65 |
| | - | 0 | 701 | 701 |
| Σ | | 63 | 703 | 766 |

Sensitivity 1
Specificity 0.997155
PPV 0.969231
r-squared 0.9665
false results 0.002611

| | | DQ8 ITA | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 10 | 2 | 12 |
| | - | 0 | 180 | 180 |
| Σ | | 10 | 182 | 192 |

Sensitivity 1
Specificity 0.99
PPV 0.83
r-squared 0.8242
false results 0.01

| | | DQ8 ALL | | |
|---|---|---|---|---|
| | | + | - | Σ |
| SNP | + | 56 | 4 | 60 |
| | - | 2 | 1024 | 1026 |
| Σ | | 58 | 1028 | 1086 |

Sensitivity 0.965517
Specificity 0.996109
PPV 0.933333
r-squared 0.8142
false results 0.005525

**Suplementary table 3. Results of the logistics regression analysis to assess the celiac diseae risk conferred by different HLA genotypes.**

| | | | 95% CI | | |
|---|---|---|---|---|---|
| Population | Genotype | OR | Lower | Upper | p-value |
| Finland | DQ2.5/DQ2.5 | NA | NA | NA | **3.03E-08** |
| | DQ2.5/DQ2.2 | NA | NA | NA | 9.98E-01 |
| | DQ2.5/DQ8 | 329.33 | 31.78 | 3413.33 | **1.18E-06** |
| | DQ2.5/DQ7 | 494 | 41.72 | 5848.96 | **8.71E-07** |
| | DQ2.2/DQ7 | 380 | 20.58 | 7016.00 | **6.53E-05** |
| | DQ2.5/DQX | 490.83 | 65.17 | 3696.81 | **1.80E-09** |
| | DQ2.2/DQ2.2 | NA | NA | NA | 9.99E-01 |
| | DQ8/DQ2.2 | 76 | 5.99 | 963.92 | **8.34E-04** |
| | DQ8/DQ8 | 76 | 4.71 | 1226.16 | 2.27E-03 |
| | DQ8/DQ7 | 76 | 2.53 | 2282.27 | 1.26E-02 |
| | DQ8/DQX | 32.57 | 4.05 | 262.16 | 1.06E-03 |
| | DQ2.2/DQX | 17.88 | 1.88 | 170.26 | 1.21E-02 |
| | DQ7/DQ7 | 4.7E-08 | 0.00 | . | 1.00E+00 |
| | DQ7/DQX | 9.5 | 0.81 | 111.22 | 7.29E-02 |
| Hungary | DQ2.5/DQ2.5 | NA | NA | NA | 9.97E-01 |
| | DQ2.5/DQ2.2 | NA | NA | NA | 9.97E-01 |
| | DQ2.5/DQ8 | NA | NA | NA | 9.97E-01 |
| | DQ2.5/DQ7 | NA | NA | NA | 9.97E-01 |
| | DQ2.2/DQ7 | NA | NA | NA | 9.97E-01 |
| | DQ2.5/DQX | NA | NA | NA | 9.97E-01 |
| | DQ2.2/DQ2.2 | NA | NA | NA | 9.97E-01 |
| | DQ8/DQ2.2 | 1.00 | 0 | . | 1.00E+00 |
| | DQ8/DQ8 | NA | NA | NA | 9.97E-01 |
| | DQ8/DQ7 | 1.00 | 0 | . | 1.00E+00 |
| | DQ8/DQX | 1.00 | 0 | . | 1.00E+00 |
| | DQ2.2/DQX | 1.00 | 0 | . | 1.00E+00 |
| | DQ7/DQ7 | NA | NA | NA | 9.98E-01 |
| | DQ7/DQX | NA | NA | NA | |
| Italy-Milan | DQ2.5/DQ2.5 | 246.86 | 69.40 | 878.12 | **1.76E-17** |
| | DQ2.5/DQ2.2 | 137.68 | 55.22 | 343.30 | **4.33E-26** |
| | DQ2.5/DQ8 | 26.41 | 8.44 | 82.64 | **1.85E-08** |
| | DQ2.5/DQ7 | 39.62 | 16.08 | 97.59 | **1.25E-15** |
| | DQ2.2/DQ7 | 64.25 | 27.63 | 149.40 | **4.12E-22** |
| | DQ2.5/DQX | 44.80 | 19.24 | 104.29 | **1.15E-18** |
| | DQ2.2/DQ2.2 | 27.43 | 6.27 | 120.04 | **1.10E-05** |
| | DQ8/DQ2.2 | 50.29 | 12.72 | 198.72 | **2.30E-08** |
| | DQ8/DQ8 | 27.43 | 3.92 | 191.68 | **8.43E-04** |
| | DQ8/DQ7 | 4.57 | 1.32 | 15.81 | 1.63E-02 |
| | DQ8/DQX | 8.83 | 3.27 | 23.82 | **1.71E-05** |
| | DQ2.2/DQX | 2.74 | 0.98 | 7.72 | 5.59E-02 |
| | DQ7/DQ7 | 0.00 | 0.00 | . | 9.98E-01 |
| | DQ7/DQX | 0.78 | 0.27 | 2.27 | 6.43E-01 |
| Italy-Trieste | DQ2.5/DQ2.5 | 103.33 | 11.56 | 923.64 | **3.32E-05** |
| | DQ2.5/DQ2.2 | 108.50 | 12.16 | 967.80 | **2.70E-05** |
| | DQ2.5/DQ8 | 72.33 | 6.51 | 803.11 | **4.91E-04** |
| | DQ2.5/DQ7 | 40.69 | 5.01 | 330.55 | **5.26E-04** |
| | DQ2.2/DQ7 | 41.33 | 5.06 | 337.82 | **5.16E-04** |
| | DQ2.5/DQX | 35.59 | 4.55 | 278.44 | **6.65E-04** |
| | DQ2.2/DQ2.2 | 1.92E-08 | 0 | . | 9.99E-01 |
| | DQ8/DQ2.2 | 31 | 1.90 | 506.77 | 1.60E-02 |
| | DQ8/DQ8 | 31 | 1.02 | 941.00 | 4.86E-02 |
| | DQ8/DQ7 | 10.33 | 0.80 | 132.96 | 7.32E-02 |
| | DQ8/DQX | 17.22 | 1.78 | 166.98 | 1.41E-02 |
| | DQ2.2/DQX | 3.26 | 0.28 | 38.48 | 3.48E-01 |
| | DQ7/DQ7 | 1.92E-08 | 0 | . | 9.98E-01 |
| | DQ7/DQX | 1.92E-08 | 0 | . | 9.98E-01 |

OR odds ratio; CI confidence interval

# Six new celiac disease loci replicated in an Italian population confirm association with celiac disease

Jihane Romanos [1,*], Donatella Barisani [2,*], Gosia Trynka [1], Alexandra Zhernakova [3], Maria Teresa Bardella [4,5], Cisca Wijmenga [1].

[1] Department of Genetics, University Medical Center Groningen, University of Groningen, the Netherlands; [2] Department of Experimental Medicine, University of Milano Bicocca, Italy; [3]Complex Genetics Section, Department of Medical Genetics, University Medical Center Utrecht, the Netherlands;[4] Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy;[5] Department of Medical Sciences, University of Milan, Italy.

* These authors contributed equally to this work.

*J Med Genet. 2009; 46(1):60-3.*

# CHAPTER 3

# Abstract

**Background and aims:** The first genome wide association study on coeliac disease (CD) and its follow- up have identified eight new loci that contribute significantly towards CD risk. Seven of these loci contain genes controlling adaptive immune responses, including *IL2/IL21* (4q27), *RGS1* (1q31), *IL18RAP* (2q11–2q12), *CCR3* (3p21), *IL12A* (3q25–3q26), *TAGAP* (6q25) and *SH2B3* (12q24). **Methods:** We selected the nine most associated single nucleotide polymorphisms to tag the eight new loci in an Italian cohort comprising 538 CD patients and 593 healthy controls. **Results:** Common variation in IL2/IL21, RGS1, IL12A/ SCHIP and SH2B3 was associated with susceptibility to CD in our Italian cohort. The LPP and TAGAP regions also showed moderate association, whereas there was no association with CCR3 and IL18RAP. **Conclusion:** This is the first replication study of six of the eight new CD loci; it is also the first CD association study in a southern European cohort. Our results may imply there is a genuine population difference across Europe regarding the loci contributing to CD.

**Keywords:** coeliac disease, replication, association study, meta-analysis.

# Introduction

Coeliac disease (CD) is a chronic disorder of the small intestine, resulting from an aberrant cellular response to gluten peptides; it affects as much as 1% of the European population. The only treatment is a lifelong gluten-free diet. In the past decade, tremendous progress has been achieved in unravelling the genetic aetiology of CD. Twin and family based studies clearly show a strong genetic component to CD development.[1] The clearly identified genetic risk factors for this disease are the HLA-DQ2 and HLA-DQ8 molecules. These are estimated to explain ~40% of the heritability of CD.[2] The other 60% of the genetic susceptibility to CD is shared between an unknown number of non-HLA genes, each of which is estimated to contribute only a small risk effect. Linkage screens and candidate gene studies have led to the discovery of several susceptibility loci and genes, such as the CELIAC2 locus (5q), *MYO9B* and *CTLA4*.[3–5]

The first genome wide association study (GWAS) in CD was recently performed in 778 CD cases and 1422 population controls from the UK.[6] The only locus other than the HLA region showing genome wide significance was 4q27, a ~ 500 kb block of linkage disequilibrium (LD) containing the *IL2* and *IL21* genes. We established independent replication of single nucleotide polymorphisms (SNPs) from the IL2/IL21 region in both Dutch and Irish cohorts of coeliac patients and healthy controls. Moreover, the same region was found to be associated with type 1 diabetes and rheumatoid arthritis, suggesting it is a common autoimmune locus.[7] Both IL2 and IL21 molecules are widely expressed cytokines important for T cell maturation and proliferation, and they are therefore attractive candidates for CD pathogenesis.

In a more extensive follow-up of 1020 top GWAS associated single nucleotide polymorphisms (SNPs) in several independent cohorts from the UK, Dutch and Irish populations, Hunt et al identified seven new risk regions that meet a genome wide significance threshold in 7238 samples (p value overall ,561027).[8] Six of these new CD loci contain genes controlling adaptive immune responses, including *RGS1* (1q31), *IL18RAP* (2q11–2q12), *CCR3* (3p21), *IL12A* (3q25– 3q26), *TAGAP* (6q25) and *SH2B3* (12q24). The seventh associated locus is located on 3q28 and harbours the *LPP* gene, which might play a role in maintaining cell adhesion and motility. Three of these loci have also been associated with other inflammatory and autoimmune disorders: the CCR3 and SH2B3 loci with type 1 diabetes and the IL18RAP locus with Crohn's disease.[9 10]

We set out to replicate the associations found with the seven new loci and the IL2/IL21 locus to CD in an Italian CD cohort.

# METHODS



## Subjects and controls

DNA isolated from whole blood was available from 538 patients diagnosed by a referral centre for CD (Centro per la prevenzione e diagnosi della malattia celiaca, Fondazione IRCCS Ospedale Maggiore Policlinico) and from 593 healthy controls from the north of Italy. The average age of onset was 24.7 years (range 1–78 years). All the affected individuals were diagnosed according to the revised ESPGHAN criteria showing a Marsh III lesion.[11] In addition, patients' serum samples tested positive for both anti-transglutaminase and anti-endomysium antibodies. Only 1.3% of the affected individuals had no HLA-DQ2 and/or HLA-DQ8 risk alleles, which is in accordance with published data.[12] Written informed consent was obtained from all individuals before enrolment in the study. The study was approved by the ethics committee of the Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy.

## Genotyping

In order to tag the eight loci, we selected the nine most associated SNPs reported by Hunt et al.[8] For the IL2/Il21 locus, the most associated SNP rs13119723 was discarded as this SNP showed bad clustering. Therefore, we selected the second most associated SNP rs6822844 to tag the IL2/IL21 locus. For the IL12A/SCHIP locus, we genotyped two SNPs, rs17810546 and rs9811792, since they were reported to be independently associated. SNPs were genotyped using TaqMan probes and primers, using assays developed by Applied Biosystems, and an ABI 7900HT system (Applied Biosystems, Nieuwerkerk a/d IJssel, the Netherlands). Genotyping was performed following the manufacturer's specifications. DNA samples were processed in 384 well plates and each plate with patients' and control DNA contained eight negative controls and 16 genotyping controls (four duplicates of four different samples obtained from the Centre d'Etude du Polymorphisme Humain (CEPH)). There was no discordance in the genotypes of any of the CEPH samples. Laboratory staff were blind to the disease status of each sample.

## Statistical analysis

The genotype frequencies were tested for Hardy–Weinberg equilibrium (HWE) with a value of $p<0.05$ considered as not being in HWE. Allele frequencies were determined in patients and controls. Difference in allele distribution between patients and controls and association analysis were performed using two tailed x2 analysis while meta-analysis of our Italian cohort, the UK GWAS of van Heel et al and the Irish, Dutch and UK2 cohorts of Hunt et al was performed using the Maentel–Haenszel method.[6 8] Odds ratios (OR) and confidence intervals (CI) were calculated using Woolf's method with Haldane's correction. Power calculations

were performed using the genetic power calculator (http://pngu.mgh.harvard.edu/,purcell/gpc/) assuming allele frequencies of 0.2 and 0.3.

# RESULTS

Replicating genetic findings in several populations is an important step in establishing a genetic effect on disease predisposition. We therefore genotyped nine associated SNPs tagging the eight CD susceptibility loci identified by Van Heel et al and Hunt et al.[6 8] Our study had ~ 80% power to detect an odds ratio of 1.45, while it had ~28% power to detect an odds ratio of 1.2. We observed association for six of the loci in our Italian cohort. Table 1 summarises the genotyping results and case–control association analysis at the single SNP level. All SNPs were in HWE in our control population (data not shown).

The first GWAS in CD in a UK cohort identified 4q27 region as a susceptibility region for CD.[6] We genotyped SNP rs6822844 which was the most associated SNP identified by meta-analysis in the first GWAS and the second most associated one reported in the follow-up.[6 8] We saw a decrease in frequency of the rs6822844*A allele in Italian cases (8.1%) compared to controls (10.9%); this association was significant (p = 0.025) and in the same direction as described earlier (OR 0.72, 95% CI 0.54 to 0.96).

In our cohort, the most associated locus was located on chromosome 3q25–3q26. Two SNPs were tested in this block due to independent

**Table 1.** Genotyping results and case-control association analyses at the single SNP level.

| | dbSNP | Chr | alleles | MAF | MAF in cases | MAF in controls | allelic p-value | OR | 95%CI |
|---|---|---|---|---|---|---|---|---|---|
| RGS1 | rs2816316 | 1 | A:C | C | 0.128 | 0.166 | **0.0123** | 0.74 | 0.58-0.94 |
| IL18RAP | rs917997 | 2 | G:A | A | 0.244 | 0.227 | 0.3498 | 1.1 | 0.90-1.34 |
| CCR3 | rs6441961 | 3 | G:A | A | 0.372 | 0.368 | 0.8277 | 1.02 | 0.85-1.21 |
| IL12A,SCHIP1 | rs17810546 | 3 | A:G | G | 0.109 | 0.067 | **0.0004** | 1.71 | 1.26-2.31 |
| IL12A | rs9811792 | 3 | A:G | G | 0.467 | 0.421 | **0.0313** | 1.21 | 1.02-1.43 |
| LPP | rs1464510 | 3 | A:C | A | 0.519 | 0.474 | **0.0348** | 1.2 | 1.20-1.42 |
| IL2/IL21 | rs6822844 | 4 | A:C | A | 0.081 | 0.109 | **0.0253** | 0.72 | 0.54-0.96 |
| TAGAP | rs1738074 | 6 | A:G | A | 0.454 | 0.412 | **0.0495** | 1.19 | 1.00-1.41 |
| SH2B3 | rs3184504 | 12 | G:A | A | 0.554 | 0.494 | **0.005** | 1.27 | 1.07-1.50 |

Chr chromosome; MAF minor allele frequency; OR odds ratio; CI confidence interval; allelic p-value was calculated by 2x2 two-sided c-square test; bold indicates p-value < 0.05.

**Table 2**. Differences among five celiac disease populations.

| Population (number of cases) | | | | UK GWAS (778) | | UK2 (719) | | Irish (416) | |
|---|---|---|---|---|---|---|---|---|---|
| | SNP | CHR | minor allele | MAF cases/controls | p-value | MAF cases/controls | p-value | MAF cases/controls | p-value |
| RGS1 | rs2816316 | 1 | C | 0.136/0.174 | **1.16E-03** | 0.129/0.188 | **1.05E-06** | 0.161/0.187 | 0.103 |
| IL18RAP | rs917997 | 2 | A | 0.267/0.215 | **9.06E-05** | 0.245/0.208 | **4.78E-03** | 0.214/0.204 | 0.549 |
| CCR3 | rs6441961 | 3 | A | 0.341/0.297 | **2.84E-03** | 0.342/0.302 | **6.31E-03** | 0.343/0.318 | 0.201 |
| IL12A, SCHIP1 | rs17810546 | 3 | G | 0.162/0.123 | **3.39E-04** | 0.146/0.127 | 0.076 | 0.184/0.13 | **2.47E-04** |
| IL12A | rs9811792 | 3 | G | 0.486/0.439 | **2.78E-03** | 0.483/0.44 | **7.44E-03** | 0.493/0.446 | **0.023** |
| LPP | rs1464510 | 3 | A | 0.52/0.457 | **7.65E-05** | 0.516/0.446 | **1.13E-05** | 0.483/0.448 | 0.086 |
| IL2/IL21 | rs6822844 | 4 | A | 0.126/0.179 | **4.80E-06** | 0.138/0.175 | **1.73E-03** | 0.15/0.199 | **2.43E-03** |
| TAGAP | rs1738074 | 6 | A | 0.472/0.422 | **1.56E-03** | 0.459/0.428 | **0.049** | 0.519/0.468 | **0.014** |
| SH2B3 | rs3184504 | 12 | A | 0.541/0.489 | **1.08E-03** | 0.522/0.472 | **1.54E-03** | 0.505/0.477 | 0.172 |

| Population (number of cases) | | | | Dutch (508) | | Italian (538) | | Meta-analysis | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SNP | CHR | minor allele | MAF cases/controls | p-value | MAF cases/controls | p-value | p-value | OR | 95% CI |
| RGS1 | rs2816316 | 1 | C | 0.141/0.189 | **1.09E-03** | 0.128/0.166 | **0.012** | **1.05E-12** | 0.72 | 0.66-0.79 |
| IL18RAP | rs917997 | 2 | A | 0.293/0.215 | **3.61E-06** | 0.244/0.227 | 0.35 | **1.50E-09** | 1.26 | 1.17-1.36 |
| CCR3 | rs6441961 | 3 | A | 0.38/0.322 | **1.95E-03** | 0.372/0.368 | 0.828 | **1.52E-06** | 1.18 | 1.10-1.26 |
| IL12A, SCHIP1 | rs17810546 | 3 | G | 0.171/0.124 | **6.37E-04** | 0.109/0.067 | **4.23E-04** | **4.95E-12** | 1.39 | 1.26-1.52 |
| IL12A | rs9811792 | 3 | G | 0.497/0.439 | **3.17E-03** | 0.467/0.421 | **0.031** | **4.80E-09** | 1.21 | 1.14-1.29 |
| LPP | rs1464510 | 3 | A | 0.521/0.5 | 0.293 | 0.519/0.474 | **0.035** | **5.58E-10** | 1.22 | 1.15-1.31 |
| IL2/IL21 | rs6822844 | 4 | A | 0.13/0.186 | **1.14E-04** | 0.081/0.109 | **0.025** | **2.35E-14** | 0.7 | 0.64-0.77 |
| TAGAP | rs1738074 | 6 | A | 0.46/0.395 | **8.18E-04** | 0.454/0.412 | **0.049** | **9.39E-09** | 1.21 | 1.13-1.29 |
| SH2B3 | rs3184504 | 12 | A | 0.528/0.479 | **0.013** | 0.554/0.494 | **5.04E-03** | **2.76E-09** | 1.21 | 1.14-1.30 |

CHR chromosome; MAF minor allele frequency; OR odds ratio; CI confidence interval; bold indicates p-value < 0.05.

**Key points**

- Eight new coeliac loci were identified by a genome wide association study.
- Six of the eight coeliac loci are also associated in an Italian population.
- There is genetic heterogeneity among populations.

association reported by Hunt et al; the rs17810546 showed convincing association in the Italian samples with the same allele as reported in the GWAS follow- up study (p allele = 4.23E-04; OR 1.71, 95% CI 1.26 to 2.31).[8] The second SNP in the same block, rs9811792, showed a moderate association (p = 0.031; OR 1.21, 95% CI 1.02 to 1.43). These two SNPs may also represent an independent association signal (D' = 0.97; r2 = 0.113) in our cohort. This region harbours two potentially interesting genes, *IL12A* (interleukin-12A) and *SCHIP1* (schwannomin interacting protein 1). The second most associated SNP rs3184504 (p = 5.04E-03; OR 1.27, 95% CI 1.07 to 1.50) mapped on chromosome 12q24, in the vicinity of *SH2B3* and *ATNX2* genes.

SNP rs2816316 was the most significant SNP outside the HLA and IL2/IL21 loci identified by Hunt et al.[8] It is located on chromosome 1q31, in a ~70 kb LD block containing the *RGS1* gene (regulator of G-protein signalling 1). We also found association for this SNP in our cohort (p = 0.012; OR 0.74, 95% CI 0.58 to 0.94). Moderate association (p = 0.035; OR 1.2, 95% CI 1.20 to 1.42), consistent with previous findings, was found for SNP rs1464510, located on 3q28, in a ,70 kb LD block harbouring the *LPP* gene. SNP rs1738074, located on chromosome 6q25, showed a trend towards association with CD in our cohort (p = 0.05). This SNP is a ~200 kb LD block containing *TAGAP* (T cell activation GTPase activating protein). This 6q25 region was also found to be linked to CD in a large Dutch family.[13] All associations were observed with the same allele as in the original study.[8]

We saw no association for two SNPs (rs917997 and rs6441961) which were located on chromosome 2q11–2q12 (IL18RAP locus) and 3p21 (CCR3 locus), respectively (table 1).

# DISCUSSION

In the last decade, our understanding of CD pathogenesis was mainly based on the binding of HLA-DQ2/DQ8 to gluten peptides, the role of tissue transglutaminase, and the identification of immunological dominant T cell epitopes. However, advances in genetics now allow us to identify novel genes involved in the susceptibility to CD, thereby helping us to understand the pathogenesis of this disorder. We have performed the first replication

of eight new loci identified in the first GWAS performed in CD.[6 8] We found a positive association for six of these loci: IL2/IL21, RGS1, IL12A/SCHIP, LPP, TAGAP and SH2B3 (table 1). These loci harbour candidate genes involved mainly in the Th1 pathway: *IL2* and *IL21* are important in T cell activation, while *RGS1* regulates chemokine receptors' signalling and is involved in B cell activation and proliferation. Not much is known about the *SCHIP1* gene but the *IL12A* gene, located in the same LD block, encodes the IL12p35 of IL12 subunit, which is important for T cells and natural killer cells, both of which are involved in the Th1 pathway. *LPP* shows a very high expression in the small intestine and may play a structural role in maintaining cell shape and motility at sites of cell adhesion. The *TAGAP* gene is interesting since it is expressed in activated T cells and has a Rho-GAP domain similar to *MYO9B*, another CD associated gene.[5] The *SH2B3* gene is a good candidate for CD since it is expressed in the small intestine, mainly in monocytes and dendritic cells and to a lesser extent in resting B, T and natural killer cells. Moreover, the SNP that we have tested could be a causal variant since it is a non-synonymous SNP leading to an amino acid change R262W in an important domain of the protein. Fine mapping and deep sequencing of these regions, and functional studies, are needed to identify the true causal genes and their role in the disease process.

We were unable to detect association between the CCR3 locus or the IL18RAP locus and CD in our Italian cohort. This might be due to clinical heterogeneity although this seems highly unlikely given that our Italian cohort is a mixture of adult and paediatric patients. It might also be due to the power being too low to detect an odds ratio of 1.2, or even genetic heterogeneity among populations. For the SNP in the *CCR3* gene, there was no significant difference between the frequencies of the minor allele in patients compared to controls (37.2% and 36.8%, respectively). We looked at the separate results of the four populations analysed for the GWAS and noticed that this region was only associated in the UK GWAS, UK2 and Dutch samples, whereas in the Irish cohort, the frequency between patients and controls was not significantly different (table 2). Similarly, the IL18RAP locus was not associated in the Irish cohort.

These results support the hypothesis of risk genes in complex disorders being population specific. Genetic heterogeneity is recognised for the CD associated HLA risk alleles within Europe. In southern Europe, individuals carrying DQ2.5 in trans are more prevalent than in the north, where individuals carry DQ2.5 in cis more frequently.[14] In addition, risk allele HLA-DQ8 is more frequent in southern Europe and accounts for 6–10% of the CD patients there.[15] Differences within European populations due to regional founder effects were also suggested for the inflammatory bowel disease associated genes, *NOD2* and *DLG5*,[16 17] while allele frequencies for NOD2 risk alleles were reported to vary significantly between European populations.[18] Another reason for discrepancies among populations could be the complex interaction between marker allele frequencies and founder mutations. Since the linkage disequilibrium varies between distinct populations, the causative variant could be in less LD with the tested SNP. Fine mapping with a dense SNP set is necessary to exclude these

genes as disease causing variants in the Italian patients.

In conclusion, our study confirms the association of six of the eight new loci with CD and may point to heterogeneity among European populations.



**Acknowledgements:** We thank all the patients and controls who participated in this study. We thank Agata Szperl, Eleonora AM Festen and Cleo van Diemen for their help in the laboratory and Jackie Senior for critically reading the manuscript. Funding: The study was supported by grants from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW) and KP6 EU grant 036383 (PREVENTCD).

**Competing interests**: None. Patient consent: Obtained.

## REFERENCES




1. *Greco L, Romino R, Coto I, Di CN, Percopo S, Maglio M, Paparo F, Gasperi V, Limongelli MG, Cotichini R, D'Agate C, Tinto N, Sacchetti L, Tosi R, Stazi MA. The first large population based twin study of coeliac disease. Gut 2002;50:624–8.*
2. *Bevan S, Popat S, Braegger CP, Busch A, O'Donoghue D, Falth-Magnusson K, Ferguson A, Godkin A, Hogberg L, Holmes G, Hosie KB, Howdle PD, Jenkins H, Jewell D, Johnston S, Kennedy NP, Kerr G, Kumar P, Logan RF, Love AH, Marsh M, Mulder CJ, Sjoberg K, Stenhammer L, Walker-Smith J, Marossy AM, Houlston RS. Contribution of the MHC region to the familial risk of coeliac disease. J Med Genet 1999;36:687–90.*
3. *Babron MC, Nilsson S, Adamovic S, Naluai AT, Wahlstrom J, Ascher H, Ciclitira PJ, Sollid LM, Partanen J, Greco L, Clerget-Darpoux F. Meta and pooled analysis of European coeliac disease data. Eur J Hum Genet 2003;11:828–34.*
4. *Djilali-Saiah I, Schmitz J, Harfouch-Hammoud E, Mougenot JF, Bach JF, Caillat- Zucman S. CTLA-4 gene polymorphism is associated with predisposition to coeliac disease. Gut 1998;43:187–9.*
5. *Monsuur AJ, de Bakker PI, Alizadeh BZ, Zhernakova A, Bevova MR, Strengman E, Franke L, van't SR, van Belzen MJ, Lavrijsen IC, Diosdado B, Daly MJ, Mulder CJ, Mearin ML, Meijer JW, Meijer GA, van OE, Wapenaar MC, Koeleman BP, Wijmenga C. Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nat Genet 2005;37:1341–4.*
6. *Van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CJ, McManus R, McGinnis R, Cardon LR, Deloukas P, Wijmenga C. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 2007;39:827–9.*
7. *Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, van der SG, Radstake TR, Barrera P, Roep BO, Koeleman BP, Wijmenga C. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet 2007;81:1284–8.*
8. *Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 2008;40:395–402.*

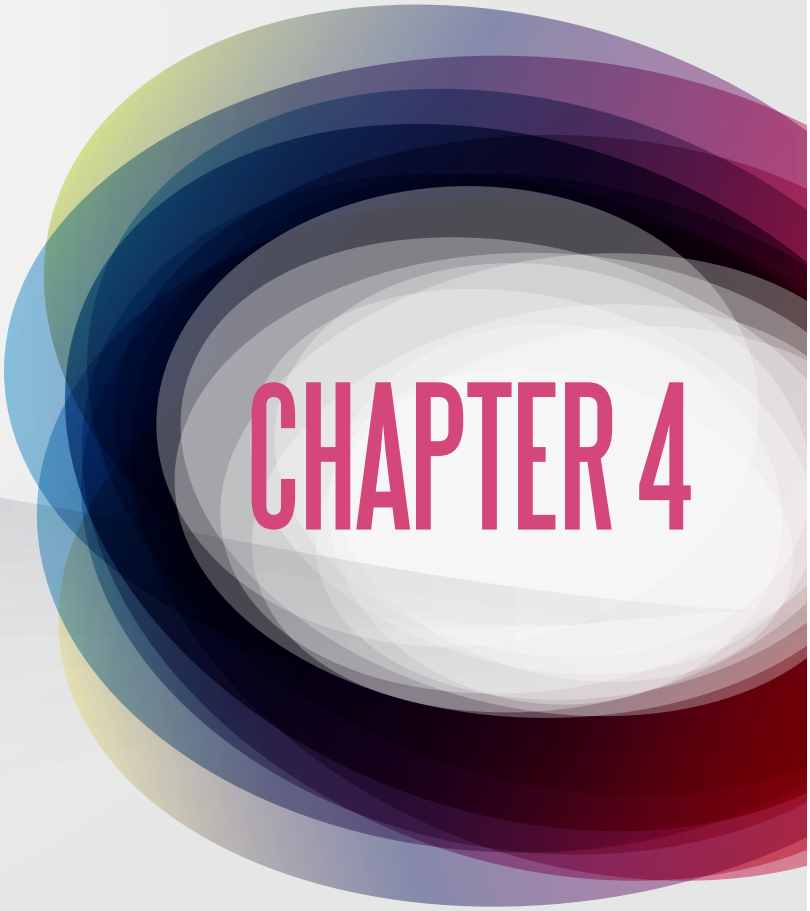
9. *Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgoviste C, Simmonds MJ, Heward JM, Gough SC, Dunger DB, Wicker LS, Clayton DG. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39:857–64.*
10. *Zhernakova A, Festen EM, Franke L, Trynka G, van Diemen CC, Monsuur AJ, Bevova M, Nijmeijer RM, van 't SR, Heijmans R, Boezen HM, van Heel DA, van Bodegraven AA, Stokkers PC, Wijmenga C, Crusius JB, Weersma RK. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. Am J Hum Genet 2008;82:1202–10.*
11. *Working Group of the European Society of Paediatric Gastroenterology and Nutrition. Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. Arch Dis Child 1990;65;909–11.*
12. *Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, Partanen J; European Genetics Cluster on Celiac Disease. 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;4:469–77.*
13. *Van Belzen MJ, Meijer JW, Sandkuijl LA, Bardoel AF, Mulder CJ, Pearson PL, Houwen RH, Wijmenga C. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology 2003;125:1032–41.*
14. *Margaritte-Jeannin P, Babron MC, Bourgey M, Louka AS, Clot F, Percopo S, Coto I, Hugot JP, Ascher H, Sollid LM, Greco L, Clerget-Darpoux F. HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease. Tissue Antigens 2004;63:562–7.*
15. *Torres MI, Lopez Casado MA, Rios A. New aspects in celiac disease. World J Gastroenterol 2007;13:1156–61.*
16. *Arnott ID, Nimmo ER, Drummond HE, Fennell J, Smith BR, MacKinlay E, Morecroft J, Anderson N, Kelleher D, O'Sullivan M, McManus R, Satsangi J. NOD2/CARD15, TLR4 and CD14 mutations in Scottish and Irish Crohn's disease patients: evidence for genetic heterogeneity within Europe? Genes Immun 2004;5:417–25.*
17. *Tenesa A, Noble C, Satsangi J, Dunlop M. Association of DLG5 and inflammatory bowel disease across populations. Eur J Hum Genet 2006;14:259–60.*
18. *Cavanaugh J. NOD2: ethnic and geographic differences. World J Gastroenterol 2006;12:3673–7.*

# Predicting susceptibility to celiac disease by genetic risk profiling

Jihane Romanos & Cisca Wijmenga

Department of Genetics, University Medical Center Groningen, University of Groningen, the Netherlands.

*Annals of Gastroenterology and Hepatology. 2010;1(1):11-18.*



# CHAPTER 4

# Abstract

Celiac disease (CD) is a common, immune-mediated, intestinal disorder with a prevalence of approximately 1% in Caucasians. An important and necessary genetic risk factor is HLA-DQ2. Dietary gluten is the triggering environmental factor, and a lifelong gluten-free diet is currently the only treatment for CD. CD is officially diagnosed by serology followed by a small intestinal biopsy. However, the majority of CD patients go undiagnosed as the symptoms associated with CD can be rather subtle, and many patients remain at a silent or latent stage. As CD is associated with increased morbidity and mortality, it puts a severe socio-economic burden on patients, their families, and society. Improved diagnosis of CD and early intervention would alleviate, or reverse, these negative effects. The recent identification of part of the genetic risk for CD may help in diagnosing individuals at high risk for CD before the disease manifests. In this review, we show how genetic knowledge can be applied as a diagnostic or screening tool to prevent comorbidity and long-term complications. We envision a two-step approach. First, based on human leukocyte antigen (HLA) typing, we can exclude individuals with no HLA-DQ2/DQ8 as they have no risk of developing CD. Second, we can combine the presence of HLA-DQ2/DQ8 with the non-HLA genetic risk factors and classify the remaining individuals into low (,0.1%), intermediate (0.1–7%), and high (.7%) risk groups. Individuals in both the intermediate- and high-risk groups should undergo serology and biopsy testing. Our prediction model for CD will lead to improved diagnostic and prevention strategies.

**Keywords:** celiac disease, genetic profile, predictive test, human leukocyte antigen (HLA), gluten intolerance, gluten sensitive enteropathy, celiac sprue

# Introduction



Genome-wide association studies (GWAS) were the scientific breakthrough of 2007, with large numbers of individuals being genotyped for hundreds of thousands of common genetic variants (single nucleotide polymorphisms or SNPs) [1, 2]. The success of this new technology has led to a wave of discoveries of several disease-associated variants for complex diseases, such as type 2 diabetes, breast cancer, Crohn's disease, ulcerative colitis, celiac disease, and many others [3–9]. These new findings have provided important and novel insights into the diseases' biology, raised hopes of being able to identify high-risk individuals based on their genetic profiles, and promoted new approaches for potential therapy and prevention.

The variants identified by the GWA studies are common in the general population and, individually, most variants have only a small effect on disease risk, with odds ratios between 1.1 and 1.5. In addition, known associations usually account for a rather limited part of the heritability (often ~10%). Yet, several companies, such as deCODE genetics and 23andme, have already begun to use SNPs for predicting the risk of developing complex genetic disorders (http://www.decode. com/; https://www.23andme.com/). It is evident that predictive testing based on a single SNP is of limited value, as complex diseases are caused by many different genetic variants [10, 11]. Therefore, genetic profiles based on combining information from many risk variants are likely to be much better predictors of disease risk, even if the full repertoire of susceptibility alleles is still unknown [12].

Lately, the use of risk scores based on combining informa- tion from different disease-associated SNPs was applied to a number of complex diseases including celiac disease, type 2 diabetes, and Crohn's disease [13–19]. These studies showed that affected individuals carry, on average, more disease risk alleles than control individuals [13, 14, 20]. Using logistic regression, it was also shown that individuals carrying an increasing number of risk alleles have an increased risk of developing the disease under study, consistent with an independent multiplicative model [13, 14, 17, 21]. The area under the receiver operator characteristic curve (AUC) can be used to assess how different risk prediction models compare, irrespective of choosing specific specificity and sensitivity requirements. The AUC of genotype scores showed a slight improvement on top of the clinical risk prediction [15, 17, 18, 22–24], but in most studies, the improvement was not statistically significant [16, 25]. However, risk profiling was seen to be more useful in classifying individuals into high-risk and low-risk groups, particularly in the context of population screening [16, 17, 19, 26]. To discriminate between the power of clinical risk factors and genetic variants, Lyssenko et al [17] investigated predictions for type 2 diabetes with an increasing duration of follow-up. They observed an increase in AUC for the genetic risk model with a longer duration of follow-up, whereas the AUC decreased for the clinical risk model. This suggests that assessing the genetic risk profile is clinically more meaningful the earlier in life it is measured. Therefore,

implementing a risk prediction in a clinical context could potentially be of major economic benefit to general population health [27].

## Celiac disease: an important health problem



One of the diseases for which genetic risk prediction would be extremely important is celiac disease (CD) (OMIM #212750), as this might help in its early diagnosis or allow prevention of disease by early intervention. CD is one of the most common food intolerances in western populations, with population screening revealing a prevalence of approximately 1%. However, around 86% of CD patients currently go undiagnosed as they present with atypical, silent, or latent CD [28–31]. The broad spectrum of symptoms (often subtle or non-specific, e.g., tiredness, diarrhea, feeling ‘‘unwell’’) and the fact that they vary considerably between individuals, and even in a single individual over time, often results in a delayed or missed diagnosis. CD is characterized by a chronic inflammation of the small intestinal mucosa that may result in atrophy of intestinal villi, malabsorption, and a variety of clinical manifestations, which may begin either in childhood or in adult life. The only treatment is a lifelong, gluten-free diet which, in most cases, leads to complete remission of the small intestine symptoms and the disappearance of other symptoms.

The health burden of CD is considerable. Several studies have shown an increased risk of morbidity and mortality in patients with both undiagnosed and untreated CD, as well as in those diagnosed later in life due to associated conditions such as type 1 diabetes [32–35]. The mortality rate in undiagnosed CD was associated with nearly a fourfold increase in the USA [34]. Moreover, a retrospective cohort study revealed a modest increase in death in CD patients, as well as in patients with only intestinal inflammation or latent CD patients [33]. Classical CD is frequently found in conjunction with other autoimmune disorders such as type 1 diabetes (3–7%), autoimmune thyroiditis (5%), autoimmune hepatitis, asthma (24.6%), and systemic lupus erythematosus (2.4%) [36, 37]. The risk of developing another autoimmune disease with CD was shown to be higher in patients with a family history of autoimmune disease and with a diagnosis of CD made in childhood or young adulthood [35]. This risk was reported to diminish by a factor of two upon adopting a gluten-free diet [35].

CD is a clear example of an immune-mediated disease for which early diagnosis followed by dietary treatment can prevent its severe and sometimes life-threatening complications, such as reduced fertility, gut malignancy, and osteoporosis. The disease has a great economic and social impact, especially in reducing the individual’s quality of life [38]. Therefore, primary prevention by inducing oral tolerance of the disorder is also a major focus for researchers nowadays (Box 1).

**Box 1 |** Genetic risk profiling in CD

A. Diagnostic tool

Who: Individuals with classical or atypical CD.

Aim: Dietary treatment to prevent severe and long-term complications.

How: 1. Excluding CD as possibility in cases not carrying HLA-DQ2/DQ8 molecules.

2. Reduction of number of serology re-testing and biopsy.

3. Early diagnosis since early detection may be difficult on a clinical basis.

B. Screening tool

Who: Newborns and families with positive history of CD or other autoimmune disease.

Aim: Early intervention and early treatment for prevention.

How: 1. Excluding CD as possibility in cases not carrying HLA-DQ2/DQ8 molecules.

2. Identifying higher risk individuals for close follow-up and early diagnosis.

3. Early introduction of gluten to induce oral tolerance in at-risk newborns.

4. Classify patients into subgroups using molecular diagnosis for different therapies.

# Diagnosis of celiac disease

CD is an autoimmune disease that can potentially affect many organs and not only, as previously thought, the gastrointestinal tract [39]. The classic clinical symptoms, which are observed in only a minority of patients, are steatorrhea, abdominal distension, edema, malabsorption, and failure to thrive [40]. Silent or asymptomatic CD patients may have no gastrointestinal symptoms at all, or may present with atypical symptoms such as iron-deficient anemia, osteoporosis, dermatitis herpetiformis, infertility, and others [40–42].

A clinically suspected CD patient is initially tested for the presence of specific antibodies (antigliadin, anti-endomysial (EmA) and antihuman tissue transglutaminase (tTG)). The most commonly used tests are to detect IgA-EmA and IgA- tTG antibodies. Both are targeted at tissue transglutaminase autoantigen and have a similar sensitivity (86–100% for EmA, 77–100% for tTG) and specificity (90–100% for EmA, 91– 100% for tTG) [43]. However, these serological markers are not accurate in patients with selective IgA deficiency, patients with only IgG class autoantibodies and normal serum IgA, or patients with positive EmA antibodies and an absence of tTG antibodies [44]. Thus, a combination of tTG, EmA, and total IgA serum level should be tested first, and it should be complemented with IgG-tTG or IgG-EmA testing in the case of IgA deficiency [45]. A serious problem in serological screening is the fluctuation of antibody levels in children. Simell et al [46] showed that a large proportion of antibody- positive children are only transiently antibody positive, proving that the antibodies quite

commonly disappear spontaneously without changes in their gluten exposure.

When one or more antibodies are positive, a small intestinal biopsy is taken. Pathological changes in duodenal biopsy, which is the gold standard for diagnosing CD, range from intraepithelial lymphocytosis with normal villous architecture to total villous atrophy [47]. According to Marsh, individuals presenting with significant villous atrophy are classified as CD Marsh stage III, whereas normal villi but an increased number of intraepithelial lymphocytes are classified as Marsh I or II [42, 48]. However, there are several causes of intraepithelial lymphocytosis without villous atrophy in addition to CD, such as Helicobacter pylori infection and tropical sprue [49]. It is therefore difficult to diagnose CD in the setting of intestinal inflammation if villous atrophy is absent [50]. Moreover, this biopsy procedure is invasive, expensive, and carries a risk of complications.

A third form of CD is latent CD, which is diagnosed in patients with positive autoantibodies and the typical HLA- predisposing genotype (HLA-DQ2 and/or -DQ8; discussed below), but who have a normal, or minimally abnormal, mucosal architecture with an increased number of intrae- pithelial lymphocytes [40]. This form has been linked to unexplained neurological or psychiatric disorders such as cerebellar ataxia, schizophrenia, and autism [51–53].

Earlier estimates of the prevalence of CD used to rely on classical symptomatic cases with confirmed biopsies. However, large-scale antibody screening followed by biopsy confirmation revealed a higher prevalence lying between 1:200 and 1:70 in the USA and most western and Middle Eastern countries [29, 39, 54–56]. The prevalence appears to increase with age, as was shown in a recent study in Finland where a prevalence of 1:47 was reported in randomly selected subjects older than 52 years [57].

## Genetics of celiac disease

In the past decade, substantial resources have been invested in understanding the genetic etiology of CD. Familial aggregation was found in 5–15% of CD patients, while a high concordance rate of 83–86% was observed among monozygotic twin pairs, and 16.7–20% between dizygotic twins [36, 58, 59]. This indicates that genetic components play a major role in the induction and manifestation of CD. HLA-DQ2 and HLA-DQ8 heterodimers are well known to be important genetic risk factors for CD. They are constructed from alfa and beta chains encoded by the HLA-DQA1 and HLA-DQB1 genes respectively. HLA-DQ2 includes the sub- groups HLA-DQ2.5 and HLA-DQ2.2 encoded by DQA1*0501- DQB1*0201 and DQA1*0201-DQB1*0202, respectively, whereas HLA-DQ8 is encoded by DQA1*03-DQB1*0302. These HLA molecules can be encoded in cis (i.e., alfa and beta chains encoded by DQA1 and DQB1 of the same chromosome) or in trans (i.e., alfa and beta chains encoded by DQA1 and DQB1 of different chromosomes). For example, the HLA-DQ2.5 molecule can be formed in trans from the HLA-DQ2.2 haplotype (DQA1*0201-DQB1*0202) and the HLA-DQ7 haplotype (DQA1*0505-

**Table 1.** Classification of individuals in three groups based on HLA genotypes and absolute HLA risk

| Risk category | HLA genotypes | Absolute HLA risk (%) |
|---|---|---|
| Low risk | DQ7/DQ7 | 0.0000 |
| | DQX/DQX | 0.0433 |
| | DQ7/DQX | 0.0470 |
| Intermediate risk | DQ2.2/DQX | 0.1661 |
| | DQ8/DQ7 | 0.2765 |
| | DQ8/DQX | 0.5326 |
| | DQ2.5/DQ8 | 1.5769 |
| | DQ2.2/DQ2.2 | 1.6366 |
| | DQ8/DQ8 | 1.6366 |
| | DQ2.5/DQ7 | 2.2587 |
| | DQ2.5/DQX | 2.6194 |
| | DQ8/DQ2.2 | 2.9600 |
| | DQ2.2/DQ7 | 3.7232 |
| High risk | DQ2.5/DQ2.2 | 7.7079 |
| | DQ2.5/DQ2.5 | 12.8137 |



DQB1*0301). HLA-DQ2.5 was shown to have the strongest association with a predisposition to CD, explained by its high affinity to binding gluten proteins. HLA-DQ2.2 contributes less to the risk of CD, as it differs in one amino acid (a phenylalanine in place of tyrosine), which leads to a lower binding stability [60]. Moreover, the risk of CD is shown to be higher in individuals homozygous for the HLA-DQ2.5 or HLA-DQ2.5/DQ2.2 genotypes compared with those homozygous for HLA- DQ2.2 or heterozygous for HLA-DQ2.5 or DQ2.2 [61, 62]. Around 95–98% of CD patients carry these risk molecules compared with 30–40% of the general population, indicating that other, non-HLA genes also play a role in the pathogen- esis of CD. The contribution of HLA to the development of CD is essential, but it represents less than 40% of the total genetic risk, so the remaining risk is likely to be conferred by non-HLA genes [65–67].

In the last few years, candidate gene approaches, genome- wide linkage studies, and genome-wide association studies (GWAS) have mapped several non-HLA genes in CD. However, a major breakthrough in CD genetics came from the first GWAS on a UK cohort, which initially identified an association with the IL2/IL21 region on chromosome 4q27, a region later found to be associated with type 1 diabetes and rheumatoid arthritis [63, 64]. In follow-up studies, 12 additional CD risk loci were discovered and replicated in several other populations [6, 65–68]. A second GWAS was recently performed on CD cohorts from the UK, the Netherlands, Italy, and Finland (Dubois et al, unpublished). In this GWAS, all the known loci were replicated, and 13 new loci reached genome-wide significance ($<5x10^{-8}$). Thus, at the end of 2009, we have found 26 non-HLA loci

associated with CD, which, together with HLA, can explain around 50% of the disease heritability. Interestingly, many of the susceptibility loci include genes that play a role in T-cell differentiation, immune cell signaling, the innate immune response, and tumor necrosis factor (TNF) signaling. Moreover, several of these genes have also been found to be associated with other immune-related disorders [70].



## Risk model for celiac disease

The genetic risk for CD is an important component of the overall risk for developing CD, in addition to non-genetic factors such as gluten consumption, socio-economic factors, and behavioral risk factors. Predicting the risk for developing CD before clinical manifestation of the disease could be a powerful tool in early diagnosis and preventing the disease by early intervention with a gluten-free diet. In the future, it might even become possible to alleviate the disease by inducing oral tolerance to dietary gluten. Except for HLA, the other known genetic risk variants involved in CD have very modest effect sizes and thus a limited value in predicting an individual's risk for disease development. as the true causal variants have not yet been identified, and most associated SNPs are proxies correlated with the true causal variants, the true effect sizes may well be underestimated. We have recently created a genetic scoring system, which could help to classify individuals into low-, intermediate-, and high-risk groups based on their HLA dose–effect genotypes (Table 1) [13]. CD is unique among the complex diseases, as one of its genetic risk factors (i.e., HLA-DQ2/DQ8) is certainly necessary for the disease to develop but is not sufficient in itself; therefore, individuals with no HLA-DQ2/DQ8 have practically zero risk of developing CD. Hence, the negative predictive value of HLA genotyping is close to 100% (Figure 1). However, among individuals positive for HLA-DQ2/DQ8, only 3% will go on to develop clinically recognized CD [70]. Based on HLA dosage, we could classify individuals with a double dose of HLA-DQ2 as high risk (i.e., DQ2.5/DQ2.5 and DQ2.5/DQ2.2) and the others as intermediate risk.

The non-HLA risk alleles are normally distributed in cases and controls, although CD patients carry, on average, more non-HLA risk alleles than healthy people [13]. Moreover, an increasing number of risk alleles were also shown to be associated with an increased risk for CD [13]. This risk increases sixfold in individuals carrying 13 or more non-HLA risk alleles compared with those carrying five or fewer risk alleles. Based on HLA genotypes only, the AUC was 85.4%, whereas the AUC based on both HLA and non-HLA risks improved to 87.4%. When we considered non-HLA risk alleles in addition to HLA, we observed that individuals with an intermediate HLA risk and who carry 13 or more non-HLA risk alleles had the same odds ratio for developing CD as individuals with a high HLA risk and five or fewer non-HLA risk alleles. Thus, these individuals were reclassified into the high-risk group.

Most studies on genetic risk profiling, including our own, have reported the risk using odds ratios,

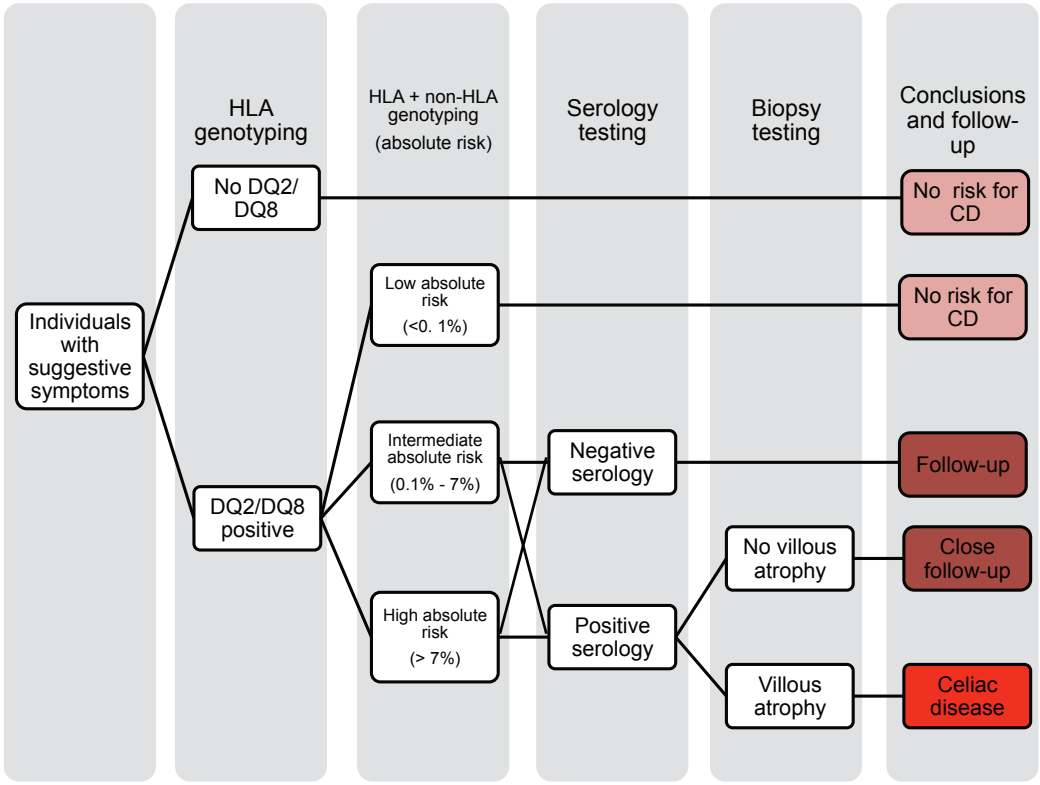



**Figure 1.** Diagnostic flow chart for celiac disease

| HLA genotype | Cases | Controls | Absolute risk (%) |
|---|---|---|---|
| DQ7/DQ7 | 0 | 39 | 0 |
| DQX/DQX | 5 | 128 | 0.04 |
| DQ7/DQX | 7 | 165 | 0.05 |
| DQ2.2/DQX | 9 | 60 | 0.17 |
| DQ8/DQ7 | 5 | 20 | 0.28 |
| DQ8/DQX | 14 | 29 | 0.53 |
| DQ2.5/DQ8 | 13 | 9 | 1.58 |
| DQ2.2/DQ2.2 | 6 | 4 | 1.64 |
| DQ8/DQ8 | 3 | 2 | 1.64 |
| DQ2.5/DQ7 | 50 | 24 | 2.26 |
| DQ2.5/DQX | 97 | 40 | 2.62 |
| DQ8/DQ2.2 | 11 | 4 | 2.96 |
| DQ2.2/DQ7 | 129 | 37 | 3.72 |
| DQ2.5/DQ2.2 | 128 | 17 | 7.71 |
| DQ2.5/DQ2.5 | 53 | 4 | 12.81 |
| TOTAL | 530 | 582 | |

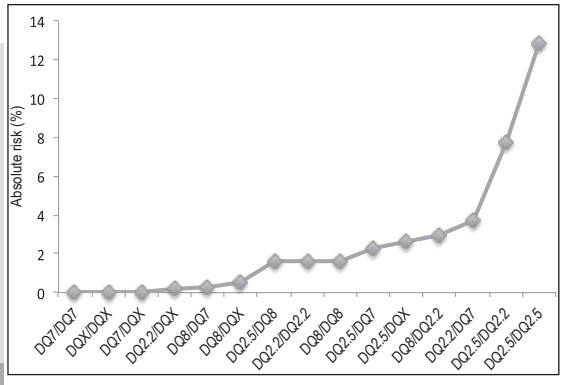



**Figure 2.** Absolute risk of is diferent HLA genotypes

relative risks, or hazard ratios, which compare the risks of disease with a reference risk, i.e., with individuals who carry few risk alleles [72]. But for individuals who undergo genetic testing, an absolute risk for disease is more interesting depending on their genetic profile. For this review, we have calculated the absolute risk for different HLA genotypes, assuming a general population prevalence of CD of 1%. Figure 2 shows the different risks for different HLA genotypes in an Italian cohort that was used for validating the genetic risk model described by Romanos et al [13]. As expected, individuals with no HLA-DQ2 and/or no HLA-DQ8 have practically no risk of developing CD (absolute risk <0.1%), whereas individuals with a double dose of HLA-DQ2 (DQ2.5/DQ2.5 and DQ2.5/ DQ2.2) have high risks of 8% and 13% respectively. The other HLA genotypes confer an intermediate risk (Figure 2). Combining HLA and non-HLA alleles allows reclassification of individuals into three risk groups: low absolute risk (<0.1%), intermediate absolute risk (0.1–7%), and high absolute risk (>7%), which correspond to the categories based on HLA genotypes (Table 1 and Figure 1). An example of reclassification is individuals in the intermediate-risk group (green dots in Figure 3) who were moved to the low- risk group, whereas others were moved to the high-risk group, indicating that combining HLA and non-HLA genotypes provides better accurate risk predictions. If we were to apply this risk model to a large cohort of CD patients, it would lead to 7.5% of intermediate-risk individuals being reconsidered as at high risk for developing CD. The current CD genetic risk model is based on a case–control cohort and should be validated in a prospective cohort in order

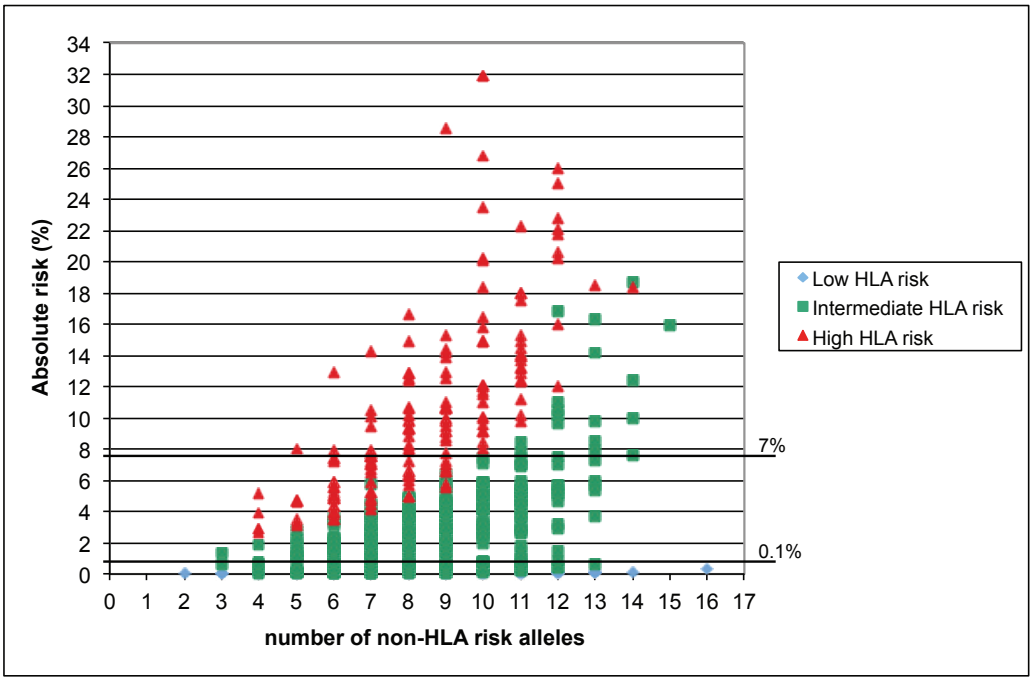



**Figure 3.** Level of individual absolute risk after including HLA and non-HLA risk alleles. The black lines at 0.1% and 7% risk mark the cut-off between low intermediate, and high risk based on HLA alone.

to estimate the positive and negative predictive values more accurately [73].

In addition, prospective testing is required to investigate whether individuals in the low-risk group are indeed at low absolute risk of developing CD. If so, we can envisage a future scenario in which only individuals in the intermediate- and high-risk groups would undergo serological testing, followed by a biopsy in the serology-positive individuals (Figure 1). As such, genetic profiling could limit the number of individuals who need to undergo serological testing and could also be used to select those individuals who should be followed up more closely.



# Application of genetic testing

As with all other complex diseases, CD affects a relatively large proportion of the human population, and puts a major burden on healthcare systems. Although the use of genetic risk prediction in clinics and mass population screening is still being debated in the literature due to the ethical and social concerns, our risk prediction model for CD could already be used to assist in better diagnosis (diagnostic tool) and prevention (screening tool) strategies (Box 1) [27, 41, 74]. For CD, early diagnosis means early intervention with treatment and prevention of long-term complications, including the development of severe and irreversible phenotypes and of other autoimmune disorders [75]. In diagnostic work, HLA genotyping is already used by many clinics as it can help to exclude the disease in individuals with atypical CD but no HLA-DQ2/DQ8. By combining the HLA and non-HLA risks, individuals could be better classified into low-, intermediate-, and high-risk groups and thus fewer indivi- duals would need to undergo serological testing and biopsy (Figure 1). Finally, risk profiling could help to detect and follow up individuals with non-specific symptoms or suggestive serology results.

Another important consideration in risk profiling is screening (Box 1). Currently, the prediction of risk for complex diseases in families is based mainly on pedigree analysis and some clinical measurements, but this approach yields predictions that are of low precision: for instance, the same risk is given to full siblings without offspring although the genomes of siblings can be so different. Estimation of risk based on genetic profiling could increase precision by differentiating between two siblings. Thus, the current genetic risk model for CD might be suited to high-risk families that already have a CD patient. The prevalence of CD in relatives of CD patients was shown to be increased in both first-degree (2.6–17.2%) and second-degree (2.6–19.5%) relatives [36]. Profile screening would permit early diagnosis of high-risk family members without symptoms or with unclear symptoms. Another application of genetic profiling would be in individuals and members of families with other immune-related disorders. Several immune-related diseases have been shown to share susceptibility variants with CD, and a modified model that includes all the shared genetic variants might identify individuals at high risk of developing an immune disorder [13, 70]. Patients with different diseases but overlapping biological

pathways might benefit from the same therapy [70].

Last but not least, genetic screening in newborns might help to identify those with a high risk for CD and could reduce the number of babies that need to be closely monitored for minor symptoms and have repeated antibody testing and early biopsy. The manifestation of CD in infants is mainly around 2 years of age, after the introduction of gluten into their diets, so screening newborns for evidence of antibodies specific to CD is mostly unhelpful. Genetic risk factors could be used to classify newborns into distinct biological pathways. This provides a rationale for different therapies being employed in the future for different patients based on their genetic information rather than by trial and error. One type of therapy could be primary prevention, which might be attained through the introduction of small doses of gluten most probably between the age of 4 and 6 months [76]. This would increase the chance that such infants could develop an oral tolerance to gluten, and might possibly promote the maintenance of tolerance throughout life [77].



# Future perspectives

The rapidly increasing knowledge on the genetic back- ground of CD has not only yielded important insights into the pathogenesis of the disease, but is also slowly entering daily clinical practice. So far, GWA studies have identified variants that explain at most 50% of the genetic heritability of CD. Larger sample sizes or a combination of several studies are needed to make better predictions of genetic risk [12]. The increasing number of associated loci and the future identification of the true genetic risk variants might enable us to identify high-risk CD patients. In the long term, the complete sequencing of variants in each person's genome might eventually be used to predict the individual's risk of developing celiac disease.

# Conclusion

CD is an important health problem for the individual and society. Identifying high-risk individuals would help in detecting CD at an early stage, and in initiating treatment before too much damage has occurred. Early intervention would reduce the risks that are currently faced by genetically susceptible individuals. Testing for the HLA-DQ2 and HLA- DQ8 genes has already yielded benefit by excluding individuals who have practically no risk of developing CD from further testing. Using HLA and non-HLA loci can better classify individuals into low-, intermediate-, and high-absolute-risk groups. The clinical value of a genetic test necessarily depends on the ease of intervention and its effect; in this respect, a gluten-free diet is a safe and feasible intervention. As such, it has now become attractive to screen children born into families with a high risk for CD or other autoimmune diseases fairly early in life for their celiac risk.

**Funding**: The study was supported by grants from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW), the Netherlands Organization for Scientific Research (NWO-VICI grant 918.66.620 to CW), and a KP6 EU grant 036383 (PREVENTCD).

**Acknowledgments**: We thank Jackie Senior for critical reading of the manuscript, and Roan Kanninga and Lude Franke for creating a program in Java to calculate the absolute risks.



## References

1. *Pennisi E. Breakthrough of the year. Human genetic variation. Science. 2007;318(5858):1842–1843.*
2. *Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008;118(5):1590–1605.*
3. *Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881– 885.*
4. *Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007; 316(5829):1331–1336.*
5. *Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006; 314(5804):1461–1463.*
6. *Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genet. 2008;40(4):395–402.*
7. *Barrett JC, Clayton DG, Concannon P, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genet. 2009 May 10. [Epub ahead of print]*
8. *Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genet. 2008;40(8):955–962.*
9. *Barrett JC, Lee JC, Lees CW, et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nature Genet. 2009;41(12):1330–4.*
10. *Holtzman NA, Marteau TM. Will genetics revolutionize medicine? N Engl J Med. 2000;343(2):141–144.*
11. *Janssens AC, Gwinn M, Valdez R, et al. Predictive genetic testing for type 2 diabetes. BMJ. 2006;333(7567): 509–510.*
12. *Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev. 2008;18(3):257–263.*
13. *Romanos J, van Diemen CC, Nolte IM, et al. Analysis of HLA and non- HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology. 2009;137(3):834–840.*
14. *Weersma RK, Stokkers PC, van Bodegraven AA, et al. Molecular prediction of disease risk and severity in a large*

*Dutch Crohn's disease cohort. Gut. 2009;58(3):388–395.*

15. *Lin X, Song K, Lim N, et al. Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score - the CoLaus Study. Diabetologia. 2009;52(4):600–608.*
16. *Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008; 359(21):2208–2219.*
17. *Lyssenko V, Jonsson A, Almgren P, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med. 2008; 359(21):2220–2232.*
18. *van Hoek M, Dehghan A, Witteman JC, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes. 2008;57(11):3122–3128.*
19. *Lango H, Palmer CN,Morris AD, et al.Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes. 2008;57(11):3129–3135.*
20. *Young RP, Hopkins RJ, Hay BA, et al. Lung cancer susceptibility model based on age, family history and genetic variants. PLoS ONE. 2009;4(4): e5302.*
21. *Soranzo N, Spector TD, Mangino M, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature Genet. 2009;41(11):1182–1190.*
22. *Young RP, Hopkins RJ, Hay BA, et al. A gene-based risk score for lung cancer susceptibility in smokers and ex-smokers. Postgrad Med J. 2009; 85(1008):515–524.*
23. *Balkau B, Lange C, Fezeu L, et al. Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes Care. 2008;31(10):2056– 2061.*
24. *De Jager PL, Chibnik LB, Cui J, et al. Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurol. 2009;8(12):1111–1119.*
25. *Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med. 2008; 358(12):1240–1249.*
26. *Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008; 358(26):2796–2803.*
27. *Khoury MJ, Jones K, Grosse SD. Quantifying the health benefits of genetic tests: the importance of a population perspective. Genet Med. 2006;8(3):191–195.*
28. *Mearin ML. Celiac disease among children and adolescents. Curr Probl Pediatr Adolesc Health Care. 2007;37(3):86–105.*
29. *Fasano A, Catassi C. Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum. Gastroenterology. 2001;120(3):636– 651.*
30. *Csizmadia CG, Mearin ML, von Blomberg BM, et al. An iceberg of childhood coeliac disease in the Netherlands. Lancet. 1999;353(9155):813– 814.*

31. *Catassi C, Fabiani E, Ratsch IM, et al. The coeliac iceberg in Italy. A multicentre antigliadin antibodies screening for coeliac disease in school-age subjects. Acta Paediatr Suppl. 1996;41:229–235.*
32. *Logan RF, Rifkind EA, Turner ID, Ferguson A. Mortality in celiac disease. Gastroenterology. 1989;97(2):265–271.*
33. *Ludvigsson JF, Montgomery SM, Ekbom A, et al. Small-intestinal histopathology and mortality risk in celiac disease. JAMA. 2009;302(11): 1171–1178.*
34. *Rubio-Tapia A, Kyle RA, Kaplan EL, et al. Increased prevalence and mortality in undiagnosed celiac disease. Gastroenterology. 2009;137(1):88– 93.*
35. *Cosnes J, Cellier C, Viola S, et al. Incidence of autoimmune diseases in celiac disease: protective effect of the gluten-free diet. Clin Gastroenterol Hepatol. 2008;6(7):753–758.*
36. *Dube 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(4 Suppl 1):S57–S67.*
37. *Murray JA. Celiac disease in patients with an affected member, type 1 diabetes, iron-deficiency, or osteoporosis? Gastroenterology. 2005;128(4 Suppl 1):S52–S56.*
38. *van Doorn RK, Winkler LM, Zwinderman KH, et al. CDDUX: a disease- specific health-related quality-of-life questionnaire for children with celiac disease. J Pediatr Gastroenterol Nutr. 2008;47(2):147–152.*
39. *Green PH, Jabri B. Coeliac disease. Lancet. 2003;362(9381):383–391.*
40. *Dewar DH, Ciclitira PJ. Clinical features and diagnosis of celiac disease. Gastroenterology. 2005;128 (4 Suppl 1):S19–S24.*
41. *Mearin ML, Ivarsson A, Dickey W. Coeliac disease: is it time for mass screening? Best Pract Res Clin Gastroenterol. 2005;19(3):441–452.*
42. *Hill ID, Dirks MH, Liptak GS, et al. Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. J Pediatr Gastroenterol Nutr. 2005;40(1):1–19.*
43. *Hill ID. What are the sensitivity and specificity of serologic tests for celiac disease? Do sensitivity and specificity vary in different populations? Gastroenterology. 2005;128(4 Suppl 1):S25–S32.*
44. *Green PH, Barry M, Matsutani M. Serologic tests for celiac disease. Gastroenterology. 2003;124(2):585–586.*
45. *National Institutes of Health Consensus Development Conference Statement on Celiac Disease, June 28–30, 2004. Gastroenterology. 2005; 128(4 Suppl 1):S1–S9.*
46. *Simell S, Hoppu S, Hekkala A, et al. Fate of five celiac disease-associated antibodies during normal diet in genetically at-risk children observed from birth in a natural history study. Am J Gastroenterol. 2007;102(9): 2026– 2035.*
47. *Marsh MN, Swift JA, Williams ED. Studies of small-intestinal mucosa with the scanning electron microscope. BMJ. 1968;4(5623):95–96.*

48. *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(1):330–354.*
49. *Memeo L, Jhang J, Hibshoosh H, et al. Duodenal intraepithelial lymphocytosis with normal villous architecture: common occurrence in H. pylori gastritis. Mod Pathol. 2005;18(8):1134–1144.*
50. *Kakar S, Nehra V, Murray JA, et al. Significance of intraepithelial lymphocytosis in small bowel biopsy samples with normal mucosal architecture. Am J Gastroenterol. 2003;98(9):2027–2033.*
51. *Cascella NG, Kryszak D, Bhatti B, et al. Prevalence of celiac disease and gluten sensitivity in the United States Clinical Antipsychotic Trials of Intervention Effectiveness Study Population. Schizophr Bull. 2009 Jun 3. [Epub ahead of print]*
52. *Burk K, Farecki ML, Lamprecht G, et al. Neurological symptoms in patients with biopsy proven celiac disease. Mov Disord. 2009 Dec 15; 24(16):2358–62.*
53. *Genuis SJ, Bouchard TP. Celiac disease presenting as autism. J Child Neurol. 2010;25(1):114–9.*
54. *Green PH, Cellier C. Celiac disease. N Engl J Med. 2007;357(17):1731– 1743.*
55. *Di Sabatino A, Corazza GR. Coeliac disease. Lancet. 2009;373(9673): 1480–1493.*
56. *Lohi S, Mustalahti K, Kaukinen K, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther. 2007;26(9):1217–1225.*
57. *Vilppula A, Kaukinen K, Luostarinen L, et al. Increasing prevalence and high incidence of celiac disease in elderly people: a population-based study. BMC Gastroenterol. 2009;9:49.*
58. *Greco L, Romino R, Coto I, et al. The first large population based twin study of coeliac disease. Gut. 2002;50(5):624–628.*
59. *Nistico L, Fagnani C, Coto I, et al. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut. 2006;55(6):803–808.*
60. *Fallang LE, Bergseng E, Hotta K, et al. Differences in the risk of celiac disease associated with HLA-DQ2.5 or HLA-DQ2.2 are related to sustained gluten antigen presentation. Nature Immunol. 2009;10(10): 1096–1101.*
61. *Vader W, Stepniak D, Kooy Y, et al. The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten- specific T cell responses. Proc Natl Acad Sci USA. 2003;100(21):12390–12395.*
62. *Congia M, Cucca F, Frau F, et al. A gene dosage effect of the DQA1*0501/ DQB1*0201 allelic combination influences the clinical heterogeneity of celiac disease. Hum Immunol. 1994;40(2):138–142.*
63. *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. Nature Genet. 2007;39(7):827–829.*
64. *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.*

*Am J Hum Genet. 2007;81(6):1284–1288.*

65. *Romanos J, Barisani D, Trynka G, et al. Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. J Med Genet. 2009;46(1):60–63.*
66. *Trynka G, Zhernakova A, Romanos J, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut. 2009;58(8):1078–1083.*
67. *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(26): 2767–2777.*
68. *Garner CP, Murray JA, Ding YC, et al. Replication of celiac disease UK genome-wide association study results in a US population. Hum Mol Genet. 2009;18(21):4219–4225.*
69. *Dubois PCA, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature Genet. 2010; **42**(4): 295-302.*
70. *Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature Rev Genet. 2009;10(1):43–55.*
71. *Liu E, Rewers M, Eisenbarth GS. Genetic testing: who should do the testing and what is the role of genetic testing in the setting of celiac disease? Gastroenterology. 2005;128(4 Suppl 1):S33–S37.*
72. *Janssens AC, van Duijn CM. Genome-based prediction of common diseases: methodological considerations for future research. Genome Med. 2009;1(2):20.*
73. *Kraft P, Wacholder S, Cornelis MC, et al. Beyond odds ratios— communicating disease risk based on genetic profiles. Nature Rev Genet. 2009;10(4):264–269.*
74. *Munnich A. Is genetics inhumane? J Med Genet. 2008;45(10):632–634.*
75. *Ventura A, Neri E, Ughi C, et al. Gluten-dependent diabetes-related and thyroid-related autoantibodies in patients with celiac disease. J Pediatr. 2000;137(2):263–265.*
76. *Norris JM, Barriga K, Hoffenberg EJ, et al. Risk of celiac disease autoimmunity and timing of gluten introduction in the diet of infants at increased risk of disease. JAMA. 2005;293(19):2343–2351.*
77. *Troncone R, Ivarsson A, Szajewska H, Mearin ML. Review article: future research on coeliac disease - a position report from the European multistakeholder platform on coeliac disease (CDEUSSA). Aliment Pharmacol Ther. 2008;27(11):1030–1043.*

# Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease

Jihane Romanos [1,*], Cleo C van Diemen [1,*], Ilja M Nolte [2], Gosia Trynka [1], Alexandra Zhernakova [3], Jingyuan Fu [1,2], Maria Teresa Bardella [4,5], Donatella Barisani [6], Ross McManus [7], David A van Heel [8], Cisca Wijmenga [1].

1 Department of Genetics, University Medical Center of Groningen, University of Groningen, Groningen, the Netherlands; [2] Department of Epidemiology, University Medical Center of Groningen, University of Groningen, Groningen, the Netherlands; [3] Complex Genetics Section, Department of Medical Genetics, University Medical Center Utrecht, the Netherlands; [4] Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy; [5] Department of Medical Sciences, University of Milan, Italy; [6] Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca, Monza, Italy; [7] Department of Clinical Medicine, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland; [8] Institute of Cell and Molecular Science, Barts and the London School of Medicine and Dentistry, London, UK.

* These authors contributed equally to this work.

*Gastroenterology 2009; 137(3):834-40, 840.e1-3.*

# CHAPTER 5

# Abstract

**BACKGROUND & AIMS**: Celiac disease (CD) is a common chronic disorder of the small intestine, resulting from aberrant cellular responses to gluten peptides, and often remains undiagnosed. It is a complex genetic disorder, although 95% of the patients carry the risk heterodimer human leukocyte antigen (HLA)-DQ2. Genome-wide association studies on CD have identified 9 non-HLA loci that also contribute to CD risk, most of which are shared with other immune-related diseases. Our aim is to predict the genetic risk for CD using HLA and non-HLA risk alleles. **METHODS**: We selected 10 independent polymorphisms in 2,308 cases and 4,585 controls from Dutch, UK, and Irish populations and categorized the individuals into 3 risk groups, based on their HLA-DQ2 genotype. We used the summed number of non-HLA risk alleles per individual to analyze their cumulative effect on CD risk, adjusting for gender and population group in logistic regression analysis. We validated our findings in 436 Italian cases and 532 controls. **RESULTS**: CD cases carried more non-HLA risk alleles than controls: individuals carrying <13 risk alleles had a higher CD risk (odds ratio, 6.2; 95% confidence interval, 4.1–9.3) compared with those carrying 0–5 risk alleles. Combining HLA and non-HLA risk genotypes in one model increases sensitivity by 6.2% compared with using only HLA for identification of high-risk individuals with slight decrease in specificity. **CONCLUSIONS**: We can use non-HLA risk factors for CD to improve identification of high-risk individuals. Our risk model is a first step toward better diagnosis and prognosis in high-risk families and population-based screening.

**Keywords:** celiac disease, risk prediction, SNP.

# Introduction

Celiac disease (CD) is a chronic, inflammatory disease of the small intestine induced by dietary proteins in wheat, rye, and barley. It is among the most common genetically determined conditions in humans occurring in 0.5%–1.0% of European populations.[1] Only about one third of the identified patients presents with diarrhea; another third is diagnosed upon targeted screening, and one fifth presents with nonspecific, recurrent abdominal pain.[2,3] Evidently, many cases remain undiagnosed and these carry the risk of long-term complications, including growth failure, anemia, osteoporosis, infertility, and cancer.[1,4] Moreover, they may be at increased risk for a number of CD-associated autoimmune disorders, like type 1 diabetes, dermatitis herpetiformis, autoimmune thyroiditis, and autoimmune hepatitis.[5]



CD is a multifactorial disorder. Several genetic factors combined with an environmental trigger are necessary for the disease to develop. Genetic predisposition to CD includes the human leukocyte antigen (HLA)-DQ2 and HLA-DQ8 heterodimers as major risk factors; these are estimated to explain some 40% of the disease heritability. The remaining 60% of the genetic susceptibility to CD is shared between an unknown number of non-HLA genes, each of which is estimated to contribute only a small risk.[6] Recently, the first genome-wide association study on CD and its follow-up have identified 9 non-HLA loci that contribute to CD risk.[7–9]

Because CD is a major health problem for both the individual and in the community, a test to identify individuals at high risk would be useful where the diagnosis remains uncertain despite an intestinal biopsy or as part of a long-term screening strategy for asymptomatic indi viduals at familial risk. Our aim is to predict the genetic risk that individuals carry for CD and to assess whether the non-HLA genes can lead to greater risk prediction than HLA alone. This test could be used by clinicians in the prevention, diagnosis, and prognosis of CD.

# Methods

## Study Populations

Our study included 2 cohorts of CD cases and controls for whom genotype data of HLA and non-HLA risk loci were available. All were of self-reported European ancestry and came from Dutch, UK, Irish, and Italian populations. The first cohort was used for creating the risk model: Dutch, 508 CD cases and 888 controls; UK, 1486 CD cases and 2983 controls; and Irish, 416 CD cases and 957 controls. The second cohort was used for validation: Italian, 538 CD cases and 593 controls (Table 1). We excluded individuals for whom genotyping failed for ≥ 1 single nucleotide polymorphisms (SNPs), which left us with 2308 and 436 cases and 4585 and 532 controls from the first and second cohorts, respectively, eligible for the study. Female gender

**Table 1.** Characteristics of controls and celiac disease.

| Characteristics | Celiac cases | Controls |
|---|---|---|
| **Number of subjects** | | |
| 1. Dutch[1] | 508 | 888 |
| 2. UK[1] | 1486 | 2983 |
| 3. Irish[1] | 416 | 957 |
| 4. Italian[2] | 538 | 593 |
| **Subjects with valid genotypes** | | |
| Number of subjects used in the initial cohort | 2308 | 4585 |
| Number of subjects used in the validation cohort | 436 | 532 |
| **Female (%)** | | |
| 1. Dutch | 66.4 | 39.1 |
| 2. UK | 74.5 | 57.9 |
| 3. Irish | 66.8 | 70.5 |
| 4. Italian | 74.5 | 61.8 |
| **Median age (range)** | | |
| 1. Dutch (age of diagnosis) | 39 (0-83) | - |
| 2. UK (age of diagnosis) | 42 (0-84) | - |
| 3. Irish | - | - |
| 4. Italian (age of onset) | 25 (1-78) | - |
| **HLA Genotypes in initial cohort** | | |
| *High risk group* | | |
| HLA-DQ2.5/DQ2.5 | 458 | 108 |
| HLA-DQ2.5/DQ2.2 | 617 | 184 |
| *Intermediate risk group* | | |
| HLA-DQ2.2/DQ2.2 | 9 | 48 |
| HLA-DQ2.5/DQX | 956 | 996 |
| HLA-DQ2.2/DQX | 167 | 789 |
| *Low risk group* | | |
| HLA-DQX/DQX | 101 | 2460 |
| **HLA Genotypes in the validation cohort** | | |
| *High risk group* | | |
| HLA-DQ2.5/DQ2.5 | 44 | 4 |
| HLA-DQ2.5/DQ2.2 | 112 | 17 |
| *Intermediate risk group* | | |
| HLA-DQ2.2/DQ2.2 | 6 | 4 |
| HLA-DQ2.5/DQX | 131 | 61 |
| HLA-DQ2.2/DQX | 116 | 123 |
| *Low risk group* | | |
| HLA-DQX/DQX | 27 | 323 |

[1] Population used to build the prediction model; [2] Population used for validation of the model

**Table 2.** Selected SNPs associated with CD

| SNP | Locus | Chr. | bp position | Alleles[a] | Odds Ratio | *P-value* | Model chosen | Reference |
|---|---|---|---|---|---|---|---|---|
| rs2816316 | RGS1 | 1 | 190803436 | **A**:C | 1.41 | 2.58E-11 | Additive | 7 |
| rs917997 | IL18RAP | 2 | 102437000 | **A**:G | 1.27 | 8.49E-10 | Additive | 7 |
| rs6441961 | CCR | 3 | 46327388 | **A**:G | 1.21 | 3.14E-07 | Additive | 7 |
| rs17810546 | IL12A/SCHIP | 3 | 161147744 | **G**:A | 1.34 | 1.07E-09 | Dominant | 7 |
| rs9811792 | IL12A | 3 | 161179692 | **G**:A | 1.21 | 5.24E-08 | Additive | 7 |
| rs1464510 | LPP | 3 | 189595248 | **A**:C | 1.21 | 5.33E-09 | Additive | 7 |
| rs6822844 | IL2/IL21 | 4 | 123728871 | **A**:C | 1.41 | 2.82E-13 | Recessive | 7 |
| rs2327832 | OLIG3-TNFAIP3 | 6 | 138014761 | **G**:A | 1.25 | 1.31E-08 | Additive | 9 |
| rs1738074 | TAGAP | 6 | 159385965 | **A**:G | 1.21 | 6.71E-08 | Additive | 7 |
| rs3184504 | SH2B3 | 12 | 110368991 | **A**:G | 1.19 | 1.33E-07 | Additive | 7 |

Chr, Chromosome; bp, base pair; [a]Bold indicates the risk alleles.



and HLA-DQ genotypes are described as percentages in Table 1. Data on age at diagnosis was available for Dutch (median, 39 years) and UK (median, 42 years), whereas for the Italian cohort, we had data on age of onset (median, 25 years). Patients were diagnosed according to the European Society for Paediatric Gastroenterology, Hepatology, and Nephrology criteria.[10] The collection of the cohorts has been described previously.[7,8,11] The collection of patient and control materials was approved by the Medical Ethical Committee of the University Medical Centre Utrecht, Oxfordshire Research Ethics Committee B, or East London and the City Research Ethics Committee 1, the Institutional Ethics Committee of St. James's Hospital, Dublin, and the Ethical Committee of the Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan. All participants provided written, informed consent.

## Genotyping

The primary HLA association in most CD patients is with DQ2 (more specifically HLA-DQ2.5) and in 5%– 10% with DQ8.[12,13] The risk molecule HLA-DQ2, includes HLA-DQ2.5 (ie, the DQA1*05/DQB1*0201 haplotype) and HLA-DQ2.2 (DQA1*0201/DQB1*0202). We had available genotype data of 3 tag SNPs (rs2395182, rs7775228, and rs2187668) to predict whether an individual had 0, 1, or 2 HLA-DQ2.5 and/or DQ2.2 haplotypes using the tagging SNP approach described by Monsuur et al.[14] DQ8 was not taken into account in this study owing to unavailability of SNP data to predict this haplotype. In addition, from the first genome-wide association study in CD and its follow-up, we included genotype data of 10 SNPs from 9 genomic regions with a genome-wide significant P-value (< $5x10^{-7}$; Table 2).[7–9] To assess the non-HLA genotype score, we selected the SNP with the strongest evidence for association at each locus, except for the IL12A/SCHIP locus, for which we included 2 significant but independent SNPs (rs17810546 and rs9811792).

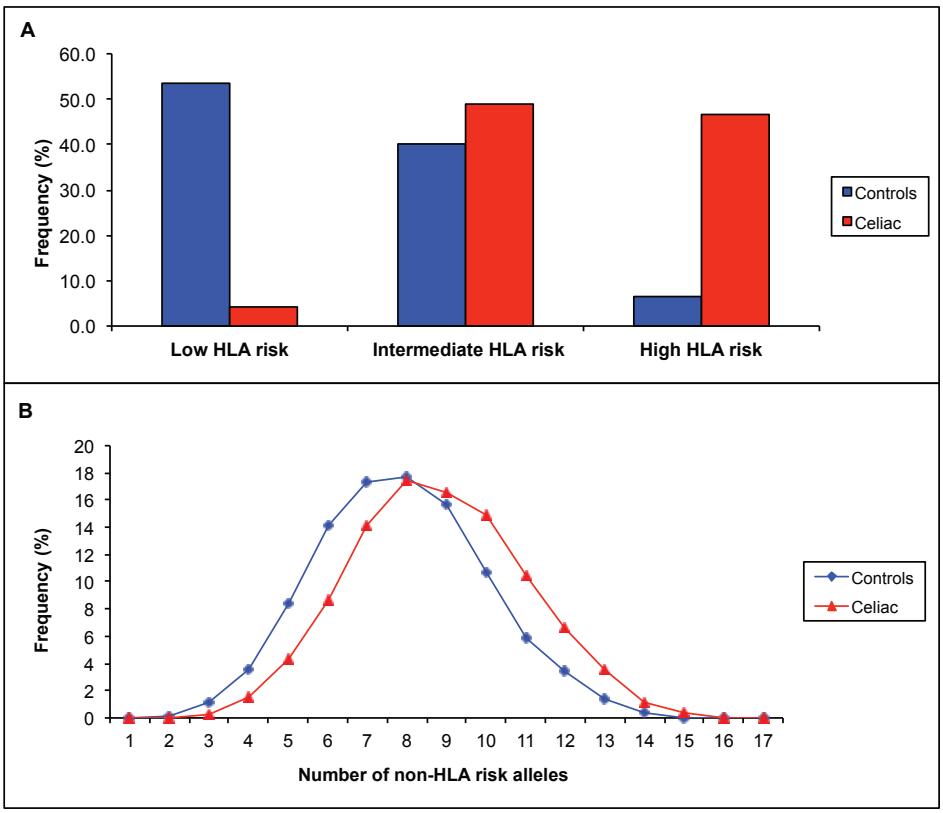



**Figure 1.** Frequency distribution of HLA (A) and non-HLA loci (B) in cases and controls.

## Data Analysis

We coded genotypes as 0 for the non-risk homozygous genotype, 1 for the heterozygous genotype, and 2 for the homozygous risk genotype. We determined the inheritance model (dominant, recessive, or additive) for each individual SNP by analyzing the β-coefficients (log [odds ratio (OR)]) of the genotypes as categorical variables in logistic regression models adjusted for gender and population origin (Supplemental Table 1). In the dominant model, heterozygous genotypes were recoded as homozygous risk genotypes (weight of 2); in the recessive model, heterozygous genotypes were recoded as homozygous non-risk (0), whereas homozygous risk genotypes were recoded as 1. In the additive model, the risk scores remained unchanged. Hunt et al[7] showed that all SNPs have an independent association with CD, have similar ORs, and show no evidence of interaction between SNPs, so we summed the number of non-HLA risk alleles to obtain a total allele score per individual (Table 2).

Several studies have shown that individuals homozygous for HLA-DQ2.5 or HLA-DQ2.5/DQ2.2

genotypes have an increased risk for CD compared with those homozygous for HLA-DQ2.2, or heterozygous for HLA-DQ2.5 or for HLA-DQ2.2.[6,15,16] Therefore, we categorized individuals as having a low risk if they were HLA-DQ2 negative (ie, neither HLA- DQ2.5 nor HLA-DQ2.2), an intermediate risk if they were homozygous for HLA-DQ2.2, or heterozygous for HLA-DQ2.5 or for HLA-DQ2.2, or a high risk for those homozygous for HLA-DQ2.5 or HLA-DQ2.5/DQ2.2.

*Risk models.* We estimated the risk for each total allele score in logistic regression models using the group of individuals with a total allele score of 0 and up to 5 as a reference group and adjusting for the HLA risk group, gender, and population group. To evaluate the overall discrimination of our genetic model, we calculated the area under the receiver operating characteristic curve (AROC). For this analysis, we calculated the risk for HLA and each non-HLA SNP independently in logistic regression analyses. The β-coefficients from these analyses reflect the weighted risk per genotype. The sum of these β-coefficients per individual was used as a weighted total risk score including HLA and non-HLA SNPs.

Moreover, we selected only HLA-DQ2–positive individuals (ie, the intermediate- and high-risk groups) and estimated the risk per total allele score to assess whether the non-HLA genes would lead to better CD risk prediction than HLA alone. We used a cutoff value of the OR of the HLA risk between the intermediate- and high-risk individuals to reclassify individuals as being at high risk using non-HLA risk genotypes. In other words, we reclassified subjects with intermediate HLA risk to high risk when they had a risk load of non-HLA alleles that was similar to the OR between the intermediate- and high-risk HLA group. All analyses were performed using SPSS version 16.0.

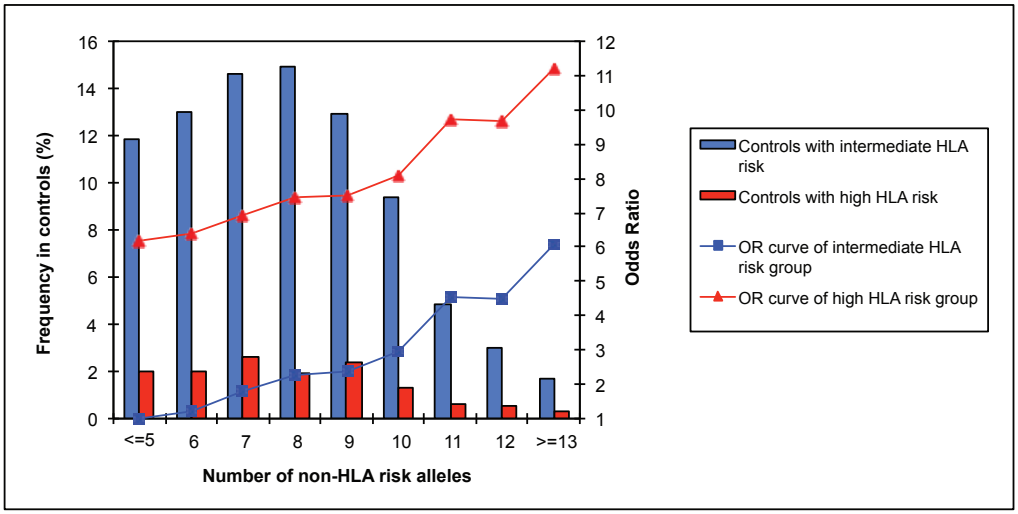
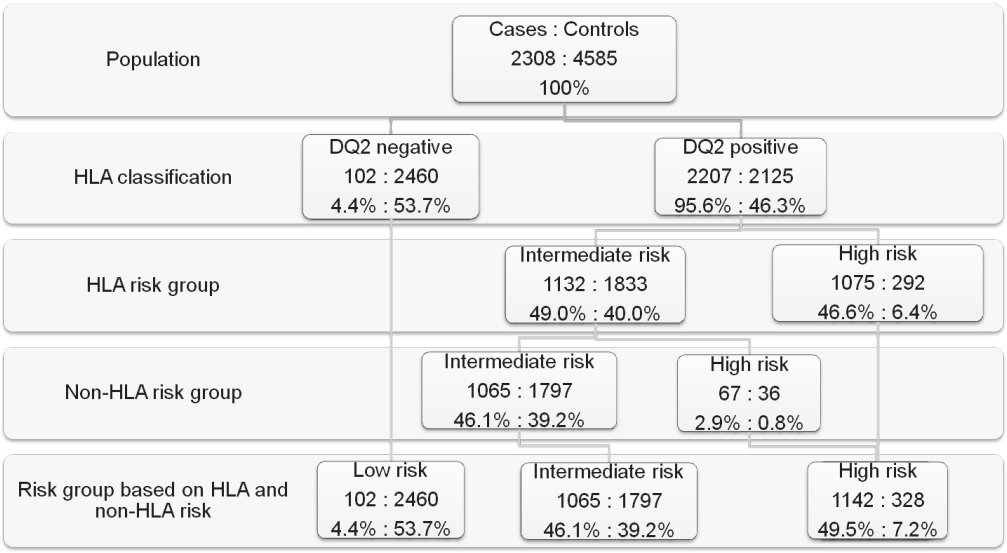



**Figure 2.** Risk and frequency distribution of non-HLA loci in DQ2 positive individuals.

A. Model




B. Validation

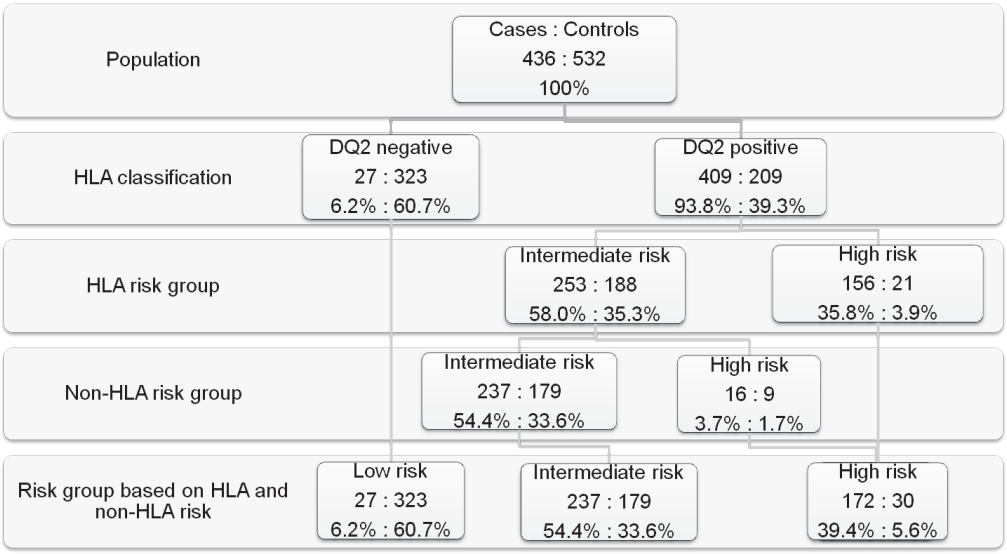



**Figure 3.** Reclassification of individuals at intermediate risk to high risk for CD.

*Age of onset.* Data on age of diagnosis was available for Dutch and UK cases; therefore, we categorized 941 CD patients as early (≤ 12 years; n = 172) or late onset (≥ 18 years; n = 745). Subjects with age of onset between 12 and 18 years (n = 24) were not included in this analysis. Using logistic regression, we estimated the effect of the risk from the total allele score in both groups, adjusting for gender and population group.

*Effect of additional risk alleles.* Adding risk alleles to the model may increase the accuracy of risk prediction. Thus, we tested whether including additional SNPs from the same risk locus improved the accuracy of prediction. To estimate the effect of knowing more risk loci, we also performed different simulation models by adding extra simulated loci to the 10 SNPs already selected: (1) adding simulated risk genotypes with similar frequencies and effect sizes; and (2) adding simulated risk genotypes with lower frequencies and effect sizes.



# Results

## Distribution of HLA and Non-HLA Alleles

We divided the first study population into 3 groups: low risk (4.4% of cases, 53.7% of controls), intermediate risk (49.0% of cases, 40.0% of controls), and high risk (46.6% of cases, 6.4% of controls), based on the dosage effect of the HLA-DQ2 molecule. Figure 1A shows the distribution of the HLA risk groups in cases and controls from the 3 combined populations (Dutch, UK, and Irish). Thus, similar to the literature, around 90% of cases are HLA-DQ2.5 carriers versus 28.1% of controls.[12–14,17] Because we know that HLA is necessary but not sufficient for disease development, we created a genetic risk model using the 10 non-HLA SNPs recently identified as independently associated to CD. Both cases and controls display a normal distribution of the total number of risk alleles.

## Genetic Risk Model

Logistic regression analysis, adjusted for population, gender, and HLA, shows that an increasing number of risk alleles is associated with an increased risk for CD (Supplemental Figure 1). Individuals carrying ≥13 risk alleles (5.2% of cases vs 1.9% of controls) had a significantly greater CD risk ($P = 1.7x10^{-18}$; OR, 6.2; 95% confidence interval [CI], 4.1–9.3) compared with a reference group carrying 0–5 alleles (6.1% of cases, 13.2% of controls). The reference group is the 5th percentile of the cumulative risk alleles distribution of cases and controls together. When stratifying on HLA-DQ2 positivity, logistic regression analysis in the intermediate- and high-risk groups showed that the individuals carrying a higher number of non-HLA risk alleles were at greater risk for CD compared with the reference group. An individual with a high HLA risk and zero non-HLA risk has an OR of 6.2 (95% CI, 5.3–7.2), which is very similar to the risk of individuals with intermediate HLA risk and

13 non- HLA risk alleles (OR, 6.1; 95% CI, 3.9–9.4; Figure 2; Supplemental Table 2). When testing the genetic risk score on age of diagnosis, we found no significant difference in risk between having an early or late diagnosis of CD.

## Reclassification of Individuals From the Intermediate- to the High-Risk Category

To evaluate our genetic model, we calculated the $A_{ROC}$ for the 2 models, with and without non-HLA risk alleles. For the prediction based on only HLA risk, the $A_{ROC}$ is 85.4% (95% CI, 84.5–86.3), whereas the $A_{ROC}$ based on HLA and non-HLA risks improved to 87.4% (95% CI, 86.6 – 88.3). Furthermore, we used a reclassification approach to examine the usefulness of our model in diagnosis.[18,19] We determined how many individuals from the intermediate-risk group could be reclassified into the high-risk group using the cutoff value of the additional HLA risk between the intermediate- and high-risk groups (OR, 6.1; 95% CI, 5.2–7.1). From 2,965 subjects (1132 cases, 1833 controls) with intermediate HLA risk, we were able to reclassify 103 individuals (67 cases, 36 controls) as high risk after including the non-HLA risk alleles (Figure 3A). Thus, 7.5% of HLA-DQ2–positive in dividuals with intermediate a priori risk should be reclassified into the high-risk group for CD. We compared our model results for only HLA genotypes versus those for both HLA and non-HLA risk alleles: The sensitivity increased by 6.2% (from 46.6% [95% CI, 44.5– 48.6] to 49.5% [95% CI, 47.4 –51.5]), whereas the specificity decreased slightly from 93.6% (95% CI, 92.9 –94.3) to 92.8% (95% CI, 92.1–93.6).



## Validation of the Prediction Model

For the validation of the prediction model, we used an independent Italian cohort of 436 cases and 532 controls with valid genotypes. Based on only the HLA genotypes, we classified 156 cases and 21 controls in the high-risk group, 253 cases and 188 in the intermediate-risk group, and 27 cases and 323 controls in the low-risk group to develop CD. When we included non-HLA risk alleles with the HLA, we were able to reclassify 16 cases and 9 controls of intermediate risk into high-risk group (an increase of 14.1%; Figure 3B).

## Effects of Adding Risk Alleles to the Prediction Model

We hypothesized that the risk prediction would improve by adding more risk alleles from the same locus. We therefore included additional SNP data from the IL2/IL21 and LPP loci to the model. However, this did not increase the ORs for these loci.

By using a simulated dataset of risk genotypes with different effect sizes and genotype frequencies, it was evident that with 10 extra risk genotypes, the OR for the group with the greatest number of risk alleles increased to 12.1. When we included extra genotypes with lower effect sizes and lower frequencies, the OR was 11.6 for the group with the highest number of risk genotypes compared with the reference group (Supplemental Figure 2).

Using these models, the number of individuals that could be reclassified from intermediate to high risk was approximately 10%, which was similar for each simulation. Thus, it seems that there is a plateau for the number of individuals who can be reclassified, no matter how many new risk loci are identified or what type they are.

# Discussion



The recent genome-wide association studies have identified many non-HLA loci for immune-related complex diseases.[20] These loci generally confer only moderate risks with an OR ranging from 1.1 to 1.5, which has shown by simulation to be sufficient to predict individuals at high risk in the population.[21] In this study, we show that by combining the individual risk alleles, the risk of developing CD increases, reaching an OR of 6.2 for individuals carrying ≥13 risk alleles. Moreover, using non-HLA risk loci leads to a 7.5% increase (from 31.6% to 33.9%) in the classification of DQ2-positive individuals at high risk for CD compared with taking only their HLA genotype into account. By testing the risk model in an independent population, we have shown that our risk model is valid. Genetic risk prediction should assist in better diagnosis and prognosis strategies; for CD, this would include the early diagnosis of at-risk family members of CD patients who often have unclear symptoms.[4] Identifying newborns at risk for CD by genetic testing would help to target the appropriate group to be followed up more closely by repetition of antibody testing and early biopsy. An early diagnosis means early intervention with a gluten-free diet, which may prevent the development of a more severe disease phenotype or of comorbidities, like other immune-related diseases.[22]

The first step in classifying CD using genetics should be based on HLA-typing, which identifies 30%– 40% of the general population at risk. This group can be further divided into high risk and intermediate risk by combining both HLA and non-HLA risk alleles. With the currently identified non-HLA loci, 33.9% of the HLA-DQ2– positive individuals will fall into the high-risk group (1,142 cases, 328 controls). These non-CD individuals should be followed more closely by serologic screening and, if necessary, by intestinal biopsy because they can be undiagnosed cases in the general population.

Currently, 9 CD risk loci outside HLA have been identified, together accounting for 5% of the explained genetic risk. In future genome-wide association studies, more genes will be identified that will increase the predictive power for genetic tests. However, the question is how many more loci need to be identified to optimize the risk prediction. Our simulation of adding new risk loci with different characteristics highlighted 2 aspects. First, with 10 additional risk loci, we can already reclassify 10% of individuals at intermediate risk to the high-risk group. Second, the predictive value of the number of risk alleles is limited, because the number of individuals that can be reclassified does not increase much with more risk loci. The largest gain in risk prediction is thus gathered by identifying the loci with the largest effect size. The reasons why we will not

reach a 100% inclusion of all individuals at risk are the involvement of small, undetectable genetic factors and triggering environmental factors like infections that are not included in our risk model.

The results of genome-wide association studies have shown that immune-related diseases, including CD, type 1 diabetes, and Crohn's disease, share a large genetic component as seen in the disease-overlapping genes. Approximately 25 genes are now known to be shared between ≥ 2 immune-related diseases, such as *IL2/IL21* and *IL18RAP*.[7,8,11,20,23–26] Moreover, both the shared as well as the disease-specific genes fall into functional categories related to T-cell biology and innate immunity, indicating that these diseases have a shared genetic and functional pathogenesis. Immune-related diseases affect approximately 5%–10% of the general population and often co-occur in patients and in families.[5] We believe that generating a modified risk model to include all the genes shared by immune-related diseases will efficiently identify high-risk individuals, in particular in families with only 1 individual affected by an immune-related disease. We have already made such a risk model in which we included only the genes shared with other immune-related diseases (*IL18RAP, IL2/IL21, RGS1, SH2B3, TAGAP, CCR3, OLIG3-TNFAIP3*).[7,8,11,20,23–25] In this model, the group carrying ≥ 9 risk alleles (3.8% of controls) had a higher risk for CD (OR, 4.8; 95% CI, 3.5–6.5). Soon, we will be able to apply such risk models to other diseases for which prospective cohort studies are available. Because HLA plays a less prominent role in other immune-related diseases than in CD (shown by lower heritability), a risk model based on non-HLA genes may prove to be an even more powerful tool than in CD.



**Acknowledgments**

Jihane Romanos and Cleo C. van Diemen contributed equally to this work. The authors thank all the patients and controls who participated in this study. We thank Jackie Senior and Hermien de Walle for critically reading the manuscript.

**Conflicts of interest**

The authors disclose the following: David A. van Heel, Cisca Wijmenga and Ross McManus declare competing financial interests. A patent application by Queen Mary University of London is in progress. The remaining authors disclose no conflicts.

**Funding**

Supported by grants from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW), the Netherlands Organisation for Scientific Research (NWO-VICI grant 918.66.620 to CW) and KP6 EU grant 036383 (PREVENTCD).

## References


1. *Van Heel DA, West J. Recent advances in celiac disease. Gut 2006;7:1037–1046.*
2. *Ravikumara M, Tuthill DP, Jenkins HR. The changing clinical presentation of celiac disease. Arch Dis Child 2006;12:969– 971.*
3. *Beattie RM. The changing face of celiac disease. Arch Dis Child 2006;12:955–956.*
4. *Mearin ML, Catassi C, Brousse N, et al. European multi-centre study on celiac disease and non-Hodgkin lymphoma. Eur J Gastroenterol Hepatol 2006;2:187–194.*
5. *Somers EC, Thomas SL, Smeeth L, et al. Autoimmune diseases co-occurring within individuals and within families: a systematic review. Epidemiology 2006;2:202–217.*
6. *Dubois PC, van Heel DA. Translational mini-review series on the immunogenetics of gut disease: immunogenetics of celiac disease. Clin Exp Immunol 2008;2:162–173.*
7. *Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 2008;4:395–402.*
8. *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. Nat Genet 2007;7:827–829.*
9. *Trynka G, Zhernakova A, Romanos J, et al. Celiac disease associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 2009; 58(8): 1078-83..*
10. *Revised criteria for diagnosis of celiac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. Arch Dis Child 1990;8:909–911.*
11. *Romanos J, Barisani D, Trynka G, et al. Six new celiac disease loci replicated in an Italian population confirm association to celiac disease. J Med Genet 2008;46:60–63.*
12. *Sollid LM. Celiac disease: dissecting a complex inflammatory disorder. Nat Rev Immunol 2002;9:647–655.*
13. *Kaukinen K, Collin P, Maki M. Natural history of celiac disease. In: Fasano A, Troncone R, Branski D, eds. Frontiers in celiac disease. Basel: Karger; 2008:12–17.*
14. *Monsuur AJ, de Bakker PI, Zhernakova A, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PLoS ONE 2008;5: e2270.*
15. *Vader W, Stepniak D, Kooy Y, et al. The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. Proc Natl Acad Sci U S A 2003;21:12390–12395.*
16. *Congia M, Cucca F, Frau F, et al. A gene dosage effect of the DQA1*0501/DQB1*0201 allelic combination influences the clini- cal heterogeneity of celiac disease. Hum Immunol 1994;2:138– 142.*
17. *Karinen H, Karkkainen P, Pihlajamaki J, et al. Gene dose effect of the DQB1*0201 allele contributes to severity of celiac disease. Scand J Gastroenterol 2006;2:191–199.*
18. *PencinaMJ,D'AgostinoRB,Sr.,D'AgostinoRB,etal.Evaluatingthe added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;2:157–172.*
19. *Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction.*

*Circulation 2007;7:928 –935.*

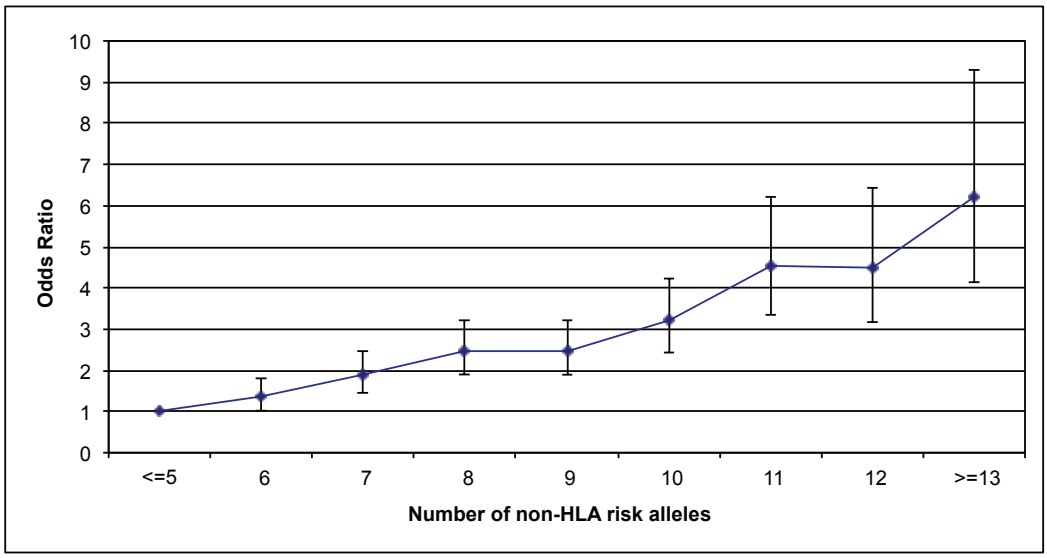
20. *Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related dis- eases. Nat Rev Genet 2009;1:43–55.*
21. *Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007;17:1520–1528.*
22. *Ventura A, Magazzu G, Greco L. Duration of exposure to gluten and risk for autoimmune disorders in patients with celiac dis- ease. SIGEP Study Group for Autoimmune Disorders in Celiac Disease. Gastroenterology 1999;2:297–303.*
23. *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.*
24. *Xavier RJ, Rioux JD. Genome-wide association studies: a new window into immune-mediated diseases. Nat Rev Immunol 2008; 8:631– 643.*
25. *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. Am J Hum Genet 2007;6:1284–1288.*
26. *Festen EA, Goyette P, Scott R,et al. Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut 2009; 58:799 – 804.*

## Supplementary Data

**Supplemental Table 1.** Beta of Heterozygous and Homozygous Risk Alleles in the General Model Adjusted for Gender and population origin and using Dutch, UK and Irish Cohorts.

| | | Risk per genotype | | | General model | | |
|---|---|---|---|---|---|---|---|
| SNP | Locus | 0 | 1 | 2 | ß1 | ß2 | Model chosen |
| rs2816316 | RGS1 | CC | AC | AA | 0.307 | 0.645 | Additive |
| rs917997 | IL18RAP | GG | AG | AA | 0.213 | 0.465 | Additive |
| rs6441961 | CCR | GG | AG | AA | 0.145 | 0.345 | Additive |
| rs17810546 | IL12A/SCHIP | AA | AG | GG | 0.383 | 0.434 | Dominant |
| rs9811792 | IL12A | AA | AG | GG | 0.200 | 0.385 | Additive |
| rs1464510 | LPP | CC | AC | AA | 0.244 | 0.409 | Additive |
| rs6822844 | IL2/IL21 | AA | AC | CC | 0.089 | 0.500 | Recessive |
| rs2327832 | OLIG3-TNFAIP3 | AA | AG | GG | 0.197 | 0.476 | Additive |
| rs1738074 | TAGAP | GG | AG | AA | 0.187 | 0.379 | Additive |
| rs3184504 | SH2B3 | GG | AG | AA | 0.156 | 0.406 | Additive |

**Supplemental Figure 1.** Increase risk of CD with increase number of non-HLA risk alleles, with the reference group being <5

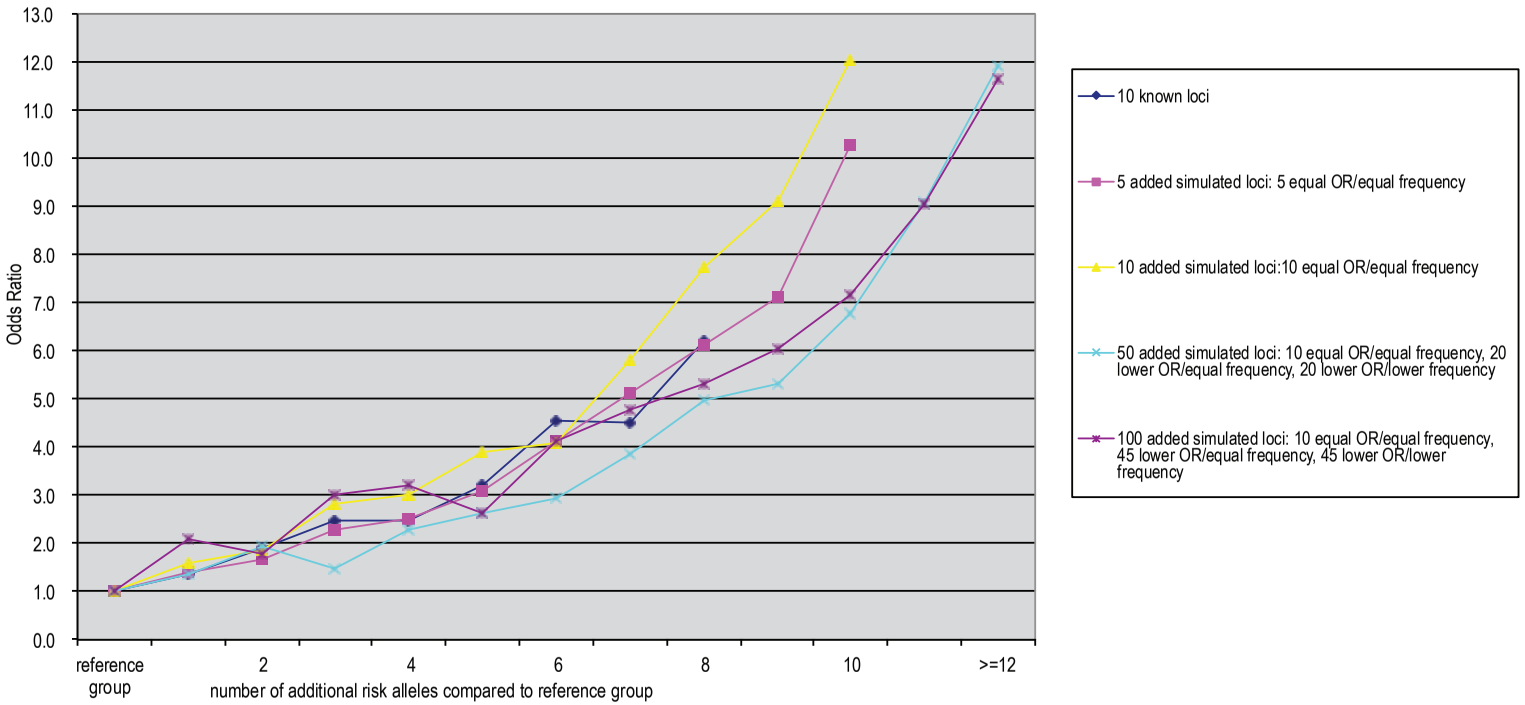


**Supplementary Table 2**: Risk of CD of non-HLA loci in the group of HLA-DQ2 positive individuals

| Groups | Odds Ratio | 95.0% CI | | p-value |
|---|---|---|---|---|
| | | Lower | Upper | |
| HLA (intermediate vs high) | 6.2 | 5.3 | 7.2 | 4.15E-116 |
| No. of non-HLA risk alleles | | | | |
| <=5 | 1.0 | n.a | n.a | 3.56E-35 |
| 6 | 1.2 | 0.9 | 1.6 | 1.84E-01 |
| 7 | 1.8 | 1.4 | 2.4 | 3.96E-05 |
| 8 | 2.3 | 1.7 | 3.0 | 4.10E-09 |
| 9 | 2.4 | 1.8 | 3.1 | 9.70E-10 |
| 10 | 3.0 | 2.2 | 3.9 | 2.11E-13 |
| 11 | 4.5 | 3.3 | 6.3 | 1.38E-19 |
| 12 | 4.5 | 3.1 | 6.6 | 3.95E-15 |
| >=13 | 6.1 | 3.9 | 9.4 | 4.04E-16 |

[a] Reference group are individuals with 0 to 5 risk alleles.

**Supplemental Figure 2.** Simulation of additional risk alleles to the genetic risk model.




| | 10 known loci | 5 added loci | 10 added loci | 50 added loci | 100 added loci |
|---|---|---|---|---|---|
| total number of controls | 4585 | 4585 | 4585 | 4585 | 4585 |
| number of controls with high HLA risk | 292 | 292 | 292 | 292 | 292 |
| number of controls with intermediate HLA risk | 1833 | 1833 | 1833 | 1833 | 1833 |
| number of controls with intermediate HLA and high non-HLA risk* | 36 | 87 | 189 | 132 | 189 |
| total % of controls at high risk | 7.2 | 8.3 | 10.5 | 9.2 | 10.5 |

*individuals are at high non-HLA risk when their OR of non-HLA is similar to that of OR between intermediate and high HLA risk

# Adding non-HLA variants to HLA risk model for celiac disease classifies individuals into more appropriate risk categories

Jihane Romanos [1,*], Anna Rosén [2,3,*], Gosia Trynka [1], Lude Franke [1], Agata Szperl [1], Javier Gutierrez-Achury [1], Cleo C. van Diemen [1], Roan Kanninga [1], Soesma Medema-Jankipersadsing[1], Andrea K. Steck [4], George S. Eisenbarth [4], David A. van Heel [5], Bozena Cukrowska[6], Luigi Greco [7], Maria Cristina Mazzilli [8], Concepción Núñez [9], Jose Ramon Bilbao [10], M. Luisa Mearin [11], Donatella Barisani [12], PreventCD Group [§], Marian Rewers [4], Jill M. Norris [13], Anneli Ivarsson [2], Marieke H. Boezen [14,#], Edwin Liu [4,#], Cisca Wijmenga [1,#]

[1] Department of Genetics, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands. [2] Department of Public Health and Clinical Medicine, Epidemiology and Global Health, Umeå University, Umeå, Sweden. [3] Department of Medical Biosciences, Clinical and Medical Genetics, Umeå University, Umeå, Sweden. [4] Barbara Davis Center for Childhood Diabetes, University of Colorado Denver, Aurora, Colorado, USA. [5] Institute of Cell and Molecular Science, Barts and the London School of Medicine and Dentistry, London, UK. [6] Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland. [7] European Laboratory for Food-Induced Disease, University of Naples Federico II, Naples, Italy. [8] Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy. [9] Immunology Department, Hospital Clínico S. Carlos, Instituto de Investigación Sanitaria San Carlos IdISSC, Madrid, Spain. [10] Immunogenetics Research Laboratory, Hospital de Cruces, Barakaldo 48903 Bizkaia, Spain. [11] Department of Pediatrics, Leiden University Medical Centre, Leiden, The Netherlands. [12] Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca, Monza, Italy.
[13] Epidemiology Department, Colorado School of Public Health, Aurora, USA.
[14] Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

[§] See appendix
[*] These authors contributed equally.
[#] These authors jointly directed this project

*Submitted for publication*

# CHAPTER 6

# Abstract

**Background:** At present, more than 80% of celiac disease (CD) patients are not being properly diagnosed and therefore remain untreated, leading to a greater risk of developing CD-associated complications and increased mortality. Thus, genetic testing for CD could help prioritize high-risk individuals for closer follow-up. The major genetic risk for CD, HLA-DQ2 and DQ8, is already used in clinical practice for excluding the disease. However, approximately 30% of the population carries these alleles and never develop CD. We recently identified a total of 57 non-HLA SNPs that contribute to CD development. Here, we explore if CD risk prediction can be improved by adding non-HLA susceptible variants, and assess how well this can be transferred to other cohorts. **Methods:** We developed an average weighted genetic risk score (GRS) with 10, 26 and 57 SNPs in 2,675 cases and 2,815 controls and assessed the improvement in risk prediction provided by the non-HLA SNPs. Moreover, we assessed the transferability of the genetic risk model with 26 non-HLA variants to a nested case-control population (n=1,709) and a prospective cohort (n=1,244) and finally tested how well this model predicts CD outcome for 985 independent individuals. **Findings:** Adding 57 non-HLA variants to HLA showed a statistically significant improvement compared to scores from HLA only, HLA+10 SNPs, and HLA+26 SNPs. The area under the ROC curve reached 0.854 and 11.1% of individuals were reclassified to a more accurate risk group. Moreover, the risk model with HLA+26 SNPs was shown to be useful in independent populations, mainly for HLA-DQ2- and DQ8-positive individuals. **Interpretation:** Personalizing our risk assessment by more detailed genetic profiling improved the identification of potential CD patients. This sets out a possible role for combined HLA and non-HLA genetic testing in diagnostic work for CD.

**Keywords:** celiac disease, genetic risk prediction, SNP, weighted genetic risk score

# Introduction

Celiac disease (CD) is a genetically driven, autoimmune disorder of the small intestine triggered by gluten, a protein common in a normal Western diet. CD is a major health problem affecting around 1% of Western populations but with a poorly understood etiology.[1] Several countries have shown an alarming increase in the prevalence of CD recently and some European countries are already reporting a prevalence of 2-3%.[2–4] Age at onset ranges from infancy to late adulthood, and clinical presentation of the disease can be highly variable, from classical gastrointestinal manifestations like diarrhea, stomach ache, and malabsorption to more severe and non-reversible presentations such as osteoporosis and infertility.[5] Due to its broad spectrum of symptoms, the disease is severely under-recognized: on average only 1 out of 7 patients is being properly diagnosed.[6] In particular, family members of CD patients or those suffering from another immune-mediated disease are at an increased risk of developing CD. Unrecognized and therefore untreated CD patients have a greater risk of developing CD-associated complications or other immune-mediated diseases (such as type 1 diabetes, autoimmune hepatitis, or thyroid disease), and they show a markedly higher mortality.[4,7] The antibodies used as markers for CD have a relatively high sensitivity and specificity mainly in patients with severe intestinal lesions, however, those with mild lesions, partial villous atrophy, or children younger than 2 years can be missed by these tests.[8,9]



The importance of genetic testing is highlighted by the revised guidelines for the diagnosis of CD recently proposed by The European Society of Pediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN), which include HLA-DQ2 and DQ8 as one of the diagnostic criteria.[10] These heterodimers are known to be the major genetic risk factors for CD and have an almost 100% negative predictive value, but their positive predictive value is low since approximately 30% of the population carries one or both of these alleles.[11] In the past few years, two genome-wide association studies (GWAS) and one fine-mapping project have identified 10 non-HLA SNPs (9 loci), 26 non-HLA SNPs (26 loci), and 57 non-HLA SNPs (39 loci) which contribute to CD susceptibility.[12–15] In 2009, we published a genetic risk model for CD using HLA and the 10 non-HLA risk variants from the first GWAS.[16] That model was shown to be efficient in improving the identification of individuals at high-risk of developing CD. Now, with the increased number of associated loci known for CD, our aim was to firstly test if the genetic risk model could be improved by adding the new variants, secondly assess how well the risk model is transferable to other cohorts, and thirdly evaluate how well this risk profiling can be used in clinical practice.

**Table 1.** The different datasets included in this study: a discovery set for single SNP odds ratio calculation, a derivation set to create the risk models, two validation set to validate the risk model, and a test set to evaluate the model in clinical practice.

| Cohorts | Discovery set: case/control | | Derivation set: case/control | | Validation set 1: nested case/control | | Validation set 2: Prospective | | test set: case/control | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cases | Controls | Cases | Controls | Cases | Controls | CDA* | noCDA* | Cases | Controls |
| Italy | 695 | 635 | 693 | 635 | | | | | 99 | 219 |
| Netherlands | 535 | 586 | 535 | 583 | | | | | 61 | 175 |
| Poland | 235 | 270 | 236 | 269 | | | | | 50 | 67 |
| Spain1 | 242 | 171 | 242 | 170 | | | | | 34 | 122 |
| Spain2 | 268 | 160 | 269 | 159 | | | | | 33 | 125 |
| UK | 700 | 1000 | 700 | 999 | | | | | | |
| Sweden | | | | | 306 | 1403 | | | | |
| Non-Hispanic white American | | | | | | | 70 | 1174 | | |
| sub-total | 2675 | 2822 | 2675 | 2815 | 306 | 1403 | 70 | 1174 | 277 | 708 |
| Total | 5497 | | 5490 | | 1709 | | 1244 | | 985 | |

***CDA celiac disease autoimmunity.**



# METHODS

## Study populations

Our study includes four groups (Table 1): (1) a discovery set of 2,675 CD cases and 2,822 healthy controls in which we calculated the odds ratio (OR) for each SNP after having identified the mode of inheritance; (2) a derivation set of 2,675 cases and 2,815 controls in which we created the risk model; (3) two validation sets for validating the risk model and which include a total of 1,709 nested case-control collection (validation set 1), and a prospective cohort of 1,244 individuals (validation set 2); and (4) a test set of 985 independent individuals on whom we applied the risk model.

The discovery and derivation case-control samples were previously included in our CD meta-analysis and incorporated cohorts from the Netherlands, Italy, Poland, Spain and the UK.[15] To prevent over-fitting of the model, we randomly selected 50% of the cases and controls to form a discovery dataset in which we calculated the OR, while and the other half became the derivation set to create the risk model (Table 1). The samples were evenly distributed across the different populations, except for the UK cohort from which we randomly selected 700 cases and 1,000 controls to obtain sample sizes equal to the other populations.

The first validation set included cases and matched controls from a Swedish cross-sectional CD-screening of 12-year-olds. In 2005-2006 and in 2009-2010, all sixth graders attending schools in five Swedish towns (Lund, Växjö, Norrköping, Norrtälje and Umeå) were invited to participate in the study. The majority of these children were born in 1993 or in 1997 and belong to birth cohorts that differ with respect to infant feeding practices. The 1993 birth cohort was found to have a CD prevalence of 3%.[2,17] The two birth cohorts together contain 306 CD patients from which DNA was available. A group of controls, matched for gender and age, was

randomly selected among those with normal levels of CD markers (total of 1,403 healthy controls). Since there was no difference between the frequencies of SNPs in the two cohorts, we treat the 1993 and 1997 cohorts as one collection in our analysis.

The second validation set included 1,244 non-Hispanic, white American children from a prospective population-based cohort from Denver, Colorado, USA; they are being followed annually from birth up to 20 years of age for development of transglutaminase auto-antibodies (TGA) and CD (the DAISY study).[18] This cohort contains two prospective groups of high-risk children: (a) a group of newborns born in 1993 selected from the general population based on the presence of the type 1 diabetes associated HLA-DQ risk allele, and (b) young non-diabetic children with a sibling or parent with type 1 diabetes but unselected for their HLA genotype.



The test set included 985 parents of high-risk CD children (those with a first-degree relative with CD) from the Netherlands, Italy, Poland and Spain, which were collected as part of the PreventCD project.[19] This is an ongoing European study, which aims to develop strategies for preventing CD and other autoimmune diseases by optimizing infant feeding practices.

The derivation, validation 1 and 2, and test datasets were each collected for a different purpose by different investigators and are independent of each other. All samples had Caucasian self-reported ancestry and have been described elsewhere.[15,17–19] CD patients in the discovery, derivation and test sets were diagnosed according to the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) criteria. In the validation 1 set, CD diagnosis required villous atrophy or intraepithelial lymphocytosis in combination with symptoms/signs of celiac disease. In the validation 2 set, CD was defined as having an ever very high level of transglutaminase auto-antibodies or proven by biopsy, and so we refer to this group as celiac disease autoimmunity (CDA).[20]

## Genotyping and statistical analysis

HLA-DQ2 and DQ8 are the major and necessary genetic risk factors for CD development. Several studies have shown that individuals homozygous for HLA-DQ2.5 or HLA-DQ2.5/DQ2.2 genotypes have an increased risk for CD compared with those homozygous for HLA-DQ2.2 or DQ8, or heterozygous for HLA-DQ2.5, DQ2.2 or DQ8, while individuals with no DQ2/DQ8 have practically no risk for CD.[16,21–23] To predict whether an individual has 0, 1 or 2 HLA-DQ2 and/or DQ8 alleles, we genotyped six tagging single-nucleotide polymorphisms (SNPs).[24] We then categorized the individuals into three risk groups: low risk (coded as 0) if they were HLA-DQ2/DQ8 negative (i.e. neither HLA-DQ2.5, DQ2.2 nor DQ8), high risk (coded as 2) for those homozygous for HLA-DQ2.5 or HLA-DQ2.5/DQ2.2, and intermediate risk (coded as 1) for all other combinations.[16]

To assess if the new susceptibility variants improve genetic risk prediction, we compared three genetic

**Table 2.** SNPs and weights used for the three different risk models: GRS_10, GRS_26 and GRS_57.

| Locus | Chr. | GRS_10 | | | GRS_26 | | | GRS_57 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SNP | Risk allele | ß=ln(OR) | SNP | Risk allele | ß=ln(OR) | SNP | Risk allele | ß=ln(OR) |
| *C1orf93, TNFRSF14, MMEL1* | 1 | | | | rs3748816 | A | 0.06217 | rs4445406 | T | 0.07356 |
| *RUNX3* | 1 | | | | rs10903122 | G | 0.08316 | rs72657048 | G | 0.087 |
| *FASLG* | 1 | | | | | | | rs12068671 | T | 0.1605 |
| | | | | | | | | rs859637 | T | 0.1825 |
| *RGS1* | 1 | rs2816316 | A | 0.2142 | rs2816316 | A | 0.2142 | **rs72734930**## | A | 0.3622 |
| | | | | | | | | rs1359062 | G | 0.206 |
| *KIF21B, C1orf106* | 1 | | | | rs296547 | C | 0.1103 | rs10800746 | C | 0.1155 |
| *PUS10* | 2 | | | | rs13003464 | G | 0.09383 | rs13003464 | G | 0.09383 |
| *PLEK, FBX048* | 2 | | | | rs17035378 | T | 0.09479 | rs10167650 | T | 0.1098 |
| *IL18RAP, IL18R1* | 2 | rs917997 | G | 0.2253 | rs917997 | G | 0.2253 | rs990171 | A | 0.2231 |
| *ITGA4, UBE2E3* | 2 | | | | rs4667121 | C | 0.05532 | rs1018326 | C | 0.09917 |
| *STAT4* | 2 | | | | | | | **rs6715106**# | A | 0.3624 |
| | | | | | | | | rs12998748 | G | 0.1372 |
| | | | | | | | | rs6752770 | G | 0.06202 |
| *CTLA4, ICOS, CD28* | 2 | | | | | | | **rs10207814**## | T | 0.1509 |
| | | | | | | | | rs1980422 | C | 0.1568 |
| | | | | | rs4675374 | T | 0.1544 | rs34037980 | A | 0.05851 |
| *CCR4, GLB1* | 3 | | | | rs13314993 | G | 0.1289 | rs4678523 | C | 0.1435 |
| *CCR1-3, LTF* | 3 | rs6441961 | T | 0.1255 | rs13098911 | T | 0.1845 | rs7616215 | C | 0.1031 |
| | | | | | | | | rs2097282 | C | 0.1299 |
| | | | | | | | | **rs60215663**# | A | 0.2788 |
| *ARHGAP1* | 3 | | | | rs11712165 | G | 0.03233 | rs61579022 | A | 0.03245 |
| *SCHIP1, IL12A* | 3 | rs9811792 | C | 0.1556 | | | | rs1353248 | C | 0.1579 |
| | | rs17810546 | G | 0.2642 | rs17810546 | G | 0.2642 | imm_3_161120372 | A | 0.2937 |
| | | | | | | | | rs2561288 | T | 0.1773 |
| *LPP* | 3 | rs1464510 | A | 0.1794 | rs1464510 | A | 0.1794 | rs2030519 | A | 0.2107 |
| *KIAA1109, ADAD1, IL2, IL21* | 4 | | | | | | | **rs62323881**# | A | 0.2048 |
| | | rs6822844 | G | 0.3911 | rs13151961 | A | 0.3658 | rs13132308 | A | 0.3895 |

Chr. chromosome; SNP single nucleotide polymorphism; OR odds ratio. The genes mentioned are the most plausible genes reported. # SNPs with MAF = 5-10%, ## SNPs with MAF <5%.

**Table 2 (continued).** SNPs and weights used for GRS_10 (A), GRS_26 (B) and GRS_57 (C).

| *Locus* | Chr. | GRS_10 | | | GRS_26 | | | GRS_57 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SNP | Risk allele | ß=ln(OR) | SNP | Risk allele | ß=ln(OR) | SNP | Risk allele | ß=ln(OR) |
| *IRF4* | 6 | | | | | | | rs12203592 | C | 0.1242 |
| | | | | | | | | rs1050976 | C | 0.09018 |
| *BACH2* | 6 | | | | rs10806425 | A | 0.1215 | rs7753008 | C | 0.1386 |
| *PTPRK* | 6 | | | | rs802734 | G | 0.1338 | rs55743914 | T | 0.1858 |
| | | | | | | | | rs72975916 | C | 0.2027 |
| *TNFAIP3, OLIG3* | 6 | rs2327832 | G | 0.2037 | rs2327832 | G | 0.2037 | imm_6_138043754 | G | 0.1317 |
| | | | | | | | | rs17264332 | G | 0.2298 |
| *TAGAP* | 6 | rs1738074 | T | 0.1545 | rs1738074 | T | 0.1545 | rs182429 | A | 0.1566 |
| | | | | | | | | **rs1107943**[#] | C | 0.1342 |
| *ELMO1* | 7 | | | | | | | 1kg_7_37384979 | G | 0.1715 |
| *PVT1* | 8 | | | | rs9792269 | A | 0.08532 | rs10808568 | A | 0.08465 |
| *PFKFB3, PRKCQ* | 10 | | | | | | | rs2387397 | C | 0.08623 |
| *ZMIZ1* | 10 | | | | rs1250552 | A | 0.1472 | rs1250552 | A | 0.1472 |
| *POU2AF1, C11orf93* | 11 | | | | | | | rs7104791 | T | 0.1471 |
| *TREH, DDX6* | 11 | | | | | | | rs10892258 | G | 0.1234 |
| *ETS1* | 11 | | | | rs11221332 | T | 0.1378 | rs61907765 | T | 0.1415 |
| *SH2B3, ATXN2* | 12 | rs3184504 | T | 0.1516 | rs3184504 | T | 0.1516 | rs3184504 | T | 0.1516 |
| *ZFP36L1, C14orf181* | 14 | | | | | | | rs11851414 | C | 0.1599 |
| *CLK3, CSK* | 15 | | | | | | | rs1378938 | A | 0.1226 |
| *CIITA* | 16 | | | | | | | rs6498114 | G | 0.1431 |
| *SOCS1, PRM1, PRM2* | 16 | | | | rs12928822 | C | 0.1128 | rs243323 | A | 0.1001 |
| | | | | | | | | **imm_16_11281298**[##] | G | 0.002075 |
| | | | | | | | | rs9673543 | G | 0.003185 |
| *PTPN2* | 18 | | | | rs1893217 | G | 0.2063 | rs11875687 | C | 0.1934 |
| | | | | | | | | **rs62097857**[##] | A | 0.1095 |
| *UBASH3A* | 21 | | | | | | | rs1893592 | A | 0.07879 |
| *ICOSLG* | 21 | | | | rs4819388 | C | 0.1541 | rs58911644 | A | 0.118 |
| *UBE2L3, YDJC* | 22 | | | | | | | rs4821124 | C | 0.1328 |
| *HCFC1, TMEM187, IRAK1* | X | | | | | | | rs13397 | A | 0.08922 |

Chr. chromosome; SNP single nucleotide polymorphism; OR odds ratio. The genes mentioned are the most plausible genes reported. [#] SNPs with MAF = 5-10%, [##] SNPs with MAF <5%.

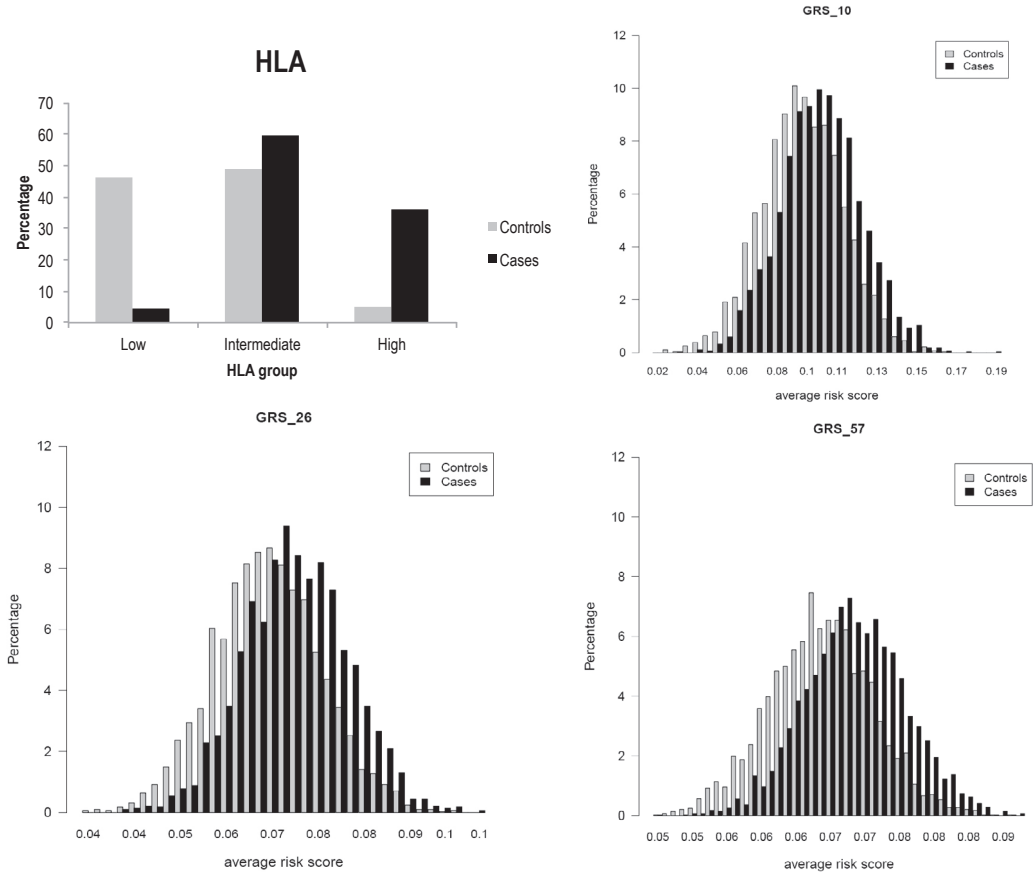



**Figure 1.** Distribution of HLA group and average risk scores of the GRS_10, GRS_26 and GRS_57 models in 2,675 cases and 2,815 controls.

risk scores calculated using: (1) the 10 non-HLA SNPs identified by the first GWAS, (2) the 26 non-HLA SNPs identified by the second GWAS, and (3) the 57 non-HLA SNPs identified by the fine-mapping project (Table 2).[12–15] All these SNPs were reported at genome-wide significance ($p<5x10^{-8}$) in each study.

For the discovery study and derivation set, genotype data were acquired as part of our previous fine-mapping project using Immunochip, a custom-made platform from Illumina.[25] A stringent quality control check was performed on these samples and has been described elsewhere.[15] Samples in validation sets 1 and 2, and in the test sets were genotyped on Illumina 48-plex VeraCode technology for the 26 SNPs identified in the second GWAS and the six HLA tagging SNPs, following the manufacturer's protocol. Genotyping data analysis and clustering was performed in Illumina's GenomeStudio. Genotype clusters were manually investigated and adjusted if necessary. All plates included one duplicate sample and one positive control. One SNP corresponding to *IL18RAP* locus (imm_2_102429801) was not present on VeraCode, so we used a perfect proxy for it (rs917997, $r^2=1$, D'=1) (Table 2).

Using the derivation cohort, we coded each SNP genotype as 0 for the non-risk homozygous genotype, 1 for the heterozygous genotype, and 2 for the homozygous risk genotype, and then determined the inheritance mode (co-dominant, dominant, recessive, overdominant or log-additive) by analyzing the genotypes as categorical variables in logistic regression and adjusting for HLA group, gender and population origin. Comparing the Akaike Information Criterion (AIC) from each model, we saw no major differences between the inheritance models and thus we used the log-additive model, which was the best-fit model for most SNPs. The allelic odds ratio ranged from 1.002 to 1.476 for the individual loci. In order to account for a difference in risk contribution from each SNP, we used a weighted method and calculated an average genetic risk score (GRS) for each individual. First, we multiplied the β-coefficients in Table 2 by the number of risk alleles (0, 1, 2) for each SNP per individual, took the sum across 10, 26 or 57 non-HLA SNPs depending on the model, and then divided the total by the number of alleles included in the model to obtain an average weighted risk score per allele. Only individuals with a defined HLA genotype and with more than 95% of genotypes available were included in the analysis. We used an averaged GRS per allele in order to be able to compare GRS from different datasets with different numbers of SNPs that passed the quality control. Then, the GRS were categorized in quintiles of the control population. The controls in validation set 1 were healthy individuals who had a negative screening result for CD; we used both cases and controls to calculate the quintiles. For validation set 2, we had genotype data from 986 non-Hispanic, white American individuals from the general population, which we used to calculate the quintiles. In each validation set, we estimated the risk for each category of the GRS in a logistic regression using the third quintile (p40-p60) as a reference group and adjusting for HLA group, gender and population group.



To evaluate the overall discrimination of our genetic model, we calculated the area under the receiver-operating characteristic curve (AUC) for only HLA (model 1) and combining HLA and the genetic risk score of non-HLA variants (model 2). To assess whether the GRS would lead to better CD risk prediction than using HLA alone, we calculated the net-reclassification improvement (NRI). This method requires predefined risk categories, so we also used the integrated discrimination improvement (IDI) method that does not have this requirement. A two-tailed p-value less than 0.05 indicated statistical significance. All analyses were performed using PLINK v1.07, the R package PredictABEL, and SPSS version 16.0.[26,27]

# RESULTS

To prevent over-fitting, we calculated an average weighted GRS for the derivation, validation 1 and 2, and test cohorts using the effect sizes obtained from the discovery cohort. Figure 1 shows the distribution of HLA and the three GRS in the large derivation set of 2,675 CD cases and 2,815 controls. The cases are shifted towards

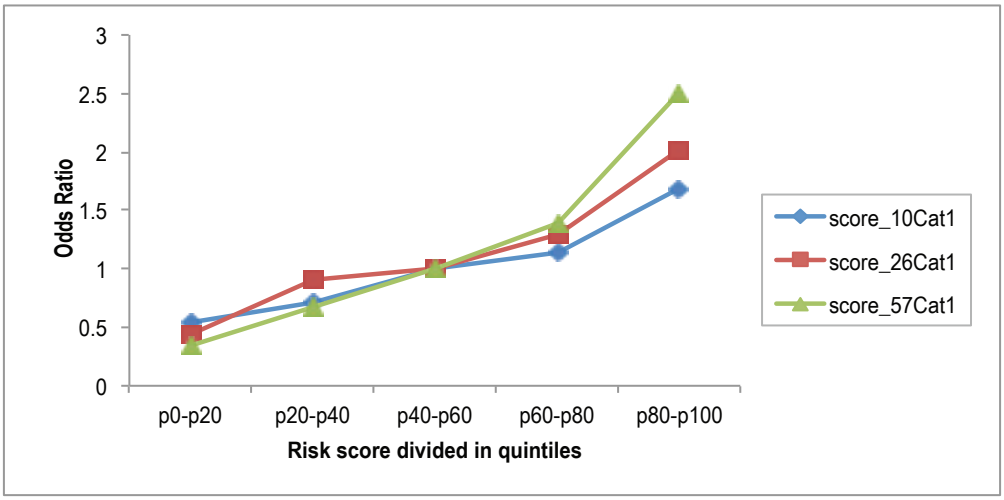



**Figure 2.** Plot of odds ratio for celiac disease associated with quintiles of the GRS_10, GRS_26 and GRS_57 models in a nominal logistic regression, using the third quintile as reference and HLA group, gender and population origin as covariates.

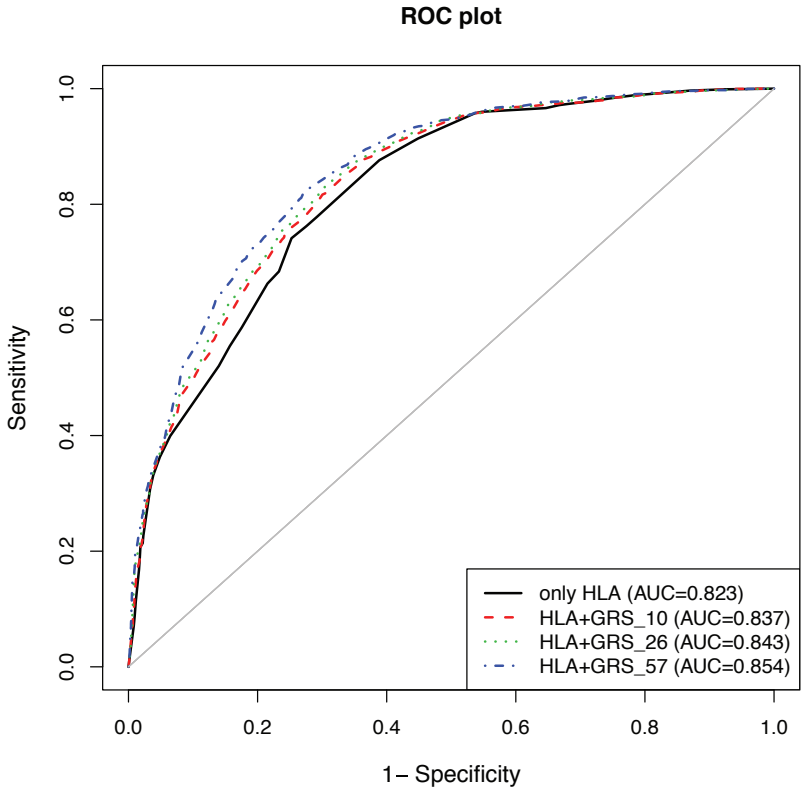



**Figure 3.** Receiver-operator characteristic (ROC) curves and area under the curve (AUC) for the HLA-only model (AUC=0.823), and combined HLA+GRS_10 (AUC=0.837), HLA+GRS_26 (AUC=0.843) and HLA+GRS_57 (AUC=0.854) models.

a higher genetic risk score in all three models. GRS_10, GRS_26 and GRS_57 show a clear separation of distribution between cases and controls with the mean (SD) in cases (0.103 (0.020), 0.071 (0.009), 0.069 (0.006), respectively) being statistically different to the mean (SD) in controls (0.095 (0.020), 0.067 (0.009), 0.066 (0.006), respectively) ($p=2.71x10^{-45}$, $3.41x10^{-67}$, $3.2x10^{-111}$ respectively (independent sample 2-tailed t-test)).

To make it easier to interpret the results of an average weighted genetic risk score, we divided participants into five categories defined as quintiles of the control populations. The third quintile (p40-p60) was used as the reference category (OR=1) since it included the median of the controls, which can be regarded as representative of the mean risk in the general population. Figure 2 shows the increase of OR with increasing risk score for all three GRS models. Interestingly, the genetic risk score with 57 additional variants performs better than GRS_26 and GRS_10 mainly in the top quintile (p80-p100). Individuals in the top quintile of GRS_57 were estimated to have 2.5 times higher risk (95% CI=2.1-3.0) than those with a mean genetic risk score, and 7.2 times higher risk (95% CI=5.7-9.2) than those in the bottom quintile.



Figure 3 shows the receiver-operator characteristic (ROC) curves for HLA only, HLA+GRS_10, HLA+GRS_26 and HLA+GRS_57. The AUC estimates were improved with an increasing number of susceptibility variants used in the model. Combining HLA with 57 non-HLA SNPs showed the best discrimination, with an AUC reaching 0.854 (95% CI=0.844-0.864), while HLA+GRS_10 had an AUC of 0.837 (95% CI=0.827-0.848), and HLA+GRS_26 an AUC of 0.843 (95% CI=0.832-0.853) compared to HLA alone, which had an AUC of 0.823 (95% CI=0.812-0.834). The improvement between only the HLA model and the models with GRS was statistically highly significant, with p-value <<0.0001 for all comparisons.

To confirm that adding non-HLA risk variants to the HLA group improved risk prediction, we tested the ability of the combined HLA and GRS models to reclassify individuals into predefined risk groups based on HLA. The individuals could be grouped into three categories: low (predicted risk <25%), intermediate (predicted risk =25-75%), and high risk (predicted risk >75%) and thus we used the same cut-offs to classify individuals using the models with HLA and non-HLA risk scores (Figure 4). Table 3 shows the reclassification results in the derivation cohort for all three models and the NRI and IDI results (these are measures of discrimination that are commonly used to show the predictive ability of a test). Among the cases, 241 (15.1%) individuals, who had been classified as intermediate risk based on their HLA, were moved into the high risk category (>75%) when their GRS with 57 variants was added to their HLA. Similarly, 25 (18.2%) of the controls who were first classified as high risk (>75%) were moved to the intermediate risk category and 212 (15.4%) were moved from the intermediate to low risk category (<25%). NRI and IDI were statistically significant for all models (Table 3). Even when we used 20% and 80%, or 30% and 70% as the cut-off, the NRI and IDI were still significant. The model with 57 SNPs performed the best by reclassifying 11.1% of the individuals into a more accurate risk

group, while the model with 26 SNPs reclassified 7.1% and that with only 10 SNPs reclassified 4.1%.

To assess if such genetic risk profiling can be transferred to other populations and whether it performs similarly in other study designs, we tested the GRS with 26 SNPs in a nested case-control population from Sweden (validation set 1) and in a prospective cohort from the USA (validation set 2), which had not been assessed in previous gene discoveries.

In validation set 1, the mean 0.068 (SD=0.0099) of GRS_26 in healthy individuals was statistically different from the mean 0.071 (SD=0.0097) of affected individuals (independent sample 2-tailed t-test = $1.28 \times 10^{-5}$). Based on HLA genotypes, we first categorized the individuals into three groups and identified only one CD case in the low-risk group (no HLA-DQ2 or DQ8), indicating the high negative predictive value of HLA typing to exclude CD risk. We further focused our test on 1035 individuals positive for DQ2 and/or DQ8 (intermediate



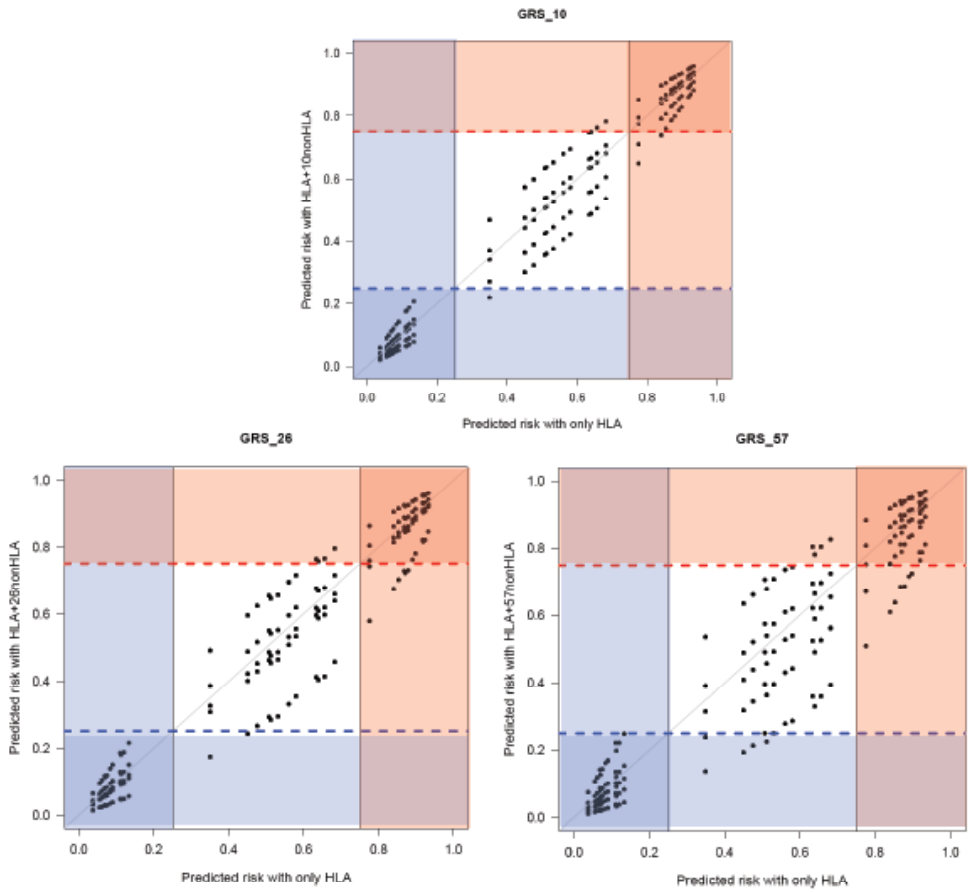



**Figure 4.** Plot of predicted risk using HLA-only model vs. HLA and GRS models showing how individuals can be shifted from one risk group to another. The GRS_57 model shows the number of individuals who were reclassified. All models were adjusted for gender and five population origin. The black vertical line defines the three groups based on HLA (low <25%, intermediate =25-75%, high >75%), while the blue dashed line is the 25% predicted risk and the red dashed line is the 75% predicted risk based on HLA+ non-HLA variants.

**Table 3**. Reclassification table for HLA-only vs. combined HLA and GRS_10, grs_26 and GRS_57

| HLA only | | HLA and GRS_10 | | | | HLA and GRS_26 | | | | HLA and GRS_57 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | < 25% | 25% - 75% | > 75% | Reclassified% | < 25% | 25% - 75% | > 75% | Reclassified% | < 25% | 25% - 75% | > 75% | Reclassified% |
| < 25% | Total | 1419 | 0 | 0 | 0 | 1419 | 0 | 0 | 0 | 1419 | 0 | 0 | 0 |
| | case | 114 | 0 | 0 | 0 | 114 | 0 | 0 | 0 | 114 | 0 | 0 | 0 |
| | Control | 1305 | 0 | 0 | 0 | 1305 | 0 | 0 | 0 | 1305 | 0 | 0 | 0 |
| 25% - 75% | Total | 64 | 2710 | 189 | 0.09 | 104 | 2562 | 297 | 0.14 | 261 | 2389 | 313 | 0.19 |
| | case | 12 | 1444 | 134 | 0.09 | 16 | 1354 | 220 | 0.15 | 49 | 1300 | 241 | 0.18 |
| | Control | 52 | 1266 | 55 | 0.08 | 88 | 1208 | 77 | 0.12 | 212 | 1089 | 72 | 0.21 |
| > 75% | Total | 0 | 39 | 1069 | 0.04 | 0 | 81 | 1027 | 0.07 | 0 | 77 | 1031 | 0.07 |
| | case | 0 | 24 | 947 | 0.02 | 0 | 52 | 919 | 0.05 | 0 | 52 | 919 | 0.05 |
| | Control | 0 | 15 | 122 | 0.11 | 0 | 29 | 108 | 0.21 | 0 | 25 | 112 | 0.18 |
| NRI [95% CI] | | 0.041 [ 0.029 - 0.053 ] ; p-value<< 0.0001 | | | | 0.071 [ 0.055 - 0.087 ] ; p-value<< 0.0001 | | | | 0.111 [ 0.093 - 0.129 ] ; p-value<< 0.0001 | | | |
| IDI [95% CI] | | 0.021 [ 0.018 - 0.025 ] ; p-value<< 0.0001 | | | | 0.031 [ 0.027 - 0.036 ] ; p-value<< 0.0001 | | | | 0.054 [ 0.048 - 0.060 ] ; p-value<< 0.0001 | | | |

NRI net-reclassification index; IDI integrated discrimination improvement; CI confidence interval.

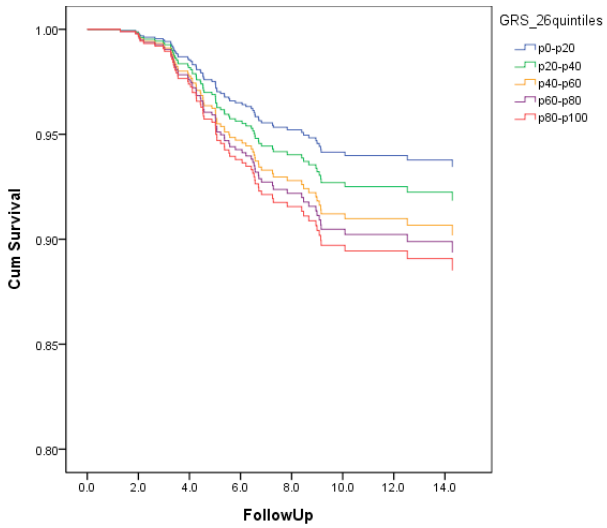



**Figure 5.** Survival analysis using GRS_26 and adjusting for gender, recruitment group and HLA group in DQ2- and DQ8-positive individuals. None of the groups showed statistical significance with the lowest group (p0-p20) taken as reference group.

and high-risk groups based on HLA). The predicted risk based on only HLA ranged from 23.57% to 27.74% for the intermediate HLA group, and 60.95% to 66.02% for the high risk HLA group. Using the lowest ranges as a cut-off for reclassification (23% and 60%), 215/695 (30.9%) of the healthy individuals in the intermediate group (23-60%) were moved to the low risk group (<23%). The NRI of HLA only versus combining HLA and the weighted genetic score of 26 SNPs was 0.116 (95% CI=0.051-0.180; p-value=0.00042), while IDI was 0.013 (95% CI=0.006-0.020; p-value=0.0004) (data not shown).

In the prospective cohort (validation set 2), we categorized individuals based on quintiles calculated from a general population cohort and used the lowest quintile (p0-p20) as a reference group. Based on HLA, there were no CDA cases in the lowest group and thus we continued our analysis with 1,116 individuals who were DQ2 and/or DQ8 positive. Using the COX proportional hazard model adjusted for gender, recruitment group and HLA, we observed an increase in hazard ratio with increasing risk score category (Figure 5). Although this was not statistically significant, it showed a trend of association, with the top group reaching a hazard ratio of 1.8 (95% CI=0.81- 3.98) compared to individuals in the bottom quintile.

To test how well this risk profiling can be used in clinical practice, we calculated a predicted risk for 985 independent individuals (test set) before unraveling their status using the OR calculated in validation set 1 and presented in Table 4. We then grouped the individuals into the low, intermediate and high-risk categories defined earlier. After we checked the CD status of individuals, we compared their classification from using only HLA in the model to using HLA+GRS_26. Combining HLA and 26 non-HLA variants in the model led to 14.6% of the individuals being reclassified in more appropriate categories (Table 5).

# DISCUSSION

Our study demonstrates that by including non-HLA risk variants for CD, the performance of our risk model increases. HLA-DQ testing is well known to have a high negative predictive value for CD, since individuals who are HLA-DQ2 or DQ8 negative have practically no risk of developing the disease. However, the majority of individuals who are positive for these molecules do not develop CD i.e. the HLA-DQ-testing has a poor positive predictive value. By developing a risk profile that combines HLA and non-HLA variants, the diagnostic accuracy of genetic testing for CD is increased. In our previous study, we showed an improved classification with a simple count model of 10 non-HLA variants identified by the first GWAS.[16] In the current study, we have further developed the model by including up to 57 non-HLA SNPs and then compared four genetic risk models for CD including gender and population origin: (1) only HLA, (2) HLA and GRS for 10 non-HLA SNPs from the first GWAS, (3) HLA and GRS for 26 non-HLA SNPs from the second GWAS, and (4) HLA and GRS for 57 non-HLA SNPs from the fine-mapping project. We used a weighted GRS to account for the differences in odds ratio of each allele (this is spread wider with the increasing number of susceptibility variants discovered). All three GRS were associated with CD in our case/control derivation set, with individuals in the top quintile having 1.68, 2.00 and 2.50 times higher risk of CD compared to those in the middle quintile. On the other hand, individuals



**Table 4.** Odds ratio calculated from a logistic regression for models with HLA only and HLA+GRS_26 using the derivation set.

| Parameters | model HLA only OR (95% CI) | model HLA+GRS_26 OR (95% CI) |
|---|---|---|
| (Intercept) | 0.04 (0.03-0.05) | 0.03 (0.03-0.05) |
| Male | 1 | 1 |
| Female | 1.69 (1.48-1.93) | 1.71 (1.49-1.96) |
| UK | 1 | 1 |
| Italy | 2.11 (1.77-2.51) | 1.96 (1.64-2.35) |
| Netherlands | 1.52 (1.27-1.83) | 1.52 (1.26-1.83) |
| Poland | 1.96 (1.53-2.50) | 1.88 (1.46-2.42) |
| Spain1_Basque | 1.91 (1.47-2.48) | 1.96 (1.50-2.55) |
| Spain2_Madrid | 2.37 (1.83-3.07) | 2.343 (1.80-3.05) |
| Low HLA | 1 | 1 |
| Intermediate HLA | 13.92 (11.31-17.13) | 14.07 (11.40-17.37) |
| High HLA | 89.12 (68.18-116.48) | 91.34 (69.56-119.94) |
| p0-p20 (Quintile 1) | NA | 0.44 (0.35-0.55) |
| p20-p40 (Quintile 2) | NA | 0.91 (0.74-1.13) |
| p40-p60 (Quintile 3) | NA | 1 |
| p60-p80 (Quintile 4) | NA | 1.30 (1.06-1.59) |
| p80-p100 (Quintile 5) | NA | 2.00 (1.65-2.43) |

OR of 1 indicates the reference group. OR odds ratio; CI confidence interval.

in the bottom quintiles have 0.54, 0.44, and 0.45 times less risk of developing CD compared to someone with a mean GRS from the general population.

Adding non-HLA variants to the HLA prediction improved not only the discriminatory power as assessed by the ROC curves, but also the reclassification of individuals into more appropriate risk categories with the increase in NRI and IDI. Compared to other genetically complex diseases like multiple sclerosis and type 2 diabetes, where AUC reached 0.769 and 0.74 respectively, GRS in CD performs very well.[28,29] Our best AUC reached 0.854 for the GRS_57 model (HLA and 57 non-HLA risk variants). This is in range to the Framingham Risk Score for coronary heart disease (AUC ~ 0.8), which is considered to be clinically useful.[30] Moreover, our risk model developed from a case-control discovery sample does seem to be applicable in clinical practice and transferable to other populations, mainly in individuals who are positive for HLA-DQ2 and/or DQ8.



Among the 57 SNPs, four variants were rare, with a frequency of less than 5% and four were of low frequency (5-10%). To assess the effect of rare and low frequency variants on GRS, we compared the risk profiling using 49 common SNPs (frequency >10%), 53 common and low frequency SNPs (frequency >5%), and all 57 SNPs. We found no difference between them (data not shown). In addition, it is possible that the GRS with 57 susceptibility variants performs much better than GRS_26 and GRS_10 because the SNPs were identified from a fine-mapping study and are probably better proxies of the causative variants. We calculated a GRS_10iChip and GRS_26iChip using the variants identified by the fine-mapping project but corresponding to the loci found to be associated in the first and second GWAS, respectively. There was, however, no difference between GRS_10 and GRS_10iChip, or between GRS_26 and GRS_26iChip, indicating that the SNPs identified by GWAS are indeed tagging the causative variants (data not shown).

The ultimate aim of genetic studies is two-fold: i) to learn more about the etiology and pathogenesis in order to develop treatment with high efficacy and low toxicity, and ii) to find genetic markers that can be used as diagnostic tool to identify high-risk individuals for early treatment or intervention, if such means are available. There is an ongoing study to evaluate whether early intervention - by introducing small quantities of gluten at four to six months of age - can prevent disease onset in high-risk infants (PreventCD).[19] The design of the study includes all HLA-DQ2 and/or DQ8 children, many of whom will never develop CD as they do not carry the other risk factors required. If the study proves that oral tolerance can be induced by changing infant feeding practices, it will become important to target the group of newborns who will benefit from this intervention. Our risk model will help classify individuals into high- and low-risk groups more accurately by using HLA and other genetic factors. Since only one in seven patients are being properly diagnosed, a more appropriate CD risk model could lead to more efficient strategies to identify potential CD patients early on, and thereby help avoid potentially health complications of the disease. They may even benefit from early intervention preventing the

**Table 5.** Reclassification table for HLA-only vs. combined HLA and GRS_26 in the test set of 985 individuals.

| HLA only | | HLA and GRS_26 | | | |
|---|---|---|---|---|---|
| | | < 25% | 25% - 75% | > 75% | Reclassified% |
| < 25% | Total | 243 | 0 | 0 | 0 |
| | case | 17 | 0 | 0 | 0 |
| | Control | 226 | 0 | 0 | 0 |
| 25% - 75% | Total | 9 | 477 | 78 | 0.15 |
| | case | 0 | 102 | 48 | 0.32 |
| | Control | 9 | 375 | 30 | 0.09 |
| > 75% | Total | 0 | 5 | 173 | 0.03 |
| | case | 0 | 1 | 109 | 0.01 |
| | Control | 0 | 4 | 64 | 0.06 |
| NRI [95% CI] | | 0.146 [0.093 - 0.199]; p-value<< 0.0001 | | | |
| IDI [95% CI] | | 0.025 [0.014 - 0.037]; p-value<< 0.0001 | | | |

NRI net-reclassification index; IDI integrated discrimination improvement; CI confidence interval.



disease onset. Combining HLA and non-HLA variants could be a first step towards identifying high-risk groups in the clinical setting and/or on a population level. However, such an approach carries extensive ethical and practical considerations that need to be well scrutinized before implementation.

**Role of funding source:** The study sponsors had no role in the study design, collection, analysis, or interpretation of the data. The corresponding author had full access to all the data in the study and final responsibility for the decision to submit this manuscript for publication.

**Acknowledgements:** We thank all the patients and controls who participated in this study, Mathieu Platteel and Astrid Maatman for helping in preparing samples and genotyping, and all the doctors who contributed in collecting the DNA. We acknowledge the use of DNA from CEGEC (Spanish Consortium on the Genetics of Celiac Disease). We thank Jackie Senior for critically reading the manuscript. This study was supported by grants from the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW), the Netherlands Organisation for Scientific Research (NWO-VICI grant 918.66.620 to CW), by the European Union-supported project FP6-2005-FOOD-4B-36383-PreventCD and by grant number 5R1DK084568-02 from National Institutes of Health (NIH).

**Conflict of interest:** No conflicts of interest exist.

# Appendix

## The PreventCD study group

Belgium: Association of European Coeliac Societies (AOECS): C Scerri, T Koltai; Croatia: University Children's Hospital Zagreb: S Kolaček, Z Mišak, S Abdović; Germany: Dr. v. Haunersches Kinderspital, Ludwig Maximilians University: S Koletzko, G Osiander, K Werkstetter; Phadia GmbH: E Mummert; Hungary: Coeliac Disease Center of Heim Pál Children's Hospital: IR Korponay-Szabo, J Gyimesi; Israel: M Berant Rambam Health Care Campus, Haifa, Israel; Schneider Children's Medical Center, Sackler faculty of Medicine, Tel-Aviv University: R Shamir, C Hartman; Italy: Eurospital SpA: E Bravi, M Poles; European Laboratory for the Investigation of Food-Induced Diseases (ELFID), University Federico II: R Auricchio, G Limongelli, V Bruno, G Limongelli, R Troncone; Spedali Civili, Brescia: V Villanacci; The Netherlands: Danone Research BV: JG Bindels; Leiden University Medical Center: R Brand, CE Hogen Esch, EG Hopman, YL Kasim Ragab-Volders, F Koning, EMC Kooy-Winkelaar, MC te Marvelde, HB Nguyen, E Stoopman, H Putter; Norway: University of Oslo: LM Sollid, M Ráki; Poland: Akademia Medyczna w Warszawie: A Chmielewska, P Dziechciarz, G Piescik, H Szajweska; Spain: Hospital Universitari de Sant Joan de Reus / Universitat Rovira i Virgili: G. Castillejo, J. Escribano, A. Josa, Instituto de Biomedicina de Valencia (CSIC): A. Capilla and Hospital Sant Joan de Deu Barcelona: V.Varea; La Fe University Hospital: C Ribes-Koninckx, A Lopez; La Paz University Hospital: E Martinez, I Polanco; Sweden: Linköping University: L Högberg, L Stenhammar; Lund University A Carlsson, C Webb; Norrtälje Hospital: L Danielsson, S Hammarroth; Umeå University: O Hernell, D Holmberg, A Hörnell, C Lagerqvist, A Myléus, K Nordyke, F Norström, C Olsson, O Sandström, S Wall; Växjö Hospital: E Karlsson.

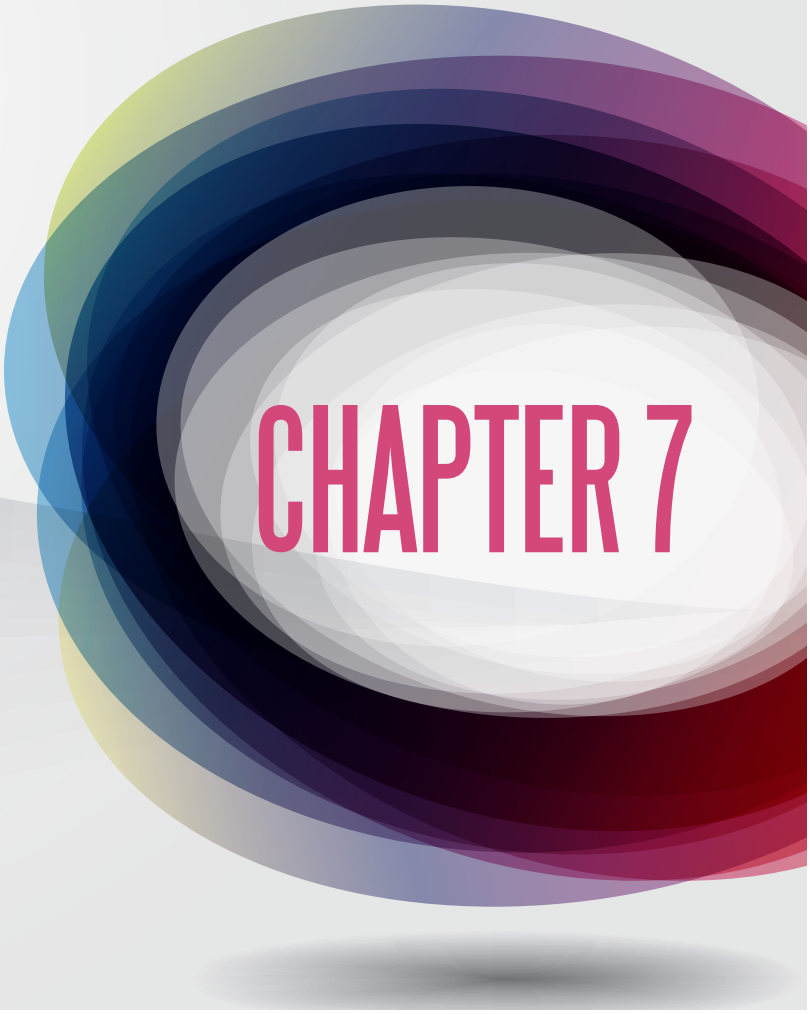
## References


1. *Mearin ML, Ivarsson A, Dickey W. Coeliac disease: is it time for mass screening? Best Pract Res Clin Gastroenterol 2005; 19(3): 441-52.*
2. *Myleus A, Ivarsson A, Webb C, Danielsson L, Hernell O, Hogberg L, et al. Celiac disease revealed in 3% of Swedish 12-year-olds born during an epidemic. J Pediatr Gastroenterol Nutr 2009; 49(2): 170-6.*
3. *Lohi S, Mustalahti K, Kaukinen K, Laurila K, Collin P, Rissanen H, et al. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther 2007; 26(9): 1217-25.*
4. *Rubio-Tapia A, Kyle Ra, Kaplan EL, Johnson DR, Page W, Erdtmann F, et al. Increased prevalence and mortality in undiagnosed celiac disease. Gastroenterology 2009; 137(1): 88-93.*
5. *Tack GJ, Verbeek WHM, Schreurs MWJ, Mulder CJJ. Small-intestinal histopathology and mortality risk in celiac disease. Nat Rev Gastroenterol Hepatol 2010; 7(4): 204-13.*
6. *Walker MM, Talley NJ. Clinical value of duodenal biopsies - Beyond the diagnosis of coeliac disease. Pathol Res Pract 2011; in press.*
7. *Byass P, Kahn K, Ivarsson A. The global burden of childhood coeliac disease: a neglected component of diarrhoeal mortality? PLoS One 2011; 6(7): e22774.*
8. *Sweis R, Pee L, Smith-Laing G. Discrepancies between histology and serology for the diagnosis of coeliac disease in a district general hospital: is this an unrecognised problem in other hospitals? Clin Med 2009; 9(4): 346-8.*
9. *Sanders DS, Hurlstone DP, McAlindon ME, Hadjivassiliou M, Cross SS, Wild G, et al. Antibody negative coeliac disease presenting in elderly people--an easily missed diagnosis. BMJ 2005; 330(7494): 775-6.*
10. *Ribes-Koninckx C, Mearin M, Korponay-Szabo I, Shamir R, Husby S, Ventura A, et al. Coeliac disease diagnosis: espghan 1990 Criteria or need for a change? Results of a questionnaire. J Pediatr Gastroenterol Nutr 2011; in press.*
11. *Hadithi M, von Blomberg BM, Crusius JB, Bloemena E, Kostense PJ, Meijer JW, et al. Accuracy of serologic tests and HLA-DQ typing for diagnosing celiac disease. Ann Intern Med 2007; 147(5): 294-302.*
12. *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(4): 395-402.*
13. *Trynka G, Zhernakova a, Romanos J, Franke L, Hunt Ka, Turner G, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 2009; 58(8): 1078-83.*
14. *Dubois PCA, 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(4): 295-302.*

15. *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; in press.*
16. *Romanos J, van Diemen C,C., Nolte IM, Trynka G, Zhernakova A, Fu J, et al. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology 2009; 137(3): 834,40, 840.e1-3.*
17. *Webb C, Halvarsson B, Norstrom F, Myleus A, Carlsson A, Danielsson L, et al. Accuracy in celiac disease diagnostics by controlling the small-bowel biopsy process. J Pediatr Gastroenterol Nutr 2011; 52(5): 549-53.*
18. *Norris JM, Barriga K, Hoffenberg EJ, Taki I, Miao D, Haas JE, et al. Risk of celiac disease autoimmunity and timing of gluten introduction in the diet of infants at increased risk of disease. JAMA 2005; 293(19): 2343-51.*
19. *Hogen Esch CE, Rosen A, Auricchio R, Romanos J, Chmielewska A, Putter H, et al. The PreventCD Study design: towards new strategies for the prevention of coeliac disease. Eur J Gastroenterol Hepatol 2010; 22(12): 1424-30.*
20. *Liu E, Bao F, Barriga K, Miao D, Yu L, Erlich HA, et al. Fluctuating transglutaminase autoantibodies are related to histologic features of celiac disease. Clin Gastroenterol Hepatol 2003; 1(5): 356-62.*
21. *Dubois PC, van Heel Da. Translational mini-review series on the immunogenetics of gut disease: immunogenetics of coeliac disease. Clin Exp Immunol 2008; 153(2): 162-73.*
22. *Vader W, Stepniak D, Kooy Y, Mearin L, Thompson A, van Rood J,J., et al. The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. Proc Natl Acad Sci U S A 2003; 100(21): 12390-5.*
23. *Congia M, Cucca F, Frau F, Lampis R, Melis L, Clemente MG, et al. A gene dosage effect of the DQA1*0501/DQB1*0201 allelic combination influences the clinical heterogeneity of celiac disease. Hum Immunol 1994; 40(2): 138-42.*
24. *Monsuur AJ, de Bakker PI, Zhernakova A, Pinto D, Verduijn W, Romanos J, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PloS one 2008; 3(5): e2270.*
25. *Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011; 13(1): 101.*
26. *Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81(3): 559-75.*
27. *Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. Eur J Epidemiol 2011; 26(4): 261-4.*
28. *Wang JH, Pappas D, De Jager PL, Pelletier D, de Bakker PI, Kappos L, et al. Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data. Genome Med 2011; 3(1): 3.*
29. *Imamura M, Maeda S. Genetics of type 2 diabetes: the GWAS era and future perspectives. Endocr J 2011; 58(9): 723-39.*

30. *Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998; 97(18): 1837-4.*

# Discussion and Conclusions



# CHAPTER 7

Celiac disease (CD) is one of the well-studied immune-related diseases. The environmental factor contributing to disease susceptibility is gluten, a storage protein found in wheat, barley and rye. This was discovered in the 1950s by Dicke and colleagues who showed that removal of wheat and rye from the diet improved the condition of affected patients.[1] Several years later, the major genetic factor for CD was cloned. It is the human leukocyte antigen (HLA) class II genes, particularly the genes HLA-DQA1 and HLA-DQB1 coding the alpha and beta of HLA-DQ2 and DQ8 heterodimers, respectively.[2, 3] Now, it is well known that HLA-DQ2 and HLA-DQ8 molecules bind to gluten protein and cause an immune reaction, which leads to inflation of the small intestine and destruction of its epithelial cells. These molecules are very common in the general population but luckily not all carriers develop CD, which indicates that other genetic factors must contribute to disease susceptibility.



The prevalence of CD in Western populations has been increasing steadily due to improvement in diagnosis of the atypical and silent CD forms.[4] According to the latest consensus report on CD, small bowel biopsies are considered to be the gold standard and are used to confirm the diagnosis.[5] However, gastroenterologists use serological tests, including tissue transglutaminase (tTGA) and endomysial antibodies (EMA), as a pre-screening tool to minimize the number of patients who have to undergo the invasive biopsy operations. Lately, professionals have been demanding the modification of the diagnostic criteria to avoid biopsy in some cases, for instance, symptomatic children who are HLA-DQ2 and/or DQ8 positive with high EMA or tTGA levels.[5] This shows the first step in using genetic factors in the diagnosis of CD.

In the past five years, two genome-wide association studies (GWAS) and one fine-mapping project had a tremendous success in identifying regions contributing to CD susceptibility.[6-10] They found in total 39 non-HLA loci in addition to HLA. Unlike HLA-DQ2/DQ8, all these non-HLA genetic risk factors have only a slight effect on the susceptibility to CD (OR ranging from 1.09 to 1.7), however, the risk alleles in these loci are common in people of European ancestry, with allele frequencies of 0.1-0.9.

With advances in technology, the focus of researchers has shifted to translating GWAS findings into advances in clinical care. In this thesis, I have presented a general overview on CD and its complications, showed the use of HLA as a diagnostic tool, the difference among populations in the non-HLA associations, and showed that a risk model combining HLA and non-HLA risk variants can be used in diagnosis and has proved to reclassify more than 10% of individuals in more appropriate categories. Although this genomic profiling is not yet being used in clinical practice, it can still be helpful in identifying individuals at high risk for CD.

## Who can benefit from genetic profiling of CD?

CD is often assumed to be a childhood disease, but it has now become apparent that adults also develop CD.[11] Clinical manifestations vary from common gastrointestinal symptoms like diarrhea, abdominal

distension, and failure to thrive, to more silent and irreversible symptoms like anemia, osteoporosis, and infertility. Moreover, untreated CD patients are at increased risk of developing other immune-mediated diseases, such as type 1 diabetes or autoimmune thyroiditis.[12] Biagi and Corazza hypothesized that there is a set quantity of gluten that, once exceeded, complicates CD and thus triggers refractory CD (the irreversible form of CD) and lymphoma.[13] With this hypothesis of "lymphoma triggering amount of gluten", it is thought that if the amount of gluten is exceeded before a diagnosis of CD is made, patients will develop lymphoma and die quickly, regardless of how strictly they follow a gluten-free diet after diagnosis. In order to prevent such severe and life-threatening complications and to reduce chances of mortality, it is important to identify individuals at high-risk early in life. Populations that are at an increased risk of developing CD are mainly first-degree relatives of a CD patient and individuals with immune-related disease. Using genetic profiling in this high-risk group would help to stratify individuals into low, intermediate and high-risk groups and thus diminish the number of individuals who have to undergo serological testing and later a biopsy.



CD was previously considered to be an unavoidable disease in genetically susceptible individuals who consumed gluten-containing food. However, this theory has been proven untrue by several studies which showed that infant nutrition may have a significant effect on the risk of developing CD.[14-16] A study by Norris et al. showed that introducing gluten to babies between the ages of 4-6 months lowered the risk of developing CD.[14] Moreover, data from the Swedish epidemic in the mid-1980s suggest that CD might have been prevented in several individuals if there had been no change made in infant nutrition guidelines.[16, 17] Based on this, the ongoing European CD research project, PreventCD, aims to develop strategies for preventing CD. Their hypothesis is that it is possible to induce a tolerance to gluten by exposing newborns to small quantities of gluten between 4 and 6 months of age.[18] Infants with HLA-DQ2 and/or DQ8 positives and with a first-degree relative with CD are enrolled in the study and blindly randomized to placebo or gluten intervention. The outcome of the study is still unknown but since our group is part of this research project, we calculated a predicted risk using HLA alone and HLA+26 non-HLA risk variants for each newborn and classified them into low, intermediate and high-risk groups (as defined in chapter 6). Figure 1 shows the reclassification of some newborns into new risk groups when adding non-HLA risk variants. Three out of the 24 affected newborns who are classified as intermediate based on their HLA genotypes were reclassified as high-risk when adding non-HLA variants while only one out of 20 affected newborns from high-risk is reclassified as intermediate risk. Follow-up of the rest of newborns is needed to see how well our risk model could predict the disease. We hypothesize that intervention and induction of gluten tolerance would be mainly successful in individuals with intermediate risk and much less in those with high risk, since they already have a high genetic predisposition to developing CD. In a couple of years, the results of the intervention will be clearer and thus genetic profiling might become a tool to screen and identify newborns who can benefit from early intervention to prevent or delay CD development.

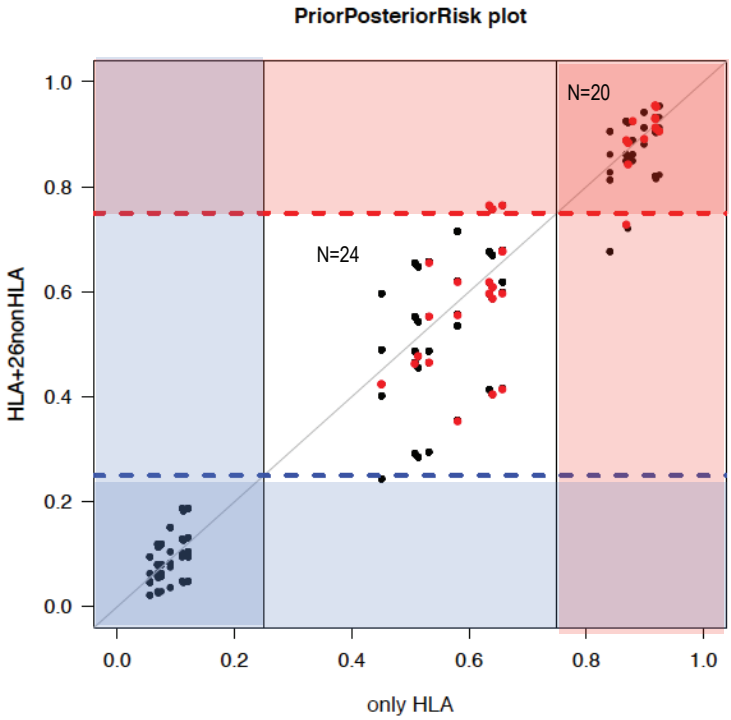



**Figure 1.** Predicted risk plot showing the classification of 1215 newborns in low (blue), intermediate (white) and high (red) risk groups based on HLA alone (x-axis) versus HLA+26 non-HLA variants (y-axis). Red dots indicate the 44 newborns that are currently diagnosed with celiac disease.

# How to improve the accuracy of risk prediction?

Variants identified by GWAS have a low associated risk and explain only a small proportion of familial clustering. CD is one of the rare disorders, in which HLA explains slightly less than half of the genetic heritability (~40%), however the 39 non-HLA loci identified to date explain only another 13.7% together.[6] In order to use genetic profiling widely in clinical practice, we need to improve the accuracy of risk prediction by identifying additional susceptibility variants and by demonstrating its benefits in better medical care (see Box 1).

## Identification of causative variants

GWAS represent an important advance in complex diseases. In the past five years, GWAS have identified over 1200 genome-wide associated loci for 210 traits (NHGRI GWA Catalog, www.genome.gov/GWAStudies). One drawback of these available genotyping arrays is that the vast majority of the associated variants do not cause directly susceptibility to disease by, for example, disrupting the expression or function of a

protein. They are more tagging SNPs, which tag a whole region, and thus it is hard to pinpoint the causal variant or causal gene. Narrowing an implicated locus down to a single causative variant by fine mapping might improve the risk prediction. For autoimmune disorders, a consortium was set up and created an immunochip custom-made platform, which was aimed at fine mapping over 180 established immune-related loci. On average, it has a 12-15x greater marker density than the Hap550 platform (Illumina). It contains re-sequencing information, either from the 1000Genomes project or disease-specific case-control re-sequencing efforts, at 186 loci associated with 12 immune-mediated diseases. Because of this, 34% of Immunochip consists of low/rare frequency variants.[6] In addition to the fine-mapping content, it also consists of variants contributing to the longer tail of disease associations ($p>5x10^{-8}$). All together this makes Immunochip a platform with immunity-focused genetic content with possibly causative variants. Using this platform, we have identified 13 additional loci bringing the total number of susceptibility loci to 39.[6] Using all 57 independent risk variants from 39 non-HLA loci in chapter 6, we saw an improvement in risk prediction (Table 1 and Figure 3). With a simple count of non-HLA risk alleles (resulting in a possible score of 0-114 risk alleles), cases had a median of 55 (mean of average risk score = 0.069) non-HLA risk alleles, while controls had a median of 53 (mean of average risk score = 0.066) (t-test p-value = $1.83x10^{-111}$). Based on HLA genotypes, individuals could be classified into three major groups (low-risk = no-DQ2/DQ8; high-risk = DQ2.5/DQ2.5 and DQ2.5/DQ2.2; or intermediate-risk = all other combinations). Figure 2 shows the frequency distribution of HLA (panel A) and non-HLA risk alleles (panel B) in 2678 cases and 2824 controls from Dutch, Polish, Spanish, Italian and English populations. Compared with the model of 10 SNPs and the model of 26 SNPs, the risk model with more variants showed a better prediction of disease (Table 1, Figure 3 and chapter 6).



**Box 1 |** What measurements are needed to assess the net benefit of a genetic test?

| Measurements | Definition |
|---|---|
| Sensitivity | proportion of people who are correctly classified as high risk among those who will develop the disease |
| Specificity | proportion of people who are correctly classified as low risk among those who will not develop the disease |
| Positive predictive value (PPV) | proportion of people who will develop the disease among those classified as high risk |
| Negative predictive value (NPV) | proportion of people who will not develop the disease among those classified as low risk |
| Absolute risk | probability of developing a disease over a time-period |
| Area under the receiver operating characteristic curve ($A_{ROC}$) | measure of how well a parameter can distinguish between individuals who will develop the disease and those who will not develop the disease |
| Net reclassification index (NRI) | the number of individuals reclassified as high risk |
| Integrated discrimination improvement (IDI) | measure of the separation between people who develop the disease and those who do not in terms of the average |

**Table 1.** Statistics of risk models of only HLA, HLA and genetic risk score with 10 SNPs (GRS_10), 26 SNPs (GRS_26) and 57 SNPs (GRS_57), using the same 2675 cases and 2815 controls

| | Only HLA model | HLA+GRS_10 | HLA+GRS_26 | HLA+GRS_57 |
|---|---|---|---|---|
| Median of the number of risk alleles in controls | - | 9 | 24 | 53 |
| Median of the number of risk alleles in cases | - | 9 | 26 | 55 |
| Mean of the average genetic risk score in controls | - | 0.095 | 0.067 | 0.066 |
| Mean of the average genetic risk score in cases | - | 0.103 | 0.071 | 0.069 |
| T-test for difference between in cases and controls | - | $1.57x10^{-45}$ | $1.74x10^{-67}$ | $1.83x10^{-111}$ |
| Reclassification of intermediate HLA risk in high risk group (cases/controls) | - | 134/55 | 220/77 | 241/72 |
| Sensitivity for high risk group (%) | 36.3 | 40.4 | 42.6 | 43.3 |
| Specificity for high risk group (%) | 95.1 | 93.7 | 93.4 | 93.5 |
| Positive predictive value for high risk group (%) | 87.6 | 85.9 | 86 | 86.3 |
| Negative predictive value for high risk group (%) | 61.1 | 62.3 | 63.1 | 63.5 |
| $A_{ROC}$ [95% confidence interval] | 0.823 [0.812-0.834] | 0.837 [0.827-0.848 ] | 0.843 [0.832-0.853] | 0.854 [0.844-0.864] |
| Net-reclassification improvement (NRI) | - | 0.0409 | 0.071 | 0.111 |
| Integrated discrimination improvement (IDI) | - | 0.0214 | 0.0314 | 0.0543 |

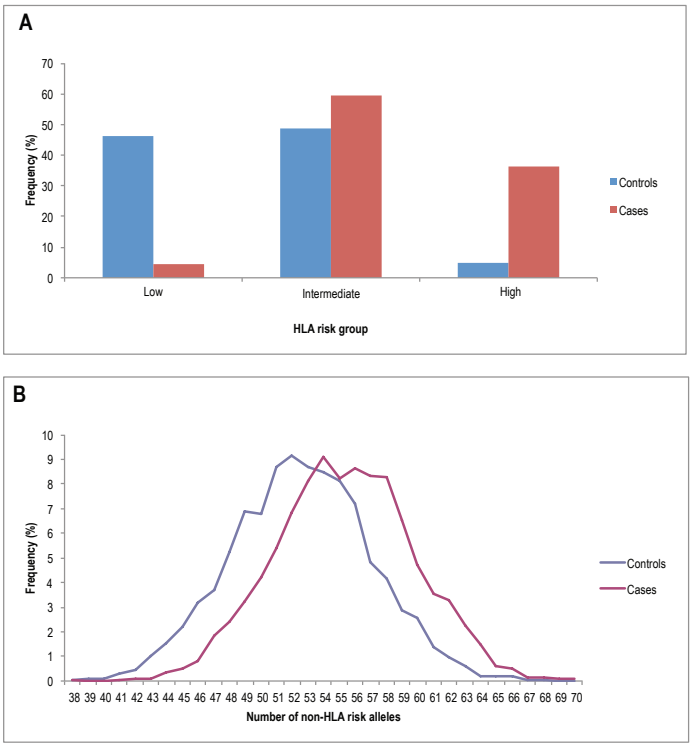



**Figure 2.** Frequency distribution of (A) HLA group and (B) count model using 57 non-HLA loci.



## Heterogeneity among populations

Rare variants (that occur in less than 5% of the general population) tend to be population-specific.[19] Moreover, human genomes are rich in structural diversity and the discovery and genotyping of these variants is not straightforward but has lagged behind those of SNPs. Large variations like deletions, duplication and insertions are individually rare but are collectively frequent, and can pinpoint to the causative genes. The effect sizes of rare variants are usually greater than those observed for common variants.[20] Based on the hypothesis that a population or even individuals might have rare variants causing the disease, a risk model per population might be needed in order to better determine a person's risk of getting a complex disease or even in predicting their response to a particular treatment. We showed in chapter 3 that common variants might not replicate in all populations and thus some loci might be more involved in disease susceptibility in one population than another. It is also possible that the tagSNP that is used on these platforms does not tag the risk variant in that population. This problem is mainly seen in replication studies where they fail to replicate the most associated SNP, especially in non-European populations. Shriner et al. showed that instead of taking the most associated SNP in Europeans and replicating it in non-European populations, it is better to use all SNPs that are in high LD with the top SNP.[21] For CD, we tried to replicate the findings of the 26 European non-HLA loci in 497 celiac

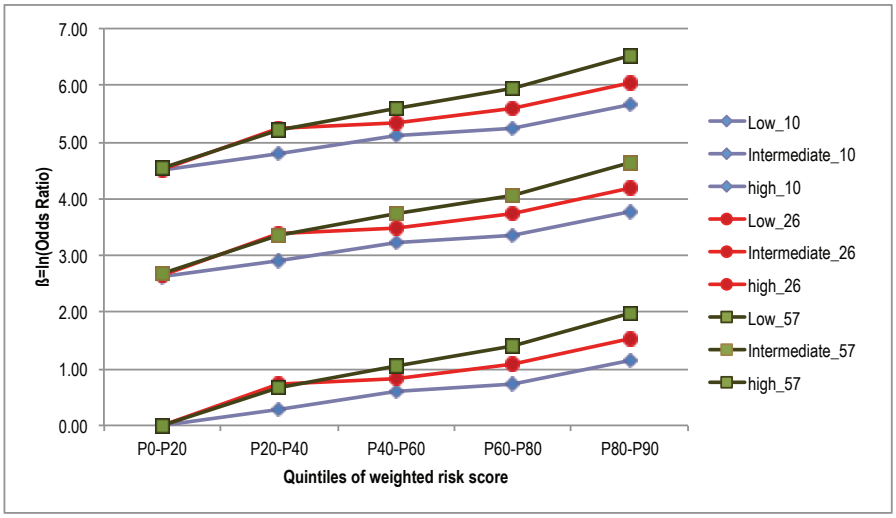





**Figure 3.** Comparison of models with 10, 26 and 57 risk alleles using gender, population origin and HLA groups as covariates.

cases and 736 controls from North India (Senapati et al., submitted). When we performed an 'exact' analysis by directly testing for association of the index SNP, we were able to replicate three loci. Using the transferability method by testing the exact European SNP along with the variants in LD, we were able to transfer 12 loci to the Indian population. Together, 50% of the non-HLA loci showed a significant association in this North Indian population. We applied similar strategies to an inbred population, the Saharawi of North Africa, which has the highest prevalence of CD known so far (5.6%).[22, 23] We found 7 out of 26 European loci to be transferable (26.9%). Testing the association of only top-associated European SNPs replicated 2 out of 26 SNPs, while testing the association of all pruned SNPs ($r^2 < 0.5$) identified two loci to be significantly replicated in the Saharawi after correction for multiple testing per locus (13 loci had at least one SNP with $p<0.05$). In total, 8 out of 26 loci seemed to be replicated in at least one approach. This indicates that GWAS discoveries should not be generalized too easily between different populations. Replicating an association in a non-European population might not identify the causative regions, genes or variants in that population and it is therefore recommended to perform a GWAS for each population, and certainly for those with a high prevalence of the disease, like the Saharawi in the case of CD.[24] Creating a risk profile in one population and applying it in another European population is not yet a major issue with common SNPs, but it will be a problem when we start including rare variants, which are population-specific. Thus, instead of taking the meta-analyzed risk variants and their odds ratios, we would need to have a discovery and a validation cohort from the same population to test the model. This means we will need bigger datasets and larger cohorts.

## Endo-phenotypes analysis

Currently, GWAS of CD had included all individuals with CD, irrespective of differences in disease manifestation or the extent of their intestinal lesions. CD has a variable spectrum of symptoms and thus individuals can be symptomatic (presenting with classical gastrointestinal symptoms), atypical (presenting with non-classical symptoms like anemia and osteoporosis), silent (no symptoms but flat mucosa), latent (no symptoms but positive serology), or have refractory CD (irreversible CD), and in addition there is different age of onset (pediatric and adult), isolated cases, familial cases, or having one or more autoimmune-related diseases to take into account. For example, the majority of patients who have dermatitis herpetiformis (a form of skin disease related to celiac disease) or those who have silent CD (no symptoms present) are likely to have diffuse intestinal lesions ranging from Marsh 0 to 3 (where Marsh 3 is having complete villous atrophy).



Studies have shown that family history in addition to the known genotype of first-degree relatives improves the risk prediction of the index individual. Being from a family with CD, the a priori risk is around 10% compared to 1% for the general population. Taking into account the HLA genotype of the parents, affected sibling, and the index individual, Bourgey et al estimated the risk for a sibling of a CD index patient to range from 0.1% to 29%.[25] Ruderfer et al developed a family-based model for risk prediction where they incorporated the genotype data from both the index individual and a relative of known phenotype.[26] They showed that a lower genetic load of known risk variants in an affected relative tended to increase the index's risk of disease over the level of risk predicted by the index's own genotype, which is because the affected sibling's genotype acts as a surrogate for all the other unmeasured risk factors. For example, if a sibling has the low-risk genotype but is still affected, he or she is likely to have higher rate of other, unobserved risk factors, either genetic or environmental.

One of the populations at particularly high risk for developing CD include individuals with other autoimmune disorders, such as type 1 diabetes (T1D), Addison's disease, and thyroid disease. Approximately 10% of patients with T1D, and 25% of those with the HLA-DQ2/DQ2 genotype, in fact have CD.[27] Inversely, 20% of individuals with CD will have other autoimmune diseases - a risk suggested to increase with the duration of gluten exposure;[28] these include diabetes (5-10%) and thyroid disease (14%) as the most common. GWAS and cross-disease studies have identified the same regions or even the same SNPs as being associated to both T1D and CD.[29] These include the shared association to *HLA, RGS1, TAGAP, IL18RAP, TNFAIP3/OLIG3, SH2B3, CTLA4, CCR, IL2/21, BACH2, PRKCQ* and *PTPN2* loci. We tested the genetic differences between individuals developing both diseases and those having only T1D or only CD. In a pilot study, we genotyped 48 established CD and/or T1D associated SNPs in 1115 non-Hispanic white American (NHWA) T1D patients, in 803 Dutch CD patients (confirmed by biopsy) and in 260 NHWA T1D patients with CD-autoimmunity (CDA&T1D group), defined as persistent tissue-transglutaminase positivity on two consecutive visits. Our analysis showed

that individuals with both diseases accumulated private T1D and CD risk variants. To improve the study, we genotyped 196,524 SNPs present on Immunochip in 257 American and 4 Dutch CDA&T1D patients, and in 1146 Dutch CD patients. Association analysis was performed using logistic regression, adjusting for gender and the first two components from the multidimensional scaling to adjust for the different ethnic backgrounds of our samples. We saw an enrichment of association signals in the autoimmune disease regions, but not for 1753 SNPs associated to bipolar disorder, which are also present on the Immunochip platform. This indicates that the association signals might be true and that our data set showed no stratification. We calculated a platform-specific significance threshold to be $1.9x10^{-6}$ at a 5% false discovery rate and identified two regions reaching this threshold. Another three loci were suggestive at p-value $<1x10^{-5}$. These regions include the *PTPN22*, *CTLA4* and *INS* loci, which are known to be associated to T1D. We realize that the CDA&T1D group was small, so we are now expanding it by adding cases from the Type 1 Diabetes Genetics Consortium samples. We are also performing the same analysis to compare the genetic background of individuals with only T1D to those who develop both T1D and CDA. Despite its limitations, this study suggests that individuals with both T1D and CDA do carry the CD variants as well as private T1D risk variants, but do not have enrichment for those that are shared between both diseases.



## Parent of origin effect

Several rare diseases, such as Prader-Willi syndrome, are related to defects in the imprinting region. The mechanism underlying imprinting is not fully understood but is known to involve epigenetic processes, including DNA methylation and histone acetylation. Recently, Kong et al. showed that variants for cancer and type 2 diabetes confer risk only when inherited from a specific parent.[30] Moreover, they also showed that a variant could either confer risk or reduce the risk of type 2 diabetes depending on the parent of origin. Type 1 diabetes is another complex disease where a parental effect has been reported.[31] A variant in an imprinted region that includes a functional candidate gene DLK1 showed robust evidence for a paternally inherited risk for type 1 diabetes. Although the power to detect such associations is low and these effects might be more prevalent with rare variants, they should not be overlooked. Taking this into account in a risk profiling might well improve the accuracy of risk prediction for offspring.

## Pathway-based association analysis

Single nucleotide markers have limitations since some genes may be involved in disease risk but may not reach a stringent genome-wide significance threshold in any GWAS. To overcome this, other strategies have been developed in recent years: association tests using multiple SNP markers, using imputed genotypes, incorporating linkage information, and pathway-based association approaches.[32] Multiple related genes in the

same functional pathway may work together to confer disease risk, but not all of them will reach a genome-wide significance due to limited power. A good example is the IL-12-IL-23 pathway in Crohn's disease where only three genes showed genome-wide significance and another three were confirmed as susceptibility genes in a replication study.[33] Identifying a susceptible pathway might help in finding drug targets more easily since the most associated gene might not be the best candidate for therapeutic intervention.

# Conclusions

Personalized medicine already exists for monogenetic disorders such as Huntington's disease, phenylketonuria (PKU), and some hereditary forms of cancers. Recent successes in identifying susceptibility variants that underlie many important biomedical phenotypes have increased confidence that this information can be translated into clinically beneficial improvements in medical care. Several companies, like 23andMe, deCODEme, and Navigenics, have begun offering direct-to-consumer testing that uses the SNPs identified as carrying risk by GWAS to predict the risk of an individual of developing a complex disorder. Although the majority of identified risk-marker alleles confer very small risk and have low discriminatory and predictive ability, several researchers argue that some information is better than none at all.



**Box 2 |** What factors can affect an individual's risk prediction?

**Genetic factors**
Parents' genotypes
Affected/unaffected sibling's genotype

**Non-genetic factors**
Gender
Age when gluten introduced into feeding scheme
Amount of gluten consumption
Breast feeding received
Family history
Presence of other immune-related diseases in the individual
Presence of other immune-related diseases in family
Infections (rotavirus)
Ethnicity (e.g. Saharawi with prevalence of 5.6%)
Symptoms
Related diseases (e.g. Turner syndrome, Down syndrome, bipolar disorder)
Mode of delivery at birth

We know now that CD is a complex genetic disorder that involves HLA and non-HLA genes, adaptive and innate immunity, and environmental factors. However, many questions still remain to be answered including: (1) Are we ready to diagnose CD via immune (anti-TTG, EMA) and immunogenetic analyses without confirming the presence of the disease with a biopsy? (2) Do we have the tools for the immunological treatment of CD that would avoid the need for a difficult lifelong gluten-free diet? (3) Are all the different types of CD manifestations of one disease or of several diseases?

Genetic risk profiling for CD is not yet applicable in clinics but it can already help in identifying individuals at high risk, like the first-degree relatives of CD patients or individuals with immune-related diseases, with the aim of lowering the number of individuals needing to undergo a biopsy. Identification of additional susceptibility variants, pathways and causative genes, and non-genetic factors is a key step in improving our understanding of the disease mechanism and in designing effective strategies for risk assessment and targeted treatments (Box 2). Since a substantial proportion of individual differences in disease susceptibility are known to be due to genetic factors, it is important to pinpoint the causative variant, the mode of inheritance, and the interaction with environmental factors. Moreover, GWAS on sub-phenotypes of CD or for using on non-European populations would improve the identification of new variants that have higher penetrance.



With advances in technology, genomic profiling might be the future of personalized medicine for CD. Using one genotyping or sequencing chip, individuals can be easily screened for genetic risk factors early in life and categorized as having a low, intermediate or high risk of developing CD. Depending on their genetic background, they could then benefit from early intervention to prevent CD, be screened regularly for increase in serological markers or be given new drug treatments that would target specific pathways.

# 10 points to remember from this thesis

1. Undiagnosed celiac diseases patients are at increased risk of developing irreversible complications, thus there is a great need to improve the diagnosis and identification of high-risk individuals.
2. Screening for HLA-DQ2 and DQ8 alleles can already help discard celiac disease from a diagnosis.
3. The use of six tag single nucleotide polymorphisms (SNPs) is a sensitive and cheap tool for screening the general population and predicting the presence or absence of one or two copies of HLA-DQ2 and DQ8 alleles.
4. Failing to replicate genome-wide association findings in new populations can be due to the sample size, but also to differences in their genetic backgrounds.
5. The exciting findings of genome-wide association studies have helped refine our model and reclassify some individuals positive for DQ2 and/or DQ8 from an intermediate risk based on their HLA genotypes into a high-risk group.
6. New variants associated to celiac disease have improved the classification of 11% of individuals to more accurate categories.
7. Diagnosis of celiac disease should combine several parameters, starting with HLA screening, calculating the non-HLA genetic risk score, serological typing and finally biopsy testing.
8. Genetic risk profiling for celiac disease would have a higher sensitivity and specificity with the inclusion of more specific genetic factors, such as rare and/or population-specific variants.
9. Genetic testing for celiac disease is not yet suitable for clinical use as it still needs further refining by including non-genetic factors like family history, the presence of other immune-related diseases, time and amount of gluten introduction during weaning, and the duration of breast feeding.
10. One day, newborns will be first screened for HLA and non-HLA variants, categorized into CD risk groups and then treated based on their genetic profile.

## References


1. *van Berge-Henegouwen, ,G.P. & Mulder, C. J. Pioneer in the gluten free diet: Willem-Karel Dicke 1905-1962, over 50 years of gluten free diet. Gut 34, 1473-1475 (1993).*
2. *Sollid, L. M. et al. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J. Exp. Med. 169, 345-50 (1989).*
3. *Spurkland, A., Sollid, L. M., 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. 35, 188-192 (1992).*
4. *Armstrong, M. J., Robins, G. G. & Howdle, P. D. Recent advances in coeliac disease. Curr. Opin. Gastroenterol. 25, 100-109 (2009).*
5. *Ribes-Koninckx, C. et al. Coeliac disease diagnosis: ESPGHAN 1990 criteria or need for a change? Results of a questionnaire. J. Pediatr. Gastroenterol. Nutr. Epub June 28 (2011).*
6. *Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. in press (2011).*
7. *Dubois, P. C. A. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295-302 (2010).*
8. *Trynka, G. et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 58, 1078-83 (2009).*
9. *Hunt, K. A. et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395-402 (2008).*
10. *van Heel, D. A. et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827-9 (2007).*
11. *Vilppula, A. et al. Increasing prevalence and high incidence of celiac disease in elderly people: a population-based study. BMC Gastroenterol. 9, 49 (2009).*
12. *Somers, E. C., Thomas, S. L., Smeeth, L. & Hall, A. J. Autoimmune diseases co-occurring within individuals and within families: a systematic review. Epidemiology 17, 202-217 (2006).*
13. *Biagi, F. & Corazza, G. R. Mortality in celiac disease. Nat. Rev.Gastroenterol. Hepatol. 7, 158-62 (2010).*
14. *Norris, J. M. et al. Risk of celiac disease autoimmunity and timing of gluten introduction in the diet of infants at increased risk of disease. JAMA 293, 2343-2351 (2005).*
15. *Ziegler, A. G., Schmid, S., Huber, D., Hummel, M. & Bonifacio, E. Early infant feeding and risk of developing type 1 diabetes-associated autoantibodies. JAMA 290, 1721-1728 (2003).*
16. *Ivarsson, A. The Swedish epidemic of coeliac disease explored using an epidemiological approach--some lessons to be learnt. Best Pract. Res.Clin. Gastroenterol. 19, 425-40 (2005).*
17. *Ivarsson, A. et al. Epidemic of coeliac disease in Swedish children. Acta Paediatr. 89, 165-71 (2000).*
18. *Hogen Esch, C. E. et al. The PreventCD Study design: towards new strategies for the prevention of coeliac*

*disease. Eur. J. Gastroenterol. Hepatol. 22, 1424-1430 (2010).*

19. *Gravel, S. et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. 108, 11983-11988 (2011).*
20. *Mefford, H. C. & Eichler, E. E. Duplication hotspots, rare genomic disorders, and common disease. Curr. Opin. Genet. Dev. 19, 196-204 (2009).*
21. *Shriner, D. et al. Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4, e8398 (2009).*
22. *Teresi, S. et al. Celiac disease seropositivity in Saharawi children: a follow-up and family study. J. Pediatr. Gastroenterol. Nutr. 50, 506-509 (2010).*
23. *Catassi, C. et al. Why is coeliac disease endemic in the people of the Sahara? Lancet 354, 647-648 (1999).*
24. *Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. Nature 475, 163-165 (2011).*
25. *Bourgey, M. et al. HLA related genetic risk for coeliac disease. Gut 56, 1054-9 (2007).*
26. *Ruderfer, D. M., Korn, J. & Purcell, S. M. Family-based genetic risk prediction of multifactorial disease. Genome Med. 2, 2 (2010).*
27. *Bao, F. et al. One third of HLA DQ2 homozygous patients with type 1 diabetes express celiac disease-associated transglutaminase autoantibodies. J. Autoimmun. 13, 143-8 (1999).*
28. *Mearin, M. L., Ivarsson, A. & Dickey, W. Coeliac disease: is it time for mass screening? Best Pract. Res. Clin. Gastroenterol. 19, 441-52 (2005).*
29. *Smyth, D. J. et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med. 359, 2767-77 (2008).*
30. *Kong, A. et al. Parental origin of sequence variants associated with complex diseases. Nature 462, 868-874 (2009).*
31. *Wallace, C. et al. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat. Genet. 42, 68-71 (2010).*
32. *Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11, 843-854 (2010).*
33. *Barrett, J. C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40, 955-962 (2008).*

# English Summary

# Summary

Around 1 in 100 individuals cannot eat pasta, bread or cookies because they have a condition called celiac disease (CD). CD is caused by one of the most common food intolerances seen in Western populations. It is an immune-related disorder where gluten, a protein present in wheat, barley and rye, causes an immune reaction leading to damage and flattening of the small intestine in genetically predisposed individuals. The main clinical symptoms are diarrhea, abdominal pain, distension, constipation, fatigue and weight loss. However, these symptoms vary widely among patients, with some of them having no symptoms at all. Around 80% of patients are not properly diagnosed and therefore remain untreated, which means they are at high risk of developing irreversible complications like anemia, premature osteoporosis, and unexplained infertility, in addition to being at high risk of developing other immune-related disorders like type 1 diabetes and thyroiditis. The only treatment is a life-long gluten-free diet; this is a safe diet but is not easy to adopt since it is more expensive and socially-restricting.

CD is a complex genetic disease meaning that both environmental and genetic factors play a role in its development. Gluten is the major environmental factor and is the main trigger of the disease in combination with the heterodimers HLA-DQ2 and/or DQ8. These molecules are coded by the genes HLA-DQA1 and HLA-DQB1, which are the major genetic risk factors for CD. More than 95% of affected individuals carry the HLA-DQ2 and/or DQ8 molecules, but around 30-40% of the general population also carry these molecules and never develop CD. This indicates that other genetic factors are needed for the disease to develop.

In this thesis, we focus on describing the genetics of celiac disease, the differences among populations, and how these findings can be translated into clinically useful tests.

In chapter 1, we introduce CD and describe a molecular approach to the diagnosis of this disorder. The chapter describes the history, clinical features, treatment, diagnosis, epidemiology, pathogenesis and the genetic etiology known up to the first genome-wide association study (GWAS) performed in 2007. Until then, only a few regions of the genome had been found to be involved in CD pathogenesis using candidate gene approaches and linkage studies. Unfortunately, these findings were not always found to hold true when tested in different populations, which indicated that there might be genetic differences between populations. Finding the causal gene or causative variants in each region proved to be very difficult since they need to occur frequently enough in the general population, to co-exist in one individual and cause the disease. The only variants for which the function and involvement in CD were known were HLA-DQ2 and HLA-DQ8. By 2007, it was already becoming apparent that genetics could prove useful in diagnosis of complex diseases by identifying individuals at high-risk. For CD, HLA had already been shown to have a high negative predictive value and it could thus be used as a first screening to exclude the disease. The new era of single nucleotide polymorphism tags

(tagSNPs) was helpful in using only a few variants to predict a specific haplotype. The method is fast, easy and needs a very small amount of DNA. Thus, we developed a tagSNP method to predict the most important HLA haplotypes for CD (DQ2.5, DQ2.2, DQ8 and DQ7) in Dutch, English, Spanish and Italian populations. Since it is a prediction method based on linkage disequilibrium, it needs to be tested in each population before being used as a screening tool among high-risk individuals or at the population level. In chapter 2, we describe validating this method in Finnish, Hungarian and two independent Italian populations. The specificity and sensitivity of this test ranges from 95% to 100%. The added value of the test is that it not only indicates the presence or absence of an allele like the traditional-HLA typing methods, but it also predicts the homozygosity and heterozygosity of a CD risk haplotype. This allows us to conduct studies on genetic risk effects and to separate individuals for risk calculations. For example, people who have only one HLA-DQ2 haplotype carry an intermediate risk, while those with two HLA-DQ2 haplotypes carry a much higher risk.

In 2008, an extensive follow-up of the first genome-wide association on CD identified eight new loci that contributed significantly towards CD risk in three independent cohorts from the UK, Netherlands and Ireland. To establish that a genetic region is truly associated to a disease and not found by chance, it is important to replicate the findings in several new populations. In chapter 3, we selected the nine most associated single nucleotide polymorphisms tagging the eight regions identified by the first GWAS and its follow-up, and tested them for association in 538 celiac cases and 593 controls from Italy. Four of the eight loci were found to be significantly associated to CD in the Italian cohort, two more showed moderate association and two had no association. Being from the south of Europe compared to the populations in our initial studies, this result may imply that there is a genuine population difference across Europe regarding the genetic regions contributing to CD. However, a study with larger sample sizes is needed to confirm this.

With all the differences between populations and the small number of susceptibility variants that do not contribute much to disease heritability, several studies have shown that genetic profiling for complex diseases can improve diagnosis, prevention or even treatment of a disease. We introduce this approach in chapter 4, showing the importance of genetic risk profiling in identifying high-risk individuals for CD and reducing the number of individuals who need to undergo serological testing and having a small intestinal biopsy taken to confirm the disease. We envisage a two-step approach to applying genetic knowledge as a diagnostic or screening tool to prevent co-morbidity and long-term complications. First, based on HLA typing, individuals with no HLA-DQ2 or DQ8 can be excluded as they have practically no risk of developing CD. For the rest, by combining their non-HLA variants with their HLA-DQ2 and/or DQ8, individuals can be classified into low (absolute risk < 0.1%), intermediate (absolute risk 0.1% - 7%) or high (absolute risk > 7%) risk groups. Only those in the intermediate and high-risk groups would then undergo serology and biopsy testing. In chapter 5, we describe the first study on risk profiling for CD and how it statistically improves the distinction between

cases and controls. Using the 10 non-HLA variants identified in the first GWAS and its follow-up, we calculated a risk score for each individual by summing the number of risk alleles in 2308 cases and 4585 controls from the Netherlands, UK and Ireland. As expected, we found CD cases carried more non-HLA risk alleles than controls. In addition, individuals with 13 or more risk alleles had a 6.2 times increased risk compared to those carrying fewer than five risk alleles. This was validated in an independent Italian cohort. Combining HLA and non-HLA variants improved the sensitivity of identifying high-risk individuals from 46.6% using only HLA to 49.5%, although the specificity decreased slightly from 93.6% to 92.8%. In the same study, using simulation data we showed that adding risk alleles to the prediction model improved the identification and classification of high-risk individuals. In chapter 6, we showed that this is also true using real genotyping data of 2675 CD cases and 2822 controls from the Netherlands, Italy, Spain, Poland and UK. We compared average weighted genetic risk scores using 10, 26 and 57 variants identified by the first CD GWAS, a second GWAS and a fine-mapping study, respectively. Adding non-HLA variants to risk profiling improves the identification and distinction between cases and control. This is seen by the increase in the area under the receiver-operating cure (AUC), which rose from 82.3% (only–HLA model) to 83.2% (model with 10 variants), 84.3% (model with 26 variants) and 85.4% (model with 57 variants). In addition, the net reclassification improvement (NRI), which is a measure of how much better individuals are re-classified in the correct categories compared to the model with only HLA, improved from 4.1% (model with 10 variants), to 7.1% (model with 26 variants) to 11.1% (model with 57 variants).

The idea of genetic testing is not new. As early as the 1960s, doctors were urging that newborn babies should be tested for rare diseases like phenylketonuria (PKU) that causes mental retardation. PKU can be prevented with a special diet if it is detected early in life. The tests for PKU and other rare but treatable diseases are now performed routinely soon after a baby is born. In chapter 7, we discuss the use of genetic testing for CD. Individuals from families with a first-degree relative affected with CD, or those who have an immune-related disease such as type 1 diabetes, are at high risk of developing CD and thus could be the first group to benefit from genetic screening. In addition, if an intervention treatment was found to be effective in newborns, then genetic profiling could be used to identify the individuals who would benefit from early intervention to prevent or delay CD development. However, the model we propose in this thesis can still be improved by the identification of more rare or population-specific risk variants, pathways and causative genes and by including non-genetic factors such as family history, time and amount of gluten introduction during weaning, and duration of breast feeding.

Finally, genetic profiling might soon be used for other common complex diseases. Here lies the future of personalized medicine. People who learn early in life that they are genetically predisposed to a disease like CD can benefit by knowing what symptoms to look out for and by recognizing the disease in its early stages.

They may also be able to change aspects of their lifestyle and environment, or benefit from early intervention to prevent the disease onset. One day, people will be able to visit their doctors, have a blood sample drawn, and find out more about their health risks for several diseases, including CD. However, before this vision becomes a reality, we have a long way to go and a lot to learn about genes, transcripts, proteins, metabolites, gene-gene interactions and gene-environment interactions.

## 10 points to remember from this thesis

1. Undiagnosed celiac diseases patients are at increased risk of developing irreversible complications, thus there is a great need to improve the diagnosis and identification of high-risk individuals.
2. Screening for HLA-DQ2 and DQ8 alleles can already help discard celiac disease from a diagnosis.
3. The use of six tag single nucleotide polymorphisms (SNPs) is a sensitive and cheap tool for screening the general population and predicting the presence or absence of one or two copies of HLA-DQ2 and DQ8 alleles.
4. Failing to replicate genome-wide association findings in new populations can be due to the sample size, but also to differences in their genetic backgrounds.
5. The exciting findings of genome-wide association studies have helped refine our model and reclassify some individuals positive for DQ2 and/or DQ8 from an intermediate risk based on their HLA genotypes into a high-risk group.
6. New variants associated to celiac disease have improved the classification of 11% of individuals to more accurate categories.
7. Diagnosis of celiac disease should combine several parameters, starting with HLA screening, calculating the non-HLA genetic risk score, serological typing and finally biopsy testing.
8. Genetic risk profiling for celiac disease would have a higher sensitivity and specificity with the inclusion of more specific genetic factors, such as rare and/or population-specific variants.
9. Genetic testing for celiac disease is not yet suitable for clinical use as it still needs further refining by including non-genetic factors like family history, the presence of other immune-related diseases, time and amount of gluten introduction during weaning, and the duration of breast feeding.
10. One day, newborns will be first screened for HLA and non-HLA variants, categorized into CD risk groups and then treated based on their genetic profile.

# Dutch Summary

# Samenvatting

# Samenvatting

Ongeveer 1 op de 100 mensen kan geen pasta, brood of koekjes eten, omdat ze lijden aan de ziekte coeliakie. Coeliakie wordt veroorzaakt door een van de meest voorkomende voedselintoleranties in de westerse wereld. Het is een immuun-gerelateerde aandoening waarbij gluten, een eiwit in tarwe, gerst en rogge, een immuunreactie veroorzaakt die leidt tot schade en afvlakking van de dunne darm in genetisch gepredisponeerde individuen. De belangrijkste klinische symptomen zijn diarree, buikpijn, opgezette buik, constipatie, vermoeidheid en gewichtsverlies. Echter, deze symptomen variëren sterk tussen patiënten, waarbij een aantal van hen helemaal geen symptomen heeft. Ongeveer 80% van de patiënten is niet goed gediagnosticeerd en zal dus onbehandeld blijven; dit betekent dat ze een verhoogd risico hebben op het ontwikkelen van onomkeerbare complicaties zoals bloedarmoede, voortijdige osteoporose, en onverklaarde onvruchtbaarheid. Bovendien hebben ze een verhoogd risico op het ontwikkelen van andere immuun- gerelateerde aandoeningen, zoals type 1 diabetes en thyroïditis. De enige behandeling is een levenslang glutenvrij dieet; dit is een veilig dieet, maar is niet makkelijk om vol te houden, omdat het duurder is en sociaal beperkend.

Coeliakie is een genetisch complexe ziekte, wat betekent dat zowel omgevingsfactoren en genetische factoren een rol spelen in het ontstaan van de ziekte. Gluten is de voornaamste omgevingsfactor en is de belangrijkste aanzet tot de ziekte in combinatie met de heterodimeren HLA-DQ2 en/of DQ8. Deze moleculen worden gecodeerd door de genen HLA-DQA1 en HLA-DQB1, de belangrijkste genetische risicofactoren voor coeliakie. Meer dan 95% van de aangedane individuen draagt de HLA-DQ2 en / of DQ8 moleculen tegenover ongeveer 30-40% van de algemene bevolking die nooit coeliakie ontwikkelt. Dit geeft aan dat andere genetische factoren nodig zijn om de ziekte te ontwikkelen.

In dit proefschrift richten we ons op het beschrijven van de genetica van coeliakie, de verschillen tussen populaties, en hoe deze bevindingen vertaald kunnen worden in klinisch bruikbare testen

In hoofdstuk 1 leiden we coeliakie in en beschrijven we een moleculaire benadering voor de diagnose van deze aandoening. Dit hoofdstuk beschrijft de geschiedenis, de klinische kenmerken, behandeling, diagnose, epidemiologie, pathogenese en de genetische etiologie voor zover bekend tot aan de eerste genoombrede associatie studie (GWAS) uitgevoerd in 2007. Tot dan toe waren met behulp van de kandidaat-gen-benadering en linkage studies slechts enkele regionen van het genoom gevonden die betrokken zijn bij de pathogenese van coeliakie. Helaas werden deze bevindingen niet altijd herhaald in andere populaties, waaruit bleek dat er wellicht genetische verschillen bestaan tussen populaties. Het vinden van het causale gen of de oorzakelijke varianten in elke regio bleek erg moeilijk, omdat ze vaak genoeg moeten voorkomen in de algemene bevolking, naast elkaar moeten bestaan binnen een individu en de ziekte veroorzaken. HLA-DQ2 en HLA-DQ8 waren de enige varianten waarvoor de functie en betrokkenheid bij coeliakie bekend waren. In 2007 werd al duidelijk dat

de genetica nuttig kan zijn bij de diagnose van complexe ziekten door het identificeren van individuen met een hoog risico. HLA heeft een hoge negatief-voorspellende waarde voor coeliakie en kan dus gebruikt worden als een eerste screenings methode om de ziekte uit te sluiten. Door het nieuwe tijdperk van “single nucleotide polymorphism tags” (tagSNPs) hoeven slechts een paar varianten bekend te zijn om een specifiek haplotype te voorspellen. De methode is snel, eenvoudig en behoeft slechts een zeer kleine hoeveelheid DNA. We ontwikkelden een tagSNP methode om de belangrijkste HLA-haplotypes te voorspellen voor coeliakie (DQ2.5, DQ2.2, DQ8 en DQ7) in Nederlandse, Engelse, Spaanse en Italiaanse populaties. Aangezien deze methode gebaseerd is op linkage disequilibrium, moest deze worden getest in elke populatie alvorens te worden gebruikt als een screening hulpmiddel bij hoog-risico individuen of op bevolkingsniveau. In hoofdstuk 2 beschrijven we de validatie van deze methode in een Finse, Hongaarse en twee onafhankelijke Italiaanse populaties. De specificiteit en gevoeligheid van deze test varieert van 95% tot 100%. De toegevoegde waarde van de test is dat het niet alleen de aanwezigheid of afwezigheid van een allel aangeeft, zoals de traditionele HLA-typering methoden, maar ook voorspelt het homozygositeit en heterozygositeit van een coeliakie risico haplotype. Dit stelt ons in staat om studies uit te voeren op de genetische effecten en mensen te onderscheiden op basis van risico berekeningen. Bijvoorbeeld, mensen met slechts een HLA-DQ2 haplotype dragen een intermediair risico, terwijl mensen met twee HLA-DQ2 haplotypes een veel hoger risico dragen.

In 2008 zijn door middel van een extensieve follow-up van de eerste GWAS 8 nieuwe loci geïdentificeerd voor coeliakie in drie onafhankelijke cohorten uit Groot-Brittannië, Nederland en Ierland. Om vast te stellen dat een genetisch gebied werkelijk is gekoppeld aan een ziekte en niet bij toeval is gevonden, is het van belang de bevindingen te repliceren in verscheidene nieuwe populaties. In hoofdstuk 3 hebben we de negen sterkst geassocieerde SNPs geselecteerd die de acht met coeliakie geassocieerde loci taggen en hebben deze getest voor associatie in 538 coeliakie patiënten en 593 controles uit Italië. Vier van de acht loci bleken significant geassocieerd met coeliakie in het Italiaanse cohort, twee hadden een matige associatie en twee hadden geen associatie. Aangezien deze populatie in tegenstelling tot de populatie in de initiële studie uit het zuiden van Europa komt, kan dit betekenen dat er een werkelijk populatie verschil is in Europa ten aanzien van de genetische regio’s die bijdragen aan coeliakie. Er is echter een studie met een grotere steekproefomvang nodig om dit te bevestigen.

Hoewel er verschillen bestaan tussen populaties en er een klein aantal risicovarianten bestaat die weinig bijdragen aan de erfelijkheid van de ziekte, hebben verschillende studies aangetoond dat genetische profielen voor complexe ziekten de diagnose, preventie of zelfs behandeling van een ziekte kunnen verbeteren. Wij introduceren deze aanpak in hoofdstuk 4, waarin we het belang aantonen van genetische risicoprofielen voor het identificeren van hoog-risico individuen voor coeliakie en voor het verminderen van het aantal individuen die serologie testen en een dunne darm biopsie moeten ondergaan om de ziekte vast te

stellen. We stellen een aanpak in twee fasen voor om genetische kennis toe te passen als een diagnostische of screenings methode om co-morbiditeit en complicaties op lange termijn te voorkomen. Ten eerste kunnen op basis van HLA-typering personen zonder HLA-DQ2 of DQ8 worden uitgesloten aangezien deze praktisch hebben geen kans hebben op coeliakie. Verder kunnen mensen door het combineren van hun niet-HLA-varianten met hun HLA-DQ2 en/of DQ8 worden ingedeeld in laag (absoluut risico <0.1%), gemiddeld (absoluut risico 0,1% - 7%) of hoog (absoluut risico op > 7%) risico groepen. Alleen die personen in de gemiddeld en hoog-risicogroep ondergaan serologie en biopsie testen. In hoofdstuk 5 beschrijven we de eerste studie over risicoprofielen voor coeliakie en hoe het statistisch het onderscheid tussen patiënten en controles verbetert. Met behulp van de 10 non-HLA-varianten uit de eerste GWAS en de follow-up hebben we een risico-score voor elk individu berekend door het optellen van het aantal risico allelen in 2308 gevallen en 4585 controles uit Nederland, Groot-Brittannië en Ierland. Zoals verwacht hadden personen met coeliakie meer niet-HLA risico-allelen dan de controlegroep. Daarnaast hadden personen met 13 of meer risico-allelen een 6,2 keer verhoogd risico in vergelijking met degenen die minder dan vijf risico-allelen hadden. We hebben dit vervolgens gevalideerd in een onafhankelijk Italiaans cohort. De combinatie van HLA en niet-HLA-varianten verbeterde de gevoeligheid voor het identificeren van hoog-risico individuen van 46,6% met alleen HLA tot 49,5%, waarbij de specificiteit licht daalde van 93,6% naar 92,8%. In dezelfde studie toonden we met gesimuleerde data aan dat de toevoeging van meer risico-allelen aan het model de identificatie en classificatie van hoog-risico individuen verbeterd. In hoofdstuk 6 laten we zien dat dit ook geldt voor werkelijke genotypering gegevens van 2675 coeliakie patiënten en 2822 controles uit Nederland, Italië, Spanje, Polen en Groot-Brittannië. We vergeleken de gemiddelde gewogen genetisch risico-scores van 10, 26 en 57 varianten resulterend uit respectievelijk de eerste coeliakie GWAS, een tweede GWAS en een fine-mapping studie. Het toevoegen van niet-HLA-varianten aan het risicoprofiel verbetert de identificatie van en het onderscheid tussen de patiënten en controles. Dit blijkt uit de toename van de oppervlakte onder de receiver-operationele curve, die van 82,3% (alleen-HLA-model) steeg tot 83,2% (model met 10 varianten), 84,3% (model met 26 varianten) en 85.4 % (model met 57 varianten). Daarnaast is de netto reclassificatie verbetering (een maat voor hoeveel beter individuen juist worden gecategoriseerd vergeleken met het model met alleen HLA) verbetert van 4,1% (model met 10 varianten), tot 7,1% (model met 26 varianten) tot 11,1% (model met 57 varianten).

Het idee van genetische testen is niet nieuw. Al in 1960 drongen artsen aan op screening van pasgeboren baby's op zeldzame ziekten zoals fenylketonurie (PKU) dat mentale retardatie veroorzaakt. PKU kan voorkomen worden met een speciaal dieet als het vroeg in het leven ontdekt wordt. De testen voor PKU en andere zeldzame, maar behandelbare ziekten, worden nu routinematig snel na de geboorte van een baby uitgevoerd. In hoofdstuk 7 bespreken we het gebruik van genetische testen voor coeliakie. Personen uit gezinnen met een eerstegraads familielid met coeliakie of een andere immuun-gerelateerde ziekte zoals type 1

diabetes, hebben een hoog risico op het ontwikkelen van coeliakie. Deze groep kan profiteren van genetische screening. Wanneer er een effectieve behandeling zal zijn bij pasgeborenen, dan kunnen genetische profielen worden gebruikt om die personen te identificeren die zouden profiteren van zo'n vroege interventie om de ziekte te voorkomen of te vertragen. Echter, het genetische model dat we opgesteld hebben in dit proefschrift moet nog worden verbeterd door de identificatie van meer zeldzame of populatie-specifieke risico-varianten, "pathway" en oorzakelijke genen en door het opnemen van niet-genetische factoren zoals familiegeschiedenis, het tijdspunt en de hoeveelheid gluten bij introductie tijdens het spenen, en duur van de borstvoeding.

Tenslotte zouden genetische profielen binnenkort kunnen worden gebruikt voor andere veel voorkomende complexe ziekten. Hier ligt de toekomst van de gepersonaliseerde geneeskunde. Mensen die al vroeg in hun leven te weten komen dat ze een genetisch verhoogd risico hebben om een ziekte als coeliakie te ontwikkelen kunnen van deze kennis profiteren door te letten op symptomen zodat de ziekte in een vroeg stadium herkend kan worden. Zij kunnen ook aspecten van hun levensstijl en milieu veranderen, of profiteren van vroegtijdige interventie om het begin van de ziekte te voorkomen. Eens zullen mensen naar hun arts gaan, zullen een bloedmonster af laten nemen, en meer te weten komen over hun gezondheidsrisico's voor verschillende ziekten, waaronder coeliakie. Echter, voordat deze visie werkelijkheid wordt, hebben we een lange weg te gaan en hebben we nog veel te leren over genen, gen-transcripten, eiwitten, metabolieten, gen-gen interacties en gen-omgeving interacties.

# 10 punten om te onthouden uit dit proefschrift

1. Niet gediagnosticeerde coeliakie patiënten hebben een verhoogd risico op het ontwikkelen van irreversibele complicaties, daarom is er een grote behoefte om de diagnose en identificatie van hoog-risico individuen te verbeteren.
2. Screening voor HLA-DQ2 en DQ8 allelen kan nu al een coeliakie diagnose helpen uitsluiten.
3. Het gebruik van zes tag SNPs is een gevoelig en goedkoop hulpmiddel voor het screenen van de algemene bevolking en het voorspellen van de aanwezigheid of afwezigheid van een of twee kopieën van HLA-DQ2 en DQ8 allelen.
4. Het niet repliceren van genoombrede associatie bevindingen in nieuwe populaties kan te wijten zijn aan de steekproefgrootte, maar ook aan verschillen in genetische achtergronden.
5. De bevindingen van genoombrede associatie studies hebben ons genetisch risico model verfijnd en maken het mogelijk individuen positief voor HLA DQ2 en/of DQ8 te reclassificeren van een intermediair risico op basis van alleen HLA naar een hoog-risico groep.
6. Nieuwe varianten geassocieerd met coeliakie hebben de classificatie verbeterd van 11% van de personen tot nauwkeurigere risico categorieën.
7. De diagnose van coeliakie zou een combinatie van parameters moeten omvatten, te beginnen met HLA screening, dan het berekenen van de niet-HLA genetisch risico score, vervolgens serologie typeren en uiteindelijk biopsie testen.
8. Genetische risicoprofielen voor coeliakie zullen een hogere gevoeligheid en specificiteit hebben door het opnemen van meer specifieke genetische factoren, zoals zeldzame en/of populatie-specifieke varianten.
9. Genetische testen voor coeliakie zijn nog niet geschikt voor klinisch gebruik, omdat ze verder verfijnd dienen te worden met niet-genetische factoren als familiegeschiedenis, de aanwezigheid van andere immuun-gerelateerde ziekten, het tijdspunt en de hoeveelheid bij introductie van gluten tijdens het spenen, en de duur van de borstvoeding.
10. Eens zullen pasgeborenen eerst worden gescreend voor HLA en niet-HLA-varianten, dan worden onderverdeeld in coeliakie risicogroepen en dan worden behandeld op basis van hun genetisch profiel.

# Portuguese Summary

# Resumo

# Resumo

Cerca de 1 em 100 indivíduos não podem comer pão, macarrão ou biscoitos, pois eles têm uma condição chamada de doença celíaca (DC). DC é causada por uma das intolerâncias alimentares mais comuns observados em populações ocidentais. É uma doença imune-relacionada, onde o glúten, uma proteína presente no trigo, cevada e centeio, provoca uma reação imune, levando a danos e achatamento do intestino delgado em indivíduos geneticamente predispostos. Os principais sintomas clínicos são diarréia, dor abdominal, distensão, constipação, fadiga e perda de peso. No entanto, esses sintomas variam amplamente entre os pacientes, com alguns deles tendo nenhum sintoma. Cerca de 80% dos pacientes não são diagnosticados corretamente e, portanto, permanecem sem tratamento, o que significa que eles estão em alto risco de desenvolver complicações irreversíveis, como anemia, osteoporose precoce e infertilidade sem causa, além de estar em alto risco de desenvolver outras doenças relacionadas ao sistema imunológico, como diabetes tipo 1 e tireoidite. O único tratamento é uma dieta ao longo da vida livre de gluten, esta é uma dieta segura, mas não é fácil de adotar, uma vez que é mais caro e com restrição socialmente .

DC é uma doença genética complexa o que significa que os fatores ambientais e genéticos desempenham papel no seu desenvolvimento. O glúten é o principal fator ambiental e é o principal desencadeador da doença em combinação com os heterodímeros HLA-DQ2 e/ou DQ8 Estas moléculas são codificadas pelos genes HLA-DQA1 e HLA-DQB1, que são os principais fatores de risco genético para DC. Mais de 95% dos indivíduos afetados carregam as moleculas HLA-DQ2 e/ou DQ8, mas em torno de 30-40% da população geral também carregam estas moléculas e nunca desenvolvem DC. Isso indica que outros fatores genéticos são necessários para que a doença se desenvolver.

Nesta tese, nos concentramos em descrever a genética da doença celíaca, as diferenças entre as populações, e como estes resultados podem ser traduzidos em testes clinicamente útil.

No capítulo 1, introduzimos a DC para descrever uma abordagem molecular para o diagnóstico desta doença. O capítulo descreve a história, as características clínicas, tratamento, diagnóstico, epidemiologia, patogênese e etiologia genética conhecida até o primeiro estudo de associação do genoma (GWAS) realizado em 2007. Até então, apenas algumas regiões do genoma havia sido encontrado para ser envolvido na patogênese DC usando abordagens de gene candidato e estudos de ligação. Infelizmente, esses achados nem sempre foram encontradas verdadeiros quando testados em diferentes populações, o que indica que pode ser diferenças genéticas entre as populações. Encontrar o gene causador ou variantes causativas em cada região provou ser muito difícil, pois eles precisam ocorrer com freqüência suficiente na população geral, a co-existir em um indivíduo e a causar a doença. As únicas variantes para o qual a função e o envolvimento em DC eram conhecidos foram HLA-DQ2 e HLA-DQ8. Em 2007, ele já estava se tornando aparente que a genética pode

ser útil no diagnóstico de doenças complexas, identificar indivíduos de alto risco. Para DC, HLA já havia sido demonstrado que têm um alto valor preditivo negativo e pode, portanto, ser usado como uma primeira triagem para excluir a doença. A nova era das tags de polimorfismo de nucleotídeo único (tagSNPs) foi útil em utilizar apenas algumas variantes para prever um haplótipo específico. O método é rápido, fácil e precisa de uma quantidade muito pequena de DNA. Assim, desenvolvemos um método de tagSNPs para prever os haplótipos mais importante para a DC (HLA-DQ2.5, DQ2.2, DQ8 e DQ7) em populações Holandês, Inglês, Espanhol e Italiana. Como esse método e de previsão baseado no desequilíbrio de ligação, o método precisa ser testado em cada população antes de ser usado como uma ferramenta de triagem entre os indivíduos de alto risco ou ao nível da população. No capítulo 2, descrevemos a validação deste método em finlandês, húngaro e dois independentes populações italianas. A especificidade e sensibilidade do teste varia de 95% a 100%. O valor acrescentado do teste é que ele não indica apenas a presença ou ausência de um alelo como o tradicional - métodos de tipagem HLA, mas também prevê a homozigose e heterozigose de um haplótipo de risco. Isso nos permite realizar estudos sobre os efeitos de risco genéticos e separar os indivíduos para cálculos de risco. Por exemplo, pessoas que têm apenas um haplótipo HLA-DQ2 carregam um risco intermediário, enquanto aqueles com dois haplótipos HLA-DQ2 carregam um risco muito maior.

Em 2008, um acompanhamento do primeiro estudo de associação do genoma para DC identificou oito novas regiões que contribuíram significativamente para o risco em três populações independentes do Reino Unido, Holanda e Irlanda. Para estabelecer que uma região genética é realmente associada a uma doença e não encontrados por acaso, é importante replicar os resultados em várias novas populações. No capítulo 3, foram selecionados os nove mais associados polimorfismos de nucleotídeo único dos oito regiões identificadas pela primeira GWAS e o acompanhamento, e testou-os para associação, em 538 casos celíacos e 593 controles da Itália. Quatro dos oito regiões foram encontrados a ser significativamente associadas à DC da população italiana, duas mostraram associação moderada, e duas não apresentaram associação. Sendo a partir do sul da Europa em comparação com as populações em nossos estudos iniciais, esse resultado pode sugerir que há uma diferença de população genuína em toda a Europa em relação às regiões genéticas que contribuem para DC. No entanto, um estudo com maior numero de amostras são necessárias para confirmar isso.

Com todas as diferenças entre as populações e o pequeno número de variantes de susceptibilidade que não contribuem muito para heritabilidade da doença, vários estudos têm demonstrado que a caracterização genética de doenças complexas podem melhorar a prevenção, o diagnóstico ou mesmo o tratamento de uma doença. Introduzimos esta abordagem no capítulo 4, mostrando a importância do perfil de risco genético na identificação de indivíduos de alto risco para DC e reduzindo o número de pessoas que precisam passar por testes sorológicos e ter uma biópsia do intestino delgado tomadas para confirmar a doença. Prevemos uma

abordagem em duas etapas para aplicação do conhecimento genético como uma ferramenta de diagnóstico ou para evitar co-morbidades e complicações a longo prazo. Primeiro, com base na tipagem HLA, indivíduos sem HLA-DQ2 ou DQ8 podem ser excluídos por ter praticamente nenhum risco de desenvolvimento de DC. Quanto ao resto, combinando a sua não-HLA com suas variantes HLA-DQ2 e/ou DQ8, os indivíduos podem ser classificados em baixo (risco absoluto <0,1%), intermediária (risco absoluto de 0,1% - 7%) ou alto (risco absoluto > 7%) grupos de risco. Somente aqueles nos grupos intermediários e de alto risco se submeteriam à sorologia e teste de biópsia. No capítulo 5, descrevemos o primeiro estudo sobre o perfil do risco para DC e como ela melhora estatisticamente a distinção entre casos e controles. Usando os 10 não-HLA variantes identificadas no primeiro GWAS e seu seguimento, foi calculado um escore de risco para cada indivíduo através da soma do número de alelos de risco em 2308 casos e 4585 controles da Holanda, Reino Unido e Irlanda. Como esperado, encontramos casos com DC tem alelos de risco não-HLA mais do que os controles. Além disso, os indivíduos com 13 ou mais alelos de risco teve um 6,2 vezes maior risco em comparação com as que transportam menos de cinco alelos de risco. Este foi validado em uma população independente italiana. Combinando HLA e não-HLA variantes melhorou a sensibilidade de identificar indivíduos de alto risco de 46,6% utilizando apenas HLA para 49,5%, embora a especificidade diminuiu ligeiramente de 93,6% para 92,8%. No mesmo estudo, usando dados de simulação que mostrou que a adição de alelos de risco para o modelo de previsão melhorou a identificação e classificação dos indivíduos de alto risco. No Capítulo 6, mostramos que isso é verdade também usando dados de genotipagem real de 2.675 casos de DC e 2822 controles da Holanda, Itália, Espanha, Polónia e Reino Unido. Foram comparados os escores de risco média ponderada genética utilizando 10, 26 e 57 variantes identificadas pelo primeiro GWAS, o segundo GWAS e o estudo de 'fine-mapping' respectivamente. Acrescentando não-HLA variantes para perfis de risco melhora a identificação e distinção entre casos e controle. Isto é visto pelo aumento na área debaixo a curva de receptor-operacional (AUC), que passou só de 82,3% (modelo com HLA) para 83,2% (modelo com HLA+10 variantes), 84,3% (modelo com HLA+26 variantes) e 85,4 % (modelo com HLA+57 variantes). Além disso, a melhoria reclassificação líquido (net-reclassification index, NRI), que é uma medida da quantos indivíduos são re-classificados nas categorias corretas em relação ao modelo com apenas HLA, melhorou de 4,1% (modelo com HLA+10 variantes), para 7,1% (modelo com HLA+26 variantes) para 11,1% (modelo com HLA+57 variantes).

A idéia dos testes genéticos não é nova. Já em 1960, os médicos estavam pedindo que os bebês recém-nascidos devem ser testados para as doenças raras, como fenilcetonúria (PKU), que causa retardo mental. PKU pode ser prevenida com uma dieta especial, se for detectada precocemente na vida. Os testes para fenilcetonúria e outras doenças raras, mas tratável agora são realizadas rotineiramente logo após o bebê nascer. No capítulo 7, discutimos o uso dos testes genéticos para DC. Indivíduos de famílias com um parente de primeiro grau afetados com DC, ou aqueles que têm uma doença imune-relacionados, tais como diabetes

tipo 1, estão em risco elevado de desenvolver DC e, portanto, poderia ser o primeiro grupo a se beneficiar da avaliação genética. Além disso, se um tratamento de intervenção foi encontrado para ser eficaz em recém-nascidos, em seguida, a caracterização genética poderia ser usada para identificar os indivíduos que se beneficiariam de uma intervenção precoce para prevenir ou retardar o desenvolvimento da DC. No entanto, o modelo que propomos nesta tese ainda pode ser melhorado pela identificação de variantes de risco mais raros ou população específica, os caminhos e os genes causadores e pela inclusão de fatores não-genéticos, tais como histórico familiar, período e quantidade de introdução de glúten durante o desmame , da duração e da amamentação.

Finalmente, o perfil genético pode em breve ser usado para outras doenças comuns complexas. Isto e o futuro da medicina personalizada. Pessoas que aprendem nos primeiros anos de vida que eles são geneticamente predispostas a uma doença como a DC podem se beneficiar por saber quais são os sintomas a procurar e ao reconhecer a doença em seus estágios iniciais. Podem também ser capaz de mudar aspectos de seu estilo de vida e meio ambiente, ou beneficiar de intervenção precoce para prevenir o aparecimento da doença. Um dia, as pessoas vão poder visitar os seus médicos, têm uma amostra de sangue tirada, e descobrir mais sobre os riscos a sua saúde de várias doenças, incluindo DC. No entanto, antes que essa visão se torne uma realidade, temos um longo caminho a percorrer e muito a aprender sobre genes, transcrições, proteínas, metabólitos, interações gene-gene e interações gene-ambiente.

# 10 pontos para se lembrar desta tese

1. Pacientes diagnosticadas como infectada de doenças celiacas  estão em risco aumentado de desenvolver complicações irreversíveis, por isso, há uma grande necessidade de melhorar o diagnóstico e a identificação de indivíduos de alto risco.
2. Triagem para os alelos HLA-DQ2 e DQ8 já pode ajudar a descartar a doença celíaca a partir de um diagnóstico.
3. O uso de seis polimorfismo de nucleotídeo único (tagSNPs) é um instrumento sensível e barato para a triagem da população geral e prever a presença ou ausência de um ou dois alelos  HLA-DQ2 e DQ8.
4. Não reproduzir os achados de associação do genoma inteiro em novas populações pode ser devido ao tamanho da amostra, mas também às diferenças em suas origens genéticas.
5. As descobertas emocionantes dos estudos de associação do genoma inteiro tem ajudado a redefinir o nosso modelo e a re-classificar alguns indivíduos positivos para HLA-DQ2 e/ou DQ8 de risco intermediário baseado em seus genótipos HLA em um grupo de alto risco.
6. Novas variantes associadas à doença celíaca melhoraram a classificação de 11% dos indivíduos com mais categorias precisas.
7. Diagnóstico da doença celíaca deve combinar vários parâmetros, a começar com a triagem do HLA, o cálculo do não-HLA escore de risco genético, tipagem serológica e, finalmente, o teste de biópsia.
8. Perfis de risco genético para a doença celíaca teria uma maior sensibilidade e especificidade com a inclusão de mais específicos fatores genéticos, como raras e/ou população específicas variantes.
9. Testes genéticos para a doença celíaca ainda não é adequado para o uso clínico, mas ainda precisa de refinação incluindo fatores não-genéticos, como história familiar, a presença de outras doenças ligadas à imunidade, período e quantidade de introdução de glúten durante o desmame, e a duração do aleitamento materno.
10. Um dia, os recém-nascidos serão primeiro selecionados para genotipagem do HLA e variantes não HLA, divididos em grupos de risco para DC e, em seguida, tratados com base em seu perfil genético.

# French Summary

# Résumé

Environ 1 sur 100 personnes ne peut pas manger des pâtes, du pain ou des biscuits parce qu'elles souffrent d'une maladie appelée maladie coeliaque (MC). Cette maladie est l'une des intolérances alimentaires les plus communes présentées dans les populations occidentales. La caractéristique de cette maladie auto-immune est que le gluten, une protéine présente dans le blé, l'orge et le seigle, provoque une réaction immunitaire conduisant à des dommages de la paroi de l'intestin grêle chez les individus génétiquement prédisposés. Les principaux symptômes cliniques sont la diarrhée, les douleurs abdominales, ballonnements, constipation, fatigue et perte de poids. Cependant, ces symptômes varient considérablement selon les patients, certains d'entre eux ne présentant pas de symptômes du tout. Environ 80% des patients ne sont pas correctement diagnostiqués et restent donc non traités, ce qui signifie qu'il y a un fort risque qu'ils développent des complications irréversibles telles que l'anémie, l'ostéoporose précoce, l'infertilité inexpliquée, et d'autres maladies auto-immunes, comme le diabète du type 1 et la thyroïdite. Le seul traitement consiste à définitivement exclure le gluten de l'alimentation. C'est un régime faisable mais difficile à adopter, car il est coûteux et restrictif.

MC est une maladie génétique complexe dont le développement est influencé par des facteurs génétiques et environnementaux. Le gluten est le principal facteur environnemental ainsi que le principal déclencheur de la maladie en combinaison avec les molecules HLA-DQ2 et/ou DQ8. Ces molécules sont codées par les gènes *HLA-DQA1* et *HLA-DQB1*, qui sont les principaux facteurs de risque génétiques pour la MC. Plus de 95% des personnes touchées sont porteuses des allèles HLA-DQ2 et/ou DQ8, mais environ 30-40% de la population générale également porteuses ne développent jamais cette maladie. Ceci indique que d'autres facteurs génétiques sont nécessaires au développement de la maladie.

Dans cette thèse, nous décrivons la génétique de la maladie coeliaque, les différences entre les populations, et comment ces résultats peuvent être traduits en tests cliniques utiles. Dans le chapitre 1, nous introduisons la MC et nous essayons de décrire un approche moléculaire pour le diagnostic de ce trouble. Ce chapitre décrit l'histoire, les caractéristiques cliniques, le traitement, le diagnostic, l'épidémiologie, la pathogenèse et l'étiologie génétique connue jusqu'à la première étude d'association sur le génome entier (genome-wide association study - GWAS) réalisée en 2007. Jusque-là, les résultats basés sur des approches gènes-candidats et des études de liaison n'avaient impliqué que quelques régions du génome dans la pathogenèse de la MC. Cependant, ces résultants n'étaient pas toujours reproduisibles dans des populations différentes, ce qui indiquait d'éventuelles différences génétiques entre les populations. Trouver le gène causal ou les variantes causales de chaque région s' est avéré très difficile car ils doivent à la fois se produire assez fréquemment dans la population générale, coexister chez un individu et provoquer la maladie. Les seules variantes dont la fonction et l'implication dans la MC sont connues, sont HLA-DQ2 et HLA-DQ8. En

2007, il devenait apparent que la génétique pouvait être utile afin de diagnostiquer des maladies complexes par l'identification de sujets à haut risque. Dans le cas de MC, HLA a déjà montré une valeur prédictive négative élevée et donc pourrait être utilisé comme premier critère d'exclusion de la maladie. La nouvelle ère de polymorphisme d'un seul nucléotide (tagSNP) a été utiles pour prédire un haplotype spécifique en utilisant seulement quelques variantes. La méthode est rapide, facile et nécessite une très petite quantité d'ADN. Ainsi, nous avons développé une méthode 'tagSNP' afin de prédire les haplotypes du HLA les plus important pour la MC (DQ2.5, DQ2.2, DQ8 et DQ7) dans les populations néerlandaise, anglaise, espagnole et italienne. Cette méthode de prédiction étant basée sur le déséquilibre de liaison, elle doit être testée dans chaque population, avant d'être utilisée comme un outil de dépistage chez les personnes à risque élevé ou au niveau de la population. Dans le chapitre 2, nous décrivons la validation de cette méthode dans les populations finlandaise, hongroise ainsi qu'une deuxième population italiennes indépendantes. La spécificité et la sensibilité de ce test varient de 95% à 100%. La valeur ajoutée de ce test est que non seulement il indique la présence ou l'absence d'un allèle, comme le traditionnel typage du HLA, mais il prévoit également l'homozygotie et l'hétérozygotie d'un risque haplotype de la MC. Cela nous permet de mener des études sur les effets de risques génétiques et de séparer les individus pour les calculs de risques. Par exemple, les personnes qui n'ont qu'un seul haplotype HLA-DQ2 ont un risque intermédiaire, tandis que ceux avec deux haplotypes HLA-DQ2 ont un risque beaucoup plus élevé.

En 2008, un suivi de la première étude d'association génomique sur la MC a identifié huit nouvelles régions contribuant de manière significative au risque de la MC dans trois groupes indépendant provenant du Royaume-Uni, des Pays-Bas et d'Irlande. Pour établir qu'une région génétique est réellement associée à une maladie et non pas découverte par hasard, il est important de reproduire les résultats dans d'autres populations. Dans le chapitre 3, nous avons sélectionné les neuf associés polymorphisme (SNP) des huit régions identifiées par le premier GWAS et son suivit, et les ont testés pour l'association dans 538 cas de coeliaque et 593 contrôles de l'Italie. Quatre des huit régions sont significativement associés a la MC, deux ont une association modérée et deux n'avait aucune association. Ces résultats provenant de populations du sud de l'Europe par rapport aux populations dans nos études initiales, cela peut signifier qu'il y a une réelle différence entre les populations européennes en ce qui concerne les régions génétiques contribuant à la MC. Cependant, une étude avec un plus grand nombre d'individus est nécessaire pour confirmer cette hypothèse.

Avec toutes les différences entre les populations et le petit nombre de variantes qui ne contribuent pas beaucoup à l'héritabilité des maladies, plusieurs études ont montré que le profilage génétique pour les maladies complexes peut améliorer le diagnostic, la prévention ou même le traitement d'une maladie. Nous introduisons cette approche dans le chapitre 4, montrant l'importance du profilage de risque génétique pour identifier les individus à haut risque pour la MC et en réduisant le nombre de personnes qui ont besoin de subir

un test sérologique et une biopsie intestinale pour confirmer la maladie. Nous envisageons une approche en deux étapes qui consiste à utiliser les connaissances génétiques comme outil de diagnostic ou de dépistage afin d'empêcher la comorbidité et les complications à long terme. Dans la première étape, basée sur le typage du HLA, les individus négative pour HLA-DQ2 ou DQ8 peuvent être exclus, car ils présentent un risque pratiquement nulle de développer la MC. Pour le reste, en combinant leurs variantes non-HLA avec leurs HLA-DQ2 et/ou DQ8, les individus peuvent être classés en groupes à risque faible (risque absolu <0,1%), intermédiaire (risque absolu de 0,1% - 7%) ou élevé (risque absolu > 7%). Seuls ceux dans les groupes intermédiaires et à haut risque seraient alors soumis à des tests sérologiques et une biopsie. Dans le chapitre 5, nous décrivons la première étude sur le profilage de risque pour la MC et la façon dont elle améliore statistiquement la distinction entre les cas et les contrôles. En utilisant les 10 non-HLA variantes identifiées dans le premier GWAS et son suivi, nous avons calculé un score de risque pour chaque individu en additionnant le nombre d'allèles de risque dans 2308 cas et 4585 contrôles provenant des Pays-Bas, Royaume-Uni et l'Irlande. Comme prévu, les patients avaient en moyenne plus d'allèles de risque que les contrôles. En outre, les personnes ayant 13 ou plus d'allèles de risque avaient un risque 6,2 fois plus élevé par rapport à ceux qui avaient moins de cinq allèles de risque. Cela a été validé dans un groupe indépendant de sujets italiens. Combiner des variantes HLA et non HLA améliore la sensibilité de l'identification des individus à haut risque de 46,6% à 49,5% comparé à l'utilisation unique du HLA, bien que la spécificité diminue légèrement, passant de 93,6% à 92,8%. Dans la même étude, en utilisant des données de simulation, nous avons montré que l'ajout d'allèles de risque dans le modèle améliore l'identification et la classification des individus à haut risque. Dans le chapitre 6, nous avons montré que cela est également vrai en utilisant les données de génotype réel de 2675 patients et 2822 contrôles des Pays-Bas, d'Italie, d'Espagne, de Pologne et du Royaume-Uni. Nous avons comparé les moyennes pondérées des scores de risque génétique de 10, 26 et 57 variantes identifiées par le premier GWAS, le deuxième GWAS et 'fine-mapping' respectivement. L'ajout de variantes non-HLA pour un profil de risque améliore l'identification et la distinction entre patients et contrôles. Cela est démontré par l'augmentation de l'espace sous la caractéristique de fonctionnement du récepteur (AUC), qui passe de 82,3% (modèle avec HLA seulement) à 83,2% (modèle avec HLA+10 variantes), 84,3% (modèle avec HLA+26 variantes) et 85,4 % (modèle avec HLA+57 variantes). En outre, l'indice d'amélioration de reclassement net (net-reclassification index, NRI), qui mesure le nombre de personnes qui sont mieux reclassées dans les catégories correctes par rapport au modèle avec seulement HLA, augmente, passant de 4,1% (modèle avec HLA+10 variantes), à 7,1% (modèle avec HLA+26 variantes) à 11,1% (modèle avec HLA+57 variantes).

L'idée d'un test génétique n'est pas nouvelle. Dès les années 1960, les médecins demandaient que les nouveau-nés soient testés pour des maladies rares comme la phénylcétonurie (PCU) qui entraîne à un retard mental. PCU peut être évitée avec un régime alimentaire spécial si elle est détectée tôt dans la vie. Les

tests de la phénylcétonurie et d'autres maladies rares mais traitables sont maintenant effectués régulièrement, peu après la naissance du bébé. Au chapitre 7, nous discutons de l'utilisation des tests génétiques pour la MC. Les individus issus de familles avec un parent affecté au premier degré par la MC, ou ceux qui ont une maladie auto-immune, tels que diabète du type 1, présentent un risque élevé de développer la MC et pourraient donc être le premier groupe à bénéficier du dépistage génétique. En outre, si un traitement d'intervention s'avère être efficace chez les nouveaux nés, le profilage génétique pourrait être utilisé pour identifier les personnes qui pourraient bénéficier de cette intervention précoce pour prévenir ou retarder le développement de la MC. Cependant, le modèle que nous proposons dans cette thèse peut encore être améliorée par l'identification de variantes de risque plus rares ou spécifiques aux populations, les voies et les gènes en cause et en incluant des facteurs non génétiques tels que les antécédents familiaux, le temps et la quantité d'introduction du gluten lors du sevrage, et la durée de l'allaitement maternel.

Finalement, le profilage génétique pourrait bientôt être utilisé pour d'autres maladies complexes. C'est là que réside l'avenir de la médecine personnalisée. Les personnes qui apprennent au début de leur vie qu'ils sont génétiquement prédisposés à une maladie comme la MC peuvent bénéficier en sachant quels sont les symptômes à surveiller et en reconnaissant la maladie à ses débuts. Ils peuvent également être en mesure de modifier des aspects de leur mode de vie et de l'environnement, ou bénéficier d'une intervention précoce pour prévenir l'apparition de la maladie. Un jour, les gens seront en mesure de consulter leur médecin, avoir un échantillon de sang, et en savoir plus sur leurs risques de santé liés à plusieurs maladies, y compris la MC. Toutefois, avant que cette vision ne devienne réalité, nous avons un long chemin à parcourir et beaucoup à apprendre sur les gènes, transcrits, protéines, métabolites, interactions gène-gène et interactions gène-environnement.

# 10 points à retenir de cette thèse

1. Non diagnostiqués, les patients atteints de maladies cœliaque ont un risque augmenté de développer des complications irréversibles, pour cela il est nécessaire d'améliorer le diagnostic et l'identification des individus à haut risque.
2. Le dépistage des allèles HLA-DQ2 et DQ8 peut déjà aider à éliminer la maladie cœliaque à partir d'un diagnostic.
3. L'utilisation de six polymorphisme simple nucléotide (SNP) est un outil sensible et à bon marché pour le dépistage de HLA dans la population générale et pour prédire la présence ou l'absence d'un ou deux copies des allèles HLA-DQ2 et/ou DQ8.
4. L'échec de la reproduction des résultats d'association sur le génome entier dans nouvelles populations peut être dû non seulement au nombre d'individu inclus, mais aussi aux différences génétiques.
5. Les résultats passionnants des études de l'association du génome entier ont aidé à affiner notre modèle et à reclasser certaines personnes positives pour HLA-DQ2 et/ou DQ8 d'un risque intermédiaire en fonction de leur génotype HLA dans un groupe à haut risque.
6. De nouvelles variantes associées à la maladie coeliaque ont amélioré la classification de 11% des individus dans des catégories plus précises.
7. Le diagnostic de maladie coeliaque doit combiner plusieurs paramètres, à commencer par le dépistage HLA, le calcul du score de non-HLA variantes de risque génétiques, typage sérologique et finalement le test de biopsie.
8. Le profil de risque génétique pour la maladie coeliaque aurait une plus grande sensibilité et spécificité avec l'inclusion de facteurs génétiques plus spécifiques, comme les variantes rares et / ou spécifiques aux populations.
9. Les testes génétiques de la maladie coeliaque ne sont pas encore approprié pour l'utilisation clinique car il y a encore besoin d'études qui incluent des facteurs non génétiques, comme les antécédents familiaux, la présence d'autres maladies auto-immunes, la période et la quantité d'introduction du gluten, et la durée de l'allaitement maternel.
10. Un jour, les nouveaux nés seront d'abord examinés pour les variantes HLA et non HLA, classées en groupes à risque pour la maladie coeliaque et ensuite traitée en fonction de leur profil génétique.

# Arabic Summary

# ملخص

حوالي واحد الى مئة شخص لا يستطيعون تناول الباستا، الخبز أو الكعكة المحلاة لأنهم يعانون من وضع يسمى مرض التجويفي البطني وهو كناية عن اضطرابات هضمية بسبب الحساسية المفرطة في الغذاء وهو الأكثر شيوعا لدى معظم السكان الغربيين. انه اضطراب يتعلق بالمناعة المتصلة بالغلوتين وهو بروتيين موجود في القمح، الشعير والجاودار، ويسبب ردة فعل جهاز المناعة التي تؤدي الى اضرار وتسطيح في الامعاء الدقيقة لدى الأفراد المعرضين جينيا. العوارض السريرية الرئيسة هي الاسهال، ألام البطن، الانتفاخ، الامساك، التعب وفقدان الوزن. غير أن هذه الأعراض تتفاوت على نطاق واسع بين المرضى. بعضهم لا يشكو من أي عوارض اطلاقا. حوالي 80 ٪ من المرضى لا يمكن تشخيصهم بصورة صحيحة وبالتالي يظلون بدون علاج وهذا يعني أنهم في خطر كبير من تطور المضاعفات لا يمكن تداركها: كفقر الدم، ترقق العظم المبكر والعقم غير المبرر، علاوة الى الخطر الكبير في نمو أمراض مناعية أخرى مثل السكري نوع 1 والغدة الدرقية. العلاج الوحيد هو اتباع حمية/نظام غذائي بدون غلوتين على مدى الحياة؛ هي حمية آمنة ولكن ليس من السهل تبنيها لأنها أكثر كلفة وغير متاحة اجتماعيا. المرض التجويفي البطني هو مرض جيني مركب يعني أن العوامل الوراثية والجينية يلعبان دورا في تطوره. الغلوتين هو العامل الأساسي البيئي والمحفز الأساسي للمرض في تركيبته مع الهيتروداميرز ( HLA-DQ2/DQ8) التي يتم ترميزها بواسطة جينات HLA-DQA1 وHLA-DQB1 ، والتي هي أهم عوامل الخطر الجينية للمرض التجويفي البطني. أكثر من 95 ٪ من المصابين يحملون HLA-DQ2/DQ8 ، ولكن حوالي 40-30 ٪ من السكان عموما ، يحملون كذلك هذه الجزيئات و لا يطورون أبدا المرض التجويفي البطني. يشير ذلك الى أن هناك عوامل وراثية أخرى للمرض كي يتطور.

نركز في هذه الأطروحة على وصف الجينيات في المرض التجويفي البطني ، والاختلافات بين السكان ، وكيف يمكن ترجمة هذه النتائج الى اختبارات مفيدة سريريا. في الفصل الأول نقدم المرض التجويفي البطني ونصف مقاربة الجزيئية بتشخيص هذا الاضطراب. يصف هذا الفصل التاريخ ، والمظاهر السريرية، العلاج، التشخيص ، علم الأوبئة ، والمسببات المرضية الوراثية والجينية المعروفة بدراسة مشتركة ضمن نطاق الجينوم (GWAS) التي أجريت حتى عام 2007، حتى ذلك الحين ، لم يتم العثور إلا على بضع من مناطق الجينوم متدخلة في مرض التجويف البطني باستخدام مقاربات ترشيحية ودراسات ربط، للأسف ، لم تكن هذه النتائج التي وجدت دائما صحيحة عند اختبارها في مجموعات سكانية مختلفة، مما يدل على أنه قد يكون هناك اختلافات جينية بين السكان. ثبت أنه من الصعب جدا العثور على الجينة المسببة للمرض أو المتغيرات في كل منطقة لأنها للضروري أن تحدث المتغيرات بشكل متكرر لدى السكان بشكل عام ، للتواجد في فرد واحد وتسبب المرض. المتغيرات المعروفة في الوظيفة والتدخل في المرض التجويفي البطني قد عرفت في HLA- DQ2 و HLA-DQ8. بحلول عام 2007 ، أصبح من الواضح أن الوراثة يمكن أن تكون مفيدة في تشخيص الأمراض المعقدة عن طريق تحديد الأفراد المعرضين لمخاطر عالية. للمرض التجويفي البطني، تبين أن HLA قيمة تنبؤية سلبية وبالتالي يمكن أن يستخدم كفرز أولي لاستبعاد المرض. كانت الحقبة الجديدة من علامات تعدد أشكال النوكليوتيدات (SNPs) مفيدة في استخدام عدد قليل فقط من المتغيرات للتنبؤ بنمط فردي محدد. الطريقة سريعة وسهلة وتحتاج الى كمية صغيرة جدا من الحمض النووي.وهكذا، طورنا طريقة لتنبؤ HLA بأهم النسخ المتنوعة للمرض التجويفي البطني (DQ7 ، DQ8 ، DQ2.2 ، DQ2.5) لدى السكان الهولنديين، الانكليز، الاسبان والايطاليين. نظرا لأنه أسلوب تنبؤ مبني على أساس ربط الخلل، يحتاج الى اختباره في كل مجموعة من السكان قبل أن يتم استخدامه كأداة فحص عالية المخاطر بين الأفراد أو على مستوى السكان. في الفصل الثاني نصف التحقق من صحة هذا الأسلوب لدى السكان في فنلندا، هنغاريا والايطاليين. تتراوح خصوصية وحساسية هذا الاختبار بين 95 ٪ الى 100 ٪ والقيمة المضافة لهذا الاختبار أنه لا يشير فقط الى وجود أو عدم وجود جينة تقليدية – traditional HLA، لكنها تتوقع أيضا لقاحية متجانسة ولقاحية غير متجانسة في المرض التجويفي البطني. وهذا يسمح لنا باجراء دراسات عن آثار المخاطر الجينية للأفراد وحسابات منفصلة للخطر. على سبيل المثال ، الأشخاص الذين لديهم واحد فقط HLA - DQ2 والذين يحملون مخاطر متوسطة ، في حين أن أولئك مع اثنين من HLADQ2 يحملون مخاطر أعلى من تلك بكثير.

عام 2008، جرت ابحاث أولى على مستوى الجينوم على مرض التجويفي البطني حددت ثمان مواضع جديدة ساهمت بصورة فعّالة في اتجاه تحديد مخاطر المرض التجويفي البطني في ثلاثة محاور مستقلة من المملكة المتحدة، هولندا وايرلندا، لإثبات ارتباطها حقا بالمنطقة الجينية ولم توجد عن طريق الصدفة. من المهم تكرار هذه

النتائج مع لسكان جدد. في الفصل 3، اخترنا تسعة SNP الأكثر ارتباطا بأشكال متعددة من تركيبة واحدة متخذين هدفا المناطق الثمانية التي حددها GWAS الأول ومتابعتها واختبارها لاشتراكها ب 538 حالة من المرض التجويفي البطني و593 دون المرض من إيطاليا. 4 من المناطق الثمانية بدت مرتبطة الى حد كبير بالمرض التجويفي البطني في المجموعة الإيطالية. أظهر اثنان آخران ارتباطا معتدلا واثنان لم يكن لديهما أي ارتباط. حيث أنه من جنوب أوروبا مقارنة مع السكان في دراستنا الأولية . هذا يعني أنه بالنتيجة هناك فرق حقيقي بين السكان في مختلف أنحاء أوروبا في ما يتعلق بالمناطق الجينية التي تساهم في المرض التجويفي البطني. ومع ذلك ، هناك حاجة إلى دراسة أكبر حجما لتأكيد ذلك. مع كل الاختلافات بين السكان والعدد القليل من المتغيرات القابلة الحصول والتي لا تساهم كثيرا في توريث المرض. أظهرت دراسات عدة أن التنميط الجيني للأمراض المعقدة يمكن أن يحسن عملية التشخيص والوقاية أو حتى العلاج من المرض. نقدم تلك المقاربة في الفصل 4، لاظهار أهمية التنميط الجيني في تحديد المخاطر العالية لدى الأفراد من المرض التجويفي البطني وتخفيض عدد الأفراد الذين يحتاجون للخضوع لفحوصات مصلية، وبعدها أخذ خزعة من الامعاء الدقيقة للتأكد من المرض. نتوخى مقاربة من خطوتين لتطبيق المعارف الوراثية الجينية كأداة تشخيصية أو أداة فرز لتفادي المشاركة في المراضة والمضاعفات على المدى الطويل. أولا، على أساس نوع HLA لدى الأفراد. مع عدم وجودHLA-DQ2/DQ8 يمكن استبعادالمرض بما أنه من الناحية العملية ليس هناك خطر لتطور المرض التجويفي البطني. للباقي، من خلال الجمع بين هذه المتغيرات غير HLA ، يمكن تصنيف الأفراد إلى خطر منخفض (خطر مطلق <0,1 ٪) المتوسط (خطرمطلق 0,1 ٪ - 7 ٪) أو خطر عالي (خطر مطلق >7 ٪). من مجموعات الخطر فقط ان المجموعات التوسطو والعالية المخاطر تخضع لاختبار الأمصال والخزعة. نصف في الفصل 5 ، الدراسة الأولى حول مخاطر التنميط للمرض التجويفي البطني، وكيف يمكن إحصائيا التمييز بين الحالات والضوابط. باستخدام 10 من المتغيرات غير HLA التي تم تحديدها في دراسة GWAS الأولى، احتسبنا درجة المخاطرة لكل فرد عن طريق جمع عدد مخاطر الأليلات في 2308 حالات و 4585 ضوابط من هولندا، المملكة المتحدة وايرلندا. كما هو متوقع وجدنا عدد مخاطر الأليلات في الحالات أكثر من الأليلات في الضوابط. أضف إلى ذلك ان الأفراد الذين كانوا يحملون أليلات مخاطر أكثر من 13 قد تزايد خطرهم 6,2 مرة مقارنة مع أولئك الذين يحملون أليلات مخاطر أقل من خمس. تم اقرار ذلك مع فوج ايطالي مستقل. طور الجمع بين HLA والمتغيرات غير HLA بتحديد الأفراد ذات الخطورة العالية من 46,6 ٪ يستخدمون فقط HLA إلى 49,5 ٪ ، مع أن الخصوصية قد انخفضت بشكل طفيف من 93,6 ٪ إلى٪92,8. في الدراسة نفسها، وذلك باستخدام بيانات المحاكاة أظهرنا أن إضافة الأليلات خطرا على نموذج تحسين التنبؤ تصنف الأفراد الذين يعانون من خطر عال. في الفصل6، أظهرنا أن هذا صحيح أيضا بالنسبة للتنميط الجيني باستخدام بيانات حقيقية في قضايا المرض التجويفي البطني وضوابطه من 2675 حالة مرض تجويفي بطني و2822 ضوابط من هولندا، ايطاليا، اسبانيا، بولندا والمملكة المتحدة. قمنا بمقارنة متوسط الأرقام القياسية للخطر الجيني باستخدام 10 و 26 و57 من المتغيرات التي حددها GWAS ، ودراسة ثانية ل GWAS ولدراسة الخريطة الدقيقة بذلك على التوالي. مضيفين متغيرات غير HLA الى خطر التنميط مما يحسن التعرف والتمييز بين الحالات و الضوابط. ويعتبر ذلك من خلال الزيادة في منطقة خاضعة لعلاج متلقي التشغيل (AUC) ، والذي ارتفع من 82,3 ٪ (فقط HLA) إلى 83,2 ٪ (نموذج ب 10 متغيرات، 84,3 ٪ (نموذج ب 26 متغيرات و 85,4 ٪ (نموذج ب 57 متغيرا). بالإضافة إلى ذلك، تحسن صافي إعادة التصنيف (NRI) الذي هو مقياس لمدى اعادة تصنيف افضل للأفراد في الفئات الصحيحة مع مقارنة بنموذج HLA فقط ، تحسن ذلك من 4,1 ٪ (نموذج مع المتغيرات 10) ، إلى 7,1 ٪ (نموذج مع المتغيرات 26) إلى 11,1 ٪ (نموذج مع المتغيرات 57).

فكرة الاختبارات الجينية ليست جديدة. في وقت مبكر من سنة 1960، كان الأطباء قد حثوا على ضرورة اختبار الأطفال حديثي الولادة لأمراض نادرة مثل الفينيل كيتون (PKU) الذي يسبب التخلف العقلي. ويمكن منع PKU مع اتباع نظام غذائي خاص إذا تم الكشف عنه في وقت مبكر من الحياة. وتجرى حاليا اختبارات للPKU وغيرها من الأمراض النادرة ولكن يمكن علاجها بشكل روتيني قريبا بعد ولادة الطفل. في ألفصل 7، نناقش استخدام الاختبارات الجينية للمرض التجويفي البطني لدى أفراد من عائلات لديهم قريب من الدرجة الأولى مصابا بمرض التجويفي البطني، أو أولئك الذين لديهم أمراض المناعة ذات الصلة مثل داء السكري نوع 1 . فهم معرضون لمخاطر عالية لتطوير المرض التجويفي البطني، وبالتالي يمكن للمجموعة الأولى الاستفادة من الفحص الجيني. أضف إلى ذلك ، إذا تم العثور على علاج للتدخل يكون فعالا في الأطفال حديثي الولادة ، يمكن عندئذ استخدام التنميط الجيني لتحديد الأشخاص الذين يمكنهم الاستفادة من التدخل المبكر لمنع أو تأخير تطور المرض التجويفي البطني. ومع ذلك، النموذج الذي نقترحه في هذه الأطروحة لا يزال بالامكان تحسينه لتحديد متغيرات مخاطر أكثر ندرة بين السكان

والجينات المسببة المسارات وبما في ذلك العوامل غير الوراثية مثل التاريخ العائلي ، والوقت وإدخال كمية الغلوتين أثناء الفطام ومدة الرضاعة الطبيعية.

أخيرا ، وربما قريبا يمكن استخدام التنميط الجيني لغيره من الأمراض المعقدة المشتركة. هنا يكمن مستقبل الطب الشخصي. يمكن للأشخاص الذين يعلمون مبكرا في حياتهم أنهم مهيئون وراثيا لمرض مثل المرض التجويفي البطني الاستفادة من معرفة الأعراض التي ينبغي البحث عنها وبمعرفة المرض في مراحله المبكرة. قد يكونوا أيضا قادرون على تغيير نمط جوانب حياتهم وبيئتهم، أو الاستفادة من التدخل المبكر لمنع حدوث المرض. سيكون الناس قادرين على زيارة أطبائهم ، وسحب عينة من دمهم للمعرفة أكثر حول المخاطر الصحية التي تواجههم في العديد من الأمراض، بما في ذلك المرض التجويفي البطني. ومع ذلك ، قبل أن تصبح هذه الرؤية حقيقة واقعة، لدينا طريق طويل لنقطعه ، ولمعرفة الكثير عن الجينات والبيانات، والبروتينات ، الأيضات ، وتفاعلات الجينات الوراثية والجينات وتفاعلات البيئة.

# 10 نقاط لا بد في تذكرها من هذه الأطروحة

1. المرضى الذين لا يقومون بتشّخيص أمراض الجهاز الهضمي هم في خطر متزايد لمضاعفات لا رجعة فيها . وبالتالي هناك حاجة كبيرة لتحسين التشخيص وتحديد هوية الأفراد ذات المخاطر العالية.
2. يمكن للكشف عن أليلات HLA - DQ2/DQ8 أن يساعد بالفعل على تجاهل الداء الزلاقي (المرض التجويفي البطني) في التشخيص.
3. ان استعمال ستة أنماط لأشكال النوكليوتيداتSNP هي وسيلة دقيقة حساسة وقليلة الكلفة للفحص العام للسكان وتوقع وجود أو عدم وجود واحد من HLA - DQ2/DQ8
4. ان عدم مضاعفة مشاركة جينوم واسعة في السكان الجدد يمكن أن يكون سببه حجم العينة . وأيضا إلى اختلافات في خلفيات وراثية.
5. وقد ساعدت هذه النتائج المثيرة للمشاركة في دراسات الجينوم على نطاق صقل نموذجنا وإعادة تصنيف بعض الأفراد إيجابا لDQ2 و / أو DQ8 من مخاطر وسيطة تقوم على أساس المورثات HLA في مجموعة المخاطر العالية.
6. لقد تحسنت المتغيرات الجديدة المرتبطة بأمراض الجهاز الهضمي بنسبة 11 ٪ بين الأفراد إلى فئات أكثر دقة.
7. ينبغي أن يجمع تشخيص الداء الزلاقي العديد من المعلمات. بدءا من الفرز HLA. احتساب درجة المخاطر الجينية غيرHLA . أخذ خزعة مصلية وأخيرا القيام باختبارها.
8. خطر التنميط الجيني لأمراض الجهاز الهضمي قد يكون لديه احساسا أعلى ودقة مع إدراج العوامل الوراثية الأكثر تحديدا . مثل المتغيرات النادرة و/أو المعينة.
9. الاختبارات الجينية لمرض الاضطرابات الهضمية ليست بعد مناسبة للاستخدام السريري لأنها لا تزال بحاجة إلى مزيد من البحث بادخال عوامل غير وراثية مثل التاريخ العائلي. ووجود أمراض مناعية أخرى ذات صلة. والوقت وإدخال كمية الغلوتين أثناء الفطام . ومدة الرضاعة الطبيعية.
10. ذات يوم. سيتم فحص الأطفال حديثي الولادة أولا لجهة المتغيرات HLA وغير HLA المصنفة في الفئات المعرضة لخطر المرض التجويفي البطني ومن ثم علاجهم بالاستناد الى البيانات الشخصية الوراثية الخاصة بهم.

# Acknowledgements

# Acknowledgements

Preparing a PhD for four years would not have been easy without the friends and colleagues that I met along the way. I didn't only learn about complex genetics but also about many cultures and backgrounds. Now, I have friends and colleagues from probably all over the world (well maybe not ALL but most of the world!). As everyone I met shaped my vision in different ways and actually contributed directly or indirectly to this thesis, now it is time to acknowledge them and express my gratitude.

Dear Cisca, it all started with meeting you at the central station in Utrecht and you offering me the opportunity to join your group in Groningen. Looking back at the last four years, I can't imagine doing a better PhD. I appreciate all your scientific input, your great leadership, your positive attitude even when I screwed up a big expensive experiment, your efficient discussions, and the time you took for your PhD students even when it was last minute. It is your motivation and dedication to this work that motivated me. Thank you for all the opportunities, your encouragement, guidance, trusts and support from the initial to the final level. It was a pleasure to be your PhD student and part of your group.

I am grateful to Prof. Peter Pearson for introducing me to Cisca and for beleiving in me.

I am thankful to Prof. A.J. Oldehinkel, Prof. G.H. de Bock and Prof. H.J. Verkade for agreeing to be part of the reading committee for my thesis, taking the time to read it and for giving their approval for the defense.

I owe my deepest gratitude to Jackie, this thesis would have not seen the light without your help. Thank you for all the editing during these four years: my thesis, my presentations, my CV and much more. Thank you for making space in your agenda even when it was last minute and for your constructive comments, for going through my papers on Scopus one by one to select mine as there are apparently many "Romanos J" and calculating my H-index and citations.

I would like to thank all those I collaborated with and with whom I exchanged many, many e-mails and had fruitfull discussions. The PreventCD Group (preventing celiac disease) was the first project I worked on and was the starting point of meeting and working with multi-cultural people from Leiden, Naples, Warsaw, Madrid, Valencia, Barcelona, Tel Aviv, Zagreb, Budapest, Munich, Trieste, Umeå and Oslo. I will not cite all the names, as I might miss someone. Thank you all for the pleasant collaboration, great progress meetings, nice dinners and fruitful discussions. I am always amazed how much you are actively involved in the project knowing every single person in the study. A special thank you to Luisa and Yvonne for all your help and organization in leading this project. I hope the efforts of these 4 years will lead to great and positive results in preventing celiac disease. I look forward to reading about the outcome of this study in 2014. Dear Marike, thank you for your time and patience in helping me with my projects. You were always willing to squeeze me into your tight

agenda and were constructive in responding to my e-mails. Dear Ilja, thanks for the time you took to explain power calculation among other statistics and run some of my analyses. Dear Martin W, thank you for your help getting me up and running in the lab when I arrived. It was a pleasure to write my first review under your supervision. Dear Donatella, Maria Teresa, Maria Cristina, Suraksha, Mohammad and Sjoerd thank you for the collaboration, constant willingness to provide me with more DNA samples and digging out information from your databases. Dear Lotte and Paivi, thank you for the successful collaboration that resulted in a fine manuscript we wrote together. Dear Carlo, the Saharawi population was a very interesting and challenging project for me; I hope we will be able to publish the results soon. Dear Ed, Jill, Marian, George, Andrew, Kathy and Michelle, it was a great pleasure to meet you in Denver and collaborate on the Daisy project. Dear Ed and Tania, thank you for the great month I spent in Denver, for the nice sightseeing, the dinners, and for teaching me how to play tennis. Dear Thelma, Garima, Senapati, Kirti, Thenral, Shruti, Michael, Suman, Eipshita and Mitashree (may she rest in peace), thank you for the amazing time I had in Delhi, India. You showed me the culture of your country, with a lot of colors, customs and flavors. It was a pleasure to meet all of you and work with you for a week.

I cannot forget to express my appreciation to all the staff of the Genetics Department. Dear Marina, Edwin and Hayo, thank you for taking care of the finances, solving the computer problems and responding to the administration concerns. Dear Bote, Mentje, Ria, Joke and Hélène, thank you for all your help during these years in sending and receiving all the DNA packages and taking care of documents. A special thank you to Hélène for all your help with arranging the documents for the thesis and your willingness to help whenever you could. I am also very thankful to all my colleagues in the Genetics Department. Thank you for making the lab such an enjoyable place to work. Dear Helga, Gerben, Annemieke, Peter, Joerike, Tjakko, thank you for being great roommates during my first year and your help in translating Dutch letters. Dear Yunia, Paul, Mats, Omid, Rajendra, Eva, Olga, Justyna, Anna P, Céline, Anna D, Christine, Marcel W, it was great to meet you all, to hang out with you at lunches, meeting you in the corridors and at the lab bench, sharing the nicest moments of my PhD as well as the stressful ones. Thank you for your support and exchange of experiences. Yunia and Gerben, wrapping up our PhD at the same time was very helpful, good luck with your defense and all the best for the future. Dear Bahram, Pieter and Mariska, thank you for your help with my VeraCode experiment, especially when the BeadExpress crashes and I stress out. I have a lot of good memories in the "old" Illumina lab. Dear Elvira, you left this world very young, thank you for teaching me how to perform Infinium experiments, I will always remember you. Dear Marten H, thank you for letting me use the taqman machine in your department. Dear Robert, Ellen, Dineke, Gerard, thank you for the discussions and comments to improve my research. I can truly say I was in a great environment for the past four years.

Dear friends and colleagues who are or have been part of the celiac and IBD group, when I started

here, we were able to have our Monday morning meetings in Cisca's room but since we are growing, we need now a bigger table than the one in the 'Pob' room. Dear Agata, Gosia, Asia, Barbara, Javier, Rodrigo, Isis, Cleo, Vinod, Sebo, Karin, Suzanne, Mitja, Rutger, Senapati, Rinse, Lude, Harm-Jan, Juha, Dasha, Jingyuan, Mathieu, Soesma, Astrid, Morris, Flip, Freerk, Alex, Laurent, Patrick, Marc Jan, Roan, Noortje, Monique, Ania R, Sasha, Marcel B, Ron and Krista K, thank you for all the work discussions, suggestions and help. Dear Soesma, Mathieu and Astrid, being able to do several projects at the same time and in a short period of time was only possible with your help and efficient work. I had the honor to work with each one of you. Soesma, you were the best student I have ever had. It is so easy to work with you, especially since we work so much alike - it is scary sometimes. Mathieu, they say that men cannot multi-task, well, they're wrong. I have never seen anyone as you doing as many tasks at once as possible. Keep up this energy and enthusiasm! Astrid, your organizational skills and detailed notes added a lot to this group. Since you are from a different field, you tried to use your experience to improve our lab organization, while at the same time you were open to learn new things. Dear Cleo, I had the pleasure to work with you on my top 'risk model' paper. I owe the most of my statistical background in this field to you. Thank you for all the time we sat in front of SPSS, discussing things in the whole way. Big thanks to you (and your two lovely daughters) for taking the time to translate my summary into Dutch while you were on maternity leave. Dear Lude, thank you for your help with the statistics. Although you did not always understand my projects, you still tried to help me out. Dear Jingyuan, thank you for your advice and comments on the risk model and for writing very helpful R scripts. Dear Javier (Dr. Gutierrezzz), thanks for your help with the computer and programming, and for your patience in replying to all my computer questions. Dear Rodrigo, it was nice to have a brazilian colleague who reminded me of the joyful and warm culture. it was a pleasure to work with you and muito obrigada pela correcao da traducao do meu resumo em portugues. Dear Barbara, thank you for your help with DNA isolation and measurement. Dear Isis, it was a pleasure to work with you in the last few months. I have enjoyed the nice parties, especially dancing with you. I will come to Groningen especially to teach you more belly dancing ;) Dear Ania, you were one of the first people I met in Groningen and hung out with. It was a pleasure to write my first review with you and I had a great time with you in Maribor. Dear Noortje, they told me that it takes a long time before you become friends with a Dutch person, it wasn't true. I remember you inviting me for a dinner at your place just a few weeks after I arrived! Thanks for your help with Taqman and the nice trip we took to Kiel. Dear Sasha, I'm always impressed with how you manage to be a great scientist and an amazing mother and wife. Thank you for teaching me the basics of complex genetics, how to read the Taqman plots, and for the good time we had at the ESHG meeting in Vienna. Dear Monique, I enjoyed very much our talks about life and what we believe in. It seems that the Dutch and Lebanese cultures are closer than I thought. Dear Gosia, Agata and Asia, I look back at these four years, and I think to myself: " wow, time really goes fast". It was a pleasure meeting you, being your roommate, colleague

and friend, sharing the joys and worries of this PhD, having dinners together, partying and traveling. Brussels was the greatest trip, I mainly remember that I didn't stop laughing. We talked about everything from scientific topics to personal life. I tell you, it wasn't easy to be in a group where I'm the only person who doesn't speak Polish ;) but you tried to switch to English every time I was there. I promise my next language course will be Polish :)

Dear Claudia, thank you for making the cover of my thesis and the layout. You are a great artist and soon to be the greatest mother. I am sure someone will recognize your talent sooner then you think. Dear Jana, thanks for the nice chats and your trusting me to babysit Masha. (Masha, I wish you all the best in your studies and going to Japan).

I made many friends at work and also outside work. I shared a lot of dinners, parties and birthdays with many of them. Dear Maria, Kaushal, Mateusz, Anna F, Pedro, Susana, Roberta, Bispo, Romy, Raquel, Hassiba, Maxi, Camiel, Fany, Elena, Ildiko, Lara, Bruno, Ryan, Juan, Marcos, Thomas and Nai-hua, it was a pleasure meeting you all. I had a great time partying with you. Thank you for all your support and friendship, I hope we will keep in touch. Dear Ildiko, thanks for the weekends and long nights spent working together on our thesis. Good luck with your PhD! Dear Elena, you are the most sociable person I have ever met. I think you know all the people in Groningen ;) It is so easy to talk to you, get to know you and become your friend. Thank you for organizing all the theme parties, dinners in new restaurants, and trips in Holland … and a big thanks for being my paranimf and helping me with all the arrangements.

Dear Adour, Robert, Alina, Mark, Alex and Sami H, it was a pleasure to know you and always good to meet you whenever I was in the South of Holland. Dear Nareen, we met in Lebanon a long time ago, then we became friends at AUB, and now we ended up living in the same country. It was great to feel I have an old friend close by. Thank you for hosting me in your house in Amsterdam several times. Dear Anya and Max, my second home in Amsterdam, thanks for letting me stay at your place every time I had to travel early or arrive late and wasn't able to catch the train. Anya, we became friends very fast as we found we had a lot in common. I enjoyed shopping with you, talking about cosmetics, fashion and life.

Dear Saleh, Mona, Nada, Mazen, Chirine M and little Amy. You were my Lebanese family in Groningen. We only met two years ago, but it did not take long before we came good friends. Every time we met, I felt like I was at home, you made my nostalgia for home easier.

Dear Mara, Bine and Fany, without you I would have never survived the five months in the student house. Bine, you were our best cook as you could create the best food from anything. if you ever decide to quit science, please become a cook. Mara, I admire your courage to go to Tanzania for a full year to teach English and learn their language. Keep up the good will and if you stay there, I will definetly visit you. Fany, I don't know where to start with thanking you. You were my neighbor, my friend, and my family here. We shared a lot of

moments together. Thank you for always being there. I enjoyed our long dinners, nights out, all our discussions on life, love and men … Thanks for being my paranimf, correcting the French summary, and helping me with all the arrangements.

Dear friends around the world, Carola, Fernanda, Chirine B, Lea, Mia, Mounia, Dina, Hala, Dana, Sahar, Lana, Amine, Georges, Leticia, Andre, Rita, Helvio and the colleagues and friends from USP, thank you for being there for me even though you are far. Every time we meet, it was as if I have never left. Little Sophie, Marina and Patricio, can't wait to see you!

Dear Simon, Nicole, Wissam, Hani and Samy, I had a great time visiting you in the weekends and conducting long discussions about careers, studies and hobbies. It was so good to know that I had a family I could rely on when needed.

Coming from a Lebanese family, which means a big family divided between Lebanon, Brazil, Panama and USA, I would like to thank all my uncles, aunts, cousins, tios, tias, primos, oncles, tantes, cousins/cousines, khalos, 3amos, 3amtos, for believing in me and supporting me in one way or another.

Jido and Teita, although you are gone, I felt you were always with me, protecting me and guiding me from above. I hope I have made you proud.

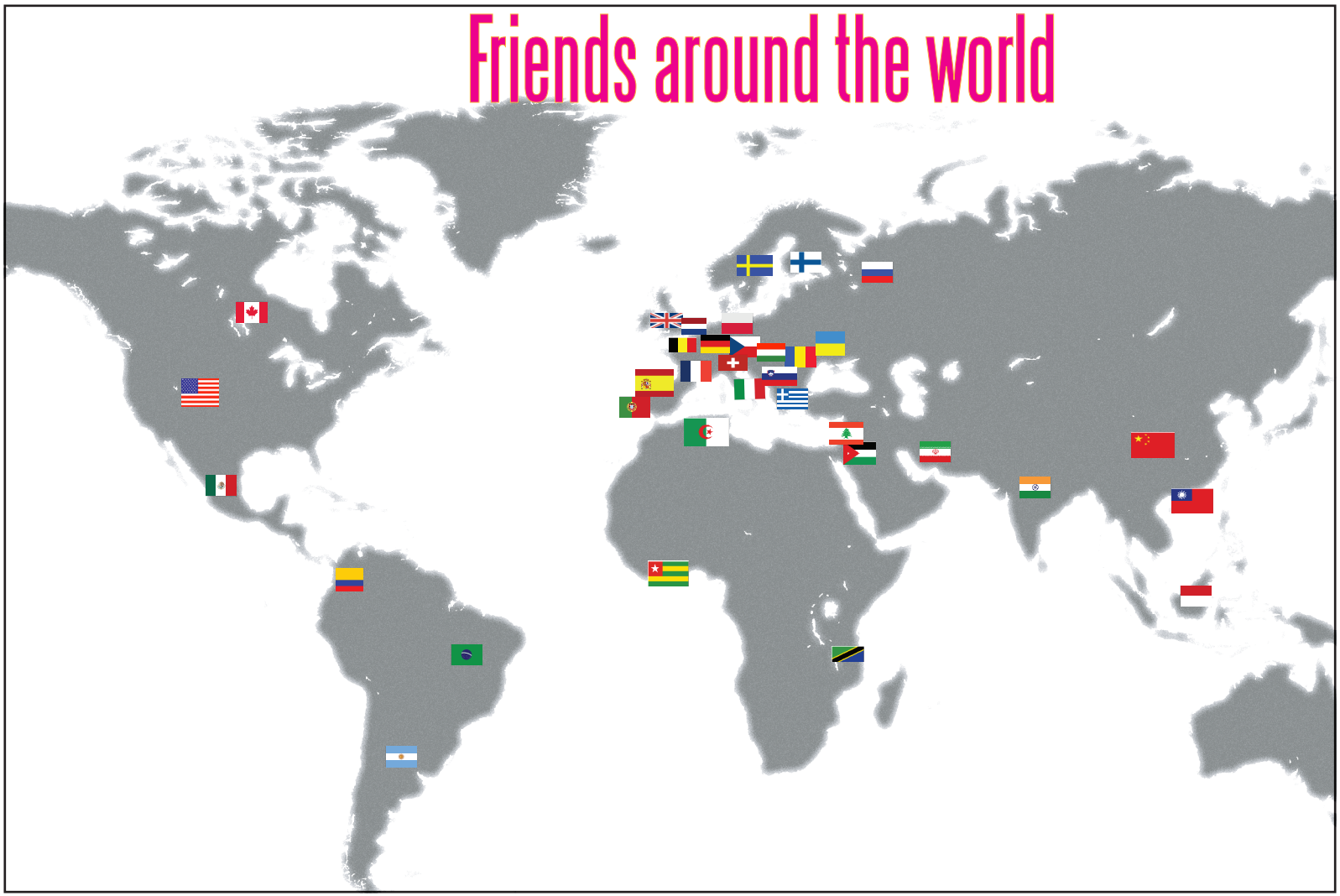
Dear family, Mami, Papi, Kristy, Julie, Ziad and Teta Vola, thank you for all your support and always helping me to achieve the best. Chère Teta Vola, c'est grace a tes prières et ton amour que je suis arrivée ou je suis aujourd'hui. Merci pour tout, je t'aime! Juju and Ziz, it wasn't easy to be far away when you were preparing the nicest day of your lives, your wedding. It was an honor to be your maid of honor. Kouki, being the youngest, never meant you know less. Dearest sisters, thank you for all your advice, support and believing in me.

Last but not least, chers mami et papi, je ne peux jamais vous remercier pour tous ce que vous avez fait pour moi. Mamin, je sais combien c'était dificile de m'avoir loin mais comme tu savais que c'était mon reve de faire ce PhD, tu m'as laisse partir, tu as essayé de comprendre et bien qu'on était dans deux pays different, tu m'appelais chaque jour, surtout durant les moments dificiles. Papi, sans ton encouragement et ta motivation pour que j'étudie et je realise mes reve, je ne serais jamais arrivé ici. Tu étais toujours present a m'aider à conquérir ce que je voulais. Vous êtes mes idoles et tous ce travaille vous est dedié.

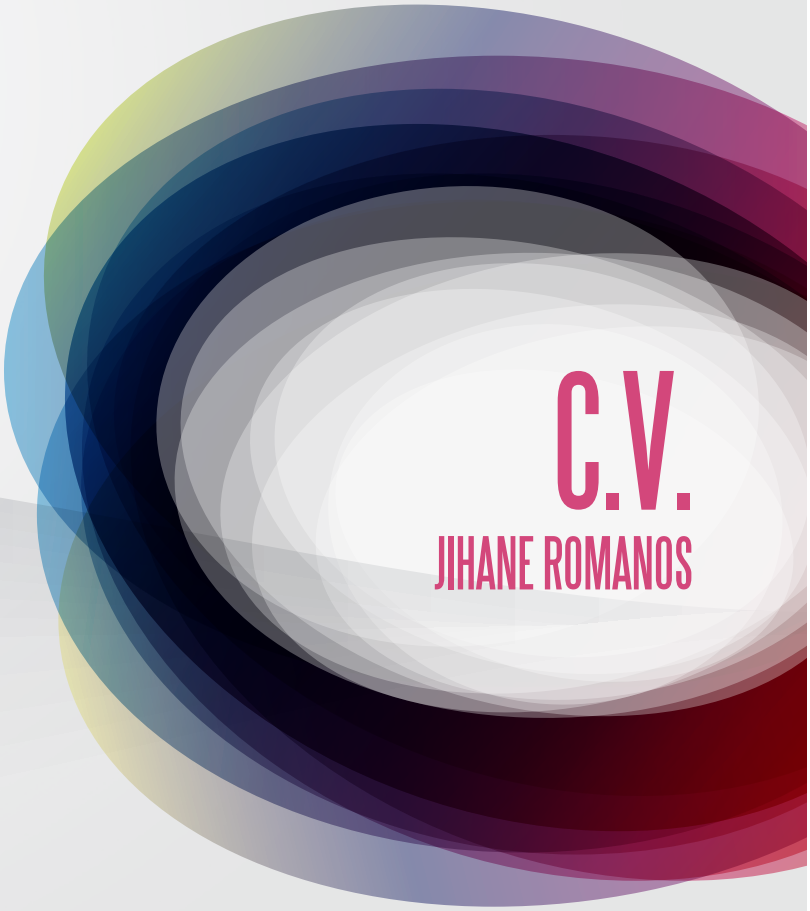
THANK YOU … MERCI … DANK JE … OBRIGADA … and CHUKRAN!
I learned a lot from each one of you and I hope to keep in touch long after my PhD defense.

Cheers,
Jihane

# Friends around the world



# CURRICULUM VITAE



# C.V.

## JIHANE ROMANOS

# Curriculum Vitae

Jihane Romanos was born on 23rd January 1982 in Haret-Sakhr, Lebanon. She lived in Sao Paulo, Brazil, for two years before moving back to Lebanon where she grew up and went to school at the Soeur des Saint Coeur college in Kfarhbab. After taking the French baccalaureate in 2000, she went to the American University of Beirut where she majored in Biology (BSc). In 2003 Jihane moved back to Brazil, where she followed a three-month training course at the genetics department of the University of Sao Paulo. In 2004, she started her MSc degree in that department, working on the screening of mutations in the *OTOF* gene in patients with hearing impairment and its relation with auditory neuropathy, under the supervision of Prof. Regina Celia Mingronino-Netto, PhD. She graduated in 2006 with an MSc in Biology, specializing in genetics. In 2007, she started her PhD work at the Department of Genetics, University Medical Center Groningen, the Netherlands, under the supervision of Prof. Cisca Wijmenga, PhD. Her main topic has been the genetics of celiac disease and assessing its diagnostic value. As someone holding both the Brazilian and Lebanese nationalities, Jihane has excellent language skills. She is fluent in Arabic, French, English and Portuguese, and has also learned some basic Dutch.

# Publications (H-index in Sept 2011 = 8)

*Both authors contributed equally to the manuscript


1. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha EG, Almeida RC, Dias KM, van Diemen CC, Dubois PCA, Duerr RH, Edkins S, Franke L, Fransen K, Gutierrez J, Heap GAR, Hrdlickova B, Hunt S, Plaza Izurieta L, Izzo V, Joosten LAB, Langford C, Mazzilli MC, Mein CA, Midah V, Mitrovic M, Mora B, Morelli M, Nutland S, Núñez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Ribes-Koninckx C, Ricaño-Ponce I, Rich SS, Rybak A, Santiago JL, Senapati S, Sood A, Szajewska H, Troncone R, Varadé J, Wallace C, Wolters VM, Zhernakova A, CEGEC (Spanish Consortium on the Genetics of Coeliac Disease), PreventCD Study Group, Wellcome Trust Case Control Consortium, Thelma B.K., Cukrowska B, Urcelay E, Bilbao JR, Mearin ML, Barisani D, Barrett JC, Plagnol V, Deloukas P, Wijmenga C, van Heel DA. **Dense genotyping reveals and localises multiple common and rare variant association signals in celiac disease.** *Nat. Genet.* 2011; in press

2. Sperandeo MP, Tosco A, Izzo V, Tucci F, Troncone R, Auricchio R, Romanos J, Trynka G, Auricchio S, Jabri B, Greco L. **Potential celiac patients: a model of celiac disease pathogenesis.** *PLoS One*. 2011; 6(7):e21281.

3. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, Westra HJ, Fehrmann RS, Kurreeman FA, Thomson B, Gupta N, Romanos J, McManus R, Ryan AW, Turner G, Brouwer E, Posthumus MD, Remmers EF, Tucci F, Toes R, Grandone E, Mazzilli MC, Rybak A, Cukrowska B, Coenen MJ, Radstake TR, van Riel PL, Li Y, de Bakker PI, Gregersen PK, Worthington J, Siminovitch KA, Klareskog L, Huizinga TW, Wijmenga C, Plenge RM. **Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci.** *PLoS Genet*. 2011; 7(2):e1002004.

4. Romanos J and Wijmenga C. **Predicting susceptibility to celiac disease by genetic risk profiling.** *Annals of Gastroenterology and Hepatology.* 2010; 1(1):11-18.

5. Hogen Esch CE, Rosén A, Auricchio R, Romanos J, Chmielewska A, Putter H, Ivarsson A, Szajewska H, Koning F, Wijmenga C, Troncone R, Mearin ML. **The PreventCD Study design: towards new strategies for the prevention of coeliac disease.** *Eur J Gastroenterol Hepatol.* 2010; 22(12):1424-1430.

6. Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CG, Franke L, Oosting M, Barisani D, Bardella MT; Finnish Celiac Disease Study Group, Joosten LA, Saavalainen P, van Heel DA, Catassi C, Netea MG, Wijmenga C. **Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection.** *Am J Hum Genet.* 2010; 86(6):970-7.

7. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, Bardella MT, van den Berg LH, Bockett NA, de la Concha EG, Dema B, Fehrmann RS, Fernández-Arquero M, Fiatal S, Grandone E, Green PM, Groen HJ, Gwilliam R, Houwen RH, Hunt SE, Kaukinen K, Kelleher D, Korponay-Szabo I, Kurppa K, MacMathuna P, Mäki M, Mazzilli MC, McCann OT, Mearin ML, Mein CA, Mirza MM, Mistry V, Mora B, Morley KI, Mulder CJ, Murray JA, Núñez C, Oosterom E, Ophoff RA, Polanco I, Peltonen L, Platteel M, Rybak A, Salomaa V, Schweizer JJ, Sperandeo MP, Tack GJ, Turner G, Veldink JH, Verbeek WH, Weersma RK, Wolters VM, Urcelay E, Cukrowska B, Greco L, Neuhausen SL, McManus R, Barisani D, Deloukas P, Barrett JC, Saavalainen P, Wijmenga C, van Heel DA. **Multiple common variants for celiac disease influencing immune gene expression.** *Nat Genet.* 2010; 42(4):295-302.

8. Romanos J*, van Diemen CC*, Nolte IM, Trynka G, Zhernakova A, Fu J, Bardella MT, Barisani D, McManus R, van Heel DA, Wijmenga C. **Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease.** *Gastroenterology* 2009; 137(3):834-40, 840.e1-3.

9. Romanos J, Kimura L, Fávero ML, Izarra FA, de Mello Auricchio MT, Batissoco AC, Lezirovitz K, Abreu-Silva RS, Mingroni-Netto RC. **Novel OTOF mutations in Brazilian patients with auditory neuropathy.** *J. Hum. Genet.* 2009; 54(7):382-5.

10. Romanos J*, Koskinen L*, Kaukinen K, Mustalahti K, Korponay-Szabo K, Barisani D, Bardella M.T, Ziberna F, Vatta S, Széles G, Pocsai Z, Karell K, Haimila K, Ádány R, Not T, Ventura A, Mäki M, Partanen J, Wijmenga C, Saavalainen P. **Cost-effective HLA-typing with tagging SNPs predicts celiac disease risk haplotypes in Finnish, Hungarian and Italian populations.** *Immunogenetics* 2009; 61(4):247-56.

11. Romanos J and Wijemenga C. **Letter to the editor in response to the article "Two single nucleotide polymorphisms identify highest-risk diabetes human leukocyte antigen genotype: potential for rapid screening" by Barker et al.** *Diabetes* 2009; 58(1):e1; author reply e2.

12. Romanos J, Barisani D, Trynka G, Zhernakova A, Bardella MT, Wijmenga C. **Six new celiac**

**disease loci replicated in an Italian population confirming association to celiac disease.** *J Med Genet.* 2009; 46(1):60-3.

13. Trynka G, Zhernakova A, Romanos J, Franke L, Hunt KA, Turner G, Bruinenberg M, Heap GA, Platteel M, Ryan AW, de Kovel C, Holmes GK, Howdle PD, Walters JR, Sanders DS, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, Kelleher D, Barisani D, Bardella MT, McManus R, van Heel DA, Wijmenga C. **Coeliac disease associated risk variants in TNFAIP3 and REL implicate altered NF-κB signalling.** *Gut* 2009; 58(8):1078-83.

14. Monsuur AJ, de Bakker PI, Zhernakova A, Pinto D, Verduijn W, Romanos J, Auricchio R, Lopez A, van Heel DA, Crusius JB, Wijmenga C. **Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms.** *PLoS One* 2008; 28;3(5):e2270.

15. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GKT, Howdle PD, Walters JRF, Sanders DS, Playford RJ, Trynka G, Mulder CJJ, Mearin ML, Verbeek WHM, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA. **Newly identified genetic risk variants for celiac disease related to the immune response.** *Nat Genetics* 2008; 40:395-402.

16. Romanos J*, Rybak A*, Wijmenga C, Wapenaar MC. **Molecular diagnosis of celiac disease: Are we there yet?** *Expert Opin. Med. Diagn*. 2008; 2(4):399-416.

17. Abreu-Silva RS, Batissoco AC, Lezirovitz K, Romanos J, Rincon D, Auricchio MT, Otto PA, Mingroni-Netto RC. **Correspondence regarding Ballana et al., "Mitochondrial 12S rRNA gene mutations affect RNA secondary structure and lead to variable penetrance in hearing impairment".** *Biochem Biophys Res Commun.* 2006; 343(3):675-6.

18. Fávero ML, Romanos J, Mingroni-Netto RC, Balieiro CR, Donini TS, Spinelli M. **Auditory neuropathy due to mutations in OTOF gene.** *Arq. Otorrinolaringol*. 2005; 9(4):325-30.

19. Viana-Morgante AM (The Human Cytogenetics Study Group including Romanos J). **The ratio of maternal to paternal UPD associated with recessive diseases.** *Hum Genet*. 2005; 117(2-3):288-90.