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.