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):

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.
CELIAC DISEASE: MOVING FROM GENETIC ASSOCIATIONS TO CAUSAL VARIANTS

Barbara Hrdlickova¹, Harm-Jan Westra¹, Lude Franke¹, Cisca Wijmenga¹

¹ Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands

ABSTRACT

Genome-wide association studies are providing insight into the genetic basis of common complex diseases: more than 1,150 genetic loci (2,165 unique SNPs) have recently been associated to 159 complex diseases. The hunt for genes contributing to immune-related diseases has been particularly successful in celiac disease, for example, with 27 genome-wide significantly associated loci identified so far. One of the current challenges is how to move from a genetic association with a disease to finding disease-associated genes and causal variants, as a step towards understanding the underlying disease process. About 50% of disease-associated SNPs affect the expression of nearby genes (so-called expression quantitative traits or eQTLs) and these can provide clues for finding causal variants. Although eQTLs can be useful, fine-mapping and sequencing are required to refine the association signal. Ultimately, sophisticated study designs will be needed to find the causal variants involved in complex diseases. In this review we use celiac disease as an example to describe the different aspects that need to be considered on the path from genetic association to disease-causing variants.
GENERAL INTRODUCTION

Complex or multifactorial diseases, such as celiac disease, type 2 diabetes and asthma, are determined by the interaction of multiple genes with environmental factors. As a consequence, these diseases are fairly common and can cluster in families [1, 2], yet they do not display a clear-cut pattern of inheritance.

Each complex disease has a unique genetic architecture which can be defined by (1) the number of disease susceptibility loci, (2) the distribution of the effect sizes of each of the loci, and (3) the interaction between loci with environmental factors [3]. Yet the similarities in the genetic architecture are far stronger than the differences as most diseases are determined by a large number of loci with modest effect and only a few genes with a stronger effect. The latter are rarer in the population and only a few have been identified so far. Very few, if any, structural variants have been identified. Recent studies show there is a high overlap in the disease loci for immune-mediated diseases [4]. The genetic architecture can be influenced by gene and genotype frequencies, population history, the distribution of environmental factors, and biological features such as age and sex. Historically, the architecture of complex traits in modern humans probably reflects the architecture and evolution of the human genome. The oldest human alleles originated in Africa and are shared by all human populations. Together, these age-old alleles account for approximately 90% of the genetic variation seen in humans. An exponential population growth, together with the independent development of agriculture and urbanization in the different regions over the past 10,000 years, has possibly resulted in the selection of alleles [5] that confer a large degree of the present genetic variation. The genetic architecture underlying complex diseases is expected to follow the same evolutionary path and is likely to include both common variants with modest effect sizes and rare variants that are presumably more population-specific and that exert stronger effect sizes.

Recent technological developments have resulted in the identification of hundreds of genetic loci that contribute to complex diseases [6]. In particular, the availability of genome-wide association studies (GWAS) has been a major step forward (see Box 1). GWAS provide an unbiased, in vivo human association of a gene or genetic pathway in the disease pathogenesis: there is no prior biological knowledge required on the biochemical or molecular function of disease genes. GWAS therefore often identify genes and pathways that were not previously known to play a role in the disease of interest. This means we can generate new hypotheses, which open up new avenues for investigation. So far, GWAS have mainly identified common genetic variants with rather modest effect sizes (usually with odds ratios (OR) <1.5) and, as such, these loci can in general only explain a small percentage of the genetic disease risk (usually less than 20%).

IMMUNE-RELATED DISEASES WITH SPECIAL EMPHASIS ON CELIAC DISEASE

Immune-related diseases affect approximately 3-5% of the population worldwide [7]. Celiac disease, an inflammatory disorder of the small intestine, is one of the most common immune-related diseases. The symptoms are often aspecific and it has a prevalence of 0.7% to 2% of undetected cases in the general population [8]. A necessary triggering environmental factor in celiac disease is dietary gluten, a storage protein present in several staple grains such as wheat, barley and rye. The enzyme tissue transglutaminase modifies the gluten protein, which can trigger a cascade of innate and adaptive
immune responses in the intestine in genetically susceptible individuals on exposure to gluten. The immune response leads to flattening of the intestinal mucosa, resulting in malabsorption of nutrients, which causes a wide-range of symptoms including diarrhea, steatorrhea, weight loss, fatigue, anemia, ataxia and infertility [8].

A common feature of celiac disease is the central role of T-cells in causing tissue inflammation. The main genetic factor in celiac disease, HLA-DQ2/8, orchestrates a pro-inflammatory T-helper 1 (Th1) response against gluten. These gluten-specific CD4+ T-cells are the hallmark of the disease and are thought to be responsible for disease development. Celiac disease has become a model disease for studying the underlying biological mechanisms of immune-related disorders because (1) there is a clear involvement of the major histocompatibility complex (MHC), (2) both innate and adaptive immune responses are observed, (3) there is a high co-morbidity with other immune-related diseases, and (4) the triggering environmental factor is known.

The hunt for genetic loci contributing to immune-related diseases has been fairly successful, with the discovery of genome-wide significant associations of 27 celiac disease loci (and an additional 13 suggestively associated loci), 99 inflammatory bowel disease (IBD) loci (including loci for both Crohn’s disease (n=71) and ulcerative colitis (n=47)), 31 rheumatoid arthritis loci and 53 type 1 diabetes loci (Table 1), [9-17]. Many of these loci are not unique to one disease but are shared by other immune-related diseases, including both autoimmune and inflammatory disorders [11, 18-21].

**COMMON AND RARE VARIANTS IN COMPLEX DISEASES**

For the purpose of this review, we define ‘common’ variants as those that are present in a population at a frequency of over 5%, ‘low-frequency’ variants as those with frequencies between 1–5%, and ‘rare’ variants as those with a frequency less than 1%. The ‘common disease, common variant’ model, which assumes that complex diseases derive
The International HapMap Project [67] catalogued the common SNPs and linkage disequilibrium (LD) structure across the human genome by analyzing more than three million SNPs in different populations. It created a public database, providing a genetic resource to allow researchers easy, user-friendly access to the data on a gene of interest or on the whole genome [45, 68].

The 1000 Genomes Project [54] is an ongoing international project aiming to perform whole-genome sequencing of approximately 100 samples from 27 different populations; it will provide a complete overview of human genetic variation (MAF≥0.01). The project is expected to create a more detailed catalogue of genetic variation (SNPs, insertions, deletions, inversions, copy number variants) and a human reference sequence, which will be freely accessible through public online databases. The data from the 1000 Genomes Project can be used by researchers 1) in combination with data from GWAS for imputing genotypes of many samples for additional rare variants without extra cost, and 2) for comparing allele frequencies and linkage disequilibrium patterns across populations. The 1000 Genomes Project was started in January 2008 and the finalization phase is planned for the end of 2011. There is a first publication [69] and some data is already available online [54] from the pilot studies containing samples from the HapMap Project.

All the new strategies and advanced technologies in genetics have increased the amount of genetic information known about individuals from different backgrounds. Collecting this information, which is often extremely sensitive, and the freely accessible public databases have highlighted the importance of research ethics in genetics. It is crucial to establish common principles to guide the use and storage of such data and how the personal information on individuals should be separated from their genetic data to avoid abuse of such information. Other ethical questions regarding the informed consent needed to permit the use of biological material in future research has arisen from the expanding use of biobanks and tissue repositories. The Genetic Information Nondiscrimination Act [70, 71] and the International Declaration on Human Genetic Data from UNESCO [72, 73] aim to reduce apprehension in this area.

Table 1. Number of associated loci found for immune-related diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Manifestation</th>
<th>Number of identified loci</th>
<th>Percentage of genetic risk explained</th>
<th>Refs.</th>
<th>Number of loci with an eQTL effect*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celiac disease</td>
<td>Chronic inflammation of the small intestine</td>
<td>27</td>
<td>40%</td>
<td>[9, 10, 27]</td>
<td>27</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>Chronic, inflammatory bowel disease, which</td>
<td>71</td>
<td>23%</td>
<td>[12]</td>
<td>39 (55%)</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Chronic, systemic inflammation of synovial tissue and joint destruction</td>
<td>31</td>
<td>16%</td>
<td>[13]</td>
<td>23 (74%)</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Autoimmune destruction of insulin-producing β-cells of the pancreas</td>
<td>53</td>
<td>45%</td>
<td>[14, 61]</td>
<td>32 (60%)</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>Chronic, inflammatory bowel disease, with ulcers in the upper layer of the lining of the large intestine</td>
<td>47</td>
<td>16%</td>
<td>[15-17]</td>
<td>18 (38%)</td>
</tr>
</tbody>
</table>

* eQTL mapping was performed in 1,469 peripheral blood samples (FDR<0.05, 100 permutations, 250 kb distance between SNP and mid-probe position), [9]. To remove non-genetic variation in gene expression, we removed 50 principal components by linear regression. Disease SNPs were downloaded from the GWAS catalog from Manolio et al., accessed 2011-01-17 [31].

from the additive and/or multiplicative effects of a combination of common variants, now seems unlikely to explain the entire genetic predisposition to disease. Therefore, the focus in recent human genetics research is shifting towards the notion that low-frequency and rare variants may also be important contributors to complex diseases, resulting in re-sequencing and specialized chip efforts (discussed below). Manolio and colleagues [22] suggested a model of three risk categories: high risk-rare alleles (with frequencies below 0.5%, causing Mendelian diseases), moderate risk-low frequency alleles (with frequencies between 0.5–5%, increasing disease risk two- to threefold, but without clear Mendelian segregation) and low risk-common alleles (with frequencies above 5% and very modest effects on the disease).

To date, celiac disease is one of the best understood complex diseases with 27 associated loci, which determine about 40% of its genetic susceptibility [9,10,23,24], (Figure 1).
Figure 1. Frequency and risk distribution of celiac disease susceptibility loci. This graph shows the distribution of both the frequency and the odds ratios of the known celiac disease-associated loci as described by Dubois et al. [9]. Risk alleles are colored red, protective alleles green. Shaded alleles did not reach genome-wide significance.

Box 2. Expression QTL (eQTL) mapping and celiac disease

By treating gene expression as a quantitative trait, it is possible to correlate genotypes of disease-associated SNPs with transcriptional levels of the genes from the same region (expression quantitative trait loci, eQTLs), [74]. The eQTL mapping does not prove the causality of the gene, but is a useful approach for identifying a SNP with a functional effect on an adjacent gene and to give preference to genes for follow-up functional studies. eQTLs have already been associated with several diseases, such as celiac disease, Crohn’s disease and asthma [9, 10, 33, 75]. After identifying a genetic locus associated with a disease, the next step is to determine if such a variant alters the gene expression of a nearby transcript.

Variability in gene expression can arise at different levels, such as transcription, mRNA stability, splicing, and translation efficiency, which either cause changes in the mRNA sequence or changes at the level or sequence of regulatory RNAs [43, 44, 76]. It is expected that the underlying variability affecting each of these levels will be located at different positions with respect to the transcript (Figure 2). Although genetic variability at a locus may affect very distal genes (and even those on another chromosome), in most examples the most proximal genes are investigated (so-called cis-eQTL effects), [66].

So far, only one study investigating the influence of common variants associated with celiac disease on the gene expression levels of nearby genes has been published [9]. In this study, Dubois et al. included the 26 non-HLA celiac loci to associate the SNP genotypes to expression levels of genes located within a region of 250 kb around these SNPs, obtained from 1,469 human whole blood samples of random individuals [9]. The direction of these associations can be positive, negative, or not present. They found that the direction of the association was not always identical for all the genes in the disease-associated loci, as a SNP can have both negative and positive effects on some eQTL regions. One example is the 12q24.12 locus, which contains 13 different genes and an association peak with SNP rs653178. Two of these genes (TMEM116, ALDH2) showed a negative direction with the rs653178 disease allele, whereas two other genes (ATXN, SH2B3) in the locus showed a positive direction. For the remaining nine genes, no significant difference in gene expression level was obtained [9, 20]. This may imply that celiac disease-associated variants can have different effects on multiple pathways, by modulating the gene expression of different sets of genes within one locus. In total, approximately 71% of celiac disease eQTL SNPs showed a negative effect on the gene expression, indicating that most of the celiac disease-risk SNPs evoke a down-regulation of gene expression [20].
**Figure 2.** Expression QTL (eQTL) mapping. A) eQTL links genotyping data from genetic markers identified by GWAS with gene expression data from microarray analysis obtained from transcripts present in a target cell or tissue. The direction of these associations can be positive, negative, or not present. B) Example of an eQTL for one of the celiac disease-associated loci. The SNP rs917997 genotype is strongly correlated with *IL18RAP* gene expression levels (p-value=1.1x10^{-133}), with the lowest level of expression for carriers homozygous for the rs917997 allele A. *IL18RAP* encodes the β-chain of the IL-18 receptor. The SNP-gene expression correlation is shown for a dataset consisting of 1,240 samples hybridized to Illumina HT-12 arrays. C) When compared to the GG genotype of rs917997, the AA genotype is associated with a 9-fold decrease in expression.
Altogether 26 loci [all outside the HLA locus] fit in the ‘low risk, common alleles’ group. The HLA DQA1/HLA-DQB1 locus constitutes a new category, as this is a moderate risk variant (OR>4) which occurs at high frequency in the population (>20%). In general, a frequency reduction of high-risk variants is expected to occur when the variant contributes to decreased reproductive fitness or to negative selection [25]. However, there are exceptions, such as the non-synonymous SNP rs3184504 in the SH2B3 gene, which is associated to both celiac disease and type 1 diabetes. The disease-associated allele (rs3184504*A) shows consistent signs of positive selection in all European populations, indicating that the allele associated with autoimmune disease may have increased as a by-product of natural selection in European populations [26]. In this case, Zhernakova and colleagues [27] used the Integrated Haplotype Score method [28] to estimate that this selective switch probably happened as a result of an infectious disease outbreak that occurred 1,200-1,700 years ago. This could have been the Justinian plague, a pandemic that afflicted the Eastern Roman Empire in 541–542 AD.

The current spectrum of disease-associated alleles is biased towards common alleles, since low-frequency and rare alleles are usually not present on GWAS genotyping arrays. Identifying such alleles requires sequence analysis of large series of cases and controls, or association studies using custom-designed genotyping platforms containing low-frequency and rare alleles such as the Immunochip [29], the Metabochip or the Cardiochip [30].

The Immunochip is a custom-made Illumina Infinium HD array, designed to densely genotype immune-related disease loci. It contains common and rare variants from 186 distinct GWAS loci associated to 12 different immune-related diseases (ankylosing spondylitis, autoimmune thyroid disease, celiac disease, Crohn’s disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). The variants present on the Immunochip were derived from the pilot project for the 1000 Genomes Project and from disease-specific re-sequencing data. Hence, the Immunochip will assist in identifying additional variants (both rare and common), in detailed haplotype analysis, and in refining association signals.

FROM GENETIC LOCI TO GENETIC VARIANTS

It is important to realize that most genetic associations in complex diseases are to loci and not to single genes, unlike Mendelian disorders. This is due to several factors: firstly, the genotyping arrays that have been used in GWAS preclude an accurate localization of the association signals, as they contain approximately 500,000 single nucleotide polymorphisms (SNPs) on average. Secondly, the tag SNPs can be in strong linkage disequilibrium (LD) with another SNP or also with a common structural variant which could, in fact, be more associated with the disease and not be typed by the chip. Thirdly, a substantial number of the loci implicated in complex diseases contain more than one gene (58%), or no gene at all (16%), when looking 0.1 cM upstream and downstream of disease-associated SNPs in the recombination map of HapMap 2, release 22 [31]. Finally, approximately 43% of trait-associated SNPs are located in intergenic regions and 45% are located in non-coding introns [6, 32]. As an effect, it is often difficult to identify the causal variant from disease-associated loci.

Approximately 50% of the loci associated to immune-mediated diseases influence the level of transcription of nearby genes (i.e. eQTL effects), [9, 17, 33, 34], (Box 2, Figure 2). This suggests that these disease-associated variants probably influence the function of
disease genes by regulating the gene transcription levels rather than by altering the coding sequence. It is tempting to speculate that genetic associations to loci that are devoid of genes might also affect levels of gene transcription, through mechanisms such as long-range transcription regulation, enhancers, microRNAs (miRNAs), etc. In animals, miRNAs regulate gene expression by imperfect hybridization to the 3’ UTR of target mRNA, thereby inhibiting protein synthesis or causing mRNA degradation [35, 36]. MiRNA is increasingly being recognized as an important element in the development of the innate and adaptive immune systems, and changes in miRNA expression are described in many immune-related diseases such as inflammatory bowel disease, rheumatoid arthritis or systemic lupus erythematosus [37-41]. For example, Zhao et al. [42] demonstrated significant overexpression of miRNA-126 and its inverse correlation with protein levels of DNA methyl transferase 1 [DNMT1] in systemic lupus erythematosus CD4+ T-cells in comparison to cells from healthy controls. When they overexpressed miRNA-126, a lupus-like autoreactivity was observed in CD4+ T-cells from healthy controls, which was shown to be caused by demethylation and upregulation of autoimmunity genes (in particular CD11 and CD70), [42].

In contrast, mutations in Mendelian diseases with a clear recessive or dominant inheritance pattern usually affect the function or the level of the disease-causing protein. These mutations, for example, alter the coding sequence, by changing splice sites and thereby deleting essential parts of the protein, or by introducing a premature stop-codon resulting in mRNA decay [43, 44], (Figure 3). As a consequence, many of the Mendelian mutations result in a loss-of-function of the allele carrying the mutation. Since our preliminary insights into the molecular mechanisms of complex diseases suggest a different class of disease-associated

Figure 3: Mutation mechanisms and their functional consequences. Mutations can occur in different parts of a gene: A) wild-type situation without a mutation, B) mutation in a coding region (exons) resulting in an alteration of the protein or its function via the substitution of an amino acid, leading to loss-of-function or haploinsufficiency (left), or by alternative splicing leading to exon skipping (right). C) mutation in a non-coding region (promoter, 3’UTR, 5’UTR, intron) resulting in changes in the amount of protein. The gene structures have a color depending on their type, e.g., exons are green, introns are yellow and UTRs are grey.
Moving from Associated SNP to Causal Gene or Mutation

An important next step in genetic research will be to identify the causative genes and gene variants from the GWAS loci. Often the associated SNPs map to regions of strong linkage disequilibrium which contain more than one gene [45].

The process required for moving from an associated SNP found in GWAS results to a disease-causing gene and gene variant (Figure 4) includes different steps. Fine mapping is usually an essential first step to refine the area of association by analyzing a high density of genetic markers across the region of LD, assuming that the causal gene and gene variant are located near the most associated genetic variant. The next step includes targeted re-sequencing of the refined candidate region to identify rare mutations that are independent of the common associated SNPs and that preferably alter protein function, as these are the easiest to recognize and interpret. The underlying hypothesis is that a disease-causing gene harbors multiple risk variants and at least one of them will be common or in strong LD with a common SNP (identified by GWAS), while the others will be rare. Such a strategy has recently been proven to work in Crohn’s disease [47], for which low-frequency coding variants were identified in the IL23R gene, one of the loci identified by GWAS. However, the identification of true disease-causing mutations remains a major issue. Filtering results against public databases like the 1000 Genomes Project may assist in this, as well as determining segregation of causal variants in large affected families.

Fine-mapping followed by targeted re-sequencing of GWAS regions can be laborious when there are many associated loci to investigate. A more holistic approach that circumvents the step of fine-mapping is to apply whole exome- or whole genome sequencing.
Whole exome sequencing is efficient for detecting causal variants residing in coding regions. However, if some 50% of the causal variants are non-coding variants influencing gene transcription levels, whole exome sequencing might not be the best strategy to find disease-causing variants in complex diseases. A recent study by Lehne et al. indicated that a substantial number of disease-causing risk variants may be found in coding parts of genes [48]. They analyzed the distribution of association signals based on GWAS data from seven complex diseases (bipolar disease, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, and type 1 and type 2 diabetes) from the Wellcome Trust Case Control Consortium and found a significant disease-dependent concentration of association signals in exons and genes [48]. Based on work in Crohn's disease [47], Hirschsprung disease [49] and lipid traits [50, 51], we can expect more results from these types of studies in the near future.

Whole exome sequencing has been proven to be extremely successful for Mendelian disorders and has led to the identification of the mutations causing Miller syndrome and Kabuki syndrome, for example [52, 53].

The ultimate level of resolution is whole genome sequencing and this would also reveal all the variants outside the coding part of the genome. Although such sequencing is still too expensive to apply to large series of patient and control samples (as in GWAS), we expect to see a rapid increase in fully sequenced genomes as costs decrease. While the costs of whole genome sequencing are falling, the coverage of variants GWAS may be increased by imputation using the 1000 Genomes Project data as a reference dataset (since this dataset is an extensive catalogue of genetic variants with MAF≥0.01), [54].

Interpreting non-coding sequence variants requires extensive follow-up in case-control series, using bioinformatics approaches and in-depth functional studies, as recently shown by Musunuru [51]. Whole genome- as well as whole exome sequencing approaches will benefit greatly from sophisticated study designs since interpreting such large datasets remains a major challenge. Both a family-based design [with multiple affected individuals] [51] or sequencing of the extremes of the trait distribution [50] may increase the chance of finding causal mutations [55].

What is common to the different types of sequencing techniques is the need for bioinformatics approaches to process the massive amounts of data generated and to interpret the findings. As the individual genes conferring susceptibility to complex disease are expected to be connected to each other in molecular networks, computational approaches should also be directed towards defining such networks. On the one hand, these biological pathways will indicate causal genes and, on the other hand, any gene with a causal mutation can be part of the pathway. Proof that the true disease-causing variant has been identified will require in-depth functional studies.

**DISCUSSION**

The genetic mechanism underlying complex diseases is one of the fundamental questions in biomedical research today and is partly driven by the results obtained from GWAS studies. Complex diseases are expected to be caused by tens to hundreds of loci, both rare-low-frequency loci and common ones, many of which may have an impact on the expression of nearby genes. There is much debate on the extent of rare, highly penetrant, mutations associated to complex diseases, but they are not expected to explain the majority of the missing heritability of complex diseases [56]. This suggests that many of the causal risk variants may be rather common and therefore difficult to recognize. This may even be true for the coding parts of genes, as was
recently shown for a synonymous change in the Multidrug resistance 1 [MDR1] gene that replaces a frequent codon by a rare codon and thereby affects the protein folding [57].

The challenge is how to connect GWAS findings to genes and function. Current technologies allow for fine-mapping of disease loci followed by targeted re-sequencing or, alternatively, by sequencing the entire genome or just the coding part of the genome. Despite these technological advances, the interpretation of the results remains a challenge. Fine-mapping will at least zoom in on the truly-associated region and applying more sophisticated study designs should help further refine the regions of association, for example by taking different racial groups into account [58], or by studying different phenotypes associated to the same locus [18].

The real problem arises with the management and interpretation of the huge amounts of data generated by whole genome sequencing. The combination of sequencing results with eQTL mapping data, which establishes links between genetic markers found in GWAS and gene expression levels, can help to mark and confirm the causal mutation. Several in silico analyses are unavoidable in this step. Previous studies have shown that eQTL effects are enriched around the transcriptional start sites and within 250 bp upstream of transcription end sites [59]; they are often within 20 kb of the relevant gene [60]. When trying to identify the causal variants, we could focus on the beginning of these regions, containing binding sites for different transcriptional factors, the 3’ UTR regions, or their relation to RNA stability, to name just a few factors. One example of where a non-coding variant directly influences the phenotype is the SORT1 gene. The common non-coding variant rs12740374 in the 1p13 locus creates a binding site for C/EBP transcription factor and altering the hepatic expression of SORT1, which modifies the level of plasma low-density lipoprotein cholesterol [51].

With the revolution in sequence analysis and large-scale projects like the 1000 Genomes Project, the 10,000K Project in the UK, and the Genome of the Netherlands, we will learn a lot more about our genome and how to interpret it. In the meantime, GWAS findings provide good starting points towards determining the disease-associated genes, while new bioinformatics approaches will help pinpoint the true causal genes.

ACKNOWLEDGEMENTS

The authors thank Jackie Senior and members of the Wijmenga lab for critically reading the manuscript. The authors are supported by grants from the Celiac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009 to CW), by the Netherlands Organization for Scientific Research (NWO, VICI grant 918.66.620 to CW and VENI grant 916.10.135 to LF), and by a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031 to LF).

REFERENCES


