From genome-wide association studies to disease mechanisms

Kumar, Vinod; Wijmenga, Cisca; Withoff, Sebo

Published in:
Seminars in immunopathology

DOI:
10.1007/s00281-012-0312-1

Publication date:
2012

Semin Immunopathol (2012) 34:567–580
DOI 10.1007/s00281-012-0312-1

REVIEW

From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases

Vinod Kumar · Cisca Wijmenga · Sebo Withoff

Received: 1 March 2012 / Accepted: 10 April 2012 / Published online: 14 May 2012

Abstract Celiac disease is characterized by a chronic inflammatory reaction in the intestine and is triggered by gluten, a constituent derived from grains which is present in the common daily diet in the Western world. Despite decades of research, the mechanisms behind celiac disease etiology are still not fully understood, although it is clear that both genetic and environmental factors are involved. To improve the understanding of the disease, the genetic component has been extensively studied by genome-wide association studies. These have uncovered a wealth of information that still needs further investigation to clarify its importance. In this review, we summarize and discuss the results of the genetic studies in celiac disease, focusing on the "non-HLA" genes. We also present novel approaches to identifying the causal variants in complex susceptibility loci and disease mechanisms.

Keywords Celiac disease · Autoimmune disease · Immune-related disease · Genome-wide association studies · GWAS · Pathway analysis

Introduction

Immune-related diseases range from autoimmune diseases such as celiac disease (CD), rheumatoid arthritis (RA), multiple sclerosis (MS), and type I diabetes (T1D), to more chronic inflammatory disorders such as asthma and inflammatory bowel disease (IBD). Together these disorders now account for 5–10 % of all disease cases in Western countries (see elsewhere in this issue).

Celiac disease is one of the best-understood immune-related diseases. It is the most common food intolerance in humans, affecting at least 1 % of the Western population. It is a multifactorial disease caused by many different genetic factors that act in concert with non-genetic causes. A genetic association between CD and the HLA class II genes in the major histocompatibility complex (MHC) was documented almost 40 years ago [1]. One of the most important triggering factors is dietary gluten, a storage protein present in wheat and related grains (hordein in barley, secalin in rye, and avedin in oats) (see elsewhere in this issue).

CD is an excellent model for studying the contribution of genetic factors to immune-related disorders because: (1) the environmental triggering factor is known (gluten), (2) as in other autoimmune diseases, specific HLA types (*HLA-DQA1* and *HLA-DQB1* in the case of CD) are critically involved (see elsewhere in this issue), (3) there is involvement of non-HLA disease-susceptibility loci, many of which are shared with other autoimmune diseases, (4) there is an elevated incidence of other immune-related diseases both in family members and individuals, and (5) both the innate and the adaptive immune responses play a role in CD [2].

Prior to genome-wide association studies (GWAS) the genetics of CD included candidate gene studies in case-control cohorts and linkage studies in multi-generation families and affected sibpairs [3]. None of these studies have convincingly resulted in the identification of genetic factors beyond the well-established HLA-DQA1 and HLA-DQB1 genes. With the introduction of GWAS, the number of genetic factors implicated in CD has increased and 54 % of its heritability can now be explained. However, the

This article is published as part of the Special Issue on *Celiac Disease*.

V. Kumar · C. Wijmenga (✉) · S. Withoff
Department of Genetics, University Medical Hospital Groningen, University of Groningen,
PO Box 30001, 9700 RB Groningen, the Netherlands
e-mail: c.wijmenga@umcg.nl

methods for calculating the heritability are currently under debate [4], but CD remains the immune-related disorder with the best-characterized genetic component (e.g., MS 20 %, RA 16 %, CrD 23 %, UC 16 %, T1D 45 %) [5, 6].

GWAS in CD: yielding only the tip of the iceberg

GWA studies provide an unbiased approach for identifying genes and pathways involved in a certain phenotype, as they are not based on prior biological knowledge of the genes that they identify. Indeed, GWAS frequently identify genes and/or pathways that were not previously implicated in the phenotype of interest, for example, the unexpected role of the autophagy pathway in IBD [7]). Such an unbiased approach is highly beneficial as it generates new hypotheses that open up new avenues for investigation. Nevertheless, we must be careful in interpreting GWAS findings, as it is sometimes difficult to pinpoint the primary target of the genetic association. It is important to realize that the gene names of disease-associated loci are merely signposts. Often it is difficult to identify the single gene or gene variant providing risk or protection to a disease, because disease-associated loci often contain multiple genes and potential risk variants. Since individual genetic risk variants are usually common and have only a modest effect on disease risk, and because the cell or a sample of the tissue where the disease manifests is difficult to obtain for research purposes, it is difficult to investigate the consequence of the true causal risk variant. Despite these hurdles, GWAS have uncovered hundreds of loci associated to immune-related disorders, although these may represent only the tip of the iceberg [8–10]. This wealth of information will serve to formulate hypotheses that can be tested using experimental studies. Moreover, GWAS data can also be subjected to bioinformatic analysis to obtain more details about the tip of the iceberg and to reveal what still remains under the surface (see later sections in this review). To appreciate the complexity of GWAS, it is important to fully grasp the statistics involved. The interested reader can find an extensive description of the analytical methods in a review by Balding [11]. Here, we will describe how GWAS have contributed to our understanding of the genetics of CD.

The first GWAS for CD was performed in 2007 on a relatively small cohort consisting of 778 CD patients and 1,422 controls, all from the UK [12]. The subjects were tested for association to some 300,000 genetic variants in the human genome (so-called single nucleotide polymorphisms or SNPs) and the top 1,500 most associated SNPs were followed-up in replication cohorts consisting of 1,643 cases and 3,406 controls. Besides HLA, 13 regions in the genome were identified as harboring genes and genetic variants associated to CD [12–14]. Interestingly, the majority of the identified regions contained genes controlling immune responses, such as the *IL2-IL21* locus on 4q27, thereby suggesting, for the first time, the potential role of IL2, a cytokine important for the homeostasis and function of T cells, and of IL21, a new member of the type 1 cytokine superfamily which regulates many other immune and non-immune cells. This first GWA study also revealed the phenomena of pleiotropy, i.e., genetic variants associated to CD are also associated with other immune-related diseases. For example, the *IL2-21* locus is now a well-established disease susceptibility locus for T1D, RA, UC, MS, and systemic lupus erythematosus (SLE) [2, 15–22].

A much larger GWAS on CD included more than 4,500 CD patients and nearly 11,000 controls from four different populations (UK, Italy, Finland, the Netherlands) and 550,000 SNPs [23]. After replicating the most-significant 131 SNPs in seven follow-up cohorts of European descent, comprising almost 5,000 CD patients and more than 5,500 controls, 13 new regions in the genome were found to be associated with CD, bringing the total number of non-HLA associated loci to 26. The study by Dubois et al. [23] also showed that about 50 % of CD-associated SNPs affect the expression of nearby genes (so-called expression quantitative traits loci or eQTLs), indicating that the mechanism underlying CD is governed by a deregulation of gene expression.

More recently, the number of loci associated to CD was raised to 39 [24] when the Immunochip platform became available [25] (see fine-mapping approaches).

The "resolution" of GWAS heavily depends on the number of samples included. One way to circumvent this limitation is to combine datasets and to perform a meta-analysis, as was done by Dubois et al. [23]. Given the pleiotropic nature of the genetics underlying immune-related diseases, it also became possible to conduct cross-disease meta-analyses aimed at identifying additional shared susceptibility loci, as has been successfully demonstrated for CD. Two published GWAS datasets, one on CD [23] and one on RA [16], were pooled and the data obtained from the primary analysis was replicated using 2,169 CD cases (and 2,255 controls) and 2,845 RA cases (and 4,944 controls). In this meta-analysis, eight SNPs were replicated, including four SNPs mapping to loci that had not previously been associated with either disease (*CD247*, *UBEL3*, *DDX6*, and *UBASH3A*) and another four SNPs mapping to loci that had previously only been established in one of the diseases (*SH2B3*, *8q24.2*, *STAT4*, and *TRAF1-C5*). The identification of these eight loci, together with six known loci (*MMEL1/TNFRSF14*, *REL*, *ICOS/CTLA4*, *IL2/IL21*, *TNFAIP3*, and *TAGAP*), brought the total number of non-HLA susceptibility loci shared between CD and RA to 14 [17]. A similar study was performed for CD and CrD and identified four shared susceptibility loci [21]. Although meta-analysis can

help identify shared risk loci, it is important to realize that it is also possible to obtain contradictory data. Sometimes the association to the same loci is more complex and observed with different SNPs, or with identical SNPs but with the opposite allele. For example, the A allele of SNP rs917997 in *IL18RAP* is increased in frequency in CD cases, while the same allele is decreased in frequency in T1D patients [26]. This could mean that the SNP is protective in one disease and a risk factor in the other.

Fine-mapping approaches

One of the problems associated with GWAS is that the genome is not necessarily covered at a high resolution. The early GWAS chips used in CD studies contained 300,000–550,000 SNPs, while the human genome consists of 3 billion basepairs, of which at least 1-2 % is polymorphic in any given individual. Many loci are therefore not covered densely enough with SNPs, resulting in the association with disease of regions that can contain multiple genes. This complicates the interpretation of the GWAS results, but one of the most straightforward approaches to address this problem is to fine-map disease-associated loci by zooming in on specific collections of SNPs that cover defined gene-sets at high density. A recent genetic study aimed at fine-mapping CD GWAS loci was performed on the Immunochip platform [24]. The Immunochip [25] is a custom Illumina Infinium HD array, which was specifically designed by the Immunochip Consortium to densely fine-map existing GWAS loci and to replicate loci that had not yet reached genome-wide significance. The approximately 200,000 SNPs on the Immunochip array consist of SNP variants that were present in public databases at the time of production (September 2009), including variants described in the European samples sequenced as part of the 1000 Genomes Project pilot phase I. The Immunochip covers: (1) the 186 loci associated with autoimmune or inflammatory diseases meeting genome-wide significance criteria ($P<5\times10^{-8}$), from 12 immune-mediated diseases (autoimmune thyroid disease, ankylosing spondylitis, CD, CrD, IgA deficiency, MS, primary biliary cirrhosis, psoriasis, RA, SLE, T1D, and UC), (2) the MHC and KIR/LILR loci, (3) the most significant SNPs from GWAS loci with sub-significant *P* values awaiting deep replication, and (4) a small proportion of SNPs of investigator-specific undisclosed content. In the case of CD, the Immunochip was used to genotype more than 12,000 CD patients and a similar number of controls from seven different populations [24]. The platform revealed a total of 39 genome-wide significant loci (Fig. 1a), but upon conditional analysis 13 loci were found to include more than one independent association signal, resulting in a total of 57 independent non-HLA signals. These 57 SNPs are in general rather common, with frequencies above 5 % and modest effect sizes with an odds ratio between 1.124 and 1.360 (compared to an odds ratio of >5 for HLA) (Fig. 1b). Because of the higher density of SNPs for each of the loci, it was possible to refine the association signal to a single gene for 29 loci (Fig. 1a).

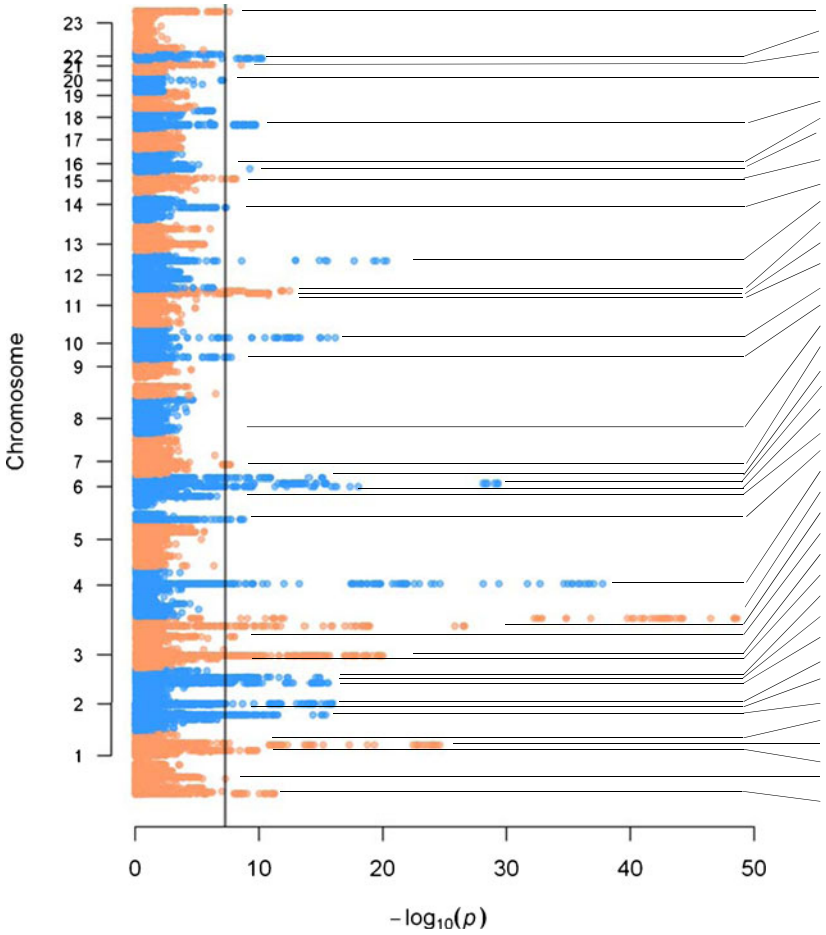
One of the most surprising findings from this fine-mapping study was the observation that the *PTPRK* gene is the causal gene in the *THEMIS*/*PTPRK* locus [23]. Immunological publications on the function of the *THEMIS* gene had suggested that it could be a very interesting candidate risk gene for CD, as it is an important regulator of thymic T cell selection [27]. This observation suggested an important role for the thymus; this is an attractive theory given the lack of oral tolerance present in CD. However, there is only limited literature on the *PTPRK* gene, but knock-out of the *Ptprk* gene in rats leads to a Th cell deficiency [28]. This example shows that GWAS results can easily be misinterpreted if attractive candidates are chosen without performing further validation. Immunochip analysis also identified 147 non-CD autoimmune disease loci with intermediate *p* values (in GWAS only SNPs with a $P<5\times10^{-8}$ are considered true associations as they have reached "genome-wide significance"). It cannot be ruled out that these SNPs play a role in the disease process but that the study was underpowered to unequivocally prove involvement of these SNPs, suggesting that there might be dozens more genes contributing to CD.

Another approach for fine-mapping is imputation [29, 30]. Imputation is an in silico process in which the allelic combinations of non-genotyped SNPs in an individual are inferred (though not directly assayed) based on the haplotype structure present in large reference datasets, such as the ones provided by the 1000 Genomes Project (2010) and the International HapMap project [31–33]. A haplotype is the combination of alleles at adjacent locations (loci) on the chromosome that are transmitted together. After imputation, each dataset typically contains information on 2.5–4 million SNP variants per individual, including low-frequency variants that are not covered on a typical GWAS array [34]. Subsequent association analysis on imputed genotypes may narrow down the region of association and help pinpoint the causative variant. As imputation is merely an in silico prediction of unknown genotypes based on the haplotype structure of a reference population, sufficient quality control measures are needed to exclude badly imputed SNPs and then the predicted genotypes need to be validated by other genotyping techniques or direct sequencing.

Genetic architecture of celiac disease

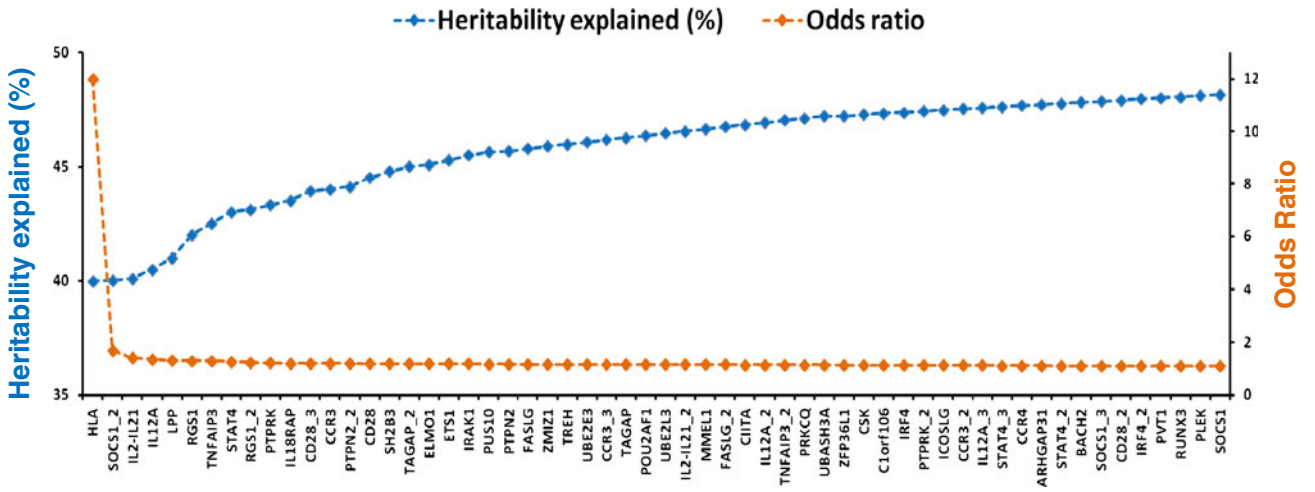
The studies conducted thus far (Fig. 2) suggest that the genetic architecture of CD follows the common disease-

a



CD candidate genes	OR	shared AID
HCFC1, TMEM187, IRAK1	1.18	--
UBE2L3, YDJC	1.16	RA, SLE
ICOSLG	0.89	UC, CrD
UBASH3A	0.88	T1D, VL, RA
PTPN2	1.17	T1D, CrD, RA
SOCS1, PRM1, PRM2	0.92	--
CIITA and others	1.14	UC
CLK3 and others	1.13	--
ZFP36L1	1.13	CrD, MS
SH2B3, ATXN2	1.19	T1D, RA
E1ST	1.18	SLE
TREH, DDX6	0.86	RA
POU2AF1 and others	1.16	--
ZMIZ1	0.86	MS, IBD, VL, CrD
PFKFB3, PRKCQ	0.88	RA, T1D
PVT1	0.91	MS
ELMO1	1.18	--
TAGAP	1.16	CrD, MS
OLIG3, TNFAIP3	1.29	RA, SLE, PSO
PTPRK	1.21	MS
BACH2	1.10	T1D, CrD, MS
IRF4	0.89	--
KIAA1109, ADAD1, IL2, IL21	0.71	T1D, RA, AA, UC
LPP	0.76	VL
SCHIP1, IL12A	1.36	PBC, MS
ARHGAP31	1.11	--
CCR1-3, LTF	1.20	--
CCR4, GLB1	1.11	--
CD28, CTLA4, ICOS	1.19	T1D, RA, AA, GD
STAT4	0.79	SLE, RA, PBC
UBE2E3, ITGA4	1.16	--
IL18R1, IL18RAP	1.20	CrD
PLEK, FBX048	0.92	MS
PUS10	1.17	UC, CrD
C1orf106	0.89	UC, MS
RGS1	0.77	MS
FASLG, TNFSF18	0.87	--
RUNX3	0.92	ASP
C1orf93, MMEL1, TTC34	0.87	RA, UC

b



CD genes (HLA and non-HLA) arranged according to odds ratio

Fig. 1 Overview of the celiac disease loci. **a** Manhattan plot showing the CD susceptibility loci identified by Immunochip. The *x*-axis displays the − log10 *P* values and the *y*-axis displays the chromosomes. Candidate genes from 39 loci are shown in the first text column. At three loci (*IRAK1, SH2B3*, and *MMEL1*), the most significant SNPs at each locus are in absolute linkage with coding variants. Next, the odds ratios (OR) of all CD SNPs are displayed. In the last column 28 CD loci are also shown to be susceptibility regions for other autoimmune diseases (the shared disease associations are extracted from the GWAS catalogue (www.genome.gov/gwastudies)). *AA* alopecia areata, *AID* autoimmune disease, *ASP* ankylosing spondylitis, *CrD* Crohn's disease, *IBD* inflammatory bowel disease, *MS* multiple sclerosis, *PBC* primary biliary cirrhosis, *PSO* psoriasis, *RA* rheumatoid arthritis, *SLE* systemic lupus erythematosus, *T1D* type I diabetes, *UC* ulcerative colitis, *VL* vitiligo. **b** Odds ratios (OR) and cumulative heritability associated with each locus. Along the *x*-axis all the CD risk loci are arranged according to decreasing OR. Multiple independent signals at one locus are depicted as "gene name"_2 or "gene name"_3 (e.g., *SOCS1_1, SOCS1_2*, and *SOCS1_3* indicate three independent signals at the SOCS1 locus). We assumed a CD heritability of 89 % [75] and CD prevalence of 1.5 % to estimate the cumulative heritability explained. The OR of 12 for HLA [74] and the ORs of the non-HLA CD loci [24] were published previously

common variant (CD-CV) hypothesis [35–37]. To date, approximately 54 % of the genetics of CD can be explained

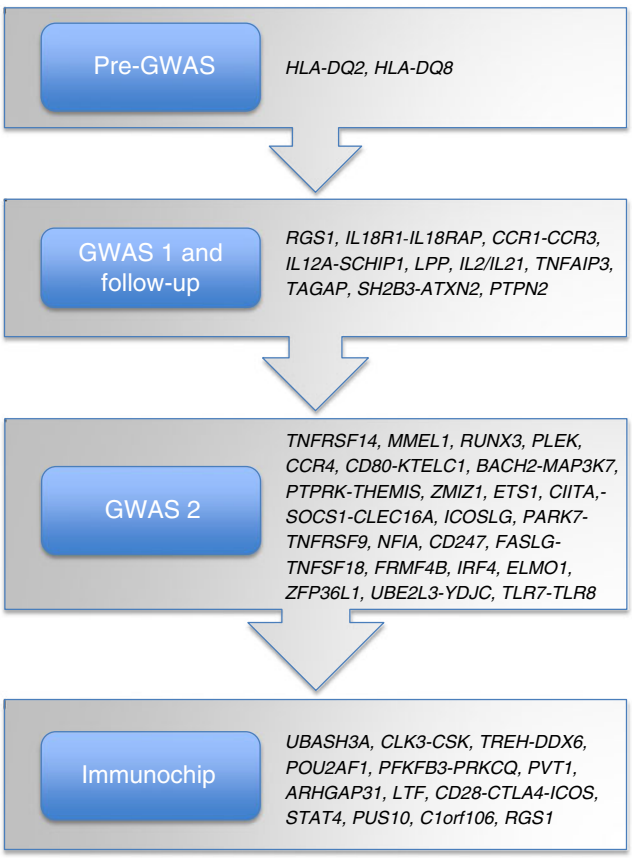


Fig. 2 History of celiac disease genetics. The final Immunochip analysis increased the number of independent non-HLA CD susceptibility SNPs to 57 (see text for further details)

by HLA plus the 57 non-HLA SNPs. The CD-CV hypothesis suggests that the remainder of the CD iceberg will also consist of common variants with very small effect sizes. Since identifying more of these variants would require extremely large cohort sizes, this would be very difficult to realize. There are several reasons why it is conceivable that the design of the current studies inhibits identification of less common genetic variants (with allele frequencies between 1–5 %): (1) the GWAS genotyping platforms are skewed towards covering common variants; (2) rare variants tend to be more population-specific but the studies conducted with the Immunochip for instance—which does contain low-frequency variants—did not take separate populations into account, thereby probably missing population-specific effects. The ultimate way to identify low frequency (allele frequency 1–5 %) and rare variants (allele frequency <1 %) requires different technologies, such as deep sequencing. The advent of whole genome sequencing is expected to reveal much of the landscape of rare variation [38], but for large population studies this approach is currently still too expensive. Another option is testing for the existence of rare variants with high effect sizes, but this requires a different strategy. Each population should be investigated separately for the disease-associated haplotype, which then needs to be resequenced to identify all the possible variants on it. However, the CD GWAS cohorts studied so far mostly consisted of populations of European descent, which limits the variation in predisposing haplotypes. To capture the vast majority of potential disease-causing rare variants would thus require as many different (multi-ethnic) CD cohorts as possible. Comparing haplotypes across different populations also has some additional advantages and may result in even more refinement of established association signals or help in identifying population-specific risk haplotypes/variants. For example, genotyping tag-SNPs at *TNFAIP3*, one of the autoimmune risk loci, in an African-American SLE cohort revealed a novel African-derived risk haplotype that was in linkage disequilibrium (LD) with a non-synonymous coding SNP [39], whereas in another study [40] re-sequencing of the same region in Europeans and Koreans revealed a deletion of T, followed by a T > A transversion in a non-coding region that showed much stronger odds ratio in Koreans than Europeans for SLE (odds ratio = 2.54 versus 1.7 in Europeans). Thus, the use of multi-ethnic disease cohorts for fine-mapping the disease-associated regions can be a powerful approach.

Now that a plethora of CD susceptibility factors has been identified, the challenge is to pinpoint the causal variants from each locus, and to prove that these causal variants affect the function of tissues and cell types involved in CD. Meeting this challenge requires a multidisciplinary approach, involving the generation and integration of bioinformatic, genetic, immunological and cell biological experimental data and clinical data. Below we will discuss the strategies that can be employed to

meet this challenge, while focusing on the non-HLA CD susceptibility loci.

Regulatory regions

Until recently, the focus of genetic studies on autoimmune diseases has been on protein coding genes and many investigators expected to find SNPs that alter protein sequences, and thereby protein function. One of the most surprising findings from the recent study by Trynka et al. [24] is that only three of the 57 independent SNPs appear to affect protein sequences (in *MMEL1*, *SH2B3*, and *IRAK1*). A careful inspection of the finely mapped loci indicates that many SNPs map to either 5′ or 3′ untranslated regions (UTRs), to introns, or to intergenic regions (Fig. 3a). The *RUNX3*, *RGS1*, *ETS1*, *TAGAP*, and *ZFP36L1* genes show association with CD in the 5'-UTR region (i.e., 1st exon and 10 kb upstream of it), suggesting that the transcriptional regulation of these genes is affected by the CD-risk SNPs. There are different ways in which 5′-UTR SNPs can exert an effect on transcription, for example by altering or creating binding sites for transcription factors, or by modifying the binding sites for chromatin-modifying protein complexes, which in turn can affect DNA methylation and/or histone modification ('epigenetic effects') [41, 42]. The association of CD to *IRF4*, *PTPRK*, and *ICOSLG* seems to affect 3′-UTR sequences which, theoretically, could lead to a decrease in stability or increased degradation of the respective mRNAs, or to inhibition of translation by, for example, altering binding sites for RNA-stabilizing/destabilizing proteins or by affecting miRNA binding sites. In the *PTPRK*

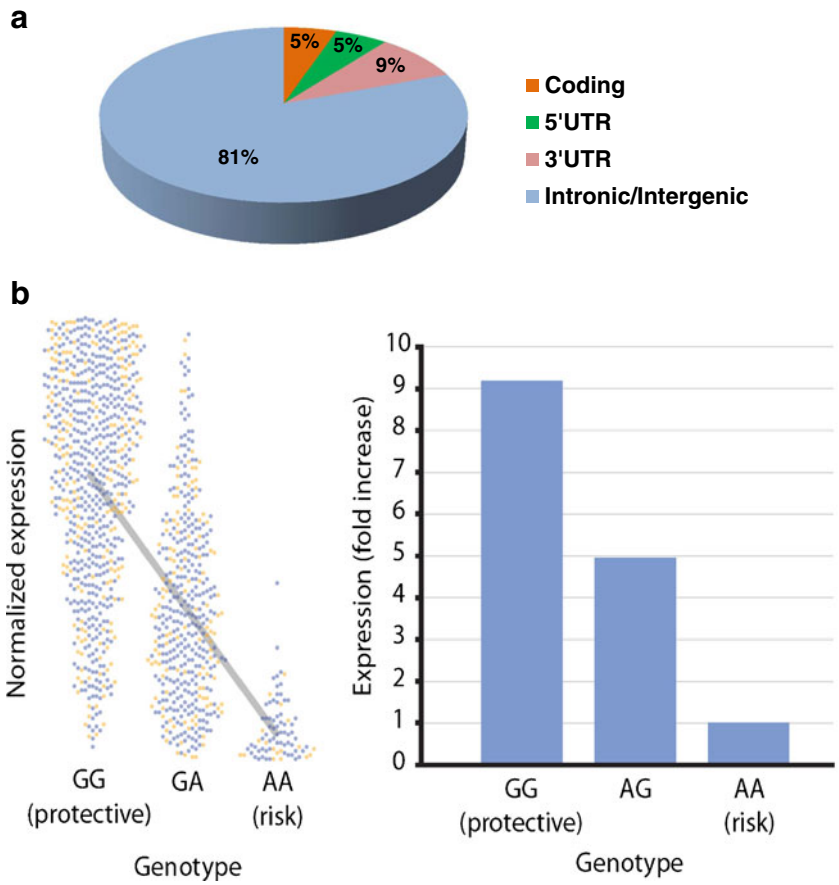


Fig. 3 Location and effect of CD risk SNPs. **a** Genomic location of the SNPs. Proxy SNPs (R^2>0.8) for 57 CD top SNPs were extracted using the 1000 Genomes Project CEU population. Only three (5 %) of the 57 SNPs were in linkage with coding variants. About 5 and 9 % are located in the 5'-UTR and the 3'-UTR regions, respectively. This leaves 81 % of the variants to be located in non-coding regions of the genome (intergenic or intronic). The latter SNPs could be involved in the regulation of gene expression or they could affect non-coding RNA species. **b** Expression quantitative trait loci (eQTL) analysis at SNP rs917997. The figure shows the association of the risk genotype with a lower expression of *IL18RAP* ($P=1.1\times10^{-133}$). The *left panel* displays the distribution of the normalized expression levels of *IL18RAP* mRNA according to the genotypes at rs917997. The *blue* and *orange dots* indicate samples from male and female volunteers, respectively. The *right panel* displays the foldchange in the levels of *IL18RAP* mRNA.

gene, one of the SNPs is located in a potential binding site for hsa-miR-1910 [24]. Furthermore, the CD SNP rs7559479 in the *IL18RAP* locus alters the binding efficiency of hsa-miR-140-3p, hsa-miR-212, and hsa-mir-27a, while another SNP in the same area (rs7603250) affects the binding of hsa-miR-668. It is important to note that rs7559479 also creates a potential binding site for has-miR-136 (as predicted by snpinfo.niehs.nih.gov). We consider it interesting that the *IL18RAP* gene displays the strongest e-QTL effect [6], corroborating the hypothesis that miRNAs may affect the expression of *IL18RAP*. Moreover, some of the CD-risk "top-SNPs" (i.e., the SNPs with the lowest *P* values) show overlap with genes encoding non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), long intergenic non-coding RNAs (lincRNAs), or small nucleolar RNAs (snoRNAs) (unpublished results), indicating that additional layers of gene regulation and gene-splicing are involved in the disease mechanism. This finding should not be surprising as about 16 % of the loci associated with complex diseases do not harbor protein-coding genes [43, 44]. Altogether, it has become clear that ~95 % of the CD-risk SNPs are located in regulatory regions (Fig. 3a). The fine mapping of CD loci is ongoing and more light will be shed on the role of these regions in the etiology of CD.

Expression QTL analysis can help to identify the causative gene in a locus with multiple candidates

It is difficult to identify the causal gene in a disease-associated locus that contains multiple candidate genes. The fact that the disease-associated SNP may not be the causal SNP, in strong LD with the true causal variant, it adds to the problem of identifying the causal gene. An elegant strategy that can be applied to narrow down the causal gene in a locus is to correlate genotypes with expression data. This approach has been coined expression QTL analysis [45–47]. Although eQTL analysis does not prove that the gene is the causal one in the locus, it can help in prioritizing genes for follow-up studies.

eQTLs come in two flavors: (1) *cis*-eQTLs in which SNPs affect expression of nearby genes [48], and (2) *trans*-eQTLs in which SNPs affect the expression of genes far away on the same chromosome or even on another chromosome [48]. Dubois et al. [23] used a dataset consisting of genome-wide gene expression data and genome-wide SNP data of 1,469 human primary blood leukocytes to perform an eQTL analysis in CD. They showed that 20 out of the 38 (53 %) non-HLA CD susceptibility loci they investigated displayed significant eQTL effects. The most impressive eQTL effect was found for SNP rs917997 in the *IL18RAP* gene ($P=7.4\times10e-87$) causing a 9-fold difference of *IL18RAP* expression between carriers of two wild-type alleles versus carriers of two risk alleles [6]. This helped to pinpoint *IL18RAP* as the likely causal gene in a locus also harboring *IL18R1*, *IL1RL1*, and *IL1RL2,* since the latter three did not display a *cis*-eQTL effect. Altogether these findings indicate that the mechanism underlying CD is governed by a deregulation of gene expression. Other immune-related diseases show similar numbers of eQTLs for disease-associated SNPs, suggesting that this is a more general phenomenon [6]: for example, 39 out of 71 CrD loci (55 %) show an eQTL effect [49], and 32 out of 53 T1D loci (60 %) [15].

The identification of eQTL effects in *trans* (*trans*-eQTLs) is much more difficult, presumably since these are more tissue specific and cell specific [50]. *Trans*-eQTLs are of interest because they implicate biological processes by linking disease SNPs to the expression pattern of many genes, thereby potentially revealing disease-associated pathways. As an example, Fehrmann et al. [48] described the *trans*-eQTL effects of 1,167 published trait- or disease-related SNPs on gene expression in peripheral blood mononuclear cells (PBMCs) of 1,469 unrelated individuals. T*rans*-eQTL effects were observed on 113 genes, of which 46 could be replicated in a dataset obtained from monocytes of 1,490 different individuals, and 18 could be replicated in a dataset generated from subcutaneous adipose, visceral adipose, liver and muscle tissue from the same replication cohort. In addition, they identified 18 unlinked SNP pairs, associated with a single phenotype and affecting the regulation of the same *trans*-gene. The fact that singular genes are regulated in *trans* by multiple SNPs could indicate the importance of the *trans*-gene in the disease mechanism. In the same study, they also found that HLA SNPs are 10-fold enriched for *trans*-eQTL effects [48].

Applying pathway analysis to zoom in on gene function and disease mechanisms

Although the GWAS approach has its shortcomings, for instance it cannot pinpoint the causal gene in all loci, the approaches described above can help suggest causal candidate genes. A significant subset of the CD susceptibility loci can be associated with T cell biology, including *REL, TNFAIP3, THEMIS/PTPRK, ETS1, RUNX3, TLR7/TLR8, BACH2,* and *IRF4* [19], but it is likely that other cell types are affected as well. Yet another strategy that can be applied to GWAS results is pathway analysis and quite a number of pathway analysis tools are now publicly available [51, 52]. In some of the pathway analysis approaches, human datasets have successfully been intersected with results obtained from model organisms such as yeast, worms and flies, to infer functional and physical interaction networks [53]. Pathway analysis algorithms predict pathways based on

connections between the genes in the query list that can be distilled from literature co-citation, gene ontology terms, co-expression, protein-protein interaction data, possession of common regulatory motifs or domains, tissue-specific co-expression, subcellular co-localization, and phenotypic profiling. All of these sources of information have been shown to provide useful data on biological function. Using these data and insights, systems biology approaches [54] can then be applied to unravel the role of the immune system in CD. While these approaches have so far been less often applied in mammalian systems, the recent availability of relevant datasets in humans and mice will facilitate such strategies.

In a recent review Wang et al. outlined the development of pathway-based approaches for GWAS and discussed their practical use and caveats [51]. Many of the available tools examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest. Gene Relationships Among Implicated Loci (GRAIL) is a computational tool that takes a list of GWAS regions and predicts the likely causal gene in each locus using information from 250,000 PubMed abstracts [55]. GRAIL can predict new loci and was successfully applied to RA, where it identified CD28, PRDM1, and CD2/CD58 as involved in the disease [56]. Functional relationships between genes and their products can also be obtained from the Kyoto Encyclopedia of Genes and Genomes [57], the Biomolecular Interaction Network Database [58], the Human Protein Reference Database [59], the Gene Ontology (GO) Database [60], predicted (tissue-specific) phenome-interactome/expression networks [61, 62], the CCSB Interactome Database [63], and microarray co-expression datasets (GEMMA; http://www.chibi.ubc.ca/Gemma).

Functionally related genes tend to be co-regulated transcriptionally, although the regulatory mechanisms can be extremely complex [64]. Despite this complexity, it is feasible to predict the function of a gene based on its "co-expressed gene signature". As an example, GEMMA was used to acquire the gene set that is co-expressed with PTPRK, a gene with an unknown function. Subsequently, a commercially available pathway analysis suite—MetaCore-GeneGO (www.genego.com/metacore.php)—was used to search for significant enrichment terms to suggest a function for PTPRK. The enrichment analysis suggested that PTPRK is involved in B cell activation (Fig. 4). Although this observation has not yet been followed up, this illustrates that these kinds of approaches can be readily applied to generate novel hypotheses.

When performing pathway analyses it is important to identify the correct tissue or cell type in which the disease gene probably operates [65]. For this, public databases such as BioGPS [66] can be used. The generally accepted view on CD pathogenesis is that CD is a T cell-mediated enteropathy in which T cells are major players in recognizing gluten epitopes in the context of HLA alleles and inducing

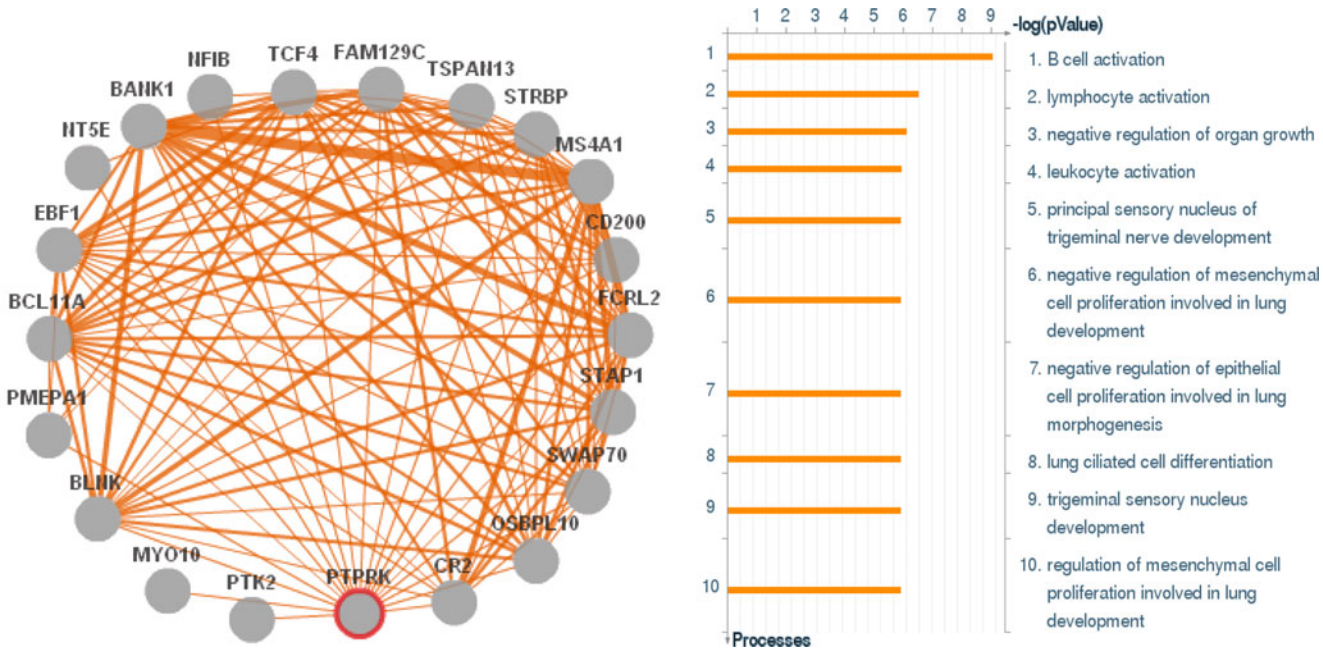


Fig. 4 Co-expression analysis to predict the function of *PTPRK* gene. The *left panel* lists the genes showing co-expression with *PTPRK* in at least 15 different microarray datasets (extracted from the GEMMA co-expression database) and depicts the presence of interactions between those genes. The *width of the lines* represents the number of datasets (ranging from 15 to 25) containing evidence for the interaction. The *right panel* displays the results of an enrichment analysis performed on the *PTPRK* co-expressed genes, using the MetaCore GeneGo tool (see text). The *x*-axis displays significance for each of the biological processes plotted on the *y*-axis

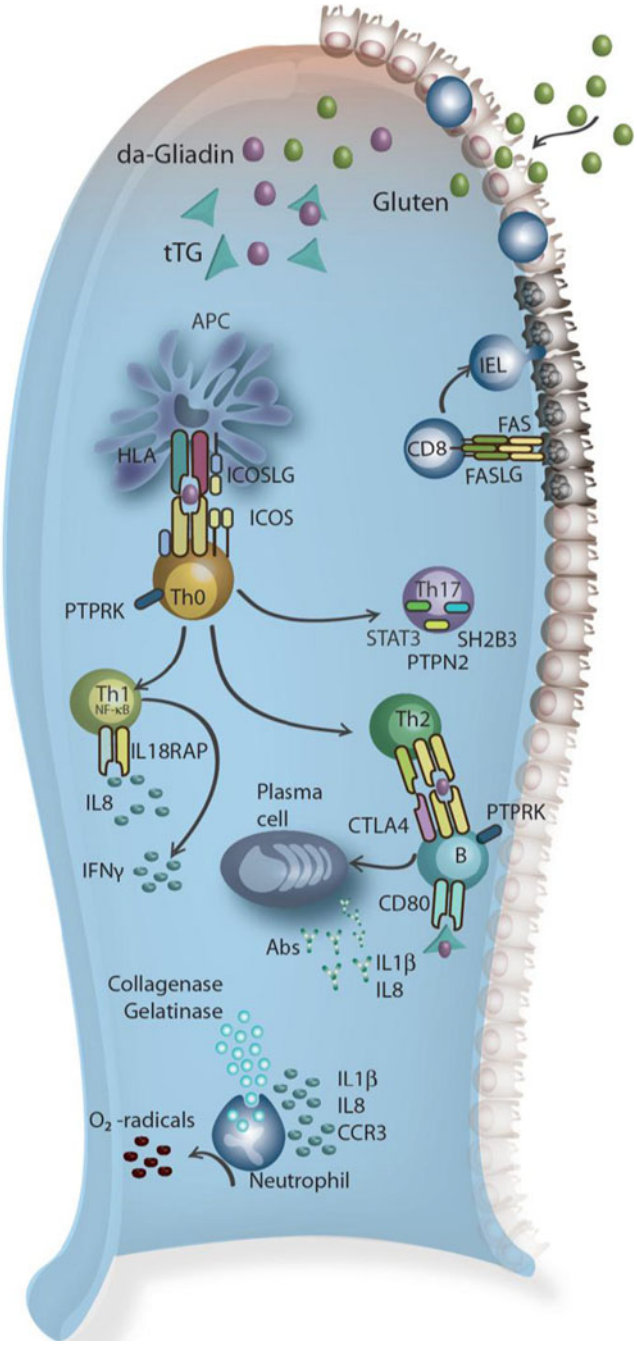


Fig. 5 Immune cell types implied to be involved in celiac disease by pathway analyses. Gluten molecules, the environmental trigger of CD, are degraded into gliadins which in turn are modified by tissue transglutaminase (*tTG*) into deamidated gliadin (*da-Gliadin*). The latter peptides are presented to the immune system, resulting in activation of various immune cell types (according to pathway analyses, see text). For a more detailed description of the genes involved in these processes, see the text and reviews by Trynka et al. [14] and Abadie et al. [67]. *Abs*, antibodies; *FASLG*, *FAS* ligand; *ICOSLG*, *ICOS* ligand; *IEL*, intraepithelial lymphocytes

anti-gluten T cell responses [67]. However, a BioGPS analysis using a human expression dataset [68] associates the CD loci not only with T cells (*TAGAP, TNFRSF14, CCR4, CTLA4, UBASH3A,* and *CD28*), but also with NK cells (*UBE2E3, RUNX3, FASLG, PTPN2,* and *IL18RAP*), neutrophils (*PLEK* and *CCR3*), and B cells (*BACH2, SOCS1, POU2AF1, ICOSLG, IRF4, CIITA, ZFP36L1,* and *CSK*) (Fig. 5). The results indicate that this tissue- or cell-specific approach can assist in generating new biological hypotheses. The BioGPS results suggest a role for NK cells in CD, while it has previously been shown that an impaired distribution of intraepithelial NK cells induces permanent loss of tolerance to gliadin [69] and that a deficiency of NK cells is involved in CD [70]. On the other hand, it is possible that the affected "NK cell genes" do not affect NK cell function, but that they are involved in the pathology mediated by intraepithelial lymphocytes (IELs) associated with CD, as it has been reported that CD IELs are derived from CD8 T cells but that they start expressing NK cell effector molecules [67]. Neutrophils may cause impaired intestinal

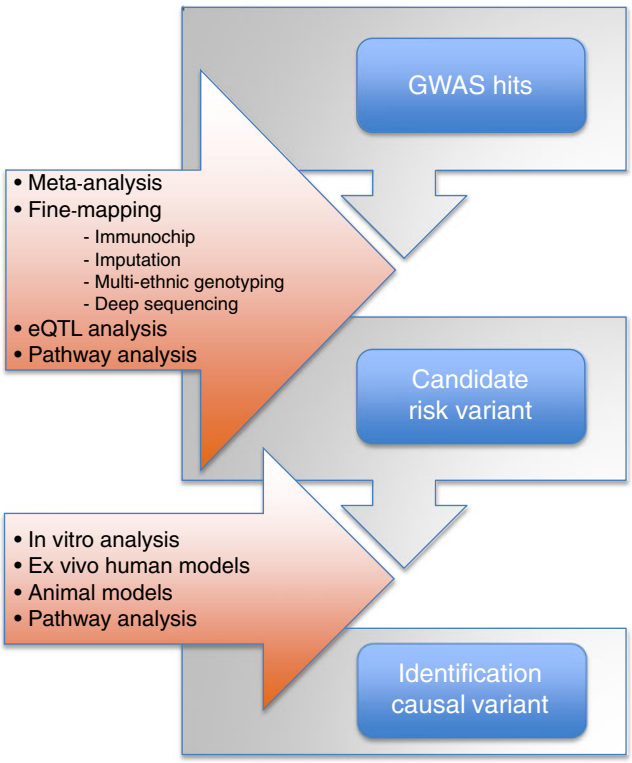


Fig. 6 Summary of strategies to identify causal variants and disease mechanisms. GWAS association signals can be followed up by meta-analysis and/or fine-mapping to identify specific causal variants. Pathway and eQTL analyses can be applied to prioritize the causative genes and to generate hypotheses to explain the biological link between a causal gene and disease. Identified causal variants and genes can in turn be followed up by experiments by, for instance, ex vivo stimulation experiments using human or animal immune cells or by experiments with inflammation models in whole animals

barrier function by inducing a chronic inflammation and could thereby contribute to CD pathogenesis [71]. Lastly, B cells could be involved in presenting gluten to T cell receptors and thus contribute to the amplification of the anti-gluten T cell response [72]. Altogether these results suggest that this kind of pathway analysis yields clinically relevant information about the contribution of multiple immune cell types to CD pathology.

It has to be kept in mind that pathway analysis is based on the use of databases that contain experimental data and that the quality of this data is not equally high for every dataset included. Moreover, these tools favor the well-defined pathways [73] and lesser-studied genes may not be taken into account, making it more difficult to identify lesser known genes and pathways involved in disease etiology. Despite these shortcomings, pathway analysis approaches are becoming a mainstay in medical research and they have already demonstrated their usefulness in generating new hypotheses that can subsequently be tested.

Conclusions

Despite decades of research on CD, we still do not understand the exact mechanisms underlying this disease. However, the recent GWAS and follow-up studies have started to uncover the genetic components contributing to this disease. Although on the genetic level immune-related diseases still show differences in, for example, the number of disease susceptibility loci, the effect sizes associated to each locus, and the environmental factors involved in the various diseases [10], it is also clear that there is a remarkable overlap of susceptibility factors between various immune-related diseases [2, 15–22]. This overlap clearly implies the involvement of shared pathways in multiple autoimmune diseases and, most importantly, suggests that general treatment modalities might be feasible for some immune-related diseases. However, not all of the results obtained so far can be readily interpreted as the resolution of the SNP analyses is, in many cases, still not high enough. Many susceptibility loci—also shared loci—still contain multiple genes. Several strategies can be applied to pinpoint the causal variants in these loci (Fig. 6) and it can be expected that, in the near future, combinations of these approaches, which involve the integration of complex datasets containing different levels of information, will identify novel causal variants associated with immune-related diseases. The elucidation of these novel components has immediate clinical relevance, as they can be included in genetic-risk modeling approaches [74]. Moreover, they might represent novel biomarkers for celiac disease, enabling physicians to diagnose all at-risk patients, preferably before the onset of symptoms, which would greatly reduce the overall cost to society and the burden on patients. Most importantly, the causal variants, or other molecules that have been identified as playing a role in the same pathway, represent new potential therapeutic targets, not only for celiac disease but also other autoimmune diseases.

Acknowledgments We thank members of the Department of Genetics of the University Medical Center in Groningen for fruitful discussions. We would also like to thank Claudia M. González Arévalo and Harm-Jan Westra for help with the graphics, and Jackie Senior for editing the final text. This work was made possible by grants to CW from the Celiac Disease Consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009), from the Netherlands Organization for Scientific Research (*NWO*, VICI grant 918.66.620), and from the Dutch Digestive Diseases Foundation (*MLDS*, WO 11-30).

References

1. Keuning JJ, Peña AS, van Leeuwen A, van Hooff JP, va Rood JJ (1976) HLA-DW3 associated with coeliac disease. Lancet 1(7958):506–508
2. Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 10(1):43–55
3. Van Heel DA, Hunt K, Greco L, Wijmenga C (2005) Genetics in coeliac disease. Best Pract Res Clin Gastroenterol 19(3):323–339
4. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109(4):1193–1198
5. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, Edkins S, Gray E, Booth DR, Potter SC, Goris A, Band G, Oturai AB, Strange A, Saarela J, Bellenguez C, Fontaine B, Gillman M, Hemmer B, Gwilliam R, Zipp F, Jayakumar A, Martin R, Leslie S, Hawkins S, Giannoulatou E, D'alfonso S, Blackburn H, Martinelli Boneschi F, Liddle J, Harbo HF, Perez ML, Spurkland A, Waller MJ, Mycko MP, Ricketts M, Comabella M, Hammond N, Kockum I, McCann OT, Ban M, Whittaker P, Kemppinen A, Weston P, Hawkins C, Widaa S, Zajicek J, Dronov S, Robertson N, Bumpstead SJ, Barcellos LF, Ravindrarajah R, Abraham R, Alfredsson L, Ardlie K, Aubin C, Baker A, Baker K, Baranzini SE, Bergamaschi L, Bergamaschi R, Bernstein A, Berthele A, Boggild M, Bradfield JP, Brassat D, Broadley SA, Buck D, Butzkueven H, Capra R, Carroll WM, Cavalla P, Celius EG, Cepok S, Chiavacci R, Clerget-Darpoux F, Clysters K, Comi G, Cossburn M, Cournu-Rebeix I, Cox MB, Cozen W, Cree BA, Cross AH, Cusi D, Daly MJ, Davis E, de Bakker PI, Debouverie M, D'hooghe MB, Dixon K, Dobosi R, Dubois B, Ellinghaus D, Elovaara I, Esposito F, Fontenille C, Foote S, Franke A, Galimberti D, Ghezzi A, Glessner J, Gomez R, Gout O, Graham C, Grant SF, Guerini FR, Hakonarson H, Hall P, Hamsten A, Hartung HP, Heard RN, Heath S, Hobart J, Hoshi M, Infante-Duarte C, Ingram G, Ingram W, Islam T, Jagodic M, Kabesch M, Kermode AG, Kilpatrick TJ, Kim C, Klopp N, Koivisto K, Larsson M, Lathrop M, Lechner-Scott JS, Leone MA, Leppä V, Liljedahl U, Bomfim IL, Lincoln RR, Link J, Liu J, Lorentzen AR, Lupoli S, Macciardi

F, Mack T, Marriott M, Martinelli V, Mason D, McCauley JL, Mentch F, Mero IL, Mihalova T, Montalban X, Mottershead J, Myhr KM, Naldi P, Ollier W, Page A, Palotie A, Pelletier J, Piccio L, Pickersgill T, Piehl F, Pobywajlo S, Quach HL, Ramsay PP, Reunanen M, Reynolds R, Rioux JD, Rodegher M, Roesner S, Rubio JP, Rückert IM, Salvetti M, Salvi E, Santaniello A, Schaefer CA, Schreiber S, Schulze C, Scott RJ, Sellebjerg F, Selmaj KW, Sexton D, Shen L, Simms-Acuna B, Skidmore S, Sleiman PM, Smestad C, Sørensen PS, Søndergaard HB, Stankovich J, Strange RC, Sulonen AM, Sundqvist E, Syvänen AC, Taddeo F, Taylor B, Blackwell JM, Tienari P, Bramon E, Tourbah A, Brown MA, Tronczynska E, Casas JP, Tubridy N, Corvin A, Vickery J, Jankowski J, Villoslada P, Markus HS, Wang K, Mathew CG, Wason J, Palmer CN, Wichmann HE, Plomin R, Willoughby E, Rautanen A, Winkelmann J, Wittig M, Trembath RC, Yaouanq J, Viswanathan AC, Zhang H, Wood NW, Zuvich R, Deloukas P, Langford C, Duncanson A, Oksenberg JR, Pericak-Vance MA, Haines JL, Olsson T, Hillert J, Ivinson AJ, De Jager PL, Peltonen L, Stewart GJ, Hafler DA, Hauser SL, McVean G, Donnelly P, Compston A, Consortium International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2 (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476(7359):214–219

6. Hrdlickova B, Westra HJ, Franke L, Wijmenga C (2011) Celiac disease: moving from genetic associations to causal variants. Clin Genet 80(3):203–313
7. Stappenbeck TS, Rioux JD, Mizoguchi A, Saitoh T, Huett A, Darfeuille-Michaud A, Wileman T, Mizushima N, Carding S, Akira S, Parkes M, Xavier RJ (2011) Crohn disease: a current perspective on genetics, autophagy and immunity. Autophagy 7 (4):355–374
8. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363(2):166–176
9. Cho JH, Gregersen PK (2011) Genomics and the multifactorial nature of human autoimmune disease. N Engl J Med 365(17):1612–1623
10. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7–24
11. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10):781–791
12. Van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CJ, McManus R, McGinnis R, Cardon LR, Deloukas P, Wijmenga C (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39(7):827–829
13. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40(4):395–402
14. Trynka G, Zhernakova A, Romanos J, Franke L, Hunt KA, Turner G, Bruinenberg M, Heap GA, Platteel M, Ryan AW, de Kovel C, Holmes GK, Howdle PD, Walters JR, Sanders DS, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, Kelleher D, Barisani D, Bardella MT, McManus R, van Heel DA, Wijmenga C (2009) Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 58(8):1078–1083
15. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS, Consortium Type 1 Diabetes Genetics Consortium (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6):703–707
16. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJ, Costenbader KH, Criswell LA, Crusius JB, Cui J, de Bakker PI, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TW, Kastner DL, Ke X, Lee AT, Liu X, Martin P, Morgan AW, Padyukov L, Posthumus MD, Radstake TR, Reid DM, Seielstad M, Seldin MF, Shadick NA, Steer S, Tak PP, Thomson W, van der Helm-van Mil AH, van der Horst-Bruinsma IE, van der Schoot CE, van Riel PL, Weinblatt ME, Wilson AG, Wolbink GJ, Wordsworth BP, Wijmenga C, Karlson EW, Toes RE, de Vries N, Begovich AB, Worthington J, Siminovitch KA, Gregersen PK, Klareskog L, Plenge RM, BIRAC Consortium, YEAR Consortium (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 42(6):508–514
17. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, Westra HJ, Fehrmann RS, Kurreeman FA, Thomson B, Gupta N, Romanos J, McManus R, Ryan AW, Turner G, Brouwer E, Posthumus MD, Remmers EF, Tucci F, Toes R, Grandone E, Mazzilli MC, Rybak A, Cukrowska B, Coenen MJ, Radstake TR, van Riel PL, Li Y, de Bakker PI, Gregersen PK, Worthington J, Siminovitch KA, Klareskog L, Huizinga TW, Wijmenga C, Plenge RM (2011) Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet 7(2):e1002004
18. Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, van der Steege G, Radstake TR, Barrera P, Roep BO, Koeleman BP, Wijmenga C (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet 81 (6):1284–1288
19. Trynka G, Wijmenga C, van Heel DA (2010) A genetic perspective on coeliac disease. Trends Mol Med 16(11):537–550
20. Festen EA, Goyette P, Scott R, Annese V, Zhernakova A, Lian J, Lefèbvre C, Brant SR, Cho JH, Silverberg MS, Taylor KD, de Jong DJ, Stokkers PC, Mcgovern D, Palmieri O, Achkar JP, Xavier RJ, Daly MJ, Duerr RH, Wijmenga C, Weersma RK, Rioux JD (2009) Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut 58(6):799–804
21. Festen EA, Goyette P, Green T, Boucher G, Beauchamp C, Trynka G, Dubois PC, Lagacé C, Stokkers PC, Hommes DW, Barisani D, Palmieri O, Annese V, van Heel DA, Weersma RK, Daly MJ, Wijmenga C, Rioux JD (2011) A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. PLoS Genet 7(1):e1001283
22. Gutierrez-Achury J, Coutinho de Almeida R, Wijmenga C (2011) Shared genetics in coeliac disease and other immune-mediated diseases. J Intern Med 269(6):591–603
23. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, Bardella MT, van den Berg LH, Bockett NA, de la Concha EG, Dema B, Fehrmann RS, Fernández-Arquero M, Fiatal S, Grandone E, Green PM, Groen HJ, Gwilliam R, Houwen RH, Hunt SE, Kaukinen K, Kelleher D, Korponay-Szabo I, Kurppa K, MacMathuna P, Mäki M, Mazzilli MC, McCann OT, Mearin ML, Mein CA, Mirza MM, Mistry V, Mora B, Morley KI, Mulder CJ, Murray JA, Núñez C, Oosterom E, Ophoff RA, Polanco I, Peltonen L, Platteel M, Rybak

A, Salomaa V, Schweizer JJ, Sperandeo MP, Tack GJ, Turner G, Veldink JH, Verbeek WH, Weersma RK, Wolters VM, Urcelay E, Cukrowska B, Greco L, Neuhausen SL, McManus R, Barisani D, Deloukas P, Barrett JC, Saavalainen P, Wijmenga C, van Heel DA (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42(4):295–302

24. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha EG, de Almeida RC, Dias KR, van Diemen CC, Dubois PC, Duerr RH, Edkins S, Franke L, Fransen K, Gutierrez J, Heap GA, Hrdlickova B, Hunt S, Izurieta LP, Izzo V, Joosten LA, Langford C, Mazzilli MC, Mein CA, Midah V, Mitrovic M, Mora B, Morelli M, Nutland S, Núñez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Ribes-Koninckx C, Ricaño-Ponce I, Rich SS, Rybak A, Santiago JL, Senapati S, Sood A, Szajewska H, Troncone R, Varadé J, Wallace C, Wolters VM, Zhernakova A, Spanish Consortium on the Genetics of Coeliac Disease (CEGEC), PreventCD Study Group, Wellcome Trust Case Control Consortium (WTCCC), Thelma BK, Cukrowska B, Urcelay E, Bilbao JR, Mearin ML, Barisani D, Barrett JC, Plagnol V, Deloukas P, Wijmenga C, van Heel DA (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43(12):1193–1201
25. Cortes A, Brown MA (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13(1):101
26. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JH, Howson JM, Stevens H, McManus R, Wijmenga C, Heap GA, Dubois PC, Clayton DG, Hunt KA, van Heel DA, Todd JA (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med 359(26):2767–2777
27. Allen PM (2009) Themis imposes new law and order on positive selection. Nat Immunol 10(8):805–806
28. Asano A, Tsubomatsu K, Jung CG, Sasaki N, Agui T (2007) A deletion mutation of the protein tyrosine phosphatase kappa (Ptprk) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat. Mamm Genome 18(11):779–786
29. Shea J, Agarwala V, Philippakis AA, Maguire J, Banks E, Depristo M, Thomson B, Guiducci C, Onofrio RC, Kathiresan S, Gabriel S, Burtt NP, Daly MJ, Groop L, Altshuler D, Myocardial Infarction Genetics Consortium (2011) Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. Nat Genet 43(8):801–805
30. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11(7):499–511
31. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320
32. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747–753
33. Buchanan CC, Torstenson ES, Bush WS, Ritchie MD (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc 19(2):289–294
34. Jostins L, Morley KI, Barrett JC (2011) Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. Eur J Hum Genet 19(6):662–666
35. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405(6788):847–856
36. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17(9):502–510
37. Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145
38. Singleton AB, Hardy J, Traynor BJ, Houlden H (2010) Towards a complete resolution of the genetic architecture of disease. Trends Genet 26(10):438–442
39. Lodolce JP, Kolodziej LE, Rhee L, Kariuki SN, Franek BS, McGreal NM, Logsdon MF, Bartulis SJ, Perera MA, Ellis NA, Adams EJ, Hanauer SB, Jolly M, Niewold TB, Boone DL (2010) African-derived genetic polymorphisms in TNFAIP3 mediate risk for autoimmunity. J Immunol 184(12):7001–7009
40. Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, Bates JS, Hu Y, Kelly JA, Kaufman KM, Guthridge JM, Alarcón-Riquelme ME, Anaya JM, Bae SC, Bang SY, Boackle SA, Brown EE, Petri MA, Gallant C, Ramsey-Goldman R, Reveille JD, Vila LM, Criswell LA, Edberg JC, Freedman BI, Gregersen PK, Gilkeson GS, Jacob CO, James JA, Kamen DL, Kimberly RP, Martin J, Merrill JT, Niewold TB, Park SY, Pons-Estel BA, Scofield RH, Stevens AM, Tsao BP, Vyse TJ, Langefeld CD, Harley JB, Moser KL, Webb CF, Humphrey MB, Montgomery CG, Gaffney PM, Networks BaG (2011) Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat Genet 43(3):253–258
41. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Sci 312(5777):1215–1217
42. Stefan M, Jacobson EM, Huber AK, Greenberg DA, Li CW, Skrabanek L, Conception E, Fadlalla M, Ho K, Tomer Y (2011) Novel variant of thyroglobulin promoter triggers thyroid autoimmunity through an epigenetic interferon alpha-modulated mechanism. J Biol Chem 286(36):31168–31179
43. Hindorff LA, Gillanders EM, Manolio TA (2011) Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis 32(7):945–954
44. Frazer K, Ballinger D, Cox D, Hinds D, Stuve L, Gibbs R, Belmont J, Boudreau A, Hardenbol P, Leal S, Pasternak S, Wheeler D, Willis T, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel S, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio R, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye M, Tsui S, Xue H, Wong J, Galver L, Fan J, Gunderson K, Murray S, Oliphant A, Chee M, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J, Phillips M, Roumy S, Sallée C, Verner A, Hudson T, Kwok P, Cai D, Koboldt D, Miller R, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L, Mak W, Song Y, Tam P, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird C, Delgado M, Dermitzakis E, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger B, Whittaker P, Bentley D, Daly M, de Bakker P, Barrett J, Chretien Y, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter D, Sabeti P, Saxena R, Schaffner S, Sham P, Varilly P, Stein L, Krishnan L, Smith A, Tello-Ruiz M, Thorisson G, Chakravarti A, Chen P, Cutler D, Kashuk C, Lin S, Abecasis G, Guan W, Li Y, Munro H, Qin Z, Thomas D, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon L, Clarke G, Evans D, Morris A, Weir B, Mullikin J, Sherry S, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer D, Suda E, Rotimi C, Adebamowo C, Ajayi I, Aniagwu T, Marshall P, Nkwodimmah C, Royal C, Leppert M, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole I, Knoppers B, Foster M, Clayton E, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock G, Yakub I, Birren B, Wilson R, Fulton L,

Rogers J, Burton J, Carter N, Clee C, Griffiths M, Jones M, McLay K, Plumb R, Ross M, Sims S, Willey D, Chen Z, Han H, Kang L, Godbout M, Wallenburg J, L'Archevêque P, Bellemare G, Saeki K, An D, Fu H, Li Q, Wang Z, Wang R, Holden A, Brooks L, McEwen J, Guyer M, Wang V, Peterson J, Shi M, Spiegel J, Sung L, Zacharia L, Collins F, Kennedy K, Jamieson R, Stewart J, International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861

45. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nat Genet 39(10):1202–1207
46. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448 (7152):470–473
47. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Sci 315(5813):848–853
48. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, Fu J, Deelen P, Groen HJ, Smolonska A, Weersma RK, Hofstra RM, Buurman WA, Rensen S, Wolfs MG, Platteel M, Zhernakova A, Elbers CC, Festen EM, Trynka G, Hofker MH, Saris CG, Ophoff RA, van den Berg LH, van Heel DA, Wijmenga C, Te Meerman GJ, Franke L (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 7(8):e1002197
49. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42(12):1118–1125
50. Rotival M, Zeller T, Wild PS, Maouche S, Szymczak S, Schillert A, Castagné R, Deiseroth A, Proust C, Brocheton J, Godefroy T, Perret C, Germain M, Eleftheriadis M, Sinning CR, Schnabel RB, Lubos E, Lackner KJ, Rossmann H, Münzel T, Rendon A, Erdmann J, Deloukas P, Hengstenberg C, Diemert P, Montalescot G, Ouwehand WH, Samani NJ, Schunkert H, Tregouet DA, Ziegler A, Goodall AH, Cambien F, Tiret L, Blankenberg S, Cardiogenics Consortium (2011) Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet 7(12):e1002367
51. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11(12):843–854
52. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12(9):628–640
53. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, Auluck PK, Geddie ML, Valastyan JS, Karger DR, Lindquist S, Fraenkel E (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet 41(3):316–323
54. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser ID (2011) Systems biology in immunology: a computational modeling perspective. Annu Rev Immunol 29:527–585
55. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ, Consortium IS (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 5(6):e1000534
56. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, Catanese JJ, Xie G, Stahl EA, Chen R, Alfredsson L, Amos CI, Ardlie KG, Barton A, Bowes J, Burtt NP, Chang M, Coblyn J, Costenbader KH, Criswell LA, Crusius JB, Cui J, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TW, Kastner DL, Ke X, Kurreeman FA, Lee AT, Liu X, Li Y, Martin P, Morgan AW, Padyukov L, Reid DM, Seielstad M, Seldin MF, Shadick NA, Steer S, Tak PP, Thomson W, van der Helm-van Mil AH, van der Horst-Bruinsma IE, Weinblatt ME, Wilson AG, Wolbink GJ, Wordsworth P, Altshuler D, Karlson EW, Toes RE, de Vries N, Begovich AB, Siminovitch KA, Worthington J, Klareskog L, Gregersen PK, Daly MJ, Plenge RM, BIRAC Consortium, YEAR Consortium (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nature genetics 41(12):1313–1318
57. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
58. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, Hogue CW (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33(Database issue): D418–D424
59. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, Niranjan V, Harsha HC, Talreja N, Vrushabendra BM, Ramya MA, Yatish AJ, Joy M, Shivashankar HN, Kavitha MP, Menezes M, Choudhury DR, Ghosh N, Saravana R, Chandran S, Mohan S, Jonnalagadda CK, Prasad CK, Kumar-Sinha C, Deshpande KS, Pandey A (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32(Database issue):D497–D501
60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
61. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S (2008) A large-scale analysis of tissue-specific pathology and gene expression of

human disease genes and complexes. Proc Natl Acad Sci U S A 105(52):20870–20875
62. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25(3):309–316
63. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437 (7062):1173–1178
64. Brazhnik P, de la Fuente A, Mendes P (2002) Gene networks: how to put the function in genomics. Trends Biotechnol 20(11):467–472
65. Hu X, Kim H, Stahl E, Plenge R, Daly M, Raychaudhuri S (2011) Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am J Hum Genet 89(4):496–506
66. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, Su AI (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10(11):R130
67. Abadie V, Sollid LM, Barreiro LB, Jabri B (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu Rev Immunol 29:493–525
68. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101 (16):6062–6067
69. Hadziselimovic F, Emmons LR, Schaub U, Signer E, Bürgin-Wolff A, Krstic R (1992) Occurrence of large granular lymphocytes and natural killer cells in the epithelium of the gut distinguishes two different coeliac diseases. Gut 33(6):767–772
70. Grose RH, Thompson FM, Cummins AG (2008) Deficiency of 6B11+ invariant NK T-cells in celiac disease. Dig Dis Sci 53 (7):1846–1851
71. Diosdado B, van Bakel H, Strengman E, Franke L, van Oort E, Mulder CJ, Wijmenga C, Wapenaar MC (2007) Neutrophil recruitment and barrier impairment in celiac disease: a genomic study. Clin Gastroenterol Hepatol 5(5):574–581
72. Kumar V, Wijmenga C (2011) Celiac disease: update from the 14th International Celiac Disease Symposium 2011. Expert Rev Gastroenterol Hepatol 5(6):685–687
73. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33(5):419–431
74. Romanos J, van Diemen CC, Nolte IM, Trynka G, Zhernakova A, Fu J, Bardella MT, Barisani D, McManus R, van Heel DA, Wijmenga C (2009) Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology 137(3):834–840
75. Nisticò L, Fagnani C, Coto I, Percopo S, Cotichini R, Limongelli MG, Paparo F, D'Alfonso S, Giordano M, Sferlazzas C, Magazzù G, Momigliano-Richiardi P, Greco L, Stazi MA (2006) Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55(6):803–808