Beyond genome wide association studies in celiac disease by exploring the non-coding genome

de Almeida, Rodrigo

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Appendices

Summary
Samenvatting
Resumo
Acknowledgements
Curriculum Vitae
**Summary**

Our body is built from 1 quadrillion (1,000,000,000,000,000) individual cells. It is the genetic matter — our genome — in the nucleus of each of these cells, that determines not only traits like eye color or length, but also harbors the susceptibility of an individual to get a disease. Our genome consists of deoxyribonucleic acid (DNA). The function of DNA is based on the order — the sequence — of only 4 building blocks called nucleotides. The full name of these nucleotides are adenine, thymine, cytosine, and guanine which are always abbreviated as A, T, C and G. Human DNA is organized in 23 units called chromosomes that together are 3.3 billion (3,300,000,000) nucleotides long.

Since the original determination of the human DNA sequence in 2001 technologies have been developed that allow for the rapid determination of the nucleotide sequence of individuals. One of the major findings from these studies is that human DNA differs 0.1% between different (unrelated) individuals. This means that every person differs on 3 million nucleotide position from for instance their neighbour. The sequencing of hundreds of thousands of individuals enabled the mapping of the varying nucleotides (also called single nucleotide polymorphisms (SNPs)) throughout the human genome. Surprisingly enough it also showed that SNPs close to each other show very strong correlation and lead to a so-called haplotype block structure of the genome. Consequently, the 3 million different SNPs are not needed to pinpoint genetic factors associated with traits and diseases, but roughly 500,000 SNPs that identify the different haplotype blocks are often sufficient. This new perspective on a more simplified map of human genome variation led in 2007 to a revolution in genetics as it allowed for the first time to perform genetics association studies across the whole human genome in thousands of individuals at relatively low cost. These studies are coined genome wide association studies (GWAS). As of today it has unraveled the genetic underpinnings of more than 400 different genetic traits and diseases ranging from coffee consumption and height to schizophrenia and autoimmune diseases.

GWAS are case-control studies, in which the frequencies of a set of 500,000 genome wide SNPs are compared between thousand of control samples and thousands of samples from individuals diagnosed with e.g. a particular disease. To date two successful GWAS have been performed in celiac disease which have provided novel insights into underlying biological pathways. CeD is the most prevalent food intolerance affecting around 1% of the Western population. Individuals genetically predisposed for CeD develop intolerance to gluten peptides
derived from proteins present at high levels in grains that are the basis of the common western diet. Consequently CeD patients cannot eat food prepared (or potentially contaminated) with gluten, such as bread, cookies and pasta. CeD associates very strongly with two genes that can make up the major histocompatibility complex (MHC), specifically the HLA-DQ2 and HLA-DQ8 molecules. Carriage of the HLA-DQ2 genotypes alone already accounts for 35% of the genetic risk of CeD. As approx. 25% of the general population carries the CeD associated HLA genotype it is not sufficient to develop CeD.

GWAS in CeD have identified 26 other factors, besides HLA, that are also important for disease development. A custom-made platform (Immunochip) was designed to fine-map the 26 non-HLA loci but also allowed our group to identify another 13 non-HLA loci to be associated to CeD, making the total number of loci 40 (HLA plus 39 non-HLA).

When I started with my PhD research the immune-genetics field was facing a big hurdle that needed to be taken. As mentioned above the regions associated with disease by GWAS are often rather large. This implied that novel approaches needed to be developed and applied to prioritize the causal SNPs and genes in these regions. Another issue that emerged was that in many loci the SNPs that associated the strongest with the disease were mapping to non-coding regions. In some cases there were no known protein encoding located in the locus at all and such loci were described to be “gene deserts”. A big development during the course of my studies was the completion of the Encyclopedia of DNA Elements (ENCODE) project. This project aimed to describe the “functional genome” and has not only exposed thousands of novel transcripts but also has mapped hundreds of thousands regulatory regions that control gene expression.

A clinical problem in CeD that I tried to address during my studies problem is that it is challenging to diagnose CeD. It has been estimated that only one out of 8 CeD patients is correctly diagnosed with CeD. Recently, a novel class of non-coding RNAs is emerging as promising biomarker candidates, the so-called circulating micro-RNAs (miRNAs). MiRNAs are small non-coding RNAs with about 22 nucleotides that regulate the expression of specific RNA targets. It was found that miRNAs are very stable in many body fluids and are therefore detectable in blood, cerebrospinal fluid, urine and even tears. Moreover it has been shown that circulating miRNA profiling can uncover disease and even disease-stage specific miRNAs in circulation. This feature makes them excellent biomarker candidates.

In this thesis research I addressed two major aims: (1) The identification of functional SNPs and genes in the loci exposed by the CeD Immunochip study, and (2)
investigate whether circulating miRNA are biomarkers for CeD.

In Chapter 1 I provide an overview of the genetic factors identified by pre-GWAS and GWAS studies that that were associated with CeD. Moreover, I discuss the CeD-associated genes and pathways that are shared with other immune-mediated diseases.

In Chapter 2 it is shown that by applying an integrative functional genomics and pathway analysis approach using novel and publicly available data we can prioritize functional SNPs, genes and pathways that contribute to CeD. One of the findings was that 4 poorly described genes might play a role in intestinal barrier function, a process which is known to be disturbed in CeD. In addition, we describe a transcriptional connection between Interferon- with CeD susceptibility genes, which sheds light on how IFN- is dysregulated in CeD, although the IFNG locus is not directly affected by CeD-associated SNPs.

The strongest associated non-HLA locus in CeD is the LPP locus. In Chapter 3 I present a fine-mapping approach that reduced the size of the association region. By applying haplotype analysis and integrating functional genomics approaches on publicly available data, we were able to refine this from a 70 kb region to a region of only 2.8 kb. SNP prioritization using ENCODE data suggest that rs4686484 is potentially the functional SNP in this region.

Recent meta-analyses in other diseases identified SNPs in the LPP region to have an expression quantitative trait locus (eQTL) effect on BCL6 which is located 658.7 kb away from the LPP gene well outside the actual LPP locus. In Chapter 4 I explored the possibility that SNPs in the LPP locus associated with CeD affect gene expression of genes other than LPP using RNA-seq data. This approach allowed us to reveal an eQTL effect of one of the SNPs in the locus on the long non-coding RNA (lncRNA) LPP-AS1. This is the first time that a lncRNA has been implicated in CeD. Cell type specific expression analysis and pathway enrichment analysis suggested a role for this lncRNA in ubiquitination. Functional studies are needed to fully proof that LPP-AS1 is one of the disease-causing genes.

In Chapter 5 I provide an overview on what is known about how genetic variation might impact regulatory non-coding RNAs, focusing on IncRNAs and miRNAs.

In Chapter 6 I used next generation sequencing (NGS) to profile circulating miRNAs in serum of CeD patients that were enrolled in the PreventCD cohort. The PreventCD project, is a prospective study in which samples were collected at fixed time points after birth of individuals at high risk for CeD. We found a panel of 45 miRNAs that are differentially expressed at time of CeD diagnosis when compared to 3 months after birth. A subset of these are already changing before diagnosis and some
appear to normalize upon the start of a gluten-free diet. These miRNAs provide a catalogue of potential biomarker candidates for CeD.

In Chapter 7 I describe two miRNA profiles in samples obtained from blood and small intestinal biopsies from another cohort of pediatric CeD patients and control individuals. This analysis exposed a panel of 49 circulating miRNAs in plasma and 109 miRNAs in biopsies that were differentially expressed when comparing CeD patients with the controls. Furthermore, we have identified miRNAs which are differentially expressed in the small intestines of CeD patients which could be involved in the pathogenesis of CeD.

In Chapter 8 I place the results of the work presented in this thesis in the context of current knowledge and point out the challenges arising in the genomics field, where the availability is growing exponentially.

The main conclusions of this thesis are:

- Different layers of publicly available functional genomics data can be used to prioritize SNPs, genes and pathways in CeD;
- CeD-associated SNPs are mostly located within genomic regions with regulatory function;
- Haplotype analysis in stratified populations is a powerful method to fine map disease associated regions;
- CeD associated SNPs can affect the expression of IncRNAs;
- Circulating miRNAs are potential biomarkers candidates to detect CeD and adherence to gluten-free diet;
- Even in complex cell mixtures such as biopsies CeD specific miRNA can be detected. Some of these might be involved in CeD pathogenesis.
Appendices

Summary
Samenvatting
Resumo
Acknowledgements
Curriculum Vitae
Samenvatting
Ons lichaam bestaat uit een biljard (1.000.000.000.000.000) individuele cellen. Het is de verzameling van al ons genetisch materiaal, het genoom, dat niet alleen de kleur van onze ogen bepaalt en hoe lang we zullen worden, maar ook hoe vatbaar men is voor het ontwikkelen van bepaalde aandoeningen. Het genoom bestaat uit de stof desoxyribonucleïnezuur (DNA). DNA is opgebouwd uit vier verschillende bouwstenen die nucleotiden genoemd worden: Adenine (A), Thymine (T), Guanine (G) en Cytosine (C). Hoe het DNA zijn werk doet is afhankelijk van de volgorde, de sequentie, van deze vier verschillende nucleotiden. In onze cellen is het DNA opgeslagen in 23 verschillende eenheden, de chromosomen, die in totaal zijn opgebouwd uit 3,3 miljard (3.300.000.000) nucleotiden.

Sinds de nucleotide volgorde van het menselijke genoom voor het eerst is gepubliceerd in 2001, zijn er technieken ontwikkeld om deze volgorde, de sequentie, steeds preciezer en sneller te bepalen. Het gevolg van deze ontwikkelingen was dat er steeds meer onderzoek gedaan kon worden naar het menselijk genoom. Deze onderzoeken leverden zeer belangrijke informatie op. Zo werd bekend dat niet verwante individuen maar 0,1% van elkaar verschillen op DNA niveau. Dit betekent dat het verschil tussen u en uw buurman/buurvrouw ‘maar’ ongeveer 3,3 miljoen nucleotiden is. Door van honderdduizenden mensen de sequentie van hun genoom te bepalen, konden de nucleotide verschillen nauwkeurig worden bepaald. Deze specifieke inter-persoon “variabele” nucleotiden worden ook wel “single nucleotide polymorphisms” (SNPs) genoemd. De sequencing studies hebben verrassenderwijs ook laten zien dat SNPs die dicht bij elkaar liggen vaak een sterke onderlinge correlatie hebben en zo blokken van SNPs vormen, de zogenaamde “haplotype-blocks”. Als je weet welk haplotype een individu draagt, kun je met grote zekerheid zeggen dat deze persoon alle SNPs in dit haplotype-block met zich mee draagt. Dit betekent dat om het genoom goed in kaart te brengen men niet alle 3,3 miljoen SNPs hoeft te bepalen, maar dat men gebruik kan maken van indicator SNPs die representatief zijn voor een haplotype block (ongeveer 500.000). Dit nieuwe inzicht heeft in 2007 tot een revolutie in de genetica geleid. Voor het eerst kon het hele humane genoom in kaart worden gebracht tegen een relatief lage prijs. Dit leidde tot de opkomst van de “genome-wide association studies” (GWAS), waarin er gezocht kon worden naar een relatie tussen de aanwezigheid van SNPs en fenotypes. Tot op heden zijn er meer dan 400 verschillende fenotypes met behulp van GWAS onderzocht, variërend van het drinken van koffie en lengte, tot schizofrenie en auto-immuun ziektes.

In GWAS worden met behulp van ~500.000 SNPs alle “haplotype
blocks” in kaart gebracht (en daarmee dus het hele genoom). Vervolgens wordt de frequentie van elke individuele SNP in de controlegroep van gezonde individuen vergeleken met de frequentie van dezelfde SNP in de “testgroep” (bijvoorbeeld een cohort mensen met een bepaalde aandoening). Tot nu toe zijn er twee succesvolle GWAS gedaan met patiënten met de ziekte coeliakie (CeD), die nieuwe inzichten gaven in de biologische processen die een rol spelen in deze aandoening.

CeD is de meest voorkomende voedselintolerantie en komt bij ongeveer 1% van de westere populatie voor. Mensen met CeD zijn intolerant voor een bepaalde set eiwitten, de gluten, die voorkomt in verschillende graanoorten die de basis vormen van het normale westere dieet. CeD patiënten kunnen geen voedingswaren eten die gluten bevatten, en zelfs geen etenswaren die gecontamineerd zijn met gluten, waaronder bijvoorbeeld brood, koekjes, bier en pasta.

Twee genen die sterk geassocieerd zijn met CeD coderen voor eiwitten van het humane “major histocompatibility complex” (MHC): HLA-DQ2 en HLA-DQ8. Het hebben van het HLA-DQ2 genotype is verantwoordelijk voor 35% van de overerfbaarbaarheid van CeD. Echter gezien 25% van de algemene populatie drager is van HLA-DQ2, terwijl maar 1% CeD ontwikkelt, betekent dit dat HLA-DQ2 niet voldoende voor het ontwikkelen van CeD is. Met behulp van GWAS zijn naast HLA nog 26 andere genetische factoren gevonden die belangrijk zijn voor het ontwikkelen van CeD. Daarnaast is de Immunochip beschikbaar gekomen. De Immunochip is een speciaal genotyperings-platform dat specifiek is ontwikkeld om autoimmuunerelateerde SNPs op te sporen. Met behulp van de Immunochip heeft onze onderzoeksgroep niet alleen de 26 niet-HLA loci verder weten te specificeren, maar zijn er ook nog 13 extra loci opgespoord die geassocieerd zijn met CeD. In totaal zijn er dus 40 genetische CeD loci bekend (HLA locus plus 39 niet-HLA loci).

Toen ik in 2010 begon met mijn PhD onderzoek, stond de immuno-genetica voor een grote uitdaging. De regio’s op het genoom die door GWAS gedefinieerd werden als geassocieerd met de ziekte zijn namelijk vaak groot en konden soms tientallen SNPs en meerdere genen omvatten. Het was dus erg moeilijk om de causale SNPs en genen te prioriteren. Er moesten eerst nieuwe technieken worden ontwikkeld die dit wel konden doen. Wat ook duidelijk werd is dat verreweg de meeste SNPs (>90%) die zijn geassocieerd met CeD, niet in eiwit coderende delen van het genoom lagen. Soms lagen er zelfs helemaal geen eiwit-coderende genen in het betreffende locus en daarom werden deze loci “gene deserts” genoemd. In de loop van mijn PhD,
in 2012, was er nog een belangrijke nieuwe ontwikkeling in de genetica namelijk het bekend worden van de resultaten van het “Encyclopedia of DNA Elements” (ENCODE) project. Het doel van dit project was om alle functionele elementen in het genoom te beschrijven en deze analyses hebben niet alleen tienduizenden nieuwe transcripten opgeleverd, maar ook honderdduizenden regio’s die genexpressie reguleren.

Het is lastig om op een gemakkelijke manier de diagnose CeD te stellen en het is geschat dat slechts een van de acht CeD patiënten juist gediagnostiseerd is. Ik heb me tijdens mijn PhD ook met dit probleem bezig gehouden. De definitieve CeD diagnose kan alleen worden gesteld door biopent te nemen van de dunne darm, een invasieve en belastende methode. In 2011 kwamen er aanwijzingen voor nieuwe veelbelovende biomarker kandidaten in de vorm van circulerende micro-RNAs (miRNAs). Dit zijn kleine stukjes RNA (ongeveer 22 nucleotiden lang), die niet voor eiwitten coderen maar die kunnen zorgen voor de verlaging van specifieke messenger RNAs (mRNAs). Onderzoek liet zien dat miRNAs uitzonderlijk stabiel zijn en voorkomen in onder andere bloed, liquor, urine en zelfs in traanvocht. In het profiel van circulerende miRNAs in andere ziektes zijn miRNAs gevonden die specifiek bleken te zijn voor bepaalde ziektes en deels zelfs ziektestadium-specifiek. Deze eigenschappen maken miRNAs uitstekende kandidaten om als biomarkers te dienen.

Het onderzoek gebundeld in dit proefschrift had twee hoofddoelen: (1) om de functionele SNPs en genen te prioriteren die in de CeD loci gelegen zijn, zoals die door de CeD Immunochip studie waren gedefinieerd, en (2) om te bepalen of circulerende miRNAs goede biomarkers zijn voor coeliakie.

Opbouw van het proefschrift

In hoofdstuk 1 geef ik een overzicht van de genetische factoren die zowel voor en tijdens het GWAS tijdperk zijn gedefinieerd. Daarnaast bespreek ik hier de CeD-geassocieerde genen en onderliggende mechanismen die worden gedeeld met andere immuun-gemedieerde aandoeningen.

In hoofdstuk 2 laten we zien dat de functionele SNPs, genen en onderliggende mechanismen die bijdragen aan het ontstaan van CeD, gevonden kunnen worden door middel van een integratieve “functional genomics” en “pathway analysis” aanpak. Hiervoor zijn zowel nieuwe als publiekelijk beschikbare data gebruikt. Een van onze bevindingen was dat vier slecht beschreven genen een rol zouden kunnen spelen in de functie van de intestinale barrière waarvan al eerder was aangetoond dat deze verstoord is in CeD. Ook beschrijven we de relatie tussen SNPs in CeD kandidaatgenen en de verhoogde expressie van Interferon-. Dit heeft inzicht gegeven in hoe IFN- verstoord kan zijn in CeD, zonder dat er zich een SNP in het IFNG locus bevindt.
Van alle niet-HLA CeD geassocieerde loci, is de relatie van het LPP locus met CeD het sterkst. In hoofdstuk 3 presenteer ik een “fine-mapping” methode, waarmee het mogelijk bleek om de grootte van de geassocieerde regio te verkleinen. Door haplotype analyse en gentegreerde “functional genomics” methode werd het gebied van 70 kb naar 2,8 kb gereduceerd. Door ENCODE data te gebruiken, kon ik de SNPs in LPP sorteren op waarschijnlijkheid van causaliteit. Uit deze analyses kwam naar voren dat rs4686484 een zeer goede kandidaat is voor de causale SNP in het LPP locus.

Uit recente meta-analyses in andere ziektes kwam dat SNPs in de LPP regio een “expression quantitative trait locus” (eQTL) effect hebben op het BCL6 gen. Dit gen ligt 658,7 kb van het LPP gen af, buiten het eigenlijke LPP locus. In hoofdstuk 4 heb ik gekeken naar het effect van SNPs in het LPP locus op de expressie van andere genen, met behulp van RNA-sequencing data. Hierbij vonden we een eQTL effect van een van deze SNPs op LPP-AS1, een zogenaamd “long non-coding RNA” (lncRNA). Dit is de eerste keer dat een lncRNA gemplcieerd werd in CeD. Co-expressie analyse liet zien dat dit lncRNA een rol zou kunnen spelen in ubiquitinering. Verdere functionele studies zijn nu nodig om te bewijzen dat LPP-AS1 een van de causale genen is in CeD.

In hoofdstuk 5 geven we een overzicht van wat er bekend is over de invloed van genetische variatie op regulerende niet-coderende RNAs, met als focus IncRNAs en miRNAs. In hoofdstuk 6 heb ik “next generation sequencing” gebruikt voor het bepalen van miRNA profielen in serum van CeD patiënten uit het PreventCD cohort. PreventCD is een prospectieve studie waarbij van kinderen met een hoog (familiair) risico op CeD direct na de geboorte samples genomen zijn op vaste tijdpunten tot op het moment dat sommige van deze kinderen inderdaad CeD ontwikkelden. We vonden 45 miRNAs die differentieel tot expressie kwamen op het moment van diagnose ten opzichte van de profielen in de samples die 3 maanden na de geboorte zijn genomen. Ook vonden we dat een aantal miRNAs al veranderden voordat de diagnose gesteld werd, en dat sommige van deze miRNAs weer lijken te normaliseren na de start van een glutenvrij dieet. Deze miRNAs kunnen worden toegevoegd aan de lijst van potentiële biomarker kandidaten voor CeD.

In hoofdstuk 7 beschrijf ik miRNA profielen in het bloed en in dunne darm biopsieën die afkomstig zijn van een ander cohort van pediatrische CeD patiënten en controle individuen. Deze analyse liet een panel zien van 49 circulerende miRNAs en 109 miRNAs in biopsieën die differentieel tot expressie kwamen in CeD patiënten in vergelijking met controles. Onder de in de dunne darm tot expressie komende miRNAs kunnen miRNAs zijn die betrokken zijn bij de pathogenese
van CeD.

In hoofdstuk 8 plaats ik de resultaten van het onderzoek, gepresenteerd in dit proefschrift, in de context van de actuele kennis. Ook belicht ik de uitdagingen waarvoor de wereld van de huidige genetica op dit moment staat, waarin de hoeveelheid beschikbare genetische data exponentieel aan het groeien is.

De zes hoofdconclusies van dit proefschrift zijn:

- Verschillende lagen van publiekelijk beschikbare gegevens ("functional genetics data"), kunnen worden gebruikt om causale SNPs, genen en signaleringsroute’s te identificeren in CeD gassocieerde loci.
- CeD SNPs liggen meestal in regio’s van het genoom die een regulerend effect op genexpressie hebben.
- Haplotype analyse in specifieke bevolkingsgroepen is een krachtige methode om ziekte-geassocieerde loci te verkleinen.
- CeD-geassocieerde SNPs kunnen de expressie van IncRNAs beïnvloeden.
- Circulerende miRNAs zijn potentiële kandidaten voor biomarkers die kunnen worden gebruikt in CeD diagnose of in het monitoren hoe goed patiënten zich houden aan het glutenvrije dieet.
- Zelfs in weefsels met een rijk scala aan verschillende cellen, zoals intestinale biopten, kunnen CeD-specifieke miRNAs worden gedetecteerd. Sommige van deze miRNAs zouden een rol kunnen spelen in de pathogenese van CeD.
Appendices

Summary
Samenvatting
Resumo
Acknowledgements
Curriculum Vitae
Resumo
Nosso corpo é constituído de aproximadamente 1 quadrilhão (1.000.000.000.000.000) de células individuais. O material genético - nosso genoma - no núcleo de cada uma dessas células que determina não só as características como a cor ou o tamanho de nossos olhos, mas que também abrange a nossa susceptibilidade individual para desenvolver determinada doença. Nosso genoma consiste de ácido desoxirribonucleico (DNA). A função do DNA é baseada na ordem - a sequência - de apenas quatro blocos chamados nucleotídeos. O nome completo desses nucleotídeos é adenina, timina, citosina e guanina, os quais são sempre abreviados como A, T, C e G, respectivamente. O DNA humano é organizado dentro de 23 unidades chamadas cromossomos que juntas formam 3.3 bilhões (3.300.000.000) de nucleotídeos.

Desde a total determinação da sequência do DNA humano em 2001, novas tecnologias vem sendo desenvolvidas que permitem uma rápida averiguação da sequência de um nucleotídeo individualmente. Uma das maiores descobertas desses tipos de estudos foi que o DNA humano difere em apenas 0.1% entre indivíduos diferentes (não aparentados). Isso significa que cada indivíduo difere do seu vizinho, por exemplo, em 3 milhões de posições de nucleotídeos. A sequência do DNA de centenas de milhares de indivíduos tem possibilitado o mapeamento de variantes de nucleotídeos (também chamados polimorfismos de nucleotídeo único (SNPs - do inglês single nucleotide polymorphisms) através do genoma humano. Surpreendentemente, este tipo de mapeamento também mostrou que SNPs próximos uns dos outros têm uma correlação muito forte entre eles levando o genoma a possuir uma estrutura de blocos chamada de haplótipos. Consequentemente, não são necessários os 3 milhões de diferentes SNPs para localizar fatores genéticos associados a características e doenças, mas cerca de 500.000 SNPs que identificam os blocos de haplótipos diferentes já são suficientes. Em 2007, esta nova perspectiva de uma forma mais simplificada do mapa da variação do genoma humano levou a uma revolução no campo da genética, pois pela primeira vez permitiu pesquisadores realizarem estudos genéticos de associação por todo o genoma humano em milhares de indivíduos a relativamente baixo custo. Estes estudos são chamados de estudos genômicos de associação (GWAS- do inglês genome-wide association studies). Até o momento, esse tipo de estudo tem revelado os fundamentos genéticos de mais de 400 diferentes doenças e traços genéticos, variando desde consumo de café e altura até esquizofrenia e doenças autoimunes.

GWAS são estudos do tipo
caso-controle, nos quais a frequência de um conjunto de 500.000 SNPs em todo o genoma são comparados entre milhares de amostras controles e milhares de indivíduos carregando uma característica em particular ou doença. Até o presente momento, dois GWAS bem sucedidos foram realizados em coortes de pacientes com doença celíaca (DC), estes estudos forneceram novas percepções em vias biológicas da doença.

DC é a intolerância alimentar que apresenta a maior prevalência conhecida atualmente, afetando cerca de 1% da população Ocidental. Indivíduos geneticamente predispostos para DC desenvolvem uma intolerância aos peptídeos do glúten derivados de proteínas presentes em altos níveis em grãos que estão na base da dieta normal Ocidental. Consequentemente, pacientes com DC não podem comer alimentos preparados (ou potencialmente contaminados) com glúten, como por exemplo pão, biscoitos e macarrão. DC está fortemente associada a dois genes do complexo principal de histocompatibilidade (MHC - do inglês *major histocompatibility complex*), especificamente as moléculas do HLA-DQ2 e HLA-DQ8. Indivíduos portadores do genótipo do HLA-DQ2 apresentam cerca de 35% do risco genético de desenvolver a DC. No entanto, aproximadamente 25% da população geral carrega o genótipo HLA associado a DC, então somente o HLA não é suficiente para um indivíduo desenvolver DC.

GWAS na DC identificaram 26 outros fatores genéticos além do HLA, os quais também são importantes para o desenvolvimento da doença. Uma plataforma de genotipagem customizada (Immunochip) foi desenvolvida para refinar o mapeamento de loci associados a doenças autoimunes, porém o uso dessa plataforma também permitiu que o nosso grupo de pesquisa identificasse 13 loci não pertencentes ao sistema HLA associados a DC, fazendo com que o número total de loci genéticos conhecidos aumentasse para 40 (locus HLA e mais 39 loci não pertencentes ao HLA).

Quando eu iniciei meu doutorado em 2010, estudos genéticos em doenças autoimunes estavam enfrentando um grande obstáculo: as regiões genômicas descobertas pelos GWAS que foram associadas a doenças geralmente são bastante grandes e em consequência disso, foi preciso desenvolver novas abordagens afim de que em estudos futuros pesquisadores fossem capazes de priorizar SNPs e genes com maior probabilidade de serem os causais nesses loci associados. Outro assunto que emergiu em seguida, foi que em muitos loci, os SNPs que estavam mais fortemente associados com a doença se encontravam
mapeados em regiões não codificantes de proteína. Em alguns casos, não existia absolutamente nenhum gene que codificava proteína localizado no locus e estes loci foram descritos como “desertos de genes”. Um dos principais desenvolvimentos durante o curso de meus estudos foi a realização do projeto Enciclopédia de Elementos do DNA (ENCODER - do inglês Encyclopedia of DNA Elements) em 2012. O objetivo dessa enciclopédia foi descrever e mapear o “genoma funcional”: isto não apenas revelou milhares de novos transcritos, como também mapeou centenas de milhares de regiões reguladoras que controlam a expressão gênica.

Um dos problemas clínicos da DC que eu abordei durante meus estudos foi o desafio em diagnosticar a DC facilmente e com acurácia. Atualmente, um diagnóstico definitivo é realizado por um procedimento invasivo para obter uma biópsia intestinal. Tem sido estimado que apenas um em cada oito pacientes com DC é corretamente diagnosticado. Em 2011, uma nova classe de RNAs não codificantes emergiu como candidatos promissores para biomarcadores, até então sendo chamados de micro-RNAs (miRNA) circulantes. MiRNAs são pequenos RNAs não codificantes de cerca de 22 nucleotídeos que regulam a expressão de específicos mRNAs alvos. Foi encontrado que miRNAs são extraordinariamente estáveis em diferentes fluidos do corpo e que eles podem ser detectados em sangue, líquido cefalorraquidiano, urina e até mesmo em lágrimas. Além disso, tem sido mostrado que o perfil de miRNAs circulantes pode revelar alguns desses miRNAs que são específico para doenças e ainda específicos para estágios da doença. Essas características fazem dos miRNAs circulantes excelentes candidatos para serem usados como biomarcadores.

Na pesquisa descrita nesta tese eu abordei dois objetivos principais: (1) identificar SNPs e genes funcionais nos loci descobertos pelo estudo do Immunochip em DC, e (2) determinar se miRNAs circulantes podem atuar como bons biomarcadores para doença celíaca.

Esboço da tese

Eu forneci uma visão geral de fatores genéticos identificados por estudos pre-GWAS e por GWAS que foram associados a DC no Capítulo 1. Posteriormente, eu discuto os genes associados a DC, assim como vias biológicas que são compartilhadas com outras doença autoimunes.

No Capítulo 2, nós mostramos que aplicando uma abordagem integrativa de genômica funcional e análises de vias biológicas, usando novos dados disponíveis para o público, nós podemos priorizar SNPs funcionais, genes e vias que contribuem para DC. Um de nossos achados foi que quatro genes que estavam pobremente descritos anteriormente, podem ter um papel
na função da barreira intestinal, um processo que é conhecido por estar alterado em DC. Nós também descrevemos uma conexão transcricional entre Interferon-γ (IFN-γ) e genes de suscetibilidade a DC, o que esclarece como IFN-γ está desregulado na DC, embora o locus IFNG não esteja diretamente afetado pelos SNPs associados a DC.

O locus fora do HLA mais associado a DC é o LPP locus. No Capítulo 3, eu apresentei uma abordagem de mapeamento fino que reduziu o tamanho da região associada. Aplicando análises de haplótipos e métodos de integração de genômica funcional em dados disponíveis para o público, nós fomos capazes de reduzir esta região de 70 kb para apenas 2,8 kb. A priorização de SNPs usando dados do ENCODE sugeriu que o SNP rs4686484 poderia ser o SNP funcional nessa região.

Recente estudos de meta-análises em outras doenças identificaram SNPs na região do LPP apresentando um efeito na expressão gênica (eQTL, do inglês expression quantitative trait locus) no gene BCL6, o qual está localizado 658,7 kb distante do gene LPP e fora do vigente locus LPP. No Capítulo 4, usando dados de RNA-seq eu explorei a possibilidade de que SNPs no locus LPP poderiam estar afetando a expressão de outros genes. Esta abordagem nos permitiu determinar um efeito eQTL de um desses SNPs no locus LPP em um RNA longo não codificante (IncRNA, do inglês long-non-coding RNA) LPP-AS1. Esta é a primeira vez que um IncRNA foi implicado na DC. Análises de expressão específica para o tipo de célula, e análises de enriquecimento de vias sugeriram que este IncRNA tem um papel em ubiquitinação. Estudos funcionais são necessários no momento para provar que o LPP-AS1 é um dos genes causais da DC.

No Capítulo 5, eu forneço uma revisão sobre como variações genéticas podem ter um impacto regulatório em RNAs não-codificantes, com um foco principalmente em IncRNAs e miRNAs.

No Capítulo 6, eu usei o sequenciamento de nova geração (NGS, do inglês next generation sequencing) para realizar um perfil de miRNAs circulantes em soros de pacientes com DC que estavam participando da coorte do projeto PreventCD. O PreventCD é um estudo prospectivo em que amostras foram coletadas em séries de tempos fixos depois do nascimento da criança sabendo que apresentava alto risco (familiar) de desenvolver a DC. Nós encontramos um painel de 45 miRNAs diferencialmente expressos na hora do diagnóstico da DC comparando com o perfil de expressão visto em crianças com 3 meses de idade. Nós também encontramos um grupo desses miRNAs que já estão alterados antes da doença ser diagnósticada, enquanto alguns miRNAs aparecem normais no início de uma dieta livre de...
glúten. Estes miRNAs fornecem um catálogo de candidatos promissores a biomarcadores para a DC.

No Capítulo 7, eu descrevo dois perfis de miRNAs em amostras de sangue e de biópsias do intestino delgado obtidas de outra coorte pediátrica de pacientes com DC e indivíduos controles. Esta análise revelou um painel de 49 miRNAs circulantes em plasma e 109 miRNAs em biópsias que estavam diferentemente expressos em pacientes com DC comparados a controles. Nós também identificamos miRNAs que estão diferencialmente expressos no intestino delgado de pacientes com DC que podem estar envolvidos na patogênese da DC.

No Capítulo 8, eu apresento os resultados de pesquisa desta presente tese no contexto do conhecimento atual e ainda aponto os desafios emergentes no campo da genômica, onde a disponibilidade de dados está crescendo exponencialmente.

As seis principais conclusões desta tese são:

- Diferentes camadas de dados de genômica funcional disponíveis podem ser usadas para priorizar SNPs, genes e vias na DC.
- SNPs associados a DC em sua maioria estão localizados em regiões que tem uma função reguladora.
- Análise de haplótipos em populações estratificadas é um método poderoso para mapeamento fino de regiões associadas a doenças.
- SNPs associados a DC podem afetar a expressão de IncRNAs. miRNAs circulantes são potencialmente candidatos a biomarcadores para detectar a DC e como o paciente adere a dieta livre de glúten. Até mesmo em complexos de misturas de células como biópsias intestinais, miRNAs específicos para DC podem ser detectados. Alguns destes miRNAs podem estar envolvidos na patogênese da doença.
Appendices

Summary
Samenvatting
Resumo
Acknowledgements
Curriculum Vitae
Looking back on these four years, I realized how important it was to leave Brazil and go abroad and do my PhD. Besides the science part, going abroad was a whole life changing experience for me. I was really lucky to have so many good people around me in Groningen. The work environment was amazing, full of discussion and friendly collaboration and support from different colleagues with different backgrounds. This definitely made a big difference in my PhD journey. I can say that I was in the right place at the right time and most importantly, with the right people. I would like to thank all the people who were involved directly or indirectly in my PhD. Thank you all.

Dear Cisca, I still remember the day in 2010 I got your email saying that my sandwich PhD project could fit into one of the ongoing projects in your group. It was a very exciting moment for me. Further, you gave me the opportunity to collaborate in different projects in which I could learn and develop new skills, which I didn’t even imagine that I could do them. Thanks for helping me to build a more critical, scientific way of thinking. Among a bunch of other things, I would like to specially thank you for you kindness, guidance, mentorship, leadership and teaching.

Dear Sebo, thanks for show me the microRNA world and its possibilities, for all our discussions and your teaching me about biology. I believe that all the discussions we had laid a new brick in my scientifical path. Thanks for your support and guidance, my PhD would not have been done on time without your help and involvement.

Dear Vinod, you were a kind of guru for me. With your singular patience, you showed me how to analyze the data without too much stress, using a step-by-step manner that changed my vision of data analysis. I am glad that someone with such wisdom crossed my PhD path. Thanks a lot.

I would like to thank my supervisor, Riccardo Pratesi, and my co-supervisor, Lenora Gandolfi, from University of Brasília, Brazil. Without both of you this work could not have progressed the way it did. “Obrigado mais uma vez pelo apoio e carinho.”

It is with immense gratitude that I acknowledge the support and help of my parents, Cícero e Terezinha. “Obrigado pela confiança, carinho e amor.”

Dear Gosia, thanks for show me how big data analysis works and for all your insights into the LPP project. You really made me believe I could do it. I still remember your face when I was giving a presentation at the ESHG in Amsterdam, I think you were as nervous as I was, but in fact, it really made a difference for me back then. Besides, you always knew how to
combine fun and work, thanks for being part of this PhD. I consider it an honor to have worked with you.

Dear Jihane, I still remember when I met you for the first time in São Paulo and asked you about the group in the Netherlands. I remember you said “Do you like computers?” Without your sincerity I would never have been able to decide to go and do a PhD in the Netherlands. “Obrigado!”

Dear Isis, you are a sister that I met in Groningen and I will care for my whole life. There is no doubt why I chose you to be one of my paranymphs. Thanks for your true friendship, kindness and hospitality. I will never forget the uncountable dinners and parties in your house, I felt as if I was in a real Latin family on those days. I also would like to thank you for all the discussions we had through these years, especially your collaboration on the LPP paper. My PhD and my life in Groningen would not have been the same if you hadn't been part of it. And, of course, I cannot forget to thank you for helping to improve my Spanish, or should I say my Portu-ol. “Muchas gracias Chiquitina y calma chiquita, calma.”

Dear Javier, thanks for your friendship, since the beginning you were a good friend for me. I will never forget the day we met and you immediately invited me to go to your house and have lunch. That really made me feel part of your family. Apart from this, thank you for all your help with statistics and programming, you really showed me how to get it done. I really enjoyed all the discussions we had. I still remember how much fun we had writing a review together. I cannot imagine my PhD without you Javi, gracias amigo.

Dear Olga, you are my Portuguese/Chinese friend who is always there for me when I need someone. Thanks for your true friendship, I really appreciate all your efforts and everything you have done for me. I was always overwhelmed by how you can quickly organize stuff in your life, it’s a special gift you have that I admire. I’m thankful that I had you to speak Portuguese to, especially in the last year in Eriba when we met for a coffee, this made a real difference in the last year of my PhD. Olguinha, you are a friend for life, my paranimf, minha amiga, muito obrigado.

I would like to express my gratitude to the two editors who helped me to put this PhD into the right words, Jackie and Kate. Special thanks to Jackie, for being patient and kind all the time. I cannot find words to express my gratitude to you. Thank you.

Dear Urmo, thanks for teaching me how to customize my microRNA analysis in R. You came to Groningen at the right time for me. I’m glad we also had a lot of fun outside the lab. Thanks for your guitar talent.
Dear Harm-Jan, thanks for all your eQTL mapping help. Thanks for opening up your bat cave full of fun song and many others musical instruments to play with music. It was really fun to make that blues in your house and also to have that Scary Movie session. Thanks for sharing your skills and friendship.

Dear Juha, thanks for all the philosophical discussions we had during these years. I still remember that trip we made to Slovenia, when I thought we would miss the plane because we were delayed on the train, and you just looked really calm, like nothing was happening, and said “No need for hurry”, indeed you made me calm down a bit. Thanks for being part of this PhD journey.

Dear Sasha and Lude, I’m really thankful to have the two of you in my PhD life. Thanks for sharing all of your thoughts and work discussions that made this thesis a valuable piece of work.

I am indebted to my many colleague bioinformaticians who supported me during my PhD. Patrick, thanks for your help and patient to show me the big data analysis of RNA-seq. Marc Jan, thanks for showing me many R scripts and helping me to understand this amazing tool. Dasha, thanks for sharing some of your scripts and helping me to get the right tools to do the alignments of microRNAs. Marijke thanks for helping me with the expression analysis. Jingyuan, thanks for helping with the cluster analysis. Alexandros, thanks for your help with the imputation and sharing your Greek culture with me. Jennifer, thanks for helping me with the microRNA scripts. Pieter and Freerk, thanks for all your help with the Next Generation Sequencing analysis, and many other bioinformatics stuff. Yang Li, thanks for all your help with the RNA-seq analysis, and your patience to explain it all for me.

Dear Barbara, thanks for all the discussions and talks we had in these four years, for sharing your happiness and yoga moods with me.

Dear Omid, you were really the boss back in those days in the office in the old building. Thanks for sharing a bit of your amazing culture with me, this is what makes the department such a great international environment to work.

It is with immense gratitude that I acknowledge the support and help of all the staff members of the Department of Genetics. Hélène, Mentje, Joke, Marina and Edwin. Special thanks to Hélène, for all patience and work efficiency that helps me to get things done.

I am indebted to all the members of the Genomics Analysis Facility. Thanks Pieter, Jelkje, Cleo and Bahram for helping with the NGS experiments. Special thanks to Bahram, for sharing
your culture with me and all the smart talks we had during coffee breaks.

Dear Rutger, I owe my deepest gratitude to you. Without all your efforts and help, this thesis would not have been completed. Thanks for making my work environment in the lab more fun than ever. You’re the best, thanks.

It gives me great pleasure to acknowledge all the technicians. Mathieu, Soesma, Astrid, Michiel, Ludoff, Kim, Krista and Jan, thank you very much for all support and help during these four years.

I would like to thank all the IBD group for all their help and friendly collaboration. Rinse, Suzanne, Karen, Marijn, Noortje, Rudi, Floris and Lieke, thank you all.

Special thanks to my Spanish-speaking friends. Alejandro, Anna Duari, Mamem, Luz Maria, Maria Zorro, Genaro, thanks for all the fun we had. Without you all, my PhD life in Groningen would not have been the same and I would not have had so much fun as I had.

Dear Cleo Smeets, thanks for all the fun time, drinks and talks at O’Malley’s or Pacific. I will never forget your attitude to go to concerts and festivals and have a lot of fun there. I’m glad you were part of my PhD life. Thanks sweetie.

Dear Claudia, thanks for the design and figures you did for my thesis. Your design is spectacular, you’re a true artist.

Dear Mitja, it was really fun to have you around in the first year. I will never forget my stay in Slovenia and your interesting culture, thanks for showing me.

Dear Senapati, thanks for your friendship, it means a lot for me. I really enjoyed the time when you were in Groningen. I’m glad that I had such a nice person as part of my PhD life.

I would like to thank my Polish “mafia” colleagues, Asia, Agata, Justyna, Zuzanna and Ania. Special thanks to Anna to all your help with the miRNA analysis.

I would like to thank my colleagues, Ette, Dineke, Esther, Jackie Dekens, Yunia, Rajendra, and Iris Jonkers, for their support and help.

Lieve Hendrika, thanks for all your support and help to finish this PhD. Thanks for being patient and always understanding at the most stressful moments I had. You were a real friend and partner in all these moments, it helped me a lot to go through this entire journey.

I would like to thank all the Merkus family who made me feel part of their family. Thank you all for your kindness. I cannot forget to thank my dear
friends from Brazil. Patrícia and Eduardo, I really appreciated it when you came to visit me in Groningen. I will never forget our discussions back in the days at the UnB. Obrigado. All the people from the Pediatric Lab in Brazil, Rita, Lilian, Yanna, Danilo and Lucas, thanks to all. My Bang ou Tec friends, Paulo Rená, Diego, Pedro, Nathália and Raphael, thanks for all the chats we had in the four years I spent in the Netherlands, it really made me feel that I still had friends in Brazil. Special thanks to Marcos and Drica for all your kindness and true friendship. I really enjoyed my visits to you in Sweden and when you came all the way to visit me in Groningen. It meant a lot to me.

A very special thanks to Juliana Mazzeu from the University of Brasília, who in 2009 told me about a friend who was working in a great group in the Netherlands. This turned out to be Jihane. Ju, thanks for you’re your advice, without you I would not have thought about going abroad for my PhD.

I cannot express my gratitude enough to all my family, who supported and believed in me. Specialy my sisters Juliana, Gabriela and Fernanda, my brother Bernardo, my cousins, Marina, Luisa, Pablo, Arthur and all the others “Netos da Didica”. Thanks to all my tios, tias and my avó Nadyr. Special thanks to my aunt Marilucia, I really appreciated your visit to me. My aunt Terezinha, thanks for all the inspiration and for believing in me. Thank you all.

I could not forget to thank the people in my dear student house on Blekeslaan. Thank you all for being part of my life. Special thanks to Ana Barreto, my Portuguese friend who made my life much more fun when I arrived in that place. Rene, my student manager, I really enjoyed our trip to Finland. Steffen, thanks for supporting me. Special thanks also to Fisnik and Aimé, guys you are friends for life. Elles, you are my Brazilian/Dutch friend who never let me down. Thanks for your friendship in all these years.

Thank you all, dank je wel, gracias, obrigado!
Appendices

Summary
Samenvatting
Resumo
Acknowledgements
Curriculum Vitae
Rodrigo Coutinho de Almeida was born in 9th November 1980 in Bom Jesus da Lapa, Brazil. In 2004 he graduated in Biology at the University of Mato Grosso, Brazil. In 2008 he finished his MSc in Health Science at the University of Brasilia, Brazil, where he worked on the epidemiology of celiac disease in Brazil. In 2011 he started a ‘sandwich PhD’ between the University of Brasilia and the University of Groningen, focusing on the genetics of celiac disease at the Department of Genetics, University Medical Center Groningen, the Netherlands, under the supervision of Prof. Cisca Wijmenga. In 2012 he was awarded a De cock foundation grant to work with microRNA profiling. Recently, he has been awarded a Young Talent fellowship from the Brazilian federal government to work as a post-doc at the University of Paraná, Curitiba, where he will be involved with next-generation sequencing, microRNAs and gene expression in complex diseases.

Publications

*Both authors contributed equally to the manuscript*


