Human Molecular Genetics, 2015, Vol. 24, No. 2 **397–409**
doi:10.1093/hmg/ddu453
Advance Access published on September 4, 2014

Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels

Vinod Kumar[1,*], Javier Gutierrez-Achury[1,†], Kartiek Kanduri[4,†], Rodrigo Almeida[1,†], Barbara Hrdlickova[1], Daria V. Zhernakova[1], Harm-Jan Westra[1], Juha Karjalainen[1], Isis Ricaño-Ponce[1], Yang Li[1], Anna Stachurska[1], Ettje F. Tigchelaar[1], Wayel H. Abdulahad[2], Harri Lähdesmäki[5], Marten H. Hofker[3], Alexandra Zhernakova[1], Lude Franke[1], Riitta Lahesmaa[4], Cisca Wijmenga[1] and Sebo Withoff[1,*]

[1]Department of Genetics, [2]Department of Rheumatology and Clinical Immunology and, [3]Department of Pediatrics, Molecular Genetics Section, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, [4]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku 20520, Finland and [5]Department of Information and Computer Science, Aalto University School of Science, Espoo 02150, Finland

Received December 5, 2013; Revised and Accepted September 1, 2014

Although genome-wide association studies and fine mapping have identified 39 non-HLA loci associated with celiac disease (CD), it is difficult to pinpoint the functional variants and susceptibility genes in these loci. We applied integrative approaches to annotate and prioritize functional single nucleotide polymorphisms (SNPs), genes and pathways affected in CD. CD-associated SNPs were intersected with regulatory elements categorized by the ENCODE project to prioritize functional variants, while results from *cis*-expression quantitative trait loci (eQTL) mapping in 1469 blood samples were combined with co-expression analyses to prioritize causative genes. To identify the key cell types involved in CD, we performed pathway analysis on RNA-sequencing data from different immune cell populations and on publicly available expression data on non-immune tissues. We discovered that CD SNPs are significantly enriched in B-cell-specific enhancer regions, suggesting that, besides T-cell processes, B-cell responses play a major role in CD. By combining eQTL and co-expression analyses, we prioritized 43 susceptibility genes in 36 loci. Pathway and tissue-specific expression analyses on these genes suggested enrichment of CD genes in the Th1, Th2 and Th17 pathways, but also predicted a role for four genes in the intestinal barrier function. We also discovered an intricate transcriptional connectivity between CD susceptibility genes and interferon-γ, a key effector in CD, despite the absence of CD-associated SNPs in the IFNG locus. Using systems biology, we prioritized the CD-associated functional SNPs and genes. By highlighting a role for B cells in CD, which classically has been described as a T-cell-driven disease, we offer new insights into the mechanisms and pathways underlying CD.

INTRODUCTION

Celiac disease (CD) is one of the best-understood complex diseases; it is the only disease for which both the triggering environmental and major genetic predisposing factors are known. The chronic inflamed condition of the small intestine in CD is triggered by gluten peptides mainly derived from dietary wheat, rye and barley (1). The only known therapy for

*To whom correspondence should be addressed at: Department of Genetics, University Medical Center Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands. Tel: +31 503617245; Fax: +31 503617230. Email: s.withoff@umcg.nl (S.W.); Email: v.kumar@umcg.nl or vinodkmagadi@gmail.com (V.K.)
†J.G.-A., K.K. and R.A. authors share the second authorship.

CD is a gluten-free diet, but this can induce complications such as nutrient deficiency and a decreased quality of life (2,3). The identification of susceptibility genes for CD may lead to alternative therapeutic options. In the last decade, genome-wide association studies (GWAS) have identified thousands of common variants for hundreds of complex human phenotypes (4). Along with the well-characterized HLA subtypes *HLA-DQ2* and/or *HLA-DQ8*, GWAS have pinpointed many CD-associated loci (5–7). Using the Immunochip platform, Trynka *et al*. (8) identified a total of 39 genome-wide significant (GWS) non-HLA loci, harboring 57 independent single nucleotide polymorphisms (SNPs) that are associated with CD. Their study aimed to fine-map GWAS loci and indeed decreased the average size of CD susceptibility loci from ∼300 to ∼50 kb. Although GWAS results have contributed significantly to our understanding of the genetic architecture and risk profile of CD patients (9), the inherent nature of the human genome—specifically its linkage disequilibrium (LD) structure (10)—complicates identification of the disease-predisposing genes and variants.

Remarkably, only 5% of the CD-associated variants are localized in protein-coding exons, while the others are located in the non-coding part of the genome that is known to harbor gene regulatory motifs or non-coding genes (11,12). Moreover, it was shown that SNPs associated to disease can alter the expression of long non-coding RNAs (13). These insights provide a novel perspective on the mechanisms modulated by 'regulatory SNPs'. New layers of regulatory information, for instance provided by the ENCODE project (14), should therefore be interrogated to fully understand the effects of CD-associated SNPs localizing to non-coding regions. Since pinpointing the true causal genes either by expression quantitative trait loci (eQTL) mapping, or by pathway or gene regulatory network analyses, has helped us understand the pathogenesis of several complex diseases (15–19), we aimed to apply these methods to identify CD-associated causal genes.

We integrated different layers of functional data, such as eQTL mapping results, tissue-specific expression data, DNase I hypersensitivity (DHS) analysis results, DNase I footprints, chromatin immunoprecipitation followed by sequencing (ChIP-seq) data and results from network and pathway analyses, to prioritize the plausible functional SNPs and genes from the 39 known CD-associated loci. We show that of all these SNPs, 95% are located within regulatory regions and are enriched in B-cell-specific enhancers.

Pathway analysis on prioritized CD genes not only showed their enrichment for immune pathways, but also suggested that four poorly characterized genes *(LPP*, *C1orf106*, *ARHGAP31* and *PTPRK)* are involved in intestinal barrier function, a process known to be affected in CD (20,21).

Finally, we examined the expression of the prioritized CD genes for correlation with the expression of the interferon-γ (IFNγ) gene. We have previously shown that increased expression of IFNγ correlates with tissue damage in CD, but had not yet found any evidence for a susceptibility to CD being due to genetic variations in the IFNG locus (22). This had suggested that high IFNγ levels were merely a consequence of the disease. Our new analysis presents a possible genetic explanation for the elevated expression of IFNγ in CD.

RESULTS

Functional annotation indicates CD SNPs to be located more often within gene regulatory regions

The 57 GWS SNPs at the 39 CD loci are not necessarily the disease-predisposing ones, since there are often SNPs for each locus very strongly correlated to GWS SNPs due to high or perfect LD. Therefore, we also extracted 620 proxy SNPs ($r^2 \geq 0.9$) to the 57 GWS CD-SNPs resulting in 677 SNPs for functional annotation using RegulomeDB (23). Our detailed, step-wise annotation and prioritization process is depicted in Figure 1A and B and Supplementary Material, Figure S1 and explained in the Supplementary Materials. Overall, we found that 66% (448/677) of the variants in 37 CD loci overlapped with at least one layer of ENCODE data, which is a 'minimum-evidence' criterium for a regulatory function (Fig. 1B). Moreover, 26 SNPs in 19 loci overlapped with three or more layers of functional information (e.g. a DHSplus a TF-binding site and a matched TF motif); these were prioritized as functional SNPs with 'maximum evidence' (Fig. 1B and Supplementary Material, Table S1). By setting criteria for a functional SNP to show either maximum or medium evidence, 41 variants in 21 loci could be prioritized as functional. Of these 41 variants, 95% are located within DHS, more than 95% changed TF-binding motifs and 68% are located within enhancers (Fig. 2A and Supplementary Material, Table S2). These findings indicate that the majority of CD SNPs are likely to influence disease through mechanisms regulating gene expression.

CD SNPs are significantly enriched in B-cell-specific enhancers

Our SNP prioritization analysis showed that 68% of the prioritized CD SNPs are associated with enhancers. Since enhancers are gene elements that are important regulators of cell-type-specific transcription control, we tested whether CD SNPs function in a 'cell-type-specific' manner by performing enhancer-enrichment analysis on 677 CD SNPs using the HaploReg database (24,25). We observed a significant enrichment ($P \leq 7.38 \times 10^{-5}$) of three cell enhancers for 677 CD SNPs [embryonic stem cell- (H1), B-lymphoblastoid cell- (GM12878) and erythrocytic leukemia cell- (K562) enhancers]. We then compared the fold enrichment of enhancers in these three cell lines to investigate whether CD SNPs are located significantly more often in cell-type-specific enhancers than any random set of SNPs present within 1 Mb around GWAS CD SNPs (see Supplementary Methods). Interestingly, we observed a more than 2.5-fold enrichment of B-cell-specific enhancers for CD SNPs ($P = 0.029$; Fig. 2B), suggesting that a subset of CD SNPs may specifically regulate gene expression in B cells.

eQTL and network analyses prioritize immune genes as functional genes in CD loci

Although 25 single genes were located nearest to or within LD blocks fine mapped by Immunochip (Fig. 3), many of the CD-associated loci were found to harbor intriguing candidate genes, which may lead to bias in pinpointing the true functional gene. We therefore applied eQTL and network analyses to

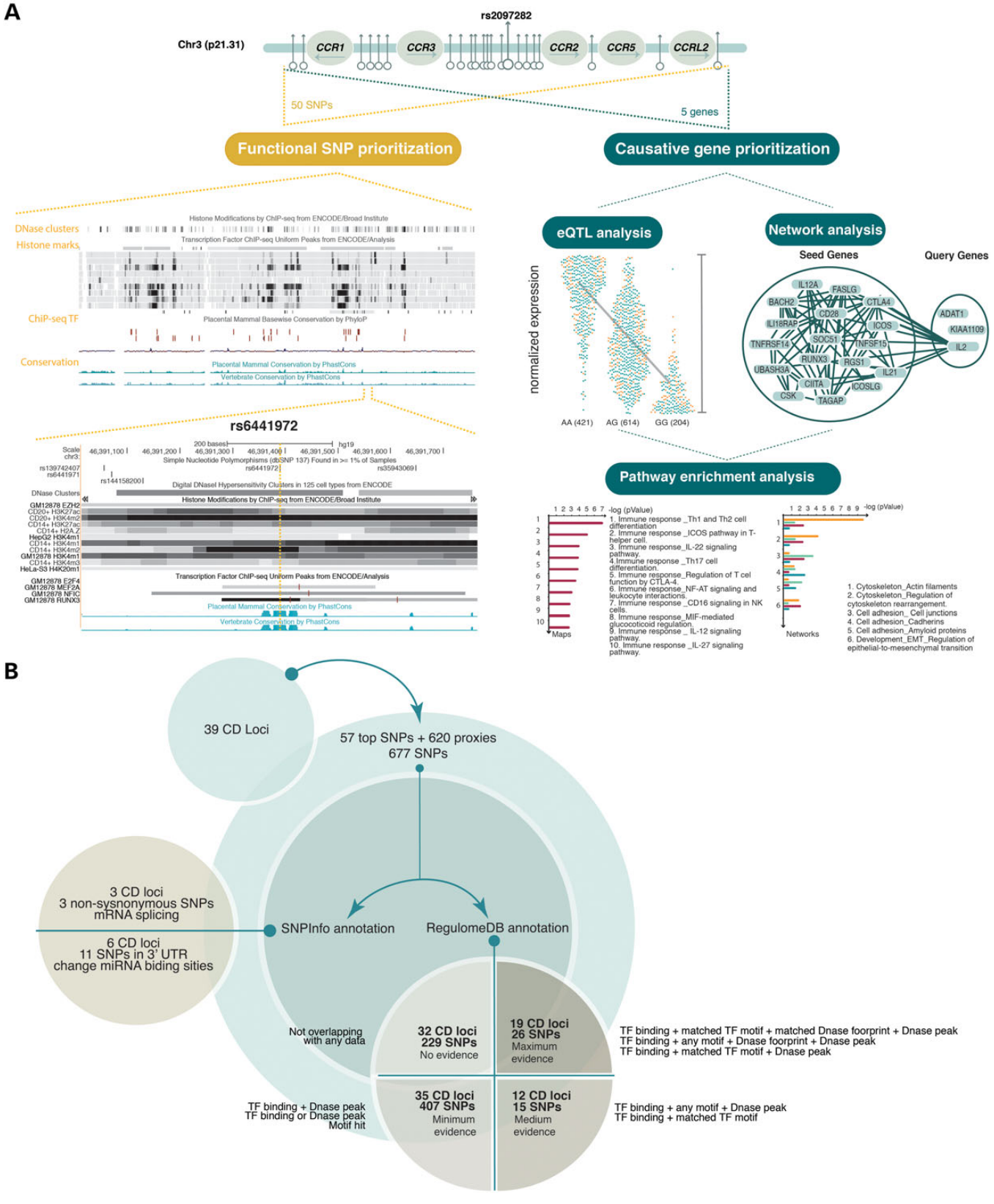

Figure 1. A detailed summary of our integrative approach used to prioritize functional SNPs and functional genes. (**A**) All 50 SNPs that are in LD ($R^2 \geq 0.9$) with CD-associated rs2097282 were intersected with functional data to prioritize rs6441972 as the most likely functional SNP. Five genes (*CCR1*, *CCR3*, *CCR2*, *CCR5* and *CCRL2*) in the associated LD region were queried initially by eQTL mapping and *CCR3* was prioritized as the functional gene from this locus. Genes without eQTL were selected as query genes to be prioritized based on further network analysis. Finally, the prioritized genes from both eQTL and network analyses were included in a pathway enrichment analysis. (**B**) A detailed annotation and prioritization of 677 CD-associated SNPs.

A

Locus	cis-eQTL	Lead SNP + Proxies	RegulomeDB functional annotation				SNPinfo functional annotation		
			Maximum evidence	Medium evidence	Minimum eveidence	No evidence	Non-synonymous	microRNA binding	Splicing
UBE2L3 -rs4821124		60	4	1	40	15	0	2	1
RUNX3 -rs72657048		47	2	2	34	9	0	0	0
MMEL1/TNFRSF14 -rs4445406		54	2	1	44	7	1	0	3
CSK -rs1378938		15	2	1	11	1	0	1	1
RGS1 -rs1359062		24	1	2	17	4	0	0	0
PTPN2 -rs11875687		20	1	2	8	9	0	0	0
CCR3 -rs2097282		50	1	1	26	22	0	1	0
ZFP36L1 -rs11851414		5	1	1	3	0	0	0	1
ICOSLG -rs58911644		15	1	0	12	2	0	0	0
TREH/DDX6 -rs10892258		19	1	0	10	8	0	0	0
UBE2E3 -rs1018326		13	1	0	6	6	0	0	0
SOCS1 -rs243323		7	1	0	6	0	0	0	0
TNFAIP3 -rs17264332		6	1	0	4	1	0	0	0
PVT1 -rs10808568		7	1	0	3	3	0	0	0
ETS1 -rs61907765		7	1	0	3	3	0	0	0
CIITA -rs6498114		5	1	0	3	1	0	0	0
IRF4 -rs1050976		4	1	0	3	0	0	4	0
IL18RAP -rs990171		62	0	1	26	35	0	2	0
STAT4 -rs6715106		11	0	1	8	2	0	0	0
PRKCQ -rs2387397		1	0	1	0	0	0	0	0
ELMO1 -rs79758729		48	0	0	24	24	0	0	0
LPP -rs2030519		10	0	0	8	2	0	0	0
FASLG -rs12068671		14	0	0	7	7	0	0	0
PLEK -rs10167650		14	0	0	7	7	0	0	0
BACH2 -rs7753008		11	0	0	5	6	0	0	0
PUS10/AHSA2 -rs13003464		7	0	0	5	2	0	0	0
CD28/CTLA4 -rs1980422		7	0	0	5	2	0	0	0
ARHGAP31 -rs61579022		8	0	0	4	4	0	0	0
PTPRK -rs55743914		5	0	0	3	2	0	1	0
SH2B3 -rs3184504		4	0	0	3	1	1	0	1
C1orf106 -rs10800746		3	0	0	3	0	0	0	0
TAGAP -rs182429		5	0	0	2	3	0	0	1
IL2/IL21 -rs13132308		4	0	0	2	2	0	0	0
CCR4 -rs4678523		3	0	0	2	1	0	0	0
IL12A -rs76830965		2	0	0	2	0	0	0	0
IRAK1 -rs13397		2	0	0	2	0	1	0	0
ZMIZ1 -rs1250552		1	0	0	1	0	0	0	0
POU2AF1 -rs7104791		1	0	0	0	1	0	0	0
UBASH3A -rs1893592		1	0	0	0	1	0	0	0

B

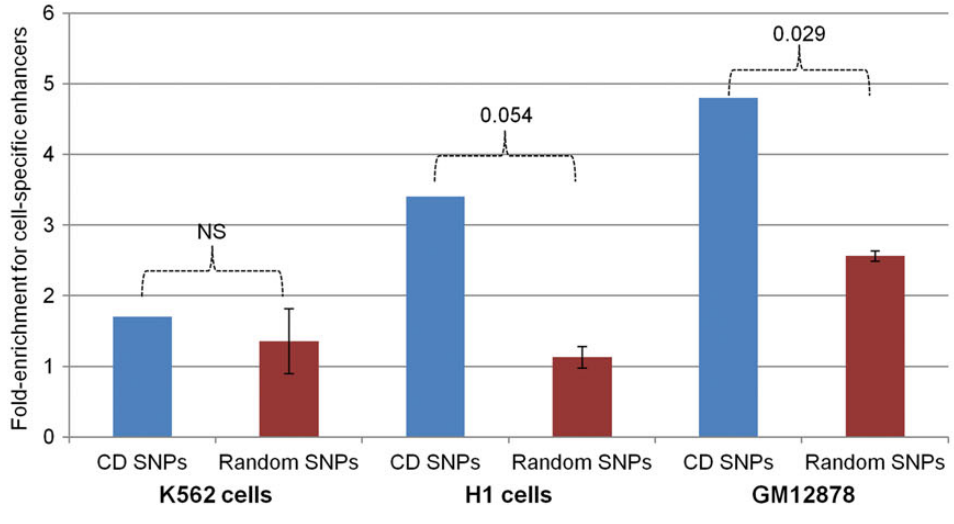

Figure 2. Functional annotation of CD-associated SNPs. (**A**) The 57 GWS SNPs from 39 CD loci and their 620 proxy SNPs were intersected with functional annotation data using RegulomeDB and SNPinfo database; the results are shown separately. The number and length of the color bars indicate the total number of SNPs in each functional category. The loci are ordered according to the highest number of SNPs showing maximum evidence for being functional SNPs. CD-associated SNPs are often located within a cell-type-specific enhancer. (**B**) A plot to show the difference in fold enrichment of enhancers for CD SNPs compared with a random set of SNPs. [Embryonic stem cell- (H1), B-lymphoblastoid cell- (GM12878) and erythrocytic leukemia cell- (K562) enhancers.] The significance of the difference in fold enrichment was tested using the *t*-test. The HaploReg database was used to calculate the fold enrichment of enhancers.

prioritize the most likely functional genes in all CD loci (Supplementary Materials).

We mapped *cis*-eQTL for 50/57 independent CD susceptibility SNPs (Supplementary Material, Table S3) in 1469 peripheral blood mononuclear cell (PBMC) samples and found that 19 SNPs were associated with altered expression levels of 18 genes (Table 1). Of these, 13 SNPs reached the Bonferroni threshold ($P < 5.49 \times 10^{-6}$), whereas 6 SNPs were modestly

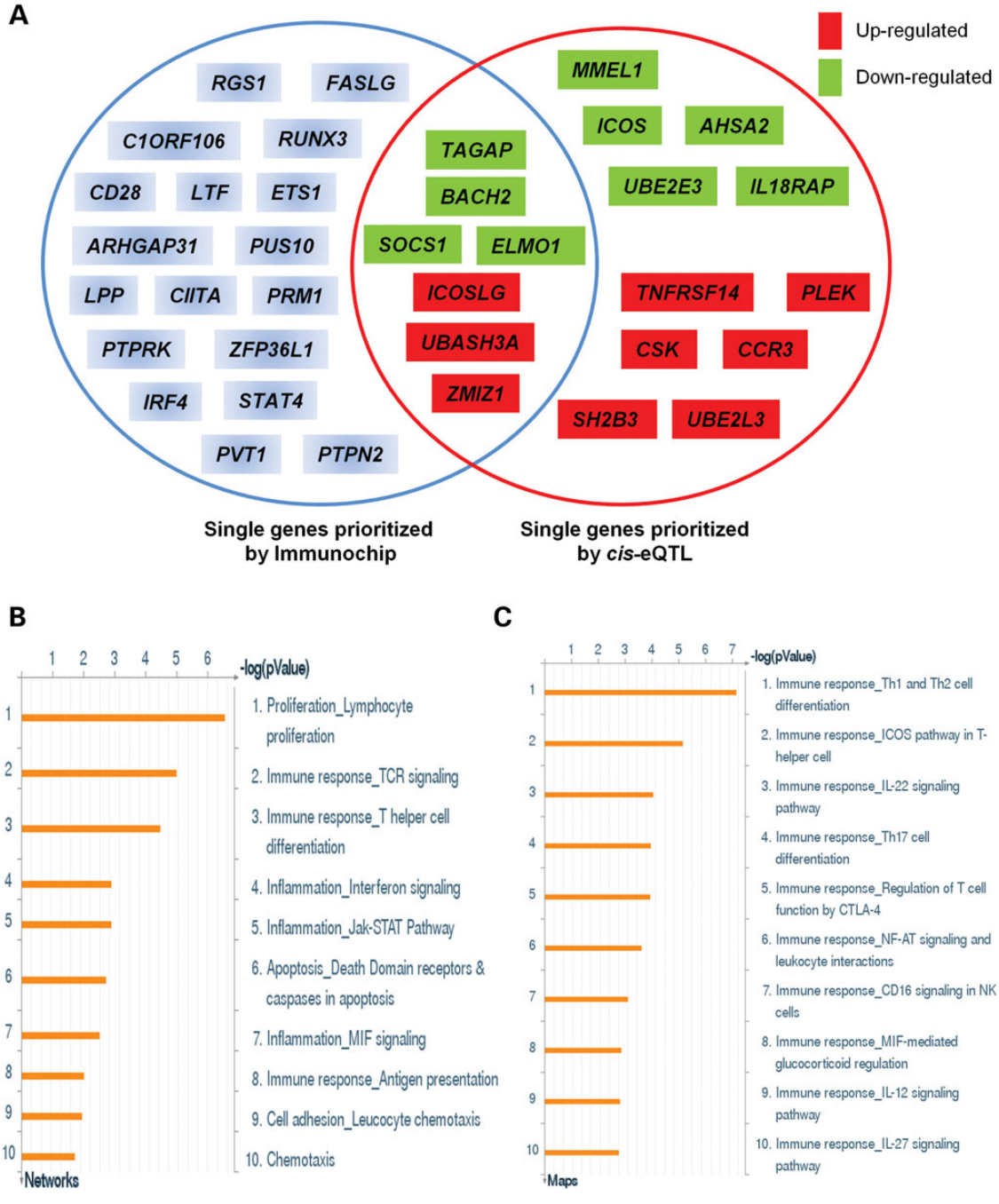

Figure 3. Pathway and network analysis on functional genes identified by Immunochip and *cis*-eQTL analyses. (**A**) The Venn diagram shows 18 single genes identified by Immunochip fine mapping alone, the 11 genes identified by eQTL analysis alone and the 7 genes identified by both approaches. These 36 genes were used as seed genes to perform network analysis to prioritize genes from other CD-associated loci. Enrichment analysis using the 49 prioritized CD genes revealed enrichment of CD genes for (**B**) the lymphocyte proliferation network, and (**C**) the T-helper cell differentiation pathways.

associated ($P < 0.05$; Table 1). We next identified the best eQTL SNP for each of these 18 genes and performed conditional analysis to test whether the effect of a CD SNP on the gene is a primary effect (i.e. both eQTL SNPs and CD SNPs are in LD with $r^2 \geq 0.8$ and/or $D' \geq 0.9$) or a secondary effect (i.e. both SNPs are not in LD but show association with gene expression levels after regressing out the effect of best eQTL SNP). We found that 13 SNP–gene pairs showed a primary effect (the CD SNPs are real eQTL SNPs), whereas 7 SNP–gene pairs showed a secondary effect (Table 1 and Supplementary Material, Table S3). Of the 18 eQTL found genes, 7 confirmed the Immunochip fine-mapping results and together the Immunochip analysis and the eQTL mapping pinpointed 36 genes (Fig. 3).

Subsequently, we performed the network analysis to identify disease-predisposing genes in the 10 loci that could not be fine-mapped by Immunochip and/or *cis*-eQTL analysis. We used the

Table 1. Results of cis-eQTL analysis in two cohorts of 1240 and 229 PBMC samples to prioritize CD susceptibility genes

Chr	GWS CD SNP	Proxy SNP for eQTL	Risk allele	Effect	*P* HT12v3[a]	*P* H8v2[b]	eQTL gene
1	rs4445406	rs3890745	C	Down	4.68×10^{-15}	7.03×10^{-6}	*MMEL1*
			C	Up	5.08×10^{-9}	5.69×10^{-2}	*TNFRSF14*
2	rs13003464		G	Down	4.38×10^{-10}	3.04×10^{-3}	*AHSA2*[c]
2	rs10167650		G	Up	8.58×10^{-23}	2.95×10^{-7}	*PLEK*
2	rs990171		A	Down	1.51×10^{-137}	3.95×10^{-21}	*IL18RAP*
2	rs1018326		C	Down	4.57×10^{-14}	0.21	*UBE2E3*[c]
2	rs34037980	rs10497873	T	Down	3.47×10^{-2}	2.05×10^{-2}	*ICOS*
3	rs6441961		T	Up	1.51×10^{-18}	1.10×10^{-5}	*CCR3*[c]
3	rs7616215		C	Down	5.09×10^{-96}	2.47×10^{-22}	*CCR3*
3	rs60215663	rs13070740	A	Down	2.44×10^{-4}	NA	*CCR3*
6	rs7753008	rs6454802	T	Down	7.40×10^{-3}	0.335	*BACH2*[c]
6	rs1107943	rs11759145	T	Down	4.12×10^{-8}	1.39×10^{-3}	*TAGAP*[c]
7	rs79758729	rs11984075	G	Down	1.12×10^{-7}	1.41×10^{-2}	*ELMO1*
10	rs1250552		G	Up	1.10×10^{-2}	0.12	*ZMIZ1*
12	rs3184504		T	Up	4.38×10^{-6}	5.12×10^{-2}	*SH2B3*
15	rs1378938		T	Up	6.91×10^{-33}	7.02×10^{-8}	*CSK*
16	rs243323		G	Down	4.82×10^{-3}	1.89×10^{-2}	*SOCS1*[c]
21	rs1893592		C	Up	7.08×10^{-44}	NA	*UBASH3A*
21	rs58911644		G	Up	2.01×10^{-2}	1.41×10^{-2}	*ICOSLG*[c]
22	rs4821124		C	Up	3.21×10^{-81}	4.77×10^{-23}	*UBE2L3*

Chr, chromosome; GWS, genome-wide significant; Effect, the risk allele is associated with either increased levels (Up) or decreased levels (Down) of a gene's expression.
[a]1240 PBMC samples assayed on HT12v3 microarray.
[b]229 PBMC samples assayed on H8v2 microarray.
[c]Identified as a secondary effect on these genes by CD SNPs. See also Supplementary Material, Table S3.

36 genes pinpointed by Immunochip analysis and eQTL mapping as seed genes (Fig. 3A). We only considered genes to predispose to disease when they were predicted by at least two out of the five pathway analysis tools (GRAIL, STRING, DAPPLE, GATHER and DAVID). This network analysis approach led to prioritization of seven more genes in seven loci (Table 2). In the remaining three loci (*FASL-TNFSF18*, *IL2-IL21* and *TREH-DDX6*), single candidate genes could not be prioritized by our analyses.

In summary, our final list of potential CD susceptibility genes contained 49 genes in 39 loci. Of the 49 genes, 36 were prioritized by Immunochip and/or eQTL analysis, 7 were prioritized by network analysis and the 6 remaining genes were derived from three loci in each of which there were 2 remaining genes (Supplementary Material, Table S4).

Pathway enrichment analysis predicts CD genes involved in lymphocyte activation

To investigate whether the 49 prioritized CD genes converge on specific biological pathways, we performed pathway analysis using MetaCore GeneGo. This implicated a subset of CD susceptibility genes (*IL2/IL21*, *CD28*, *STAT4*, *TNFRSF14*, *ICOSL*, *PRKCQ*, *IL12* and *ICOS*) in lymphocyte activation (Fig. 3B and Supplementary Material, Table S5; $P = 3.03 \times 10^{-7}$) and, more specifically, in Th1 and Th2 cell differentiation (Fig. 3C and Supplementary Material, Table S5; $P = 7.73 \times 10^{-8}$). We also observed an enrichment of CD genes for the interferon signaling network (*CIITA*, *IL18RAP*, *SOCS1*, *FASL*; $P = 1.09 \times 10^{-4}$) and for Th17 cell differentiation pathway (*CD28*, *IRF4*, *IL21*; $P = 1.24 \times 10^{-5}$). These results highlight the many CD genes that affect T-cell development and differentiation and that disregulated Th1, Th2 and Th17 pathways may be involved in CD etiology.

RNA sequencing of immune cells reveals cell type-specific expression of CD genes

To identify CD-relevant cell types, we investigated the expression profiles of the 49 prioritized CD genes (Fig. 4A) in 7 different immune cell types. Unsupervised hierarchical clustering showed 15/49 genes to be expressed in all seven different cell types (Supplementary Material, Fig. S2 and Table S6). K-means analysis identified clusters with a relatively higher, but not restricted, expression pattern in T cells (*CD28*, *ICOS*, *UBASH3A*, *RGS1*, *CTLA4*, *CCR4*), B cells (*BACH2*, *ICOSLG*, *IRF4*, *ARHGAP31*, *POU2AF1*) or NK/T cells (*PRKCQ*, *FASLG*, *PUS10*, *STAT4*, *SOCS1*, *IL18RAP*). These results not only confirm the presumed role of T cells in CD pathogenesis, but also support the postulated importance of B cells in CD (26), which had already been substantiated by our observation of a significant enrichment of CD SNPs within B-cell-specific enhancers (reported above).

Differential expression of CD genes during Th1, Th2 and Th17 cell differentiation

Pathway analysis indicated that multiple CD genes are involved in regulating the differentiation of T-helper cells (Fig. 3C). We therefore tested whether CD susceptibility genes are differentially expressed during Th1, Th2 (Supplementary Material, Fig. S3) and Th17 cell lineage differentiation (Supplementary Material, Fig. S4). For this, we performed unsupervised clustering of 39/49 prioritized CD gene expression levels using

Table 2. Network analysis results to prioritize CD susceptibility genes using five pathway analysis programs

Chr	GWS SNP rs ID	Query genes	GRAIL (*P*)	STRING (number of interactions)	DAPPLE (*P*)	GATHER (Bayes factor)	DAVID (*P*)	Candidates
1	rs859637	*FASL, TNFSF18*	*FASLG* (7.58×10^{-6}), *TNFSF18* (1.40×10^{-7})	*FASLG* (8)	NA	*FASL* and *TNFSF18* (13)	*FAS* and *TNFSF18* (3.60×10^{-4})	*FASL* and *TNFSF18*
2	rs1980422	*CD28, CTLA4*	NA	*CTLA4* (7)	*CTLA4* (1.90×10^{-3})	NA	NA	*CTLA4*
3	rs4678523	*CCR4, GLB1*	*CCR4* (1.07×10^{-7})	*CCR4* (1)	NA	*CCR4* (14)	*CCR4* (9.0×10^{-3})	*CCR4*
3	rs76830965	*SCHIP1, IL12A*	*IL12A* (2.09×10^{-9})	*IL12A* (9)	*IL12A* (1.09×10^{-3})	NA	*IL12A* (9.7×10^{-3})	*IL12A*
3	rs1353248	*SCHIP1, IL12A*	*IL12A* (2.09×10^{-9})	*IL12A* (9)	*IL12A* (1.09×10^{-3})	NA	*IL12A* (9.7×10^{-3})	*IL12A*
3	rs2561288	*SCHIP1, IL12A*	*IL12A* (2.09×10^{-9})	*IL12A* (9)	*IL12A* (1.09×10^{-3})	NA	*IL12A* (9.7×10^{-3})	*IL12A*
4	rs13132308	*KIAA1109, ADAD1, IL2, IL21*	*IL21* (6.16×10^{-10}) *IL2* (1.47×10^{-9})	*IL2* (10) *IL21* (9)	*IL21* and *IL2* (1.08×10^{-3})	*IL2* (14)	*IL21* and *IL2* (3.8×10^{-2})	*IL21* and *IL2*
4	rs62323881	*KIAA1109, ADAD1, IL2, IL21*	*IL21* (6.16×10^{-10}) *IL2* (1.47×10^{-9})	*IL2* (10) *IL21* (9)	*IL21* and *IL2* (1.08×10^{-3})	*IL2* (14)	*IL21* and *IL2* (3.8×10^{-2})	*IL21* and *IL2*
6	rs17264332	*OLIG3, TNFAIP3*	*TNFAIP3* (4.75×10^{-5})	*TNFAIP3* (6)	*TNFAIP3* (0.08)	NA	NA	*TNFAIP3*
6	rs77027760	*OLIG3, TNFAIP3*	*TNFAIP3* (4.75×10^{-5})	*TNFAIP3* (6)	*TNFAIP3* (0.08)	NA	NA	*TNFAIP3*
10	rs2387397	*PFKFB3, PRKCQ*	*PRKCQ* (6.19×10^{-5})	*PRKCQ* (6)	*PRKCQ* (0.14)	*PRKCQ* (5)	NA	*PRKCQ*
11	rs7104791	*POU2AF1, C11orf93*	*POU2AF1* (0.003)	*POU2AF1* (1)	NA	NA	*POU2AF1* (9.3×10^{-3})	*POU2AF1*
11	rs10892258	*TREH, DDX6*	NS	NS	NS	NS	NS	
16	rs80073729	*PRM1, PRM2*	NS	NS	NS	NS	NS	
X	rs13397	*HCFC1, TMEM187, IRAK1*	*IRAK1* (8.0×10^{-6})	*IRAK1* (1)	NA	NA	NA	*IRAK1*

Chr, chromosome; GWS, genome-wide significant; NA, not applicable; NS not significant.

previously published data (27–29). Our results indicated a specific time window in which CD genes are expressed during differentiation of T-cell lineages. For example, *IL18RAP*, *UBE2L3*, *FASLG*, *SH2B3*, *POU2AF1* and *IL2* are strongly expressed only after 1 h of Th2 cell differentiation and the expression is switched off after 24–48 h. Similarly, *STAT4*, *ZFP36L1*, *TAGAP*, *PLEK* and *CCR4* are expressed at low levels in the early stage of Th1 cell differentiation, whereas they were abundantly expressed during Th2 cell differentiation. During Th17 cell differentiation, the expression of *CD28*, *ETS1*, *UBASH3A*, *STAT4*, *TNFAIP3* and *UBE2L3* is restricted to an early stage (the first 6 h).

A subset of CD genes are predicted to be involved in intestinal barrier function

A subset of 15/49 CD genes could not be categorized due to a lack of published functional data (Supplementary Material, Table S7). Thus, we applied an alternative approach in an attempt to predict the function of these 15 genes. We interrogated the GeneNetwork database (www.genenetwork.nl/genenetwork, manuscript submitted) to predict their function based on co-expression data extracted from ∼80 000 Affymetrix microarray experiments. We were able to predict a significant function (Supplementary Material, Table S7) except for two genes (*PUS10* and *PVT1*). The predictions indicated that 4/13 genes (*C1ORF106*, *ARHGAP31*, *LPP* and *PTPRK*) act in the cell–cell adhesion process in the intestine.

Next, we tested whether these predictions for the four genes could be replicated using another prediction tool, GEMMA. We extracted their co-expressed genes from GEMMA and again performed a gene network enrichment analysis using MetaCore GeneGo. This showed enrichment with genes involved in cell adhesion and cell junction networks ($P = 3.67 \times 10^{-4}$), cardiac development by BMP and TGF beta signaling ($P = 1.69 \times 10^{-3}$), skeletal muscle development ($P = 3.36 \times 10^{-11}$) and cadherin-mediated cell adhesion ($P = 3.74 \times 10^{-3}$) (Fig. 4B). Interestingly, the four genes seem to play a role in the actin-cytoskeleton rearrangement and cell–cell adhesion pathways, which are crucial for maintaining the intestinal epithelial barrier function (20,21).

We then tested whether the expression levels of the four genes are differentially regulated in CD-affected intestinal tissue using microarray data from 24 duodenal biopsies (12 CD patients with villous atrophy and 12 healthy controls). We found down-regulation of the *LPP* ($P = 0.0004$), *C1orf106* ($P = 0.0005$) and *PTPRK* genes, and up-regulation of *ARHGAP31* expression in biopsies from CD cases compared with controls (Supplementary Material, Fig. S5). These findings suggest that the four genes are potential candidates involved in barrier function, which can be tested using functional studies.

Increased expression of IFN-γ may be due to altered expression of CD susceptibility genes

Pathway analysis on prioritized CD genes predicted a role for the interferon signaling network in CD (Fig. 3B). The level of IFN-γ expression is elevated in CD patients and correlates with the degree of damage to the intestinal mucosal layer (Fig. 5A) (22). We investigated whether the expression of IFNG mRNA

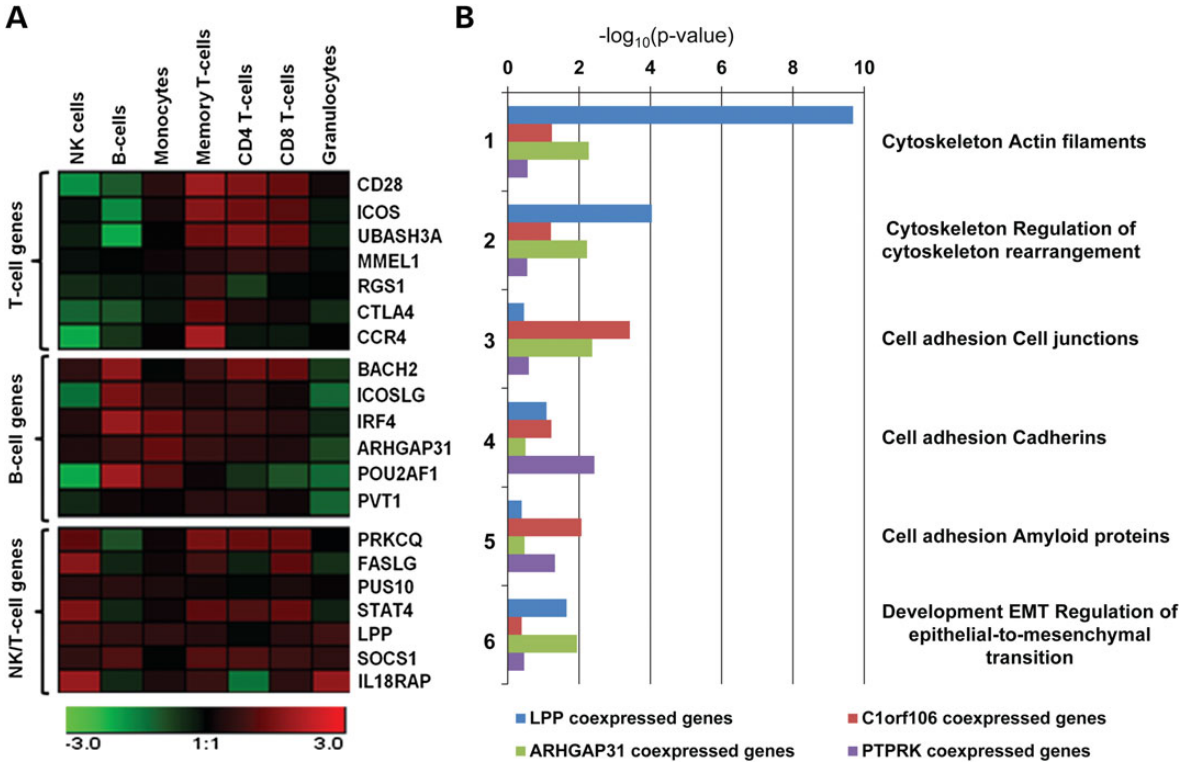

Figure 4. (**A**) K-means clustering analysis on expression of CD genes in seven cell types identified as cell-specific CD genes. (**B**) Pathway enrichment analysis on a subset of non-immune CD genes revealed a novel network involved in cell–cell adhesion and maintaining the intestinal barrier function.

correlates with expression levels of CD susceptibility genes. Using the GeneNetwork database (www.genenetwork.nl/genenetwork), we extracted all the genes significantly co-expressed with IFN-γ and built an IFN-γ co-expression network. This showed that more than 30% (15/49) of the CD genes are significantly co-expressed with IFN-γ (Fig. 5B). Because the Mouse Genome Informatics phenotype data are integrated in the GeneNetwork database, we were able to test whether mutations in these 15 IFNG co-expressed CD genes displayed any common phenotype. The assumption being that if these 15 genes are indeed connected to IFNG expression, their loss must have an impact on IFN-γ levels. In other words, we tested whether single gene knock-out mouse models of all 15 CD genes (connected to IFN-γ) would display phenotypes along with other phenotypes. We found a significant enrichment of genes in which mutations may cause differential levels of IFN-γ secretion (Fig. 5C). Additionally, we investigated whether the 15 IFNG co-expressed CD genes can be induced by IFN-γ stimulation using data generated by Fairfax *et al.* (30), who performed eQTL mapping using gene-expression data generated from monocytes that were stimulated with either IFN-γ or LPS. We extracted the expression data from their study and tested whether the 15 genes were differentially regulated upon IFN-γ stimulation. We traced the expression data for 11/15 genes and found 7 to be significantly regulated by IFN-γ stimulation (Supplementary Material, Table S8), thus supporting the hypothesis of a connection between CD genes and the IFNG pathway.

DISCUSSION

By integrating publically available data and systems biology methods for analysis in this study, we were able to zoom in on causal CD SNPs and the genes and pathways affected in CD pathology. We had four main results: (i) the SNP prioritization data confirm the notion that autoimmune disease-associated SNPs are regulatory in nature (validated by showing that these SNPs are enriched within immune cell-type-specific enhancers). We were further able to narrow down our candidate gene list to 49 genes (validated by testing their cell-type-specific expression in seven different immune cell types and in Th cell differentiation). (ii) Our analysis confirmed the classical view that T cells are key players in CD, and we report the novel link between CD-associated SNPs and the role of B cells in this disease (validated by showing that the CD SNPs are significantly enriched in B-cell-specific enhancers). (iii) Moreover, our results suggest that four, thus far poorly characterized, CD genes play a role in the intestinal barrier function, which is known to be impaired in CD (validated by showing that these genes are differentially expressed in biopsies from CD cases compared with healthy controls). Although defective barrier function is a known feature of CD, our results implicate four CD genes as being causal in this phenotype. (iv) Finally, our analysis discovered a genetic network involved in IFN-γ signaling and linked to CD (validated by showing that the IFNG co-regulated CD genes are differentially regulated in response to IFN-γ stimulation).

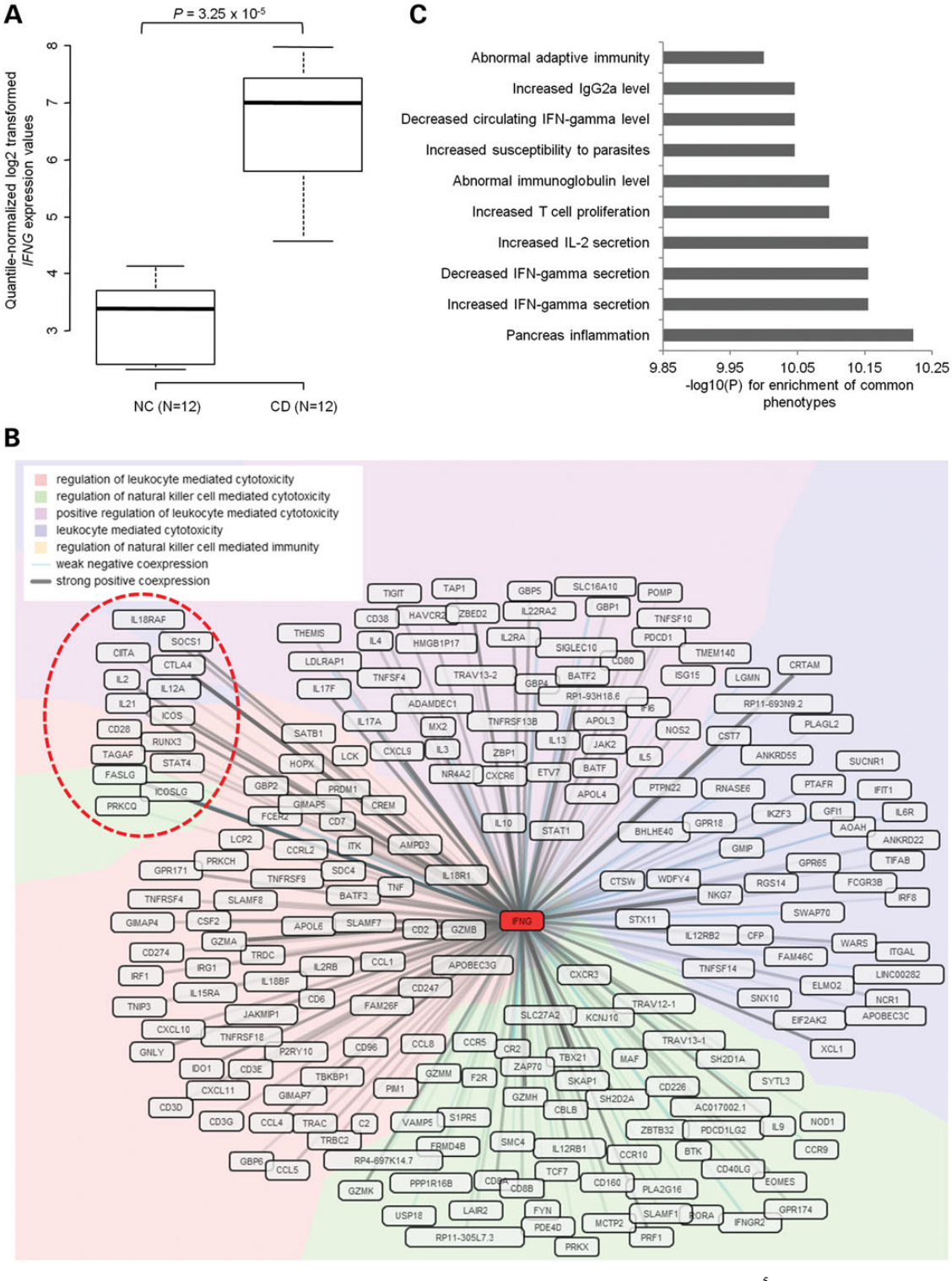

Figure 5. CD susceptibility genes regulate the expression levels of IFNG. (**A**) IFNG levels are significantly ($P = 3.25 \times 10^{-5}$) up-regulated in CD biopsies with Marsh III condition (CD; $N = 12$) compared with non-CD biopsies (NC; $N = 12$). (**B**) An IFNG co-expression gene network revealed that 15 CD susceptibility genes are co-expressed with IFNG (shown in red circle). Co-expressed genes are colored based on their GO biological process. (**C**) Common phenotype enrichment analysis in the mouse, using the Mouse Genome Informatics phenotype data for the 15 CD genes that are co-expressed with IFNG, indicated altered levels of IFNG as the most common phenotype.

The clearest proof for prioritizing disease genes is undoubtedly the identification of an eQTL SNP. The eQTL results not only help us map susceptibility genes precisely, but also to discover the direction of the effect of the CD risk alleles on gene expression. For instance, a study in mice showed that the *BACH2* gene is required to prevent the development of autoimmunity due to its

role in the development of regulatory T cells (31). Concordant with these results, we observed the association of a CD risk allele with *BACH2* down-regulation. Therefore, by looking at the direction of the effect of the eQTL on gene expression, we can predict whether the susceptibility genes are involved in the pro-inflammatory or anti-inflammatory pathways. Intriguingly, eQTL mapping identified a different disease gene than those seen in the finely mapped LD region at two loci (*AHSA2* instead of *PUS10*, and *CCR3* instead of *LTF*). These findings suggest that the susceptibility variants may affect the expression of genes located further away from the associated LD block and one must therefore be careful when pinpointing susceptibility genes in GWA studies. Furthermore, conditional eQTL analysis indicated that some of the eQTLs could be mere overlap with the effect of the best eQTL SNP (e.g. *SOCS1*) and may not be truly linked to a disease-associated SNP. Therefore, future eQTL studies using RNA-sequencing data from many different cell types and complete genotype data are required to pinpoint the true functional genes in all disease-associated loci.

It should be noted that for the other 50% of the CD SNPs, we observed no significant association between expression levels of transcripts and SNP genotypes in PBMC samples containing multiple immune cell types. Previous eQTL analyses have suggested that a considerable number of disease-associated SNPs can be tissue-specific eQTLs (32,33) as a consequence of altered gene expression by *cis*-regulatory elements in a tissue- and/or timing-specific manner. Accordingly, cell-type-specific enhancer enrichment analyses showed that CD SNPs are located not only in T-cell regulatory regions, but also within the elements that regulate different T-cell lineages, B cells and monocytes. Since B cells and monocytes constitute $<20\%$ of all the cells in a PBMC fraction, we may not be able to detect B-cell- or monocyte-specific eQTLs when mapping eQTLs in total PBMCs. Indeed, recent eQTL studies performed using gene-expression data from innate-antigen stimulated monocytes (30) or dendritic cells (34) have identified many autoimmune disease-associated SNPs as cell- and stimulation-specific eQTLs. Future studies should therefore investigate the genotype-expression correlation in cell-type-specific and stimulation-specific data sets to reveal a disease mechanism in detail.

Since, these CD functional SNPs mostly regulate gene expression by influencing the stability of binding between enhancer DNA and TFs, they are expected to cause small shifts in the expression levels of susceptibility genes. However, the cumulative effect of these small variations may have a large impact on particular biological pathways if the SNPs affect several genes involved in a common pathway. Indeed, systematic prioritization of the genes and subsequent pathway analysis enabled us to show that the majority of CD susceptibility genes converge to regulate Th cell differentiation and IFN-γ signaling pathways. A proper balance between the different T-cell lineages is crucial for maintaining a healthy adaptive immune system. It has been suggested there is an imbalance between the Th1, Th2 and Th17 lineages in CD (35,36), so the differential expression of a subset of CD susceptibility genes during Th cell differentiation may be important in maintaining balance (37). In CD, it has also been shown that dietary gluten affects Th1 and Th2 adaptive immune responses (35). The Th1 response is characterized by a high secretion of IFN-γ, which results in T-cell activation and intestinal tissue damage (22). Although the IFN-γ locus is not genetically implicated in CD (22), we have now found a strong interdependency between IFN-γ levels and the expression levels of CD genes. It is tempting to speculate that SNPs in the CD-associated genes that are connected to the IFNG network may predispose individuals to a specific, more severe form of CD (e.g. Marsh III) than other variants. This type of information might help in refining risk prediction models so that they cannot only predict an individual's risk for CD development, but also the likely progression and severity of disease.

Defects in barrier function are strongly associated with CD (45). Although several factors, such as infection, genetic predisposition and gluten intake, may increase the chances of an abnormal intestinal barrier in CD, the role of elevated levels of IFN-γ has been well established (22,38). However, it is still difficult to say whether the intestinal barrier defect is the cause or a consequence of CD. But now, for the first time, our pathway analyses have indicated that four prioritized CD genes play a role in regulating the intestinal membrane barrier function. In addition, by establishing a significant connection between CD susceptibility genes and IFNG, our study suggests that the abnormal barrier function could be a basic defect in CD. Enhancing the intestinal barrier function could offer a new lead for therapeutic intervention, while a good understanding of the functional role of CD susceptibility genes in maintaining the intestinal barrier function could provide other, alternative, therapeutic options.

In summary, our study has identified a subset of CD genes relevant for maintaining the intestinal permeability. We have also revealed an intricate transcriptional connectivity between CD susceptibility genes and IFN-γ.

MATERIALS AND METHODS

Functional annotation and prioritization of CD SNPs

The Immunochip analysis identified 57 independent CD susceptibility SNPs in 39 known CD loci. The proxies ($r^2 \geq 0.9$, 1000 Genome project, CEU population as reference) for these 57 SNPs were extracted using HaploReG v2 tool (http://compbio.mit.edu/HaploReg) (25). The potential functional consequence of each SNP was predicted using RegulomeDB (39). Although both RegulomeDB and HaploReg use the same regulatory information from the ENCODE project, RegulomeDB allows ranking of SNPs based on the amount of regulatory information which a SNP intersects with. We adopted the modified RegulomeDB scheme suggested by Schaub *et al.* (23) to prioritize the CD functional SNPs. The annotation methods are described in the Supplementary Note.

Cell-type-specific enhancer enrichment analysis for CD SNPs

We intersected 677 SNPs (57 CD SNPs + 620 proxies) for cell-type-specific enhancers using HaploReg v2 database (25). The fold enrichment of cell-type-specific enhancers was calculated as described previously (24). The uncorrected binomial *P*-value was subjected to multiple testing correction (0.05/677 SNPs) and enrichment was considered significant if the *P*-value passed a Bonferroni-corrected *P*-value $\leq 7.38 \times 10^{-5}$. To test whether this enrichment was higher than expected, we extracted 100 sets of 670 SNPs (random sets of SNPs from the

same region containing the GWAS SNPs) that were matched for allele frequency distributions of CD SNPs and no or low LD with CD SNPs ($r^2 < 0.4$). The fold enrichment of enhancers for each of the 100 sets of random SNPs was extracted from HaploReg, which allowed us to calculate the significance of enrichment as determined by fitting a normal distribution on the 100 log-transformed fold enrichment values.

Prioritization of CD susceptibility genes by eQTL mapping

As a discovery set, 1240 peripheral blood samples were investigated and the results were replicated in 229 samples, all isolated from unrelated, healthy Dutch controls. The *cis*-eQTL method is described in the Supplementary Note. Since eQTL effects can be driven by LD, rather than direct effect of the CD SNPs on gene expression, we then performed conditional analysis on SNP–gene pairs to identify whether the eQTL effect of a CD SNP on a gene is a primary effect [i.e. both eQTL SNP and CD SNP are in LD ($r^2 \geq 0.8$ and/or $D' \geq 0.9$)] or a secondary effect (i.e. both SNPs are not in LD but the CD SNP shows association with gene expression levels after regressing out the effect of the best eQTL SNP of that gene). Using the CD SNP eQTL, we first determined the top eQTL SNP for the associated gene within 250 kb. Subsequently, using linear regression, we adjusted the gene expression data for the effect of this top eQTL SNP. Finally, we assessed whether the CD SNPs showed significant eQTL effects on the residual gene expression levels, in order to determine whether the CD SNP eQTL effect is independent of the top eQTL effect.

Prioritization of CD susceptibility genes using network analysis

We used the publically available GRAIL (40), DAPPLE (41), DAVID (42,43), STRING (44) and GATHER (45) tools for pathway analysis. We hypothesized that consistent prediction of the same gene by two or more different tools points to a true functional gene. We used the default settings for these tools to predict the most likely functional gene. The websites where these tools and their methodologies can be found are provided in Supplementary Material, Table S9. The single genes implicated as susceptibility genes by either Immunochip or eQTL mapping were used as seed genes to construct biological networks associated with CD. Using each tool separately, we examined whether the query genes (multiple genes in one locus) connected significantly with the networks interconnecting the seed genes. The significance of network connectivity for each seed gene is indicated by either the *P*-value or the Bayes factor generated by the algorithms. If two or more different tools predicted that the same query gene would form networks with the seed genes, then the query gene was considered to be the functional gene in that locus.

Pathway or network enrichment analysis on all prioritized susceptibility genes

The commercially available MetaCore-GeneGO pathway analysis tool (http://www.genego.com/metacore.php) was queried to group functionally related gene networks. The significance of these analyses was evaluated based on the size of the intersection between our gene list and the set of genes/proteins corresponding to a network module or pathway. MetaCore-GeneGO indicates statistical significance of the enriched terms by a *P*-value (Fisher's exact test, corrected for multiple testing by false discovery rate).

Collection of granulocytes and peripheral blood mononuclear cell fractions

Participants were enrolled after giving informed consent, following an Institutional review board protocol approved by the University Medical Centre Groningen (Groningen, the Netherlands). The isolation and sorting methods for PBMCs are described in the Supplementary Note.

RNA isolation and library preparation

RNA was extracted using the Ambion mirVana miRNA isolation kit (Life Technologies, Bleiswijk, The Netherlands) according to the manufacturer's instructions. RNA quantity and quality were determined using the Nanodrop 1000 spectrometer (Thermo Fisher Scientific, Landsmeer, the Netherlands) and Expirion High-sensitivity RNA analysis kit (Bio-Rad, Waltham, MA, USA), respectively. RNAseq libraries were prepared from 1 μg RNA of each cell population using Illumina's TruSeq RNA kit (San Diego, CA, USA) according to the manufacturer's instructions and these libraries were subsequently sequenced on a HiSeq 2000 sequencer (Illumina).

Analysis of RNAseq reads

The sequencing reads were mapped to human reference genome NCBI build 37, using STAR v2.1.3 (46), allowing for two mismatches and retaining only uniquely mapping reads. The method is described in the Supplementary Note.

Data sets used to zoom in on specific T helper cell lineages

Gene expression microarray data measured from human T helper (Th) cells differentiating toward Th1, Th2 or Th17 lineages and from activated Th cells (Th0; controls) were taken from previous studies (27–29). The data analyses are described in the Supplementary Note.

Prediction of function of non-immune genes by co-expression and pathway analysis

We interrogated the GeneNetwork co-expression database (www.genenetwork.nl/genenetwork) that has been developed in our lab (13) (*manuscript in preparation*), the Gemma database (www.chibi.ubc.ca/Gemma/home.html), and MetaCore-GeneGO's pathway analysis tool (http://www.genego.com/metacore) to predict pathways in which prioritized CD genes are involved. The method is described in the Supplementary Note.

Analysis of expression data from biopsies

Biopsy sampling, RNA isolation and details of microarray hybridization have been previously described (47,48). Intestinal biopsies from 12 CD patients with a histological classification

of Marsh III (MIII), and duodenum biopsies from 13 healthy individuals were investigated according to the UEGW criteria. The normalized expression values were stratified according to the phenotypes (CD cases versus controls) and the significant difference was tested using the Wilcoxon rank test (implemented in R).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGMENTS

We thank Jackie Senior for editing the final version of the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This research was supported by the Netherlands Organization for Scientific Research (NWO-VENI grant 916.10.135 to L.F.), the European Research Council Advanced Grant, ERC-671274 to C.W., a fellowship to V.K. from the Dutch Digestive Diseases Foundation (MLDS WO11-30 to CW), the Academy of Finland (Centre of Excellence in Molecular Systems Immunology and Physiology Research, 2012–2017 to R.L. and H.L.), the European Commission Seventh Framework (EC-FP7-SYBILLA-201106 to R.L. and H.L.) and the Sigrid Juselius Foundation to C.W., R.L. and H.L.

REFERENCES

1. Dicke, W.K., Weijers, H.A. and Van De Kamer, J.H. (1953) Coeliac disease. II. The presence in wheat of a factor having a deleterious effect in cases of coeliac disease. *Acta Paediatr.*, **42**, 34–42.
2. Garcia-Manzanares, A. and Lucendo, A.J. (2011) Nutritional and dietary aspects of celiac disease. *Nutr. Clin. Pract.*, **26**, 163–173.
3. Bakker, S.F., Pouwer, F., Tushuizen, M.E., Hoogma, R.P., Mulder, C.J. and Simsek, S. (2013) Compromised quality of life in patients with both type 1 diabetes mellitus and coeliac disease. *Diabet. Med.*, **30**, 835–839.
4. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
5. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K. *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.*, **39**, 827–829.
6. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D. *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.*, **40**, 395–402.
7. Dubois, P.C., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A., Adany, R., Aromaa, A. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.
8. Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G. *et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193–1201.
9. Romanos, J., Rosen, A., Kumar, V., Trynka, G., Franke, L., Szperl, A., Gutierrez-Achury, J., van Diemen, C.C., Kanninga, R., Jankipersadsing, S.A. *et al.* (2014) Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut*, **63**, 415–422.
10. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, **3**, 299–309.
11. Ricano-Ponce, I. and Wijmenga, C. (2013) Mapping of immune-mediated disease genes. *Annu. Rev. Genomics Hum. Genet.*, **14**, 325–353.
12. Kumar, V., Wijmenga, C. and Withoff, S. (2012) From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. *Semin. Immunopathol.*, **34**, 567–580.
13. Kumar, V., Westra, H.J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Vosa, U. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.*, **9**, e1003201.
14. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
15. Emily, M., Mailund, T., Hein, J., Schauser, L. and Schierup, M.H. (2009) Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.*, **17**, 1231–1240.
16. Elbers, C.C., van Eijk, K.R., Franke, L., Mulder, F., van der Schouw, Y.T., Wijmenga, C. and Onland-Moret, N.C. (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.*, **33**, 419–431.
17. Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., Polman, C.H. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078–2090.
18. Askland, K., Read, C. and Moore, J. (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.*, **125**, 63–79.
19. Smeekens, S.P., Ng, A., Kumar, V., Johnson, M.D., Plantinga, T.S., van Diemen, C., Arts, P., Verwiel, E.T., Gresnigt, M.S., Fransen, K. *et al.* (2013) Functional genomics identifies type I interferon pathway as central for host defense against *Candida albicans*. *Nat. Commun.*, **4**, 1342.
20. Clemente, M.G., De Virgiliis, S., Kang, J.S., Macatagney, R., Musu, M.P., Di Pierro, M.R., Drago, S., Congia, M. and Fasano, A. (2003) Early effects of gliadin on enterocyte intracellular signalling involved in intestinal barrier function. *Gut*, **52**, 218–223.
21. Monsuur, A.J., de Bakker, P.I., Alizadeh, B.Z., Zhernakova, A., Bevova, M.R., Strengman, E., Franke, L., van't Slot, R., van Belzen, M.J., Lavrijsen, I.C. *et al.* (2005) Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. *Nat. Genet.*, **37**, 1341–1344.
22. Wapenaar, M.C., van Belzen, M.J., Fransen, J.H., Sarasqueta, A.F., Houwen, R.H., Meijer, J.W., Mulder, C.J. and Wijmenga, C. (2004) The interferon gamma gene in celiac disease: augmented expression correlates with tissue damage but no evidence for genetic susceptibility. *J. Autoimmun.*, **23**, 183–190.
23. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.
24. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
25. Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.
26. Di Niro, R., Mesin, L., Zheng, N.Y., Stamnaes, J., Morrissey, M., Lee, J.H., Huang, M., Iversen, R., du Pre, M.F., Qiao, S.W. *et al.* (2012) High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nat. Med.*, **18**, 441–445.
27. Aijo, T., Edelman, S.M., Lonnberg, T., Larjo, A., Kallionpaa, H., Tuomela, S., Engstrom, E., Lahesmaa, R. and Lahdesmaki, H. (2012) An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human T helper cell differentiation. *BMC Genomics*, **13**, 572.
28. Elo, L.L., Jarvenpaa, H., Tuomela, S., Raghav, S., Ahlfors, H., Laurila, K., Gupta, B., Lund, R.J., Tahvanainen, J., Hawkins, R.D. *et al.* (2010) Genome-wide profiling of interleukin-4 and STAT6 transcription factor regulation of human Th2 cell programming. *Immunity*, **32**, 852–862.
29. Tuomela, S., Salo, V., Tripathi, S.K., Chen, Z., Laurila, K., Gupta, B., Aijo, T., Oikari, L., Stockinger, B., Lahdesmaki, H. *et al.* (2012) Identification of

early gene expression changes during human Th17 cell differentiation. *Blood*, **119**, e151–e160.
30. Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C. *et al.* (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, **343**, 1246949.
31. Roychoudhuri, R., Hirahara, K., Mousavi, K., Clever, D., Klebanoff, C.A., Bonelli, M., Sciume, G., Zare, H., Vahedi, G., Dema, B. *et al.* (2013) BACH2 represses effector programs to stabilize T(reg)-mediated immune homeostasis. *Nature*, **498**, 506–510.
32. Fu, J., Wolfs, M.G., Deelen, P., Westra, H.J., Fehrmann, R.S., Te Meerman, G.J., Buurman, W.A., Rensen, S.S., Groen, H.J., Weersma, R.K. *et al.* (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.*, **8**, e1002431.
33. Hernandez, D.G., Nalls, M.A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., Gibbs, J.R., Ryten, M., Arepalli, S., Weale, M.E. *et al.* (2012) Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.*, **47**, 20–28.
34. Lee, J.C., Espeli, M., Anderson, C.A., Linterman, M.A., Pocock, J.M., Williams, N.J., Roberts, R., Viatte, S., Fu, B., Peshu, N. *et al.* (2013) Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell*, **155**, 57–69.
35. Castellanos-Rubio, A., Santin, I., Irastorza, I., Castano, L., Carlos Vitoria, J. and Ramon Bilbao, J. (2009) TH17 (and TH1) signatures of intestinal biopsies of CD patients in response to gliadin. *Autoimmunity*, **42**, 69–73.
36. Meresse, B., Malamut, G. and Cerf-Bensussan, N. (2012) Celiac disease: an immunological jigsaw. *Immunity*, **36**, 907–919.
37. Hawkins, R.D., Larjo, A., Tripathi, S.K., Wagner, U., Luu, Y., Lonnberg, T., Raghav, S.K., Lee, L.K., Lund, R., Ren, B. *et al.* (2013) Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity*, **38**, 1271–1284.
38. Pizzuti, D., Bortolami, M., Mazzon, E., Buda, A., Guariso, G., D'Odorico, A., Chiarelli, S., D'Inca, R., De Lazzari, F. and Martines, D. (2004) Transcriptional downregulation of tight junction protein ZO-1 in active coeliac disease is reversed after a gluten-free diet. *Dig. Liver Dis.*, **36**, 337–341.
39. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797.
40. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., Daly, M.J. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.
41. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C. and Daly, M.J. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**, e1001273.
42. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
43. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
44. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
45. Chang, J.T. and Nevins, J.R. (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics*, **22**, 2926–2933.
46. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
47. Almeida, R., Ricano-Ponce, I., Kumar, V., Deelen, P., Szperl, A., Trynka, G., Gutierrez-Achury, J., Kanterakis, A., Westra, H.J., Franke, L. *et al.* (2014) Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. *Hum. Mol. Genet.*, **23**, 2481–2489.
48. Diosdado, B., Wapenaar, M.C., Franke, L., Duran, K.J., Goerres, M.J., Hadithi, M., Crusius, J.B., Meijer, J.W., Duggan, D.J., Mulder, C.J. *et al.* (2004) A microarray screen for novel candidate genes in coeliac disease pathogenesis. *Gut*, **53**, 944–951.