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.
PART II

Chapter 8

Discussion and future perspectives
Genetics and fine mapping of celiac disease associated regions

Genome-wide association studies (GWAS) have discovered thousands of loci in a broad range of complex diseases (1). The design of GWAS (and all fine-mapping approaches) is based on the principle of linkage disequilibrium (LD) at the population level (Chapter 1). The LD pattern is critical in reducing the number of genotypes that need to be generated for hundreds of thousands of genetic markers simultaneously (instead of the millions of genetic variants present in every genome). However, the price to be paid for this simplification is that the LD structure results in association of many closely correlated single nucleotide polymorphisms (SNPs) which can make it more difficult to determine the one or few truly casual SNP(s) in the LD block. The first part of this thesis describes different ways of performing fine-mapping approaches in celiac disease (CeD) that help to zoom in on functional SNPs, as well as on genes and the pathways related to the disease.

In chapter 3 we focus on the strongest non-HLA CeD-associated loci: the LPP locus. Although CeD has a worldwide distribution, it is more prevalent in Caucasians (2). On average, the European population carries more highly disease-correlated SNPs and longer haplotype blocks than other populations (i.e. compared to African and Asian populations) (3), and these two factors enable us to zoom in on associated LD blocks in this population. By focusing only on European populations we were able to fine map this locus by application of haplotype analysis and reducing the block size of 70 kb down to 4 kb. Subsequent imputation, re-sequencing and integration approaches were used to further refine the associated area to a 2.8 kb region.

Integrating data to reveal biology

Although haplotype analysis can reduce the size of a disease-associated LD block, it needs to be combined with biological features, such as DNase hypersensitivity sites, for further prioritization of functional SNPs. After the Encyclopedia of DNA Elements (ENCODE) (4) data became publicly available, it was possible to integrate GWAS signals with functional genomics data. This provided new layers of information which showed that most GWAS SNPs are in fact concentrated in regulatory regions of the genome (5, 6). We integrated several ENCODE data layers with our own data, which helped to interpret and prioritize SNPs from the CeD GWAS loci (Chapter 2 and 3). In Chapter 2 we confirm that a large majority of CeD-associated SNPs are more likely to influence disease through regulatory mechanisms. In agreement with this, we found that the fine-mapped LPP region that we describe in Chapter 3 is enriched for
enhancers, DNase I hypersensitive sites (DHS), and histone modification signatures. Similar data integration approaches have recently been used and recommended by other studies to prioritize functional SNPs (5, 7). For example, a trans-ethnic fine-mapping study of the TNFSF4 autoimmune locus integrated genetic data with ChIP-seq data from ENCODE (4), which allowed the authors to identify a functional genetic variant that changes the sequence motif binding of the NF-kB transcription factor (8).

Another layer of information that yields biological insights into the role of candidate genes is data generated by enrichment and pathway analysis. This approach can help highlight important pathways involved in disease development and even pinpoint pathways not previously associated with the disorder under investigation (9). In Chapter 2 we prioritized CeD-genes in complex loci by performing pathway enrichment analysis on genes prioritized by expression quantitative trait locus (eQTL) mapping. By doing so, we confirmed the involvement of Th1- and Th17-related pathways in CeD. Moreover, we implicated four poorly characterized CeD genes (C1ORF106, ARHGAP31, LPP and PTPRK) as playing a role in intestinal barrier function. Although it has been established that the barrier function is impaired in CeD (10), the genetic factors contributing to this part of the phenotype were not known before our study. Lastly, we exposed an intricate transcriptional connectivity that converges on IFN- expression. This finding might explain why IFN- is dysregulated in CeD even though the IFN- locus has not, to date, been implicated in this disease (Chapter 2).

Causal SNP versus causal gene
A challenge that remains for GWAS is to confirm that an associated SNP actually affects the gene to which it maps. Although the data presented in Chapter 3 strongly suggest that the causal genetic variant contributing to CeD is located in the LPP locus, we were unable to unequivocally demonstrate that LPP is the causal gene. Interestingly, two meta-analysis studies describe allergy-associated SNPs present in the same LD block as our prioritized CeD SNP. They report that these SNPs correlate with downregulation of the BCL6 gene, which is located more than 600 kb away from the LD block (11, 12). These observations strengthen the hypothesis that causal genes can be located far from causal SNPs, as we describe in Chapters 2 and 3. Our own eQTL mapping studies for CeD-associated SNPs also revealed two of these events. A region on chromosome 2p16.1 were eQTL mapping predicts the AHSA2 gene located ~ 160 kb away from the LD-based fine-mapped region which predict the PUS10 gene, and another region on chromosome 3p21.31, were CCR3 showed an eQTL effect which is located ~ 169 kb away from the LD-based fine-mapped LTF
gene. A similar example was reported by Smemo et al. (2014) for the FTO locus associated to obesity (13) and type 2 diabetes (14). Since the first report of this genetic association in 2007, more than 400 papers have been published investigating the role of the FTO protein in obesity. It is now known that the lead SNP displays an eQTL effect on the IRX3 gene located ~1.2 megabases away from the associated FTO locus (15). This eQTL effect is only evident in brain tissue, which probably explains why the connection went unnoticed for several years. Moreover, this finding stresses the need for cell-type-specific eQTL analysis. Smemo et al. also demonstrated that IRX3 is functionally connected to obesity using mouse knock-out studies (15). Altogether, these examples support the idea that a disease-associated SNP will not necessarily affect the gene in which it is located, but instead may be regulatory in origin and affect another gene nearby or even megabases away from the genes in the LD block (potentially even on other chromosomes).

eQTL mapping to explain genetic components

eQTL mapping is a powerful approach to explain the genetic component underlying altered gene expression, and it helps to prioritize disease SNPs and genes. Complex trait-associated SNPs are more likely to be eQTLs (16). For instance, 39 out of 71 loci (55%) associated to Crohn’s disease showed an eQTL effect (17). Likewise, we observed that SNPs from 18 out of 39 CeD loci display eQTL effects (Chapter 2). The reason we do not find eQTL effects for the remaining 21 CeD loci may be the cell-type-specificity of the remaining eQTL effects. Unfortunately, there is not enough tissue-specific expression data currently available to allow for comprehensive organism-wide eQTL analysis. It has recently been reported that disease-associated SNPs are more likely to affect gene expression in a cell- or tissue-dependent manner (18, 19), further demonstrating the complexity of discovering these eQTL variants. We have shown that the SNPs associated with CeD are enriched for B-cell-specific enhancers (Chapter 2 and 3), suggesting that B cells play a major role in CeD. The importance of B cells in CeD has only recently been appreciated, and our results emphasize that this research needs to be developed further. A possible reason why B-cell-specific eQTL variants have been missed so far may be that eQTL studies have mostly been performed in complex cell mixtures, such as blood and peripheral blood mononuclear cells (PBMCs), while B cells only contribute ~5% to the cellular population in PBMCs (20). To find B-cell-specific variants, it will be important to perform eQTL mapping specifically in B cells. Fairfax et al. (2012) identified 730 GWAS genes displaying cis-eQTL effects, of which 33.5% showed B-cell specificity (21). They also
reported that certain diseases are characterized by a predominance of cell-type-specific eQTLs, such as systemic lupus erythematosus where 10 out of 16 genes displaying a cis-eQTL effect were B-cell specific (21). An important theme emerging is that many eQTL effects will only be unveiled in the context of stimulatory or activating signals (22). Thus far, most eQTL studies have been performed on resting cells or cell populations.

Another important feature of disease-associated SNPs is that ~90% of GWAS SNPs are located outside protein-coding sequences (1). In CeD, specifically, only 5% of the associated SNPs are located in coding sequences (Fig. 1), strongly suggesting that many of the CeD-associated variants affect gene regulation. How these non-coding variants affect gene regulation is still largely unknown, but it is likely that these SNPs affect gene expression through transcriptional, post-transcriptional, and post-translational mechanisms. Additionally, it has recently been shown that GWAS SNPs can also influence expression levels of long non-coding RNAs (lncRNAs) (23).

In Chapter 4, by eQTL mapping using RNA-sequencing data, we show that a CeD associated SNP downregulates a lncRNA (LPP-AS1). This suggests that this lncRNA plays a role in CeD. Pathway enrichment analysis of genes co-expressed with this lncRNA suggests a role for LPP-AS1 in ubiquitination. One of the functions that has been suggested for lncRNAs is that they could act as scaffolds to form multiprotein complexes (see Chapter 5, Fig. 5). Interestingly, the lncRNA HOTAIR was found to function as a scaffold RNA for E3-ligases and their substrates (24). We therefore hypothesize that deregulation of LPP-AS1 affects ubiquitination, which is known to be important for immune function. The finding that CeD GWAS and Immunochip analysis results also point at proteins involved in ubiquitination strengthens this assumption (25), but functional studies will be necessary to prove this hypothesis.

Little is known about eQTL effects on non-coding genes. Most eQTL mapping studies have been performed using microarray technology, which was originally designed to contain only probes for protein-coding genes. Although some lncRNAs probes are present on commercial microarrays (23), RNA-based sequencing (RNA-seq) is currently the most promising way to comprehensively profile known lncRNAs and identify novel tissuespecific lncRNAs. A wealth of RNA-seq data is now being generated from many different tissues. Novel eQTLs studies focusing on RNA-seq data are likely to identify many more genetic variants correlating with lncRNA expression.

**Non-coding elements**
The ENCODE data shows that the human genome is even more complex
than previously assumed. It further confirmed that there are thousands of regulatory RNAs, including IncRNAs and microRNAs (see Chapter 5 for an overview). Moreover, the data has been used to provide additional evidence for the fact that regulatory genomic regions can interact with target regions long distances (megabases) away, for instance if these regions are brought into contact with target promoters by DNA folding. Several high-throughput approaches can be applied to identify chromatin interactions and long-distance enhancers (Table 1). These approaches can help to link non-coding SNPs in regulatory regions to their target genes. Interestingly, a subclass of IncRNAs, the eRNAs, seems to be specifically involved in enhancer looping.

As described above, most of the CeD SNPs map to non-coding regions (Fig. 1), which is in agreement with what is observed in other complex diseases. These non-coding variants may not only regulate protein-coding genes but also affect the function of non-coding elements, such as enhancers and promoters (either directly, by changing the structure, or indirectly by changes in levels of gene

![Figure 1: Genomic location of celiac disease-associated SNPs. The majority of SNPs (81%) are located in non-protein-coding regions. About 9% are located in the 3'-UTR and 5% in the 5'-UTR. Only 5% of SNPs map to protein-coding regions (figure adapted from Kumar et al., 2012 (26)).](image)

---

### Table 1. Methods to identify regulatory elements in the genome

<table>
<thead>
<tr>
<th>Regulatory elements/events</th>
<th>Identification methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA protein interactions</td>
<td>ChIP-seq, DNase footprinting</td>
</tr>
<tr>
<td>Transcription factor binding sites</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>Open chromatin accessibility</td>
<td>DNase hypersensitivity assay and FAIRE-seq</td>
</tr>
<tr>
<td>Chromatin interactions</td>
<td>ChIA-PET, HiC, 3C, 4C, 5C and 6C</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>RRBS-seq, Methylation arrays, MeDIP-seq and MRE-seq</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>RNA-seq, RNA-PET, CAGE</td>
</tr>
<tr>
<td>Distant-acting transcriptional enhancers</td>
<td>ChIP-seq, Gro-seq, Hi-C, Pro-seq</td>
</tr>
<tr>
<td>RNA-binding proteins</td>
<td>RIP-seq</td>
</tr>
</tbody>
</table>

Table adapted from author Edwards et al., 2014 (27).
expression), or they may regulate the levels of gene expression of non-coding RNAs (ncRNAs) such as microRNAs (miRNAs) and lncRNAs.

Several classes of ncRNAs are known to be main players in the regulation of gene expression (Table 2) and regulatory RNAs belonging to different classes can even cross-talk. LncRNAs, for instance, can bind miRNAs, thus acting as competing endogenous RNAs or as natural microRNA sponges (23, 24, 25). As an example of how this process works, a muscle-specific lncRNA (lincMD1) sequesters miR-133 and miR-135, thereby effectively but indirectly regulating the expression of the two respective targets of these miRNAs, MAML1 and MEF2C, two transcription factors that activate muscle-specific gene expression (30).

**MiRNAs in celiac disease**

miRNAs are small molecules that fine tune gene expression and they can serve as targets for therapy, and function as potential biomarkers (33). None of these functions have been well explored in CeD so far.

In Chapter 5, we describe examples of how genetic variants can affect miRNAs (see Fig. 3 in Chapter 5) as we observed that 9% of CeD-associated-SNPs are located in the 3'-UTRs of genes (Fig. 1), and thereby are likely candidates to potentially affect miRNA binding sites. The 3'-UTR CeD-associated SNPs, or their proxies (r² > 0.8), can potentially destroy or create miRNA binding sites.

In Chapter 2 we predicted that 11 SNPs in the 3'-UTRs of six different CeD-associated protein-coding genes could modify miRNA binding sites. These predictions were based on the results from only one method, which might have increased the number of false positive results. Since our Chapter 2 study was carried out, more algorithms have become available.

We therefore recently decided to combine the outcome of four different algorithms designed to predict whether SNPs affect miRNA binding sites (Fig. 2 and Table 3). We only

### Table 2. Classes of non-coding RNAs

<table>
<thead>
<tr>
<th>Names</th>
<th>Size</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA</td>
<td>19-24 bp</td>
<td>Post-transcriptional silencing of protein-coding genes and ncRNAs</td>
</tr>
<tr>
<td>piRNA</td>
<td>26-31 bp</td>
<td>Transposon repression and DNA methylation</td>
</tr>
<tr>
<td>snoRNA</td>
<td>60-300 bp</td>
<td>Promote rRNA modifications</td>
</tr>
<tr>
<td>tiRNA</td>
<td>17-18 bp</td>
<td>Modulate CTCF localization and nucleosome density</td>
</tr>
<tr>
<td>lncRNA</td>
<td>&gt;200 bp</td>
<td>Broad function in transcription regulation, including X-chromosome inactivation, telomere regulation and imprinting</td>
</tr>
<tr>
<td>lincRNA</td>
<td>&gt;200 bp</td>
<td>Can act as guides for recruiting for chromatin-modifying complexes</td>
</tr>
</tbody>
</table>

Table adapted from Esteller 2011 (31) and Qureshi et al., 2012 (32). piRNA: PIWI-interacting RNA; snoRNA: small nucleolar RNA, tiRNA: transcription initiation RNA; lincRNA: long intergenic RNA.
Table 3. Description of the algorithms used to predict SNPs that affect miRNA binding

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
<th>Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRSNP</td>
<td>Applies miranda v3.3a to predict miRNA binding sites, dbSNP 135 and miRBase 18 collections</td>
<td><a href="http://cmbi.bjmu.edu.cn/mirsnp">http://cmbi.bjmu.edu.cn/mirsnp</a></td>
</tr>
<tr>
<td>microSNiper</td>
<td>Uses a custom FASTA alignment algorithm to predict miRNA targets, and dbSNP 137 and miRBase 19 data</td>
<td><a href="http://epicenter.ie-freiburg.mpg.de/services/microsniper/">http://epicenter.ie-freiburg.mpg.de/services/microsniper/</a></td>
</tr>
<tr>
<td>PolymiRTs</td>
<td>Uses TargetScan v6.2 to predict miRNAs targets sites, experimental confirmation from CLASH experiments, db SNP 137 and miRBAse 20 data</td>
<td><a href="http://compbio.uthsc.edu/miRSNP/home.php">http://compbio.uthsc.edu/miRSNP/home.php</a></td>
</tr>
<tr>
<td>miRNASNP</td>
<td>Uses TargetScan v6.2 and miranda v3.3a to predict miRNAs targets, the TarBase database for experimental confirmation, dbSNP 137 and miRBase 19 data</td>
<td><a href="http://www.bioguo.org/miRNASNP2/">http://www.bioguo.org/miRNASNP2/</a></td>
</tr>
</tbody>
</table>

58 CeD SNPs + 1075 proxies ($r^2 \geq 0.8$)

PolymiRTS  miRSNP  microSNiper  miRNASNP

#SNPs per algorithm

6 SNPs in 3'UTR overlapping in at least 3 algorithms and affecting miRNAs

Figure 2. Workflow for analyzing SNPs predicted to change miRNA binding sites. CeD SNPs and their proxies were analyzed using four different algorithms to predict potential effects on miRNA binding. Venn diagram shows number of SNPs that were predicted to affect binding site of miRNAs for each algorithm.
Discussion and future perspectives

considered SNPs that were identified by at least three of the four methods (Fig. 2). Using this more stringent strategy, we identified six SNPs in five CeD-associated protein-coding genes (IL18RAP, IL21, IRF4, PTPRK, UBE2L3) that could affect the binding of two miRNAs: miR-4252 and miR-1273g. SNPs rs7444 in the 3'-UTR of UBE2L3 is an interesting candidate for further validation. It was predicted that the risk allele (C) disrupts the binding of two miRNAs: miR-4252 and miR-1273g. The SNP is in LD with SNPs associated with other autoimmune diseases, including inflammatory bowel disease (IBD) and type 1 diabetes (T1D), suggesting that insights from CeD genetics can also provide further insights into the biology of other immune-mediated diseases. SNPs rs7444 in the 3'-UTR of UBE2L3 were identified in the eQTL data from Blood eQTL browser (http://genenetwork.nl/bloodeqtlbrowser/) (36).

Table 4. Prediction of SNPs affecting different miRNAs' binding site

<table>
<thead>
<tr>
<th>Gene</th>
<th>CeD SNP</th>
<th>Chr</th>
<th>Pos (bp)</th>
<th>r²</th>
<th>SNP in mRNA binding site</th>
<th>eQTL effect*</th>
<th>Predicted function</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL18RAP</td>
<td>rs990171</td>
<td>2</td>
<td>103068787</td>
<td>0.9</td>
<td>rs7559479</td>
<td>Down</td>
<td>Risk A allele disrupts miR-136-5p site</td>
</tr>
<tr>
<td>IL21</td>
<td>rs62323881</td>
<td>4</td>
<td>123038295</td>
<td>1</td>
<td>rs17879298</td>
<td>-</td>
<td>Risk C allele disrupts miR-380-3p site</td>
</tr>
<tr>
<td>IRF4</td>
<td>rs1050976</td>
<td>6</td>
<td>409119</td>
<td>1</td>
<td>rs9391997</td>
<td>-</td>
<td>Risk A allele disrupts miR-2115-3p site</td>
</tr>
<tr>
<td>PTPRK</td>
<td>rs55743914</td>
<td>6</td>
<td>128291199</td>
<td>0.9</td>
<td>rs3190930</td>
<td>-</td>
<td>Risk A allele creates miR-1910 site</td>
</tr>
<tr>
<td>UBE2L3</td>
<td>rs4821124</td>
<td>22</td>
<td>21976934</td>
<td>0.9</td>
<td>rs7444</td>
<td>Up</td>
<td>Risk C allele disrupts miR-4252 site</td>
</tr>
</tbody>
</table>

*eQTL results from Blood eQTL browser (http://genenetwork.nl/bloodeqtlbrowser/) (36).
and miR-1273g-3p (Table 3). The protein encoded by UBE2L3 has been described as involved in the cytotoxic function of NK cells, which represent a key cell type in the innate immune response (37). Hence, it is conceivable that if miR-4252 and/or miR-1273g-3p are expressed in NK cells, they then regulate UBE2L3 expression in this cell type. Functional assays in the correct cell type, in this case NK cells, are needed to confirm such predictions.

### Circulating miRNAs in CeD – potential clinical applications

The realization that miRNAs are stable in circulation, for example in cerebrospinal fluid, plasma and serum, makes them interesting biomarker candidates (38). In CeD, biomarkers are being sought for early disease detection and for monitoring disease progression and the effectiveness of dietary interventions. It is still not entirely clear how and why these circulating miRNAs are released from cells, but two possibilities can be envisioned. They can either end up in circulation because of tissue damage, for instance upon apoptotic or necrotic cell death, or there could be an active secretion pathway (39). Some recent studies have indicated that circulating miRNAs play a role in cell-to-cell communication (40). Furthermore, circulating miRNAs are protected from RNAses because they are encapsulated in microparticles (exosomes, apoptotic bodies, and microvesicles), in complexes with AGO2, or in lipid complexes (39).

---

**Figure 3.** Schematic representation of miRNAs (in red) and their respective targets (in blue). All but one gene (UBE2L3) contain potential binding sites for a single miRNA that can be modified by CeD-associated SNPs.
Additionally, a recent study found that 3' end post-transcriptional modification defines a distinction between miRNAs originating in cells versus those originating in exosomes (41). Furthermore, deregulation of circulating miRNAs has already been described in numerous immune-related diseases, including Crohn's disease (42) and T1D (43), as well as disease-stage-specific circulating miRNA profiles have been reported. Despite these promising aspects of circulating miRNAs, they remain technically challenging to measure because of the low concentration of cell-free miRNA in extracellular vesicles (41). In this thesis, we present two ongoing studies that focus on miRNAs in CeD. In the first study, we profiled miRNAs in CeD, further emphasizing the need for biomarkers that could help in this process (47, 48).

Table 5. Common miRNAs in the PreventCD cohort (Chapter 6) and the Italian cohort (Chapter 7).

<table>
<thead>
<tr>
<th>Common miRNAs in PreventCD cohort and Italian cohort</th>
<th>Log2Fold (PvCD)</th>
<th>P-value (PvCD)</th>
<th>Log2Fold (Italian)</th>
<th>P-value (Italian)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-150-3p</td>
<td>-1.449080462</td>
<td>0.015321755</td>
<td>-1.295542407</td>
<td>0.002244688</td>
</tr>
<tr>
<td>hsa-miR-1246</td>
<td>-1.594246402</td>
<td>0.040799068</td>
<td>-1.050415374</td>
<td>0.0058485</td>
</tr>
<tr>
<td>hsa-miR-424-5p</td>
<td>-2.08906741</td>
<td>0.009780786</td>
<td>0.996951551</td>
<td>0.0058485</td>
</tr>
<tr>
<td>hsa-miR-10a-5p</td>
<td>-1.382000345</td>
<td>0.014877126</td>
<td>-1.05484301</td>
<td>0.008991646</td>
</tr>
<tr>
<td>hsa-miR-24-3p</td>
<td>-1.044150043</td>
<td>0.031730371</td>
<td>-0.96683914</td>
<td>0.018305019</td>
</tr>
<tr>
<td>hsa-miR-100-5p</td>
<td>-1.893921902</td>
<td>0.002084626</td>
<td>-0.939389404</td>
<td>0.019285288</td>
</tr>
<tr>
<td>hsa-miR-224-5p</td>
<td>-1.399292079</td>
<td>0.045238586</td>
<td>0.970681247</td>
<td>0.039246921</td>
</tr>
<tr>
<td>hsa-miR-30b-5p</td>
<td>-1.660679689</td>
<td>0.040747629</td>
<td>1.22837621</td>
<td>0.002244688</td>
</tr>
<tr>
<td>hsa-miR-99a-5p</td>
<td>-2.105939577</td>
<td>0.000717664</td>
<td>-1.000675282</td>
<td>0.013947542</td>
</tr>
</tbody>
</table>
circulating miRNAs in children from the PreventCD project (Chapter 7), a prospective study aimed at inducing oral tolerance to gluten in children at high-risk of developing CeD (49). Serum samples were collected from the PreventCD children before positive serology and onset of CeD, which makes this cohort ideal for biomarker studies aimed at early disease detection. In addition, the children were also followed after treatment with the introduction of a gluten-free diet, which allows for biomarker studies aimed at monitoring dietary intervention. Our second study focused on a cohort of Italian children diagnosed with CeD, of whom some were also sampled after starting a gluten-free diet (Chapter 7). In the PreventCD cohort we tried to identify CeD-specific miRNAs by comparing samples obtained at month 3 with samples taken at the time of diagnosis, while in the Italian cohort we investigated differences between diagnostic samples and samples obtained from controls. Although these are different comparisons, we found nine miRNAs overlapping in both studies (Table 5) to be differentially expressed (Fig. 4). These miRNAs are interesting candidates for further follow-up, particularly because we observed little overlap with circulating miRNAs associated with other immune-related diseases, where type 1 diabetes and Crohn’s disease studies displayed the biggest overlapped (Fig. 5). These results indicate that CeD-specific miRNAs may exist. This hypothesis is consistent with previously published results describing a panel of circulating miRNAs that are differentially expressed in Crohn’s disease patients, but not deregulated in samples obtained from patients with active CeD (42).

In Chapter 6 we also prioritized three miRNA candidates (miR-15b-5p, miR-30c-5p and miR-100-5p) that were significantly differentially expressed at the time of diagnosis, but which reverted to a ‘healthy’ level 6 months after the

\[\text{PreventCD} \quad \text{Italian Cohort}\]

\[\begin{array}{ccc}
36 & 9 & 39 \\
\end{array}\]

Figure 4. Venn diagram illustrating the overlap between CeD-associated miRNAs identified in the PreventCD cohort and in the Italian cohort. Nine miRNAs are shared in these two profiles. See Chapters 6 and 7 for details.
patients started a gluten-free diet. It is worth emphasizing that these miRNAs started to be deregulated well before diagnosis (Chapter 6), making them interesting future biomarker targets. One of these miRNAs (miR-100-5p) was also differentially expressed when comparing CeD patients at diagnosis with healthy controls from the Italian cohort, supporting the validity of this specific miRNA as a true biomarker. Furthermore, on the functional level, this miRNA has been implicated in IL6 and TLR4 signaling (50).

Despite the availability of diagnostic criteria, there is still a need for additional biomarkers in CeD. Although a lifelong gluten-free diet can keep CeD at bay in the majority of cases, dietary transgression is a very rare complication in CeD patients (0.3% of diagnosed CeD patients) (51). This condition is known as refractory celiac disease. A few patients develop a hypersensitivity to extremely small quantities of gluten, which then trigger symptoms that lead to further damage of the intestinal mucosa. While antibody levels are often used to monitor adherence to the diet, antibody levels are not stable and the decrease in antibody titers can sometimes take years. Having a biomarker to monitor gluten ingestion would therefore represent a major advance in the management of CeD patients. In Chapter 7 we investigated if we could identify circulating miRNAs that can serve as specific markers to show adherence to a gluten-free diet. We compared circulating miRNAs from individuals at diagnosis and individuals following a gluten-free diet. We found 11 miRNAs that could potentially be used to monitor ingestion of gluten (Chapter 7). However, further studies with a large number of individuals are necessary to clarify if these miRNAs could indeed be gluten-specific and can therefore be used to monitor ingestion of gluten.

In Chapter 6, an additional factor...
complicating the analysis of the PreventCD cohort was the variability in patient age at diagnosis. In some cases, individuals developed CeD at 5 years of age while in one case a patient developed CeD at the age of 2 years and 6 months. This variability in age of onset will need to be taken into account when searching for miRNAs that increase before diagnosis, and is another reason for expanding the cohort under investigation.

MiRNAs in the small intestine of CeD patients

To date, only a few studies have been performed on miRNAs deregulated in the small intestine of CeD individuals, suggesting that miRNAs are involved in CeD etiology. In one study miR449a was found to be overexpressed in the small intestine of patients with active CeD, where it seems to downregulate NOTCH1 (52). Another study reported four other miRNAs (miR-192-5p, miR-31-5p, miR-338-3p, and miR-197) to be deregulated in the small intestines of CeD patients, and to display negative expression correlations with mRNA and protein levels of immune target genes including NOD2 and FOXP3 (53). The identification of specific miRNAs associated with CeD provides insight into disease biology and future therapeutic targets but the miRNAs also exhibit tissue/cell-specific expression patterns. This should be taken into account when analyzing intestinal biopsies, which represent a mixture of intestinal cell types. As a consequence, expression profiles of miRNAs in biopsies from the small intestine can be used to find functional miRNAs involved in CeD pathogenesis and possible candidates for future therapeutics approaches.

In Chapter 7 we performed miRNA expression profiling on small intestinal biopsies from active pediatric CeD patients. We found 109 miRNAs differentially expressed between cases and controls, including several immune-related miRNAs such as miR-155-5p and miR-223-3p. MiR-155-5p was found to be overexpressed in active CeD patients (Chapter 7), and the increase in this miRNA has also been reported in other autoimmune diseases including rheumatoid arthritis (RA) (54) and multiple sclerosis (MS) (55). Moreover, anti-miR-155 treatment drastically inhibited the development of MS in mice (56). This miRNA is involved in B and T-cell activation and has also been considered as a potential therapeutic target for autoimmune diseases in general. Three of our intestinal miRNAs (miR-192-5p, miR-31-5p and miR-338-3p) were also associated with CeD in an independent study of samples obtained from adult patients (53). In this study by Magni et al., three miRNAs and their target genes were validated on the gene expression and protein level. MiR-192-5p, in particular, is an interesting candidate as it is likely to be involved in cell proliferation and apoptosis. Its target genes, C-X-C motif ligand 2 (CXCL2) and nucleotide-binding oligomerization domain-containing protein 2 (NOD2), are both
Discussion and future perspectives

associated with an altered innate immune response in IBD (28, 55, 56). Magni et al. further showed that a reduction in the expression of miR-192-5p after stimulation with gliadin increases the mRNA levels of CXCL2 and NOD2 in CeD biopsies. The authors suggest that altered innate immunity might play an important role in the altered response to gluten in the intestinal mucosa of these adult CeD patients (53). We also found miR-192-5p to be downregulated in pediatric patients with active CeD. Thirteen of the miRNAs from the biopsy profiles were also found to be differentially expressed in the serum of the same individuals (Chapter 7).

Although these results are intriguing, further validation in an independent cohort is required. In summary, we found miRNA candidates in intestinal biopsies that may play a role in the disease pathogenesis. Some of these miRNAs are involved in immune processes and have already been reported as interesting therapeutic targets for other immune diseases, thus their therapeutic potential in CeD is extremely intriguing and opens up new avenues for future research.

**Future Perspectives**

**Beyond GWAS**

The GWAS era has identified thousands of genetic variants associated with complex traits and diseases. Recently, the use of
custom genotyping-arrays, such as the Immunochip (25), has helped to prioritize and identify additional disease-associated variants. However, the promise of the wealth of GWAS data remains unrealized because the biological mechanisms driven by the underlying genomic regions are largely unknown. In addition, most disease-associated variants are located outside of protein-coding regions, which makes it harder to interpret the associations. New methodologies are emerging will allow researchers to better interpret GWAS results.

The public availability of certain types of genetic data has increased exponentially over recent years. In 2010 RNA-seq data was available for fewer than 1,000 samples, while in 2013 the number of samples available had increased to nearly 10,000 (Fig.6). The explosion of annotated genomic data (e.g. from the ENCODE project (4)), enables researchers to design integrative approaches that will form the core of the post-GWAS era. These datasets contain valuable information on functional regulatory elements derived from high-throughput whole-genome assays, such as RNA-seq, ChIP-seq, DNase-seq, FAIRE-seq and proSEQ (Table 2), and which can now be integrated with GWAS results. In addition, large tissue-specific datasets are becoming available, for example through the Genotype Tissue Expression project (GTEx) (59), which aims to provide RNA-seq data from up to 41 tissues of 900 individuals (resulting in approximately 20,000 samples). The availability of this data will help us interpret GWAS results and improve our knowledge of the underlying biology of the disease.

Furthermore, methods to integrate functional data can be adapted to identify functional SNPs and provide a comprehensive and effective list of candidate variants. New generation custom genotyping-arrays should be designed for these variants in order to perform further association studies.

To move beyond GWAS and truly begin to understanding genetic disease mechanisms functional studies should be performed. Since many genes and regulatory elements are likely to be used in a cell-specific manner, such studies will be challenging, in particular when single-cell experiments are needed. Fortunately, highly multiplexed single-cell approaches are now emerging that provide new perspectives on the underlying state of an individual cell and its role in the context of disease and intervention (61). Furthermore, new functional genomics assays are also emerging that can help to link this abundant data with biological mechanisms. For instance, the clustered regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) protein 9 system enables researchers to manipulate specific genomic elements precisely, and can facilitate the elucidation of the effects of genetic variants on
Discussion and future perspectives

their target genes in biology and disease (62). Recently, a modified CRISPR/Cas9 system named CRISPR inference (CRISPRi) was developed for RNA-guided transcriptional regulation (63). This technology provides a new, highly specific tool for changing gene expression without genetically altering the target DNA sequence. The system has been used to regulate transcription of multiple genes, individually or simultaneously (63). Moreover, it was shown that CRISPRi can also silence miRNAs (64) and it thereby provides an interesting approach for follow-up of some of the miRNAs that we detected in chapters 6 and 7.

Small non-coding RNAs have great potential

miRNAs are potential biomarkers and also potential therapeutic targets. In cancer research, the use of miRNAs for diagnostic purposes has already begun. For instance, a panel of 24 miRNAs is currently monitored to identify kidney cancer (65). For any novel drug, including miRNA-targeting compounds, it will take years before it finds its way to the clinic (33). Nevertheless, several interesting miRNAs have already been tested in animal models and are currently being tested in preclinical studies. Examples of these include miR-208 in cardiac diseases (66) and miR-155 in inflammatory diseases (67).

It has been reported that exogenous miRNAs derived from plants, including common cereal grains, can be found in human serum, suggesting that genetic material from food might survive the digestive tract and enter the bloodstream. It is likely that part of the circulating miRNA spectrum (including, perhaps, those we identified) results from food intake (68). It would be interesting to investigate if some of the circulating miRNAs in CeD patients might have originated from ingested wheat and/or gluten. Hypothetically, if one could detect CeD gluten-specific miRNAs, anti-miRNA drugs could be used to inhibit these miRNAs and thereby prevent CeD development.

We can expect research investigating the role of miRNAs in CeD to only increase in the coming years.

New genomics techniques and data layers constantly drive the field forward

We are still in the early stages of understanding the full complexity of the human genome. The observation that the majority of GWAS variants map to non-coding regions opened a new perspective and placed non-coding elements, such as ncRNAs, at a higher priority for future investigations. It turns out that finding causal SNPs is not straightforward. Even if a SNP maps onto a gene, it does not necessarily mean that the effect will come from the same gene (e.g. FTO and IRX3). Even more complex situations arise when disease-associated SNPs change the expression levels of ncRNAs (Chapter 4).
Despite these hurdles, the large amount of genomic data that is becoming available will provide an exciting opportunity to explore functional regulatory elements overlapping disease-associated variants. The speed with which this new data is emerging brings new challenges to scientists in the field, who now face the problem of how to interpret such a wealth of data, although integrating different layers of genomic data with GWAS results seems to be an intelligent way to begin.

Clearly, the large number of loci identified by GWAS provides insights into disease etiology. Refining these GWAS association signals is essential to moving forward and revealing the true causative variants. Furthermore, the role of ncRNA in disease is still not fully understood and should be explored in much more detail to help understand disease mechanisms.

The constant development of novel genomics assays is driving the rapid evolution of the field of genomics. It is proving a challenge to keep up with these developments, but a new generation of scientists is now at work who can integrate and interpret big datasets to address medical biological questions.
Discussion and future perspectives

References


