Profiling RNA-Seq at multiple resolutions markedly increases the number of causal eQTLs in 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 genetic effects with changes to gene expression (eQTLs) in immune cells. 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 can be concealed. We performed integrative *cis*-eQTL analysis using association statistics from twenty autoimmune diseases (560 independent loci) and RNA-Seq data from 373 individuals of the Geuvadis cohort profiled at gene-, isoform-, exon-, junction-, and intron-level resolution in lymphoblastoid cell lines. After stringently testing for a shared causal variant using both the Joint Likelihood Mapping and Regulatory Trait Concordance frameworks, we found that gene-level quantification significantly underestimated the number of causal *cis*-eQTLs. Only 5.0-5.3% of loci were found to share a causal *cis*-eQTL at gene-level compared to 12.9-18.4% at exon-level and 9.6-10.5% at junction-level. More than a fifth of autoimmune loci shared an underlying causal variant in a single cell type by combining all five quantification types; a marked increase over current estimates of steady-state causal *cis*-eQTLs. As an example, we dissected in detail the genetic associations of systemic lupus erythematosus and functionally annotated the candidate genes. Many of the known and novel genes were concealed at gene-level (e.g. *BANK1*, *UBE2L3*, *IKZF2*, *TYK2, LYST*). By leveraging RNA-Seq, we were able to isolate the specific transcripts, exons, junctions, and introns modulated by the *cis*-eQTL - which supports the targeted design of follow-up functional studies involving alternative splicing. Causal *cis*-eQTLs detected at different quantification types were also found to localise to discrete epigenetic annotations. We provide our findings from all twenty autoimmune diseases as a web resource.

Author Summary

It is well acknowledged that non-coding genetic variants contribute to disease susceptibility through alteration of gene expression levels (known as 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 >20% 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.

54 # Introduction

55 The autoimmune diseases are a family of heritable, often debilitating, complex disorders in which
56 immune dysfunction leads to loss of tolerance to self-antigens and chronic inflammation [1]. Genome-
57 wide association studies (GWAS) have now detected hundreds of susceptibility loci contributing to risk
58 of autoimmunity [2] yet their biological interpretation still remains challenging [3]. Mapping single
59 nucleotide polymorphisms (SNPs) that influence gene expression (eQTLs) can provide meaningful
60 insight into the potential candidate genes and etiological pathways connected to discrete disease
61 phenotypes [4]. For example, such analyses have implicated dysregulation of autophagy in Crohn's
62 disease [5], the pathogenic role of $CD4^+$ effector memory T-cells in rheumatoid arthritis [6], and an
63 overrepresentation of transcription factors in systemic lupus erythematosus [7].
64
65 Expression profiling in appropriate cell types and physiological conditions is necessary to capture the
66 pathologically relevant regulatory changes driving disease risk [8]. Lack of such expression data is
67 thought to explain the observed disparity of shared genetic architecture between disease association and
68 gene expression at certain autoimmune loci [9]. A much overlooked cause of this disconnect however,
69 is not only the use of microarrays to profile gene expression, but also the resolution to which expression
70 is quantified using RNA-Sequencing (RNA-Seq) [10]. Expression estimates of whole-genes, individual
71 isoforms and exons, splice-junctions, and introns are obtainable with RNA-Seq [11–18]. The SNPs that
72 affect these discrete units of expression vary strikingly in their proximity to the target gene, localisation
73 to specific epigenetic marks, and effect on translated isoforms [18]. For example, in over 57% of genes
74 with both an eQTL influencing overall gene expression and a transcript ratio QTL (trQTL) affecting
75 the ratio of each transcript to the gene total, the causal variants for each effect are independent and
76 reside in distinct regulatory elements of the genome [18].
77
78 RNA-Seq based eQTL investigations that solely rely on whole-gene expression estimates are likely to
79 mask the allelic effects on independent exons and alternatively-spliced isoforms [16–19]. This is in part
80 due to subtle isoform switches and expression variation in exons that cannot be captured at gene-level

81 [20]. A large proportion of trait associated variants are thought to act via direct effects on pre-mRNA
82 splicing that do not change total mRNA levels [21]. Recent evidence also suggests that exon-level based
83 strategies are more sensitive than conventional gene-level approaches, and allow for detection of
84 moderate but systematic changes in gene expression that are not necessarily derived from alternative-
85 splicing events [15,22]. Furthermore, gene-level summary counts can be biased in the direction of
86 extreme exon outliers [22]. Use of isoform-, exon-, and junction-level quantification in eQTL analysis
87 also support the potential to not only point to the candidate genes involved, but also the specific
88 transcripts or functional domains affected [10,18]. This of course facilitates the design of targeted
89 functional studies and better illuminates the causative relationship between regulatory genetic variation
90 and disease. Lastly, though intron-level quantification is not often used in conventional eQTL analysis,
91 it can still provide valuable insight into the role of unannotated exons in reference gene annotations,
92 retained introns, and even intronic enhancers [23,24].
93
94 Low-resolution expression profiling with RNA-Seq will impede the subsequent identification of causal
95 eQTLs when applying genetic and epigenetic fine-mapping approaches [25]. In this investigation, we
96 aim to increase our knowledge of the regulatory mechanisms and candidate genes of human
97 autoimmune disease through integration of GWAS and RNA-Seq expression data profiled at gene-,
98 isoform-, exon-, junction-, and intron-level in lymphoblastoid cell lines (LCLs). This is firstly
99 performed in detail using association data from a GWAS in systemic lupus erythematosus, and is then
100 scaled up to a total of twenty autoimmune diseases. Our findings are provided as a web resource to
101 interrogate the functional effects of autoimmune associated SNPs (www.insidegen.com), and will serve
102 as the basis for targeted follow-up investigations.
103

104 # Results

105

106 **Gene-level expression quantification underestimates the number of causal *cis*-eQTLs**

107 Using densely imputed genetic association data from a large-scale GWAS in systemic lupus
108 erythematosus (SLE) in persons of European descent [7], we performed integrative *cis*-eQTL analysis
109 with RNA-Seq expression data profiled at five resolutions: gene-, transcript-, exon-, junction-, and
110 intron-level. The expression data are derived from the 373 healthy European donors of the Geuvadis
111 project (all individuals are included as part of the 1000 Genomes Project) profiled in lymphoblastoid
112 cell lines (LCLs) [18]. See S1 Figure and methods for a summary of how expression at the five
113 resolutions was quantified using RNA-Seq. A total of 38 genome-wide significant SLE loci (S1 Table)
114 were put forward for analysis following removal of: associated SNPs with minor allele frequency < 5%,
115 secondary associations upon conditional analysis on lead variant, and major histocompatibility complex
116 loci - owing to the known complex linkage disequilibrium (LD) patterns. To test for evidence of a single
117 shared causal variant between disease and gene expression at each of the remaining 38 SLE associated
118 loci, we employed the rigorous Joint Likelihood Mapping (JLIM) framework [9] using summary-level
119 statistics for the SLE association (primary trait) and full genotype-level data for gene expression
120 (secondary trait). Using JLIM, *cis*-eQTLs were defined if a nominal association (P<0.01) with at least
121 one SNP existed within 100kb of the SNP most associated with disease and the transcription start site
122 of the gene located within +/-500kb of that SNP (as defined by the authors of the JLIM package). JLIM
123 P-values were corrected for multiple testing as per the JLIM standards by using a false discovery rate
124 (FDR) of 5% per RNA-Seq quantification type (i.e. at exon-level, JLIM P-values were FDR adjusted
125 for total number of exons tested in *cis* to the 38 SNPs). Causal associations of the integrative *cis*-eQTL
126 SLE GWAS analysis using the JLIM package across the five RNA-Seq quantification types are
127 available in S2 Table and the full output (including non-causal associations) are available in S3 Table.
128 See S2 Figure for the distribution of JLIM P-values across the five RNA-Seq quantification types.

129

130 We found the number of *cis*-eQTLs driven by the same causal variant as the SLE disease association
131 was markedly underrepresented when considering conventional gene-level quantification (Table 1).
132 Only two of the 38 SLE susceptibility loci (5.3%) were deemed to be causal *cis*-eQTLs at gene-level
133 for three candidate genes. Interestingly, this is a similar proportion to that observed by the authors of
134 the JLIM method (*Chun et al* [9]). They found that 16 of the 272 (5.9%) autoimmune susceptibility loci
135 tested were *cis*-eQTLs driven by a shared causal variant in the Geuvadis RNA-Seq dataset using gene-
136 level quantification (based upon the seven autoimmune diseases interrogated - not including SLE).
137
138 Of note, transcript-level quantification did not increase the number of causal *cis*-eQTLs (Table 1).
139 Transcript-level analysis did, however, yield a greater number of candidate genes (seven individual
140 transcripts derived from a total of four genes). Both junction- and intron-level quantification increased
141 the number of causal *cis*-eQTLs to four (10.5% of the 38 total SLE loci). Using exon-level
142 quantification, we were able to define seven of the 38 SLE susceptibility loci (18.4%) as being
143 significant *cis*-eQTLs driven by a single shared causal variant. Exon-level analysis also produced the
144 greatest number of candidate gene targets: nine unique genes derived from 24 individual SNP-exon
145 pairs (Table 1). Therefore, even with multiple testing burden to correct for all SNP-exon *cis*-eQTL
146 pairs; we firstly conclude that exon-level analysis detects more causal *cis*-eQTLs than gene-level.
147
148 **A fifth of associated SNPs possess shared genetic effects with *cis*-eQTLs using RNA-Seq in LCLs**
149 By combining all five types of RNA-Seq quantification (gene, transcript, exon, junction, and intron) we
150 could define nine of the 38 SLE susceptibility loci (23.7%) as being driven by the same causal variant
151 as the *cis*-eQTL in LCLs (Table 1). Interestingly, this value, derived from interrogating only a single
152 cell type, is almost equal to the total number of causal autoimmune *cis*-eQTLs detected by *Chun et al*
153 [9] (~25%) when looking across the three different cell types analysed using JLIM ($CD4^+$ T-cells –
154 measured by microarray, $CD14^+$ monocytes – microarray, and LCLs – RNA-Seq gene-level).
155
156 We found that when considering the specificity of *cis*-eQTLs and target genes identified by JLIM
157 mapping across the five RNA-Seq quantification types, both gene- and transcript-level quantification

158 were redundant with respect to exon-level data; i.e. there were no causal *cis*-eQTLs or target genes
159 detected at gene- or transcript-level that were not captured by exon-level analysis (S3 Figure). Both
160 junction- and intron-level quantification captured a single causal *cis*-eQTL each that was not captured
161 by exon-level. We conclude that profiling at all resolutions of RNA-Seq is required to capture the full
162 set of potentially causal *cis*-eQTLs.
163

164 **Associated SNPs are most likely to colocalize with exon- and junction-level *cis*-eQTLs**

165 We compared the detection of *cis*-eQTLs using a standard linear-regression approach with the JLIM
166 method. To fully explore relationships within our results, a pairwise comparison was made across the
167 five RNA-Seq quantification types for matched SNP-gene *cis*-eQTL pairs (Figure 1). We only
168 considered matched SNP-gene *cis*-eQTL association pairs that had a nominal *cis*-eQTL association *P*-
169 value < 0.01 in both quantification types, and to be conservative, when multiple transcripts, exons,
170 junctions, and introns were annotated with the same gene symbol, we selected the associations that
171 minimized the difference in JLIM *P*-value between matched SNP-gene *cis*-eQTLs across RNA-Seq
172 quantification types. There were over 250 matched SNP-gene *cis*-eQTL pairs per comparison. We
173 firstly observed that the correlation of both *cis*-eQTL association *P*-values from regression and JLIM
174 *P*-values across RNA-Seq quantification types reflected the methods in which expression quantification
175 was obtained (Figure 1A). Both *cis*-eQTL and JLIM *P*-values between matched SNP-gene pairs at gene-
176 and transcript-level were highly correlated as gene-level estimates are obtained from the sum of all
177 transcript-level estimates for the same gene (see methods and S1 Figure). Exon-level and junction-level
178 associations were also highly correlated due to split-reads being incorporated into the exon-level
179 estimate. As expected, intron-level *cis*-eQTL and JLIM *P*-values for matched SNP-gene pairs were only
180 weakly correlated against other quantification types - as reads mapping to introns are not included in
181 the other quantification models. Interestingly, although *cis*-eQTL association *P*-values for matched
182 SNP-gene pairs between transcript-level and junction-level were found to be relatively high (r^2=0.70),
183 we found the JLIM *P*-values for the matched pairs to be comparatively low (r^2=0.29); suggesting that
184 whilst the strength of the *cis*-eQTL maybe similar between these quantification types, the underlying
185 causal variants driving the disease and *cis*-eQTL association are likely to be independent.

186 By plotting the JLIM *P*-values for matched SNP-gene pairs between different quantification types, we
187 found many instances of *P*-values distributed along the axes rather than on the diagonal (Figure 1B).
188 Our findings therefore suggest that often, one quantification type is more likely to explain the observed
189 disease association than the other. When we compared conventional gene-level *cis*-eQTL analysis
190 against exon-level results (Figure 1C), we found that of the 296 matched SNP-gene *cis*-eQTL
191 associations ($P<0.01$), eleven (4%) were deemed to share the same causal variant at both gene- and
192 exon-level using a nominal JLIM *P*-value threshold < 0.01. Only three of the 296 matched SNP-gene
193 *cis*-eQTL associations (1%) were captured by gene-level only - in contrast to the 26 (9% of total
194 associations) captured uniquely at exon-level. As expected, the overwhelming majority of *cis*-eQTL
195 associations (86%) did not possess a single shared causal variant at either gene- or exon-level. We
196 performed this analysis for all possible combinations of quantification types (Table 2). In each instance,
197 gene-level analysis detected only the minority of nominally causal associations for matched SNP-gene
198 association pairs (JLIM $P<0.01$). Exon-level and junction-level analysis consistently detected more
199 causal *cis*-eQTL associations than gene-, transcript-, and intron-level. In fact, when combined, exon-
200 and junction-level analysis explained the most nominally causal associations for all significant SNP-
201 gene *cis*-eQTL association pairs (23.8%).
202

203 **Leveraging RNA-Seq aids GWAS interpretation and reveals novel candidate genes**

204 We functionally dissected the 12 candidate genes taken from the nine SLE associated loci that showed
205 strong evidence of a shared causal variant with a *cis*-eQTL in LCLs. The nine, causal *cis*-eQTLs and
206 corresponding 12 candidate genes per RNA-Seq quantification type are listed in Table 3 along with
207 their *cis*-eQTL association *P*-values and related JLIM *P*-values. We systematically annotated all 12
208 genes using a combination of cell/tissue expression patterns, mouse models, known molecular
209 phenotypes, molecular interactions, and associations with other autoimmune diseases (S4 Table). We
210 found the majority of novel SLE candidate genes detected by RNA-Seq were predominately expressed
211 in immune-related tissues such as whole blood, the spleen and thymus, and the small intestine. Based
212 on our gene annotation and what is already documented at certain loci, we were sceptical on the
213 pathogenic involvement of three candidate genes (*PHTF1*, *ARHGAP30*, and *RABEP1*). Although the

214 *cis*-eQTL effect for these genes is evidently driven by the shared causal variant as the disease
215 association (defined by JLIM), it is possible that these effects of expression modulation are merely
216 passengers that are carried on the same functional haplotype as the true causal gene(s) and do not
217 contribute themselves to the breakdown of self-tolerance (detailed in S4 Table). We show the regional
218 association plots and the candidate genes detected from *cis*-eQTL analysis in S4 Figure.
219
220 The causal *cis*-eQTL rs2736340 for genes *BLK* and *FAM167A* was detected at all RNA-Seq profiling
221 types. It is well established that the risk allele of this SNP reduces proximal promoter activity of *BLK*;
222 a member of the Src family kinases that functions in intracellular signalling and the regulation of B-cell
223 proliferation, differentiation, and tolerance [26]. The allelic consequence of *FAM167A* expression
224 modulation is unknown. We found multiple instances of known SLE susceptibility genes that were
225 concealed when using gene-level quantification. For example, we defined rs7444 as a causal *cis*-eQTL
226 for *UBE2L3* at transcript- and exon-level - but not at gene-level (Table 3). The risk allele of rs7444 has
227 been associated with increased expression of *UBE3L3* (Ubiquitin conjugating enzyme E2 L3) in *ex vivo*
228 B-cells and monocytes and correlates with NF-κB activation along with increased circulating
229 plasmablast and plasma cell numbers [27]. Similarly, the rs10028805 SNP is a known splicing *cis*-
230 eQTL for *BANK1* (B-cell scaffold protein with ankyrin repeats 1). We replicated at exon-, and junction-
231 level this splicing effect which has been proposed to alter the B-cell activation threshold [28]. Again,
232 this mechanism was not detected using gene-level quantification.
233
234 *IKZF2* (detected at the exon-level only) is a transcription factor thought to play a key role in T-reg
235 stabilisation in the presence of inflammatory responses [29]. *IKZF2* deficient mice acquire an auto-
236 inflammatory phenotype in later life similar to rheumatoid arthritis, with increased numbers of activated
237 $CD4^+$ and $CD8^+$ T-cells, T-follicular helper cells, and germinal centre B-cells, which culminates in
238 autoantibody production [30]. Of note, other members of this gene family, *IKZF1* and *IKZF3*, are also
239 associated with SLE and can hetero-dimerize (S4 Table) [7]. We also believe *LYST*, *ATG4D*, and *TYK2*
240 to also be intriguing candidate genes. *LYST* encodes a lysosomal trafficking regulator [31] whilst
241 *ATG4D* is a cysteine peptidase involved in autophagy and this locus is associated with multiple

242 sclerosis, psoriasis, and rheumatoid arthritis [32]. *TYK2* is discussed in greater detail in the following
243 section.
244

245 **RNA-Seq can resolve the potential causal regulatory mechanism(s)**

246 Interestingly, for the three causal SNP-gene pairs detected at gene-level (rs2736340 – *BLK*, rs2736340
247 – *FAM167A*, and rs7444 – *CCDC116*), we found that at exon-level, all expressed exons of the stated
248 genes were deemed to possess causal associations. For example, rs2736340 is a causal *cis*-eQTL for all
249 thirteen exons of *BLK* and for all three exons of *FAM167A* (S5 Table). These data suggest that gene-
250 level analysis is capturing associations where all - or the majority of exons - are modulated by the *cis*-
251 eQTL in a causal manner.
252

253 We found that within the SLE associated loci that showed evidence of a shared causal variant with a
254 *cis*-eQTL (Table 3), there were many instances in which the proposed causal *cis*-eQTL modulated
255 expression of only a single expression element. This enabled us to resolve the potential regulatory effect
256 of the causal *cis*-eQTL to a particular transcript, exon, junction, or intron (S5 Table). We were able to
257 resolve to a single expression element in nine of the twelve candidate SNP-gene pairs. For example,
258 rs9782955 is a causal *cis*-eQTL for *LYST* at junction-level for only a single junction (chr1:235915471-
259 235916344; *cis*-eQTL P=1.3x10^{-03}; JLIM P=2.0x10^{-04}). We provide depicted examples of this isolation
260 analysis for candidate genes *IKZF2* (S5 Figure), *UBE2L3* (S6 Figure), and *LYST* (S7 Figure). Clearly
261 when only the minority of exons are effected – which we found occurred in nine of twelve association
262 pairs - gene-level analysis conceals the *cis*-eQTL association.
263

264 We provide a worked example of resolving the causal mechanism(s) using RNA-Seq for the novel
265 association rs2304256 with *TYK2* (Figure 2). The top panel of Figure 2A shows the genetic association
266 to SLE at the 19p13.2 susceptibility locus tagged by lead SNP rs2304256 (P=1.54x10^{-12}). Multiple
267 tightly correlated SNPs span the gene body and the 3′ region of *TYK2* – which encodes Tyrosine Kinase
268 2 - thought to be involved in the initiation of type I IFN signalling [33]. In the panel below, we plot the

269 gene-level association of all SNPs in *cis* to *TYK2* and show no significant association of rs3204256 with
270 *TYK2* expression (P=0.18). At exon-, and intron-level, we were able to classify rs2304256 as a causal
271 *cis*-eQTL for a single exon (chr19: 10475527-10475724; *cis*-eQTL P=2.58x10^{-09}; JLIM P<10^{-04}) and
272 single intron (chr19: 10473333-10475290; P=2.20x10^{-08}; JLIM P=2x10^{-04}) of *TYK2* respectively as
273 shown in the bottom two panels of Figure 2A. We show the exon and intron labelling of *TYK2* in further
274 detail in S8 Fig. We found strong correlation of association *P*-values of the SLE GWAS and the *P*-
275 values of *TYK2 cis*-eQTLs against at exon-level and intron-level, but not at gene-level; strengthening
276 our observation that rs2304256 is a causal *cis*-eQTL for *TYK2* at these resolutions (Figure 2B). The risk
277 allele rs2304256 [C] was found to be associated with decreased expression of the *TYK2* exon and
278 increased expression of the *TYK2* intron (Figure 2C). By plotting the *cis*-eQTL *P*-values alongside the
279 JLIM *P*-values for all exons and introns of *TYK2* against rs2304256 (Figure 2D), we clearly show that
280 only a single exon and a single intron of *TYK2* colocalize with the SLE association signal – marked by
281 an asterisk (note that rs2304256 is a strong *cis*-eQTL for many introns of *TYK2* but only shares a causal
282 variant with one intron). We show the genomic location of the affected exon and intron of *TYK2* in
283 Figure 2E (exon 8 and the intron between exons 9 and 10 – N.B that exons and introns are numbered
284 based on their inclusion in the *cis*-eQTL analysis and some maybe omitted from analysis due to no
285 expression). Intron 9-10 of *TYK2* is clearly 'expressed' in LCLs according to transcription levels
286 assayed by RNA-Seq on LCLs (GM12878) from ENCODE (Figure 2E).
287
288 Interestingly, rs2304256 (marked by an asterisk in Figure 2E) is a missense variant (V362F) within the
289 affected exon 8 of *TYK2*. The PolyPhen prediction of this substitution is predicted to be benign and, to
290 the best of our knowledge, no investigation has isolated the functional effect of this particular amino
291 acid change. We do not believe the *cis*-eQTL at exon 8 to be a result of variation at rs3204256 and
292 mapping biases, as the alignability of 75mers by GEM from ENCODE is predicted to be robust around
293 exon 8 (Figure 2E). In fact, rs3204256 [C] is the reference allele yet is associated with decreased
294 expression of exon 8.
295

296 In conclusion, we have found an interesting and novel mechanism that would have been concealed by
297 gene-level analysis that involves the risk allele of a missense SNP associated with decreased expression
298 of a single exon of *TYK2* but increased expression of the neighbouring intron. Whether the *cis*-eQTL
299 effect and missense variation act in a combinatorial manner and whether the intron is truly retained or
300 if it is derived from an unannotated transcript of *TYK2* is an interesting line of investigation.
301

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

303 We re-performed our integrative *cis*-eQTL analysis with the same Geuvadis RNA-Seq dataset in LCLs
304 using association data from twenty autoimmune diseases. This was to firstly reiterate the importance of
305 leveraging RNA-Seq in GWAS interpretation and to secondly demonstrate that our findings in SLE
306 persisted across other immunological traits. As the raw genetic association data were not available for
307 all twenty diseases, we were unable to implement the JLIM pipeline which requires densely typed or
308 imputed GWAS summary-level statistics. We therefore opted to use the Regulatory Trait Concordance
309 (RTC) method, which requires full genotype-level data for the expression trait, but only the marker
310 identifier for the lead SNP of the disease association trait (see methods for a description of the RTC
311 method). We stringently controlled our integrative *cis*-eQTL analysis for multiple testing to limit
312 potential false positive findings of overlapping association signals. To do this, we applied a Bonferroni
313 correction to nominal *cis*-eQTL *P*-values separately per disease and per RNA-Seq quantification type
314 (i.e. at exon-level, *cis*-eQTL *P*-values were corrected for the total number of exons tested in *cis* the
315 associated SNPs of the single disease in hand). A similar strategy was adopted by the authors of the
316 JLIM package who corrected separately for specific disease and cell type combinations [9]. We
317 rigorously defined causal *cis*-eQTLs, as associations with $P_{BF} < 0.05$ and RTC ≥ 0.95. An overview of
318 the analysis pipeline is depicted in S9 Figure and S10 Figure. Using an r^2 cut-off of 0.8 and a 100kb
319 limit, we pruned the 752 associated SNPs from the twenty human autoimmune diseases from the
320 Immunobase resource (S6 Table) to obtain 560 independent susceptibility loci. Again, we only
321 considered common (MAF >5%), autosomal loci outside of the MHC.
322

323 Our findings confirmed our previous results from the SLE investigation and again support the gene-
324 level study using the JLIM package from *Chun et al* [9]. As before, we found that only 5% (28 of the
325 560 loci) of autoimmune susceptibility loci were deemed to share causal variants with *cis*-eQTLs using
326 either gene- or transcript-level analysis (Figure 3A). Exon-level analysis more than doubled the yield
327 to 13% (72 of the 560 loci) with junction-, and intron-level analysis also outperforming gene-level (10%
328 and 8% respectively). When combining all RNA-Seq quantification types, we could define 20% of
329 autoimmune associated loci (110 of the 560 loci) as being candidate causal *cis*-eQTLs - which
330 corroborates our previous estimate in SLE using the JLIM package (23.7%).
331
332 By separating causal *cis*-eQTL associations out by quantification type, we found over half (65%) were
333 detected at exon-level, and considerable overlap of *cis*-eQTL associations existed between both types
334 (Figure 3B). Unlike in our SLE analysis, gene- and isoform-level analysis did capture a small fraction
335 of causal *cis*-eQTLs that were not captured at exon-level. Our data therefore suggest that although exon-
336 and junction-level, and to a lesser extent intron-level analysis, capture most candidate-causal *cis*-
337 eQTLs. It is necessary to prolife gene-expression at all quantification types to avoid misinterpretation
338 of the functional impact of disease associated SNPs.
339
340 We mapped the causal *cis*-eQTLs detected by all RNA-Seq quantification types back to the diseases to
341 which they are associated (Figure 3C). Interestingly, we observed the diseases that fell below the 20%
342 average comprised autoimmune disorders related to the gut: celiac disease (7%), inflammatory bowel
343 disease (14%), Crohn's disease (16%), and ulcerative colitis (18%). These observations are likely to be
344 a result of the cellular expression specificity of associated genes in colonic tissue and in T-cells [34].
345 Correspondingly, we observed an above-average frequency of causal *cis*-eQTLs detected in SLE (22%)
346 and primary biliary cirrhosis (37%); diseases in which the pathogenic role of B-lymphocytes and
347 autoantibody production is well documented [34]. Note that there are 60 SLE GWAS associations in
348 this analysis as these originate from three independent GWA studies (S6 Table). We further broke down
349 our results per disease by RNA-Seq quantification type (Figure 3D) and in all cases, the greatest
350 frequency of causal *cis*-eQTLs and candidate genes were captured by exon- and junction-level analyses.

351

352 **Web resource for functional interpretation of association studies of autoimmune disease**

353 We provide our analysis as a web resource (found at www.insidegen.com) for researchers to lookup
354 causal *cis*-eQTLs and candidate genes from the twenty autoimmune diseases detected across the five
355 RNA-Seq quantification types. The data are sub-settable and exportable by SNP ID, gene, RNA-Seq
356 resolution, genomic position, and association to specific autoimmune diseases.

357

358 **Causal *cis*-eQTLs localise to discrete chromatin regulatory elements**

359 The causal variants underling *cis*-eQTL associations at the five RNA-Seq quantification types were
360 often independent (Figure 1) and a previous investigation has suggested that causal variants of gene-
361 level and transcript-level *cis*-eQTLs reside in discrete functional elements of the genome [18]. We
362 therefore investigated whether this notion held true across the five RNA-Seq quantification types tested
363 in this study. To accomplish this, we selected the causal *cis*-eQTLs from the twenty autoimmune
364 diseases interrogated, and per quantification type, tested for enrichment of these SNPs across various
365 chromatin regulatory elements taken from the Roadmap Epigenomics Project in LCLs (using both the
366 Roadmap chromatin state model and the positions of histone modifications). We implemented the
367 permutation-based GoShifter algorithm to test for enrichment of causal *cis*-eQTLs and tightly correlated
368 variants (r^2>0.8) in genomic functional annotations in LCLs (see methods) [25]. Results of this analysis
369 are depicted in Figure 4. We found the 28 gene-level *cis*-eQTLs were enriched in two chromatin marks:
370 strong enhancers (P=0.036) and H3K27ac occupancy sites – a marker of active enhancers (P=0.002).
371 Transcript-level *cis*-eQTLs were also enriched in H3K27ac occupancy sites (P=0.039) but were not
372 enriched in any other marks. The 72 exon-level *cis*-eQTLs were additionally enriched in active
373 promoters (P=0.017). Interestingly, the 54 causal *cis*-eQTLs detected at junction-level were found to
374 be enriched in weak enhancers only (P=0.002); whilst the 43 intron-level *cis*-eQTLs were enriched in
375 chromatin states predicted to be involved in transcriptional elongation (P=0.001; 83% of intron-level
376 *cis*-eQTLs). Disease relevant *cis*-eQTLs detected at different expression phenotypes using RNA-Seq
377 clearly localise to largely discrete functional elements of the genome.

378

379 We quantified the number of causal *cis*-eQTLs and tightly correlated variants (r^2>0.8) per quantification
380 type that were predicted to be alter splice site consensus sequences of the target genes (assessed by
381 Sequence Ontology for the hg19 GENCODE v12 reference annotation). We found only two of the 28
382 (7%) gene-level *cis*-eQTLs disrupted consensus splice-sites for their target genes compared to the 14%
383 and 13% detected at exon- and junction-level respectively (Figure 4C). Our data suggest that although
384 exon- and junction- level analysis leads to the greatest frequency of causal *cis*-eQTLs, the majority at
385 this resolution cannot be explained directly by variation in annotated splice site consensus sequences
386 (splice region/donor/acceptor/ variants).

387

388 ## Discussion

389 Elucidation of the functional consequences of non-coding genetic variation in human disease is a major
390 objective of medical genomics [35]. Integrative studies that map disease-associated eQTLs in relevant
391 cell types and physiological conditions are proving essential in progression towards this goal through
392 identification of causal SNPs, candidate-genes, and illumination of molecular mechanisms [36]. In
393 autoimmune disease, where there is considerable overlap of immunopathology, integrative eQTL
394 investigations have been able to connect discrete aetiological pathways, cell types, and epigenetic
395 modifications, to particular clinical manifestations [2,34,36,37]. Emerging evidence however has
396 suggested that only a minority (~25%) of autoimmune associated SNPs share casual variants with basal-
397 level *cis*-eQTLs in primary immune cell-types [9].
398
399 Genetic variation can influence expression at every stage of the gene regulatory cascade - from
400 chromatin dynamics, to RNA folding, stability, and splicing, and protein translation [21]. It is now well
401 documented that SNPs affecting these units of expression vary strikingly in their genomic positions and
402 localisation to specific epigenetic marks [18]. The eQTLs that affect pre-transcriptional regulation -
403 affecting all isoforms of a gene - differ in the proximity to the target gene and effect on translated
404 isoforms than their co-transcriptional trQTL (transcript ratio QTL) counterparts. Where the effect size
405 of eQTLs generally increases in relation to transcription start site proximity, trQTLs are distributed
406 across the transcript body and generally localise to intronic binding sites of splicing factors [18,21]. In
407 over 57% of genes with both an eQTL influencing overall gene expression and an trQTL affecting the
408 ratio of each transcript to the gene total, the causal variants for each effect are independent and reside
409 in distinct regulatory elements of the genome [18]. In fact, three primary molecular mechanisms are
410 thought to link common genetic variants to complex traits. A large proportion of trait associated SNPs
411 act via direct effects on pre-mRNA splicing that do not change total mRNA levels [21]. Common
412 variants also act via alteration of pre-mRNA splicing indirectly through effects on chromatin dynamics
413 and accessibility. Such chromatin accessibility QTLs are however more likely to alter total mRNA
414 levels than splicing ratios. Lastly, it is thought that only a minority of trait associated variants have

415 direct effects on total gene expression that cannot be explained by changes in chromatin. As RNA-Seq
416 becomes the convention for genome-wide transcriptomics, it is essential to maximise its ability to
417 resolve and quantify discrete transcriptomic features so to expose the genetic variants that contribute to
418 changes in expression and isoform usage. The reasoning for our investigation therefore was to delineate
419 the limits of microarray and RNA-Seq based eQTL cohorts in the functional annotation of autoimmune
420 disease association signals.
421
422 To map autoimmune disease associated *cis*-eQTLs, we interrogated RNA-Seq expression data profiled
423 at gene-, isoform, exon-, junction-, and intron-level, and tested for a shared genetic effect at each
424 significant association. As we had densely imputed summary statistics from our SLE GWAS, we opted
425 to use the Joint Likelihood Mapping (JLIM) framework [9] to test for a shared causal variant between
426 the disease and *cis*-eQTL signals. This framework has been rigorously benchmarked against other
427 colocalisation procedures. Summary statistics were not available for the remaining autoimmune
428 diseases and therefore we implemented the Regulatory Trait Concordance (RTC) method for these
429 diseases and set a stringent multiple testing threshold to define causal *cis*-eQTLs. We found the
430 estimates of causal *cis*-eQTLs were near identical between the two methods used (Table 1 and Figure
431 3A). Exon- and junction-level quantification led to the greatest frequency of causal *cis*-eQTLs and
432 candidate genes (exon-level: 13-18%, junction-level: JLIM: 10-11%). We conclusively found that
433 associated variants were in fact more likely to colocalize with exon- and junction-level *cis*-eQTLs when
434 applying a nominal JLIM *P*-value threshold of <0.01 (Figure 1B and Table 2). Gene-level analysis was
435 thoroughly outperformed in all cases (5%). Our findings that gene-level analysis explain only 5% of
436 causal *cis*-eQTLs corroborate the findings from *Chun et al* [9] who composed and used the JLIM
437 framework to annotate variants associated with seven autoimmune diseases (multiple sclerosis, IBD,
438 Crohn's disease, ulcerative colitis, T1D, rheumatoid arthritis, and celiac disease). They found that only
439 16 of the 272 autoimmune associated loci (6%) shared causal variants with *cis*-eQTLs using gene-level
440 RNA-Seq (with the same Geuvadis European cohort in LCLs as used herein). In our investigation, we
441 argue that it is necessary to profile expression at all possible resolutions to diminish the likelihood of
442 overlooking potentially causal *cis*-eQTLs. In fact, by combining our results across all resolutions, we

443 found that 20-24% of autoimmune loci were candidate-causal *cis*-eQTLs for at least one target gene.
444 Our study therefore increases the number of autoimmune loci with shared genetic effects with *cis*-
445 eQTLs in a single cell type by over four-fold. Interestingly, using microarray data from $CD4^+$ T-cells
446 *Chun et al* classified 37 of the 272 autoimmune loci (14%) as causal *cis*-eQTLs [9] - strengthening the
447 hypothesis that autoimmune loci (especially those associated with inflammatory diseases of the gut) are
448 enriched in $CD4^+$ T-cell subsets and the cells themselves are pathogenic [25,34]. Microarray data are
449 known to underestimate the number of true causal *cis*-eQTLs [10]. If we assume that by leveraging
450 RNA-Seq we can increase the number of causal *cis*-eQTLs four-fold, we hypothesise that as many as
451 ~54% of autoimmune loci may share causal *cis*-eQTLs with gene expression at multiple resolutions in
452 $CD4^+$ T-cell populations. A large RNA-Seq based eQTL cohort profiled across many $CD4^+$ T-cell
453 subsets will therefore be of great use when annotating autoimmune-related traits. We reason that
454 although using relevant cell types and context-specific conditions will undoubtedly increase our
455 understanding of how associated variants alter cell physiology and ultimately contribute to disease risk;
456 it is clearly shown herein that we are only picking the low hanging fruit in current eQTL analyses. We
457 argue it necessary to reanalyse existing RNA-Seq based eQTL cohorts at multiple resolutions and
458 ensure new datasets are similarly dissected. Despite the severe multiple testing burden, we also argue
459 that expression profiling at multiple resolutions using RNA-Seq may be advantageous even when
460 looking for *trans*-eQTL effects. As *trans*-eQTLs are generally more cell-type specific and have a
461 weaker effect size, we decided not to perform such analyses using the Geuvadis LCL data. Large RNA-
462 Seq based eQTL cohorts in whole-blood will be more suitable for such analysis [19].
463
464 As well as biological reasons for using multiple expression phenotypes for integrative eQTL analysis,
465 there are also technical factors to consider. Gene-level expression estimates can generally be obtained
466 in two ways – union-exon based approaches [14,17] and transcript-based approaches [11,12]. In the
467 former, all overlapping exons of the same gene are merged into union exons, and intersecting exon and
468 junction reads (including split-reads) are counted to these pseudo-gene boundaries. Using this counting-
469 based approach, it is also possible to quantify meta-exons and junctions easily and with high confidence
470 by preparing the reference annotation appropriately [13,15,38]. Introns can be quantified in a similar

471 manner by inverting the reference annotation between exons and introns [18]. Of note, we found intron-
472 level quantification generated more candidate-causal *cis*-eQTLs than gene-level (Figure 3A). As the
473 library was synthesised from poly-A selection, these associations are unlikely due to differences in pre-
474 mRNA abundance. Rather, they are likely derived from either true retained introns in the mature RNA
475 or from coding exons that are not documented in the reference annotation used. Transcript-based
476 approaches make use of statistical models and expectation maximization algorithms to distribute reads
477 among gene isoforms - resulting in isoform expression estimates [11,12]. These estimates can then be
478 summed to obtain the entire expression estimate of the gene. Greater biological insight is gained from
479 isoform-level analysis; however, disambiguation of specific transcripts is not trivial due to substantial
480 sequence commonality of exons and junctions. In fact, we found only 5% of autoimmune loci shared a
481 causal variant at transcript-level.
482
483 The different approaches used to estimate expression can also lead to significant differences in the
484 reported counts. Union-based approaches, whilst computationally less expensive, can underestimate
485 expression levels relative to transcript-based, and this difference becomes more pronounced when the
486 number of isoforms of a gene increases, and when expression is primarily derived from shorter isoforms
487 [20]. The Geuvadis study implemented a transcript-based approach to obtain whole-gene expression
488 estimates. Clearly therefore, a gold standard of reference annotation and eQTL mapping using RNA-
489 Seq is essential for comparative analysis across datasets. Our findings support recent evidence that
490 suggests exon-level based strategies are more sensitive and specific than conventional gene-level
491 approaches [22]. Subtle isoform variation and expression of less abundant isoforms are likely to be
492 masked by gene-level analysis. Exon-level allows for detection of moderate but systematic changes in
493 gene expression that are not captured at gene-level, and also, gene-level summary counts can be shifted
494 in the direction of extreme exon outliers [22]. It is therefore important to note that a positive exon-level
495 eQTL association does not necessarily mean a differential exon-usage or splicing mechanism is
496 involved; rather a systematic expression effect across the whole gene may exist that is only captured by
497 the increased sensitivity. Additionally, by combining exon-level with other RNA-Seq quantification
498 types, inferences can be made on the particular isoforms and functional domains affected by the eQTL

499 which can later aid biological interpretation and targeted follow-up investigations [10]. We clearly show
500 this from our analysis of SLE candidate genes *IKZF2* (S5 Figure), *UBE2L3* (S6 Figure), *LYST* (S7
501 Figure) and *TYK2* (Figure 2). For *TYK2* we reveal a novel mechanism whereby the associated variant
502 rs2304256 [C] leads to decreased expression of a single exon and increased expression of a
503 neighbouring intron (Figure 2). By isolating particular exons, junctions, and introns, one can design
504 more refined follow-up investigations to study the functional impact of non-coding disease associated
505 variants. We show how our findings can be leveraged to comprehensively examine GWAS results of
506 autoimmune diseases. We found nine of the 38 SLE susceptibility loci were causal *cis*-eQTLs (Table
507 3) for 12 candidate genes which we later functionally annotated in detail (S4 Table).
508
509 Taken together, we have provided a deeper mechanistic understanding of the genetic regulation of gene
510 expression in autoimmune disease by profiling the transcriptome at multiple resolutions using RNA-
511 Seq. Similar analyses leveraging RNA-Seq in new and existing datasets using relevant cell types and
512 context-specific conditions (such as response eQTLs as shown in [39]) will undoubtedly increase our
513 understanding of how associated variants alter cell physiology and ultimately contribute to disease risk.
514

515 # Materials and Methods

516

517 **RNA-Sequencing expression data in lymphoblastoid cell lines**

518 RNA-Sequencing (RNA-Seq) expression data from 373 lymphoblastoid cell lines (LCLs) derived from
519 four European sub-populations (Utah Residents with Northern and Western European Ancestry, British
520 in England and Scotland, Finnish in Finland, and Toscani in Italia) of the Geuvadis project [18] were
521 obtained from the EBI ArrayExpress website under accession: E-GEUV-1. The 89 individuals of the
522 Geuvadis project from the Yoruba in Ibadan, Nigeria were excluded from this analysis. All individuals
523 were included as part of the 1000Genomes Project. Expression was profiled using RNA-Seq at five
524 quantification types: gene-, transcript-, exon-, junction-, and intron-level (the files downloaded and used
525 in this analysis have the suffix: 'QuantCount.45N.50FN.samplename.resk10.txt.gz'). Full methods of
526 expression quantification can be found in the original publication and on the Geuvadis wiki page:
527 http://geuvadiswiki.crg.es/). We have also provided a breakdown of the quantification methods in S1
528 Figure. Expression data downloaded represent quantifications that are corrected for sequencing depth
529 and gene/exon etc length (RPKM). Only expression elements quantified in >50 % of individuals were
530 kept and Probabilistic Estimation of Expression Residuals (PEER) had been used to remove technical
531 variation [40]. We transformed all expression data to a standard normal distribution.
532 In summary, transcripts, splice-junctions, and introns were quantified using Flux Capacitor against the
533 GENCODE v12 basic reference annotation [16]. Reads belonging to single transcripts were predicted
534 by deconvolution per observations of paired-reads mapping across all exonic segments of a locus. Gene-
535 level expression was calculated as the sum of all transcripts per gene. Annotated splice junctions were
536 quantified using split read information, counting the number of reads supporting a given junction.
537 Intronic regions that are not retained in any mature annotated transcript, and reported mapped reads in
538 different bins across the intron to distinguish reads stemming from retained introns from those produced
539 by not yet annotated exons. Meta-exons were quantified by merging all overlapping exonic portions of
540 a gene into non-redundant units and counting reads within these bins. Reads were excluded when the
541 read pairs map to two different genes.

542

543 **SLE associated SNPs**

544 SNPs genetically associated to systemic lupus erythematosus (SLE) were taken from the *Bentham and*
545 *Morris et al 2015* GWAS in persons of European descent [7]. The study comprised a primary GWAS,
546 with validation through meta-analysis and replication study in an external cohort (7,219 cases, 15,991
547 controls in total). Independently associated susceptibility loci taken forward for this investigation were
548 those that passed either genome-wide significance ($P<5x10^{-08}$) in the primary GWAS or meta-analysis
549 and/or those that reached significance in the replication study (q<0.01). We defined the lead SNP at
550 each locus as either being the SNP with the lowest *P*-value post meta-analysis or the SNP with the
551 greatest evidence of a missense effect as defined by a Bayes Factor (see original publication). We
552 omitted non-autosomal associations and those within the Major Histocompatibility Complex (MHC),
553 and SNPs with a minor allele frequency (MAF) < 0.05. In total, 38 independently associated SLE
554 associated GWAS SNPs were taken forward for investigation (S1 Table). Each susceptibility locus had
555 previously been imputed to the level of 1000 Genomes Phase3 using a combination of pre-phasing by
556 the SHAPEIT algorithm and imputation by IMPUTE (see original publication for full details) [7].

557

558 ***Cis*-eQTL analysis and Joint Likelihood Mapping (JLIM) of SLE associated SNPs**

559

560 ***Primary trait summary statistics file***

561 A JLIM index file for each of the 38 SLE associated SNPs was firstly generated by taking the position
562 of each SNP (hg19) and a creating a 100kb interval in both directions. Summary-level association
563 statistics were obtained form the *Bentham and Morris et al* 2015 European SLE GWAS (imputed to
564 1000Genomes Phase 3). We downloaded summary-level association data (chromosome, position, SNP,
565 *P*-value) for all directly typed or imputed SNPs with an IMPUTE info score ≥0.7 within each of the 38
566 intervals. The two-sided *P*-value was transformed into a *Z*-statistic as described by JLIM.

567

568 ***Reference LD file***

569 Genotype files in VCF format for all 373 European individuals of the Geuvadis RNA-Seq project were
570 obtained from the EBI ArrayExpress under accession: E-GEUV-1. The 41 individuals genotyped on
571 the Omni 2.5M SNP array had been previously imputed to the Phase 1 v3 release as described [18]; the
572 remaining had been sequenced as part of the 1000 Genomes Phase1 v3 release (low-coverage whole
573 genome and high-coverage exome sequencing data). Using VCFtools, we created PLINK binary
574 ped/map files for each of the 38 intervals and kept only biallelic SNPs with a MAF >0.05, imputation
575 call-rates $\geq$ 0.7, Hardy–Weinberg equilibrium *P*-value $>1x10^{-04}$ and SNPs with no missing genotypes,
576 we also only included SNPs that we had primary trait association summary statistics for. These are
577 referred to as the secondary trait genotype files. We then used the JLIM Perl script *fetch.refld0.EUR.pl*
578 to generate the 38 reference LD files from the 373 individuals (the script had been edited to include the
579 extra 95 Finnish individuals).
580
581 ***Cis-eQTL analysis***
582 We created a separate PLINK phenotype file (sample ID, normalized expression residual) for each
583 individual gene, transcript, exon, junction, and intron in *cis* (within +/-500kb) to the 38 lead SLE GWAS
584 SNPs. We only included protein-coding, lincRNA, and antisense genes in our analysis as classified by
585 Ensembl BioMart. Using the chromosome 20 genotype VCF file of the 373 European individuals (E-
586 GEUV-1), we conducted principle component analysis (PCA) and generated an identity-by-state matrix
587 using the Bioconductor package SNPRelate (S9 Figure) [41]. Based on these results, we decided to
588 include the first three principle components and the binary imputation status (as 41 individuals had been
589 genotyped on the Omni 2.5M SNP array were imputed to the Phase 1 v3 release) of the European
590 individuals (derived from Phase1 and Phase2 1000Genomes releases) in the *cis*-eQTL analysis so to
591 minimize biases derived from population structure and imputation status.
592 We used PLINK to perform *cis*-eQTL analysis using the '*--linear*' function, including the above
593 covariates, for each expression unit (phenotype file) in *cis* to the 38 loci (secondary trait genotype files).
594 We performed 10,000 permutations per regression and saved the output of each permutation procedure.
595 In *cis* to the 38 SLE SNPs were: 439 genes, 1,448 transcripts (originating from 456 genes), 3,045 exons
596 (400 genes), 2,886 junctions (332 genes), and 1,855 introns (443 genes).

597

598 ***Joint likelihood mapping (JLIM) and multiple testing correction***

599 Per RNA-Seq quantification type, a JLIM configuration file was created using the *jlim_gencfg.sh* script
600 and JLIM then run using *run_jlim.sh* – setting the r^2 resolution limit to 0.8. We merged the configuration
601 files and output files to create the final results table which included the primary and secondary trait
602 association *P*-value, the JLIM statistic, and the JLIM *P*-value by permutation. Multiple testing was
603 corrected for on the JLIM *P*-values per RNA-Seq quantification type using a false discovery rate (FDR)
604 as applied by the authors of JLIM. A JLIM *P*-value $<10^{-04}$ means that the JLIM statistic is more extreme
605 than the permutation (10,000). We classified causal *cis*-eQTLs as SLE associated variants that share a
606 single causal variant with a *cis*-eQTL based on the following: if there existed a nominal *cis*-eQTL
607 ($P<0.01$) with at least one SNP within 100kb of the SNP most associated with disease, the transcription
608 start site of the expression target was located within +/-500kb of that SNP, and the FDR adjusted JLIM
609 *P*-value of the association passed the 5% threshold. Candidate genes modulated by the causal *cis*-eQTL.

610

611 **Functional annotation of SLE associated genes from *cis*-eQTL analysis**

612 Using publically available resources, we systematically annotated the twelve SLE associated genes that
613 were classified as being modulated by causal *cis*-eQTLs. The expression profiles at RNA-level across
614 multiple cell and tissue types were interrogated in GTEx [42] and the Human Protein Atlas [43] - with
615 the top three cell/tissue types documented per gene. We noted using Online Mendelian Inheritance in
616 Man [44] any gene-phenotype relationships by caused by allelic variants and any immune-related
617 phenotypes of animal models. Protein-protein interactions of candidate genes were taken from the
618 BioPlex v2.0 interaction network (conducted in HEK293T cells) [45]. Using the ImmunoBase resource
619 (https://www.immunobase.org/), we looked up each gene and noted if the gene had been prioritized as
620 the 'candidate gene' within the susceptibility locus per publication. Finally, we counted the number
621 publications from PubMed found using the keywords 'gene name AND SLE'.

622

623 **Associated SNPs from twenty autoimmune diseases**

624 Autoimmune associated SNPs were taken from the ImmunoBase resource (www.immunobase.org).
625 This resource comprises summary case-control association statistics from twenty diseases: twelve
626 originally targeted by the ImmunoChip consortium (ankylosing spondylitis, autoimmune thyroid
627 disease, celiac disease, Crohn's disease, juvenile idiopathic arthritis, multiple sclerosis, primary biliary
628 cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes, ulcerative
629 colitis), and eight others (alopecia areata, inflammatory bowel disease, IgE and allergic sensitization,
630 narcolepsy, primary sclerosing cholangitis, Sjogren syndrome, systemic scleroderma, vitiligo).
631 The curated studies and their corresponding references used in this analysis are presented in S6 Table.
632 For each disease, we took the lead SNPs which were defined as a genome-wide significant SNP with
633 the lowest reported *P*-value in a locus. Associations on the X-chromosome and within the MHC and
634 SNPs with minor allele frequency < 5% were omitted from analysis, leaving 752 associated SNPs. We
635 pruned these loci using the '*--indep-pairwise*' function of PLINK 1.9 with a window size of 100kb and
636 an r^2 threshold of 0.8, to create an independent subset of 560 loci.
637

638 **Integrative *cis*-eQTL analysis of twenty autoimmune diseases with RNA-Seq**

639 An overview of the integration pipeline using the twenty autoimmune diseases against the Geuvadis
640 RNA-Seq cohort in 373 European LCLs is depicted in S10 Figure. Genotype data of the 373 individuals
641 were transformed and quality controlled as previously described in the above methods sections (biallelic
642 SNPs kept with a MAF >0.05, imputation call-rates $\geq$ 0.7, Hardy–Weinberg equilibrium *P*-value
643 $>1x10^{-04}$).
644 We opted to use the Regulatory Trait Concordance (RTC) method to assess the likelihood of a shared
645 causal variant between the disease association and the *cis*-eQTL signal [46]. This method requires full
646 genotype-level data for the expression trait but only the marker identifier for the lead SNP of the disease
647 association trait. SNPs within the 560 associated loci for the expression trait were firstly classified
648 according to their position in relation to recombination hotspots (based on genome-wide estimates of
649 hotspot intervals) [47]. Normalized gene expression residuals (PEER factor normalized RPKM) for
650 each quantification type were transformed to standard normal and the first three principle components

651 used as covariates in the *cis*-eQTL model as well as the binary imputation status (as previously
652 described above). All *cis*-eQTL association testing was performed using a liner regression model in R.
653 *Cis*-eQTL mapping was performed for the lead SNP and all SNPs within the hotspot recombination
654 interval against protein-coding, lincRNA, and antisense expression elements (genes, transcripts, exons
655 etc.) within +/-500kb of the lead SNP. In *cis* to the 560 loci were: 7,633 genes, 27,257 transcripts
656 (originating from 7,310 genes), 52,651 exons (5,435 genes), 48,627 junctions (4,237 genes), 34,946
657 introns (6,233 genes).
658 For each *cis*-eQTL association, the residuals from the linear-regression of the best *cis*-asQTL (lowest
659 association *P*-value within the hotspot interval) were extracted. Linear regression was then performed
660 using all SNPs within the defined hotspot interval against these residuals. The RTC score was then
661 calculated as (N_{SNPs} - $Rank_{GWAS\ SNP}$ / N_{SNPs}). Where N_{SNPs} is the total number of SNPs in the recombination
662 hotspot interval, and $Rank_{GWAS\ SNP}$ is the rank of the GWAS SNP association *P*-value against all other
663 SNPs in the interval from the liner association against the residuals of the best *cis*-eQTL.
664 We rigorously adjusted for multiple testing of *cis*-eQTL *P*-values using a Bonferroni correction per
665 quantification type (corrected for number of genes, isoforms, exons, junctions, and introns tested) and
666 per disease – as we wanted to keep our analysis as close to the authors of JLIM who themselves also
667 adjusted per cell type and per disease. We stringently defined causal *cis*-eQTLs as associations with
668 expression $P_{BF} < 0.05$ and an RTC score ≥ 0.95. Candidate genes are modulated by the *cis*-eQTL.
669

670 **Functional enrichment of causal *cis*-eQTLs in chromatin regulatory elements**

671 To test for enrichment of causal *cis*-eQTL associations in chromatin regulatory elements we
672 implemented the Genomic Annotation Shifter (GoShifter) package [25]. Chromatin regulatory elements
673 were divided into two categories: chromatin state segmentation and histone marks. The genomic
674 coordinates of the fifteen predicted chromatin state segmentations (active promoter, strong enhancer,
675 insulator etc.) for LCLs (in the GM12878 cell-line) were downloaded from the UCSC Table browser
676 (track name: wgEncodeBroadHmmGm12878HMM). Histone marks and DNase hypersensitivity sites
677 were obtained from the NIH Roadmap Epigenomics Project for LCLs (GM12878) in NarrowPeak

678 format. Sites were filtered for genome-wide significance using an FDR threshold of 0.01 and peak
679 widths harmonised to 200bp in length centred on the peak summit (as used in the GoShifter publication).

680 We obtained all SNPs in strong LD ($r^2 > 0.8$) with the causal *cis*-eQTLs by using the *getLD.sh* script
681 from GoShifter (interrogating the 1000Genomes Project for Phase3 Europeans). Per quantification type,
682 we then calculated the proportion of loci in which at least one SNP in LD overlapped a chromatin
683 regulatory element (conducted one at a time per chromatin mark). The coordinates of the chromatin
684 marks were then randomly shifted, whilst retaining the positions of the SNPs, and frequency of overlap
685 re-calculated. This was carried out over 1,000 permutations to draw the null distribution. The *P*-value
686 was calculated as the proportion of iterations for which the number of overlapping loci was equal to or
687 greater than that for the tested SNPs ($P < 0.05$ used as significance threshold).

688

689 **Data visualisation and online resource**

690 R version 3.3.1 and ggplot2 was used to create heatmaps, box-plots, and correlation plots. Genes were
691 plotted in UCSC Genome Browser [48] and regional association plots in LocusZoom [49]. To access
692 the online results table, visit www.insidegen.com and follow the link 'Lupus' then 'data for scientists'.
693 The table is under title: Expression data associated with different autoimmune diseases.

694

Acknowledgements

We thank Dr David L Morris for helpful discussions throughout this work. Philip Tombleson is employed by the Biomedical Research Centre, we thank him for his assistance with data management. The GEUVADIS 1000 Genomes RNA-Seq data was downloaded from the EBI ArrayExpress Portal (accession E-GEUV-1).

701 **References**

702 1. Fever FM. NIH Progress in Autoimmune Diseases Research. in National Institute of Health
703 Publication. 2005; 17–7576.
704 2. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and
705 complex relationships among immune-mediated diseases. Nat Rev Genet. Nature Publishing
706 Group; 2013;14: 661–73. doi:10.1038/nrg3502
707 3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential
708 etiologic and functional implications of genome-wide association loci for human diseases and
709 traits. Proc Natl Acad Sci U S A. 2009;106: 9362–9367. doi:10.1073/pnas.0903103106
710 4. Westra H-J, Franke L. From genome to function by studying eQTLs. Biochim Biophys Acta.
711 Elsevier B.V.; 2014;1842: 1896–1902. doi:10.1016/j.bbadis.2014.04.024
712 5. Klionsky DJ. Crohn's disease, autophagy, and the Paneth cell. N Engl J Med. 2009;360: 1785–
713 1786. doi:10.1056/NEJMcibr0810347
714 6. Hu X, Kim H, Raj T, Brennan PJ, Trynka G, Teslovich N, et al. Regulation of Gene
715 Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4+
716 Effector Memory T Cells. PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004404
717 7. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, et
718 al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity
719 genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. Nature Publishing
720 Group; 2015;47: 1457–1464. doi:10.1038/ng.3434
721 8. Fairfax BP, Knight JC. Genetics of gene expression in immunity to infection. Curr Opin
722 Immunol. Elsevier Ltd; 2014;30: 63–71. doi:10.1016/j.coi.2014.07.001
723 9. Chun S, Casparino A, Patsopoulos NA, Croteau-chonka DC, Raby BA, Jager PL De, et al.
724 Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-
725 associated loci in three major immune-cell types. NatGenet. 2017; doi:10.1038/ng.3795
726 10. Odhams CA, Cortini A, Chen L, Roberts AL, Viñuela A, Buil A, et al. Mapping eQTLs with
727 RNA-seq reveals novel susceptibility genes, non-coding RNAs and alternative-splicing events

728 in systemic lupus erythematosus. Hum Mol Genet. 2017;26: ddw417.
729 doi:10.1093/hmg/ddw417
730 11. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and
731 transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat
732 Protoc. 2012;7: 562–78. doi:10.1038/nprot.2012.016
733 12. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or
734 without a reference genome. BMC Bioinformatics. 2011;12: 323. doi:10.1186/1471-2105-12-
735 323
736 13. Schuierer S, Roma G. The exon quantification pipeline (EQP): a comprehensive approach to
737 the quantification of gene, exon and junction expression from RNA-seq data. Nucleic Acids
738 Res. 2016; gkw538. doi:10.1093/nar/gkw538
739 14. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput
740 sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638
741 15. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq-
742 npre20126837-2.pdf. Genome Res. 2012;12: 1088–9051. doi:10.1101/gr.133744.111
743 16. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al.
744 Transcriptome genetics using second generation sequencing in a Caucasian population.
745 Nature. Nature Publishing Group; 2010;464: 773–777. doi:10.1038/nature08903
746 17. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning
747 sequence reads to genomic features. Bioinformatics. 2014;30: 923–930.
748 doi:10.1093/bioinformatics/btt656
749 18. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen P a C, Monlong J, Rivas M a, et al.
750 Transcriptome and genome sequencing uncovers functional variation in humans. Nature.
751 2013;501: 506–11. doi:10.1038/nature12531
752 19. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing
753 the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.
754 Genome Res. 2014;24: 14–24. doi:10.1101/gr.155192.113
755 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. Li YI, Geijn B Van De, Raj A, Knowles D a, Petti A a, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352. doi:10.1126/science.aad9417

22. 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

23. 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

24. 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

25. 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

26. Guthridge JM, Lu R, Sun H, Sun C, Wiley GB, Dominguez N, et al. Two functional lupus-associated BLK promoter variants control cell-type- and developmental-stage-specific transcription. Am J Hum Genet. 2014;94: 586–598. doi:10.1016/j.ajhg.2014.03.008

27. Lewis MJ, Vyse S, Shields AM, Boeltz S, Gordon PA, Spector TD, et al. UBE2L3 polymorphism amplifies NF-κB activation and promotes plasma cell development, linking linear ubiquitination to multiple autoimmune diseases. Am J Hum Genet. The Authors; 2015;96: 221–234. doi:10.1016/j.ajhg.2014.12.024

28. Kozyrev S V, Abelson A-K, Wojcik J, Zaghlool A, Linga Reddy MVP, Sanchez E, et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. Nat Genet. 2008;40: 211–216. doi:10.1038/ng0408-484

29. Getnet D, Grosso JF, Goldberg M V., Harris TJ, Yen HR, Bruno TC, et al. A role for the transcription factor Helios in human CD4+CD25+ regulatory T cells. Mol Immunol. Elsevier

784 Ltd; 2010;47: 1595–1600. doi:10.1016/j.molimm.2010.02.001
785 30. Kim H, Barnitz RA, Kreslavsky T, Brown FD, Moffett H, Lemieux ME, et al. Stable
786 inhibitory activity of regulatory T cells requires the transcription factor Helios. Science.
787 2015;350: 334–339.
788 31. Sepulveda FE, Burgess A, Heiligenstein X, Goudin N, Ménager MM, Romao M, et al. LYST
789 Controls the Biogenesis of the Endosomal Compartment Required for Secretory Lysosome
790 Function. Traffic. 2015;16: 191–203. doi:10.1111/tra.12244
791 32. Li M, Hou Y, Wang J, Chen X, Shao ZM, Yin XM. Kinetics comparisons of mammalian Atg4
792 homologues indicate selective preferences toward diverse Atg8 substrates. J Biol Chem.
793 2011;286: 7327–7338. doi:10.1074/jbc.M110.199059
794 33. Prchal-Murphy M, Semper C, Lassnig C, Wallner B, Gausterer C, Teppner-Klymiuk I, et al.
795 TYK2 kinase activity is required for functional type I interferon responses in Vivo. PLoS One.
796 2012;7: 1–12. doi:10.1371/journal.pone.0039141
797 34. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and
798 epigenetic fine mapping of causal autoimmune disease variants. Nature. Nature Publishing
799 Group; 2015;518: 337–343. doi:10.1038/nature13835
800 35. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and
801 molecular biology. Genome Res. 2015;25: 1427–1431. doi:10.1101/gr.190983.115.
802 36. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat
803 Rev Genet. Nature Publishing Group; 2015;16: 197–212. doi:10.1038/nrg3891
804 37. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify
805 critical cell types for fine mapping complex trait variants. Nat Genet. Nature Publishing
806 Group; 2013;45: 124–30. doi:10.1038/ng.2504
807 38. Ongen H, Dermitzakis ET. Alternative Splicing QTLs in European and African Populations.
808 Am J Hum Genet. The Authors; 2015;97: 567–575. doi:10.1016/j.ajhg.2015.09.004
809 39. Kim-Hellmuth S, Bechheim M, Puetz B, Mohammadi P, Nedelec Y, Giangreco N, et al.
810 Genetic regulatory effects modified by immune activation contribute to autoimmune disease
811 associations. Nat Commun. Springer US; 2017;8: 116376. doi:10.1101/116376

812 40. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-
813 genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS
814 Comput Biol. 2010;6: 1–11. doi:10.1371/journal.pcbi.1000770
815 41. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing
816 toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:
817 3326–3328. doi:10.1093/bioinformatics/bts606
818 42. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:
819 580–585. doi:10.1038/ng.2653
820 43. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al.
821 Proteomics. Tissue-based map of the human proteome. Science. 2015;347: 1260419.
822 doi:10.1126/science.1260419
823 44. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian
824 Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic
825 Acids Res. 2005;33: 514–517. doi:10.1093/nar/gki033
826 45. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network:
827 A Systematic Exploration of the Human Interactome. Cell. 2015;162: 425–440.
828 doi:10.1016/j.cell.2015.06.043
829 46. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate
830 causal regulatory effects by integration of expression QTLs with complex trait genetic
831 associations. PLoS Genet. 2010;6: e1000895. doi:10.1371/journal.pgen.1000895
832 47. McVean GA. The fine-scale structure of recombination rate variation in the human genome.
833 Science (80-). 2004;304: 581. Available: http://dx.doi.org/10.1126/science.1092500
834 48. Kent WJ, Sugnet CW, Furey TS, Roskin KM. The Human Genome Browser at UCSC W. J
835 Med Chem. 2002;19: 1228–31. doi:10.1101/gr.229102.
836 49. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom:
837 Regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:
838 2336–2337. doi:10.1093/bioinformatics/btq419
839

840 # Figure captions

841

842 **Figure 1. Pairwise comparison of *cis*-eQTL and JLIM *P*-values for matched SNP-gene pairs**

843 This figure is complementary to the data in Table 2 and is derived from *cis*-eQTL analysis of the 38
844 SLE associated SNPs using RNA-Seq and implementation of the JLIM method to assess evidence of a
845 shared causal variant. (A) We measured the Pearson's correlation separately of all *cis*-eQTL and JLIM
846 *P*-values between matched SNP-gene *cis*-eQTL pairs across the five RNA-Seq quantification types. We
847 only considered matched SNP-gene *cis*-eQTL association pairs that had a nominal *cis*-eQTL association
848 *P*-value < 0.01 in both quantification types, and to be conservative, when multiple transcripts, exons,
849 junctions, and introns were annotated with the same gene symbol, we selected the associations that
850 minimized the difference in JLIM *P*-value between matched SNP-gene *cis*-eQTLs across RNA-Seq
851 quantification types. Note the weak JLIM *P*-value correlation of matched transcript-level and junction-
852 level *cis*-eQTLs suggesting they stem from independent causal variants. (B) Correlation plots of
853 matches SNP-gene *cis*-eQTL pairs as described above (red: *cis*-eQTL *P*-value; blue: JLIM *P*-value).
854 Note that JLIM *P*-values often aggregate on the axis rather than on the diagonal suggesting independent
855 causal variants across different quantification types. (C) An example of the sensitivity of exon-level
856 analysis relative to gene-level. The majority of nominally significant JLIM *P*-values (<0.01) for
857 matched SNP-gene pairs are captured by exon-level analysis and concealed at gene-level (green box:
858 9%).

859

860 **Figure 2. Isolation of potential causal molecular mechanism in *TYK2* by SLE *cis*-eQTL rs2304256**

861 (A) SLE GWAS association plot and *cis*-eQTL association plot around the 19p13.2 susceptibility locus
862 tagged by rs2304256. The top panel shows the association plot with SLE that spans the gene body and
863 3′ region of *TYK2* (Tyrosine Kinase 2). The haplotype block composed of highly correlated SNPs is
864 highlighted in the red block. The second panel shows the *cis*-eQTL association plot at gene-level of all
865 proximal SNPs to *TYK2* (no significant association with rs2304256 is detected). The third panel shows
866 the same regional association but at exon-level for the most associated exon of *TYK2* with rs2304256 –

867 the bottom panel is at intron-level for *TYK2* (both are highly associated). (B) Correlation of SLE GWAS
868 *P*-value and *cis*-eQTL association *P*-value for all SNPs in *cis* to *TYK2*. We show at gene-level the most
869 associated SLE SNPs are not *cis*-eQTLs (top panel). The middle and bottom panels show the same
870 correlation at exon-level and intron-level and reveal the most associated SNPs to SLE are also the most
871 associated *cis*-eQTLs to *TYK2*. (C) The direction of effect of *cis*-eQTL rs2304256 with *TYK2* at gene-
872 level (top), exon-level (middle), and intron-level (bottom panel). The risk allele is rs2304256 [C]. (D)
873 The top panel shows *cis*-eQTL association and JLIM *P*-values for all exons of *TYK2* against rs2304256.
874 Exon 8 (marked by an asterisk) is defined as having a causal association with rs2304256. The bottom
875 panel shows the intron-level *cis*-eQTL of *TYK2* against rs2304256. Note many introns are *cis*-eQTLs
876 but are not causal with rs2304256. Exons and introns are numbered consecutively from start to end of
877 gene if they are expressed (note some are not and therefore not included). (E) The genomic location of
878 the single exon and single intron of *TYK2* that are modulated by rs2304256 are highlighted (rs2304256
879 is marked by an asterisk in red). The bottom two panels show the transcription levels assayed by RNA-
880 Seq on LCLs assayed by ENCODE. Note intron 9-10 of *TYK2* is clearly expressed. The alignability of
881 75-mers by GEM is also shown to show the mapability of reads around rs2304256.
882
883 **Figure 3. Breakdown of autoimmune associated causal *cis*-eQTLs using RNA-Seq**
884 (A) Percentage and number of causal *cis*-eQTL associations detected per RNA-Seq quantification type,
885 following LD pruning of associated SNPs from twenty autoimmune diseases to 560 independent
886 susceptibly loci. The top chart shows the number of causal *cis*-eQTLs when combining all RNA-Seq
887 profiling types together (20%). (B) Sharing of causal *cis*-eQTL associations per quantification type (110
888 detected in total). Percentage of causal *cis*-eQTLs captured are shown as a percentage of the 110 total.
889 (C) Total causal *cis*-eQTLs per disease across all five levels of RNA-Seq quantification, using the 20
890 diseases of the ImmunoBase resource. In orange are disease-associated SNPs that show no shared
891 association with expression across any quantification type. In blue are the disease-associated SNPs that
892 are also causal *cis*-eQTLs. (D) Causal *cis*-eQTLs and candidate genes per disease broken down by
893 quantification type.
894

895 **Figure 4. Functional annotation of causal autoimmune *cis*-eQTLs**

896 (A) We took the causal autoimmune *cis*-eQTLs detected for each RNA-Seq quantification type and
897 performed enrichment testing for chromatin state segmentation and histone marks in LCLs taken from
898 the NIH Roadmap Epigenomics Project. We used the GoShifter algorithm to do this (see methods);
899 which takes all SNPs in strong LD (r^2>0.8) with the causal *cis*-eQTLs and calculates the proportion of
900 SNPs overlapping chromatin marks, the positions of the marks are then shuffled whilst retaining the
901 SNP positions, and the fraction of overlap recalculated over 1,000 permutations. A permutation *P*-value
902 is then generated – which is annotated in each box (*P*<0.05 deemed significant). The heat colour is
903 representative of the permutation *P*-value. Significant enrichment tests are highlighted in bold. The total
904 number of causal *cis*-eQTLs per quantification type are annotated at the bottom of the heatmap. (B) The
905 percentage of causal *cis*-eQTLs in chromatin regulatory marks per quantification type. An asterisk
906 shows that this level of enrichment is deemed to be significant as shown in panel A. (C) The percentage
907 of causal *cis*-eQTLs in chromatin regulatory marks per quantification type that are or are highly
908 correlated (r^2>0.8) with SNPs that alter splice site consensus sequences of the target genes (assessed by
909 Sequence Ontology for the hg19 GENCODE v12 reference annotation).

910

911 ## Supporting information

912

913 **S1 Table.** SLE GWAS in persons of European Descent (38 loci taken forward for *cis*-eQTL analysis).

914

915 **S2 Table.** SLE associated *cis*-eQTL associations deemed to be causal as defined by the JLIM pipeline

916 (this is the output from JLIM).

917

918 **S3 Table.** All SLE associated *cis*-eQTL associations by the JLIM pipeline – causal and non-causal

919 associations (provided as a separate XLSX).

920

921 **S4 Table.** Functional annotation of SLE candidate genes detected by *cis*-eQTL analysis using RNA-

922 Seq.

923

924 **S5 Table.** Number of expression elements that are deemed to have a causal association with the SLE

925 risk SNP.

926

927 **S6 Table.** Curated studies of the ImmunoBase Resource.

928

929 **S1 Fig.** Overview of the five quantification types used to estimate gene expression using RNA-Seq.

930

931 **S2 Fig.** Distribution of joint likelihood *P*-values across RNA-Seq quantification types with 38 SLE

932 GWAS loci.

933

934 **S3 Fig.** Specificity of *cis*-eQTLs and candidate genes identified by joint likelihood mapping using SLE

935 GWAS across the five RNA-Seq quantification types.

936

937 **S4 Fig.** Regional association plots (+/-250kb) of SLE GWAS in Europeans – showing the nine loci that
938 are causal *cis*-eQTLs and candidate genes from JLIM analysis. The full results of this analysis are in
939 Table 3 of the manuscript and the summary results from the GWAS as provided in S1 Table. Candidate
940 genes are highlighted in red.
941
942 **S5 Fig.** SLE associated SNP rs3768792 is a causal *cis*-eQTL for *IKZF2* for a single exon and a single
943 intron.
944
945 **S6 Fig.** SLE associated SNP rs7444 is a causal *cis*-eQTL for *UBE2L3* for a single transcript and a single
946 exon.
947
948 **S7 Fig.** SLE associated SNP rs9872955 is a causal *cis*-eQTL for *LYST* for a single junction.
949
950 **S8 Fig.** Exon and intron numbers for *TYK2* (corresponding to Figure 2). The transcription start site is
951 on the right of the diagram.
952
953 **S9 Fig.** Processing of genotype data and principle component analysis. Genotype data in VCF format
954 of 1000Genomes individuals were downloaded from E-GEUV1 (ArrayExpress). Insertion-deletion
955 sites were removed, and bi-allelic SNPs kept only. SNPs with HWE < 0.0001 were removed and the
956 VCF converted to 0,1,2 format using PLINK. Principle component analysis was performed on genotype
957 data using the R package SNPRelate on chromosome 20. The first 3 components were included in the
958 eQTL regression model as well as the binary imputation status (see methods).
959
960 **S10 Fig:** Overview of integrative *cis*-eQTL analysis pipeline using 20 autoimmune diseases
961

962 # Tables

963

Table 1. Number of *cis*-eQTLs driven by the same causal variant as the SLE disease association (total number of SLE loci: 38)

	Gene	Transcript	Exon	Junction	Intron	Total
Causal *cis*-eQTLs[a]	2	2	7	4	4	9[b]
% of 38 SLE GWAS loci	5.3	5.3	18.4	10.5	10.5	23.7
% of total causal eQTLs	22.2	22.2	77.8	44.4	44.4	100
Candidate genes	3	4	9	5	5	12
Expression targets[c]	2	7	24	18	13	64

The lead SNPs from the *Bentham and Morris et al 2015* GWAS in persons of European descent were functionally annotated by *cis*-eQTL analysis in the Geuvadis RNA-Seq cohort in lymphoblastoid cell lines using RNA-Seq quantification profiled at five resolutions (gene, transcript, exon, junction, and intron). Only SNPs reaching genome-wide significance, not conditional peaks, outside of the major histocompatibility complex loci, and with minor allele frequency > 5% were included leaving 38 SLE lead SNPs in total. All SLE loci were densely imputed to the 1000 Genomes Phase 3 Imputation Panel as described in methods.
All 38 loci (+/-100kb of each lead SNP) comprised a nominally significant *cis*-eQTL ($P<0.01$) for at least one gene within +/-500kb of the lead SNP at each resolution of RNA-Seq. Evidence of a single shared causal variant at each locus was assessed using the Joint Likelihood Mapping (JLIM) algorithm as described in methods.
[a]Number of loci where the disease association is consistent with a single shared effect for at least one *cis*-eQTL ($P<0.01$ and JLIM FDR adjusted $P<0.05$). [b]The total number of unique causal *cis*-eQTLs across all RNA-Seq quantification types. [c]Expression targets corresponds to the quantification type in hand (i.e. number of exons at exon-level).

964

965

Table 2. Pairwise comparison of the number of *cis*-eQTLs with a nominal JLIM *P*-value < 0.01

Quantification type X	Quantification type Y	Total matched *cis*-eQTLs (SNP ~ gene pairs $P < 0.01$)	% Shared causal variant in X and Y (JLIM $P < 0.01$)	% Shared causal variant in X only (JLIM $P < 0.01$)	% Shared causal variant in Y only (JLIM $P < 0.01$)	% No shared causal variant in X and Y (JLIM $P < 0.01$)	Correlation of JLIM *P* (X ~ Y)
Gene	Transcript	267	3.00	1.87	5.62	89.51	0.63
Gene	Exon	296	3.72	1.01	8.78	86.49	0.57
Gene	Junction	229	3.49	1.75	11.79	82.97	0.46
Gene	Intron	252	1.59	3.57	5.56	89.29	0.35
Transcript	Exon	325	3.08	5.54	9.54	81.85	0.38
Transcript	Junction	261	3.07	5.75	12.64	78.54	0.29
Transcript	Intron	279	2.15	6.45	5.73	85.66	0.24
Exon	Junction	294	6.12	7.82	9.86	76.19	0.44
Exon	Intron	314	2.87	10.83	4.78	81.53	0.34
Junction	Intron	275	3.27	13.45	5.09	78.18	0.20

This table is complementary to the data in Figure 1. We only considered matched SNP-gene *cis*-eQTL association pairs that had a nominal *cis*-eQTL association *P*-value < 0.01 in both quantification types, and to be conservative, when multiple transcripts, exons, junctions, and introns were annotated with the same gene symbol, we selected the associations that minimized the difference in JLIM *P*-value between matched SNP-gene *cis*-eQTLs across RNA-Seq quantification types. The first row for example is a pairwise comparison of matched SNP-gene pairs between gene-level and transcript-level quantification (of which there are 267 matched pairs). 3% of these are deemed nominally causal (JLIM $P < 0.01$) at both gene-level and transcript, 1.87% at gene-level only and 5.62% at transcript-level only. 89.51% of matched SNP-gene pairs between gene- and transcript-level do not possess a nominally causal *cis*-eQTL. Pearson's correlation was performed for matched SNP-gene JLIM *P*-value pairs. These data show that exon- and junction-level analysis consistently capture the majority of potentially causal cis-eQTL associations. JLIM: joint likelihood mapping.

966

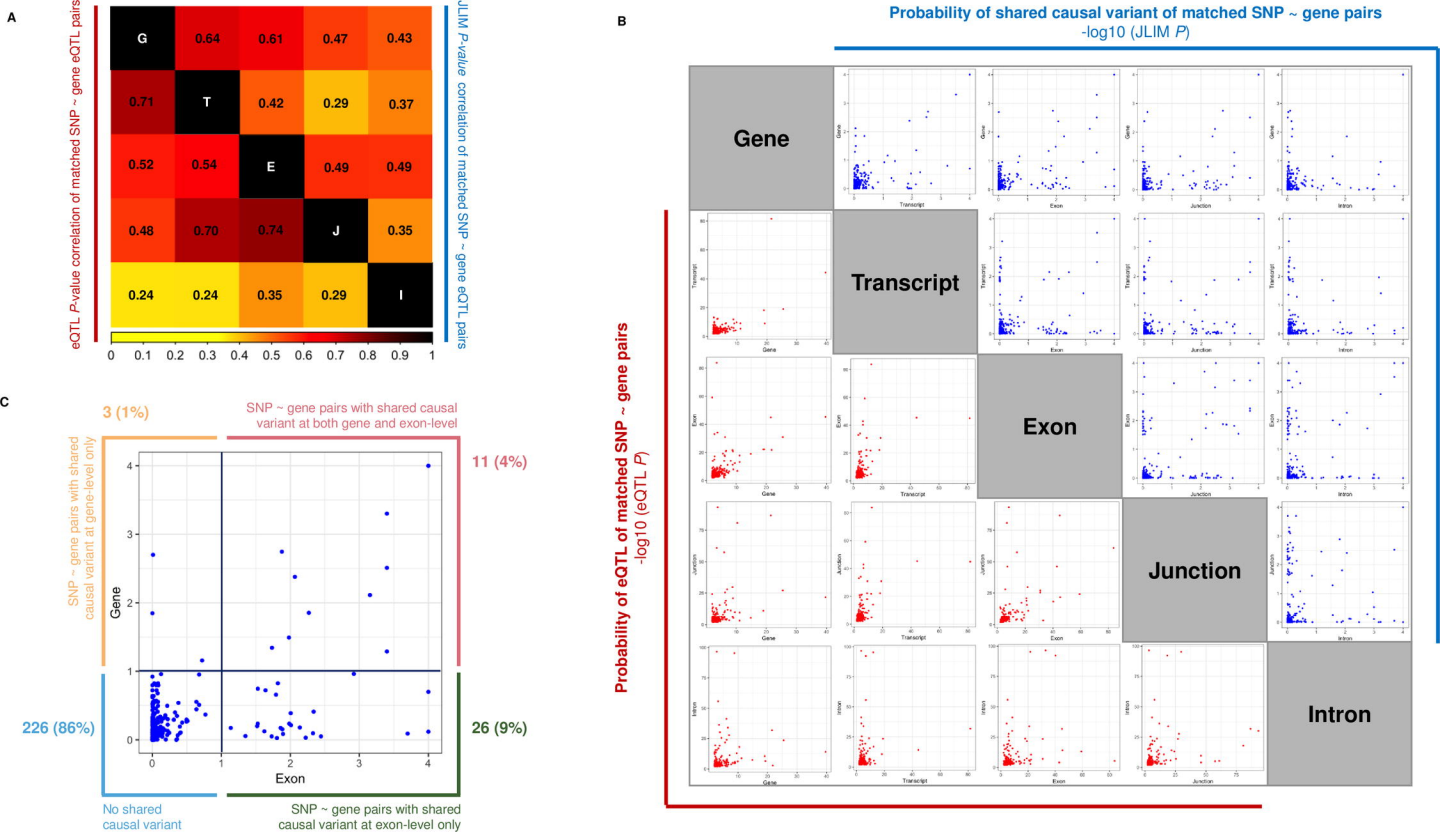
967
968

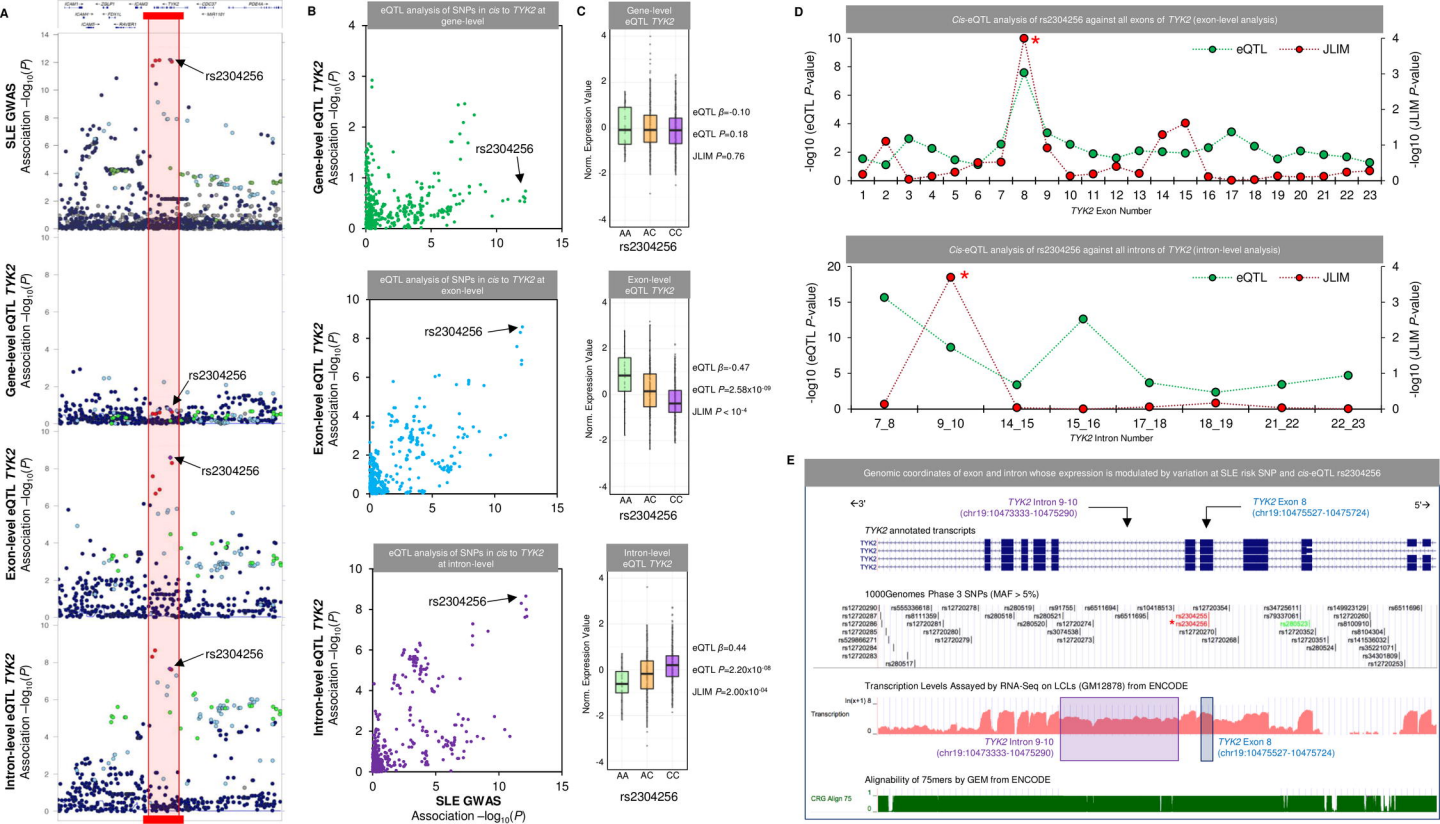
Table 3. Nine SLE loci contain *cis*-eQTLs driven by the same variant as the disease association

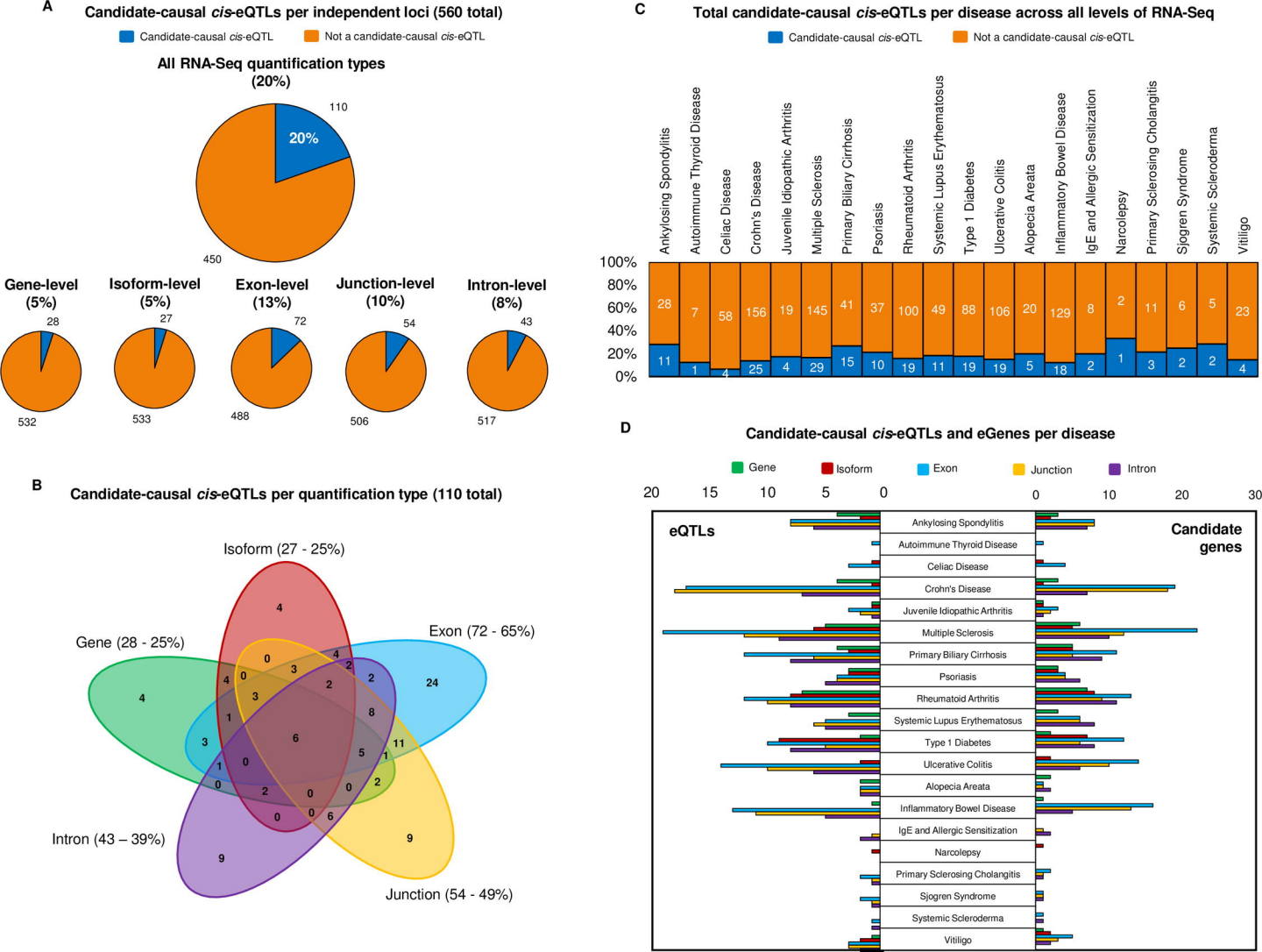
		Gene		Transcript		Exon		Junction		Intron	
Lead SNP	Gene	eQTL P^a	JLIM *P*	eQTL *P*	JLIM *P*	eQTL *P*	JLIM *P*	eQTL *P*	JLIM *P*	eQTL *P*	JLIM *P*
rs2476601	*PHTF1*	-	-	2.2×10^{-3}	6.2×10^{-1}	5.0×10^{-8}	1	8.4×10^{-47}	1	**1.4×10^{-4}**	**1.0×10^{-4}**
rs1801274	*ARHGAP30*	2.4×10^{-6}	8.1×10^{-1}	-	-	**1.1×10^{-4}**	**2.0×10^{-4}**	9.4×10^{-3}	7.4×10^{-3}	1.2×10^{-3}	4.8×10^{-1}
rs9782955	*LYST*	5.4×10^{-3}	3.90×10^{-1}	8.0×10^{-6}	9.8×10^{-1}	1.6×10^{-3}	4.6×10^{-3}	**1.3×10^{-3}**	**2.0×10^{-4}**	1.0×10^{-5}	5.0×10^{-1}
rs3768792	*IKZF2*	-	-	1.5×10^{-3}	7.7×10^{-1}	**1.9×10^{-4}**	**3.0×10^{-4}**	1.0×10^{-5}	9.0×10^{-1}	**1.1×10^{-5}**	**2.0×10^{-4}**
rs10028805	*BANK1*	1.8×10^{-3}	3.1×10^{-3}	4.9×10^{-3}	3.2×10^{-3}	**1.8×10^{-5}**	**4.0×10^{-4}**	**2.5×10^{-4}**	**2.0×10^{-4}**	1.8×10^{-4}	9.7×10^{-1}
rs2736340	*BLK*	**3.2×10^{-26}**	**$< 10^{-4}$**	**1.0×10^{-9}**	**$< 10^{-4}$**	**1.4×10^{-31}**	**$< 10^{-4}$**	**7.6×10^{-28}**	**$< 10^{-4}$**	**3.1×10^{-24}**	**$< 10^{-4}$**
	FAM167A	**2.3×10^{-40}**	**$< 10^{-4}$**	**4.4×10^{-45}**	**$< 10^{-4}$**	**5.1×10^{-46}**	**$< 10^{-4}$**	**1.5×10^{-22}**	**$< 10^{-4}$**	**7.4×10^{-15}**	**$< 10^{-4}$**
rs2286672	*RABEP1*	1.4×10^{-3}	5.1×10^{-2}	1.3×10^{-4}	9.4×10^{-1}	**7.4×10^{-5}**	**4.0×10^{-4}**	4.5×10^{-4}	7.0×10^{-4}	1.3×10^{-4}	8.5×10^{-1}
rs2304256	*TYK2*	1.2×10^{-3}	7.6×10^{-1}	9.9×10^{-6}	9.9×10^{-1}	**2.5×10^{-9}**	**$< 10^{-4}$**	1.3×10^{-4}	3.0×10^{-3}	**2.2×10^{-9}**	**2.0×10^{-4}**
	ATG4D	-	-	3.8×10^{-3}	7.2×10^{-3}	6.4×10^{-5}	3.8×10^{-3}	**3.8×10^{-4}**	**2.0×10^{-4}**	6.6×10^{-5}	9.7×10^{-1}
rs7444	*UBE2L3*	5.7×10^{-3}	2.0×10^{-1}	**5.9×10^{-14}**	**$< 10^{-4}$**	**9.9×10^{-5}**	**$< 10^{-4}$**	5.1×10^{-5}	9.5×10^{-1}	1.2×10^{-3}	9.0×10^{-1}
	CCDC116	**2.5×10^{-5}**	**5.0×10^{-4}**	1.4×10^{-6}	3.0×10^{-4}	**4.9×10^{-4}**	**4.0×10^{-4}**	-	-	-	-

Nine of the 38 SLE loci (24%) were found to be driven by the same causal variant as the disease association across all five RNA-Seq quantification types in LCLs (*cis*-eQTL $P<0.01$ and joint likelihood of shared association FDR<0.05). Bold type indicates associations that show evidence of a shared causal variant for *cis*-eQTL and disease. [a]Minimum *cis*-eQTL *P*-value for any SNP within 100 kb of the lead SNP. Dashes (–) indicate genes that were either not detected or had minimum *cis*-eQTL $P>0.01$ in the RNA-Seq quantification type in hand. JLIM *P*-values $<10^{-4}$ indicates the JLIM statistic is more extreme than permutation. JLIM: joint likelihood mapping. If multiple SNP-unit associations are deemed to be causal (i.e. one SNP shows a causal association to two exons of the same gene, the association with the smallest JLIM *P*-value is reported).

969


A
eQTL P-value correlation of matched SNP ~ gene eQTL pairs
JLIM P-value correlation of matched SNP ~ gene eQTL pairs
G 0.64 0.61 0.47 0.43
0.71 T 0.42 0.29 0.37
0.52 0.54 E 0.49 0.49
0.48 0.70 0.74 J 0.35
0.24 0.24 0.35 0.29 I
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
B
Probability of shared causal variant of matched SNP ~ gene pairs
-log10 (JLIM P)
Probability of eQTL of matched SNP ~ gene pairs
-log10 (eQTL P)
Gene
Transcript
Exon
Junction
Intron
C
3 (1%)
SNP ~ gene pairs with shared causal variant at both gene and exon-level
11 (4%)
SNP ~ gene pairs with shared causal variant at gene-level only
Gene
Exon
226 (86%)
26 (9%)
No shared causal variant
SNP ~ gene pairs with shared causal variant at exon-level only


A
SLE GWAS
Association −log10(P)
Gene-level eQTL TYK2
Exon-level eQTL TYK2
Intron-level eQTL TYK2
rs2304256
B
eQTL analysis of SNPs in cis to TYK2 at gene-level
eQTL analysis of SNPs in cis to TYK2 at exon-level
eQTL analysis of SNPs in cis to TYK2 at intron-level
Gene-level eQTL TYK2 Association −log10(P)
Exon-level eQTL TYK2 Association −log10(P)
Intron-level eQTL TYK2 Association −log10(P)
SLE GWAS Association −log10(P)
C
Gene-level eQTL TYK2
Exon-level eQTL TYK2
Intron-level eQTL TYK2
Norm. Expression Value
eQTL β=-0.10
eQTL P=0.18
JLIM P=0.76
eQTL β=-0.47
eQTL P=2.58x10-09
JLIM P < 10-4
eQTL β=0.44
eQTL P=2.20x10-08
JLIM P=2.00x10-04
AA AC CC
D
Cis-eQTL analysis of rs2304256 against all exons of TYK2 (exon-level analysis)
eQTL
JLIM
-log10 (eQTL P-value)
-log10 (JLIM P-value)
TYK2 Exon Number
Cis-eQTL analysis of rs2304256 against all introns of TYK2 (intron-level analysis)
TYK2 Intron Number
E
Genomic coordinates of exon and intron whose expression is modulated by variation at SLE risk SNP and cis-eQTL rs2304256
←3'
5'→
TYK2 Intron 9-10 (chr19:10473333-10475290)
TYK2 Exon 8 (chr19:10475527-10475724)
TYK2 annotated transcripts
1000Genomes Phase 3 SNPs (MAF > 5%)
Transcription Levels Assayed by RNA-Seq on LCLs (GM12878) from ENCODE
Transcription
Alignability of 75mers by GEM from ENCODE
CRG Align 75


A
Candidate-causal cis-eQTLs per independent loci (560 total)
Candidate-causal cis-eQTL
Not a candidate-causal cis-eQTL
All RNA-Seq quantification types
(20%)
20%
110
450
Gene-level
(5%)
28
532
Isoform-level
(5%)
27
533
Exon-level
(13%)
72
488
Junction-level
(10%)
54
506
Intron-level
(8%)
43
517
B
Candidate-causal cis-eQTLs per quantification type (110 total)
Isoform (27 - 25%)
Gene (28 - 25%)
Exon (72 - 65%)
Intron (43 – 39%)
Junction (54 - 49%)
C
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%
80%
60%
40%
20%
0%
D
Candidate-causal cis-eQTLs and eGenes per disease
Gene
Isoform
Exon
Junction
Intron
eQTLs
Candidate genes

A

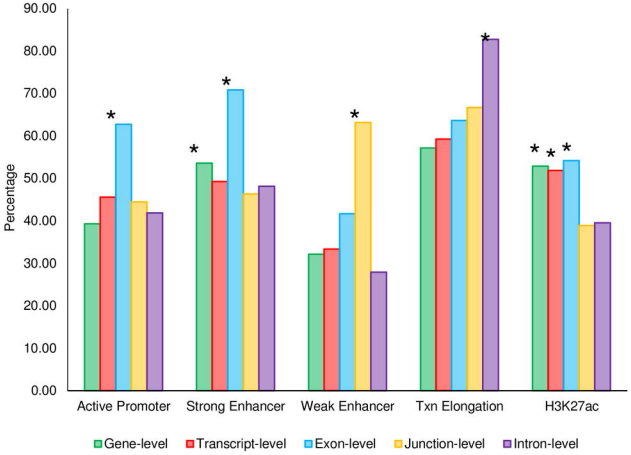
Enrichment of causal *cis*-eQTLs in chromatin regulatory elements

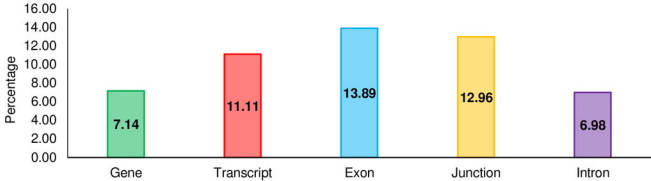
		Gene-level	Transcript-level	Exon-level	Junction-level	Intron-level
Chromatin state segmentations	Active Promoter	0.286	0.086	**0.016**	0.284	0.463
	Weak Promoter	0.944	0.668	0.450	0.41	0.364
	Poised Promoter	0.662	0.729	0.560	0.859	0.687
	Strong Enhancer	**0.036**	0.369	**0.047**	0.383	0.188
	Weak Enhancer	0.282	0.418	0.563	**0.002**	0.451
	Txn Transition	0.403	0.948	0.727	0.634	0.909
	Txn Elongation	0.339	0.346	0.340	0.386	**0.001**
Chromatin modification	DNase	0.368	0.076	0.804	0.585	0.131
	H2A.Z	0.234	0.416	0.222	0.177	0.517
	H3K27ac	**0.002**	0.039	**0.005**	0.121	0.185
	H3K36me3	0.153	0.326	0.192	0.178	0.73
	H3K4me1	0.812	0.765	0.662	0.599	0.476
	H3K4me2	0.204	0.195	0.191	0.141	0.091
	H3K4me3	0.076	0.282	0.373	0.184	0.901
	H3K79me2	0.508	0.858	0.511	1.181	0.287
	H3K9ac	0.217	0.811	0.805	0.061	0.074
	H3K9me3	1	0.937	0.617	0.884	0.815
Total causal *cis*-eQTLs		**28**	**27**	**72**	**54**	**43**

Permutation *P*-value of enrichment

B

Percentage of causal *cis*-eQTLs in chromatin regulatory elements



C

Percentage of causal *cis*-eQTLs tagged by splice SNPs