Finding the missing 'LiNCs' in celiac disease
Hrdlicková, Barbara

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):
Hrdlicková, B. (2015). Finding the missing 'LiNCs' in celiac disease [Groningen]: University of Groningen

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.

Download date: 12-01-2019
8

GENERAL DISCUSSION AND FUTURE PERSPECTIVES
PART I – A broader perspective on the work described in this thesis

Genome-wide association studies (GWAS) have identified thousands of disease-associated genetic loci in complex diseases, but to date the causality has been proven only for a small number of the associated variants. Since GWAS variants contribute only modestly to the disease phenotypes, they can only explain 10-30% of the genetic heritability in most diseases [1–4]. However, we have to keep in mind that narrowing down the size of the human genome from approximately 3 billion base pairs that contain approximately 10 million single nucleotide polymorphisms (SNPs) to a limited number of disease-associated loci and SNPs was a major step forward in the research on the genetic basis of human diseases.

Four major lessons have been learned from GWAS in multiple phenotypes [5]. Firstly, pinpointing the true causal variants using GWAS platforms based on the tag-SNP approach is challenging because GWAS provide correlations and not causations; they discover the association of tag SNPs. Often, there are numerous SNPs in strong linkage disequilibrium (LD) with these tags and the associated tag SNP does not therefore need to be the SNP causing the phenotype. Secondly, the disease-predisposing gene is not immediately evident as the vast majority of associated variants are not exonic. This means that the majority of disease-associated SNPs affect gene expression rather than gene function. Thirdly, it is becoming clearer that the effects of risk SNPs are likely to be specific to a particular cell type. Fourthly, despite all that is now known, the nature of the genetic variants that explain the remaining genetic contribution to the phenotype, the so-called ‘missing heritability’, is still unclear.

When I started my PhD research in 2010, long non-coding RNAs (lncRNAs) were just emerging as a substantial set of novel regulatory RNAs and it was proposed that these molecules could play a role in many important processes [6, 7]. Only a handful of lncRNAs were known before the publication of two key papers, one by the Rinn-Lander group and the other by the Mattick group in 2009 [6, 7]. However, in 2011 the second paper by the Rinn-Lander group described more than 3,000 human lncRNAs [8]. Soon after these publications, John Rinn et al described 8,000 human lncRNAs and their expression across 23 different human cell types and tissues [9]. I used the data from this

![Figure 1: Significant over-representation of white blood cell-specific lncRNAs in celiac disease loci [10].](image)
group and was able to show that white blood cell-specific lncRNAs are over-represented in celiac disease (CeD)-associated loci (Figure 1), which makes sense in the biological context of CeD as it is generally accepted to be a lymphocyte-driven disease.

In Part I of this thesis we describe our investigation into the possible approaches that can be used to identify more CeD-associated genetic variants. We addressed two possible approaches: (i) sampling of an extreme CeD-linked phenotype, and (ii) increasing the sample size studied. To better understand the potential effect of GWAS disease-associated variants we decided to focus on lncRNAs, which represent a part of the genome that had been largely ignored, and we deep-sequenced them to identify potential causal variants. To determine the cell types in which lncRNAs might be functional, we performed an analysis of expression profiles using RNA-sequencing data across multiple immune cell types (described in Part 2 of this thesis).

**Investigating extreme phenotypes can help to identify novel disease variants**

Up to 2010, two GWAS have been performed in CeD cohorts, pinpointing 27 loci (including the HLA-locus) as associated with CeD and together explaining ~40% of its genetic risk (heritability), [11–13]. An Immunochip study in 2011 increased the number of non-HLA loci from 26 to 39 and increased the proportion of total explained heritability up to 54% [10]. This is an exceptionally high percentage for any disease and makes CeD the autoimmune disease with the best understood genetic component. After this, the question arose as to where the missing heritability contributing to CeD genetic risk (the remaining 46%) could lie. The heritability calculation is estimated using the prevalence of disease seen in the general population and in family (twin) data [14]. We know that CeD is highly under diagnosed in general [15, 16] and this might be one of the factors contributing to the missing heritability. Novel, yet undiscovered variants (common or rare) could explain the additional part of heritability. Phenotype-associated SNPs discovered by GWAS are, by definition, common variants (minor allele frequency (MAF) >5%) with moderate effects. It has been suggested that analysis of rare (MAF<0.5%) and low-frequency (0.5%≤MAF<5%) variants with potentially large effect sizes might explain part of the additional disease risk [17]. Efforts to identify rare variants have been made by sequencing known loci and genes linked to the disease pathophysiology. Thus far, these studies have not been very successful. For instance, Hunt at al. showed that rare variants in protein-coding regions of known autoimmune disease loci play only a limited role in the susceptibility to common autoimmune diseases [18]. The results from their and other studies [18–21] suggest that the missing heritability of common diseases/quantitative traits might be mainly due to many common variants with small effects.

We hypothesized that one approach to identify more CeD variants is to focus on extreme phenotypes, such as patients with more severe forms of CeD or those suffering from complications. This hypothesis was based on the observation that in some severe, early-onset forms of certain diseases, alleles with larger effect sizes can be identified. These extreme phenotypes can therefore be explained by genotypes that are more skewed towards lower frequency variants, in comparison to milder and/or late-onset phenotypes of the same disease [22]. A strategy based on comparing patients with extreme phenotypes is cost-effective and can increase the statistical power of the study to discover risk variants. Besides these advantages, the extreme phenotype approach also has some limitations that need to be taken into account. These types of studies are characterized by a higher sensitivity for outliers;
there is also the possibility of sampling bias and the findings will not be applicable to the whole spectrum of the disease [23]. However, this approach has been successful in age-related macular degeneration, coronary heart disease, lipid levels, and in breast cancer and other types of malignancy [24–29]. These studies have shown that patients with extreme phenotypes often carry a higher load of causal variants.

In chapter 3 we describe investigations in samples from patients with refractory celiac disease type 2 (RCDII), which is considered to be an extreme subphenotype of CeD. RCDII shares phenotypic features with active, untreated CeD, such as malabsorption, villus atrophy, and infiltration of intraepithelial T-lymphocytes (IELs). However, although these symptoms disappear in CeD patients when they start on a gluten-free diet (GFD), RCDII patients gain no relief. Instead, even after GFD treatment has been started, a massive progressive intraepithelial expansion of IEL clones is observed that display an abnormal T-cell/NK cell phenotype. These IELs can ultimately transform into an aggressive non-Hodgkin lymphoma, known as enteropathy-type T-cell lymphoma (EATL), [30, 31]. To our surprise, we did not find any suggestion for a role of CeD-associated SNPs in the RCDII samples, which suggests that RCDII has an entirely different genetic background than CeD and a different pathological status. Indeed, RCDII is sometimes also classified as a low-grade intraepithelial lymphoma [31, 32]. However, we did find some interesting candidate SNPs in our RCDII cohort. One of these is located on chromosome 7 and exhibits a cis-eQTL effect on the protein-coding gene FAM188B; it has not previously been linked to RCDII nor to CeD. We predicted the potential function of this protein by “guilt-by-association” co-expression analysis, which connected FAM188B to the toll-like receptor signaling pathway, suggesting this gene plays a role in the innate immunity. To validate our association findings for RCDII, it will be essential to include more RCDII samples and increase the sample size of the study. Moreover, we need studies focusing on intestinal biology (for example, measuring cytokines (IL-15), counting aberrant T-cells, and studying how aberrant IELs evolve into low-grade EATL).

**Following up on GWAS loci with suggestive associations can uncover novel genes and pathways**

The efficiency of identifying genetic variants with lower frequency and/or with smaller effect size can be increased by performing association analyses in larger cohorts. There is a clear correlation between the sample size and number of disease-associated loci identified (correlation>0.9, Figure 2), [33]. By increasing the number of samples, the significance of previously sub-significant (or suggested) SNPs may be improved.

We pursued this approach in chapter 4 by performing a follow-up of six candidate variants that showed suggestive association in the CeD GWAS [12]. Upon genotyping of additional samples and performing a combined meta-analysis, we identified two new genetic variants, and correlated them with altered gene expression levels. One of these, *NFE2L3*, is highly expressed in stimulated monocytes.

All genetic studies in CeD conducted so far have used Caucasian (European origin) samples. Including samples from more, non-European, ethnicities is a potential avenue to enlarge the sample size, but also to enrich for low-frequency alleles, unravel population-specific disease-associated variants, and to identify causal variants (reduce the size of the disease-associated loci) due to the different LD structure across populations [44–46]. This cross-ethnic approach has been applied successfully, for instance, in a recent meta-analysis in inflammatory bowel disease (IBD) across multiple populations.
with non-European ancestry together with data from Jostins’ study [2]. This meta-analysis identified 14 new IBD loci [47]. In CeD, a proof-of-concept cross-ethnic study was performed using Indian patients and aiming at replication of the known CeD loci in a non-European CeD cohort and discovery of novel, Indian patient-specific associations [48]. Five of the known 39 CeD-associated loci could be replicated, while two novel CeD loci were proposed [48]. This study only covered around 500 cases and a similar number of controls [48], but the results imply that trans-ethnic genetic mapping studies with larger non-European CeD cohorts are likely to be successful.

Another way to increase the sample size is by pooling two different diseases with a presumed overlapping genetic architecture. For CeD this has been done successfully by combining cohorts of CeD patients with another autoimmune disease, either rheumatoid arthritis [49] or Crohn’s disease [50]. This resulted in eight and two novel associations, respectively. These cross-disease meta-analysis results suggest that similar studies using other diseases are likely to identify more shared genetic factors.

Despite all of these possibilities, at a certain point, it will become difficult to increase the sample size of CeD cohorts. In particular, for the detection of rare variants we may need tens of thousands of patients. With a presumed prevalence of 1% in the general population, 1,000,000 people from the general population would need to be screened for CeD auto-antibodies to obtain only 10,000 new patients. This approach is complicated by the fact that autoantibody levels fluctuate in CeD, suggesting that sampling at different points in time is essential to identifying new patients. Thus, this would be a costly approach. A development that might

---

**Figure 2:** Relationship between sample size and number of loci associated with autoimmune and immune-related diseases, as revealed by Immunochip studies. Abbreviations and references to respective studies: AA, alopecia areata [34]; AD, atopic dermatitis [35]; AS, ankylosing spondylitis [36]; CD, Crohn’s disease [2]; CeD, celiac disease [10]; JIA, juvenile idiopathic arthritis [37]; MS, multiple sclerosis [38]; PBC, primary biliary cirrhosis [3]; PS, psoriasis [39]; PsCh, primary sclerosing cholangitis [40]; RA, rheumatoid arthritis [41]; SJ, Sjögren’s syndrome [42]; SS, systemic sclerosis [43]; UC, ulcerative colitis [2]. (Figure adapted from Parkes et al. [33]).
help in this respect is the growing number of biobanks with samples obtained from the general population. It is possible that more CeD patients will be identified in the future, particularly if these biobanks screen individuals repeatedly over time and are linked to general practitioners and other health care resources.

Towards an understanding of the role of non-coding variants in autoimmune diseases

Although GWAS were successful in unraveling the genetic architecture of many diseases or quantitative traits, GWAS findings still need be translated into biological understanding and into applications for clinical practice. It is relatively easy to infer the causality of coding variants when they alter the amino-acid sequence of proteins, but for non-coding variants this is more challenging. The vast majority (>90%) of all disease/trait associations resulting from GWAS lie within the non-coding sequence of the genome [51]. Although these SNPs can theoretically affect regions involved in the regulation of gene expression, it will be much more difficult to annotate such variants. These regulatory GWAS SNPs may intersect with other regulatory elements such as promoters, enhancers, or silencers, and thereby disturb the binding of transcription factors, change chromatin structure, affect chromatin looping, or impact on the function of non-coding RNAs (ncRNAs), [51–54]. Regulation of gene expression is highly dependent on the cell type, developmental stage and environmental factors (such as stimulatory cues), [55]. GWAS SNPs have shown to be enriched in regulatory regions [36] and more likely to have an effect on gene expression (they are concentrated within expression quantitative trait loci (eQTL)), [56–58]. The eQTL mapping approach is a powerful way to reveal the consequences of the presence of regulatory SNPs and to identify the causal genes affected by disease SNPs [59].

It is sometimes challenging to understand the disease biology related to the GWAS results. This is exemplified by the strongest obesity-associated SNP, which is a non-coding variant but located in the FTO gene. Since it was shown to be associated, more than 450 papers have reported on the role of FTO in obesity. Only recently, an eQTL study performed using brain tissue provided clear evidence that this SNP does not affect the FTO gene at all, but instead changes the expression level of the transcription factor IRX3 located more than 500 kb away [60]. No eQTL effect was observed in blood cells and knockout studies in mice corroborated a role for IRX3 in obesity [60]. This example not only illustrates the power of eQTL mapping, but also shows the necessity of investigating the correct tissue.

One of the most surprising findings from the CeD Immunochip study was that only three out of the 57 independent association signals were present in exons of protein-coding genes [10]. The remaining 54 signals (~95%) are located in regulatory regions such as 5’ untranslated regions (5’ UTRs, 5%), 3’ UTRs (9%), in introns, and between protein-coding genes (81%) [61]. Moreover, after annotation of the CeD loci with the genes categorized by the ENCODE consortium [55, 62, 63], we observed, to our surprise, an overlap between some of the CeD-associated SNPs and non-coding RNA (ncRNA) genes such as microRNAs or long non-coding RNAs (lncRNAs). This was exciting given that ncRNAs also control gene expression. In the second part of this thesis I therefore focused on the non-coding variants and their possible involvement in disease.

Work of our group showed that approximately 10% of all GWAS SNPs associated with autoimmune and immune-related diseases (AID) map to lncRNAs [64]. Analysis of CeD-association peaks in particular suggested the involvement of lncRNAs in at
least three CeD-loci. In chapter 6 we investigated if four lncRNAs in these three loci could harbor pathogenic variants. Four of the identified sequence variants were predicted to change the secondary structure of lncRNAs, while other SNPs overlap with regulatory motifs (for transcription factor/protein-binding, promoter/enhancer histone marks, DNaseI hypersensitive sites). Eight of the identified sequence variants appear to affect the expression of protein-coding and/or lncRNA genes based on cis-eQTL mapping results. Although our proof-of-concept study needs to be extended, it is exciting to see that disease-associated variation may impact lncRNA function by changing their structure or expression level.

Cell type prioritization is essential to investigate lncRNA function

Our findings add to the growing evidence that lncRNA deregulation is involved in disease, but so far there are only a few studies that describe the effect of GWAS-associated non-coding variants on lncRNAs in complex diseases, and even less is known about the function of disease-associated lncRNAs [56, 65, 66].

One of the main characteristics of lncRNAs is their tissue-specific expression [9, 63]. Although lncRNAs were reported to be expressed at low levels, their expression can increase dramatically in relevant cell types under specific conditions (such as stimulations). Therefore, disease-relevant cell types/tissues must be gathered to enable functional studies. LncRNA expression profiling has been applied to human tumors and identified cancer-specific signatures associated with disease progression. The candidate lncRNAs can be used for diagnosis, staging, prognosis and for monitoring the response to treatment [67–70].

To elaborate on lncRNAs as candidate regulators involved in AID pathogenesis, we annotated AID-loci with coding as well as non-coding genes and performed transcriptomic profiling across 11 different immune cell types (chapter 7). Our analysis revealed that lncRNAs mapping to AID loci are more often expressed and therefore are more likely to function in immune cell types than those mapping in the whole genome. The expression profiles of disease-specific lncRNAs in particular immune cell types suggested the involvement of different cell types in most of the tested phenotypes, except in CeD. This implies that we did not capture the cell type most relevant to CeD in our analyses. More specifically, our cell panel did not include cells from the small intestine (gluten specific T-cells, intraepithelial lymphocytes, or enterocytes). Future studies will have to include these cells and we suggest that these studies should be performed by comparing resting cells with cells stimulated with the most relevant stimuli that mimic the ongoing inflammatory process in CeD.

Pathway analysis algorithms can predict the pathways in which lncRNAs are involved. For example, tissue-specific co-expression analysis can reveal pathways previously not linked to the phenotype, based on the “guilt-by-association” (co-expression between selected genes), [71]. Our group has developed a tool (RNAnetwork; www.genenetwork.nl) that predicts the function of lncRNAs based on co-expression analysis using 5,000 RNA-sequencing datasets [Karjalainen et al., manuscript in preparation]. These predictions are a good starting point for practical laboratory studies aiming to understand the function of lncRNAs. Moreover, the functional annotation of non-coding SNPs and prioritization of causal SNPs and genes will help in prioritizing the cells in which the function of lncRNAs needs to be studied. Annotating non-coding variants with publicly available signatures (such as transcription factor binding sites, epigenetic/histone modification marks, DNA methylation, or accessible chromatin regions in the context of specific tissues) identified by the ENCODE Project [55] or the Roadmap Project [72], for
example, will help to predict the effect of non-coding SNPs and prioritize the cell types in which the lncRNAs need to be investigated.

Assays that can be used to study lncRNA function are, for instance, assays that can determine interactions of lncRNAs with (a) DNA (e.g. ChIP/ChRIP [73], ChOP [74], ChiRP [75], CHART [76], ChiA-PET [77], R3C [78]), or (b) proteins (e.g. RIP [79], RIP-Chip [80], CLIP [81], iCLIP [82], HITS-CLIP/CLIP-Seq [83], protein microarrays [84]). Other state-of-the-art techniques that are applied to study lncRNAs involve overexpression or knock-down of lncRNAs (using conventional short hairpin RNA (shRNA) approaches or the new CRISPER-cas9 technology [85]), or assays that can address the structural features of lncRNAs (e.g. RNase footprinting [86], SHAPE [87], FragSeq [88], and PARS [89]).

To date, the majority of functional studies on lncRNAs have been performed in in vitro models. After a few studies, it has become clear that the mechanism of lncRNAs in cell lines (in vitro) can be very different from their roles in primary cells in vivo [90]. For instance, in vitro experiments with Evf2 lncRNA revealed enhancer activation in trans mechanism, while an in vivo model (mouse lacking Evf2) repression action in cis [90–92]. Due to the fact that lncRNAs are key players in many (if not all) crucial biologic processes, organism-wide knock-out animal models are sometimes lethal [93, 94]. This issue can be addressed by generating cell-type-specific loss-of-function models [94, 95].

PART II – Future perspective: From CeD SNPs to clinical application

Autoimmune and immune-related diseases, including CeD, are chronic, often disabling, disorders that urgently need new treatment strategies – but the translation of GWAS findings into clinical applications is still difficult. However, as we are now in the post-GWAS era, where more attention is being focused on functional follow-up studies, there is hope they will lead to better treatment options. The increase in the number of genetic loci that could be associated to CeD may not help patients immediately, but there is hope that all of the data generated in genome-wide studies and functional approaches will pinpoint and explain the crucial pathways affected in the disease. These pathways may be open to therapeutic modulation. Moreover, the increased number of risk variants will aid in designing more accurate risk prediction models, which can be applied in the clinic in the future. It can be envisioned that individuals in the “higher CeD risk” group will be followed-up more carefully than low-risk patients, to help prevent potential complications, for instance. Building better risk prediction models may also help to separate RCDII patients from those CeD patients who do not adhere well enough to the gluten-free diet (e.g. because of the presence of gluten contamination in their diet). A personalized medicine approach to CeD treatment, based on our knowledge of genetics, pathway analysis and selected markers, could be developed on targeting specific biologic pathways instead of using a traditional systemic therapy.

In the field of cancer research, lncRNAs have already been mentioned as promising cancer-biomarkers [96–99]. There is one diagnostic assay, a lncRNA-based urine test, already approved by the US Food and Drug Administration (FDA). This assay is based on detection of lncRNA PCA3 (prostate cancer antigen 3) and is used for the diagnosis of prostate cancer [96, 100]. Perhaps it will be possible, in the future, to avoid invasive endoscopical procedures to diagnose CeD and to monitor disease progression by using tests based on the detection of lncRNAs in biofluids. It is even possible that specific lncRNAs could be targeted for disease treatment, e.g. by knocking down or inhibiting harmful lncRNAs or by delivering beneficial
lncRNAs to specific cells. Firstly, down-regulation of lncRNAs could be achieved by antisense oligonucleotides (such as gapmers), [101] or by inducing the degradation of lncRNAs using the enzyme RNaseH [102, 103]. Secondly, blocking of lncRNA function could be achieved by masking the binding sites for lncRNAs in the molecules interacting with them (e.g. by using small nucleotide inhibitors [104]) or by masking the interaction sites of the lncRNAs using, for instance, antagonistic oligonucleotides [105]. Thirdly, lncRNA structure, and therefore the lncRNA function, could be disrupted with small molecule inhibitors [106, 107] or by oligonucleotides [108]. The delivery of beneficial lncRNAs might also become possible in the future if targeted delivery systems can be developed [107]. The main issue in the lncRNA field is that, although more than 15,000 human lncRNAs are currently known (Gencode version 21, released 2 October 2014), [62], the function is only known for a few dozen. This gap in our understanding of lncRNA biology and function has to be filled before we can consider using lncRNAs in disease therapy. Pinpointing the relevant disease pathways may then lead to identification of new therapeutic targets.

The lncRNA world we are entering is really very exciting and we may well be surprised by some of the findings in the (near) future. Hopefully it will be possible to translate these future findings to the clinic and take them closer to the patients who may benefit from them.

REFERENCES


