Leveraging the resolution of RNA-Seq markedly increases the number of causal eQTLs and candidate genes in human autoimmune disease

Mapping eQTLs in autoimmune disease using RNA-Seq

Christopher A. Odhams[1], Deborah S. Cunninghame Graham[1,2], Timothy J. Vyse[1,2*]

[1] Department of Medical & Molecular Genetics, King's College London, London, UK

[2] Academic Department of Rheumatology, Division of Immunology, Infection and Inflammatory Disease, King's College London, London, UK

* Corresponding author

Email: timothy.vyse@kcl.ac.uk (TJV)

Abstract

Genome-wide association studies have identified hundreds of risk loci for autoimmune disease, yet only a minority (~25%) share a single genetic effect with changes to gene expression (eQTLs) in primary immune cell types. RNA-Seq based quantification at whole-gene resolution, where abundance is estimated by culminating expression of all transcripts or exons of the same gene, is likely to account for this observed lack of colocalisation as subtle isoform switches and expression variation in independent exons are concealed. We perform integrative *cis*-eQTL analysis using association data from twenty autoimmune diseases (846 SNPs; 584 independent loci), with RNA-Seq expression from the GEUVADIS cohort profiled at gene-, isoform-, exon-, junction-, and intron-level resolution. After testing for a shared causal variant, we found exon-, and junction-level analyses produced the greatest frequency of candidate-causal *cis*-eQTLs; many of which were concealed at whole-gene resolution. In fact, only 9% of autoimmune loci shared a disease-relevant eQTL effect at gene-level. Expression profiling at all resolutions however was necessary to capture the full array of eQTL associations, and by doing so, we found 45% of loci were candidate-causal *cis*-eQTLs. Our findings are provided as a web resource for the functional annotation of autoimmune disease association studies (www.insidegen.com). As an example, we dissect the genetic associations of Ankylosing Spondylitis as only a handful of loci have documented causative relationships with gene expression. We classified fourteen of the thirty-one associated SNPs as candidate-causal *cis*-eQTLs. Many of the newly implicated genes had direct relevance to inflammation through regulation of TNF signalling (for example *NFATC2IP*, *PDE4A*, and *RUSC1*), and were supported by integration of functional genomic data from epigenetic and chromatin interaction studies. We have provided a deeper mechanistic understanding of the genetic regulation of gene expression in autoimmune disease by profiling the transcriptome at multiple resolutions.

Author Summary

It is now well acknowledged that non-coding genetic variants contribute to susceptibility of autoimmune disease through alteration of gene expression levels (eQTLs). Identifying the variants that are causal to both disease risk and changes to expression levels has not been easy and we believe this is in part due to how expression is quantified using RNA-Sequencing (RNA-Seq). Whole-gene expression, where abundance is estimated by culminating expression of all transcripts or exons of the same gene, is conventionally used in eQTL analysis. This low resolution may conceal subtle isoform switches and expression variation in independent exons. Using isoform-, exon-, and junction-level quantification can not only point to the candidate genes involved, but also the specific transcripts implicated. We make use of existing RNA-Seq expression data profiled at gene-, isoform-, exon-, junction-, and intron-level, and perform eQTL analysis using association data from twenty autoimmune diseases. We find exon-, and junction-level thoroughly outperform gene-level analysis, and by leveraging all five quantification types, we find 45% of autoimmune loci share a single genetic effect with gene expression. We highlight that existing and new eQTL cohorts using RNA-Seq should profile expression at multiple resolutions to maximise the ability to detect causal eQTLs and candidate-genes.

Introduction

The autoimmune diseases are a family of heritable, often debilitating, complex disorders whereby immune system dysfunction leads to loss of tolerance to self-antigens and chronic inflammation [1]. Genome-wide association studies (GWAS) have now detected hundreds of susceptibility loci contributing to risk of autoimmunity [2] yet their biological interpretation still remains challenging [3]. Mapping single nucleotide polymorphisms (SNPs) that influence gene expression (eQTLs) can provide crucial insight into the potential candidate genes and etiological pathways connected to discrete disease phenotypes [4]. For example, such analyses have implicated dysregulation of autophagy in Crohn's Disease [5], the pathogenic role of $CD4^+$ effector memory T-cells in Rheumatoid Arthritis [6], and an overrepresentation of transcription factors in Systemic Lupus Erythematosus [7].

Expression profiling in appropriate cell types and physiological conditions is necessary to capture the pathologically relevant regulatory changes driving disease risk [8]. Lack of such expression data is thought to explain the observed disparity of shared genetic architecture between disease association and gene expression at certain autoimmune loci [9]. A much overlooked cause of this disconnect however, is not only the use of microarrays to profile gene expression, but also the resolution to which expression is quantified using RNA-Sequencing (RNA-Seq) [10]. Expression estimates of whole-genes, individual isoforms and exons, splice-junctions, and introns are obtainable with RNA-Seq [11–18]. The SNPs that affect these discrete units of expression vary strikingly in their proximity to the target gene, localisation to specific epigenetic marks, and effect on translated isoforms [18]. For example, in over 57% of genes with both an eQTL influencing overall gene expression and a transcript ratio QTL (trQTL) affecting the ratio of each transcript to the gene total, the causal variants for each effect are independent and reside in distinct regulatory elements of the genome [18].

RNA-Seq based eQTL investigations that solely rely on whole-gene expression estimates are likely to mask the allelic effects on independent exons and alternatively-spliced isoforms [16–19]. This is in

part due to subtle isoform switches and expression variation in exons that cannot be captured at gene-level [20]. Recent evidence also suggests that exon-level based strategies are more sensitive than conventional gene-level approaches, and allow for detection of moderate but systematic changes in gene expression that are not necessarily derived from alternative-splicing events [15,21]. Furthermore, gene-level summary counts can be biased in the direction of extreme exon outliers [21]. Use of isoform-, exon-, and junction-level quantification in eQTL analysis also support the potential to not only point to the candidate genes involved, but also the specific transcripts or functional domains affected [10,18]. This of course facilitates the design of targeted functional studies and better illuminates the causative relationship between regulatory genetic variation and disease. Lastly, though intron-level quantification is not often used in conventional eQTL analysis, it can still provide valuable insight into the role of unannotated exons in reference gene annotations, retained introns, and even intronic enhancers [22,23].

Low-resolution expression profiling with RNA-Seq will impede the subsequent identification of causal eQTLs when applying genetic and epigenetic fine-mapping approaches [24]. In this investigation, we aim to increase our knowledge of the regulatory mechanisms and candidate genes of human autoimmune disease through integration of GWAS and RNA-Seq expression data profiled at gene-, isoform-, exon-, junction-, and intron-level in lymphoblastoid cell lines (LCLs). Our findings are provided as a web resource to interrogate the functional effects of autoimmune associated SNPs (www.insidegen.com), and will serve as the basis for targeted follow-up investigations.

Results

Detection of *cis*-eQTLs and candidate-genes of autoimmune disease using RNA-Seq

Using association data from twenty human autoimmune diseases, we performed integrative *cis*-eQTL analysis in lymphoblastoid cell lines (LCLs) with RNA-Seq expression data profiled at five resolutions: gene-, isoform-, exon-, junction-, and intron-level. We tested for a shared causal variant between disease and expression at each association. The 846 autoimmune-associated SNPs taken forward for analysis are documented in S1 Table and an overview of the analysis pipeline to detect candidate-causal *cis*-eQTLs and eGenes is depicted in Fig 1. Expression targets at each level of RNA-Seq quantification were interrogated in *cis* (+/-1Mb) to the 846 SNPs; comprising a total of 7,969 genes, 28,220 isoforms, 54,043 exons, 49,909 junctions, and 35,662 introns (Fig 2A).

We found that *cis*-eQTL association analysis using exon-, junction-, and intron-level quantification yielded the greatest frequency of significant ($q < 0.05$) *cis*-eQTLs and eGenes (Fig 2B). These findings persisted after testing whether each statistically significant *cis*-eQTL showed strong evidence for colocalisation with the genetic variant underlying the autoimmune disease association ($q < 0.05$ and RTC ≥ 0.95). For clarity, we define such eQTLs as candidate causal *cis*-eQTLs and we define their targets as eGenes (Fig 2C). Exon-level analysis detected the most candidate-causal *cis*-eQTLs (235) and eGenes (233) out of all quantification types, followed by junction- and intron-level quantification. Isoform- and gene-level analysis were thoroughly outperformed, with the latter detecting only 70 candidate-causal *cis*-eQTLs and 65 eGenes. In fact, we observed gene-level quantification presented the greatest dropout of significant *cis*-eQTLs that were candidate causal (Fig 2D). Only 23.8% of significant *cis*-eQTLs were candidate-causal at gene-level compared to 49.8% at exon-level; suggesting that in the autoimmune susceptibility loci tested more strongly associated *cis*-eQTLs are captured by the exon-level analysis and they are distinct from gene-level *cis*-eQTLs. Gene-level analysis under estimated candidate-causal eGenes. Our findings, highlighting the need to profile gene-expression at multiple resolutions, are summarised in Fig 2E.

Profiling at all resolutions is necessary to capture the full array associated *cis*-eQTLs

We pruned the 846 autoimmune associated SNPs using an r^2 cut-off of 0.8 and 100kb limit to create a subset of 584 independent susceptibility loci. By combining all five resolutions of RNA-Seq, we found 267 loci (45.7%) presented a shared genetic effect between disease association and gene expression (Fig 3A). Strikingly, only 9.3% of associated loci shared an underlying causal variant at gene-level, in contrast to the 29.1% classified at exon-level. We mapped the candidate-causal *cis*-eQTLs detected by RNA-Seq back to the diseases to which they are associated (Fig 3B). On average, 47% of associated SNPs per disease were classified as candidate-causal *cis*-eQTLs using all five RNA-Seq quantification types. Interestingly, we observed the diseases that fell most below this average comprised autoimmune disorders related to the gut: celiac disease (29%), inflammatory bowel disease (36%), ulcerative colitis (39%), and Crohn's disease (41%), as well as Type 1 Diabetes (37%). These observations are possibly a result of the cellular expression specificity of associated genes in colonic and pancreatic tissue. This conclusion is supported by the above-average frequency of candidate-causal *cis*-eQTLs detected in Systemic Lupus Erythematosus (50%) and Rheumatoid Arthritis (54%); diseases in which the pathogenic role of B-lymphocytes is well documented [33,34]. We further broke down our results per disease by RNA-Seq quantification type (Fig 3C) and in almost all cases, the greatest frequency of candidate-causal *cis*-eQTLs and eGenes were captured by exon- and junction-level analyses.

By separating candidate-causal *cis*-eQTL associations out by quantification type, we found over half were detected by either exon- or junction-level, and considerable overlap of *cis*-eQTL associations existed between both types (Fig 3D). The greatest correlation of effect sizes (r^2: 0.88) of candidate-causal *cis*-eQTLs between exon- and junction-level (S1 Fig). Strong correlation also existed between the effect sizes of gene- and isoform-level candidate-causal *cis*-eQTLs as expected (r^2: 0.83); yet gene-level analysis detected only 19% of all candidate-causal associations. Gene- and isoform-level analysis did however capture six and eighteen candidate-causal *cis*-eQTLs unique to their

quantification type respectively. Thus, our data suggest that although exon- and junction-level, and to a lesser extent intron-level analysis, capture the majority of candidate-causal *cis*-eQTL associations, it is necessary to prolife gene-expression at all quantification types to avoid misinterpretation of the functional impact of disease associated SNPs.

Web resource for functional interpretation of association studies of autoimmune disease

We provide our data as a web resource (www.insidegen.com) for researchers to lookup candidate-causal *cis*-eQTLs and eGenes of autoimmune diseases detected across the five RNA-Seq quantification types. Data are sub-settable and exportable by SNP ID, gene, RNA-Seq resolution, genomic position, and association to specific autoimmune diseases. Full data are also made available in S2 Table.

Functional dissection of Ankylosing Spondylitis genetic associations using RNA-Seq

We decided to apply the results of our integrative *cis*-eQTL analysis to functionally dissect the genetic associations of ankylosing spondylitis (AS). By doing so, we highlight the necessity of profiling at all resolutions of RNA-Seq to shed light on novel regulatory variants, candidate genes, and molecular pathways involved in pathogenesis. AS is a heritable inflammatory arthritis with a largely unexplained genetic contribution outside of the *HLA-B*27* allele (> 30 risk loci) [35,36]. Only a handful of loci show causative relationships with changes in gene expression [35,36]. Candidate-genes implicated by association studies however suggest discrete immunological processes such as antigen presentation, lymphocyte differentiation and activation, and regulation of the TNF/NF-κB signalling pathways are involved, and of note, strong genetic overlap exists with psoriasis, psoriatic arthritis, and inflammatory bowel disease; indicating the pathogenesis of these diseases are tightly connected [37].

Of the 31 AS associated SNPs taken forward for functional interrogation, 14 were classified as candidate-causal *cis*-eQTLs (Fig 4A; full results found at www.insidegen.com for all diseases). We

replicated the association of risk allele rs4129267 [C] with expression reduction of *IL6R* by junction-level analysis (β = -0.36; P = 1.14 x 10^{-06}) [35]. Interestingly, we found the expression of neighbouring gene, *RUSC1*, is also influenced by candidate-causal *cis*-eQTL rs4129267 where the risk allele was also reduced expression of *RUSC1* at exon-level (β = -0.24; P = 1.59 x 10^{-03}). *RUSC1* is able to polyubiquitinate *IKBKG*, a key regulator of NF-κB [38].

The effect of independently associated variants within the 5q15 locus on the expression of aminopeptidase genes *ERAP1* and *ERAP2* was also replicated (S3 Fig) [36]. This includes the association of risk allele rs30187 [T] with increased expression of *ERAP1* (β = -1.09; P = 1.60 x 10^{-71}), and the striking effect of protective allele rs2910686 [T] on the near-complete loss of *ERAP2* (β = -1.37; P = 1.95 x 10^{-175}). Again however, additional genes at this locus with no previous association to expression changes with regards to AS risk alleles were detected. *LNPEP* also belongs to the endoplasmic reticulum aminopeptidase family and has been shown to regulate the NF-κB pathway and antigen presentation via peptide trimming [39]. Interestingly, a missense variation in this gene is linked to psoriasis and is down-regulated in psoriatic lesions relative to healthy skin [40]. We found at junction-level, AS risk allele rs2910686 [C] also contributes to expression reduction of *LNPEP* (β = -0.41; P = 3.09 x 10^{-08}). Similarly, the risk allele rs30187 [T] correlated strongly with decreased expression of *CAST* (β = -0.46; P = 2.47 x 10^{-10}); encoding calpastatin, a calcium-dependent cysteine protease inhibitor. Cysteine protease activity positively correlates with the severity of arthritic lesions and degree of inflammation [41]. Our data support the notion of multiple functional effects at this locus and suggests novel pathological mechanisms including decreased expression of the inhibitor *CAST* leading to increased cysteine protease activity.

Other AS susceptibility loci contributing to expression modulation of multiple genes include rs9901869 for *TBKBP1* and *ITGB3*, and rs75301646 for *NFATC2IP* and *TUFM*. Candidate genes at the rs9901869 locus are yet to be functionally characterised [35]. Our data suggest the risk allele rs9901869 [A] increases the expression of both *TBKBP1*, which plays an active role in the NF-κB and IFN-α signalling pathways (β = 0.57; P = 2.47 x 10^{-17}) [42], and *ITGB3* - involved in the intestinal immune pathway for IgA production (β = 0.35; P = 1.53 x 10^{-6}) [43]. Similarly, we found the risk allele rs9901869 [A] increases expression of both novel candidate-causal eGenes *NFATC2IP* (β =

0.25; P = 6.18 x 10^{-4}) and *TUFM* (β = 0.60; P = 2.82 x 10^{-17}). *TUFM* has been reported as the causative gene at this locus for early onset inflammatory bowel disease [44], whereas *NFATC2IP* (Nuclear Factor of Activated T-cells 2 Interacting Protein) has clear immunological roles in the induction of IL-4 production and regulation of the TNF receptor family of proteins [45].

Our analysis has shed new light on the molecular genetics of AS and can be used in similar manner for the functional dissection of the remaining 19 autoimmune diseases (www.insidegen.com).

Functional genomic support for candidate-causal *cis*-eQTLs

The resolution of RNA-Seq can be leveraged to map candidate-genes and isolate specific exons and junctions perturbed by disease-associated variants. Functional genomic data can then be used to support potential causal associations to deduce molecular mechanisms and epigenetically prioritize causal variants.

The remaining AS associated variant rs1128905 is a candidate-causal *cis*-eQTL for both *CARD9* and *SNAPC4* (Fig 4A). The candidate-gene at this AS locus is thought be to *CARD9* [36]. Our results also draw attention to *SNAPC4* (Small Nuclear RNA-activating Complex Polypeptide 4). Using exon-level RNA-Seq, the risk allele rs1128905 [C] decreased the expression of exons 18 (β = -0.37; P = 3.65 x 10^{-07}) and 19 (β = -0.27; P = 4.80 x 10^{-04}) of the canoncical transcript of *SNAPC4* (Fig 4B). Accordingly, the exon 18-19 boundry was also significantly decreased, captured by junction-level quantifiation (β = -0.26; P = 2.74 x 10^{-04}). As rs1128905 lies over 39kb away from the transcription start site of *SNAPC4*, we used existing promoter capture Hi-C data in lymphoblastoid cell lines to assess whether rs1128905 and associated SNPs may act distally upon *SNAPC4* to influence its expression [32]. We found the bait region encompassing rs1128905 interacts with five targets with great confidence CHiCAGO score > 12 (Fig 4C) [46]. Four of these are located within the *SNAPC4* gene itself. Adding further evidence from histone marks from lymphoblastoid cell lines from the RoadMap Epigenomics Project [31], we found two SNPs in near-perfect LD with rs1128905 (r^2 > 0.95; rs10870201 and rs10870202) were localised to the peaks of H3K4me3, H3K27ac, and H3K9ac marks, and the region encompassed is predicted to be an active enhancer (S4 Fig). Our data therefore

suggest that associated SNPs rs10870201 and rs10870202 may perturb the enhancer-promoter interaction with *SNAPC4* affecting expression. In fact, rs10870201 was the best *cis*-eQTL in the 1Mb region for exons 18 and 19 of *SNAPC4*. Interestingly, although no autoimmune phenotype has been documented with *SNAPC4*, an uncorrelated SNP rs10781500 (r^2 with rs1128905 < 0.5), associated with Crohn's Disease, inflammatory bowel disease, and ulcerative colitis, has also been classified as a candidate-causal *cis*-eQTL for *SNAPC4* but not *CARD9* in *ex vivo* human B-lymphocytes (the risk allele is also correlated with reduced expression of *SNAPC4*) [47]. This effect holds true in our analysis - rs10781500 is an eQTL for *SNAPC4* but not *CARD9*.

Our data point to candidate genes and molecular mechanisms but further functional characerization is of course necessary to determine the true causative gene(s) at this locus.

Detection of autoimmune associated *trans*-eQTLs using RNA-Seq

We extended our RNA-Seq based eQTL investigation to include expression targets > 5Mb away from each of the 846 lead autoimmune GWAS variants (S3 Table). Though we were relatively underpowered for a *trans*-eQTL analysis, we were able to detect 26 *trans*-eQTLs at isoform-level, eight at exon-level, six at gene-level, three at junction-level (Fig 5A). Many of the *trans*-eQTLs detcted were only associated with one eGene, and no *trans*-eQTLs were detected at intron-level. With exon-level quantification however, we were able to identify an interesting effect of *trans*-eQTL rs7726414 - associated with Systemic Lupus Erythematosus (SLE) [7]. We found rs7726414, was a *trans*-eQTL for eight eGenes (Fig 5B). These comprise *SIPA1L2*, *PDPK1*, *IVNS1ABP*, *HES2*, *JAZF1*, *ULK4*, *RP11-51F16.8*, and *PPM1M*. We found the risk allele rs7726414 [T] was associated with increased expression of each of these eight genes (Fig 5C). We highlight, *PDPK1*, a key regulator of *IRF4* and inducer of apoptosis [48], and *JAZF1* which is genetically associated with many autoimmune diseases including SLE itself [7]. The serine/threonine-protein kinase *ULK4* is also of interest as its family member, *ULK3*, is also an SLE suseptability gene [7]. Though we did not classify rs7726414 as a candidate-causal *cis*-eQTL in our dataset, it has been documented as candidate-causal in SLE using a larger eQTL cohort profiled in lymphoblastoid cell lines for eGenes

TCF7 (Transcription Factor 7, T-Cell Specific, HMG-Box) and the ubiquitin ligase complex *SKP1* [10].

Discussion

Elucidation of the functional consequence of non-coding genetic variation in human disease is a major objective of medical genomics [49]. Integrative studies that map disease-associated eQTLs in relevant cell types and physiological conditions are proving essential in progression towards this goal through identification of causal SNPs, candidate-genes, and illumination of molecular mechanisms [50]. In autoimmune disease, where there is considerable overlap of immunopathology, integrative eQTL investigations have been able to connect discrete aetiological pathways, cell types, and epigenetic modifications, to particular clinical manifestations [2,50–52]. Emerging evidence however suggests that only a minority (~25%) of autoimmune associated SNPs share casual variants with *cis*-eQTLs in primary immune cell-types [9].

Genetic variation can influence expression at every stage of the gene regulatory cascade - from chromatin dynamics, to RNA folding, stability, and splicing, and protein translation [53]. As RNA-Seq becomes the convention for genome-wide transcriptomics, be it for differential expression or eQTL analysis, it is essential to maximise the ability to resolve and quantify discrete transcriptomic features. It is now well documented that SNPs affecting these units of expression vary strikingly in their genomic location and localisation to specific epigenetic marks [18]. The reasoning for our investigation therefore was to delineate the limits of microarray and RNA-Seq based eQTL cohorts in the functional annotation of autoimmune disease association signals. To map autoimmune disease associated *cis*-eQTLs, we interrogated RNA-Seq expression data profiled at gene-, isoform, exon-, junction-, and intron-level, and tested for a shared genetic effect at each significant association. We found exon- and junction-level quantification led to the greatest frequency of candidate-causal *cis*-eQTL and eGenes, and thoroughly outperformed gene-level analysis (Fig 2C). We argue however that it is necessary to profile expression at all possible resolutions to diminish the likelihood of overlooking potentially causal *cis*-eQTLs (Fig 3D). In fact, by combining our results across all resolutions, we found 45% of autoimmune loci were candidate-causal *cis*-eQTLs for at least one

eGene. Our findings can be used as a resource to lookup causal eQTLs and candidate genes of autoimmune disease (www.insidegen.com).

Gene-level expression estimates can generally be obtained in two ways – union-exon based approaches [14,17] and transcript-based approaches [11,12]. In the former, all overlapping exons of the same gene are merged into union exons, and intersecting exon and junction reads (including split-reads) are counted to these pseudo-gene boundaries. Using this counting-based approach, it is also possible to quantify meta-exons and junctions easily and with high confidence by preparing the reference annotation appropriately [13,15,54]. Introns can be quantified in a similar manner by inverting the reference annotation between exons and introns [18]. Conversely, transcript-based approaches make use of statistical models and expectation maximization algorithms to distribute reads among gene isoforms - resulting in isoform expression estimates [11,12]. These estimates can then be summed to obtain the entire expression estimate of the gene. Greater biological insight is gained from isoform-level analysis; however, disambiguation of specific transcripts is not trivial due to substantial sequence commonality of exons and junctions. In fact, we found only 15% of autoimmune loci shared a causal variant at transcript-level (Fig 3A). The different approaches used to estimate expression can also lead to significant differences in the reported counts. Union-based approaches, whilst computationally less expensive, can underestimate expression levels relative to transcript-based, and this difference becomes more pronounced when the number of isoforms of a gene increases, and when expression is primarily derived from shorter isoforms [20]. The GEUVADIS study implemented a transcript-based approach to obtain whole-gene expression estimates. A gold standard of eQTL mapping using RNA-Seq is essential therefore for comparative analysis across datasets.

Our findings support recent evidence that suggests exon-level based strategies are more sensitive and specific than conventional gene-level approaches [21]. Subtle isoform variation and expression of less abundant isoforms are likely to be masked by gene-level analysis. Exon-level allows for detection of moderate but systematic changes in gene expression that are not captured at gene-level, and also, gene-level summary counts can be shifted in the direction of extreme exon outliers [21]. It is therefore

important to note that a positive exon-level eQTL association does not necessarily mean a differential exon-usage or splicing mechanism is involved; rather a systematic expression effect across the whole gene may exist that is only captured by the increased sensitivity. Additionally, by combining exon-level with other RNA-Seq quantification types, inferences can be made on the particular isoforms and functional domains affected by the eQTL which can later aid biological interpretation and targeted follow-up investigations [10].

We found intron-level quantification also generated more candidate-causal *cis*-eQTLs than gene-level. As the library was synthesised from poly-A selection, these associations are unlikely due to differences in pre-mRNA abundance. Rather, they are likely derived from either true retained introns in the mature RNA or from coding exons that are not documented in the reference annotation used. We observed multiple instances where a candidate-causal *cis*-eQTL at intron-level was detected, yet a previous investigation had detected an exonic effect using a different reference annotation. For example, an intronic-effect was detected for SLE candidate eGenes *IKZF2* and *WDFY4* in this analysis (which used the GENCODE v12 basic reference annotation). Using the comprehensive reference annotation of GENCODE v12, we found these effects were in fact driven by transcribed exons located within the intronic region of the basic annotation – and were validated *in vitro* by qPCR [10]. The choice of reference annotation therefore has a profound effect on expression estimates [55]; and so again, a gold standard is necessary prevent misinterpretation and increase consistency of eQTL associations.

Lastly, we show how our findings can be leveraged to comprehensively dissect GWAS results of autoimmune diseases. We found 14 of the 31 SNPs associated with Ankylosing Spondylitis (AS) were candidate-causal *cis*-eQTLs for at least on eGene (Fig 4). The majority of these eQTLs influenced the expression of multiple eGenes which had direct relevance to biological pathways associated with autoimmunity. In fact, the majority of the candidate genes detected (for example *RUSC1*, *TBKBP1*, *NFATC2IP*, *TNFRSF1A*, and *PDE4A*) support the involvement of TNF-α and NF-κB in the pathology of AS [35]. We finally show at the *CARD9-SNAPC4* locus, how existing

functional genomic data from chromatin interaction and epigenetic modification experiments can strengthen evidence of the eQTL associations detected by RNA-Seq and allow for functional prioritization of causal variants (Fig 4C). We also highlight the benefit of exon-level analysis to also detect disease associated *trans*-eQTLs (Fig 5).

Taken together, we have provided a deeper mechanistic understanding of the genetic regulation of gene expression in autoimmune disease by profiling the transcriptome at multiple resolutions using RNA-Seq. Similar analyses in new and existing datasets using relevant cell types and context-specific conditions will undoubtedly increase our understanding of how associated variants alter cell physiology and ultimately contribute to disease risk.

Materials and Methods

Autoimmune disease associated SNPs

SNPs were taken from the ImmunoBase resource (www.immunobase.org). It comprises summary case-control association statistics from twenty diseases: twelve originally targeted by the ImmunoChip consortium (Ankylosing Spondylitis, Autoimmune Thyroid Disease, Celiac Disease, Crohn's Disease, Juvenile Idiopathic Arthritis, Multiple Sclerosis, Primary Biliary Cirrhosis, Psoriasis, Rheumatoid Arthritis, Systemic Lupus Erythematosus, Type 1 Diabetes, Ulcerative Colitis), and eight others (Alopecia Areata, Inflammatory Bowel Disease, IgE and Allergic Sensitization, Narcolepsy, Primary Sclerosing Cholangitis, Sjogren Syndrome, Systemic Scleroderma, Vitiligo). For eQTL analysis, we took the lead SNPs for each disease - defined as a genome-wide significant SNP with the lowest reported *P*-value (S1 Table). X-chromosome associations and SNPs with minor allele frequency < 5% were omitted from analysis, leaving 846 SNPs. A total of 262 SNPs were pruned using the '*--indep-pairwise*' function of PLINK 1.9 with a window size of 100kb and an r^2 threshold of 0.8, leaving 584 independent loci.

RNA-Seq gene expression data

Normalised RNA-Seq expression data of 373 lymphoblastoid cell lines from four European sub-populations (CEU, GBR, FIN, TSI) of the 1000Genomes Project (Geuvadis) [18] were obtained from EBI ArrayExpress (E-GEUV-1; full methods can be found in http://geuvadiswiki.crg.es/). In summary, transcripts, splice-junctions, and introns were quantified using Flux Capacitor against the GENCODE v12 basic reference annotation [16]. Reads belonging to single transcripts were predicted by deconvolution per observations of paired-reads mapping across all exonic segments of a locus. Gene-level expression was calculated as the sum of all transcripts per gene. Annotated splice junctions were quantified using split read information, counting the number of reads supporting a given junction. Intronic regions that are not retained in any mature annotated transcript, and reported mapped reads in different bins across the intron to distinguish reads stemming from retained introns

from those produced by not yet annotated exons. Meta-exons were quantified by merging all overlapping exonic portions of a gene into non-redundant units and counting reads within these bins [15]. Reads were excluded when the read pairs map to two different genes. Quantifications were corrected for sequencing depth and gene length (RPKM). Only expression elements quantified in > 50 % of individuals were kept and Probabilistic Estimation of Expression Residuals (PEER) was used to remove technical variation [25].

Cis and *trans*-eQTL analysis

An overview of the integration pipeline is depicted in Fig 1. Genotypes were obtained from EBI ArrayExpress (E-GEUV-1). The 41 individuals genotyped on the Omni 2.5M SNP array were previously imputed to the Phase 1 v3 release as described [18]. PCA of genotype data was performed using the Bioconductor package SNPRelate (S2 Fig) [26]. Only bi-allelic SNPs with MAF > 0.05, imputation call-rates $\geq$ 0.8, and HWE $P < 1 \times 10^{-04}$ were used. All eQTL association testing was performed with a liner-regression model in R. Normalized expression residuals (PEER factor normalized RPKM) for each quantification type were transformed to standard normal and the first three principle components used as covariates in the eQTL model as well as the binary imputation status. *Cis* and *trans*-eQTL mapping was performed for genes within +/-1Mb of the lead SNP and for genes > 5Mb from the lead SNP respectively. Adjustment for multiple testing of eQTL results per quantification type (corrected total genes, isoforms, exons, junctions, and introns) was undertaken using an FDR of 0.05 for *cis* and 0.01 for *trans* analysis (MHC associations were excluded in *trans*).

Analysis of shared causal variant

The Regulatory Trait Concordance (RTC) method was used to assess the likelihood of a shared causal variant between the GWAS SNP and the *cis*-eQTL signal [27]. SNPs were firstly classified according to their position in relation to recombination hotspots (based on genome-wide estimates of hotspot intervals) [28]. For each significant *cis*-eQTL association, the residuals from the linear-regression of the best *cis*-asQTL (lowest association *P*-value within the hotspot interval) was extracted against the

expression quantification for the expression unit in hand. Regression was then performed using all SNPs within the defined hotspot interval against these residuals. The RTC score was then calculated as (N_{SNPs} - $Rank_{GWAS\ SNP}$ / N_{SNPs}. Where N_{SNPs} is the total number of SNPs in the recombination hotspot interval, and $Rank_{GWAS\ SNP}$ is the rank of the GWAS SNP association *P*-value against all other SNPs in the interval from the liner-association against the residuals of the best *cis*-eQTL. Disease associated SNPs with statistically significant association to gene expression ($q < 0.05$) and an RTC score > 0.95 were classified as 'candidate-causal eQTLs'. Genes whose expression is modulated by the eQTL were defined as 'candidate-causal eGenes'.

Data visualisation and resources

R version 3.3.1 was used to create heatmaps, box-plots (ggplot2), and circularized chromosome diagrams (circlize). Genes were plotted in UCSC Genome Browser [29] and IGV [30]. Roadmap epigenetic data were downloaded from the web resource [31], and chromatin interaction data were taken from the CHiCP web resource [32].

Acknowledgements

We thank Dr David L Morris for helpful discussions throughout this work. We also thank Philip Tombleson for his assistance with data uploading.

The GEUVADIS 1000 Genomes RNA-Seq data was downloaded from the EBI ArrayExpress Portal (accession E-GEUV-1).

References

1. Fever FM. NIH Progress in Autoimmune Diseases Research. in National Institute of Health Publication. 2005; 17–7576.
2. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. Nature Publishing Group; 2013;14: 661–73. doi:10.1038/nrg3502
3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106: 9362–9367. doi:10.1073/pnas.0903103106
4. Westra H-J, Franke L. From genome to function by studying eQTLs. Biochim Biophys Acta. Elsevier B.V.; 2014;1842: 1896–1902. doi:10.1016/j.bbadis.2014.04.024
5. Klionsky DJ. Crohn's disease, autophagy, and the Paneth cell. N Engl J Med. 2009;360: 1785–1786. doi:10.1056/NEJMcibr0810347
6. Hu X, Kim H, Raj T, Brennan PJ, Trynka G, Teslovich N, et al. Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4+ Effector Memory T Cells. PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004404
7. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. Nature Publishing Group; 2015;47: 1457–1464. doi:10.1038/ng.3434
8. Fairfax BP, Knight JC. Genetics of gene expression in immunity to infection. Curr Opin Immunol. Elsevier Ltd; 2014;30: 63–71. doi:10.1016/j.coi.2014.07.001
9. Chun S, Casparino A, Patsopoulos NA, Croteau-chonka DC, Raby BA, Jager PL De, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. NatGenet. 2017; doi:10.1038/ng.3795
10. Odhams CA, Cortini A, Chen L, Roberts AL, Viñuela A, Buil A, et al. Mapping eQTLs with RNA-Seq Reveals Novel Susceptibility Genes, Non-Coding RNAs, and Alternative-Splicing

Events in Systemic Lupus Erythematosus. Hum Mol Genet. 2017;26: 1003–1017. doi:10.1093/hmg/ddw417

11. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7: 562–78. doi:10.1038/nprot.2012.016

12. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12: 323. doi:10.1186/1471-2105-12-323

13. Schuierer S, Roma G. The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. Nucleic Acids Res. 2016; gkw538. doi:10.1093/nar/gkw538

14. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

15. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq-npre20126837-2.pdf. Genome Res. 2012;12: 1088–9051. doi:10.1101/gr.133744.111

16. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. Nature Publishing Group; 2010;464: 773–777. doi:10.1038/nature08903

17. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30: 923–930. doi:10.1093/bioinformatics/btt656

18. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen P a C, Monlong J, Rivas M a, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–11. doi:10.1038/nature12531

19. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24: 14–24. doi:10.1101/gr.155192.113

20. Zhao S, Xi L, Zhang B. Union exon based approach for RNA-seq gene quantification: To be

or not to be? PLoS One. 2015;10: e0141910. doi:10.1371/journal.pone.0141910

21. Laiho A, Elo LL. A note on an exon-based strategy to identify differentially expressed genes in RNA-seq experiments. PLoS One. 2014;9: 1–12. doi:10.1371/journal.pone.0115964

22. Gaidatzis D, Burger L, Florescu M, Stadler MB. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat Biotech. Nature Publishing Group; 2015;33: 722–729. doi:10.1038/nbt.3269

23. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5: 621–628. doi:10.1038/nmeth.1226

24. Trynka G, Westra HJ, Slowikowski K, Hu X, Xu H, Stranger BE, et al. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. Am J Hum Genet. The Authors; 2015;97: 139–152. doi:10.1016/j.ajhg.2015.05.016

25. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6: 1–11. doi:10.1371/journal.pcbi.1000770

26. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28: 3326–3328. doi:10.1093/bioinformatics/bts606

27. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010;6: e1000895. doi:10.1371/journal.pgen.1000895

28. McVean GA. The fine-scale structure of recombination rate variation in the human genome. Science (80-). 2004;304: 581. Available: http://dx.doi.org/10.1126/science.1092500

29. Kent WJ, Sugnet CW, Furey TS, Roskin KM. The Human Genome Browser at UCSC W. J Med Chem. 2002;19: 1228–31. doi:10.1101/gr.229102.

30. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

doi:10.1093/bib/bbs017

31. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–330. doi:10.1038/nature14248

32. Schofield EC, Carver T, Achuthan P, Freire-Pritchett P, Spivakov M, Todd JA, et al. CHiCP: A web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics. 2016;32: 2511–2513. doi:10.1093/bioinformatics/btw173

33. Marston B. B cells in the pathogenesis and treatment of rheumatoid arthritis. Curr Opin Rheumatol. 2011;22: 307–315. doi:10.1097/BOR.0b013e3283369cb8.B

34. Dörner T, Giesecke C, Lipsky PE. Mechanisms of B cell autoimmunity in SLE. Arthritis Res Ther. 2011;13: 243. doi:10.1186/ar3433

35. Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, Leo P, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat Genet. 2013;45: 730–8. doi:10.1038/ng.2667

36. Reveille JD, Sims A-M, Danoy P, Evans DM, Leo P, Pointon JJ, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet. 2010;42: 123–7. doi:10.1038/ng.513

37. O'Rielly DD, Uddin M, Rahman P. Ankylosing spondylitis: beyond genome-wide association studies. Curr Opin Rheumatol. 2016;28: 337–45. doi:10.1097/BOR.0000000000000297

38. Napolitano G, Mirra S, Monfregola J, Lavorgna A, Leonardi A, Ursini MV. NESCA: A new NEMO/IKKgamma and TRAF6 interacting protein. J Cell Physiol. 2009;220: 410–417. doi:10.1002/jcp.21782

39. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet. 2009;41: 199–204. doi:10.1038/ng.311

40. Cheng H, Li Y, Zuo X-B, Tang H-Y, Tang X-F, Gao J-P, et al. Identification of a Missense Variant in LNPEP that Confers Psoriasis Risk. J Invest Dermatol. Elsevier Masson SAS; 2013;134: 1–22. doi:10.1038/jid.2013.317

41. Biroc SL, Gay S, Hummel K, Magill C, Palmer JT, Spenc DR, et al. Cysteine protease activity is up-regulated in inflamed ankle joints of rats with adjuvant-induced arthritis and decreases with in vivo administration of a vinyl sulfone cysteine protease inhibitor. Arthritis Rheum. 2001;44: 703–711. doi:10.1002/1529-0131(200103)44:3<703::AID-ANR120>3.0.CO;2-2

42. Goncalves A, Bürckstümmer T, Dixit E, Scheicher R, Górna MW, Karayel E, et al. Functional dissection of the TBK1 molecular network. PLoS One. 2011;6. doi:10.1371/journal.pone.0023971

43. Reynier F, Pachot A, Paye M, Xu Q, Turrel-Davin F, Petit F, et al. Specific gene expression signature associated with development of autoimmune type-I diabetes using whole-blood microarray analysis. Genes Immun. 2010;11: 269–278. doi:10.1038/gene.2009.112

44. Narahara M, Higasa K, Nakamura S, Tabara Y, Kawaguchi T, Ishii M, et al. Large-scale East-Asian eQTL mapping reveals novel candidate genes for LD mapping and the genomic landscape of transcriptional effects of sequence variants. PLoS One. 2014;9. doi:10.1371/journal.pone.0100924

45. Hashiguchi K, Ozaki M, Kuraoka I, Saitoh H. Establishment of a human cell line stably overexpressing mouse Nip45 and characterization of Nip45 subcellular localization. Biochem Biophys Res Commun. Elsevier Inc.; 2013;430: 72–77. doi:10.1016/j.bbrc.2012.11.020

46. Cairns J, Freire-Pritchett P, Wingett SW, Dimond A, Plagnol V, Zerbino D, et al. CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C data. Genome Biol. Genome Biology; 2016;17: 28068. doi:10.1101/028068

47. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. Nat Genet. Nature Publishing Group; 2012;44: 502–510. doi:10.1038/ng.2205

48. Chinen Y, Kuroda J, Shimura Y, Nagoshi H, Kiyota M, Yamamoto-Sugitani M, et al. Phosphoinositide protein kinase PDPK1 is a crucial cell signaling mediator in multiple myeloma. Cancer Res. 2014;74: 7418–7429. doi:10.1158/0008-5472.CAN-14-1420

49. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. Genome Res. 2015;25: 1427–1431. doi:10.1101/gr.190983.115.

50. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. Nature Publishing Group; 2015;16: 197–212. doi:10.1038/nrg3891

51. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. Nature Publishing Group; 2015;518: 337–343. doi:10.1038/nature13835

52. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. Nature Publishing Group; 2013;45: 124–30. doi:10.1038/ng.2504

53. Li YI, Geijn B Van De, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352: 600–4. doi:10.1126/science.aad9417

54. Ongen H, Dermitzakis ET. Alternative Splicing QTLs in European and African Populations. Am J Hum Genet. The Authors; 2015;97: 567–575. doi:10.1016/j.ajhg.2015.09.004

55. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16: 97. doi:10.1186/s12864-015-1308-8

Figure captions

Fig 1. *Cis*-eQTL analysis pipeline to detect candidate-causal eQTLs of autoimmune disease.

The 846 autoimmune disease associated SNPs per disease are documented in S1 Table and were LD pruned to 584 independent loci (see Methods). Genotypes of 1000Genomes individuals were quality controlled and subset to regions of recombination hotspots. If the lead GWAS SNP was found between a recombination hotspot, then all SNPs were between the recombination hotspot intervals were used in the Regulatory Trait Concordance (RTC) analysis. If the lead GWAS SNP was found within a recombination hotspot itself, then all SNPs before or after the summit (including the between summit SNPs) were used in the RTC (upper-interval and lower-interval hotspot respectively). Normalized RNA-Seq expression data at gene-, isoform-, exon-, junction-, and intron-level were obtained for the 1000Genomes individuals of the GEUVAIDS cohort in lymphoblastoid cell lines. Disease associated SNPs with statistically significant association with gene expression ($q < 0.05$) and an RTC score > 0.95 were classified as 'candidate-causal eQTLs'. Genes whose expression is modulated by the eQTL were defined as 'candidate-causal eGenes'.

Fig 2. Frequency of candidate-causal *cis*-eQTLs detected across each RNA-Seq quantification type.

Expression targets correspond to the quantification type under consideration (i.e. number of isoforms captured at isoform-level, the number of exons-captured at exon-level). (A) Number of expression targets and corresponding genes (by referencing the expression target back to the gene it belongs to) interrogated in *cis* (+/-1Mb) to the 846 autoimmune SNPs. (B) Number of significant *cis*-eQTL associations that pass an FDR q-value threshold of 0.05, comprising the number of expression targets, eQTLs, and eGenes. Only unique associations are reported (for example if two independent eQTLs act on the same eGene, the eGene is only counted once). (C) Number of candidate-causal *cis*-eQTLs ($q < 0.05$ and RTC ≥ 0.95). (D) Percentage of statistically significant *cis*-eQTLs ($q < 0.05$) that are candidate-causal ($q < 0.05$ and RTC ≥ 0.95) to show the dropout of *cis*-eQTL associations that do not

appear to share the same causal variant as disease. (E) Percentage of the 846 autoimmune associated SNPs that are candidate-causal *cis*-eQTLs, and the percentage of the 8,927 genes in *cis* that are candidate-causal eGenes.

Fig 3. Breakdown of autoimmune candidate-causal *cis*-eQTLs per RNA-Seq quantification type. (A) Percentage and number of candidate-causal *cis*-eQTLs detected per RNA-Seq quantification type, following LD pruning of associated SNPs to 584 independent susceptibly loci. (B) Total candidate-causal *cis*-eQTLs per disease across all five levels of RNA-Seq quantification (full results in found at www.insidegen.com), using the 20 diseases of the ImmunoBase resource. In orange are disease-associated SNPs that show no shared association with expression across any quantification type. In blue are the disease-associated SNPs that are also candidate-causal *cis*-eQTLs. 47% of SNPs across all diseases were candidate-causal *cis*-eQTLs on average. (C) Candidate-causal *cis*-eQTLs per disease broken down by quantification type. (D) Candidate-causal *cis*-eQTLs detected per quantification type. Percentage of candidate-causal *cis*-eQTLs captured are shown as a percentage of the 362 total.

Fig 4. Functional annotation of Ankylosing Spondylitis risk loci using RNA-Seq. (A) Heatmap of the 14 candidate-causal *cis*-eQTLs and 27 eGenes of Ankylosing Spondylitis detected across the five RNA-Seq quantification types (full results found at www.insidegen.com). Heat is relative to *P*-value of association. To normalize across quantification types, relative significance of each association per column was calculated as the $-\log_2 (P/Pmax)$; where *Pmax* is the most significant association per quantification type. If a candidate-causal association is detected at any level of quantification, it is shown and marked with an *. Associations not marked with an * are not candidate-causal. (B) Isolation of the effect of Ankylosing Spondylitis associated SNP and candidate-causal *cis*-eQTL, rs1128905, on the expression of *SNAPC4*. The risk allele is rs1128905 [C]. At exon-level, rs1128905 is a candidate-causal *cis*-eQTL for exons 18 and 19 of *SNAPC4*, and at junction-level for the exon 18-19 junction. (C) Incorporating existing functional genomic data from promoter capture Hi-C data in lymphoblastoid cell lines [32]. The bait region encompassing rs1128905 interacts

with five targets with great confidence CHiCAGO score > 12 including four interactions with *SNAPC4* which lies ~39kb away.

Fig 5. Autoimmune associated *trans*-eQTLs detected using RNA-Seq.

(A) Number of *trans*-eQTLs and *trans*-eGenes (q < 0.01) detected across all five RNA-Seq quantification types (S3 Table). (B) Genome-wide depiction of *trans*-eQTL rs7726414, which is associated with eight genes in *trans* detected using exon-level RNA-Seq. (C) Box-plots of rs7726414 on the eight *trans*-genes, the risk allele is rs7726414 [T].

Supporting information

S1 Table. Disease associated SNPs from the ImmunoBase resource taken forward for eQTL analysis. Only autosomal SNPs with MAF > 5% included (noted with 'y'). SNPs are classified as being associated with the 20 immune-related diseases with 'y' (associated) or 'n' (not associated). Abbreviations for diseases are taken from ImmunoBase (www.immunobase.org).

S2 Table. Candidate-causal *cis*-eQTLs detected across all five RNA-Seq quantification types. See tab 2 for a definition of all column headers. Only candidate-causal ($q < 0.05$ and RTC ≥ 0.95) are shown. Data are also accessible through the web-portal (www.insidegen.com).

S3 Table. *Trans*-eQTLs detected across all five RNA-Seq quantification types.

S1 Fig. Correlation of SNP-Gene association pairs across RNA-Seq quantification types. The bottom panel shows the correlation coefficient of the effect sizes (beta) of candidate-causal *cis*-eQTL associations across each RNA-Seq quantification type. The top panel shows the same data but is adjusted so to force the same direction of effect.

S2 Fig. Processing of genotype data and principle component analysis. Genotype data in VCF format of 1000Genomes individuals were downloaded from E-GEUV1 (ArrayExpress). Insertion-deletion sites were removed, and bi-allelic SNPs kept only. SNPs with HWE < 0.0001 were removed and the VCF converted to 0,1,2 format using PLINK. Principle component analysis was performed on genotype data using the R package SNPRelate on chromosome 20. The first 3 components were included in the eQTL regression model as well as the binary imputation status (see methods).

S3 Fig. Replication of *cis*-eQTL effect of rs30187 on *ERAP1* and rs2910686 on *ERAP2* expression-levels using gene-level quantification.

S4 Fig. Functional prioritization of the rs1128905 (*CARD9-SNAPC4*) locus. rs1128905, associated with Ankylosing Spondylitis, was found to be a candidate-causal *cis*-eQTL for *CARD9* and *SNAPC4*. The rs1128905 locus interacts with the *SNAPC4* locus (~39kb away) via a chromatin interaction in lymphoblastoid cell lines (Fig 4C). To find potential causal SNPs, we took all associated SNPs with in strong LD $r^2 > 0.8$ with rs1128905, and looked for colocalisation with genome-wide epigenetic marks from histone and DNase from the ENCODE project in lymphoblastoid cell lines (GM12878). Two SNPs, rs10870201 and rs10870202, were found to be in the peak summit of enhancer/promoter histone marks H3K4me3, H3K27ac, and H3K9ac, as well as in a region of DNase hypersensitivity. We hypothesize variation at these SNPs causes loss of interaction with *SNAPC4*, which reduces the expression of *SNAPC4* (Fig 4B).

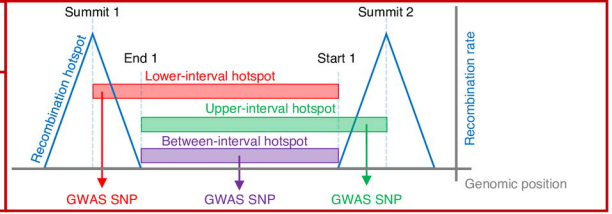
1. Autoimmune disease-associated SNPs

846 lead-SNP associations (584 independent loci), passing genome-wide significance curated by the ImmunoBase Cohort.

2. Genotype

1000 Genomes
373 Europeans (CEU, FIN, GBR, TSI)
SNPS: biallelic, HWE p > 0.0001, MAF ≥ 0.05

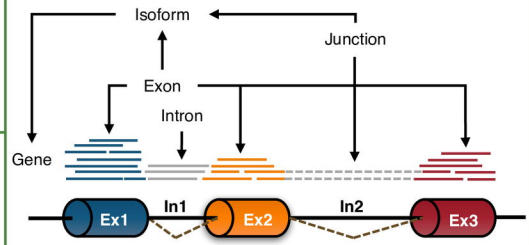
Subset and retrieve all SNPs within hotspot intervals around each GWAS SNP.
846 independent SNPs in **451** unique hotspots, corresponding to a total of **137,509** SNPs.



3. Expression

GEUVADIS Consortium
RNA-Seq of 373 Europeans
Lymphoblastoid cell lines
QC passed (expressed in > 50% of individuals)
PEER Factor normalized and SN transformed RPKM
Subset to *cis*-intervals around each GWAS SNP (+/-1Mb)

Quantification of expression targets in *cis* (+/-1Mb)
Genes (7,969 genes)
Isoforms (28,220 isoforms, 7,636 genes)
Exons (54,043 exons, 5,621 genes)
Junctions (49,909 junctions, 4,366 genes)
Introns (35,662 introns, 6,443 genes)

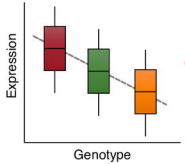


4. *Cis*-eQTL association analysis

Per quantification type, perform linear regression of all hotspot SNPs against all *cis* (+/-1Mb) normalized expression targets. Include four genotype principle components in regression model.

5. Test for shared causal variant

1) Identify best hotspot *cis*-eQTL by linear regression for each unit:



2) Perform linear regression of all hotspot *cis*-eQTLs against residuals of best then rank:

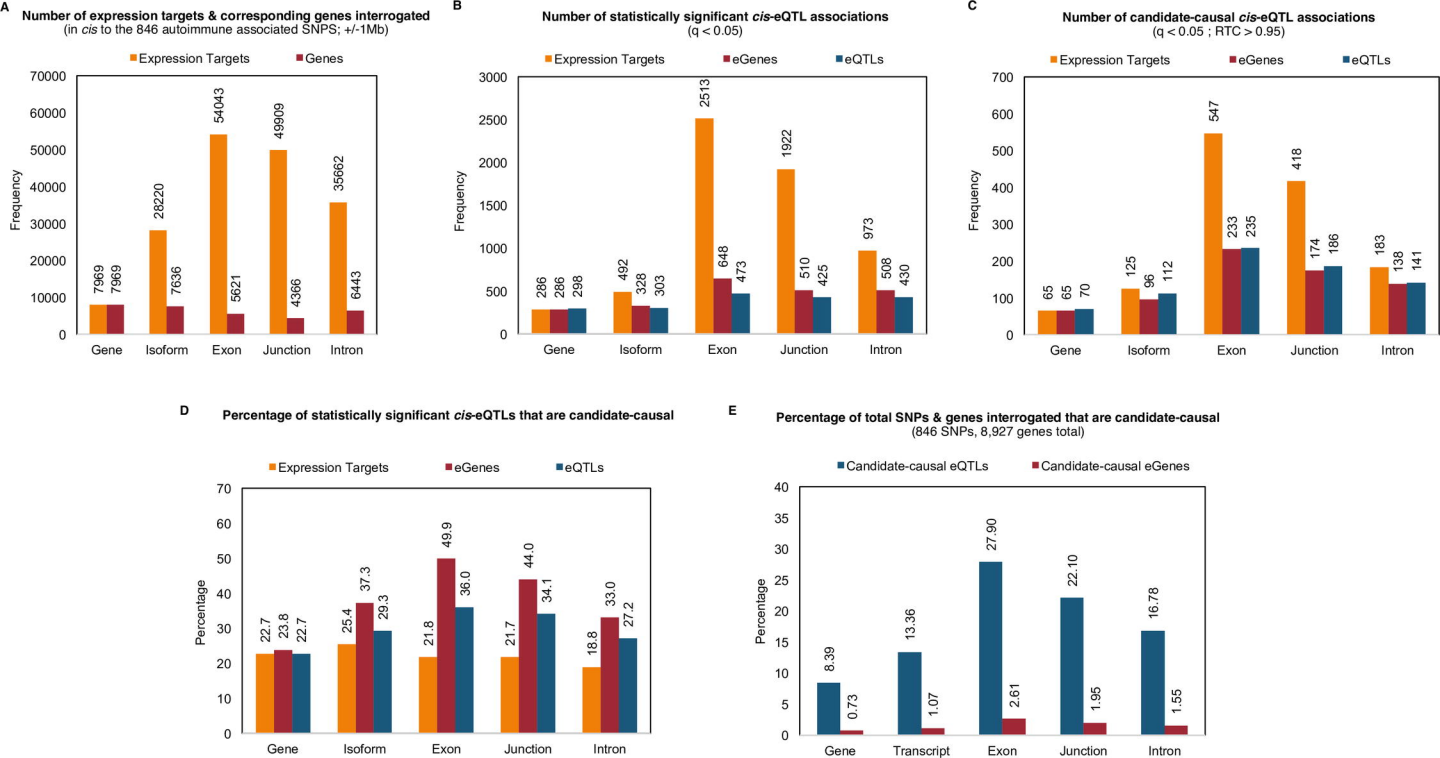
SNP	*P*-val	Rank
Best	*Least sig.*	*1*
GWAS	…	…
SNP_i	…	…
SNP_j	*Most sig.*	N_{SNPs}

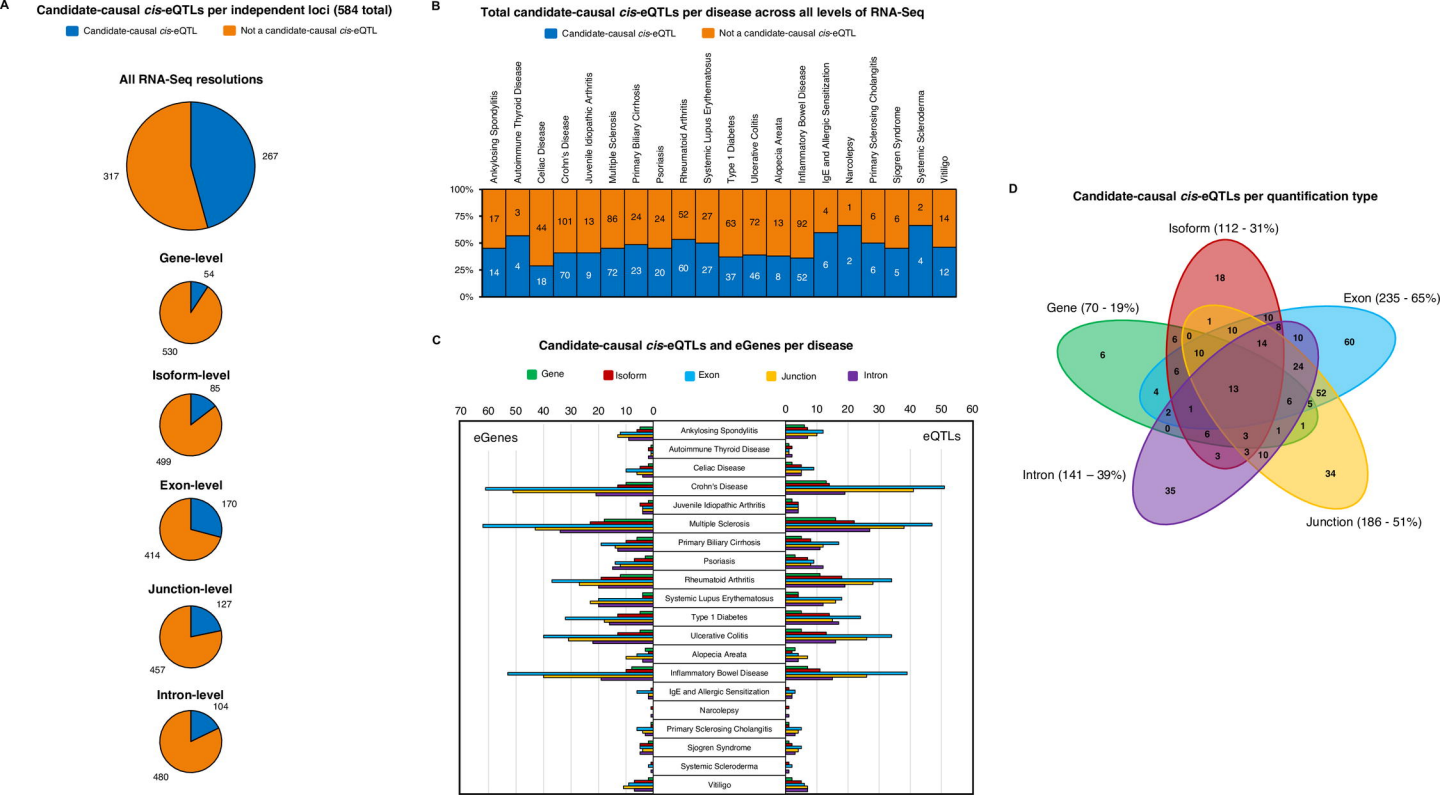
3) Calculate RTC:

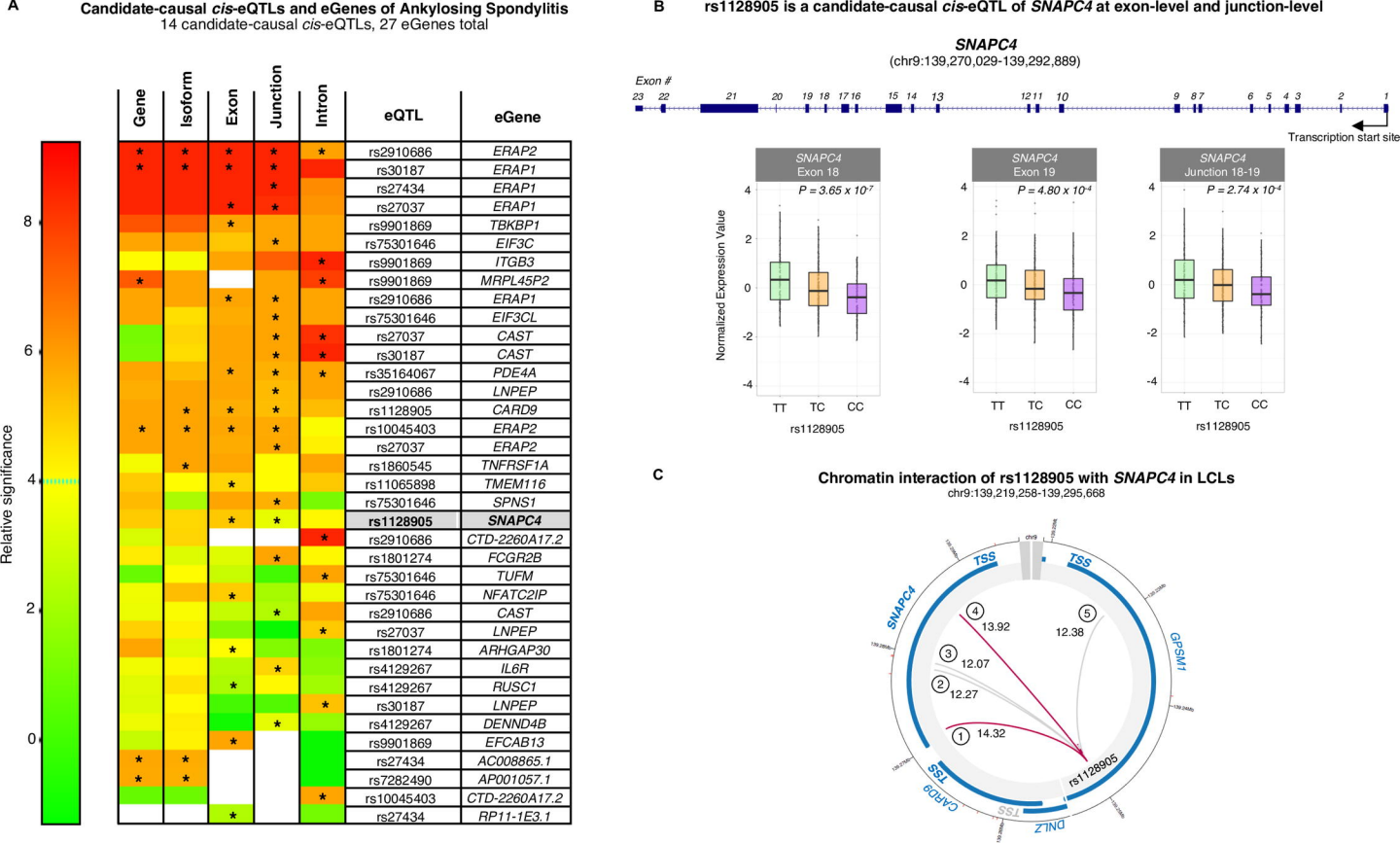
$$RTC = \frac{N_{SNPs} - Rank_{GWAS\ SNP}}{N_{SNPs}}$$

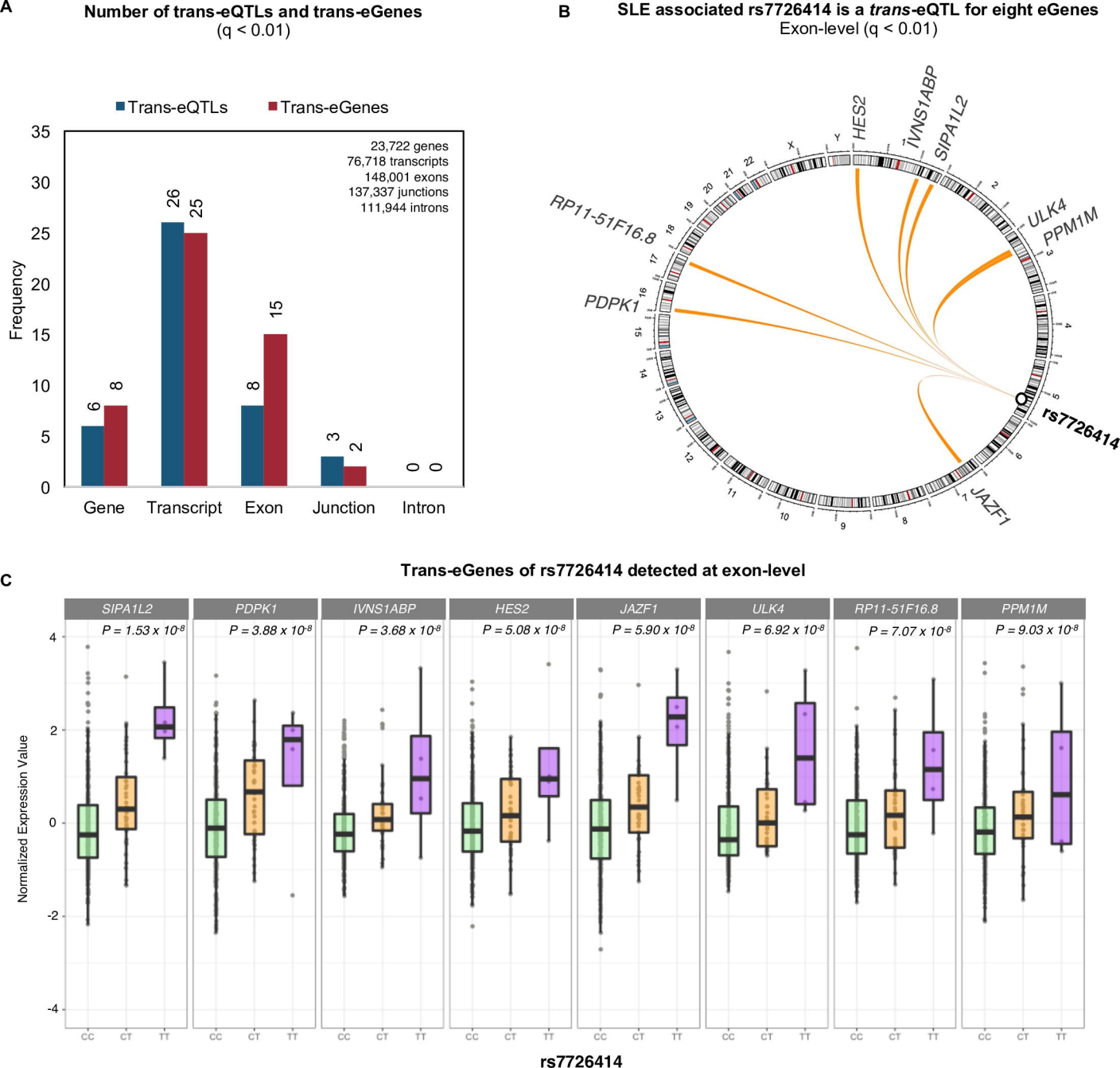
6. Define candidate-causal eQTLs and eGenes

FDR adjustment per quantification type.
q < 0.05
RTC ≥ 0.95


A
Number of expression targets & corresponding genes interrogated
(in cis to the 846 autoimmune associated SNPS; +/-1Mb)
Expression Targets
Genes
Frequency
Gene
Isoform
Exon
Junction
Intron
7969
7969
28220
7636
54043
5621
49909
4366
35662
6443
B
Number of statistically significant cis-eQTL associations
(q < 0.05)
Expression Targets
eGenes
eQTLs
Frequency
286
286
298
492
328
303
2513
648
473
1922
510
425
973
508
430
C
Number of candidate-causal cis-eQTL associations
(q < 0.05 ; RTC > 0.95)
Expression Targets
eGenes
eQTLs
Frequency
65
65
70
125
96
112
547
233
235
418
174
186
183
138
141
D
Percentage of statistically significant cis-eQTLs that are candidate-causal
Expression Targets
eGenes
eQTLs
Percentage
22.7
23.8
22.7
25.4
37.3
29.3
21.8
49.9
36.0
21.7
44.0
34.1
18.8
33.0
27.2
E
Percentage of total SNPs & genes interrogated that are candidate-causal
(846 SNPs, 8,927 genes total)
Candidate-causal eQTLs
Candidate-causal eGenes
Percentage
8.39
0.73
13.36
1.07
27.90
2.61
22.10
1.95
16.78
1.55
Gene
Transcript
Exon
Junction
Intron


A
Candidate-causal cis-eQTLs per independent loci (584 total)
Candidate-causal cis-eQTL
Not a candidate-causal cis-eQTL
All RNA-Seq resolutions
317
267
Gene-level
54
530
Isoform-level
85
499
Exon-level
170
414
Junction-level
127
457
Intron-level
104
480
B
Total candidate-causal cis-eQTLs per disease across all levels of RNA-Seq
Candidate-causal cis-eQTL
Not a candidate-causal cis-eQTL
Ankylosing Spondylitis
Autoimmune Thyroid Disease
Celiac Disease
Crohn's Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Systemic Lupus Erythematosus
Type 1 Diabetes
Ulcerative Colitis
Alopecia Areata
Inflammatory Bowel Disease
IgE and Allergic Sensitization
Narcolepsy
Primary Sclerosing Cholangitis
Sjogren Syndrome
Systemic Scleroderma
Vitiligo
100%
75%
50%
25%
0%
17 3 44 101 13 86 24 24 52 27 63 72 13 92 4 1 6 6 2 14
14 4 18 70 9 72 23 20 60 27 37 46 8 52 6 2 6 5 4 12
C
Candidate-causal cis-eQTLs and eGenes per disease
Gene
Isoform
Exon
Junction
Intron
70 60 50 40 30 20 10 0
0 10 20 30 40 50 60
eGenes
eQTLs
Ankylosing Spondylitis
Autoimmune Thyroid Disease
Celiac Disease
Crohn's Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Systemic Lupus Erythematosus
Type 1 Diabetes
Ulcerative Colitis
Alopecia Areata
Inflammatory Bowel Disease
IgE and Allergic Sensitization
Narcolepsy
Primary Sclerosing Cholangitis
Sjogren Syndrome
Systemic Scleroderma
Vitiligo
D
Candidate-causal cis-eQTLs per quantification type
Isoform (112 - 31%)
Gene (70 - 19%)
Exon (235 - 65%)
Intron (141 – 39%)
Junction (186 - 51%)
18
6
60
34
35
1
10
8
10
14
10
6
0
10
6
24
13
4
52
6
5
2
1
1
0
6
3
1
3
3
10


A
Candidate-causal cis-eQTLs and eGenes of Ankylosing Spondylitis
14 candidate-causal cis-eQTLs, 27 eGenes total
Relative significance
Gene
Isoform
Exon
Junction
Intron
eQTL
eGene
rs2910686 ERAP2
rs30187 ERAP1
rs27434 ERAP1
rs27037 ERAP1
rs9901869 TBKBP1
rs75301646 EIF3C
rs9901869 ITGB3
rs9901869 MRPL45P2
rs2910686 ERAP1
rs75301646 EIF3CL
rs27037 CAST
rs30187 CAST
rs35164067 PDE4A
rs2910686 LNPEP
rs1128905 CARD9
rs10045403 ERAP2
rs27037 ERAP2
rs1860545 TNFRSF1A
rs11065898 TMEM116
rs75301646 SPNS1
rs1128905 SNAPC4
rs2910686 CTD-2260A17.2
rs1801274 FCGR2B
rs75301646 TUFM
rs75301646 NFATC2IP
rs2910686 CAST
rs27037 LNPEP
rs1801274 ARHGAP30
rs4129267 IL6R
rs4129267 RUSC1
rs30187 LNPEP
rs4129267 DENND4B
rs9901869 EFCAB13
rs27434 AC008865.1
rs7282490 AP001057.1
rs10045403 CTD-2260A17.2
rs27434 RP11-1E3.1
B
rs1128905 is a candidate-causal cis-eQTL of SNAPC4 at exon-level and junction-level
SNAPC4
(chr9:139,270,029-139,292,889)
Exon #
Transcription start site
SNAPC4 Exon 18
P = 3.65 x 10-7
SNAPC4 Exon 19
P = 4.80 x 10-4
SNAPC4 Junction 18-19
P = 2.74 x 10-4
Normalized Expression Value
TT TC CC
rs1128905
C
Chromatin interaction of rs1128905 with SNAPC4 in LCLs
chr9:139,219,258-139,295,668
SNAPC4
GPSM1
CARD9
DNLZ
TSS
rs1128905
13.92
12.38
12.07
12.27
14.32


A
Number of trans-eQTLs and trans-eGenes
(q < 0.01)
Trans-eQTLs
Trans-eGenes
23,722 genes
76,718 transcripts
148,001 exons
137,337 junctions
111,944 introns
Frequency
Gene
Transcript
Exon
Junction
Intron
B
SLE associated rs7726414 is a trans-eQTL for eight eGenes
Exon-level (q < 0.01)
HES2
IVNS1ABP
SIPA1L2
ULK4
PPM1M
RP11-51F16.8
PDPK1
rs7726414
JAZF1
C
Trans-eGenes of rs7726414 detected at exon-level
SIPA1L2
PDPK1
IVNS1ABP
HES2
JAZF1
ULK4
RP11-51F16.8
PPM1M
P = 1.53 x 10-8
P = 3.88 x 10-8
P = 3.68 x 10-8
P = 5.08 x 10-8
P = 5.90 x 10-8
P = 6.92 x 10-8
P = 7.07 x 10-8
P = 9.03 x 10-8
Normalized Expression Value
CC
CT
TT
rs7726414