The length of the expressed 3' UTR is an intermediate molecular phenotype linking genetic variants to complex diseases

Elisa Mariella[1] Federico Marotta[1] Elena Grassi[1] Stefano Gilotto[1]
Paolo Provero[1,2,*]

[1] Dept. of Molecular Biotechnology and Health Sciences, Molecular Biotechnology Center, University of Turin, Turin, Italy
[2] Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milan, Italy
[*] Corresponding author: paolo.provero@unito.it

Abstract

In the last decades, genome wide association studies (GWAS) have uncovered tens of thousands of associations between common genetic variants and complex diseases. However, these statistical associations can rarely be interpreted functionally and mechanistically. As the majority of the disease-associated variants are located far from coding sequences, even the relevant gene is often unclear. A way to gain insight into the relevant mechanisms is to study the genetic determinants of intermediate molecular phenotypes, such as gene expression and transcript structure. We propose a computational strategy to discover genetic variants affecting the relative expression of alternative 3' untranslated region (UTR) isoforms, generated through alternative polyadenylation, a widespread post-transcriptional regulatory mechanism known to have relevant functional consequences. When applied to a large dataset in which whole genome and RNA sequencing data are available for 373 European individuals, 2,530 genes with alternative polyadenylation quantitative trait loci (apaQTL) were identified. We analyze and discuss possible mechanisms of action of these variants, and we show that they are significantly enriched in GWAS hits, in particular those concerning immune-related and neurological disorders. Our results point to an important role for genetically determined alternative polyadenylation in affecting predisposition to complex diseases, and suggest new ways to extract functional information from GWAS data.

1 Introduction

Understanding the relationship between human genotypes and phenotypes is one of the central goals of biomedical research. The first sequencing of the human genome [1, 2] and the following large-scale investigations of genetic differences between individuals by efforts such as the 1000 Genome Project Consortium [3] provided the foundation for the study of human genetics at the genome-wide level.Enrichment of trait-specific GWAS hits among apaQTL

Genome wide association studies (GWAS) examine common genetic variants to identify associations with complex traits, including common diseases. Long lists of genetic associations with disparate traits have been obtained, but their functional interpretation is far from being straightforward [4]. Indeed, because of linkage disequilibrium, GWAS identify genomic regions carrying multiple variants among which it is not possible to identify the causal ones without additional information. Furthermore, most loci identified in human GWAS are in non-coding regions, presumably exerting regulatory effects, but usually we do not known the identity of the affected gene or the molecular mechanism involved.

A possible way to gain insight into the mechanisms behind GWAS associations is to investigate the effect of genetic variants on intermediate molecular phenotypes, such as gene expression [5, 6]. Expression quantitative trait loci (eQTL) are genomic regions carrying one or more genetic variants affecting gene expression. Besides their intrinsic interest in understanding the control of gene expression, eQTL studies can be exploited for the interpretation of GWAS results, helping to prioritize likely causal variants and supporting the formulation of mechanistic hypotheses about the links between genetic variants and diseases.

Recent studies have shown that genetic variants acting on the whole RNA processing cascade are at least equally common as, and largely independent from, those that affect transcriptional activity, and that they can be a major driver of phenotypic variability in humans [7]. Therefore it is important to identify the genetic variants associated to transcript structure, including splicing and alternative UTR isoforms, besides those affecting transcriptional levels, and different approaches have been proposed to this end [8, 9]. From these studies it emerges in particular that genetic variants frequently determine changes in the length of the expressed 3' UTRs. In addition, genome-wide analyses specifically focused on alternative splicing have been performed [10, 11].

Polyadenylation is one of the post-transcriptional modifications affecting pre-mRNAs in the nucleus and involves two steps: the cleavage of the transcript and the addition of a poly(A) tail [12, 13]. The most important regulatory elements involved are the polyadenylation signal (PAS) and other cis-elements, usually located within the 3'UTR, but multiple and diversified regulatory mechanisms have been described [14, 15]. The PAS is recognized by the cleavage and

polyadenylation specificity factor (CPSF) that, together with other protein complexes, induces the cleavage of the transcript in correspondence of the downstream poly(A) site. The large majority of human genes have multiple poly(A) sites, so that alternative polyadenylation (APA) is a widespread phenomenon contributing to the diversification of the human transcriptome through the generation of alternative mature transcripts with different 3' ends. Such transcripts are translated into identical proteins, but protein level, localization, and even interactions can depend on the 3' end of the transcript [16].

APA events have been grouped into classes based on the location of the alternative poly(A) site and the type of change determined by their differential usage [12]. In this work we have taken into consideration only the simplest and most frequent mode (tandem 3'UTR APA), in which two poly(A) sites located within the same terminal exon, one in a proximal and one in a distal position, produce transcripts that differ only in the length of the 3'UTR. Such variation in 3'UTR length can have an important functional impact, for example by affecting the binding of microRNAs and RNA binding proteins and thus transcript abundance, translation and localization. Moreover, APA regulation is strongly tissue- and cell type-dependent [17, 18, 19, 20, 21] and several examples are known of altered APA regulation associated to human diseases [7, 22].

How genetic variants influence APA has not been comprehensively investigated in a large human population yet. A recent analysis of whole genome sequencing (WGS) data from [8] found hundreds of common single nucleotide polymorphisms (SNPs) causing the alteration or degradation of motifs that are similar to the canonical PAS [23], but did not extend the analysis to other possible mechanisms. Other studies found strong associations between genetic variants and APA regulation [24, 25, 26, 8, 9, 27], but a systematic investigation based on a large number of samples and variants, specifically targeted to APA rather than generically to transcript structure, and unbiased in the choice of variants to examine, is not yet available.

Here we propose a new computational strategy for the genome-wide investigation of the influence of genetic variants on the expression of alternative 3'UTR isoforms in a large population. In particular, we analyzed WGS data paired with standard RNA-Seq data obtained in 373 European (EUR) individuals [8]. Statistically, our approach is analogous to methods commonly implemented in eQTL mapping analysis and it aims to overcome the limitations illustrated above for the specific purpose of correlating variants to 3' UTR isoforms.

A central task, preliminary to the analysis of genetic variants, is thus the quantification of the alternative 3'UTR isoforms. Various strategies have been implemented to this end, from custom analysis pipelines for microarray data [28], to the development of next-generation sequencing technologies specifically targeted to the 3' end of transcripts, such as the serial analysis of gene expression (SAGE) [20] and sequencing of APA sites (SAPAs) [21], allowing also the identification of previously unannotated APA sites.

More recently, tools able to capture APA events from standard RNA-Seq data have been developed. In general, these approaches can be divided into two categories: those that exploit previous annotation of poly(A) sites [29, 30], such the ones provided by PolyA_DB2 [31] and APASdb [32], and those that instead try to infer their location from the data [17]. Although the latter approach potentially allows analyzing also previously unannotated sites, the former leads to higher sensitivity [29, 30], and was thus preferred in this study. Undoubtedly, approaches based on standard RNA-Seq are not as powerful and accurate as technologies that specifically sequence the 3' ends. However, they allow studying this phenomenon in an uncomparably larger number of samples and conditions, including the recently generated large-scale transcriptomic datasets of normal individuals that we use in this work.

2 Results

2.1 Genetic variants affect the relative expression of alternative 3'UTR isoforms of thousands of genes

In order to investigate the effect of human genetic variants on the expression of alternative 3'UTR isoforms, we developed a computational approach similar to the one commonly used for eQTL analysis (Fig. 1). It was applied to a large dataset in which WGS data paired with RNA-Seq data are available for 373 European (EUR) individuals (GEUVADIS dataset [8]). A collection of known alternative poly(A) sites [31] was used, together with a compendium of human transcripts, to obtain an annotation of alternative 3'UTR isoforms that was then combined with RNA-Seq data in order to compute, for each gene, the expression ratio between short and long isoform (m/M value) in each individual.

Linear regression was then used to identify associations between the m/M values of each gene and the genetic variants within a cis-window including the gene itself and all sequence located within 1Mbp from the transcription start site (TSS) or the transcription end site (TES). This led to the fitting of ∼30 million linear models, involving ∼6,300 genes and ∼5.3 million variants. About 190,000 models, involving 2,530 genes and ∼160,000 variants, revealed a significant association (Fig. 2, Tab. 1, Supplementary Files 3 and 4).

Our set of significant genes shows only moderate overlap with genes for which eQTLs or transcript ratio QTLs (trQTLs) were reported in Ref. [8] from the same data (Fig. 3). Alternative polyadenylation can result in changes in gene expression levels as a consequence of the isoform-dependent availability of regulatory elements affecting the stability of transcripts, such as microRNA binding sites [13]. In this case, apaQTLs should also be eQTLs. However, APA may also have effects that do not imply changes in expression levels, including the modulation of

Table 1: Number of models, genes and variants.

	Total	Significant*
Models	30,136,480	192,715
Genes	6,256	2,530
Variants	5,309,860	160,223

*corrected empirical P-value < 0.05

mRNA translation rates [33, 34] and localization [35], and protein cytoplasmic localization [36]. Similarly, a complete overlap with trQTLs is not expected, because they were identified by taking into account all the annotated alternative transcripts of a gene including alternative splicing and transcription initiation: The identification of apaQTLs for several genes for which trQTLs were not identified suggests that focusing on a specific class of transcript structure allows higher sensitivity.

These results show that a large number of genetic determinants of alternative polyadenylation can be inferred from the analysis of standard RNA-Seq data paired with the genotypic characterization on the same individuals.

2.2 apaQTLs are preferentially located within active genomic regions

Just like eQTLs, we expect apaQTLs be located within genomic regions that are active in the relevant cell type (lymphoblastoid cells for our data). In order to verify this hypothesis, we superimposed the apaQTLs to the ChromHMM annotation of the human genome for the GM12878 cell line [37], and used logistic regression, as detailed in the Methods, to determine the enrichment or depletion of apaQTLs for each chromatin state, expressed as an odds ratio (OR). As expected, significant ORs > 1 were obtained for active genomic regions, such as transcribed regions, promoters and enhancers, suggesting that genetic variants have a higher probability of being apaQTLs when they are located in active regions. Conversely, apaQTLs were depleted in repressed and inactive chromatin states. Similar results were obtained using broad chromatin states (Fig. 4), defined following [37], or all 15 chromatin states reported by ChromHMM (Supplementary Fig. 3).

As a control, the same enrichment analysis was performed with the ChromHMM annotation obtained in a different cell type, namely normal human epithelial keratinocytes (NHEK). All NHEK active chromatin states showed a reduced enrichment in apaQTLs compared with GM1278, and regions repressed in NHEK cells actually showed significant enrichment of lymphoblastoid apaQTLs (Supplementary Fig. 4 and Supplementary Fig. 5). Taken together, these results show

that genetic variants affecting alternative polyadenylation tend to be located in cell-type specific active chromatin regions.

In the following, we will divide apaQTLs in two classes: Intragenic apaQTLs are those located inside one of the genes whose isoform ratio we are able to analyze, while all other apaQTLs will be referred to as extragenic (note that these might be located inside a gene for which we are unable to perform the analysis, for one of the reasons explained in the Methods).

2.3 Intragenic apaQTLs are enriched in coding exons and 3'UTRs

Having established that genetic variants have a widespread influence of the expression of alternative 3'UTR isoforms, we turned to their putative mechanisms of action. First of all, we considered the distribution of intragenic apaQTLs among regions contributing to the mRNA vs. introns. As shown in Fig. 5, intragenic apaQTLs are enriched in coding exons and 3' UTRs, and depleted in introns and 5' UTRs. The depletion of introns suggests that most intragenic apaQTLs exert their regulatory role at the transcript level, e.g. by modulating the binding of trans-acting factors to the mRNA.

Among mRNA regions, the enrichment of 3' UTRs is expected, since these regions contain several elements involved in the regulation of both alternative polyadenylation and mRNA stability. The enrichment of coding exons could be ascribed to regulatory elements residing in these portions of the mRNAs, or to residual effects of linkage disequilibrium (LD) with variants located in the 3' UTR, notwithstanding the LD pruning procedure implemented in the enrichment analysis (see Methods). Note that while several poly(A) sites are located upstream of the last exon [38], within both intronic sequences and internal exons, such sites were not taken into account in our analysis. Finally, the depletion of 5' UTRs might be due to the distance of these elements from the polyadenylation loci, and to the fact that these regions are mostly involved in other regulatory mechanisms, such as translational regulation [39]. In the following, we examine in more detail three possible mechanisms by which intragenic apaQTLs could exert their action.

2.3.1 Creation and destruction of PAS motifs

The first possibility is direct interference with the APA regulation, favoring the production of one of the two isoforms in individuals with a particular genotype. A comprehensive atlas of high-confidence PAS has been recently reported [40]. In addition to the canonical PAS motifs (AAUAAA and AUUAAA) it contains 10 previously known signals and 6 new motifs. Exploiting this resource, we were able to identify SNPs that cause the creation or the destruction of putative functional PAS motifs and, as expected, we found that they were enriched among apaQTLs (OR

= 1.72, 95% confidence interval (CI) = 1.08 - 2.75, P-value = 0.0215). In total, 42 PAS-altering variants were found to be apaQTLs of the gene in which they reside. While expected, this result can be considered to validate our strategy.

A few examples are worth discussing in detail. SNP rs10954213 was shown by several studies [41, 24, 42] to determine the preferential production of the short isoform of the IRF5 transcription factor through the conversion of an alternative PAS motif (AAUGAA) into the canonical one (AAUAAA) in a proximal position within the 3'UTR. Consistently, we found that this variant is associated with higher prevalence of the short isoform (Fig. 6A,B). Moreover, the same variant was associated to higher risk of systemic lupus erythematosus (SLE), and higher IRF5 expression, that could be due to the loss of AU-rich elements (ARE) in the short transcript isoform [24]. Globally, these findings are in agreement with the known involvement of IRF5 in several pathways that are critical for the onset of SLE (Type I IFN production, M1 macrophage polarization, autoantibody production, and induction of apoptosis [43]).

A similar trend was detected in the case of the rs9332 variant, located within the 3'UTR of the MTRR gene, encoding an enzyme essential for methionine synthesis (Supplementary Fig. 6A). This variant was reported to be associated with a higher risk of spina bifida, along with other variants within the same gene [44]. We found that the variant is associated with the increased relative expression of the short isoform of the MTRR transcript, as a consequence of the creation of a proximal canonical PAS. We can thus speculate that, similarly to what was shown for IRF5, this post-transcriptional event could lead to a variation in the activity of the enzyme activity and ultimately to increased disease susceptibility.

The same mechanism might provide putative mechanistic explanations for associations found by GWAS studies. For example we found the variant rs5855 to be an apaQTL for the PAM gene (Supplementary Fig. 6B), essential in the biosynthesis of peptide hormones and neurotransmitters [45, 46, 47]. No eQTLs or trQTLs for this gene were revealed by the analysis of the same data reported in [8]. This variant replaces an alternative PAS motif (AGUAAA) with the canonical AAUAAA, thus presumably increasing its strength. This PAS motif is located 26 bps upstream of an APA site corresponding to a 3'UTR of ∼450 bps, instead of the ∼2,000 bps of the canonical isoform, lacking several predicted microRNA binding sites. Indeed, our analysis revealed a shortening of the 3'UTR in individuals with the alternate allele, i.e. the canonical PAS motif. Notably, the variant is in strong LD ($R^2 = 0.90$) with the intronic variant rs10463554, itself an apaQTL for PAM, which has been associated to Parkinson's disease in a recent meta-analysis of GWAS studies [48].

The creation of a distal canonical PAS motif can lead instead to 3' UTR lengthening, as shown by variant rs2842980 located in the 3' UTR of SOD2 (Fig. 6C,D). The alternate allele creates a canonical AUUAAA PAS site, whereas the reference allele is UUUAAA, 19 bps upstream of the most distal annotated poly(A) site. Also this variant is in strong LD ($R^2 = 0.93$) with an

intergenic GWAS hit, namely variant rs2842992, which has been associated to atrophic macular degeneration [49]. A mechanistic involvement of SOD2 in macular degeneration is supported by a mouse model in which the gene was deleted in the retinal pigment epithelium, inducing oxidative stress and key features of age-related macular degeneration [50].

Conversely, the destruction of a canonical, proximal PAS motif leads to shortening of the 3' UTR of BLOC1S2 (Supplementary Fig. 6C). The variant rs41290536 replaces the canonical PAS motif AAUAAA with the non-canonical one AAUGAA 17 bps upstream of a poly(A) site corresponding to a UTR length of ∼750 bps compared to the ∼2,200 of the longest isoform. The variant is in complete LD ($R^2 = 1$) with two variants that have been associated to predisposition to squamous cell lung carcinoma (rs28372851 and rs12765052) [51].

2.3.2 Alteration of microRNA binding

In an alternative scenario, genetic variants can influence the relative expression of alternative 3'UTR isoforms by acting on the stability of transcripts, for example through the creation or destruction of microRNA binding sites. For each gene with alternative 3'UTR isoforms, we divided the 3' UTR into two segments: the "PRE" segment, common to both isoforms, and the "POST" segment expressed only by the longer isoform. Variants altering microRNA binding sites located in the POST segment can result in the variation of the relative isoform expression since they affect only the expression of the long isoform.

For example, we found that the rs8984 variant is associated with an increased prevalence of the long transcript isoform of the CHURC1 gene, an effect that could be due to the destruction of a binding site recognized by microRNAs of the miR-582-5p family within the POST segment of the gene (Supplementary Fig. 7). More generally, we found that apaQTLs are enriched, albeit slightly, among the genetic variants that create or break putative functional microRNA binding sites (OR = 1.15, 95% CI = 1.02 - 1.30, P-value = 0.022). However, we could not find significant agreement between the predicted and actual direction of the change in isoform ratios for these cases. Together with the marginal significance of the enrichment, this result suggests that the alteration of microRNA binding sites is not among the most relevant mechanisms in the genetic determination of 3' UTR isoform ratios.

2.3.3 Alteration of RNA-protein binding

RNA-binding proteins (RBPs) play important roles in the regulation of the whole cascade of RNA processing, including co- and post-transcriptional events. Although many of them have not been fully characterized yet, a collection of 193 positional weight matrices (PWMs) describing

a large number of RNA motifs recognized by human RBPs has been obtained through in-vitro experiments [52]. Here we exploited this resource to identify SNPs that alter putative functional RBP binding sites. Consistently we the involvement of RBPs in the regulation of alternative polyadenylation, mRNA stability and microRNA action, we found a highly significant enrichment of RBP-altering SNPs among intragenic apaQTLs (OR = 1.48, 95% CI = 1.31 - 1.66, P-value = 8.47×10^{-11}).

Specifically, we obtained a positive and significant OR for 20 individual RBP binding motifs (Supplementary Table 1). Although in most cases the enrichment is modest, some of the enriched motifs correspond to RNA-binding domains found in RBPs with a previously reported role in polyadenylation regulation (members of the muscleblind protein family [30, 53], KHDRBS1 [54] and HNRNPC [40]). Other enriched RNA-binding motifs are associated with splicing factors (RBM5, SRSF2, SRSF9 and RBMX) and other RBPs that may be involved in RNA processing (such as members of the MEX3 protein family and HNRNPL). On the contrary, only one significant motif is associated with a RBP that may be involved in RNA degradation (CNOT4 [55]). The involvement of several splicing factors is consistent with evidence supporting a mechanistic interplay between polyadenylation and splicing, that goes beyond the regulation of the usage of intronic poly(A) sites [56, 57, 58, 59, 60].

2.4 Extragenic apaQTLs act in-cis through the perturbation of regulatory elements

Understanding the function of extragenic apaQTLs is less straightforward because, although there are few examples of DNA regulatory elements contributing to APA regulation [14], it is commonly believed that APA is mainly controlled by cis-elements located within transcripts, both upstream and downstream of the poly(A) sites [13].

To further explore this aspect we took advantage of a different annotation of active genome regions, which includes the association between regulatory regions and target genes, namely the cis-regulatory domains (CRDs) identified in lymphoblastoid cell lines in Ref. [61]. Extragenic apaQTLs were indeed found to be enriched in CRDs (OR = 1.73, 95% CI = 1.69 - 1.78, P-value $< 10^{-16}$). The 3D structure of the genome is a key aspect of gene regulation [62], as it determines physical contacts between distal regulatory regions and proximal promoters. In particular, CRDs have been described as active sub-domains within topologically associated domains (TADs), containing several non-coding regulatory elements, both proximal and distal. The perturbation of those regulatory elements by genetic variants can lead to the alteration of gene expression and perhaps interfere with other processes such as alternative polyadenylation, as suggested by our results. Importantly, CRDs have been assigned to the nearby genes they regulate. We could thus observe that extragenic apaQTLs tend to fall within CRDs that have been associated with

their target genes much more frequently than expected by chance. Indeed, this correspondence was verified for 27,527 extragenic apaQTLs, while the same degree of concordance was never obtained in 100 permutations in which each extragenic apaQTL was randomly associated to a gene in its cis-regulatory window (median number of correspondences 12,571). These results suggest an important role of genetic variants located in active, non-transcribed cis-regulatory regions in regulating alternative polyadenylation of the target genes.

2.5 A role for apaQTLs in complex diseases

Since common genetic variation is involved in complex diseases, often by affecting gene regulation, a natural question is whether apaQTLs can be used to provide a mechanistic explanation for some of the genetically driven variability of complex traits, thus adding 3' UTR length to the list of useful intermediate phenotypes. Besides the specific examples discussed above, we found an overall striking enrichment among apaQTLs of genetic variants reported in the NHGRI-EBI GWAS Catalog [63] (OR = 3.17, 95% CI = 3.01 - 3.34, P-value $< 10^{-16}$).

We also investigated the enrichment of each trait category defined by the Experimental Factor Ontology (EFO) and then for each individual trait. In line with the fact that the apaQTL mapping was performed in lymphoblastoid cells, the strongest enrichment was observed for immune system disorders (OR = 5.41, 95% CI = 4.52 - 6.45, P-value = 2.50×10^{-77}) (Fig. 7 and Supplementary Table 2). However, a strong enrichment was also detected for almost all the other tested categories, including neurological disorders (OR = 4.32, 95% CI = 3.86 - 4.83, P-value = 2.47×10^{-142}) and cancer (OR = 3.96, 95% CI = 3.36 - 4.64, P-value = 4.15×10^{-63}).

A significant enrichment was detected for 95 individual complex traits, including several diseases. Among these, the largest ORs were observed for autism spectrum disorder (OR = 42.6, 95% CI = 32.9 - 55.5, P-value = 2.36×10^{-174}), squamous cell lung carcinoma (OR = 26.1, 95% CI = 15.7 - 43.3, P-value = 1.29×10^{-36}), lung carcinoma (OR = 17.9, 95% CI = 12.7 - 25.2, P-value = 9.63×10^{-62}), schizophrenia (OR = 10.6, 95% CI = 9.01 - 12.4, P-value = 1.25×10^{-182}), and HIV-1 infection (OR = 6.51, 95% CI = 3.75 - 10.8, P-value = 2.28×10^{-12}). The complete list of enriched traits can be found in Supplementary File 5.

We observed that apaQTLs that are also GWAS hits often map to genes in the human leukocyte antigen (HLA) locus, suggesting that in at least some cases the enrichment could be mostly driven by this genomic region. Somewhat unexpectedly, this was particularly evident for neurological disorders. In order to clarify this point, we evaluated all enrichments after excluding the variants in the HLA locus. Although in some cases the OR decreased after removing HLA variants, for most GWAS categories the enrichment was still significant (Supplementary Fig. 8 and Supplementary Table 3). For example, we found 155 apaQTLs associated with autism spectrum

disorder, 116 of which affecting HLA genes. After the exclusion of HLA variants, the enrichment was still highly significant (OR = 10.66, 95% CI = 6.92 - 15.95, P-value = 7.05×10^{-29}). On the contrary, the enrichment of variants associated to pulmonary adenocarcinoma is driven by the HLA locus, and becomes non-significant after excluding HLA variants (OR = 1.35, 95% CI = 0.22 - 4.39, P-value = 0.68). The complete list of enriched traits after the exclusion of HLA variants can be found in Supplementary File 6.

2.6 The effect of genetic variants on APA can be confirmed in patients

As briefly discussed above, the rs10954213 variant is associated with a higher risk of SLE. Evidence about the related molecular mechanism arose from the analysis of cell lines derived from healthy individuals [41, 42], and the effect of the variant on IRF5 expression in blood cells was confirmed in SLE patients [64, 65]. However, direct evidence on the effect of this variant on APA regulation in SLE patients is still missing.

In order to assess whether rs10954213 affects IRF5 APA regulation in SLE patients, we analyzed RNA-Seq data derived from whole blood cells in 99 patients [66]. We detected a strong difference in IRF5 m/M values among the three rs10954213 genotypes, with the alternative allele associated with higher m/M values, i.e. shorter 3' UTR (Kruskal-Wallis test P-value = 9.21×10^{-10}; Fig. 8). Therefore the variant has, at least qualitatively, the same effect in the whole blood of SLE patients as in lymphoblastoid cell lines of normal individuals.

3 Discussion

We used a new efficient strategy to study how human genetic variants influence the expression of alternative 3'UTR isoforms. This issue has been previously investigated with different approaches [26, 25, 27, 8, 9]. The method we propose combines wide applicability, being based on standard RNA-Seq data, with the high sensitivity allowed by limiting the analysis to a single type of transcript structure variant, namely 3' UTR length. Such higher sensitivity led us to discover thousands of variants associated with 3' UTR length that were not identified in a general analysis of trascripit structure from the same data in [8]. Moreover, the significant overlap between our apaQTLs and the eQTLs identified in [8] confirms the known relevant role of 3'UTRs in gene expression regulation. However, the regulation of 3' UTR length is known to affect regulatory processes that do not directly alter mRNA abundance, such as regulation of translation efficiency, mRNA localization and membrane protein localization [12, 36]. Indeed most of the apaQTLs we found were not identified as eQTLs in [8].

The various mechanisms underlying the association between genetic variants and the relative

abundance of 3'UTR isoforms can be classified in two main classes based on whether they affect the production or degradation rates of the isoforms. The production related-mechanisms include the alteration of APA sites, of cis-regulatory elements located in promoters and enhancers, and of binding sites of RBPs involved in nuclear RNA processing; the degradation-related mechanisms include the alteration of the binding sites of microRNAs and cytoplasmatic RBPs affecting mRNA stability. Taken together, our results suggest that the genetic effects on 3'UTR isoforms act prevalently at the level of production, as shown by the strong enrichment of apaQTLs in non-transcribed regulatory regions and among the variants creating or disrupting APA sites, and by the relatively weak enrichment of variants creating or disrupting microRNA binding sites. Also the results on altered RBP binding sites confirm this picture, since most motifs altered by apaQTLs are associated to nuclear RBPs involved in nuclear RNA processing.

In particular, we identified several apaQTLs creating or destroying putative functional PAS motifs. However, it should be noted that our ability to detect these events is intrinsically limited by the motif repertoire that we used [40], which might miss some of the rarest alternative PAS motifs. For example, we found that the rs6151429 variant is associated with the increased expression of the long isoform of the transcript codified by the Arylsulfatase A (ARSA) gene (Supplementary Fig. 6D), in agreement with previous evidence [67]. However, we did not include this variant among those disrupting a PAS motif since the disrupted motif (AAUAAC) is not included in the catalog that we used. In addition, we considered only PAS-altering single nucleotide substitutions, while also other types of genetic variants can modify the PAS landscape of a gene. For example, a small deletion (rs374039502) causes the appearance of a new PAS motif within the TNFSF13B gene, and has been associated with an higher risk of both multiple sclerosis and SLE in the Sardinian population [68].

We observed a strong enrichment of apaQTLs in regulatory regions such as promoters and enhancers, as previously found for variants generically affecting transcript structure in [8]. These results point to an important role of DNA-binding cis-acting factors in the regulation of 3' UTR length, and to the existence of a widespread coupling between transcription and polyadenylation [12, 69]. Moreover, it has been shown that RBPs involved in APA regulation can interact with promoters [14].

Regarding the effect of genetic variants on mRNA stability, we focused on the perturbation of microRNA binding, taking into account both the creation and the destruction of microRNA binding sites within transcripts. The relevance of mRNA stability seemed to be confirmed by a modest enrichment of microRNA-altering SNPs among intragenic apaQTL, however the direction of their effect on microRNA binding is not statistically consistent with the expected direction of the change in 3' UTR isoform ratio. The same type of ambiguity has been previously reported with regard to the relationship between the effect of SNPs on microRNA binding and gene expression levels [70] and makes us doubt whether these microRNA-altering apaQTLs are truly

causal for the associated gene. These results suggest that the alteration of microRNA binding may not be a predominant mechanism explaining the variation of the expression of alternative 3'UTR isoforms across individuals. Limitations in the accuracy of predicted micorRNA binding sites might also contribute to this result.

Another possible mechanism of action of intragenic apaQTLs is the perturbation of the regulatory action of RBPs, as indicated by the modest but highly significant enrichment of SNPs altering RNA-binding motifs. However, the lack of strong enrichments when considering each motif individually suggests that specific RBP motifs may have a small regulatory impact on APA that may also depend on the context, as recently suggested [30]. As in the case of microRNAs, also our limited knowledge of the binding preferences of RBPs might limit our power to detect their effects: More sophisticated models should take into account the highly modular structure of RBPs that often include multiple RNA binding domains (RBDs), the emerging importance of both the binding context and the RNA structure and even more sophisticated modes of RNA binding [71, 72].

Furthermore, it is reasonable to assume that also non-canonical modes of APA regulation can be affected by genetic variants and therefore drive the detection of variable isoform expression ratios. For example, it has been recently suggested that an epitranscriptomic event, the m^6A mRNA methylation, can be associated with alternative polyadenylation [15]. In addition, recently published results suggest that genetic variants could affect APA regulation also in an indirect way, without affecting the regulatory machinery. Past studies have reported that a narrow range of 10-30nt between the PAS and the poly(A) site is required for efficient processing, however [73] suggested that also greater distances can sometimes be used thanks to RNA folding events that bring the PAS and the poly(A) site closer to each other. Therefore, we can speculate that if a genetic variant affects RNA folding in such a way as to modify the distance between the PAS and the poly(A) site, it could also influence APA regulation.

While the mechanisms discussed above act at the level of the primary or mature transcript, our results revealed a perhaps unexpectedly large number of extragenic apaQTLs, mostly located in regulatory regions. These apaQTLs point to an important role of DNA-binding elements such as transcription factors in regulating alternative polyadenylation through long-distance interactions with cleavage and polyadenylation factors. The investigation of these mechanisms is thus a promising avenue of future research.

Alternative polyadenylation can affect several biological processes, influencing mRNA stability, translation efficiency and mRNA localization [13]. Therefore, it is not surprising that its perturbation has been associated with multiple pathological conditions [7, 22]. In the present study, we detected a strong enrichment of GWAS hits among apaQTLs, supporting the idea that 3' UTR length is a useful addition to the list of intermediate molecular phenotypes that can be used for a mechanistic interpretation of GWAS hits. In particular, we identified genetic variants previously

associated to neurological disorders, such as autism, schizophrenia and multiple sclerosis, which may act by affecting the regulation of polyadenylation. The importance of post-transcriptional events in the onset of neurological diseases has been recently confirmed by two studies demonstrating that genetic variants affecting alternative splicing (sQTL) give a substantial contribution to the pathogenesis of schizophrenia [74] and Alzheimer's disease [75]. We also observed that the relevant apaQTLs often map to HLA genes, but that the enrichment is not explained by the HLA locus alone. On the other hand, examples of APA events involving HLA genes have been reported [76, 77] and genes encoding antigen-presenting molecules account for the highest fraction of genetic risk for many neurological diseases [78].

A gene-based alternative approach to the interpretation of GWAS has been recently proposed. In the original implementation of Transcriptome Wide Association Studies (TWAS) [79], eQTL data obtained in a reference dataset are used to predict the genetic component of gene expression in GWAS cases and controls, which is then correlated with the trait of interest, thus allowing the identification of susceptibility genes. More recently, the authors of [80] proposed a summary-based TWAS strategy in which the association between the genetic component of gene expression and a trait is indirectly estimated through the integration of SNP-expression, SNP-trait, and SNP-SNP correlation data. Furthermore, this kind of analysis has also been performed exploiting a collection of sQTLs, leading to the identification of new susceptibility genes for schizophrenia [81] and Alzheimer's disease [75]. In a similar way, apaQTLs could be used to discover cases in which the association between genes and diseases is driven by the alteration of the expression of alternative 3'UTR isoforms.

We are aware of some limitations of this study. First, the simple model that we used for the definition of alternative 3'UTRs isoforms limits the type of events that can be detected, because we can see only events involving poly(A) sites located within the transcript segments taken into account for the computation of the m/M values (the PRE and the POST segments). Nonetheless, the adoption of this simple model significantly reduces the computational burden and might be sufficient to indicate general trends that can be subsequently further investigated with more sophisticated models. Indeed, it has been previously shown, in a slightly different context (i.e. the comparison of APA events detected in different cellular conditions or tissues), that the results obtained with our model are comparable with those obtained exploiting a more complex model that takes into account all the possible APA isoforms of a gene, especially because also genes with multiple poly(A) sites mainly use only two or a few of them [29]. Second, our strategy depends on a pre-existing annotation of poly(A) sites: Methods that infer the location of poly(A) sites from RNA-Seq data are available, but they can have lower sensitivity in the detection of APA events [29, 30]. In addition, we examined only a single cell type (lymphoblastoid cells) to demonstrate the feasibility of apaQTL mapping analysis. A broader investigation, exploiting data such as those provided by Genotype-Tissue Expression (GTEx) consortium [82], would be particularly valuable. Indeed, APA regulation seems to be significantly tissue-specific and global

trends of poly(A) sites selection in specific human tissues have been described: For example transcripts in the nervous system and brain are characterized by preferential usage of distal PAS, whereas in the placenta, ovaries and blood the usage of proximal PAS is preferred [12].

In conclusion, we have identified thousands of common genetic variants associated with alternative polyadenylation in a population of healthy human subjects. Alternative polyadenylation is a promising intermediate molecular phenotype for the mechanisitic interpretation of genetic variants associated to phenotypic traits and diseases.

4 Material and Methods

4.1 Data sources

4.1.1 Human genome and transcriptome

The coordinates of the NCBI Reference Sequences (RefSeqs) in the human genome (hg19) were downloaded from the UCSC Genome Browser (09/04/2015) [83, 84]. The corresponding transcript-gene map was downloaded from NCBI (version 69) and the Bioconductor R package `org.Hs.eg.db v3.4.0` [85] was used to associate each Entrez Gene Id to its gene symbol. In addition, the reference sequence of the hg19 version of the human genome was downloaded from the ENSEMBL database and a collection of poly(A) sites was obtained from PolyA_DB2 (10/02/2014) [31].

ChromHMM annotations [37] were downloaded from the UCSC Genome Browser for the GM12878 and the NHEK cell lines (http://genome-euro.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm). In addition, the coordinates of Cis Regulatory Domains (CRDs) and their association with genes were downloaded for lymphoblastoid cells from ftp://jungle.unige.ch/SGX/ [61].

4.1.2 WGS and RNA-Seq data

We exploited the RNA-Seq data obtained by the GEUVADIS consortium in lymphoblastoid cell lines of 462 individuals belonging to different populations, but we considered only 373 individuals with European ancestry (EUR). BAM files were downloaded from the E-GEUV-1 dataset [8] in the EBI ArrayExpress archive (https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/processed). We also downloaded genotypic data for the same individuals (https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/genotypes/) and the results of the eQTL/trQTL mapping analyses (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/files/analysis_

results/). The downloaded VCF files include genotypes for 465 individuals: Among the 462 of them for which also RNA-Seq data are available, the large majority had been previously subjected to Whole Genome Sequencing (WGS) by the 1000 Genome Project (Phase 1) [3], but the GEUVADIS consortium additionally obtained genomic data for 41 of them through genotyping with Single Nucleotide Polymorphism (SNP) array followed by genotype imputation [8]. Furthermore, whole blood RNA-Seq data for 99 individuals affected by SLE were downloaded from the NCBI SRA database (SRP062966) [86, 66].

4.1.3 Regulatory motifs and related expression data

Different collections of regulatory motifs were downloaded. A list of 18 PAS motifs was obtained from [40], microRNA seeds were downloaded from TargetScan 7.2 [87] and Positional Weight Matrices (PWMs) describing the binding specificities of RNA-binding proteins were downloaded from the CISBP-RNA dataset [52], including both the experimentally determined motifs and those that were inferred from related proteins. In addition, the list of microRNAs and RBPs expressed in lymphoblastoid cells were obtained from the expression data available in the E-GEUV-2 dataset [8] on the EBI ArrayExpress archive (https://www.ebi.ac.uk/arrayexpress/files/E-GEUV2/GD452.MirnaQuantCount.1.2N.50FN.samplename.resk10.txt and https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/GD462.GeneQuantRPKM.50FN.samplename.resk10.txt.gz).

4.1.4 GWAS Catalog

A collection of genomic loci associated with human complex traits was obtained by downloading the NHGRI-EBI GWAS Catalog, v1.0.2 [63]. This resource is continuously updated: the version we used was downloaded on October 10th, 2018 and it was mapped to GRCh38.p12 and dbSNP Build 151. From the same website, we also downloaded a file showing the mapping of all the reported traits to the Experimental Factor Ontology (EFO) terms [88], including the parent category of each trait (the version of the downloaded file was r2018-09-30). In addition, the dpSNP Build 151 [89] collection of human genetic variants was downloaded for hg19.

4.2 Annotation of alternative 3'UTR isoforms

We considered the human transcripts included in RefSeq and associated them with the corresponding Entrez Gene Id. Moreover, we collapsed together the structures of all the transcripts assigned to a gene, using the union of all the exons of the various transcripts associated to a gene and defining the 3' or 5' UTR using respectively the most distal coding end and the most

proximal coding start. The annotation of the resulting gene structures can be found in the supplementary data (see Supplementary File 1).

The coordinates of the human poly(A) sites were converted from hg17 to hg19 using `liftOver` [90] and then combined with the gene structures defined above to define the alternative 3'UTR isoforms. For the definition of alternative 3'UTR isoforms we adopted a simple model taking into account only two alternative poly(A) sites for each gene, because previous evidence suggests that also genes with multiple poly(A) sites mainly use only two of them [29]. In particular, for each gene we selected the most proximal poly(A) site among those falling within exons, preferring those located within the 3'UTR, and the end of the gene as the distal poly(A) site. In this way we were able to define two segments of interest for each gene: the PRE segment, extending from the beginning of the last exon to the proximal poly(A) site, and the POST segment, from the proximal poly(A) site to the end of the gene. The PRE fragment is assumed to be expressed by both the long and the short isoform, while the POST segment should be expressed exclusively by the long isoform. The GTF file used for the computation of m/M values is available as Supplementary File 2.

The relative prevalence of the short and long isoforms are evaluated, as described below, based on the number of RNA-Seq reads falling into the PRE and POST regions. While the whole region from the transcription start site to the proximal poly(A) site could be taken, in principle, as the PRE region, we chose to limit it to the last exon to minimize the confounding effect of alternative splicing.

4.3 Computation of m/M values

Using the Bioconductor R package `Roar` [29], for each gene with alternative 3'UTR isoforms we obtained an m/M value in each individual. The m/M value estimates the ratio between the expression of the short and the long isoform of a gene in a particular condition and the $m/M_{\mathrm{a,i}}$ of gene a in the i_{th} individual is defined as

$$m/M_{a,i} = \frac{l_{POST_a} \times \#r_{PRE_{a,i}}}{l_{PRE_a} \times \#r_{POST_{a,i}}} - 1 \tag{1}$$

where l_{PRE_a} and l_{POST_a} are respectively the length of the PRE and POST segment of the gene a, $\#r_{PRE_{a,i}}$ and $\#r_{POST_{a,i}}$ are respectively the number of reads mapped on the PRE and the POST segment of the gene a in the i_{th} individual.

The m/M values were computed for 14,542 genes for which we were able to define alternative 3'UTR isoforms. Infinite and negative values of m/M (that happen when the POST region does not produce any reads, and when the POST region produces more reads than the PRE region

after length normalization, respectively) were considered as missing values. Then only those on autosomal chromosomes (chr1-22) and with less than 100 missing m/M values were selected for the following investigation, leaving us with 6,256 genes.

4.4 Genotypic data pre-processing

Starting from the downloaded VCF files, we extracted genotypic data for 373 EUR individuals for whom also RNA-Seq data are available using `VCFtools` [91]. In addition, only common genetic variants with Minor Allele Frequency (MAF) higher than 5% were considered in all the following analyses; the MAF value of each genetic variant was reported in the VCF files, but we recomputed them taking into account that the reference allele in the VCF file may not always be the most frequent one in the EUR population considered by itself, and conservatively attributing the most frequent homozygous genotype to individuals for which the genotype was missing.

4.5 Principal Component Analysis of genotypic data

It is known that special patterns of linkage disequilibrium (LD) can cause artifacts when a Principal Component Analysis (PCA) is used to investigate population structure [92]. We filtered out all the genetic variants falling within 24 long-range LD (LRLD) regions whose coordinates were derived from [92]. In addition, following [93], we performed an LD-pruning of the genetic variants using the `--indep-pairwise` function from `PLINK v1.9` [94] to recursively exclude genetic variants with pairwise genotypic R^2 >80% within sliding windows of 50 SNPs (with a 5-SNPs increment between windows). Also in this case `VCFtools` [91] was used to apply all these filters to the VCF files and finally `EIGENSTRAT v6.1.4` [95] was used to run the PCA on the remaining genotypic data at the genome-wide level.

4.6 apaQTL mapping

From a statistical point of view, we adopted the same strategy used in standard eQTL mapping analyses [8] to identify genetic variants that influence the expression level of the alternative 3'UTR isoforms of a gene. For each of the 6,256 examined genes, we defined a cis-window as the region spanning the gene body and 1 Mbp from both its TSS and its TSE. Then, for each gene a linear model was fitted, independently for each genetic variant within its cis-window, using the genotype for the genetic variant as the independent variable and the log2-transformed m/M

value of the gene as the dependent variable:

$$\log_2(m/M_{a,i}) = \beta_0 + \beta_1 \times g_{j,i} + \beta_2 \times I_i + \sum_{n=1}^{3} \alpha_n \times gPC_{n,i} + \epsilon_a \quad (2)$$

where $log_2(m/M_{a,i})$ is the log2-transformed m/M value computed for the a gene in the i_{th} individual, $g_{v,i}$ is the genotype of the i_{th} individual for the j_{th} genetic variant, I_i is the imputation status (0|1) of the i_{th} individual, $gPC_{n,i}$ is the value of the n_{th} Principal Component (PC) obtained from genotypic data for the i_{th} individual, β_0 is the intercept, β_1, β_2 and α_n are the fitted regression coefficients and ϵ_a is the error term for the gene a.

The fitting of the linear models was done using the CRAN R package `MatrixEQTL` [96]. Genotypes were represented using the standard 0/1/2 codification, referring to the number of alternative alleles present in each individual, and matrices with genotypic information were obtained from VCF files exploiting the Perl API (Vcf.pm) included in the VCFtools suite [91]. Following [8], in all our models we included both the imputation status of the individuals and the first three PCs obtained from genotypic data as covariates, in order to correct for possible biases due to population stratification (Supplementary Fig. 1) or genotype imputation.

The observed distribution of nominal P-values was compared with the expected one in Quantile-Quantile plots (Q-Q plots), revealing the expected inflation due to the LD issue (Supplementary Fig. 2). A permutation-based procedure was implemented [97]: all the models were fitted again after the random shuffling of the m/M values of each gene across samples; then for each gene-variant pair we counted how many times we obtained a random P-value less than its nominal P-value and divided this value by the total number of random tests done. Finally, to control for multiple testing, the empirical P-values were corrected with the Benjamini-Hochberg procedure [98] and models with a corrected empirical P-value less than 0.05 were considered statistically significant. Manhattan plots were drawn using the CRAN R package `qqman` [99].

4.7 Comparison with other molecular QTLs

In order to compare the genes for which we detected one or more apaQTLs with those for which eQTL/trQTL were reported [8], we translated the Ensembl Gene IDs (ENSG) to NCBI Entrez Gene IDs using Ensembl v67 [100] retrieved using the Bioconductor R package `biomaRt v2.30` [101, 102]. 229 ENSGs could not be translated with this procedure and were therefore excluded from this analysis.

4.8 Enrichment analyses

In order to functionally characterize the apaQTLs, we analyzed the enrichment of several features among such variants, including their genomic location, their ability to alter known regulatory motifs, and their association with complex diseases. All enrichments were evaluated through multivariate logistic regression to allow correcting for covariates. In this section we provide an overview of the method, but refer to the following subsections for details about each analysis.

For each feature we first established which genetic variants were potentially associated with the feature (for example only variants in the 3' UTR can alter microRNA binding sites). Therefore, each enrichment analysis started with the selection of the "candidate variants" that were subsequently subjected to an LD-based pruning, in order to obtain a subset of independent candidate variants (the same strategy was implemented for example in [103] to evaluate the enrichment of GWAS hits among eQTLs). LD-based pruning was always performed using `PLINK` with the same parameters used in the case of the PCA of genotypic data (see above), but applied in each case to the candidate variants only. To each candidate variant surviving pruning we attributed a binary variable indicating whether it has the feature under investigation. Finally, these variants are classified as apaQTLs (i.e. corrected empirical P-value < 0.05 for at least one gene) and null variants (i.e. nominal P-value > 0.1 in all the fitted models). We excluded the "grey area" variants with nominal P-value < 0.1 but empirical corrected P-value > 0.05 as they are likely to contain many false negatives. Finally we fitted a multivariate logistic model in which the dependent variable is the apaQTL/null status of the variant, and the independent variables are the feature of interest and covariates. The latter always include the MAF of the variant, since variants with higher MAF are more likely to be found as significant apaQTLs, and possibly other covariates depending on the feature under examination (see below).

The logistic model can thus be written as:

$$t_j = \beta_0 + \beta_1 \times Feature_j + covariates + \epsilon_j \tag{3}$$

$$Pr(apaQTL)_j = \frac{1}{1 + \exp^{-t_j}} \tag{4}$$

where $Feature_j$ is a binary variable indicating whether the genetic variant j has the feature of interest, β_0 is the intercept, β_1 is the regression coefficient for the feature, ϵ_j is the error term and $Pr(apaQTL)_j$ is the fitted probability that the genetic variant j is an apaQTL. As expected, in our models the regression coefficient of the MAF was always positive. The regression coefficient of the $Feature$ term and its associated P-value were used to establish if having the feature under investigation influences the probability of being an apaQTL, and to compute the corresponding

odds ratio (OR).

4.8.1 Chromatin states

This analysis was performed independently for two cell types (the GM12878 and NHEK cell lines). In both cases, the candidate variants were virtually all the genetic variants for which the apaQTL models were fitted, but we excluded those not associated with any chromatin state and all the structural variants, because their length can prevent them from being univocally associated with a chromatin state.

Each of the 15 chromatin states and 6 broad chromatin classes (promoter, enhancer, insulator, transcribed, repressed and inactive) defined in [37], separately for the two cell lines, was treated as a binary feature to be used as a regressor in Eq. (3), with value 1 assigned to the variants falling within a DNA region associated to the given chromatin state. Only the MAF was included in the covariates.

4.8.2 Gene regions

The candidate variants were all the intragenic variants for which the apaQTL models were fitted. We defined as intragenic all variants falling between the start and the end of the gene, plus 1,000 bps after the end (to take into account possible misannotations of the 3' UTR).

Independent enrichment analyses were performed for the following sequence classes: coding exons, introns, 5'UTR and 3'UTR. For each class the binary feature used as a regressor was assigned the value 1 for variants falling within the class and 0 otherwise. Only the MAF was included in the covariates.

4.8.3 Cis Regulatory Domains

The candidate variants were all the extragenic variants (i.e. all variants that are not intragenic according to the definition given above) for which an apaQTL model was fitted. The binary feature was given value 1 for variants falling within a CRD and 0 otherwise. Besides the MAF, the distance from the nearest gene was included as a covariate, since variants closer to a gene are more likely to be apaQTLs.

To verify that the apaQTLs tend to be included in the CRDs specifically associated to the gene on which they act, we translated the CRD-gene associations provided in [61] into Entrez Gene IDs, and we counted how many genetic variants fall within a CRD associated to at least one

gene for which the variant is an apaQTL. This number was then compared with that obtained in the same way after randomly assigning a target gene to each extragenic variant within the cis-window used for apaQTL analysis (100 independent randomizations were used).

4.8.4 Alteration of putative functional motifs

Similar strategies were implemented to investigate the alteration of different types of putative functional motifs by intragenic variants. This analysis was restricted to Single Nucleotide Polymorphism (SNPs), excluding therefore both indels and structural variants. For all SNPs we reconstructed the sequence of both the reference (REF) and the alternative (ALT) allele in the 20 bp region around each candidate genetic variant to determine whether the ALT allele creates or destroys a functional motif with respect to the REF allele. The functional motifs analyzed included PAS motifs, microRNA binding sites, and RBP binding sites.

To each candidate variant surviving LD pruning we associated, using `PLINK`, a list of tagging variants with genotypic $R^2 > 80\%$, and a binary feature value of 1 if the candidate variant itself or any of its tagging variant altered a functional motif. The enrichment of apaQTLs among motif-affecting variants was then evaluated with the logistic model described by Eq. 3. In the following, we describe the details of the logistic model for each class of functional motifs.

PAS motifs. The PAS motif is always located upstream of its target poly(A) site. It has been suggested that a narrow range of 10-30 nt is required for efficient processing, but recent work suggests that also larger distances can be functional thanks to RNA folding processes bringing the poly(A) site closer to the PAS [73]. Assuming that a PAS-altering SNP would affect the usage of its nearest poly(A) site, we associated to each intragenic SNP the nearest downstream poly(A) site, selected those for which such poly(A) site was located within the PRE/POST segments, and retained as candidate variants only those whose distance from the corresponding poly(A) site was between 10 and 100 nt. PAS-altering variants were defined as those for which a particular PAS motif was found in either the REF or the ALT sequence, but not in both (note that the interconversion between PAS motifs is considered as well, assuming that they can have different strength).

microRNA binding sites. microRNA binding sites located downstream of a poly(A) site, and hence in the POST segment, can affect the relative abundance of the long and short isoforms by allowing the selective degradation of the former by microRNAs. Therefore, we chose as candidate variants all the SNPs within the POST segment of the genes analyzed. Putative microRNA binding sites were classified, as in [87], in three classes: 8mer, 7mer-m8, and 7mer-A1 (matches classified as 6-mer were not considered). A variant was defined to alter a microRNA binding site if a putative binding site was present in either the REF or the ALT sequence, but not in both,

or if the site class was different between the REF and the ALT sequences. Moreover, altering variants were classified as creating (destroying) a binding site if only the ALT (REF) sequence contained a binding site or if the ALT (REF) sequence contained a stronger binding site than the REF (ALT), according to the hierarchy 8mer > 7mer-m8 > 7mer-A1 match. Only microRNA families conserved across mammals or broadly conserved across vertebrates and expressed in lymphoblastoid cells were considered. Following [8], each microRNA was considered expressed if its expression value was greater than 0 in at least 50% of the samples, and each microRNA family was considered expressed if at least one of its microRNAs was expressed.

RBP motifs. The candidate variants were all the intragenic SNPs. `FIMO` [104] was used to scan the REF and ALT sequences around each candidate variant, using as background the nucleotide frequencies on the sequence of all the analyzed genes. A motif was considered altered if its score was greater than 80% the score of the perfect match in only one of two alleles. As in the case of microRNAs, only motifs corresponding to RBPs expressed in lymphoblastoid cell lines were considered. Enrichment was evaluated both for SNPs altering any RBP motif, and for each expressed RBP separately.

4.8.5 GWAS hits

We considered only the GWAS catalog records referring to a single genetic variant on autosomal chromosomes for which all the fields CHR_ID, CHR_POS, SNPS, MERGED, SNP_ID_CURRENT, and MAPPED_TRAIT_URI were available, as well as the RSID. The coordinates of the selected genetic variants in hg19 were derived from dbSNP Build 151. We thus obtained 56,672 genetic variants associated with at least one complex trait. Furthermore, starting from the EFO URI(s) reported for each association, we obtained the corresponding EFO Parent URI(s) from the EFO annotation file.

All variants examined as potential apaQTLs were considered as our candidate variants. A binary feature value of 1 was attributed to each candidate variant surviving LD pruning and associated to a trait, or with a tagging variant associated to a trait, as in the case of motif-altering variants. Enrichment was evaluated for all trait-associated variants together, for each single trait, and for trait categories defined based on the EFO ontology. Only traits and trait categories associated with at least 100 GWAS hits were analyzed. The same analysis was also performed after excluding all variants within the HLA locus, as defined by The Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh370).

4.9 The rs10954213 variant in SLE patients

Since we were interested in the IRF5 gene only, RNA-Seq reads were aligned to a reduced genome comprising the gene sequence and an additional 50bp at its 3' end using Bowtie v2.2.3 [105] and TopHat v2.0.12 [106]. As genotypic data were not available for these individuals, we inferred the rs10954213 variant status from the relative proportion of A and G in the RNA-Seq reads. Individuals were considered homozygous for the reference (G) or for the alternative (A) allele when the same nucleotide was present in all the reads, and a single read with a different nucleotide was considered sufficient to call an heterozygous individual (Supplementary Fig. 9). In this way we obtained 10 homozygotes for the reference allele, 73 heterozygotes and 16 homozygotes for the alternative allele. The MAF thus obtained is consistent with that reported in the NCBI dpSNP database [89]. A Kruskal-Wallis test was then used to evaluate the differences in m/M values between genotypes. A different criterion for the assignment of genotypes (at least 20% of the reads carrying the less frequent allele required to call a heterozygous individual) gave comparable results (Supplementary Fig. 10).

References

[1] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062. URL: https://www.ncbi.nlm.nih.gov/pubmed/11237011.

[2] J. Craig Venter et al. "The Sequence of the Human Genome". In: *Science* 291.5507 (Feb. 2001), pp. 1304–1351. ISSN: 0036-8075. DOI: 10.1126/science.1058040. URL: http://www.ncbi.nlm.nih.gov/pubmed/11181995.

[3] The 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393. URL: https://www.ncbi.nlm.nih.gov/pubmed/26432245.

[4] Peter M Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation." In: *American journal of human genetics* 101.1 (July 2017), pp. 5–22. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.06.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/28686856.

[5] William Cookson et al. "Mapping complex disease traits with global gene expression". In: *Nature Reviews Genetics* 10.3 (Mar. 2009), pp. 184–194. ISSN: 1471-0056. DOI: 10.1038/nrg2537. URL: https://www.ncbi.nlm.nih.gov/pubmed/19223927.

[6] Frank W. Albert and Leonid Kruglyak. "The role of regulatory variation in complex traits and disease". In: *Nature Reviews Genetics* 16.4 (Apr. 2015), pp. 197–212. ISSN: 1471-0056. DOI: 10.1038/nrg3891. URL: https://www.ncbi.nlm.nih.gov/pubmed/25707927.

[7] Kassie S. Manning and Thomas A. Cooper. "The roles of RNA processing in translating genotype to phenotype". In: *Nature Reviews Molecular Cell Biology* 18.2 (Feb. 2017), pp. 102–114. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.139. URL: https://www.ncbi.nlm.nih.gov/pubmed/27847391.

[8] Tuuli Lappalainen et al. "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468 (Sept. 2013), pp. 506–511. ISSN: 0028-0836. DOI: 10.1038/nature12531. URL: https://www.ncbi.nlm.nih.gov/pubmed/24037378.

[9] Jean Monlong et al. "Identification of genetic variants associated with alternative splicing using sQTLseekeR". In: *Nature Communications* 5.1 (Dec. 2014), p. 4698. ISSN: 2041-1723. DOI: 10.1038/ncomms5698. URL: https://www.ncbi.nlm.nih.gov/pubmed/25140736.

[10] Hui Y Xiong et al. "RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease." In: *Science (New York, N.Y.)* 347.6218 (Jan. 2015), p. 1254806. ISSN: 1095-9203. DOI: 10.1126/science.1254806. URL: http://www.ncbi.nlm.nih.gov/pubmed/25525159.

[11] K. G. Ardlie et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235 (May 2015), pp. 648–660. ISSN: 0036-8075. DOI: 10.1126/science.1262110. URL: http://www.ncbi.nlm.nih.gov/pubmed/25954001.

[12] Ran Elkon, Alejandro P. Ugalde, and Reuven Agami. "Alternative cleavage and polyadenylation: extent, regulation and function". In: *Nature Reviews Genetics* 14.7 (July 2013), pp. 496–506. ISSN: 1471-0056. DOI: 10.1038/nrg3482. URL: https://www.ncbi.nlm.nih.gov/pubmed/23774734.

[13] Bin Tian and James L. Manley. "Alternative polyadenylation of mRNA precursors". In: *Nature Reviews Molecular Cell Biology* 18.1 (Jan. 2017), pp. 18–30. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.116. URL: https://www.ncbi.nlm.nih.gov/pubmed/27677860.

[14] Katarzyna Oktaba et al. "ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system." In: *Molecular cell* 57.2 (Jan. 2015), pp. 341–8. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2014.11.024. URL: http://www.ncbi.nlm.nih.gov/pubmed/25544561.

[15] Yanan Yue et al. "VIRMA mediates preferential m6A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation". In: *Cell Discovery* 4.1 (Dec. 2018), p. 10. ISSN: 2056-5968. DOI: 10.1038/s41421-018-0019-0. URL: https://www.ncbi.nlm.nih.gov/pubmed/29507755.

[16] Christine Mayr. "What Are 3' UTRs Doing?" In: *Cold Spring Harbor Perspectives in Biology* (Sept. 2018), a034728. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a034728. URL: http://www.ncbi.nlm.nih.gov/pubmed/30181377.

[17] Chioniso P. Masamha et al. "CFIm25 links alternative polyadenylation to glioblastoma tumour suppression". In: *Nature* 510.7505 (June 2014), pp. 412–416. ISSN: 0028-0836. DOI: 10.1038/nature13261. URL: https://www.ncbi.nlm.nih.gov/pubmed/24814343.

[18] Christine Mayr and David P. Bartel. "Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells". In: *Cell* 138.4 (Aug. 2009), pp. 673–684. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2009.06.016. URL: https://www.ncbi.nlm.nih.gov/pubmed/19703394.

[19] Rickard Sandberg et al. "Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites." In: *Science (New York, N.Y.)* 320.5883 (June 2008), pp. 1643–7. ISSN: 1095-9203. DOI: 10.1126/science.1155390. URL: http://www.ncbi.nlm.nih.gov/pubmed/18566288.

[20] Zhe Ji et al. "Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.17 (Apr. 2009), pp. 7028–33. ISSN: 1091-6490. DOI: 10.1073/pnas.0900028106. URL: http://www.ncbi.nlm.nih.gov/pubmed/19372383.

[21] Yonggui Fu et al. "Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing." In: *Genome research* 21.5 (May 2011), pp. 741–7. ISSN: 1549-5469. DOI: 10.1101/gr.115295.110. URL: http://www.ncbi.nlm.nih.gov/pubmed/21474764.

[22] Jae Woong Chang, Hsin Sung Yeh, and Jeongsik Yong. "Alternative Polyadenylation in Human Diseases." In: *Endocrinology and metabolism (Seoul, Korea)* 32.4 (Dec. 2017), pp. 413–421. ISSN: 2093-596X. DOI: 10.3803/EnM.2017.32.4.413. URL: http://www.ncbi.nlm.nih.gov/pubmed/29271615.

[23] Pedro G. Ferreira et al. "Sequence variation between 462 human individuals fine-tunes functional sites of RNA processing". In: *Scientific Reports* 6.1 (Oct. 2016), p. 32406. ISSN: 2045-2322. DOI: 10.1038/srep32406. URL: https://www.ncbi.nlm.nih.gov/pubmed/27617755.

[24] Oh Kyu Yoon et al. "Genetics and Regulatory Impact of Alternative Polyadenylation in Human B-Lymphoblastoid Cells". In: *PLoS Genetics* 8.8 (Aug. 2012). Ed. by Gene Yeo, e1002882. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002882. URL: https://www.ncbi.nlm.nih.gov/pubmed/22916029.

[25] Laurent F. Thomas and Pål Sætrom. "Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation". In: *PLoS Computational Biology* 8.8 (Aug. 2012). Ed. by Roderic Guigo, e1002621. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002621. URL: https://www.ncbi.nlm.nih.gov/pubmed/22915998.

[26] Tony Kwan et al. "Genome-wide analysis of transcript isoform variation in humans". In: *Nature Genetics* 40.2 (Feb. 2008), pp. 225–231. ISSN: 1061-4036. DOI: 10.1038/ng.2007.57. URL: https://www.ncbi.nlm.nih.gov/pubmed/18193047.

[27] Daria V. Zhernakova et al. "DeepSAGE Reveals Genetic Variants Associated with Alternative Polyadenylation and Expression of Coding and Non-coding Transcripts". In: *PLoS Genetics* 9.6 (June 2013). Ed. by Vivian G. Cheung, e1003594. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003594. URL: https://www.ncbi.nlm.nih.gov/pubmed/23818875.

[28] Antonio Lembo, Ferdinando Di Cunto, and Paolo Provero. "Shortening of 3'UTRs Correlates with Poor Prognosis in Breast and Lung Cancer". In: *PLoS ONE* 7.2 (Feb. 2012). Ed. by Jun Li, e31129. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031129. URL: https://www.ncbi.nlm.nih.gov/pubmed/22347440.

[29] Elena Grassi et al. "Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries". In: *BMC Bioinformatics* 17.1 (Dec. 2016), p. 423. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1254-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/27756200.

[30] Kevin C. H. Ha, Benjamin J. Blencowe, and Quaid Morris. "QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data". In: *Genome Biology* 19.1 (Dec. 2018), p. 45. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1414-4. URL: https://www.ncbi.nlm.nih.gov/pubmed/29592814.

[31] Ju Youn Lee et al. "PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes." In: *Nucleic acids research* 35.Database issue (Jan. 2007), pp. D165–8. ISSN: 1362-4962. DOI: 10.1093/nar/gkl870. URL: http://www.ncbi.nlm.nih.gov/pubmed/17202160.

[32] Leiming You et al. "APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals." In: *Nucleic acids research* 43.Database issue (Jan. 2015), pp. D59–67. ISSN: 1362-4962. DOI: 10.1093/nar/gku1076. URL: http://www.ncbi.nlm.nih.gov/pubmed/25378337.

[33] Noah Spies, Christopher B Burge, and David P Bartel. "3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts." In: *Genome research* 23.12 (Dec. 2013), pp. 2078–90. ISSN: 1549-5469. DOI: 10.1101/gr.156919.113. URL: http://www.ncbi.nlm.nih.gov/pubmed/24072873.

[34] Stephen N Floor and Jennifer A Doudna. "Tunable protein synthesis by transcript isoforms in human cells". In: *eLife* 5 (Jan. 2016), e10921. ISSN: 2050-084X. DOI: 10.7554/eLife.10921. URL: https://www.ncbi.nlm.nih.gov/pubmed/26735365.

[35] Juan Ji An et al. "Distinct Role of Long 3' UTR BDNF mRNA in Spine Morphology and Synaptic Plasticity in Hippocampal Neurons". In: *Cell* 134.1 (July 2008), pp. 175–187. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2008.05.045. URL: https://www.ncbi.nlm.nih.gov/pubmed/18614020.

[36] Binyamin D. Berkovits and Christine Mayr. "Alternative 3'UTRs act as scaffolds to regulate membrane protein localization". In: *Nature* 522.7556 (June 2015), pp. 363–367. ISSN: 0028-0836. DOI: 10.1038/nature14321. URL: https://www.ncbi.nlm.nih.gov/pubmed/25896326.

[37] Jason Ernst et al. "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345 (May 2011), pp. 43–49. ISSN: 00280836. DOI: 10.1038/nature09906. URL: https://www.ncbi.nlm.nih.gov/pubmed/21441907.

[38] Bin Tian, Zhenhua Pan, and Ju Youn Lee. "Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing." In: *Genome research* 17.2 (Feb. 2007), pp. 156–65. ISSN: 1088-9051. DOI: 10.1101/gr.5532707. URL: http://www.ncbi.nlm.nih.gov/pubmed/17210931.

[39] Alan G Hinnebusch, Ivaylo P Ivanov, and Nahum Sonenberg. "Translational control by 5'-untranslated regions of eukaryotic mRNAs." In: *Science (New York, N.Y.)* 352.6292 (June 2016), pp. 1413–6. ISSN: 1095-9203. DOI: 10.1126/science.aad9868. URL: http://www.ncbi.nlm.nih.gov/pubmed/27313038.

[40] Andreas J Gruber et al. "A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation." In: *Genome research* 26.8 (Aug. 2016), pp. 1145–59. ISSN: 1549-5469. DOI: 10.1101/gr.202432.115. URL: http://www.ncbi.nlm.nih.gov/pubmed/27382025.

[41] Robert R Graham et al. "Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.16 (Apr. 2007), pp. 6758–63. ISSN: 0027-8424. DOI: 10.1073/pnas.0701266104. URL: http://www.ncbi.nlm.nih.gov/pubmed/17412832.

[42] Deborah S Cunninghame Graham et al. "Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation." In: *Human molecular genetics* 16.6 (Mar. 2007), pp. 579–91. ISSN: 0964-6906. DOI: 10.1093/hmg/ddl469. URL: http://www.ncbi.nlm.nih.gov/pubmed/17189288.

[43] Elisa Lazzari and Caroline A. Jefferies. "IRF5-mediated signaling and implications for SLE". In: *Clinical Immunology* 153.2 (Aug. 2014), pp. 343–352. ISSN: 1521-6616. DOI: 10.1016/J.CLIM.2014.06.001. URL: https://www.ncbi.nlm.nih.gov/pubmed/24928322.

[44] Gary M Shaw et al. "118 SNPs of folate-related genes and risks of spina bifida and conotruncal heart defects". In: *BMC Medical Genetics* 10.1 (Dec. 2009), p. 49. ISSN: 1471-2350. DOI: 10.1186/1471-2350-10-49. URL: https://www.ncbi.nlm.nih.gov/pubmed/19493349.

[45] B A Eipper, C C Glembotski, and R E Mains. "Bovine intermediate pituitary alpha-amidation enzyme: preliminary characterization." In: *Peptides* 4.6 (1983), pp. 921–8. ISSN: 0196-9781. URL: http://www.ncbi.nlm.nih.gov/pubmed/6672794.

[46] Traci A Czyzyk et al. "Deletion of peptide amidation enzymatic activity leads to edema and embryonic lethality in the mouse." In: *Developmental biology* 287.2 (Nov. 2005), pp. 301–13. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2005.09.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/16225857.

[47] Eric D Gaier et al. "Genetic determinants of amidating enzyme activity and its relationship with metal cofactors in human serum." In: *BMC endocrine disorders* 14.1 (July 2014), p. 58. ISSN: 1472-6823. DOI: 10.1186/1472-6823-14-58. URL: http://www.ncbi.nlm.nih.gov/pubmed/25022877.

[48] Diana Chang et al. "A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci." In: *Nature genetics* 49.10 (Oct. 2017), pp. 1511–1516. ISSN: 1546-1718. DOI: 10.1038/ng.3955. URL: http://www.ncbi.nlm.nih.gov/pubmed/28892059.

[49] Lucia Sobrin et al. "Heritability and genome-wide association study to assess genetic differences between advanced age-related macular degeneration subtypes." In: *Ophthalmology* 119.9 (Sept. 2012), pp. 1874–85. ISSN: 1549-4713. DOI: 10.1016/j.ophtha.2012.03.014. URL: http://www.ncbi.nlm.nih.gov/pubmed/22705344.

[50] Manas R Biswal et al. "Conditional Induction of Oxidative Stress in RPE: A Mouse Model of Progressive Retinal Degeneration." In: *Advances in experimental medicine and biology* 854 (2016), pp. 31–7. ISSN: 0065-2598. DOI: 10.1007/978-3-319-17121-0_5. URL: http://www.ncbi.nlm.nih.gov/pubmed/26427390.

[51] James D McKay et al. "Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes." In: *Nature genetics* 49.7 (July 2017), pp. 1126–1132. ISSN: 1546-1718. DOI: 10.1038/ng.3892. URL: http://www.ncbi.nlm.nih.gov/pubmed/28604730.

[52] Debashish Ray et al. "A compendium of RNA-binding motifs for decoding gene regulation". In: *Nature* 499.7457 (July 2013), pp. 172–177. ISSN: 0028-0836. DOI: 10.1038/nature12311. URL: https://www.ncbi.nlm.nih.gov/pubmed/23846655.

[53] Yongsheng Shi and James L Manley. "The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site." In: *Genes & development* 29.9 (May 2015), pp. 889–97. ISSN: 1549-5477. DOI: 10.1101/gad.261974.115. URL: http://www.ncbi.nlm.nih.gov/pubmed/25934501.

[54] Piergiorgio La Rosa et al. "Sam68 promotes self-renewal and glycolytic metabolism in mouse neural progenitor cells by modulating Aldh1a3 pre-mRNA 3'-end processing". In: *eLife* 5 (Nov. 2016). ISSN: 2050-084X. DOI: 10.7554/eLife.20750. URL: http://www.ncbi.nlm.nih.gov/pubmed/27845622.

[55] Jason E. Miller and Joseph C. Reese. "Ccr4-Not complex: the control freak of eukaryotic cells". In: *Critical Reviews in Biochemistry and Molecular Biology* 47.4 (Aug. 2012), pp. 315–333. ISSN: 1040-9238. DOI: 10.3109/10409238.2012.667214. URL: http://www.ncbi.nlm.nih.gov/pubmed/22416820.

[56] Stefania Millevoi et al. "An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries". In: *The EMBO Journal* 25.20 (Oct. 2006), pp. 4854–4864. ISSN: 0261-4189. DOI: 10.1038/sj.emboj.7601331. URL: http://www.ncbi.nlm.nih.gov/pubmed/17024186.

[57] Stefania Millevoi et al. "A physical and functional link between splicing factors promotes pre-mRNA 3' end processing." In: *Nucleic acids research* 37.14 (Aug. 2009), pp. 4672–83. ISSN: 1362-4962. DOI: 10.1093/nar/gkp470. URL: http://www.ncbi.nlm.nih.gov/pubmed/19506027.

[58] S I Gunderson et al. "The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase." In: *Cell* 76.3 (Feb. 1994), pp. 531–41. ISSN: 0092-8674. URL: http://www.ncbi.nlm.nih.gov/pubmed/8313473.

[59] C S Lutz et al. "Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro." In: *Genes & development* 10.3 (Feb. 1996), pp. 325–37. ISSN: 0890-9369. URL: http://www.ncbi.nlm.nih.gov/pubmed/8595883.

[60] S. Liang and CS. Lutz. "p54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation". In: *RNA* 12.1 (Jan. 2006), pp. 111–121. ISSN: 1355-8382. DOI: 10.1261/rna.2213506. URL: http://www.ncbi.nlm.nih.gov/pubmed/16373496.

[61] Olivier Delaneau et al. "Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks". In: *bioRxiv* (Aug. 2017), p. 171694. DOI: 10.1101/171694. URL: https://www.biorxiv.org/content/early/2017/08/03/171694.

[62] Peter Hugo Lodewijk Krijger and Wouter de Laat. "Regulation of disease-associated gene expression in the 3D genome". In: *Nature Reviews Molecular Cell Biology* 17.12 (Dec. 2016), pp. 771–782. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.138. URL: https://www.ncbi.nlm.nih.gov/pubmed/27826147.

[63] Jacqueline MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1133. URL: http://www.ncbi.nlm.nih.gov/pubmed/27899670.

[64] Di Feng et al. "Genetic variants and disease-associated factors contribute to enhanced IRF-5 expression in blood cells of systemic lupus erythematosus patients". In: *Arthritis & Rheumatism* 62.2 (Feb. 2010), pp. 562–573. ISSN: 00043591. DOI: 10.1002/art.27223. URL: https://www.ncbi.nlm.nih.gov/pubmed/20112383.

[65] Sergey V. Kozyrev et al. "Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus". In: *Arthritis & Rheumatism* 56.4 (Apr. 2007), pp. 1234–1241. ISSN: 00043591. DOI: 10.1002/art.22497. URL: https://www.ncbi.nlm.nih.gov/pubmed/17393452.

[66] T Hung et al. "The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression." In: *Science (New York, N.Y.)* 350.6259 (Oct. 2015), pp. 455–9. ISSN: 1095-9203. DOI: 10.1126/science.aac7442. URL: http://www.ncbi.nlm.nih.gov/pubmed/26382853.

[67] V Gieselmann et al. "Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site." In: *Proceedings of the National Academy of Sciences of the United States of America* 86.23 (Dec. 1989), pp. 9436–40. ISSN: 0027-8424. URL: http://www.ncbi.nlm.nih.gov/pubmed/2574462.

[68] Maristella Steri et al. "Overexpression of the Cytokine BAFF and Autoimmunity Risk". In: *New England Journal of Medicine* 376.17 (Apr. 2017), pp. 1615–1626. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1610528. URL: https://www.ncbi.nlm.nih.gov/pubmed/28445677.

[69] Zhe Ji et al. "Transcriptional activity regulates alternative cleavage and polyadenylation." In: *Molecular systems biology* 7.1 (Sept. 2011), p. 534. ISSN: 1744-4292. DOI: 10.1038/msb.2011.69. URL: http://www.ncbi.nlm.nih.gov/pubmed/21952137.

[70] Urmo Võsa et al. "Altered Gene Expression Associated with microRNA Binding Site Polymorphisms". In: *PLOS ONE* 10.10 (Oct. 2015). Ed. by Santosh Patnaik, e0141351. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0141351. URL: https://www.ncbi.nlm.nih.gov/pubmed/26496489.

[71] Matthias W. Hentze et al. "A brave new world of RNA-binding proteins". In: *Nature Reviews Molecular Cell Biology* 19.5 (Jan. 2018), pp. 327–341. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.130. URL: https://www.ncbi.nlm.nih.gov/pubmed/29339797.

[72] Daniel Dominguez et al. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins". In: *Molecular Cell* 70 (2018), 854–867.e9. DOI: 10.1016/j.molcel.2018.05.001. URL: https://www.ncbi.nlm.nih.gov/pubmed/29883606.

[73] Xuebing Wu and David P. Bartel. "Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability". In: *Cell* 169.5 (May 2017), 905–917.e11. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2017.04.036. URL: https://www.ncbi.nlm.nih.gov/pubmed/28525757.

[74] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. "Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci". In: *Nature Communications* 8 (Feb. 2017), p. 14519. ISSN: 2041-1723. DOI: 10.1038/ncomms14519. URL: https://www.ncbi.nlm.nih.gov/pubmed/28240266.

[75] Towfique Raj et al. "Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility". In: *Nature Genetics* 50.11 (Nov. 2018), pp. 1584–1592. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0238-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/30297968.

[76] J-J. Hoarau et al. "HLA DQA1 genes generate multiple transcripts by alternative splicing and polyadenylation of the 3' untranslated region". In: *Tissue Antigens* 63.1 (Jan. 2004), pp. 58–71. ISSN: 0001-2815. DOI: 10.1111/j.1399-0039.2004.00140.x. URL: https://www.ncbi.nlm.nih.gov/pubmed/14651525.

[77] Smita Kulkarni et al. "Posttranscriptional Regulation of HLA-A Protein Expression by Alternative Polyadenylation Signals Involving the RNA-Binding Protein Syncrip." In: *Journal of immunology (Baltimore, Md. : 1950)* 199.11 (Dec. 2017), pp. 3892–3899. ISSN: 1550-6606. DOI: 10.4049/jimmunol.1700697. URL: http://www.ncbi.nlm.nih.gov/pubmed/29055006.

[78] Maneesh K. Misra, Vincent Damotte, and Jill A. Hollenbach. "The immunogenetics of neurological disease". In: *Immunology* 153.4 (Apr. 2018), pp. 399–414. ISSN: 00192805. DOI: 10.1111/imm.12869. URL: https://www.ncbi.nlm.nih.gov/pubmed/29159928.

[79] Eric R Gamazon et al. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1091–1098. ISSN: 1061-4036. DOI: 10.1038/ng.3367. URL: https://www.ncbi.nlm.nih.gov/pubmed/26258848.

[80] Alexander Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature Genetics* 48.3 (Mar. 2016), pp. 245–252. ISSN: 1061-4036. DOI: 10.1038/ng.3506. URL: https://www.ncbi.nlm.nih.gov/pubmed/26854917.

[81] Alexander Gusev et al. "Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights". In: *Nature Genetics* 50.4 (Apr. 2018), pp. 538–548. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0092-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/29632383.

[82] François Aguet et al. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (Oct. 2017), pp. 204–213. ISSN: 0028-0836. DOI: 10.1038/nature24277. URL: https://www.ncbi.nlm.nih.gov/pubmed/29022597.

[83] Jonathan Casper et al. "The UCSC Genome Browser database: 2018 update". In: *Nucleic Acids Research* 46.D1 (Jan. 2017), pp. D762–D769. DOI: 10.1093/nar/gkx1020. URL: https://www.ncbi.nlm.nih.gov/pubmed/29106570.

[84] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D733–D745. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1189. URL: https://www.ncbi.nlm.nih.gov/pubmed/26553804.

[85] Marc Carlson. *org.Hs.eg.db: Genome wide annotation for Human*. 2016.

[86] Rasko Leinonen et al. "The sequence read archive." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D19–21. ISSN: 1362-4962. DOI: 10.1093/nar/gkq1019. URL: http://www.ncbi.nlm.nih.gov/pubmed/21062823.

[87] Vikram Agarwal et al. "Predicting effective microRNA target sites in mammalian mRNAs". In: *eLife* 4.e05005 (Aug. 2015). ISSN: 2050-084X. DOI: 10.7554/eLife.05005. URL: https://www.ncbi.nlm.nih.gov/pubmed/26267216.

[88] James Malone et al. "Modeling sample variables with an Experimental Factor Ontology". In: *Bioinformatics* 26.8 (Apr. 2010), pp. 1112–1118. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btq099. URL: https://www.ncbi.nlm.nih.gov/pubmed/20200009.

[89] S T Sherry, M Ward, and K Sirotkin. "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation." In: *Genome research* 9.8 (Aug. 1999), pp. 677–9. ISSN: 1088-9051. DOI: 10.1101/GR.9.8.677. URL: http://www.ncbi.nlm.nih.gov/pubmed/10447503.

[90] A. S. Hinrichs et al. "The UCSC Genome Browser Database: update 2006". In: *Nucleic Acids Research* 34.90001 (Jan. 2006), pp. D590–D598. ISSN: 0305-1048. DOI: 10.1093/nar/gkj144. URL: https://www.ncbi.nlm.nih.gov/pubmed/16381938.

[91] P. Danecek et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (Aug. 2011), pp. 2156–2158. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330. URL: https://www.ncbi.nlm.nih.gov/pubmed/21653522.

[92] Alkes L. Price et al. "Long-Range LD Can Confound Genome Scans in Admixed Populations". In: *The American Journal of Human Genetics* 83.1 (July 2008), pp. 132–135. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2008.06.005. URL: https://www.ncbi.nlm.nih.gov/pubmed/18606306.

[93] John Novembre et al. "Genes mirror geography within Europe". In: *Nature* 456.7218 (Nov. 2008), pp. 98–101. ISSN: 0028-0836. DOI: 10.1038/nature07331. URL: https://www.ncbi.nlm.nih.gov/pubmed/18758442.

[94] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (Dec. 2015), p. 7. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/25722852.

[95] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (Aug. 2006), pp. 904–909. ISSN: 1061-4036. DOI: 10.1038/ng1847. URL: https://www.ncbi.nlm.nih.gov/pubmed/16862161.

[96] A. A. Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10 (May 2012), pp. 1353–1358. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts163. URL: https://www.ncbi.nlm.nih.gov/pubmed/22492648.

[97] G A Churchill and R W Doerge. "Empirical threshold values for quantitative trait mapping." In: *Genetics* 138.3 (Nov. 1994), pp. 963–71. ISSN: 0016-6731. URL: http://www.ncbi.nlm.nih.gov/pubmed/7851788.

[98] Yoav Benjamini and Yosef Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. 1995. DOI: 10.2307/2346101. URL: https://www.jstor.org/stable/2346101.

[99] Stephen D Turner. "qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots". In: *The Journal of Open Source Software* 3.25 (2018), p. 731. DOI: 10.21105/joss.00731org/10.21105/joss.00731. URL: https://cran.r-project.org/package=qqman..

[100] Daniel R Zerbino et al. "Ensembl 2018". In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D754–D761. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1098. URL: https://www.ncbi.nlm.nih.gov/pubmed/29155950.

[101] S. Durinck et al. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". In: *Bioinformatics* 21.16 (Aug. 2005), pp. 3439–3440. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti525. URL: http://www.ncbi.nlm.nih.gov/pubmed/16082012.

[102] Steffen Durinck et al. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nature Protocols* 4.8 (Aug. 2009), pp. 1184–1191. ISSN: 1754-2189. DOI: 10.1038/nprot.2009.97. URL: http://www.ncbi.nlm.nih.gov/pubmed/19617889.

[103] Lin Li et al. "Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma." In: *Frontiers in genetics* 4 (2013), p. 103. ISSN: 1664-8021. DOI: 10.3389/fgene.2013.00103. URL: http://www.ncbi.nlm.nih.gov/pubmed/23755072.

[104] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btr064. URL: https://www.ncbi.nlm.nih.gov/pubmed/21330290.

[105] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3 (Mar. 2009), R25. ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-3-r25. URL: https://www.ncbi.nlm.nih.gov/pubmed/19261174.

[106] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq". In: *Bioinformatics* 25.9 (May 2009), pp. 1105–1111. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btp120. URL: https://www.ncbi.nlm.nih.gov/pubmed/19289445.

[107] R. J. Pruim et al. "LocusZoom: regional visualization of genome-wide association scan results". In: *Bioinformatics* 26.18 (Sept. 2010), pp. 2336–2337. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq419. URL: https://www.ncbi.nlm.nih.gov/pubmed/20634204.

[108] James T Robinson et al. "Integrative genomics viewer". In: *Nature Biotechnology* 29.1 (Jan. 2011), pp. 24–26. ISSN: 1087-0156. DOI: 10.1038/nbt.1754. URL: https://www.ncbi.nlm.nih.gov/pubmed/21221095.

[109] Elisa Mariella et al. "The length of the expressed 3' UTR is an intermediate molecular phenotype linking genetic variants to complex diseases". In: *Mendeley Data* v1 (2019). DOI: 10.17632/6d8w2p9bzf.1.

Figures

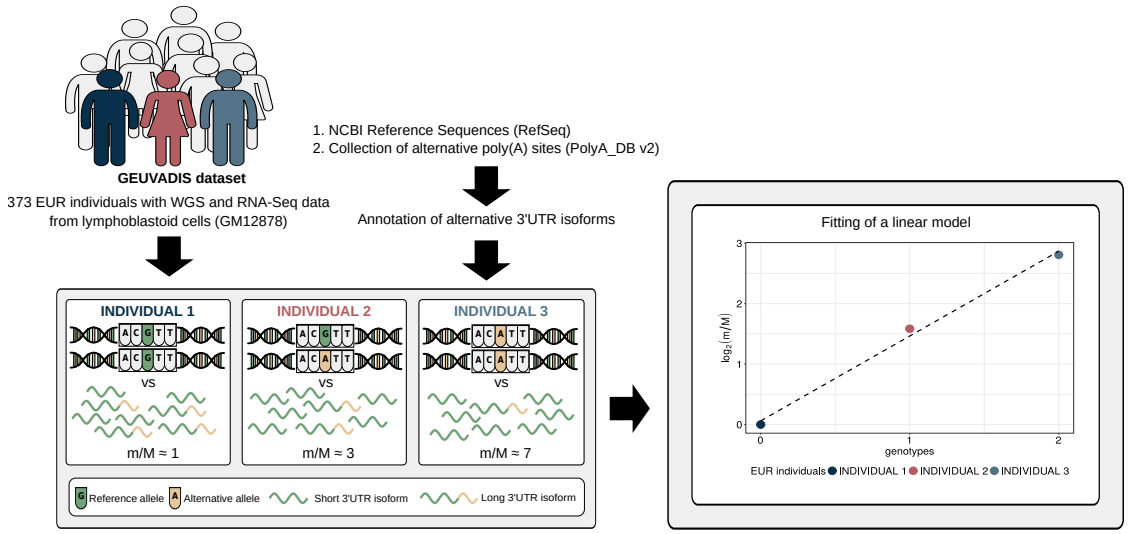


Figure 1: Schematic representation of the method. Genotypic data paired with RNA-Seq data from a large cohort of individuals are required to perform apaQTL mapping analysis. RNA-Seq data are exploited, together with an annotation of alternative 3'UTR isoforms, to compute for each gene the m/M value that is proportional to the ratio between the expression of its short and long 3'UTR isoforms. Then, the association between the m/M values of a gene and each nearby genetic variant is evaluated by linear regression. Genotypes are defined in the standard way: 0 means homozygous for the reference allele, 1 means heterozygous and 2 indicates the presence of two copies of the alternative allele.

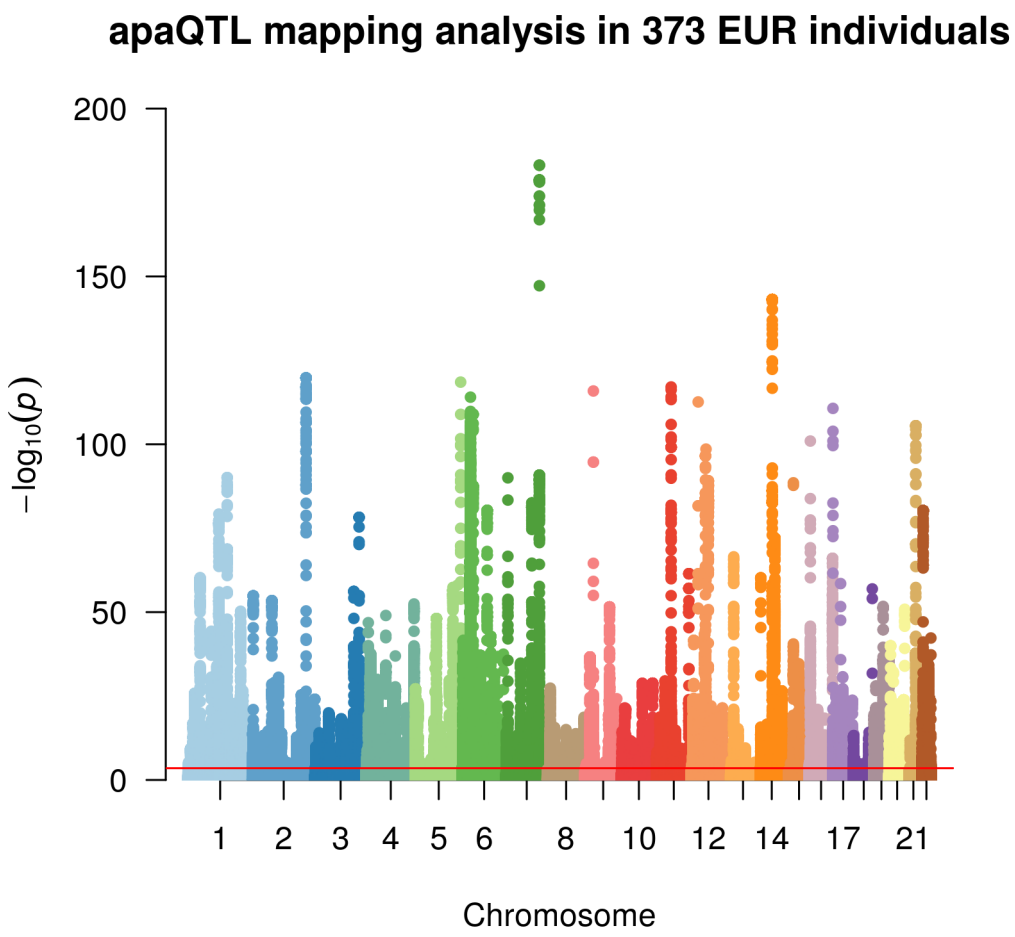


Figure 2: Manhattan plot illustrating the results of the apaQTL mapping analysis at the genome-wide level. For each fitted model, the -$\log_{10}$ P-value is shown according to the position of the tested genetic variant. The red line indicates the threshold for genome-wide statistical significance, after multiple-testing correction (nominal P-value $< 3.1 \times 10^{-4}$).

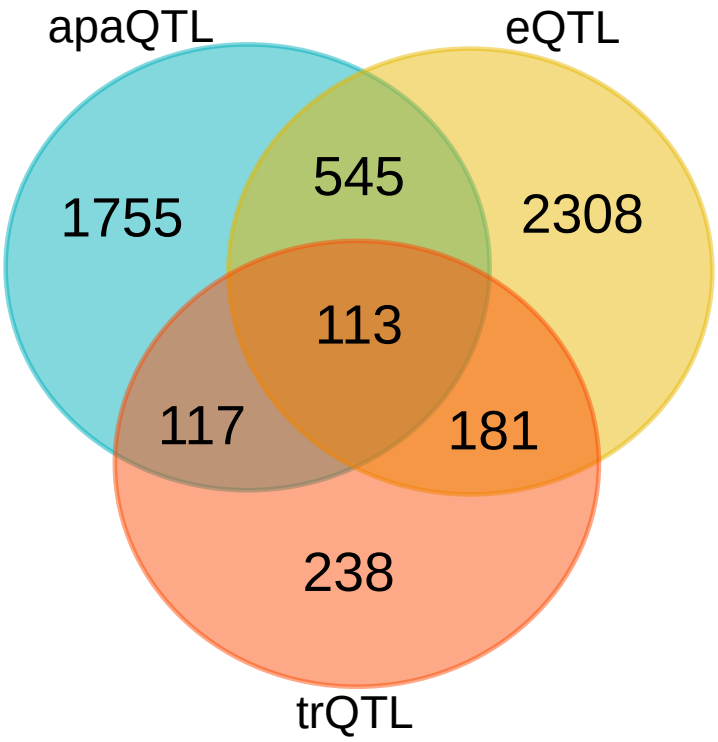


Figure 3: Overlap between genes with significant alternative polyadenylation QTL (apaQTL), expression QTL (eQTL) and transcript ratio QTL (trQTL).

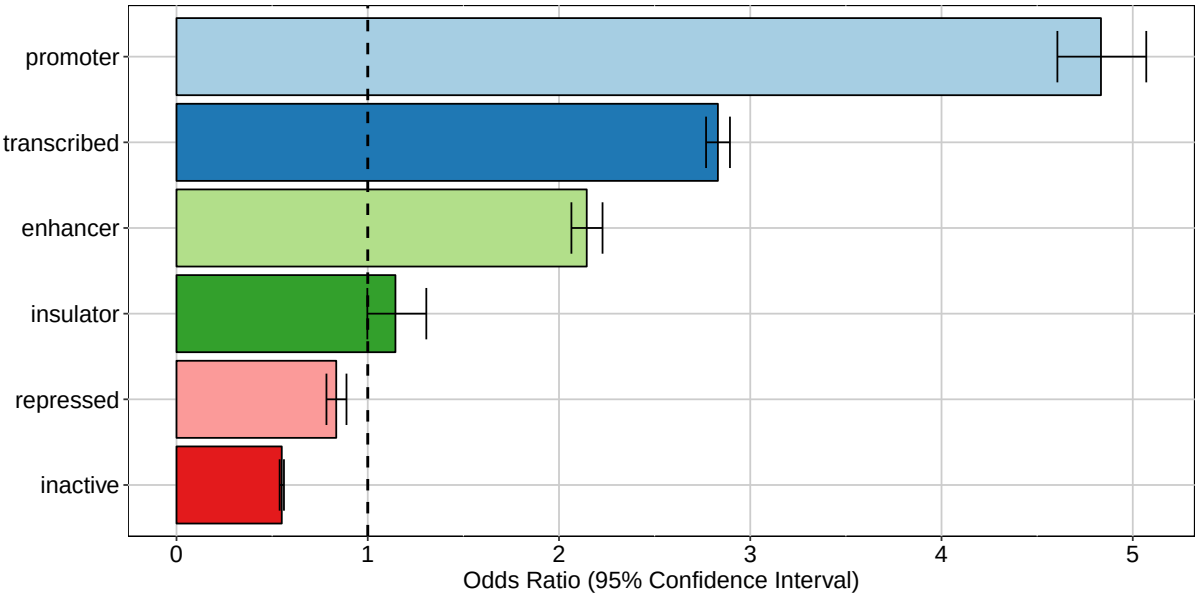


Figure 4: Enrichment of apaQTLs within broad chromatin states that were defined starting from the ChromHMM annotation. For each state, the OR obtained by logistic regression and its 95% CI are shown.

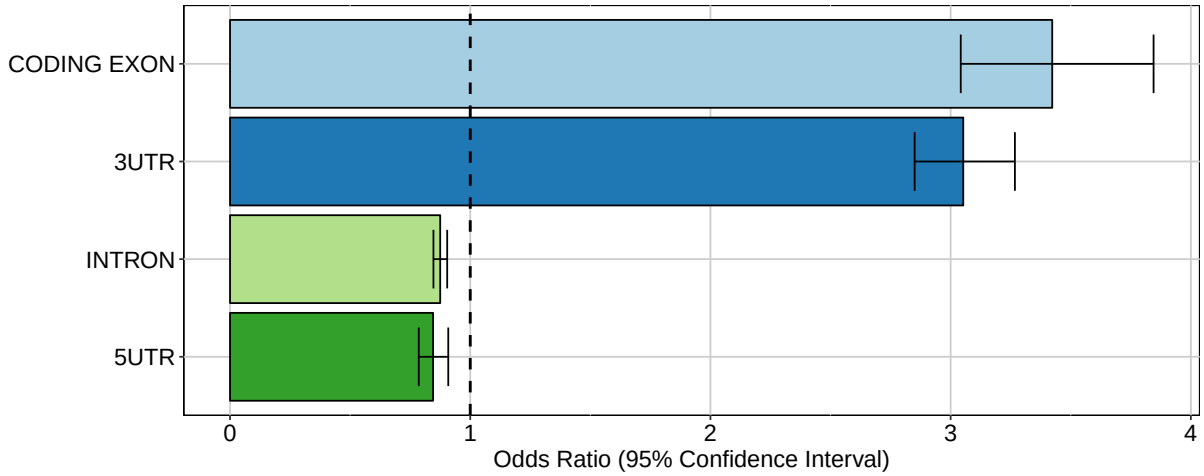


Figure 5: Enrichment of intragenic apaQTLs within coding and non-coding transcript regions. For each gene region, the OR obtained by logistic regression its 95% CI are shown.

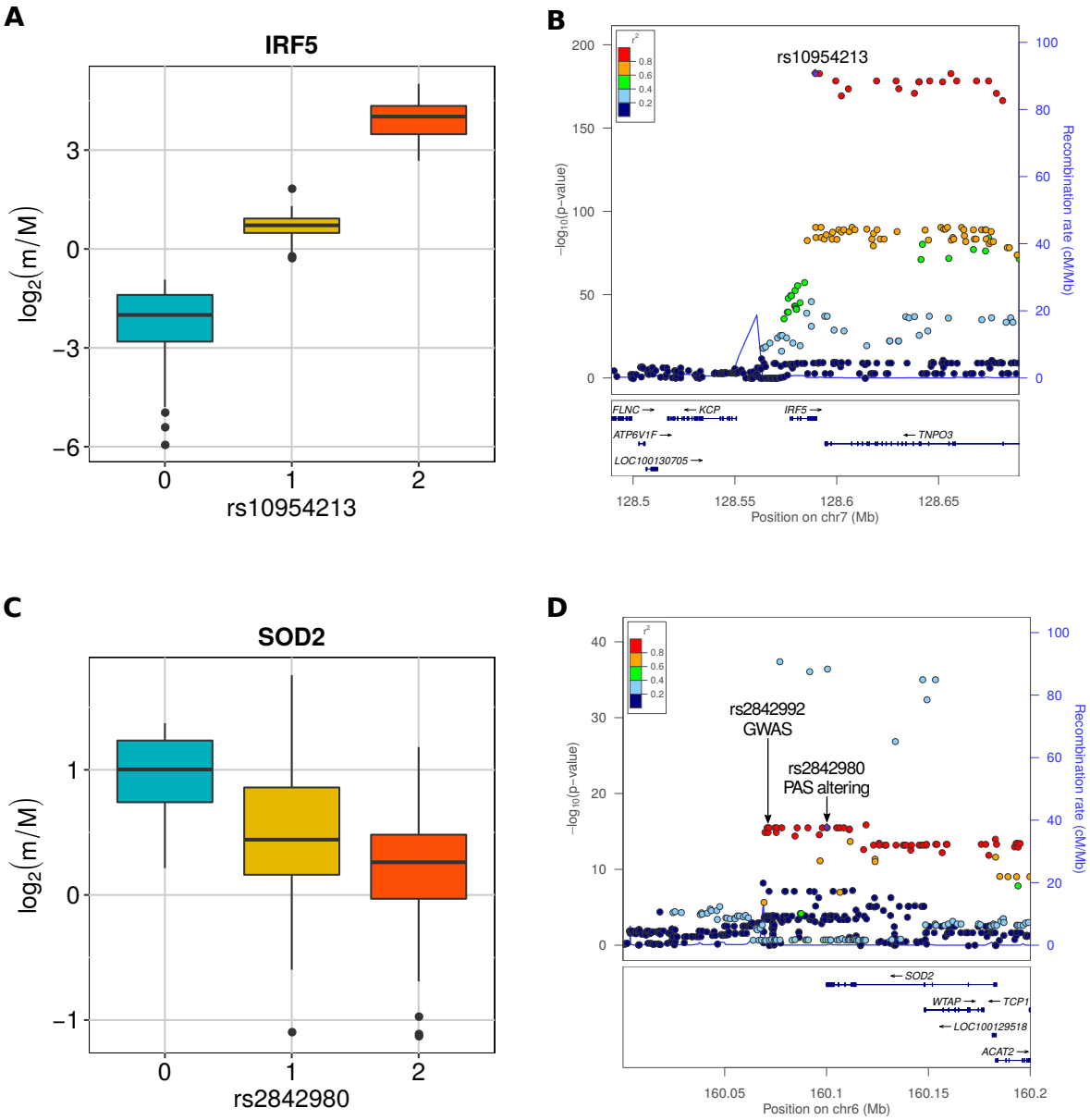


Figure 6: (A) Boxplot showing the variation of the log2-transformed m/M values obtained for IRF5 as a function of the genotype of the individuals for rs10954213. (B) Illustration of the results obtained for IRF5 in the genomic region around rs10954213 (100kb both upstream and downstream its genomic location). In the top panel each tested genetic variant was reported as a function of both its genomic coordinate and its association level with IRF5 (log10-transformed nominal P-value); the points color reflects the LD level (R^2) between rs10954213 and each of the other genetic variants in the locus. The bottom panel shows the genes and their orientation in the locus. The figure was generated by LocusZoom [107]. (C) Boxplot showing the variation of the log2-transformed m/M values obtained for SOD2 as a function of the genotype of the individuals for rs2842980. (D) LocusZoom plot illustrating the results obtained for SOD2 in the genomic region around rs2842980 (100kb both upstream and downstream its genomic location). The plot also shows an intergenic genetic variant (rs2842992) associated to a higher risk of atrophic macular degeneration.

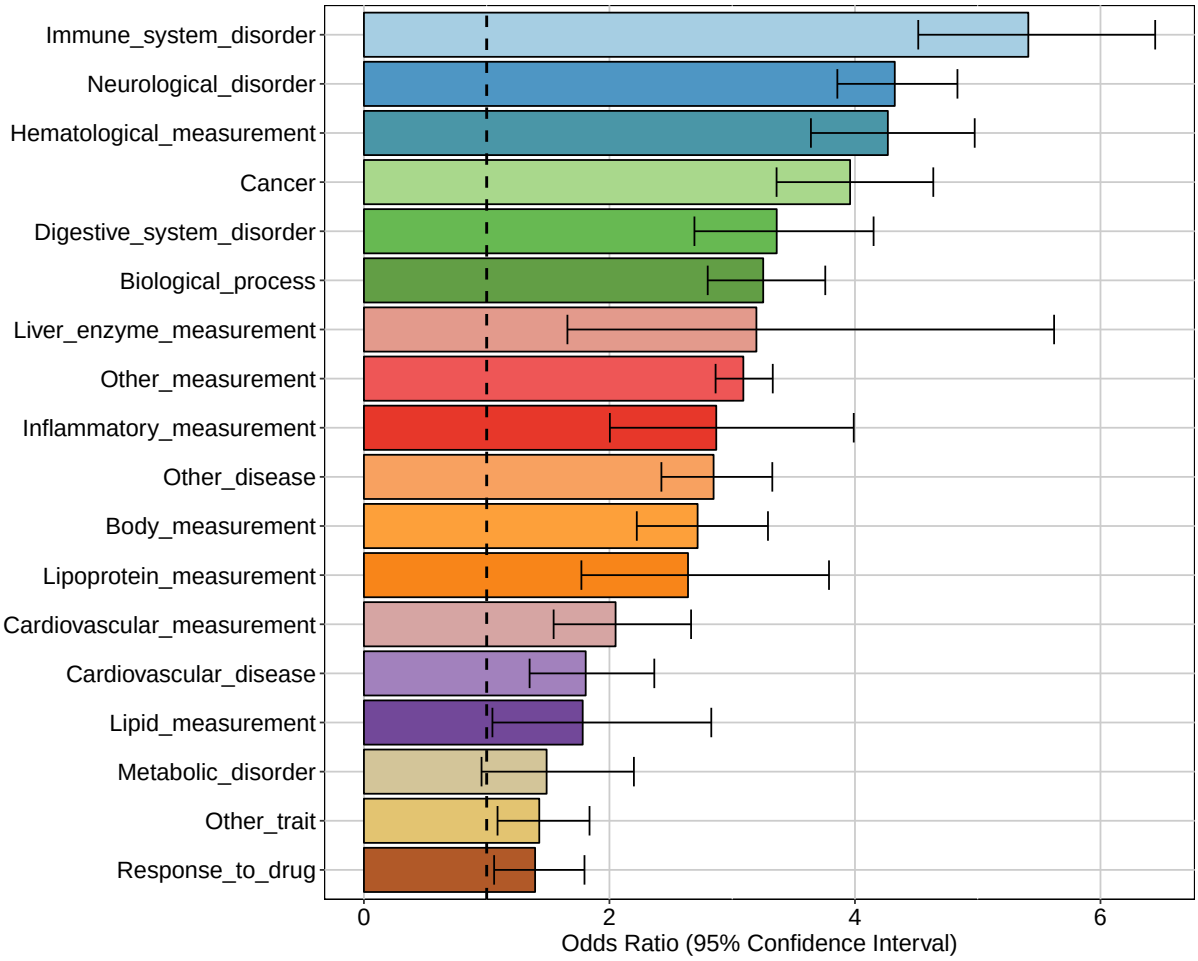


Figure 7: Enrichment of GWAS hits among apaQTLs, for different categories of complex traits. For each category, the OR obtained by logistic regression and its 95% CI are shown.

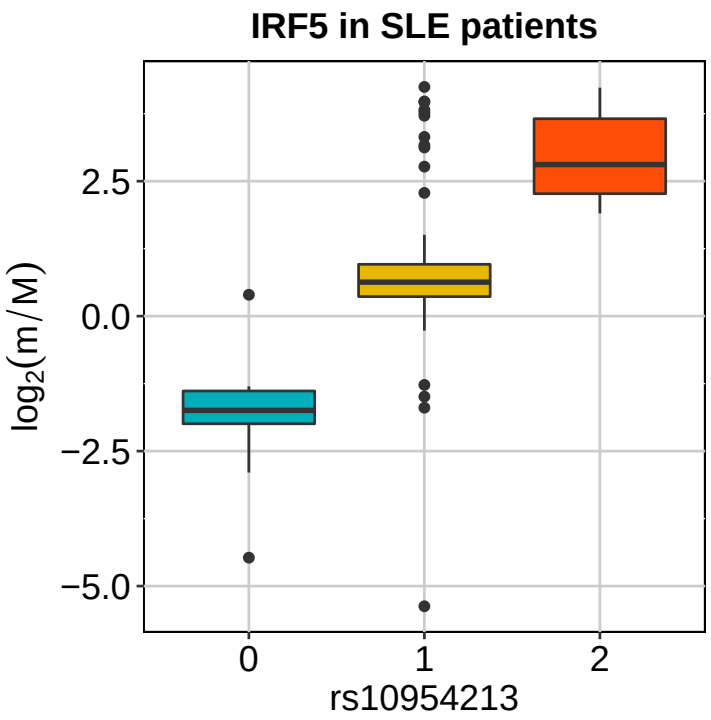
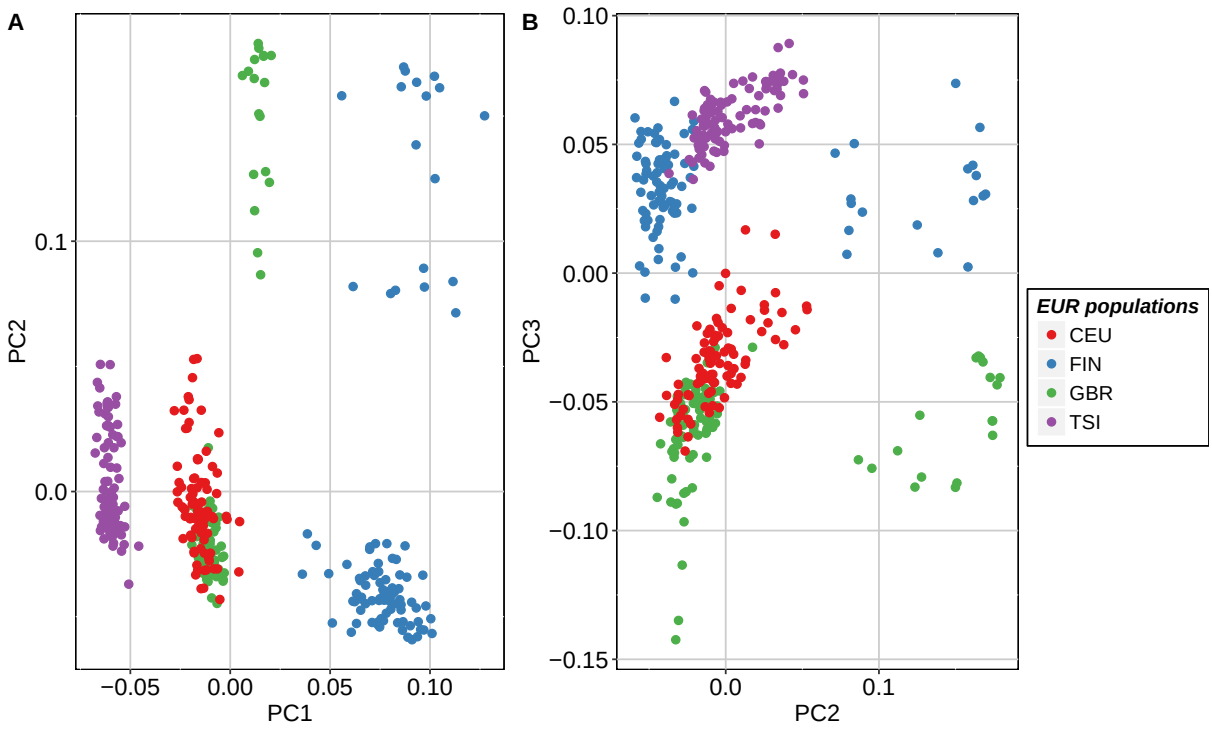
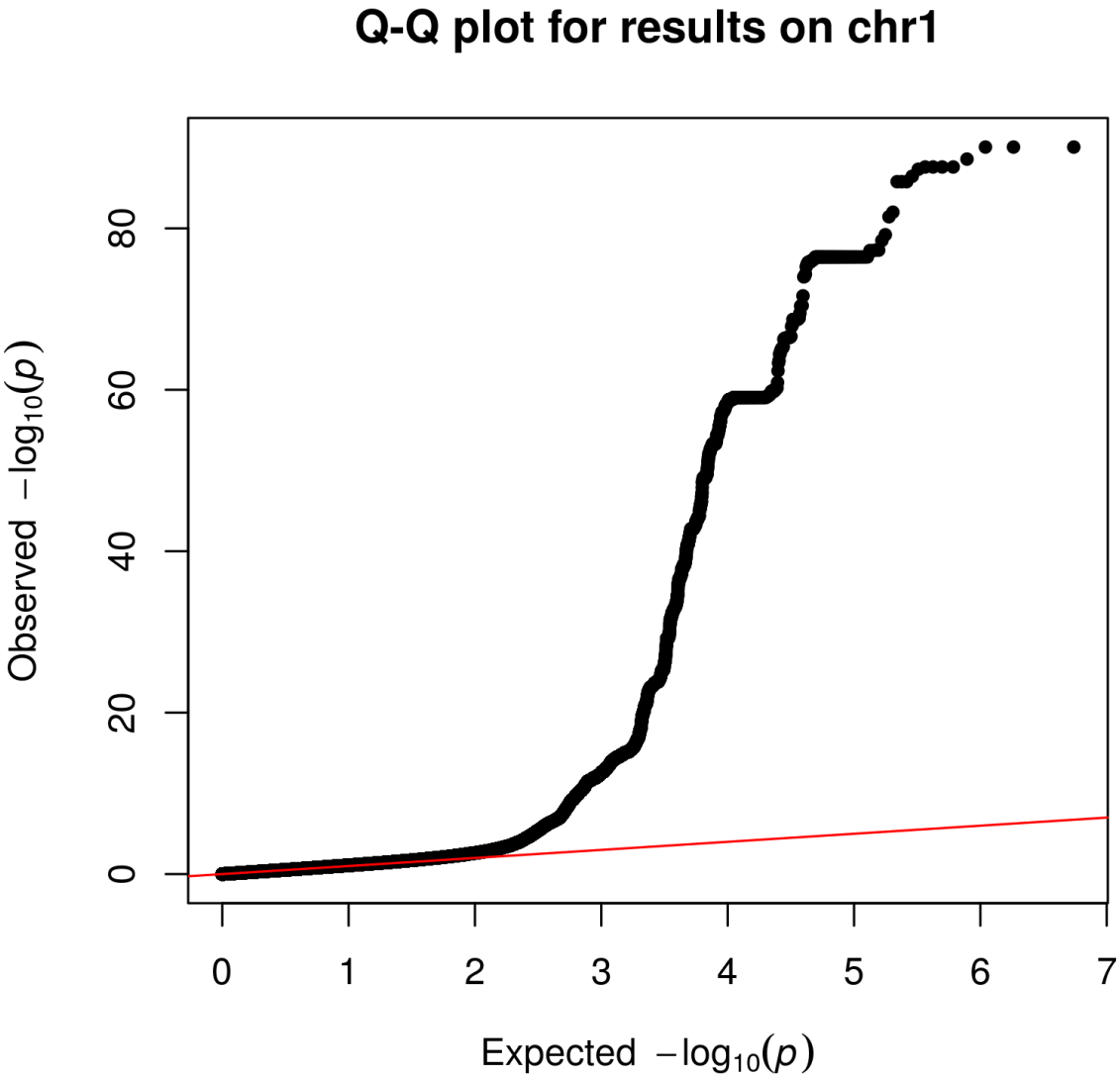


Figure 8: The effect of rs10954213 on the relative expression of the IRF5 alternative isoforms was investigated also in a small cohort of SLE patients. The boxplot show the variation of the log2-transformed m/M values obtained for IRF5 as a function of the genotype of the individuals.

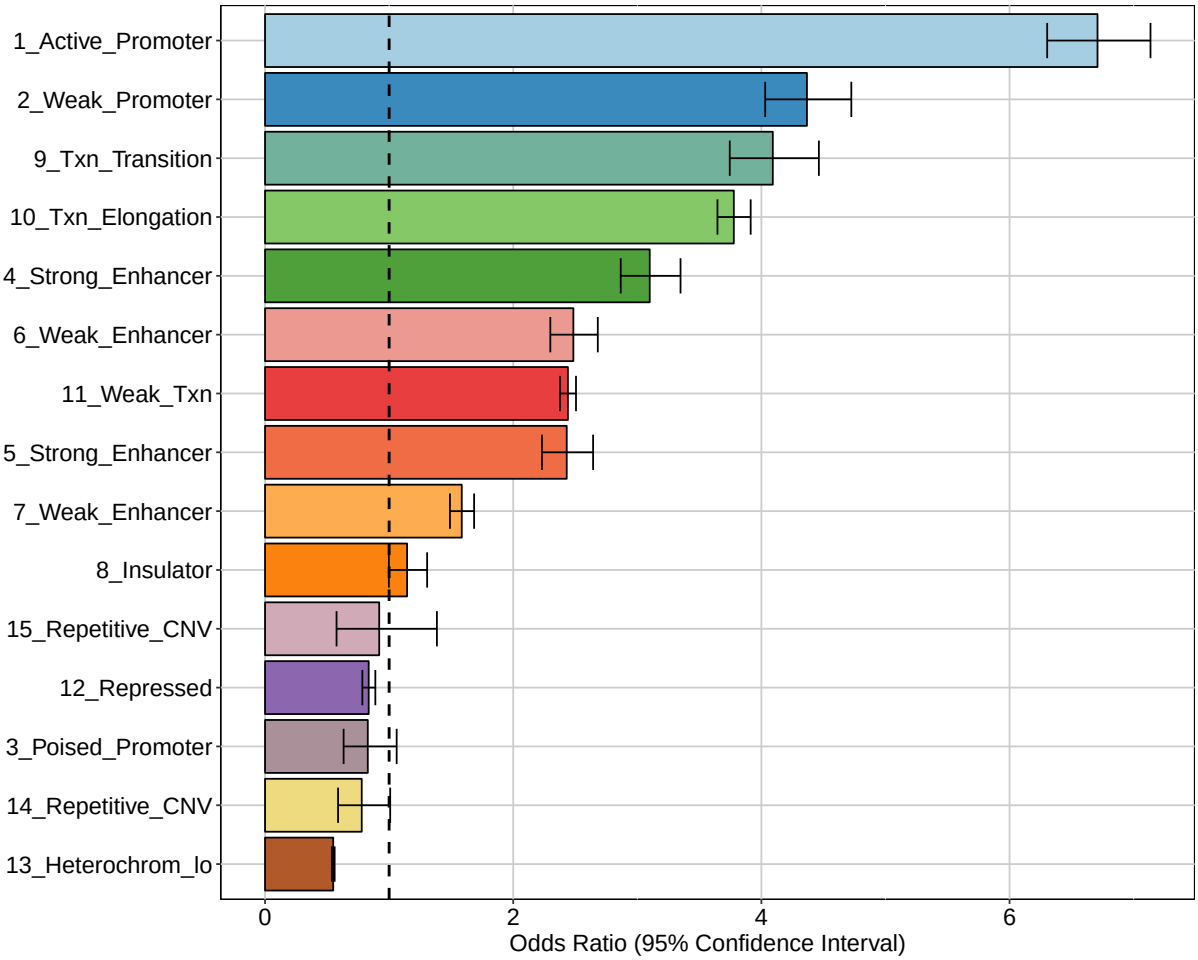
Supplementary figures



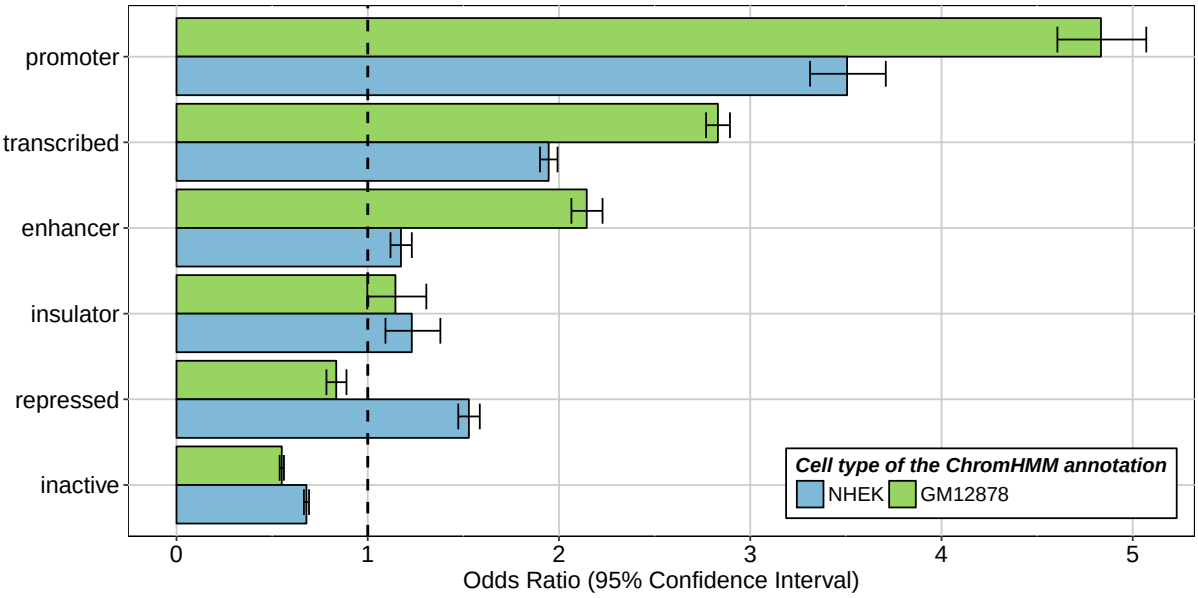
Supplementary figure 1: Principal Component Analysis (PCA) of the genotypic data of the EUR individuals were plotted. Points are colored according to the subpopulation of origin: Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR) and Toscani in Italia (TSI).



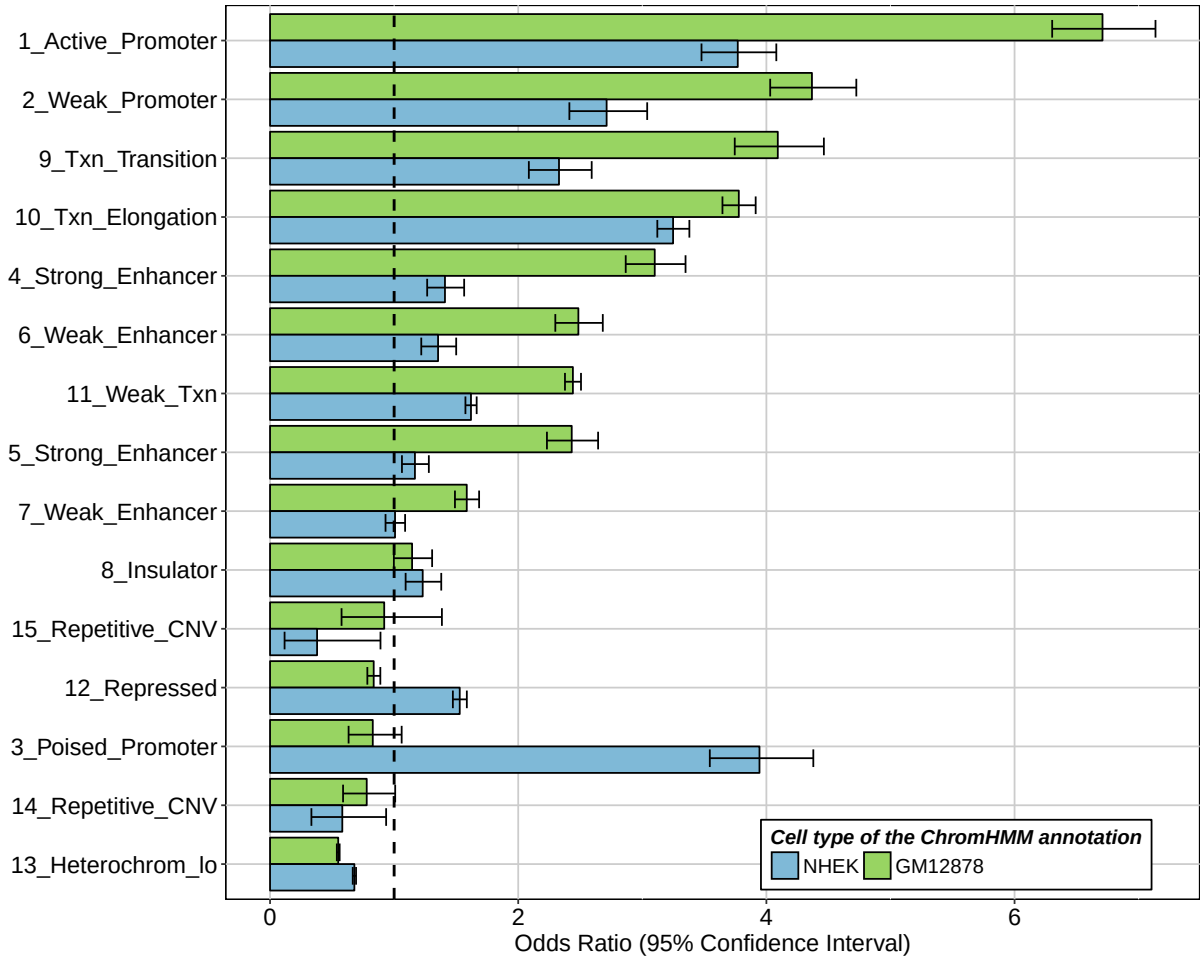
Supplementary figure 2: Q-Q plot comparing the distribution of P-values obtained fitting apaQTL models for genes on chr1 with the expected uniform distribution. It was generated by the CRAN R package `qqman`.



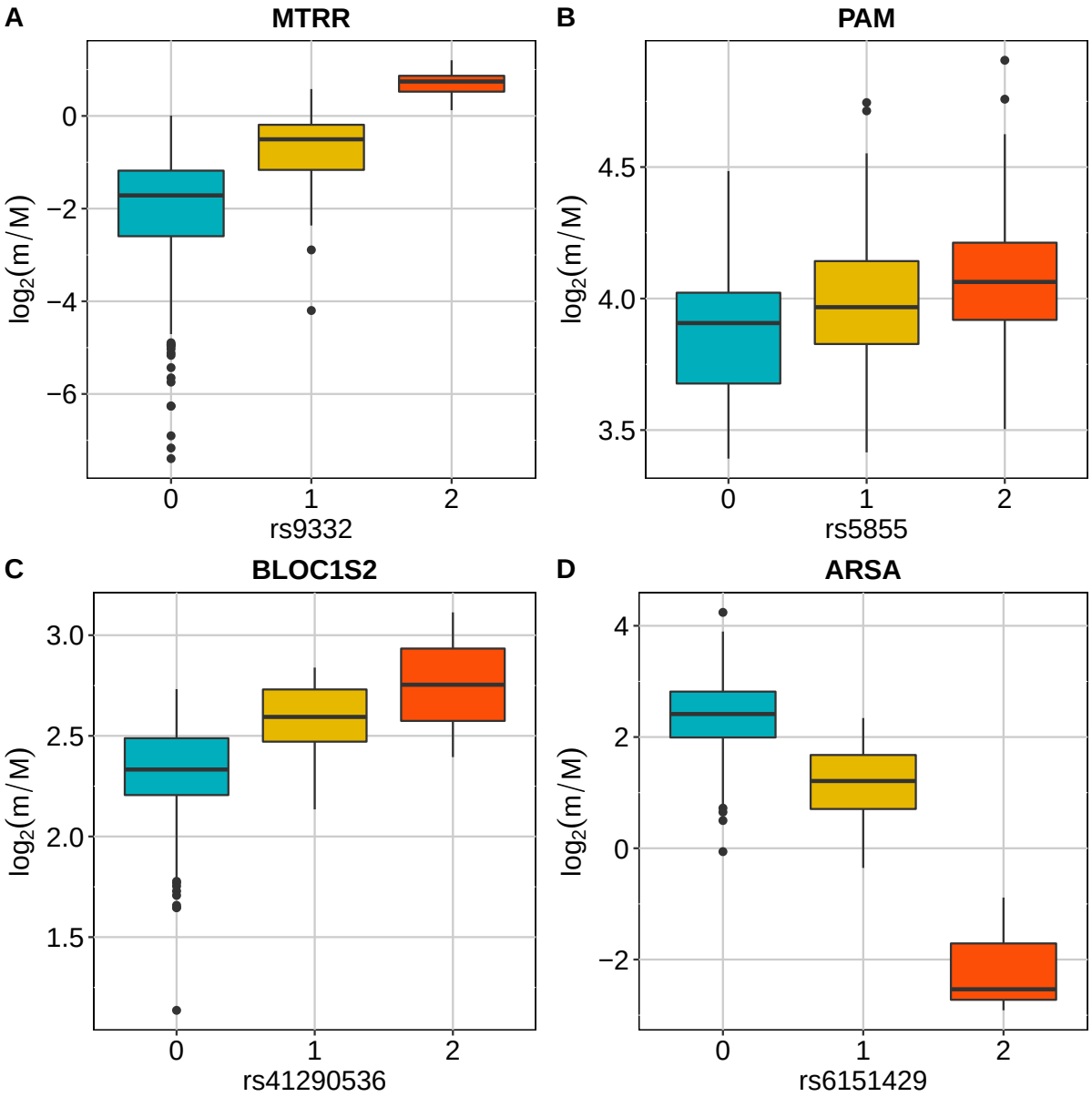
Supplementary figure 3: Enrichment of apaQTLs within chromatin states, taking into account all the 15 chromatin states reported in the ChromHMM annotation. For each of them, the OR obtained by logistic regression and its 95% CI are shown.



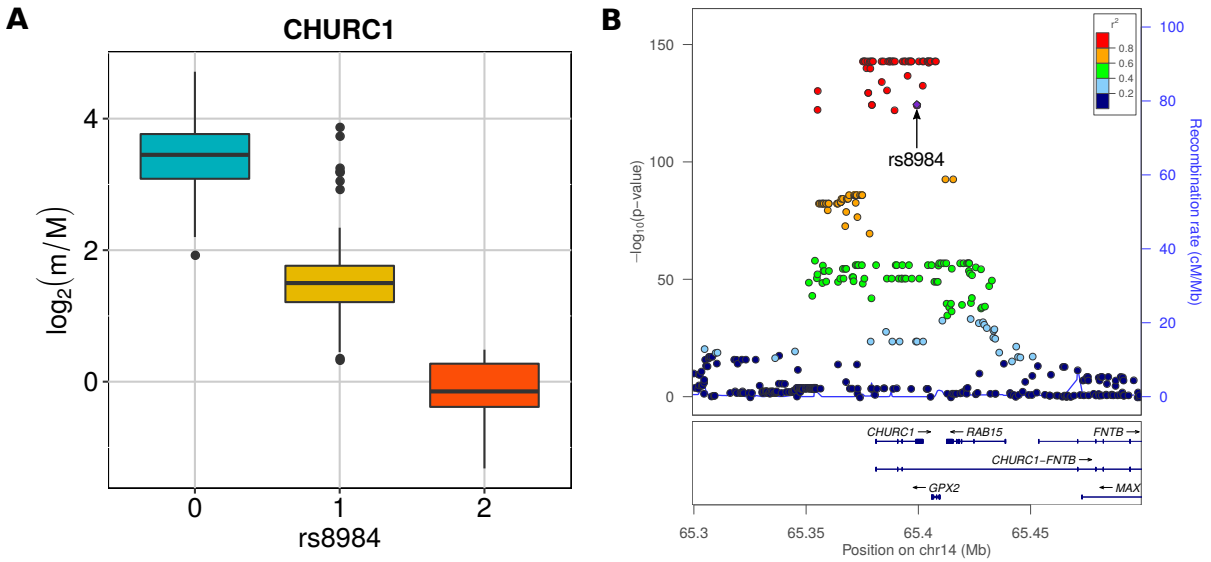
Supplementary figure 4: The results of the enrichment analysis performed with the broad chromatin states of the relevant cell type were compared with those obtained using the ChromHMM annotation of another cell type (NHEK). For each category, the OR obtained by logistic regression and the corresponding 95% CI are shown.



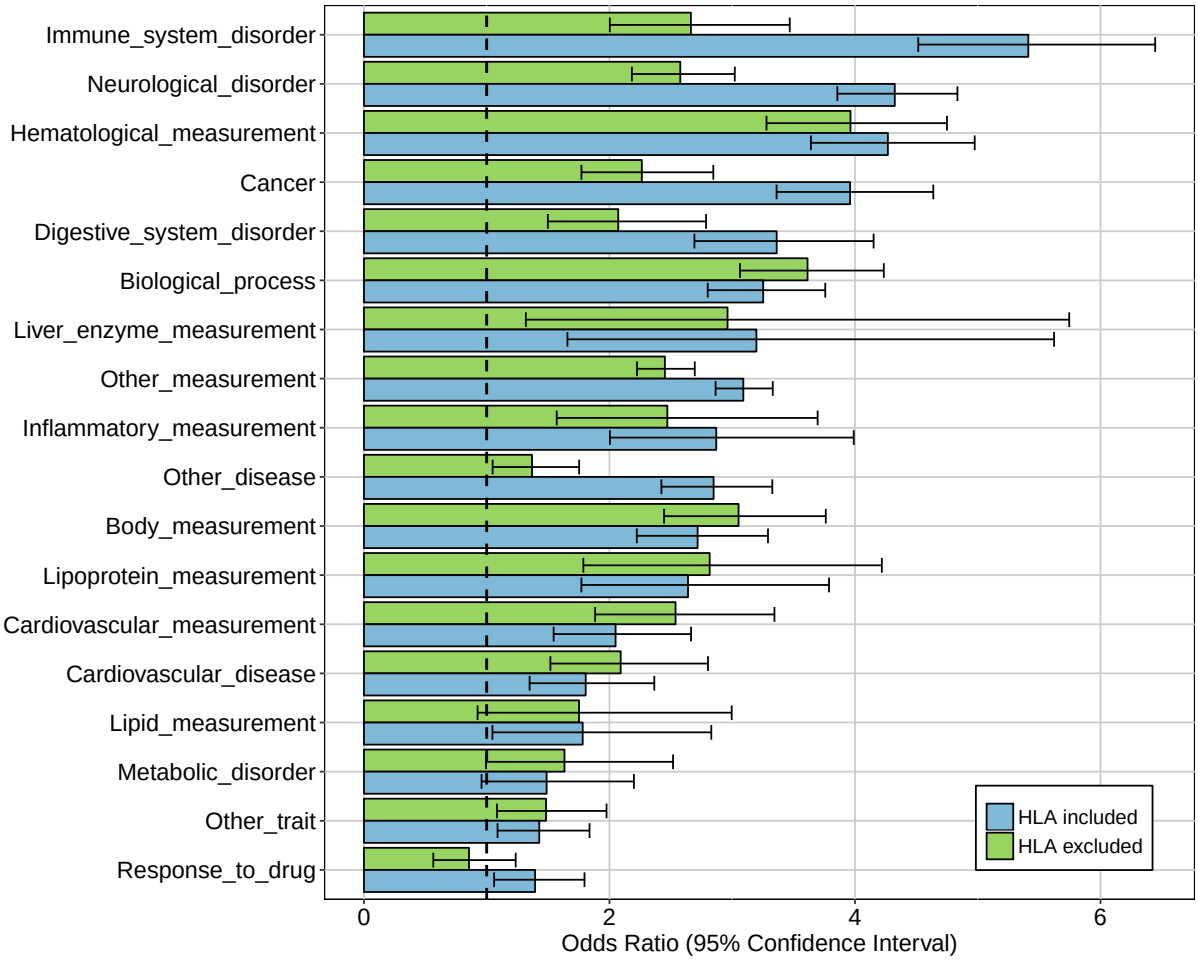
Supplementary figure 5: The results of enrichment analysis done performed with all the chromatin states of the relevant cell type were compared with those obtained using the ChromHMM annotation of another cell type (NHEK). For each category, the OR obtained by logistic regression and the corresponding 95% CIs are shown.



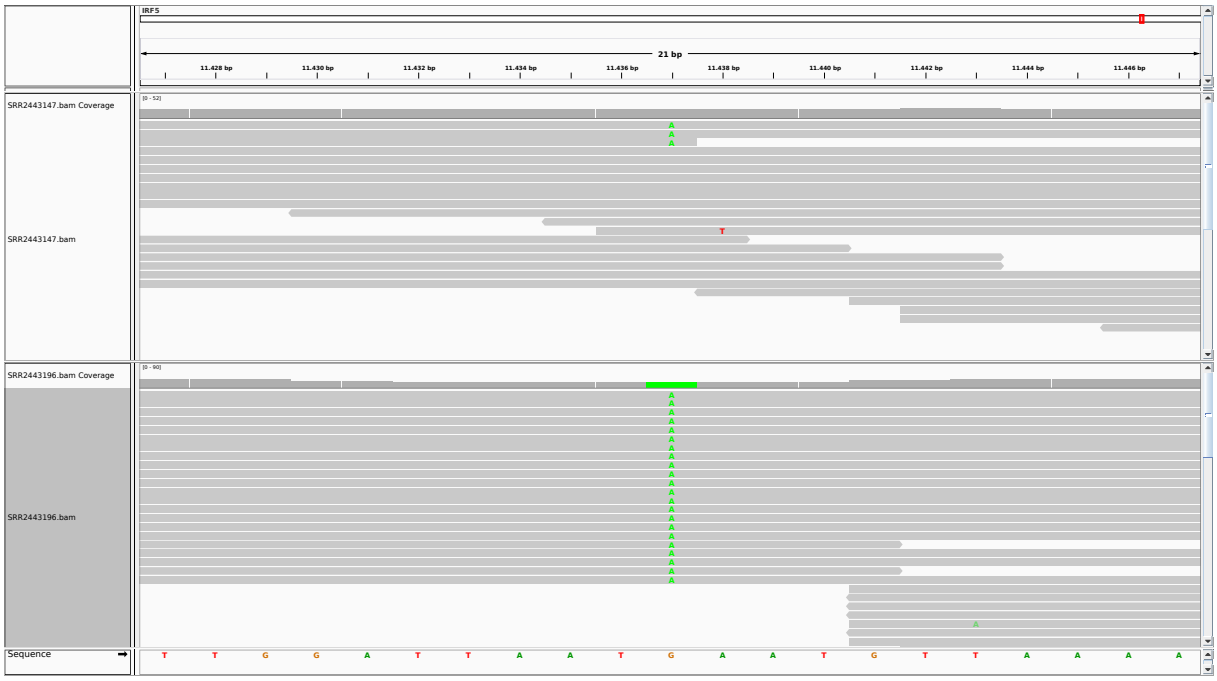
Supplementary figure 6: Boxplots showing the variation of the log2-transformed m/M values obtained for MTRR (A), PAM (B), BLOC1S2 (C) and ARSA (D), as a function of the genotype of the individuals for a single genetic variant that falls within the cis-window of the tested gene (rs9332, rs5855, rs41290536 and rs6151429, respectively).



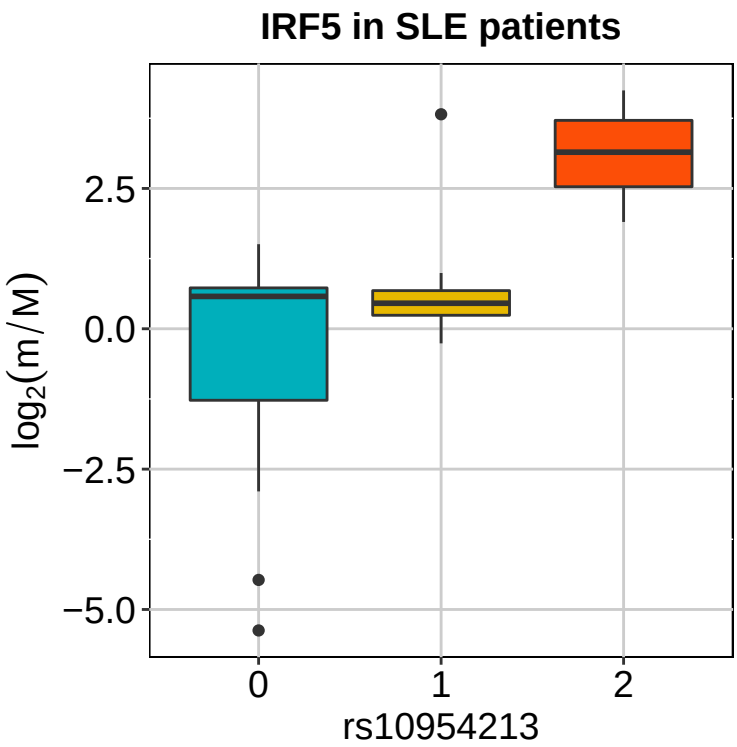
Supplementary figure 7: (A) Boxplot showing the variation of log2-transformed m/M values obtained for CHURC1 as a function of the genotype of the individuals for rs8984. (B) LocusZoom plot illustrating the results obtained for CHURC1 in the genomic region around rs8984 (100kb both upstream and downstream its genomic location).



Supplementary figure 8: Comparison of the results of the enrichment analyses performed for multiple categories of complex traits considering all the studied genetic variants (HLA included) or after having excluded those that are located within the HLA locus (HLA excluded). For each category, the OR obtained by logistic regression and the corresponding 95% CIs are shown.



Supplementary figure 9: Genotypic information was not available for the SLE patients, therefore their genotype in correspondence of the rs10954213 genetic variant was inferred from RNA-Seq data. In the main analysis, having at least one aligned read with a different nucleotide was considered sufficient to call a heterozygous individual. The figure shows the alignment of RNA-Seq reads in a region around the variant with respect to a reduced genome including only the IRF5 gene and was generated using the `Integrative Genomics Viewer (IGV)` software [108]. For example, the SRR2443147 sample (top panel) was considered heterozygous, while the SRR2443196 sample (bottom panel) was considered homozygous for the alternative allele.



Supplementary figure 10: Boxplot showing the variation of the log2-transformed values obtained for IRF5 in SLE patients as a function of the genotype of the individuals when an alternative criterium was adopted to define the genotypic classes.

Supplementary tables

1. **Supplementary Table 1: Enrichment of RBP-altering SNPs among intragenic apaQTL.**
   The table includes the following information for each RNA motif for which a significant enrichment was observed: the identifier of the RNA motif in the CISBP-RNA database (MOTIF ID), the name of the matched RBPs that are expressed in the GM12878 (RBP NAMES), the odds ratio (OR), its 95% confidence interval (95% CI), and the corresponding P-value before and after multiple testing correction by the Benjamini-Hochberg method (P-VALUE and FDR).

2. **Supplementary Table 2: Enrichment of GWAS hits for different trait categories among apaQTL**
   The table includes the following information for each trait category: URI of the trait category in the EFO database (EFO URI) and the associated name (EFO TERM), odds ratio (OR), its 95% confidence interval (95% CI), and the corresponding P-value before and after multiple testing correction by the Benjamini-Hochberg method (P-VALUE and FDR).

3. **Supplementary Table 3: Enrichment of GWAS hits for different trait categories among apaQTL, after the exclusion of genetic variants within the HLA locus.**
   The table includes the following information for each trait category:URI of the trait category in the EFO database (EFO URI) and the associated name (EFO TERM), odds ratio (OR), its 95% confidence interval (95% CI), and the corresponding P-value before and after multiple testing correction by the Benjamini-Hochberg method (P-VALUE and FDR).

Supplementary data

1. **Supplementary File 1: Annotation of gene structures**
   GTF file with the custom gene annotation that was used for all the analyses.

2. **Supplementary File 2: Annotation of alternative 3'UTR isoforms**
   GTF file used for the computation of m/M values. For each gene the coordinates of the PRE and POST segments were obtained combining its structure annotation with the poly(A) sites reported by PolyADB_2. In addition the length of each of these segments is reported.

3. **Supplementary File 3: Results of the fitted apaQTL models**
   For each model we report: the genetic variant identifier according to dpSNP137, as provided into the GEUVADIS dataset (SNP_ID), the NCBI Entrez ID (GENE_ID), the

regression coefficient (BETA), the nominal P-value (NOMINAL_PVALUE), the corresponding empirical P-value (EMPIRICAL_PVALUE) and Benjamini-Hochberg corrected empirical P-value (FDR). Multiple files listing the results obtained in each chromosome are publicly available on Mendeley Data [109].

4. **Supplementary File 4: Significant apaQTL models**
   Table listing the significant models obtained in the apaQTL mapping analysis. For each model we report: the genetic variant identifier (SNP_ID), the NCBI Entrez ID (GENE_ID), the regression coefficient (BETA), the nominal P-value (NOMINAL_PVALUE), the corresponding empirical P-value (EMPIRICAL_PVALUE) and Benjamini-Hochberg corrected empirical P-value (FDR).

5. **Supplementary File 5: Enrichment of trait-specific GWAS hits among apaQTL**
   The table includes the following information for each GWAS trait for which we found a significant enrichment: the URI of the trait in the EFO database (EFO_URI), the associated name (EFO_TERM) and its parent term in the EFO database (EFO_PARENT_TERM), the odds ratio (OR), its 95% confidence interval (95% CI) and the corresponding P-value before and after multiple testing correction by the Benjamini-Hochberg method (P-VALUE and FDR).

6. **Supplementary File 6: Enrichment of trait-specific GWAS hits among apaQTL, after the exclusion of genetic variants within the HLA locus**
   The table includes the following information for each GWAS trait for which we found a significant enrichment after the exclusion of genetic variants within the HLA locus: the URI of the trait in the EFO database (EFO_URI), the associated name (EFO_TERM) and its parent term in the EFO database (EFO_PARENT_TERM), the odds ratio (OR), its 95% confidence interval (95% CI) and the corresponding P-value before and after multiple testing correction by the Benjamini-Hochberg method (P-VALUE and FDR).

Supplementary tables

Supplementary Table 1: Enrichment of RBP-altering SNPs among intragenic apaQTL.

MOTIF ID	RBP NAMES	OR	95% CI	P-VALUE	FDR
M016_0.6	FMR1	3.72	1.01-11.3	0.0278	0.266
M025_0.6	HNRNPC	1.77	1.06-2.83	0.0212	0.265
M070_0.6	ENSG00000180771;SRSF2	3.83	1.55-8.69	0.00197	0.0521
M075_0.6	TIA1	1.81	0.994-3.1	0.0388	0.327
M081_0.6	CSDA;YB-1	3.9	1.03-12.6	0.0285	0.266
M089_0.6	HNRNPL	3.4	1.08-9.12	0.0217	0.265
M122_0.6	MEX3B;MEX3C;MEX3D	3.85	1.03-12	0.0266	0.266
M140_0.6	ENOX1;ENOX2	3.16	1.36-6.75	0.00433	0.0778
M145_0.6	RBM5	4.98	1.51-14.6	0.0044	0.0778
M147_0.6	CNOT4	2.11	0.947-4.25	0.0479	0.363
M156_0.6	TIA1	1.81	1.05-2.96	0.0246	0.266
M158_0.6	HNRNPCL1	1.77	1.06-2.83	0.0212	0.265
M160_0.6	KHDRBS1	4.15	1.54-10.2	0.00271	0.0616
M250_0.6	CSDA	2.93	0.938-7.73	0.0411	0.327
M256_0.6	ACO1	1.92	1.07-3.25	0.0209	0.265
M291_0.6	EIF4B	10.6	3.5-33.2	2.49×10^{-5}	0.00132
M292_0.6	EIF4B	1.98	1.32-2.89	0.000627	0.0249
M320_0.6	MBNL1;MBNL2;MBNL3	1.66	1.2-2.25	0.00161	0.0513
M333_0.6	SRSF9	1.8	0.99-3.07	0.0397	0.327
M344_0.6	RBMX;RBMXL1;RBMXL2	1.77	1.4-2.2	6.44×10^{-7}	5.12×10^{-5}

Supplementary Table 2: Enrichment of GWAS hits for different trait categories among apaQTL.

EFO URI	EFO TERM	OR	95% CI	P-VALUE	FDR
EFO_0000540	Immune_system_disorder	5.41	4.52-6.45	2.5×10^{-77}	1.5×10^{-76}
EFO_0000618	Neurological_disorder	4.32	3.86-4.83	2.47×10^{-142}	2.23×10^{-141}
EFO_0004503	Hematological_measurement	4.27	3.64-4.98	2.89×10^{-74}	1.3×10^{-73}
EFO_0000616	Cancer	3.96	3.36-4.64	4.15×10^{-63}	1.49×10^{-62}
EFO_0000405	Digestive_system_disorder	3.36	2.69-4.15	4.34×10^{-28}	9.76×10^{-28}
GO_0008150	Biological_process	3.25	2.8-3.76	9.69×10^{-56}	2.91×10^{-55}
EFO_0004582	Liver_enzyme_measurement	3.2	1.66-5.62	0.000167	0.000215
EFO_0001444	Other_measurement	3.09	2.86-3.33	1.22×10^{-189}	2.19×10^{-188}
EFO_0004872	Inflammatory_measurement	2.87	2-3.99	1.77×10^{-09}	3.19×10^{-09}
EFO_0000408	Other_disease	2.85	2.42-3.33	2.12×10^{-38}	5.46×10^{-38}
EFO_0004324	Body_measurement	2.72	2.22-3.29	1.59×10^{-23}	3.19×10^{-23}
EFO_0004732	Lipoprotein_measurement	2.64	1.77-3.79	5.06×10^{-07}	7.58×10^{-07}
EFO_0004298	Cardiovascular_measurement	2.05	1.55-2.66	2.29×10^{-07}	3.74×10^{-07}
EFO_0000319	Cardiovascular_disease	1.81	1.35-2.37	3.47×10^{-05}	4.8×10^{-05}
EFO_0004529	Lipid_measurement	1.78	1.05-2.83	0.0219	0.0232
EFO_0000589	Metabolic_disorder	1.49	0.958-2.2	0.0596	0.0596
EFO_0000001	Other_trait	1.43	1.09-1.84	0.00751	0.00902
GO_0042493	Response_to_drug	1.39	1.06-1.8	0.0132	0.0149

Supplementary Table 3: Enrichment of GWAS hits for different trait categories among apaQTL, after the exclusion of genetic variants within the HLA locus.

EFO URI	EFO TERM	OR	95% CI	P-VALUE	FDR
EFO_0004503	Hematological_measurement	3.96	3.28-4.75	3.45×10^{-48}	2.07×10^{-47}
GO_0008150	Biological_process	3.61	3.06-4.23	1.44×10^{-54}	1.29×10^{-53}
EFO_0004324	Body_measurement	3.05	2.45-3.76	2.99×10^{-24}	1.08×10^{-23}
EFO_0004582	Liver_enzyme_measurement	2.96	1.32-5.75	0.00338	0.00467
EFO_0004732	Lipoprotein_measurement	2.82	1.79-4.22	2.05×10^{-06}	3.99×10^{-06}
EFO_0000540	Immune_system_disorder	2.66	2-3.47	2.45×10^{-12}	7.36×10^{-12}
EFO_0000618	Neurological_disorder	2.58	2.18-3.02	3.14×10^{-30}	1.41×10^{-29}
EFO_0004298	Cardiovascular_measurement	2.54	1.88-3.34	1.83×10^{-10}	4.11×10^{-10}
EFO_0004872	Inflammatory_measurement	2.47	1.57-3.7	3.14×10^{-05}	4.71×10^{-05}
EFO_0001444	Other_measurement	2.45	2.22-2.7	5×10^{-75}	9.01×10^{-74}
EFO_0000616	Cancer	2.26	1.77-2.85	1.3×10^{-11}	3.35×10^{-11}
EFO_0000319	Cardiovascular_disease	2.09	1.52-2.8	2.22×10^{-06}	3.99×10^{-06}
EFO_0000405	Digestive_system_disorder	2.07	1.5-2.79	4×10^{-06}	6.55×10^{-06}
EFO_0004529	Lipid_measurement	1.75	0.926-3	0.0589	0.0623
EFO_0000589	Metabolic_disorder	1.63	0.994-2.52	0.0373	0.0419
EFO_0000001	Other_trait	1.48	1.08-1.98	0.0099	0.0127
EFO_0000408	Other_disease	1.37	1.05-1.75	0.0164	0.0197
GO_0042493	Response_to_drug	0.856	0.565-1.24	0.435	0.435