HHS Public Access

Author manuscript

J Autoimmun. Author manuscript; available in PMC 2017 April 14.

Published in final edited form as:

J Autoimmun. 2016 April ; 68: 62–74. doi:10.1016/j.jaut.2016.01.002.

Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs

Isis Ricaño-Ponce[a], **Daria V. Zhernakova**[a], **Patrick Deelen**[a,b], **Oscar Luo**[c], **Xingwang Li**[c], **Aaron Isaacs**[d], **Juha Karjalainen**[a], **Jennifer Di Tommaso**[a], **Zuzanna Agnieszka Borek**[a], **Maria M. Zorro**[a], **Javier Gutierrez-Achury**[a], **Andre G. Uitterlinden**[d], **Albert Hofman**[d], **Joyce van Meurs**[d], **BIOS consortium Lifelines Cohort Study**[e], **Mihai G. Netea**[f], **Iris H. Jonkers**[a], **Sebo Withoff**[a], **Cornelia M. van Duijn**[d], **Yang Li**[a], **Yijun Ruan**[c,g], **Lude Franke**[a], **Cisca Wijmenga**[a,1], and **Vinod Kumar**[a,*,1]

[a]University of Groningen, University Medical Centre Groningen, Department of Genetics, Groningen, 9700 RB, The Netherlands [b]University of Groningen, University Medical Centre Groningen, Genomics Coordination Centre, Groningen, 9700 RB, The Netherlands [c]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06030, USA [d]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Centre, Rotterdam, 3015 CE, The Netherlands [e]BIOS Consortium, Leiden, 2300 RC, The Netherlands [f]Department of Internal Medicine and Radboud Centre for Infectious Diseases, Radboud University Medical Centre, Nijmegen, 6525 GA, The Netherlands [g]Department of Genetics and Genome Sciences, University of Connecticut Health Centre, 400 Farmington Ave, Farmington, CT 06030, USA

Abstract

Genome-wide association and fine-mapping studies in 14 autoimmune diseases (AID) have implicated more than 250 loci in one or more of these diseases. As more than 90% of AID-associated SNPs are intergenic or intronic, pinpointing the causal genes is challenging. We performed a systematic analysis to link 460 SNPs that are associated with 14 AID to causal genes using transcriptomic data from 629 blood samples. We were able to link 71 (39%) of the AID-SNPs to two or more nearby genes, providing evidence that for part of the AID loci multiple causal genes exist. While 54 of the AID loci are shared by one or more AID, 17% of them do not share candidate causal genes. In addition to finding novel genes such as *ULK3*, we also implicate novel disease mechanisms and pathways like autophagy in celiac disease pathogenesis. Furthermore, 42 of the AID SNPs specifically affected the expression of 53 non-coding RNA genes. To further understand how the non-coding genome contributes to AID, the SNPs were

*Corresponding author. Department of Genetics, University Medical Centre Groningen, PO Box 30001, 9700 RB, Groningen, The Netherlands. v.kumar@umcg.nl (V. Kumar).

[1]These authors jointly directed the study.

Disclosure statement: The authors declare no competing interests.

Appendix A. Supplementary data: Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jaut.2016.01.002.

linked to functional regulatory elements, which suggest a model where AID genes are regulated by network of chromatin looping/non-coding RNAs interactions. The looping model also explains how a causal candidate gene is not necessarily the gene closest to the AID SNP, which was the case in nearly 50% of cases.

Keywords

Long non-coding RNAs; eQTLs; RNA-sequencing; Genome-wide association; Causal genes

1. Introduction

The genetic basis for autoimmune diseases (AID) has been successfully demonstrated by GWAS, which have now firmly associated more than 250 loci to 14 common AID [1]. Some of these loci are also shared by more than one AIDs. Fine-mapping studies have drastically refined the regions of association and revealed that more than 90% of the AID-associated SNPs are found in non-coding regions (summarized by Ricaño-Ponce & Wijmenga) [2]. These studies have further shown that more than 40% of AID-associated SNPs affect the expression levels of nearby protein-coding genes (cis-eQTLs) [2,3] and sometimes those of genes located elsewhere in the genome (*trans*-eQTLs) [4]. These studies indicate that causative genes can be located outside the peak of association and that fine-mapping merely points towards the location of the causal variants. Moreover, eQTL studies have focused on the protein-coding genes while completely ignoring the 65% of annotated human genome transcribed into non-coding RNAs (ncRNAs) [5]. To date, only a few ncRNAs have been implicated in diseases, but this mostly seems to reflect the difficulty of studying them. Although the function of the majority of ncRNAs is unknown, it has become clear that they are important regulators of gene expression. In a proof-of-concept study, we recently showed that disease-associated SNPs can also impact the expression of ncRNAs [6]. By annotating GWAS loci with ncRNA transcripts, we demonstrated that ncRNAs can physically overlap AID-SNPs [2]. These two observations led us to hypothesize that ncRNAs might be crucial regulators in AIDs by affecting the expression levels of protein-coding genes, either in *cis* or in *trans,* thereby providing a link between non-coding AID-SNPs and protein-coding genes.

In this study we performed *cis*-eQTL analysis on AID-SNPs using RNA-sequencing data from peripheral blood mononuclear cells (PBMCs) from 629 healthy individuals. We identified *cis*-eQTL effect on 233 genes, including 53 ncRNAs. We found that the regulation is rather more complex than a single SNP affecting the closest gene, as we observed that half of the SNPs affect more than one gene and that in more than 50% of the loci, the gene closest to the AID-SNP was not the candidate causal gene. Moreover, some of the loci shared by AID do not share causal SNPs or candidate causal genes, further contributing to the complexities of AID genetic architecture.

2. Materials and methods

2.1. Ethical statement and study cohorts

The study procedures were approved by the authorities in all the participating centres and the studies were conducted according to Dutch rules for approval by ethics committee and for gaining informed consent. Participants of the LifeLines Deep population cohort were enrolled after giving informed consent, following an institutional review board protocol approved by the University Medical Centre Groningen (Groningen, the Netherlands). The Rotterdam Study (Rotterdam cohort) was approved by the medical ethics committee according to the Dutch Population Screening Act, Rotterdam Study, as executed by the Netherlands Ministry of Health, Welfare and Sports. Written informed consent was obtained from all the participants. Blood samples for DNA isolation and subsequent genotyping analysis were collected in EDTA Vacutainer® tubes (BD Biosciences, San Jose, CA, USA). Blood samples for RNA isolation and subsequent RNA-seq analysis were collected in PAXgene tubes (PAXgene Blood RNATube, PreAnalytix GmbH, Switzerland, ref 762165). Peripheral blood mononuclear cells (PBMCs) were isolated using 2 ml of whole blood with EDTA in a cell preparation tube (CPT) containing Heparin (BD Vacutainer CPT, ref 362780), according to the manufacturer's instructions.

2.2. SNPs associated to autoimmune diseases

We collected data on 14 different AID phenotypes (Supplemental Table 1) previously genotyped by Immunochip analysis [as of February 2014] in 15 different studies (including two studies on rheumatoid arthritis). In total, we extracted 543 SNPs that showed significant association ($P < 5 \times 10^{-8}$) to any of the 14 AID (referred to as AID-SNPs); 35 SNPs are shared by at least two diseases, yielding 508 unique SNPs. After applying SNP genotype quality control filters, we obtained a final set of 460 different AID-SNPs for *cis*-eQTL mapping (Supplemental Table 2).

2.3. Genotyping and genotype imputation

DNA isolation was performed by the Qiagen robots using Autopure LS kits. Genotyping of DNA from the LifeLines Deep cohort was performed using both the HumanCytoSNP-12 BeadChip and the ImmunoChip platforms (Illumina, San Diego, CA, USA). First, SNP quality control was applied independently for both platforms. SNPs were filtered on MAF above 0.001, a Hardy-Weinberg equivalent P value $>1e^{-4}$ and a call rate of >0.98 using Plink [7]. Genotyping of the Rotterdam samples was performed with the Infinium II HumanHap 550K + 610K Quad Genotyping GenomeStudio® (Illumina). Polymorphisms were genotyped according to the manufacturer's instructions. Quality controls and the results of the genotyping have been published elsewhere [8].

The genotypes from both platforms were merged into one dataset. After merging, SNPs were again filtered on MAF 0.05 and a call rate of 0.98, resulting in a total of 379,885 genotyped SNPs. Next, this data was imputed based on the Genome of the Netherlands (GoNL) reference panel [9–11]. The merged genotypes were pre-phased using SHAPEIT2 [12] and aligned to the GoNL reference panel using Genotype Harmonizer [13] (http://www.molgenis.org/systemsgenetics/) in order to resolve strand issues. Imputation was

performed using IMPUTE2 [14] version 2.3.0 against the GoNL reference panel. We used the MOLGENIS compute imputation pipeline to generate our scripts and monitor the imputation [15].

2.4. RNA isolation and library preparation

RNA from PBMCs was extracted using the PAXgene Blood miRNA Kit (Qiagen) according to the manufacturer's instructions. RNA quantity and quality were determined using the Nanodrop 1000 spectrometer (Thermo Fisher Scientific, Landsmeer, the Netherlands) and the Expirion High-sensitivity RNA analysis kit (Bio-Rad, Waltham, MA, USA), respectively. Total RNA from whole blood was deprived of globin using Ambion's GLOBINclear kit. RNAseq libraries were prepared from 1 μg RNA of each cell population using the TruSeq RNA sample preparation kit v2 (Illumina) according to the manufacturer's instructions, and these libraries were subsequently sequenced on a HiSeq 2000 sequencer (Illumina) using paired-end sequencing of 2 × 50 bp, upon pooling of 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing.

2.5. Analysis of RNAseq reads

The sequencing reads from the LifeLines Deep data were mapped to human reference genome NCBI build 37 using STAR v2.3.1 [16], allowing for eight mismatches and five mapping positions. To reduce reference mapping bias, GoNL SNPs with MAF >1% were masked by “N”. On average, 92% of the reads were mapped, and 88% of all reads were mapped uniquely. In total, 88% of all aligned reads were mapping to exons.

Gene expression was estimated using HTSeq count [17] using Ensembl GRCh37.71 gene annotation. Only uniquely mapping reads were used for estimating expression. Before eQTL mapping, gene expression data was TMM (trimmed mean of M values), normalized [18] and log2-transformed. The expression of each gene was centred and scaled. To reduce the effect of non-genetic sources of variability, we applied principal component analysis on the sample correlation matrix and the first five components were used as covariates [19].

2.6. Cis- and trans-eQTL mapping

As a discovery set, 629 peripheral blood samples from the LifeLines Deep cohort were investigated to map *cis*-eQTLs. For *trans*-eQTL analysis, we performed a meta-analysis with 456 haematological samples downloaded from public databases (https://www.ebi.ac.uk/arrayexpress/) [20]. The eQTL mapping strategy data has been described in detail previously [19,21]. Briefly, *cis*-eQTL analysis was performed on transcript-SNP combinations for which the distance from the centre of the transcript to the genomic location of the SNP was ≤250 kb, whereas eQTLs with a distance greater than 5 Mb were defined as *trans*-eQTLs. Associations were tested by non-parametric Spearman's rank correlation test and the FDR significance thresholds ($P < 0.05$; Supplemental Table 1) were defined based on the number of SNPs associated to each disease.

We categorized the SNPs as follows: (1) SNPs showing primary effect where the index SNP is same as the SNP showing the strongest eQTL effect, or is a different SNP but in perfect

LD ($D' = 1$) or in high LD ($r^2 \geq 0.8$). (2) SNPs showing secondary effect where the index SNP and SNP showing the strongest eQTL effect show differences in allele frequency resulting in low D' values and r^2 values <0.8, but after conditioning on the eQTL SNP, the eQTL effect of index SNP is still significant. We excluded SNPs where the index SNP and SNP showing the strongest eQTL affect show differences in allele frequency resulting in low D' values and r^2 values <0.8 and where, after conditioning on the eQTL SNP, the eQTL effect is gone (n = 237).

An additional *cis*-eQTL mapping was performed to include all the SNPs, extending the mapping distance to 1 Mb up- and downstream.

2.7. Replication cohort

The *cis*-eQTL results were replicated in the Rotterdam cohort, comprising 651 PBMCs samples. RNA-seq data was obtained in the same way as LifeLines Deep samples. The adaptors identified by FastQC (v0.10.1) were clipped using cutadapt (v1.1) applying default settings. Sickle (v1.200) was used to trim low quality ends of the reads (minimum length 25, minimum quality 20). The sequencing reads were mapped to human reference genome NCBI build 37 using STAR v2.3.1 [16], allowing for eight mismatches and five mapping positions. To reduce reference mapping bias, GoNL SNPs with MAF >1% in the reference genome were masked by "N". Ensembl GRCh37.71 was used for gene annotation. The overlapping exons were merged into meta-exons and gene expression was calculated as the sum of expression values of all meta-exons of each gene. To do this, custom scripts were developed which use coverage per base from coverageBed, and intersectBed from the Bedtools suite (v2.17.0) [22] and R (v2.15.1). Before the eQTL analysis, the data was normalized as for LifeLines Deep and corrected for the first 25 principal components. In the replication set, *P* values <0.05 were considered to indicate significant eQTLs, but only if the direction of the effect was the same as in the LifeLines Deep data.

2.8. Expression data from seven cell types

We used the expression data generated from immune cell subsets from two individuals. These immune cell subsets were granulocytes, monocytes, NK cells, B cells, memory T-cells (both CD4+ and CD8+), naive CD4+ (T-helper cells) and naive CD8+ (cytotoxic T-cells). These datasets has been described previously [23].

2.9. Analysis of autophagy genes expression data in coeliac disease biopsies

Intestinal biopsies from 31 coeliac disease (CeD) patients and 12 healthy controls were investigated according to the United European Gastroenterology (UEG) criteria. Biopsy sampling, RNA isolation, details of microarray hybridization and data analyses have been described previously [24,25]. DNA from the 31 patients was genotyped using the Immunochip platform [26] and the genotype data at *ULK3* SNP was extracted. A list of 222 autophagy genes was extracted from Human Autophagy Database (HADb; http://autophagy.lu/clustering/index.html) and, for 217 of these genes, we could extract their expression data from biopsies. In total, we found 1155 out of 28,000 genes showing significant differential expression between the CeD biopsies and controls ($P < 0.05$). Among these, nearly 50% of autophagy genes (107 genes out of 217) showed differential expression.

A Fisher exact test indicated that autophagy genes are more enriched in differentially expressed genes ($P < 2.2 \times 10^{-16}$) than non-autophagy genes. In addition, a 1000 times random sampling of genes from the same dataset produced a similar result (Kolmogorov–Smirnov test; $P = 0.002$), suggesting enrichment of autophagy genes within the differentially expressed genes in the CeD biopsies. To correlate the expression data with genotypes, the normalized expression values were stratified according to the genotypes. The significant differences were tested using the t-test and *P* values <0.05 were considered significant.

2.10. Cytokine analysis

Cytokines from *Candida albicans* and LPS stimulation in human PBMCs were measured using an enzyme-linked immunosorbent assay (R&D Systems, Minneapolis, MN, USA), as previously described [27,28]. The cytokine levels were log-transformed and the correlation between cytokine production and genotypes was tested by a linear regression model, which included age and gender as co-variables. *P* values <0.05 were considered significant.

2.11. Size of the linkage disequilibrium blocks, transcription factor-and super enhancer enrichment analyses

The enrichment analysis was done to test whether there is any significant difference in the sizes of the LD block, type of transcription factor (TF) binding, or number of super enhancers encompassed in single-gene SNP and multi-gene SNP regions. Initial cis-eQTL mapping within a 500 kb cis-window identified 112 single-gene SNP and 71 multi-gene SNP regions. The size of the LD blocks for these regions was defined based on the location of their proxy SNPs ($r^2 \geq 0.8$; extracted using the CEU population in 1000 Genomes data). The difference in the size of the LD blocks between these two groups was calculated using the Wilcoxon Rank test and *P* values < 0.05 were considered significant.

Upon *cis*-eQTL mapping in a 2 Mb extended window, we obtained 90 single-gene SNP regions and 92 multi-gene SNP regions. The super enhancers within 86 cell lines were extracted from a published source [29]. (See Supplemental Table 3 for more detailed information about the cell lines.) The 90 single-gene SNP regions and 92 multi-gene SNP regions were intersected with the chromosomal coordinates of these super enhancers. The significant difference was tested by the Fisher Exact test and a *P* value of <0.05 was considered significant.

We extracted transcription factor (TF) binding data from ChIPseq experiments using RegulomeDB (http://regulomedb.org/). These regions were intersected with gSNPs and their proxies ($r^2 \geq 0.8$). We calculated the number of binding events per TF in each loci. The difference in the number of binding events per TF between multi-gene SNP and single-gene SNP regions was calculated using the Fisher Exact test and *P* values <0.05 were considered significant.

2.12. Intersecting DNase I hypersensitivity sites

The DNase I hypersensitive sites from different cell lines were extracted from the BluePrint epigenome project (http://www.blueprint-epigenome.eu/index.cfm?p=B5E93EE0-09E2-5736-A708817C27EF2DB7) and the ENCODE data from the UCSC

browser (http://ucscbrowser.genap.ca/cgi-bin/hgTrackUi?g=wgEncodeAwgDnaseUniform). These regions were intersected with gSNPs and their proxies ($r^2 \geq 0.8$). See Supplemental Table 4 for more information about the cell lines.

2.13. RNA network for gene function and pathway prediction

The RNA network is an extension of the gene network [6,30] and was built to generate functional predictions for non-coding RNA genes. The gene network database contains data extracted from approximately 80,000 microarray experiments that are publically available from the Gene Expression Omnibus. It contains data on 54,736 human, 17,081 mouse and 6023 rat Affymetrix array experiments. Principal component analysis was performed on probe-set correlation matrices of each of four platforms (two human platforms, one mouse and one rat platform), resulting in 777, 377, 677 and 375 robust principal components, respectively. Jointly these components explain between 79% and 90% of the variance in the data, depending on the species or platform. Many of these components are well conserved across species and enriched for known biological phenomena. This meant we were able to combine the results into a multi-species gene network with 19,997 unique human genes, allowing us to utilize the principal components to accurately predict gene function by using a 'guilt-by-association' procedure. Pathway prediction for each gene was performed as described by Fehrmann et al [7]. Prediction was based on the 5000 RNA-seq samples, using the Gene Ontology Biological Processes gene set for pathway definition. The resulting gene–pathway association Z-scores was compared between selected genes and all the other genes in the genome using Mann–Whitney U tests. The resulting *P* values therefore describe the predicted enrichment of each Gene Ontology term for the selected genes (manuscript in preparation). The most significant top-five terms are reported in the figures.

2.14. CHIA-PET analysis

The RNAPII mediated interactions were identified by ChIA-PET in GM12878 cells with stringent quality filtering. The ChIA-PET experimental and informatics protocols used to elucidate these chromatin interactions have been previously described [31,32]. The full dataset is published elsewhere [33].

3. Results

3.1. RNAseq based eQTL-mapping identifies 233 causal AID genes including 53 non-coding RNA genes

To systematically examine the effect of AID-SNPs on the expression of both protein-coding and ncRNA genes, we used RNA-sequencing data from PBMCs from 629 healthy individuals. We performed conditional *cis*-eQTL mapping on a final set of 460 unique SNPs encompassing 268 loci (defined as 1 Mb region around the index SNPs), which have been associated to 14 different AIDs (Supplemental Fig. 1, Supplemental Table 2). By mapping the *cis*-eQTLs at 460 AID SNPs, we identified 183 SNPs as *cis*-eQTLs (40%) that were correlated with the expression of 233 different transcripts (resulting in 326 SNP-gene pairs). Replication ($P < 0.05$) in the same direction was achieved for 84% of the SNP-gene pairs in independent RNA-sequencing data (651 PBMCs) (Supplemental Fig. 2). Our *cis*-eQTL mapping implicated 63 SNPs that affect the expression of 53 ncRNAs, and 157 SNPs that

affect the expression of 180 protein-coding genes (Supplemental Table 5). Of these, 112 were associated to the expression levels of single genes (i.e. 87 AID-SNPs were associated to 68 protein-coding genes, and 25 AID-SNPs to 22 ncRNAs), while the other 71 AID *cis*-eQTL SNPs affect the expression levels of two to seven different genes in a 500 Kb region (Fig. 1A). The percentage of *cis*-eQTLs observed for each AID varied from 25% for juvenile idiopathic arthritis to almost 70% for primary sclerosing cholangitis (Fig. 1B). To reveal the regulatory consequences of the eQTL SNPs on downstream pathways and cellular phenotypes, we investigated the expression-specificity of the eQTL-affected genes across seven different immune cell types (Supplemental Fig. 1) and predicted the gene function using co-expression analysis [30]. This was particularly necessary for the ncRNA genes since their molecular functions are largely unknown.

3.2. Integrative genomic analyses implicate novel genes for AIDs and implicate autophagy in celiac disease

Our *cis*-eQTL mapping revealed novel candidate genes in several loci, of which *GPR25* (G protein-coupled receptor 25; associated with ankylosing spondylitis (AS), coeliac disease (CeD), inflammatory bowel disease (IBD), multiple sclerosis (MS) locus on 1q32.1) and *ULK3* (unc-51-like kinase 3; CeD locus on 15q24.1) are examples (Fig. 2A, B). Previously, *C1orf106* was implicated as a causal gene for IBD at 1q32.1 [34]. While the expression level of *C1orf106* was moderately correlated with four disease-associated SNPs, our analysis unequivocally identified *GPR25* as the most significantly affected gene in this locus (Fig. 2A). We found *GPR25* to be strongly expressed in memory T-cells and NK-cells, and predicted to be involved in the positive regulation of B-cell proliferation ($P = 1.21 \times 10^{-6}$). Nonetheless, *C1orf106* is suggested to be involved in epithelial cell–cell adhesion [35], an essential process in keeping the intestinal epithelial-barrier intact in the context of IBD and CeD. Therefore, it is possible that both genes in this locus contribute to disease through different cell types. A similar observation was made for the CeD locus on 15q24 (Fig. 2B). The expression level of *ULK3* is much more strongly affected by rs1378938 ($P = 1.21 \times 10^{-46}$) than *CSK* ($P = 7.08 \times 10^{-6}$). *CSK* is involved in B-cell activation [36] and *ULK3* encodes a kinase involved in autophagy [37]. However, the autophagy pathway has never been implicated in CeD, which highlights the possibility of identifying novel pathways underlying CeD.

3.3. Studying patient biomaterials validated the functional role of ULK3 locus in autophagy in celiac disease

To investigate the functional impact of ULK3 locus on autophagy pathway in CeD, we tested whether known autophagy genes were differentially regulated in intestinal biopsies of CeD patients [35]. We indeed observed an enrichment of autophagy genes being differentially expressed (Fig. 3A) compared to a random set of genes in CeD biopsies ($P = 2.2 \times 10^{-16}$). We further confirmed that the *ULK3* affecting SNP genotype is correlated with the expression levels of autophagy genes in CeD biopsies (Supplemental Fig. 3). Autophagy is also involved in regulating inflammatory process in response to human pathogens by influencing cytokine production and secretion. We therefore tested whether ULK3 SNP influences the production of cytokines in PBMCs in response to lipopolysaccharide (LPS), a cell-wall component of Gram negative bacteria. We found a significant difference ($P =$

0.019) in the levels of IL-6 between rs1378938 CC and TT genotypes, where the risk allele T is associated with lower levels of IL-6 in response to LPS (Fig. 3B). This finding indicated that ULK3 dependent autophagy might be involved in regulation of inflammation and emphasizes the importance of studying non-gluten antigens (e.g. host–microbiome interaction) in the context of CeD pathogenesis.

3.4. AID-associated lncRNAs are involved in immune cell activation and cytokine regulation

Our study identified 27 ncRNAs at 25 AID-loci as candidate causal genes as these were the only affected transcripts in PBMCs (Table 1). Although previous microarray-based eQTL studies suggested *UBE2E3* as the causal gene at 2q31.3, a locus associated with CeD and AS [35], our analysis indicates that the AS (rs12615545)-and CeD (rs1018326)-associated SNPs are not in LD with the eQTL SNP that affects *UBE2E3* expression ($r^2 = 0.16$). Instead, both the AD and CeD index SNPs ($r^2 = 0.94$) affect the expression of long non-coding RNA (lncRNA) *AC104820.2* ($P = 9.22 \times 10^{-8}$). *AC104820.2* is strongly expressed in CD8+ T-cells (Fig. 4A) and is suggested to function in alpha-beta T-cell proliferation ($P = 4.1 \times 10^{-6}$), which is a crucial process in autoimmunity [38]. *AC104820.2* was also found to be up-regulated in intestinal biopsies of patients with active CeD [39]. Another example is lncRNA *AP002954.4* at 11q23.3, whose expression was significantly affected by three unrelated SNPs associated to three different AIDs (MS, RA and CeD). *AP002954.4* is expressed specifically in monocytes and may function in regulating cytokine responses ($P = 5.95 \times 10^{-6}$) and defence against fungal infection ($P = 3.79 \times 10^{-5}$) (Fig. 4B). Indeed, the cytokine levels produced by fungus-stimulated, human PBMCs [27,28] were dependent on SNP *rs533646*, with the risk allele G causing higher IL-6 and TNF-alpha levels (Supplemental Fig. 3). Both examples highlight the potential functional role of lncRNAs in AID.

3.5. Different AID-SNPs in shared disease loci may affect different genes

Although we found *cis*-eQTLs for 183 SNPs, they define only 120 loci (based on the 1 Mb region around the index SNPs). Among these 120 loci, 54 loci were found shared between two or more AID Supplemental Table 6). In nine of the shared loci (16.7%), we found different AID-SNPs affecting different genes. The MS and IBD associated locus at 11q13.1 provides an illustrative example. MS SNP rs694739 affects two protein-coding genes and a lncRNA gene, while the IBD SNP rs559928 affects a different lncRNA (Fig. 5). The MS SNP rs694739 and its close proxies overlap with DNAse I hypersensitivity sites (DHSs) of many immune cells, whereas one of the two proxies of the IBD SNP rs559928 specifically overlaps with DHSs in Caco-2 cells (intestinal epithelial cells), suggesting that different usage of cell-type-specific enhancers could be one mechanism by which different genes could be affected by different AID-SNPs in a shared disease locus.

3.6. Nearly 40% of AID-SNPs modulate the expression of multiple genes in cis due to extended linkage disequilibrium

About 39% of AID-SNPs (71/183 SNPs) affect the expression levels of two or more genes (multi-gene SNPs) compared to 61% (112/183 SNPs) that affect only a single gene (single-gene SNPs) (Fig. 1A). We found no difference in the average number of genes present in

these loci, the allele frequencies of the SNPs, or local patterns of co-expression of genes (Supplemental Figs. 4, 5 & 6) that could explain the differences. However, we observed significant differences in the size of the LD blocks ($P = 0.0002$), with multi-gene SNPs located within larger LD blocks (average LD block size 175 kb) than single-gene SNPs (average LD block size 100 kb) (Fig. 6A). This suggests that multi-gene SNPs may regulate gene expression in a larger *cis*-window. To test this, we extended the *cis*-window to 2 Mb and repeated the *cis*-eQTL mapping. We observed that 42% of multi-gene SNPs (30/71) affect genes that are located in the extended 2 Mb *cis*-window compared to only 19% of single-gene SNPs (21/112, $P = 0.0001$; Supplemental Fig. 7), suggesting that some AID-SNPs are located within regulatory regions that can control the expression of multiple genes spread over long distances in *cis*. In support of this we found that multi-gene SNPs regions are enriched for super-enhancers [29] ($P = 0.0018$) and CTCF binding sites (Supplemental Figs. 8 and 9), which are known to mediate chromatin looping [40,41].

3.7. Expression regulation of multiple genes in cis is partly via chromatin looping interactions

Recent studies by others suggest that lncRNAs play a role in regulating transcription of genes over longer distances, by mediating chromatin looping that brings enhancers and promoters together [42,43]. We therefore tested whether AID-SNPs affecting lncRNAs are involved in regulating multiple genes in *cis* more often than AID-SNPs affecting only protein-coding genes. We found that 69% of AID-SNPs that impact lncRNAs are also multi-gene SNPs (based on the 2 Mb *cis*-window) compared to 40% of AID-SNPs that impact only protein-coding genes ($P = 0.0002$; Fig. 6B), suggesting that lncRNAs could be one of the factors in regulating the expression of multiple *cis* genes. To confirm the observation that AID-SNPs and lncRNAs are potentially involved in looping interactions to regulate gene expression, we manually cross-referenced our lncRNA-eQTLs with genome-wide RNA polymerase II (RNAPII) interactions, mapped by ChIA-PET assay [31,32] in GM12878 (B-lymphoblastoid) [33]. We found that multi-gene SNPs are more often (64 SNPs out of 92; 70%) involved in looping interactions in B cells ($P = 0.048$) than single-gene SNPs (55%; Supplementary Tables 7 and 8). Many of these loci show very strong interaction where more than 50 independent interactions between AID SNPs and regulatory regions of eQTL genes are found (Supplementary Fig. 10a–s). We found that for some lncRNA-eQTLs, the affected lncRNAs and the promoters of the protein-coding genes are organized in the same transcription topological unit mediated by RNA Polymerase II (RNAPII). For example, SNP rs6667605 at 1q32.1 associated with UC affects the levels of transcription of four genes (*MMEL1*, *TNFRSF14*, *RP3-395M20.7* and *RP3-395M20.8*). According to the RNAPII ChIA-PET interaction data, we observed that SNP rs6667605 and the lncRNAs *RP3-395M20.7* and *RP3-395M20.8* were interacting with the promoter region of the protein-coding gene *TNFRSF14* (Fig. 6C). All three genes are expressed in lymphoblast cells and predicted to be involved in B- and T-cell activation (Supplemental Figs. 11 and 12), supporting the idea that co-regulation of *TNFRSF14* and the two lncRNAs may occur through looping interactions.

3.8. Rheumatoid arthritis lncRNA as an example of trans-regulator of genes enriched for B cell proliferation

Some AID-SNPs can affect the expression of multiple genes in *trans* [4]. We tested whether lncRNA *cis*-eQTLs can also be *trans*-eQTLs. We observed that the RA-associated SNP rs13330176 at 16q24.1 affects levels of expression of lncRNA *RP11-542M13.2* in *cis*, which is predicted to be involved in B-cell proliferation (Supplemental Fig. 13). The same SNP is also moderately associated ($P < 0.0009$) with the expression level of more than ten genes in *trans* (Supplemental Table 9), which are also involved in B-cell related processes. Although our sample size is a limiting factor in *trans*-eQTL mapping, this example illustrates the possibility of identifying downstream consequences of AID-associated lncRNAs.

4. Discussion

Identification of the correct causal genes within GWAS loci is critical not only for the proper interpretation of associated loci but also to pinpoint potential therapeutic targets. RNA-sequencing will identify causal genes in an unbiased manner as it provides quantification of the global transcriptome at high resolution to efficiently capture all transcripts including lowly abundant transcripts such as lncRNA genes, [23]. Our study indeed shows that eQTL analysis based on RNA-sequencing data is a first and important step in delineating the complex genetic architecture of AIDs. We made use of the currently largest available cohort of PMBC-based RNA-sequencing data and report, for the first time, an extensive list of 53 ncRNAs as candidate causal genes for AIDs, which highlight the added value of RNA-sequencing over conventional microarrays. Since lncRNA genes have shown to be more cell-type specific than protein-coding genes [44], further eQTL analysis in disease specific cell-types may identify additional lncRNA genes as potential causal genes for AID.

In addition to lncRNAs, we also identified 157 cis-eQTLs on protein coding genes, which helped us to implicate novel candidate causal genes and pathways in several loci. We found that a SNP associated to CeD was affecting *ULK3* that encodes a kinase involved in autophagy. By analysing intestinal biopsies of CeD patients, we observed that *ULK3* is co-regulated with genes involved in autophagy thereby implicating a possible role of the autophagy pathway in CeD pathogenesis that was not considered before [35]. Another example is a shared AID locus for MS, IBD, AS and CeD at 1q32.1 region, where our analysis identified *GPR25* as a primary causal gene. G protein-coupled receptors (GPCRs) are one of the best-studied classes of cell surface receptors and the most amenable group of proteins for novel small molecule drug discovery. GPR35 is one such example where it emerged as a potential therapeutic target through its association with IBD, type 2 diabetes and coronary heart disease [45]. The newly identified *GPR25* gene could be another attractive target to exploit therapeutically since it is associated with four different AIDs. Interestingly, both *ULK3* and *GPR25* did not turn out to be the genes nearest to the index SNPs which is rather frequently the case since only in 50% of cases the nearest gene is suggested to be the candidate causal gene (Supplemental Table 5). This finding again highlights the importance of a more careful follow up analysis of disease associated SNPs for the correct interpretation of associated loci.

Another important observation from this study is that in 40% of AID-loci multiple genes may predispose to disease. In addition, by comparing the genetics of different diseases, we show that the associated loci for many different diseases can be physically the same but yet different genes in these loci may affect different AIDs. The latter may be caused by cell type specificity, as suggested by our analysis of DNAse I hypersensitive sites. Integrating eQTL analysis with regulatory information may mark the cell-types critical to disease and techniques like ChIA-PET will be critical to eventually unmask the cell-type specific regulatory networks for different diseases. The ability to perform *trans*-eQTL mapping will facilitate the identification of the downstream effects of disease SNPs. In particular the further understanding of lncRNAs will benefit from such investigations. In this study we focused on the identification of *cis*-eQTLs because our sample size had limited power for *trans*-eQTL mapping. Since the majority of the GWAS SNPs regulate gene-expression, our systematic approach of conditional eQTL mapping and cell-type specific expression characterization of candidate causal genes for autoimmune diseases could be efficient strategy to identify novel causal genes for other complex diseases. We anticipate that our study will facilitate the understanding of lncRNA-mediated *cis* gene expression regulation in AID loci and encourage further studies with bigger sample size to identify the downstream consequences of AID-SNPs on lncRNAs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jackie Senior and Kate Mc Intyre for editing the manuscript. This work was supported by a European Research Council Advanced Grant (FP/2007–2013/ERC grant 2012-322698 to CW), the Dutch Digestive Diseases Foundation (MLDS WO11-30 to CW and VK), the European Union's Seventh Framework Programme (EU FP7) TANDEM project (HEALTH-F3-2012-305279 to CW and VK), the Netherlands Organization for Scientific Research (NWO VENI grant 916-10135 and NWO VIDI grant 917–14374 to LF; NWO VENI grant 863.13.011 to YL), and by the Dutch Multiple Sclerosis Foundation (grant 11-752 to SW). This work was performed within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.leidenlangleven.nl), the Rotterdam studies (http://www.erasmus-epidemiology.nl/rotterdamstudy) and the CODAM study (http://www.carimmaastricht.nl). We thank the participants of all aforementioned biobanks and acknowledge the contributions of the investigators to this study. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative and the Groningen Center for Information Technology (Strikwerda, W. Albers, R. Teeninga, H. Gankema and H. Wind) and Target storage (E. Valentyn and R. Williams). Target is supported by Samenwerkingsverband Noord Nederland, the European Fund for Regional Development, the Dutch Ministry of Economic Affairs, Pieken in de Delta and the provinces of Groningen and Drenthe."

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–D1006. [PubMed: 24316577]
2. Ricano-Ponce I, Wijmenga C. Mapping of Immune-mediated Disease Genes. Annual Review of Genomics and Human Genetics. 2013
3. Kumar V, Wijmenga C, Xavier RJ. Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. Curr Opin Immunol. 2014; 31C:51–57.

4. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]
5. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–1774. [PubMed: 22955987]
6. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 2013; 9:e1003201. [PubMed: 23341781]
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
8. Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet. 2008; 371:1505–1512. [PubMed: 18455228]
9. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: design, and project goals. Eur J Hum Genet EJHG. 2014; 22:221–227. [PubMed: 23714750]
10. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur J Hum Genet EJHG. 2014; 22:1321–1326. [PubMed: 24896149]
11. Whole-genome sequence variation. population structure and demographic history of the Dutch population. Nat Genet. 2014; 46:818–825. [PubMed: 24974849]
12. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat methods. 2012; 9:179–181.
13. Deelen P, Bonder MJ, van der Velde KJ, Westra HJ, Winder E, Hendriksen D, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 2014; 7:901. [PubMed: 25495213]
14. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5:e1000529. [PubMed: 19543373]
15. Byelas, H., Dijkstra, M., Neerincx, PBT., van Dijk, F., Kanterakis, A., Deelen, P., et al. Scaling bio-analyses from computational clusters to grids. Proceedings of the 5th International Workshop on Science Gateways; Zurich, Switzerland. 2013. 3–5 June, 2013, CEUR-WS.or
16. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]
17. Anders S, Pyl PT, Huber W. HTSeq – a Python framework to work with high-throughput sequencing data. Bioinformatics. 2014; 31:166–169. [PubMed: 25260700]
18. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]
19. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, et al. 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. 2011; 7:e1002197. [PubMed: 21829388]
20. Deelen P, Zhernakova D, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. 2014
21. Zhernakova DV, de Klerk E, Westra HJ, Mastrokolias A, Amini S, Ariyurek Y, et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genet. 2013; 9:e1003594. [PubMed: 23818875]
22. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

23. Hrdlickova B, Kumar V, Kanduri K, Zhernakova DV, Tripathi S, Karjalainen J, et al. Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. Genome Med. 2014; 6:88. [PubMed: 25419237]
24. Almeida R, Ricano-Ponce I, Kumar V, Deelen P, Szperl A, Trynka G, et al. Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. Hum Mol Genet. 2014; 23:2481–2489. [PubMed: 24334606]
25. Diosdado B, Wapenaar MC, Franke L, Duran KJ, Goerres MJ, Hadithi M, et al. A microarray screen for novel candidate genes in coeliac disease patho-genesis. Gut. 2004; 53:944–951. [PubMed: 15194641]
26. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011; 43:1193–1201. [PubMed: 22057235]
27. Smeekens SP, Ng A, Kumar V, Johnson MD, Plantinga TS, van Diemen C, et al. Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. Nat Commun. 2013; 4:1342. [PubMed: 23299892]
28. Kumar V, Cheng SC, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. Nat Commun. 2014; 5:4675. [PubMed: 25197941]
29. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–947. [PubMed: 24119843]
30. Fehrmann RS, Karjalainen JM, Krajewska M, Westra HJ, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015; 47:115–125. [PubMed: 25581432]
31. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]
32. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98. [PubMed: 22265404]
33. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015; 163:1611–1627. [PubMed: 26686651]
34. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011; 43:1066–1073. [PubMed: 21983784]
35. Kumar V, Gutierrez-Achury J, Kanduri K, Almeida R, Hrdlickova B, Zhernakova DV, et al. Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. Hum Mol Genet. 2014; 24:397–409. [PubMed: 25190711]
36. Manjarrez-Orduno N, Marasco E, Chung SA, Katz MS, Kiridly JF, Simpfendorfer KR, et al. CSK regulatory polymorphism is associated with systemic lupus erythematosus and influences B-cell signaling and activation. Nat Genet. 2012; 44:1227–1230. [PubMed: 23042117]
37. Young AR, Narita M, Ferreira M, Kirschner K, Sadaie M, Darot JF, et al. Autophagy mediates the mitotic senescence transition. Genes Dev. 2009; 23:798–803. [PubMed: 19279323]
38. Han A, Newell EW, Glanville J, Fernandez-Becker N, Khosla C, Chien YH, et al. Dietary gluten triggers concomitant activation of CD4+ and CD8+ alpha-beta T cells and gammadelta T cells in celiac disease. Proc Natl Acad Sci U S A. 2013; 110:13073–13078. [PubMed: 23878218]
39. Plaza-Izurieta L, Fernandez-Jimenez N, Irastorza I, Jauregi-Miguel A, Romero-Garmendia I, Vitoria JC, et al. Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease. Eur J Hum Genet EJHG. 2014; 23:1100–1105. [PubMed: 25388004]
40. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]

41. Majumder P, Gomez JA, Chadwick BP, Boss JM. The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. J Exp Med. 2008; 205:785–798. [PubMed: 18347100]
42. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature. 2013; 494:497–501. [PubMed: 23417068]
43. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol. 2014; 24:651–663. [PubMed: 25441720]
44. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]
45. Divorty N, Mackenzie AE, Nicklin SA, Milligan G. G protein-coupled receptor 35: an emerging target in inflammatory and cardiovascular disease. Front Pharmacol. 2015; 6:41. [PubMed: 25805994]

URLs

46. UCSC genome browser; http://genome-euro.ucsc.edu/index.html.
47. RNAseq data downloaded from public databases; https://www.ebi.ac.uk/arrayexpress/.
48. Access to RNA network; http://genenetwork.nl/wordpress/.
49. Access to Genotype Harmonizer; http://www.molgenis.org/systemsgenetics/.
50. Human Autophagy Database; http://autophagy.lu/clustering/index.html.

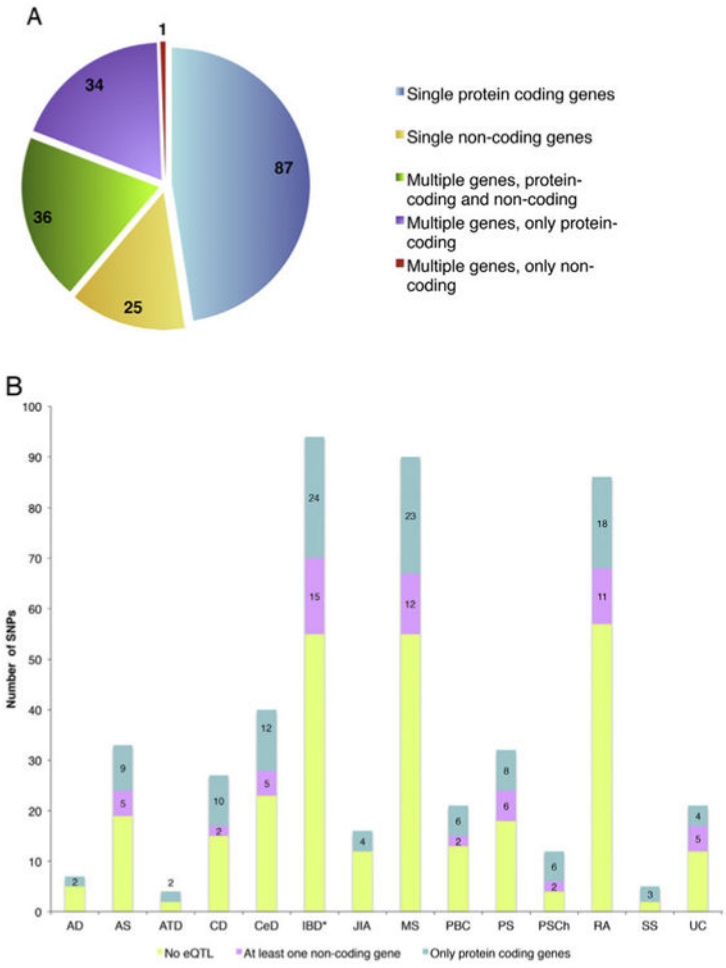

Fig. 1.

Break down of *cis*-eQTLs into coding and non-coding genes. (**A**) The pie-chart summarizes the *cis*-eQTLs we identified in which, 112 out of 183 *cis*-eQTL AID-SNPs were associated to single genes (i.e. 87 AID-SNPs to 68 protein-coding genes and 25 AID-SNPs to 22 ncRNAs), while the other 71 AID *cis*-eQTL SNPs affected the expression levels of two to seven different genes within the 500 kb region (see Supplemental Table 2). (**B**) The *cis*-eQTL mapping results per disease are shown to indicate the number of SNPs remaining for eQTL mapping (shown on top of the purple bars) as well as the number of SNPs that showed significant *cis*-eQTLs (shown on top of the orange bars). Because of the wide range in the number of SNPs associated to each of the AIDs, we defined false discovery rate (FDR) significance thresholds for each disease separately to assist in the eQTL analysis (see Methods). *These loci are shared between ulcerative colitis (UC) and Crohn's disease (CD). The non-shared loci are listed separately. Alopecia areata (AA), atopic dermatitis (AD), ankylosing spondylitis (AS), autoimmune thyroid disease (ATD), coeliac disease (CeD), inflammatory bowel disease (IBD), juvenile idiopathic arthritis (JIA), multiple sclerosis (MS), primary biliary cirrhosis (PBC), psoriasis (PS), rheumatoid arthritis (RA), primary sclerosing cholangitis (PSCh), and systemic sclerosis (SS).

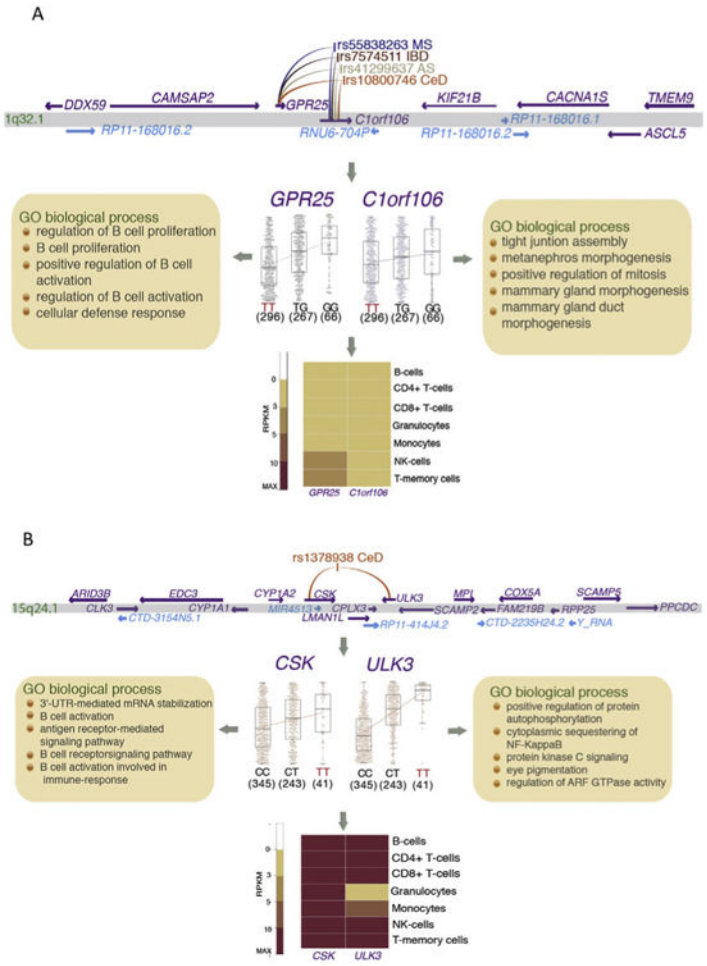

Fig. 2.

Cis-eQTL identifies novel candidate causal genes for AID. The top panel is a locus plot to show the location of all the genes tested in a 500 kb *cis*-window (hg19). The plot centred below indicates the correlation between the expression of eQTL-genes with genotypes of AID-SNPs. The risk genotypes are in red. The expression pattern for eQTL-genes across seven different immune cell types were obtained from two individuals and the average expression levels are shown as a heatmap in the lowest panel. (**A**) An example of different AID-SNPs affecting the same genes and thus representing a truly shared locus. Four different SNPs at 1q32.1 show association to four different diseases (MS, IBD, AS and CeD). All four SNPs are in absolute LD ($r^2 = 1$, D = 1) and affect two protein-coding genes. The MS-associated risk allele rs55838263*T is associated with lower expression of both *GPR25* ($P = 3.02 \times 10^{-16}$) and *C1ORF106* ($P = 0.0012$) genes. The risk alleles for the other three SNPs show similar results. (**B**) Example of a *cis*-eQTL identifying novel candidate genes and novel pathways. The CeD-associated risk allele, rs1378938*T is associated with a higher expression of both *CSK* ($P = 7.08 \times 10^{-6}$) and *ULK3* ($P = 1.21 \times 10^{-46}$). *ULK3* encodes a kinase involved in autophagy. This pathway has not been implicated in CeD so far. We found an enrichment of autophagy genes being differentially expressed compared to a random set of genes in CeD biopsies ($P = 2.2 \times 10^{-16}$). We further showed that the SNP

affecting *ULK3* is also correlated with the expression levels of autophagy genes in CeD biopsies (Supplemental Fig. 3).

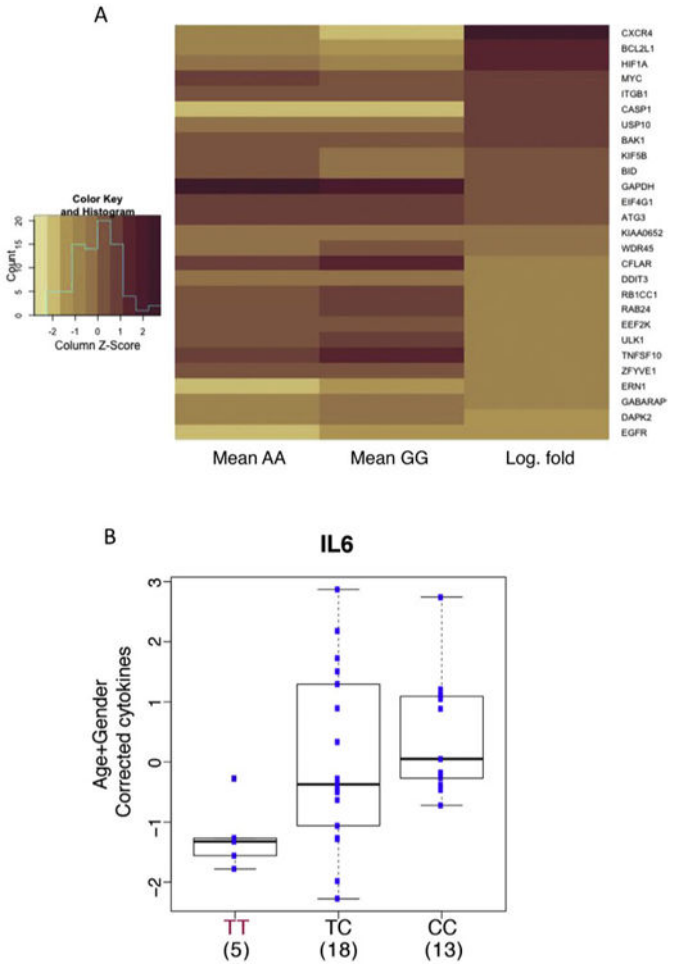

Fig. 3.

Validation of the functional role of ULK3 locus in autophagy in celiac disease. (A) Genotype-dependent expression levels of autophagy genes in coeliac disease biopsies. There were 3 AA, 15 AG and 13 GG genotypes. The expression for 217 autophagy genes could be extracted from the microarray data of coeliac disease biopsies and the genotype data at the *ULK3* SNP were extracted using Immunochip for 31 CeD patients. The heatmap shows the normalized expression values stratified according to the genotypes. The difference in gene expression between AA and GG was tested by t-test and $P < 0.05$ was considered significant. (**B**) Association of rs1378938 with interleukin 6 levels in response to lipopolysaccharide. The CeD-associated risk allele rs1378938*T (in red) results in lowered interleukin 6 ($P = 0.019$) cytokine levels upon *LPS* stimulation of primary mononuclear cells. The x-axis displays the three different genotypes and the number of individuals in each group. The y-axis presents the age and gender corrected cytokine levels.

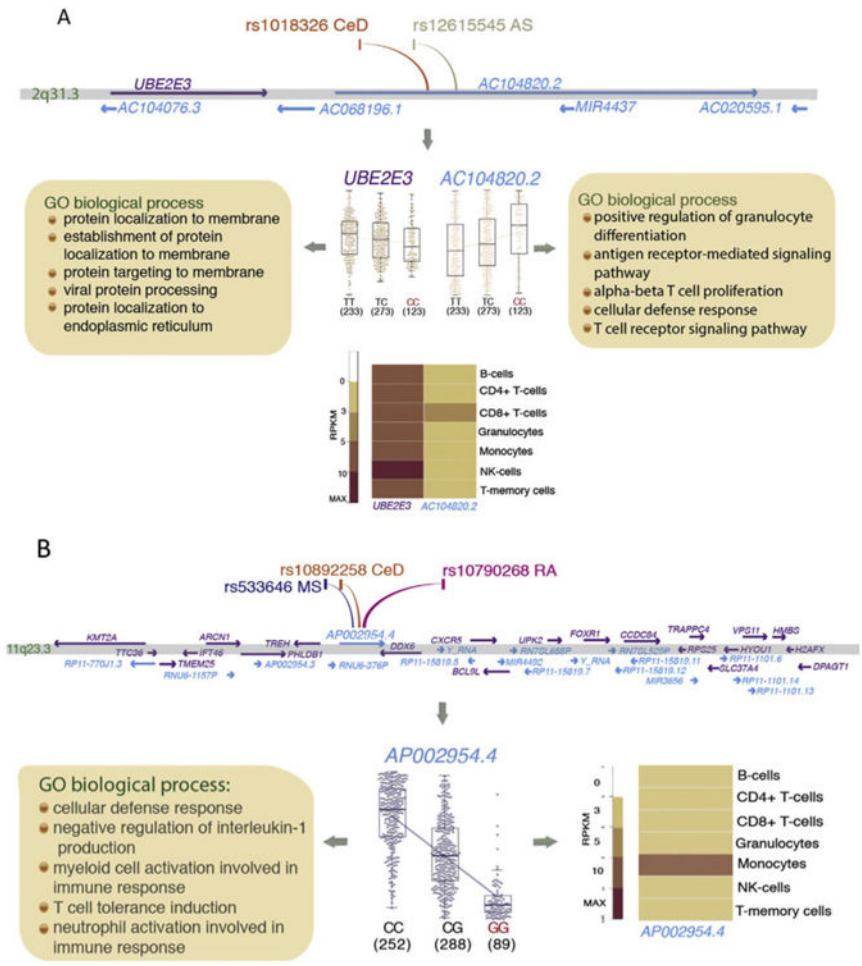

Fig. 4.

Examples of long non-coding RNAs as candidate causal genes for AIDs. (**A**) The locus on chromosome 2q32.3 is associated to AD (rs12615545) and CeD (rs1018326), and both SNPs are in strong LD ($r^2 = 0.96$, $D' = 1$). The CeD-associated risk allele, rs1018326*C is associated with higher levels of expression of the *AC104820.2* lncRNA (similar results were observed for AD risk allele at rs12615545). The function prediction based on co-expression (GO biological processes) suggested this lncRNA is involved in alpha-beta T-cell proliferation. The right-hand panel shows the expression pattern for *AC1048202* lncRNA across seven different immune cell types (obtained from two individuals and the average expression levels are shown), which indicates its strong expression in CD8+ T-cells. (**B**) The locus on chromosome 11q23.3 is associated to CeD (rs10892258), MS (rs533646) and RA (rs10790268). The MS-associated risk allele rs533646*G is associated with lower levels of expression of the *AP002954.4* lncRNA (eQTL $P = 6.41 \times 10^{-80}$; similar results were also observed for the CeD and RA risk alleles). The expression patterns across seven cell types were obtained from two individuals and the average expression levels are shown as a heatmap, which confirms the strong expression of *AP002954.4* in monocytes. The function prediction based on co-expression (GO biological processes) was obtained from the RNA network (http://genenetwork.nl/)[7].

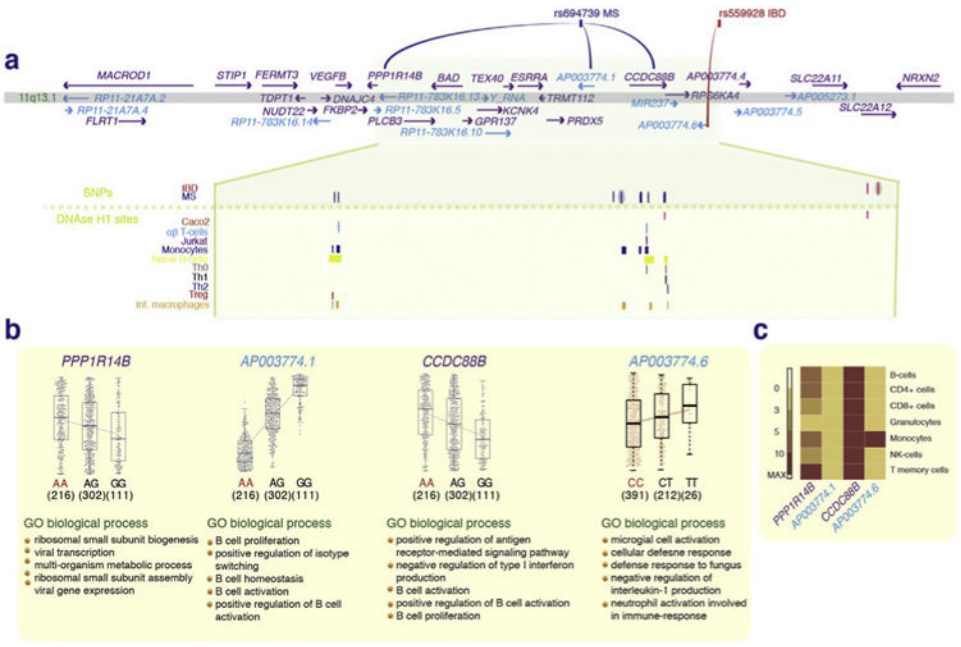

Fig. 5.

Shared disease loci may harbour different candidate causal genes for different AIDs. (**A**) The upper panel shows the regional plot of MS- (rs694739) and IBD- (rs559928) associated SNPs at 11q13.1 that affect expression of independent genes. The DNAse H1 sites (DHSs) from different cell lines were intersected with both gSNPs and their proxies. The gSNPs are highlighted with an oval shape around the line. DHSs of immune cells (alpha-beta T-cells, Jurkat T-cells, monocytes, naive B cells, T helper cells (Th0, Th1, Th2) and regulatory T-cells (Treg) and Caco-2 cells (intestinal epithelial cells) were extracted from the databases of ENCODE and the Blueprint epigenome project (**B**) The MS SNP rs694739 affects the expression levels of lncRNA *AP003774.1* ($P = 2.65 \times 10^{-130}$), followed by *CCDC88B* ($P = 7.89 \times 10^{-14}$) and *PPP1R14B* ($P = 7.08 \times 10^{-6}$), while IBD SNP rs559928 only weakly affects the expression level of lncRNA *AP003774.6* ($P = 3.21 \times 10^{-4}$). The function prediction based on co-expression (GO biological processes) was obtained from the RNA network [30]. Consistent with the DHSs pattern, the MS SNP affected genes are predicted to be involved in immune cell activation and the lncRNA affected by the IBD SNP is involved in innate immune function. (**C**) The expression pattern for eQTL genes across seven different immune cell types were obtained from two individuals and the average expression levels are shown as a heatmap.

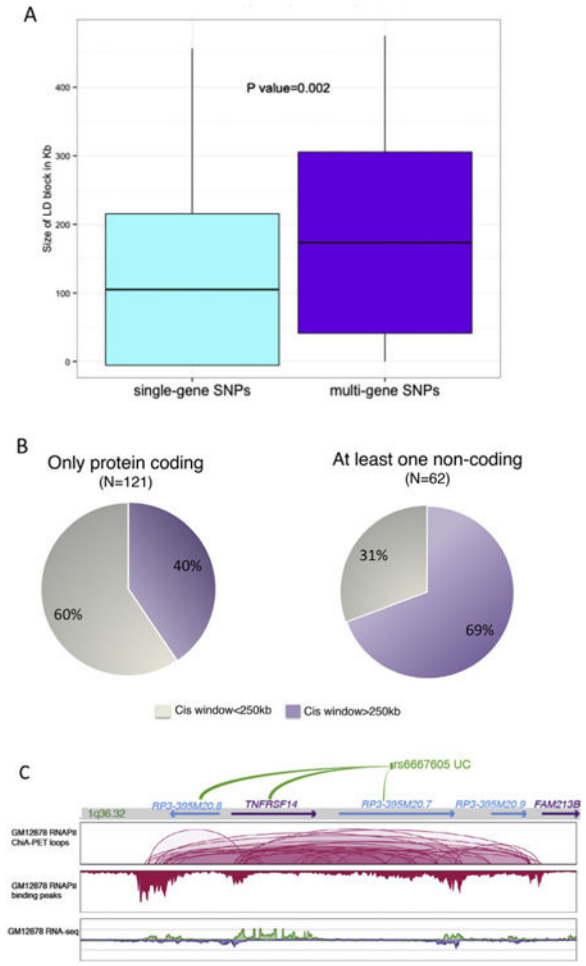

Fig. 6.

The size of the LD block and lncRNAs facilitate looping interactions to regulate multiple genes in cis. (**A**) The size of the LD blocks between SNPs that affect single genes (single-gene SNPs) and SNPs that affect multiple genes (multi-gene SNPs) was compared. The average size of the LD block for single-gene SNPs (100 kb) was significantly different from the 175 kb for multi-gene SNPs ($P = 0.0002$). The significant difference was tested using the Wilcoxon Rank test. (**B**) Mapping eQTLs for SNPs affecting lncRNA and using a 2 Mb cis-window found 69% of the AID-SNPs that impact lncRNAs are SNPs affecting multiple genes compared to 40% of AID-SNPs that impact only protein-coding genes. (**C**) Ulcerative colitis-associated SNP rs6667605 at 1q36.32 affects three genes (*TNFRSF14* is a protein-coding gene, *RP3-395M20.8* and *RP3-395M20.7* are lncRNAs). Pink loops depict the looping interactions mediated by RNAPII that lie between the UC-associated eQTL locus and the corresponding target genes in GM12878 (B-lymphoblastoid) cells. The peaks in the middle panel depict the RNAPII occupancy along this locus in GM12878 cells. The bottom panel shows the expression levels of genes in this locus. Expression signals from the + and – strands are separated into green and blue, respectively.

J Autoimmun. Author manuscript; available in PMC 2017 April 14.

Table 1

AID-SNPs that affect the expression levels of only non-coding RNAs.

AID	SNP	SNP Chr	SNP Chr position[a]	HUGO Gene ID	Gene Chr position[a]	Risk allele[b]	Direction[c]	eQTL P
CD	rs1260326	2	27730940	*AC109828.1*	27561937	T	up	1.44×10^{-4}
CD	rs1260326	2	27730940	*FTH1P3*	27615942	T	up	8.40×10^{-3}
RA	rs34695944	2	61124850	*RP11-373L24.1*	61154165	C	up	7.99×10^{-8}
IBD	rs6708413	2	103063369	*MIR4772*	103048787	G	down	8.01×10^{-15}
CeD	rs1018326	2	182007800	*AC104820.2*	182115472	C	up	9.22×10^{-8}
AS	rs12615545	2	182048452	*AC104820.2*	182115472	C	up	5.99×10^{-8}
AS	rs10045403	5	96147733	*CTD-2260A17.1*	96121092	A	down	7.89×10^{-14}
MS	rs941816	6	36375304	*MIR3925*	36590251	G	down	3.88×10^{-3}
MS	rs706015	7	27014988	*HOTAIRM1*	27137575	C	up	1.51×10^{-36}
IBD	rs4743820	9	93928416	*RP11-305L7.1*	93868412	T	up	4.86×10^{-23}
PS	rs10979182	9	110817020	*RP11-240E2.2*	110802683	A	down	5.56×10^{-5}
RA	rs2671692	10	50097819	*RP11-563N6.6*	50086380	A	up	3.90×10^{-4}
IBD	rs7097656	10	82250831	*RP11-137H2.4*	82292525	C	down	1.41×10^{-7}
MS	rs694739	11	64097233	*AP003774.1*	64094749	A	down	2.65×10^{-130}
IBD	rs559928	11	64150370	*AP003774.6*	64162558	C	up	3.21×10^{-4}
UC	rs561722	11	114386830	*RP11-64D24.2*	114154315	C	down	4.55×10^{-3}
MS	rs533646	11	118566746	*AP002954.4*	118598144	G	down	6.41×10^{-80}
CeD	rs10892258	11	118579865	*AP002954.4*	118598144	G	up	2.18×10^{-3}
RA	rs10790268	11	118729391	*AP002954.4*	118598144	G	down	4.55×10^{-3}
MS	rs12296430	12	6503500	*RP1-102E24.8*	6503653	C	up	1.01×10^{-21}
PS	rs2066819	12	56750204	*RP11-977G19.11*	56701259	C	down	2.00×10^{-3}
MS	rs6498184	16	11435990	*RP11-485G7.5*	11437367	G	down	4.92×10^{-3}
RA	rs4780401	16	11839326	*UBL5P4*	12062482	T	down	3.88×10^{-3}
RA	rs13330176	16	86019087	*RP11-542M13.2*	86016842	A	down	6.88×10^{-4}
IBD	rs727088	18	67530439	*RP11-543H23.2*	67382212	G	up	1.37×10^{-18}
PS	rs892085	19	10818092	*ILF3-AS1*	10763529	A	down	9.05×10^{-4}
UC	rs6017342	20	43065028	*RP5-1013A22.2*	43078074	C	down	3.58×10^{-3}

AID	SNP	SNP Chr	SNP Chr position[a]	HUGO Gene ID	Gene Chr position[a]	Risk allele[b]	Direction[c]	eQTL P
IBD	rs913678	20	48955424	*RP11-290F20.1*	48920358	T	up	7.99×10^{-8}
AS	rs7282490	21	45615741	*AP001057.1*	45621876	G	down	4.29×10^{-4}
IBD	rs7282490	21	45615741	*AP001057.1*	45621876	G	down	4.29×10^{-4}
IBD	rs2413583	22	39659773	*RP4-742C19.8*	39534946	C	down	3.04×10^{-3}

AS = ankylosing spondylitis, CeD = coeliac disease, CD = Crohn's disease, IBD = inflammatory bowel disease, MS = multiple sclerosis, PS = psoriasis, RA = rheumatoid arthritis and UC = ulcerative colitis.

[a] All chromosomal positions are based on UCSC/hg19 coordinates.

[b] Risk alleles are as reported in the original publications.

[c] The direction of the eQTL effect by the risk allele.