frontiers in Genetics

ORIGINAL RESEARCH
published: 16 August 2019
doi: 10.3389/fgene.2019.00714

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] and Paolo Provero[1,2]**

[1] Department of Molecular Biotechnology and Health Sciences, University of Turin, Turin, Italy, [2] Center for Tranlational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milan, Italy

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 posttranscriptional 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.

Keywords: human genetic variants, alternative polyadenylation, quantitative trait loci (QTL), whole-genome sequencing (WGS), RNA sequencing (RNA-Seq), genome-wide association studies (GWAS)

OPEN ACCESS

Edited by:
Mehdi Pirooznia, National Heart, Lung, and Blood Institute (NHLBI), United States

Reviewed by:
Ting Ni, Fudan University, China
Celso Teixeira Mendes-Junior, University of São Paulo, Brazil

***Correspondence:**
Paolo Provero
paolo.provero@unito.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 March 2019*
Accepted: *05 July 2019*
Published: *16 August 2019*

Citation:
Mariella E, Marotta F, Grassi E, Gilotto S and Provero P (2019) The Length of the Expressed 3′ UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases. Front. Genet. 10:714. doi: 10.3389/fgene.2019.00714

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 (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001) and the following large-scale investigations of genetic differences between individuals by efforts such as the 1000 Genome Project Consortium (The 1000 Genomes Project Consortium, 2015) provided the foundation for the study of human genetics at the genome-wide level.

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 (Visscher et al., 2017). 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 noncoding regions, presumably exerting regulatory effects, but usually, we do not know 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 (Cookson et al., 2009; Albert and Kruglyak, 2015). 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 (Manning and Cooper, 2017). Therefore, it is important to identify the genetic variants associated to transcript structure, including splicing and alternative untranslated region (UTR) isoforms, besides those affecting transcriptional levels, and different approaches have been proposed to this end (Lappalainen et al., 2013; Monlong et al., 2014). From these studies, it emerges in particular that genetic variants frequently determine changes in the length of the expressed 3′ UTRs, and that these variants can be located not only within the 3′ UTR itself but also in regulatory regions outside the transcript (Lappalainen et al., 2013). In addition, genome-wide analyses specifically focused on alternative splicing have been performed (Ardlie et al., 2015; Xiong et al., 2015).

Polyadenylation is one of the posttranscriptional modifications affecting pre-mRNAs in the nucleus and involves two steps: the cleavage of the transcript and the addition of a poly(A) tail (Elkon et al., 2013; Tian and Manley, 2017). 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 (Oktaba et al., 2015; Yue et al., 2018). 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 has 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 (Mayr, 2018).

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 (Elkon et al., 2013). 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 (Sandberg et al., 2008; Ji et al., 2009; Mayr and Bartel, 2009; Fu et al., 2011; Masamha et al., 2014), and several examples are known of altered APA regulation associated to human diseases (Chang et al., 2017b; Manning and Cooper, 2017).

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 Lappalainen et al. (2013) found hundreds of common single nucleotide polymorphisms (SNPs) causing the alteration or degradation of motifs that are similar to the canonical PAS (Ferreira et al., 2016) but did not extend the analysis to other possible mechanisms. Other studies found strong associations between genetic variants and APA regulation (Kwan et al., 2008; Thomas and Sætrom, 2012; Yoon et al., 2012; Lappalainen et al., 2013; Zhernakova et al., 2013; Monlong et al., 2014), 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 (Lappalainen et al., 2013). 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 (Lembo et al., 2012), 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) (Ji et al., 2009) and sequencing of APA sites (SAPAs) (Fu et al., 2011), 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 (Grassi et al., 2016; Ha et al., 2018), such the ones provided by PolyA_DB2 (Lee et al., 2007) and APASdb (You et al., 2015), and those that instead try to infer their location from the data (Masamha et al., 2014). Although the latter approach potentially allows analyzing also previously unannotated sites, the former leads to higher sensitivity (Grassi et al., 2016; Ha et al., 2018) 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.

RESULTS

Genetic Variants Affect the Relative Expression of Alternative 3′ UTR Isoforms of Thousands of Genes

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 (**Figure 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 (Lappalainen et al., 2013)]. A collection of known alternative poly(A) sites (Lee et al., 2007) 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 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 1 Mbp 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 (**Figure 2**, **Table 1**, **File S1** and **File S2**).

Our set of significant genes shows only moderate overlap with genes, for which eQTLs or transcript ratio QTLs (trQTLs) were reported in Lappalainen et al. (2013) from the same data (**Figure 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 (Tian and Manley, 2017). 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 mRNA translation rates (Spies et al., 2013; Floor and Doudna, 2016) and localization (An et al., 2008), and protein cytoplasmic localization (Berkovits and Mayr, 2015). 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.

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

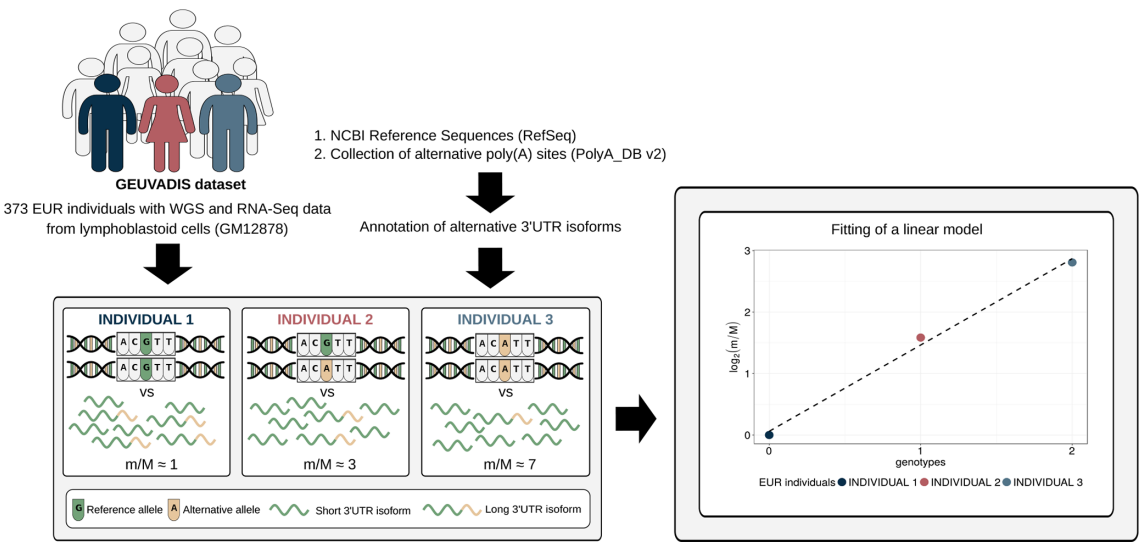

FIGURE 1 | Schematic representation of the method. Genotypic data paired with RNA-Seq data from a large cohort of individuals are required to perform alternative polyadenylation quantitative trait loci (apaQTL) mapping analysis. RNA-Seq data are exploited, together with an annotation of alternative 3′ untranslated region (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.

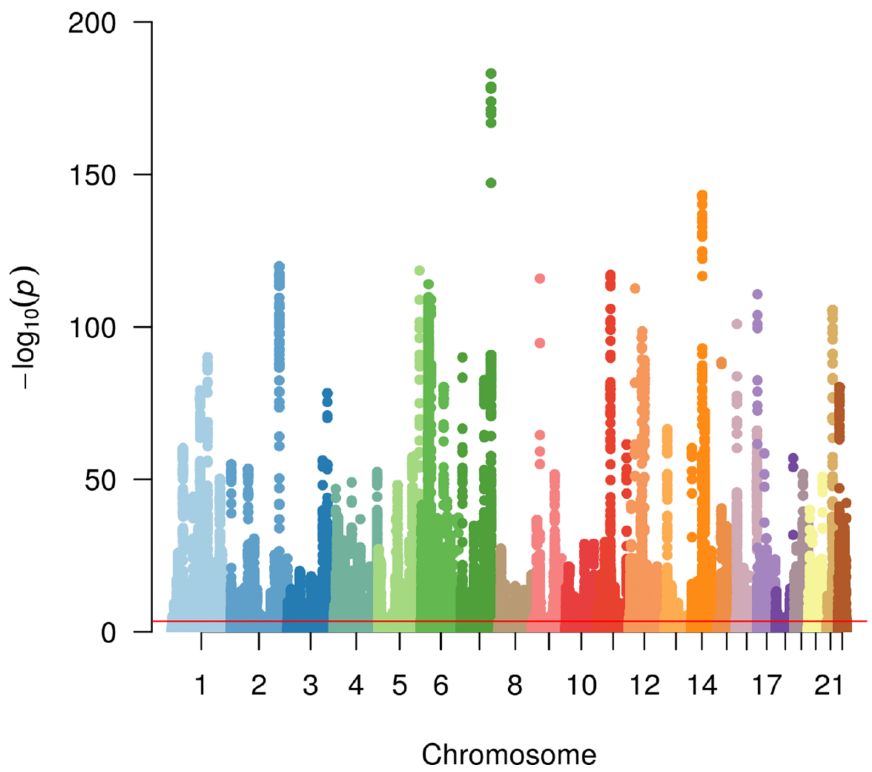

FIGURE 2 | Manhattan plot illustrating the results of the apaQTL mapping analysis. For each fitted model, the $-\log_{10}$ nominal P value is shown according to the position of the tested genetic variant. The red line indicates the threshold for genome-wide statistical significance, after multiple-testing correction (nominal $P < 3.1 \times 10^{-4}$, corresponding to corrected empirical $P < 0.05$).

TABLE 1 | Results of alternative polyadenylation quantitative trait loci (apaQTL) mapping analysis.

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

cells for our data). To verify this hypothesis, we superimposed the apaQTLs to the ChromHMM annotation of the human genome for the GM12878 cell line (Ernst et al., 2011) and used logistic regression, as detailed in the *Material and 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 (**Figure 4**), defined following Ernst et al. (2011) or all 15 chromatin states reported by ChromHMM (**Figure S1**).

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 (**Figure S2** and **Figure S3**). Taken together, these results show that genetic variants affecting alternative polyadenylation tend to be located in cell-type-specific active chromatin regions.

The detection of a significant apaQTL enrichment within promoters and enhancers suggests that also these genomic regions may be involved in the APA regulation, in agreement with the similar enrichment found, generically for trQTLs, in Lappalainen et al. (2013). However, these results could also be explained, in principle, by linkage disequilibrium between promoters or enhancers and 3′ UTR regions. To evaluate the prevalence of this phenomenon, we observed that among 2,113 (3,192) significant genetic variants surviving linkage disequilibrium (LD) pruning (see *Material and Methods*) inside promoters (enhancers), only 288 (376) are in LD ($R^2 > 0.8$) with significant genetic variants within 3′ UTRs. Furthermore, the reported enrichments remained highly significant after the exclusion of these variants, supporting the idea that promoters and enhancers have an independent role in the genetic component of APA regulation.

In the following, we will divide apaQTLs in two classes: intragenic apaQTLs are those located inside one of the genes

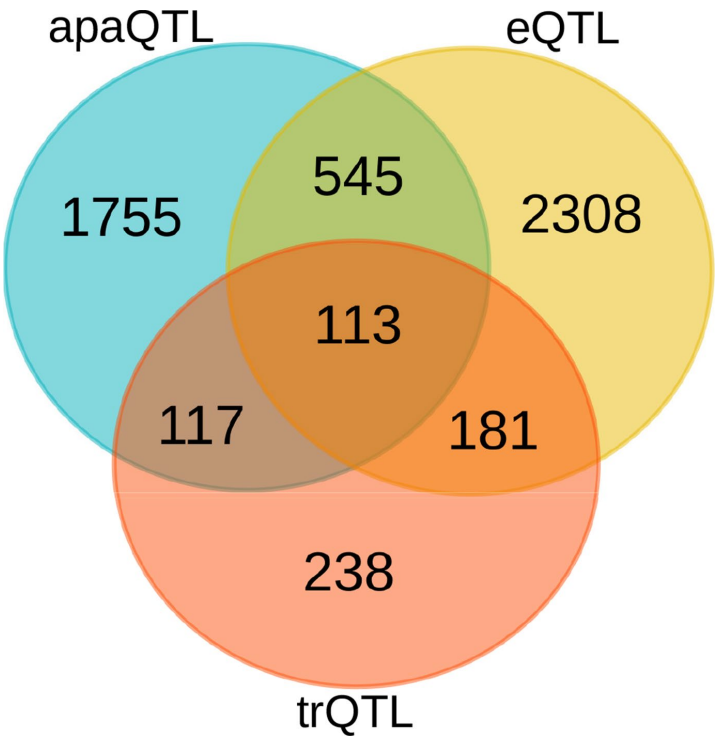

FIGURE 3 | Comparison of genes with different molecular QTLs. Overlap between genes with significant alternative polyadenylation QTL (apaQTL), expression QTL (eQTL), and transcript ratio QTL (trQTL).

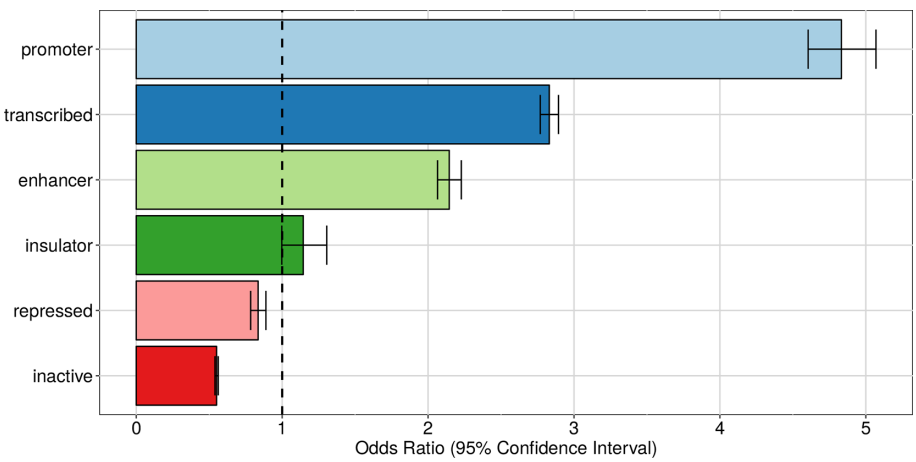

FIGURE 4 | Enrichment of apaQTLs within active genomic regions in the GM12878 cell line. For each broad state, that was defined starting from the ChromHMM annotation, the odds ratio (OR) obtained by logistic regression and its 95% CI are shown.

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 *Material and Methods*).

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 versus introns. As shown in **Figure 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 LD with variants located in the 3′ UTR, notwithstanding the LD pruning procedure implemented in the enrichment analysis (see *Material and Methods*). Note that while several poly(A) sites are located upstream of the last exon (Tian et al., 2007), 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 (Hinnebusch et al., 2016). In the following, we examine in more detail three possible mechanisms by which intragenic apaQTLs could exert their action.

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 (Gruber et al., 2016). 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 = 0.0216]. 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 (Cunninghame Graham et al., 2007; Graham et al., 2007; Yoon et al., 2012) 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 (**Figure 6**). Moreover, the same variant was associated to higher risk of systemic lupus erythematosus (SLE) and higher *IRF5* expression, which could be due to the loss of AU-rich elements (ARE) in the short transcript isoform (Yoon et al., 2012). 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 (Lazzari and Jefferies, 2014)].

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 (**Figure 7**). This variant was reported to be associated with a higher risk of spina bifida, along with other variants within the same gene (Shaw et al., 2009). 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 posttranscriptional 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 (**Figure S4**), essential in the biosynthesis of peptide hormones and neurotransmitters (Eipper et al., 1983; Czyzyk et al., 2005; Gaier et al., 2014). No eQTLs or trQTLs for this gene were revealed by the analysis of the same data reported in Lappalainen et al. (2013). 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 (Chang et al., 2017a).

Conversely, the destruction of a canonical, proximal PAS motif leads to shortening of the 3′ UTR of *BLOC1S2* (**Figure S4**). The variant rs41290536 replaces the canonical PAS motif AAUAAA with the noncanonical 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

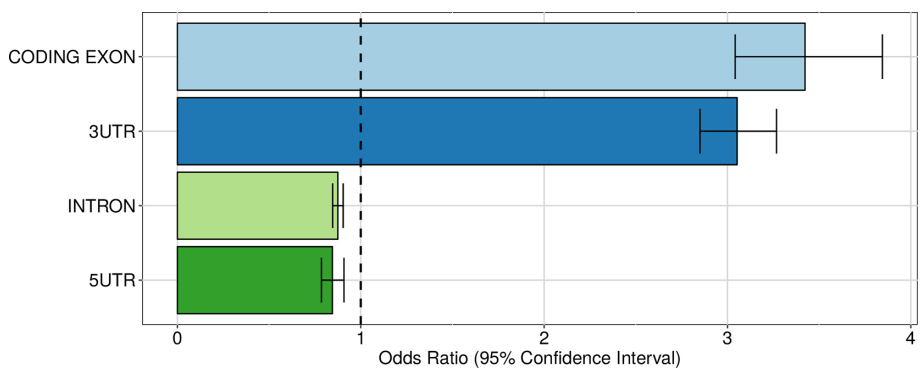

FIGURE 5 | Enrichment of intragenic apaQTLs within coding and noncoding transcript regions. For each gene region, the OR obtained by logistic regression and its 95% CI are shown.

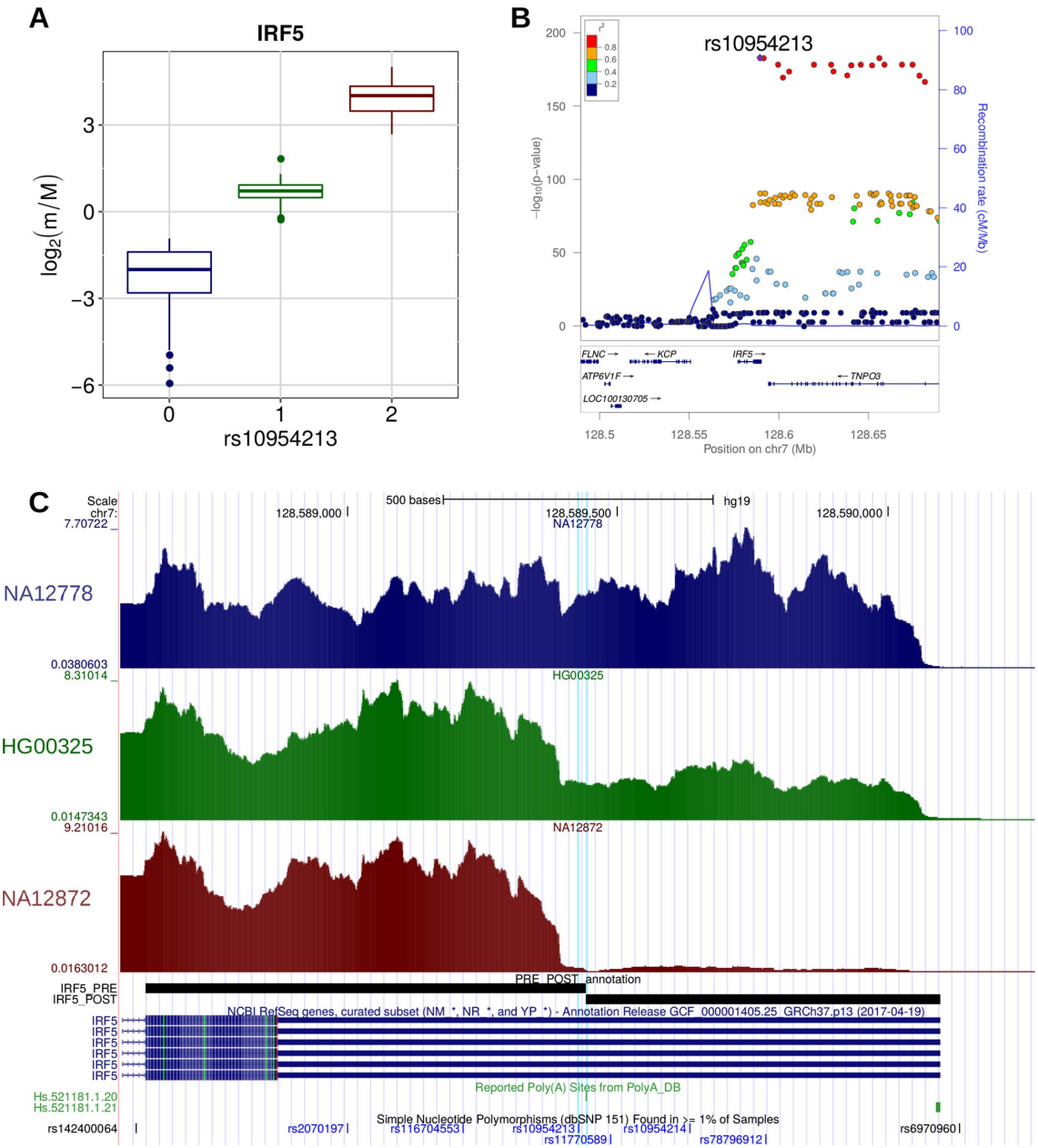

FIGURE 6 | (A) Boxplot showing the variation of the log2-transformed *m*/*M* values obtained for *IRF5* as a function of the genotype of the individuals for rs10954213. **(B)** LocusZoom plot (Pruim et al., 2010) illustrating the results obtained for *IRF5* in the genomic region around rs10954213 (100 kb 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 linkage disequilibrium (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. **(C)** Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per million mapped reads, are shown for three representative individuals: NA12778 (homozygous for the reference allele), HG00325 (heterozygous) and NA12872 (homozygous for the alternative allele). *IRF5* RefSeq, *IRF5* PRE/POST segments, poly(A) sites, and common SNPs are shown. The rs10954213 variant and the affected poly(A) site (Hs.521181.1.20) are highlighted.

variants that have been associated to predisposition to squamous cell lung carcinoma (rs28372851 and rs12765052) (McKay et al., 2017).

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, contained only in the longer isoform. Variants altering microRNA binding sites located in the POST segment

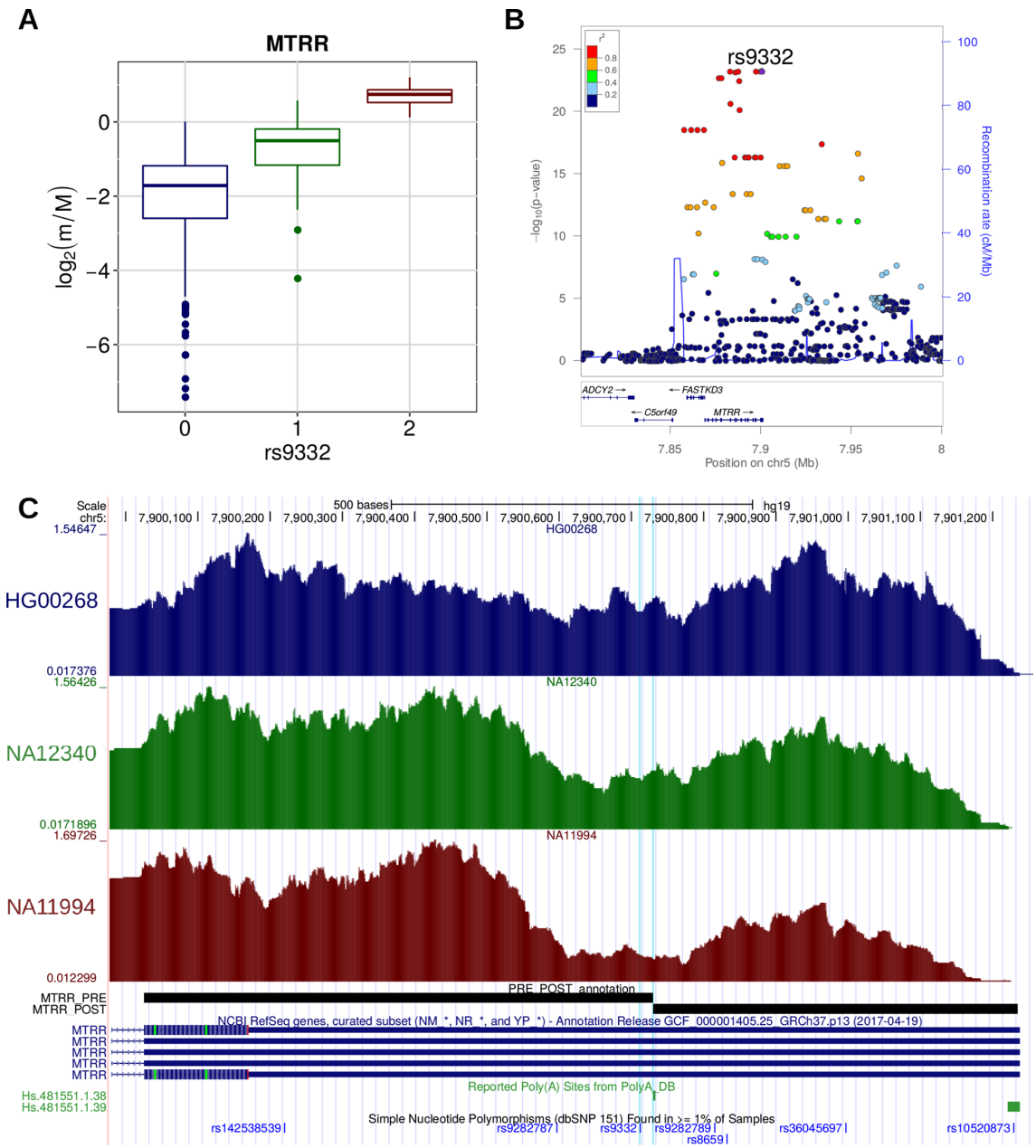

FIGURE 7 | (A) Boxplot showing the variation of the log2-transformed *m/M* values obtained for *MTRR* as a function of the genotype of the individuals for rs9332. **(B)** LocusZoom plot illustrating the results obtained for *MTRR* in the genomic region around rs9332 (100 kb both upstream and downstream its genomic location). **(C)** Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per million mapped reads, are shown for three representative individuals: HG00268 (homozygous for the reference allele), NA12340 (heterozygous), and NA11994 (homozygous for the alternative allele). *MTRR* RefSeq, *MTRR* PRE/POST segments, poly(A) sites, and common SNPs are shown. The rs9332 variant and the affected poly(A) site (Hs.481551.1.38) are highlighted.

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 (**Figure S5**). 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 = 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.

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 posttranscriptional 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 (Ray et al., 2013). Here, we exploited this resource to identify SNPs that alter putative functional RBP binding sites. Consistently with 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, $P = 8.54 \times 10^{-11}$).

Specifically, we obtained a positive and significant OR for 20 individual RBP-binding motifs (**Table S1**). 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 muscle blind protein family (Shi and Manley, 2015; Ha et al., 2018), *KHDRBS1* (La Rosa et al., 2016), and *HNRNPC* (Gruber et al., 2016)]. 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 an RBP that may be involved in RNA degradation [*CNOT4* (Miller and Reese, 2012)]. The involvement of several splicing factors is consistent with evidence supporting a mechanistic interplay between polyadenylation and splicing, which goes beyond the regulation of the usage of intronic poly(A) sites (Gunderson et al., 1994; Lutz et al., 1996; Liang and Lutz, 2006; Millevoi et al., 2006).

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 (Oktaba et al., 2015), it is commonly believed that APA is mainly controlled by cis-elements located within transcripts, both upstream and downstream of the poly(A) sites (Tian and Manley, 2017).

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 Delaneau et al. (2019). Extragenic apaQTLs were indeed found to be enriched in CRDs (OR = 1.73, 95% CI = 1.69–1.78, $P < 10^{-16}$). The 3D structure of the genome is a key aspect of gene regulation (Krijger and de Laat, 2016), as it determines physical contacts between distal regulatory regions and proximal promoters. In particular, CRDs have been described as active sub-domains within topologically associating domains (TADs), containing several noncoding 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, nontranscribed cis-regulatory regions in regulating alternative polyadenylation of the target genes.

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 (MacArthur et al., 2017) (OR = 3.17, 95% CI = 3.01–3.33, $P < 10^{-16}$).

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, $P = 2.50 \times 10^{-77}$) (**Figure 8** and **Table S2**). 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, $P = 2.47 \times 10^{-142}$) and cancer (OR = 3.96, 95% CI = 3.36–4.64, $P = 4.15 \times 10^{-63}$).

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, $P = 2.36 \times 10^{-174}$), squamous cell lung carcinoma (OR = 26.1, 95% CI = 15.7–43.3, $P = 1.29 \times 10^{-36}$), lung carcinoma (OR = 17.9, 95% CI = 12.7–25.2, $P = 9.63 \times 10^{-62}$), schizophrenia (OR = 10.6, 95% CI = 9.01–12.4, $P = 1.25 \times 10^{-182}$), and HIV-1 infection (OR = 6.51, 95% CI = 3.75–10.8, $P = 2.28 \times 10^{-12}$). The complete list of enriched traits can be found in **File S3**.

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. 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 (**Figure S6** and

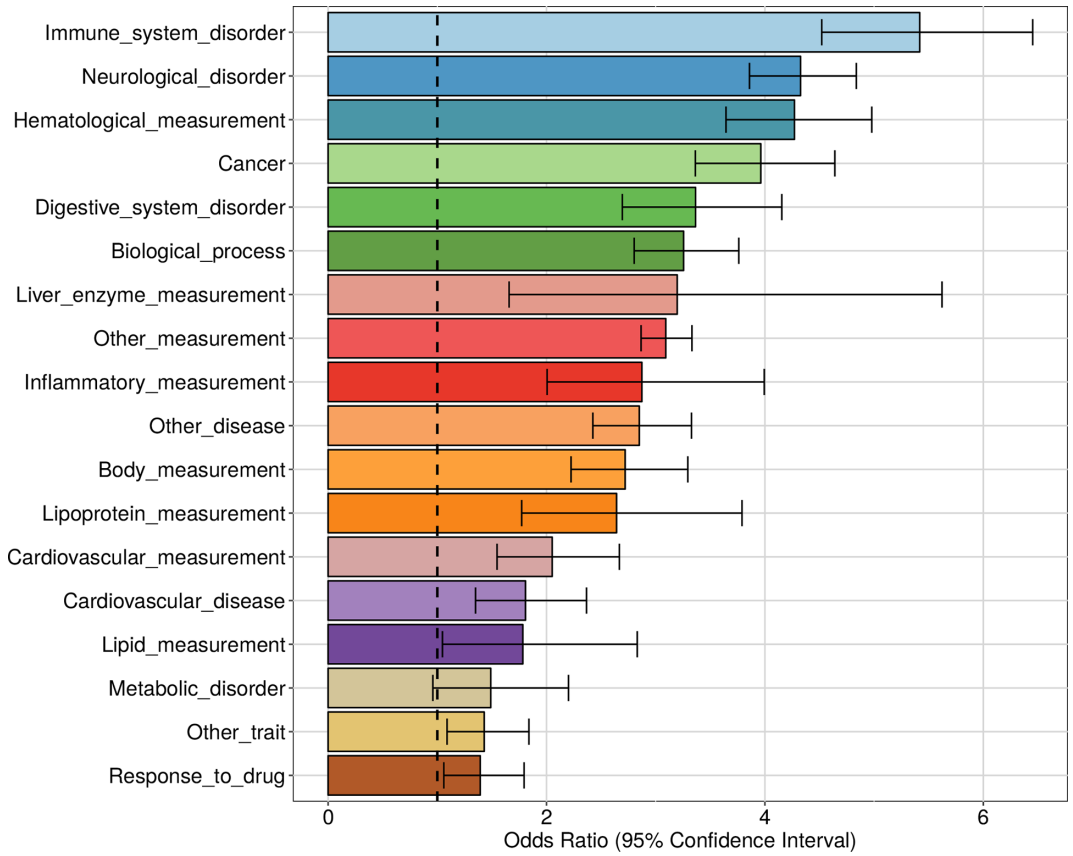

FIGURE 8 | Enrichment of genome-wide association studies (GWAS) hits among apaQTLs, for different categories of complex traits. For each category, the OR obtained by logistic regression and its 95% CI are shown.

Table S3). 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, $P = 7.05 \times 10^{-29}$). On the contrary, the enrichment of variants associated to pulmonary adenocarcinoma is driven by the HLA locus and becomes nonsignificant after excluding HLA variants (OR = 1.35, 95% CI = 0.22–4.39, $P = 0.68$). The complete list of enriched traits after the exclusion of HLA variants can be found in **File S4**.

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 (Cunninghame Graham et al., 2007; Graham et al., 2007), and the effect of the variant on *IRF5* expression in blood cells was confirmed in SLE patients (Kozyrev et al., 2007; Feng et al., 2010). However, direct evidence on the effect of this variant on APA regulation in SLE patients is still missing.

To assess whether rs10954213 affects *IRF5* APA regulation in SLE patients, we analyzed RNA-Seq data derived from whole blood cells in 99 patients (Hung et al., 2015). After the exclusion of 52 individuals whose genotype cannot be determined with certainty from RNA-Seq reads, 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 $P = 2.49 \times 10^{-8}$; **Figure 9**). 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.

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 (Kwan et al., 2008; Thomas and Sætrom, 2012; Lappalainen et al., 2013; Zhernakova et al., 2013; Monlong et al., 2014). 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 transcript structure from the same data in Lappalainen et al. (2013).

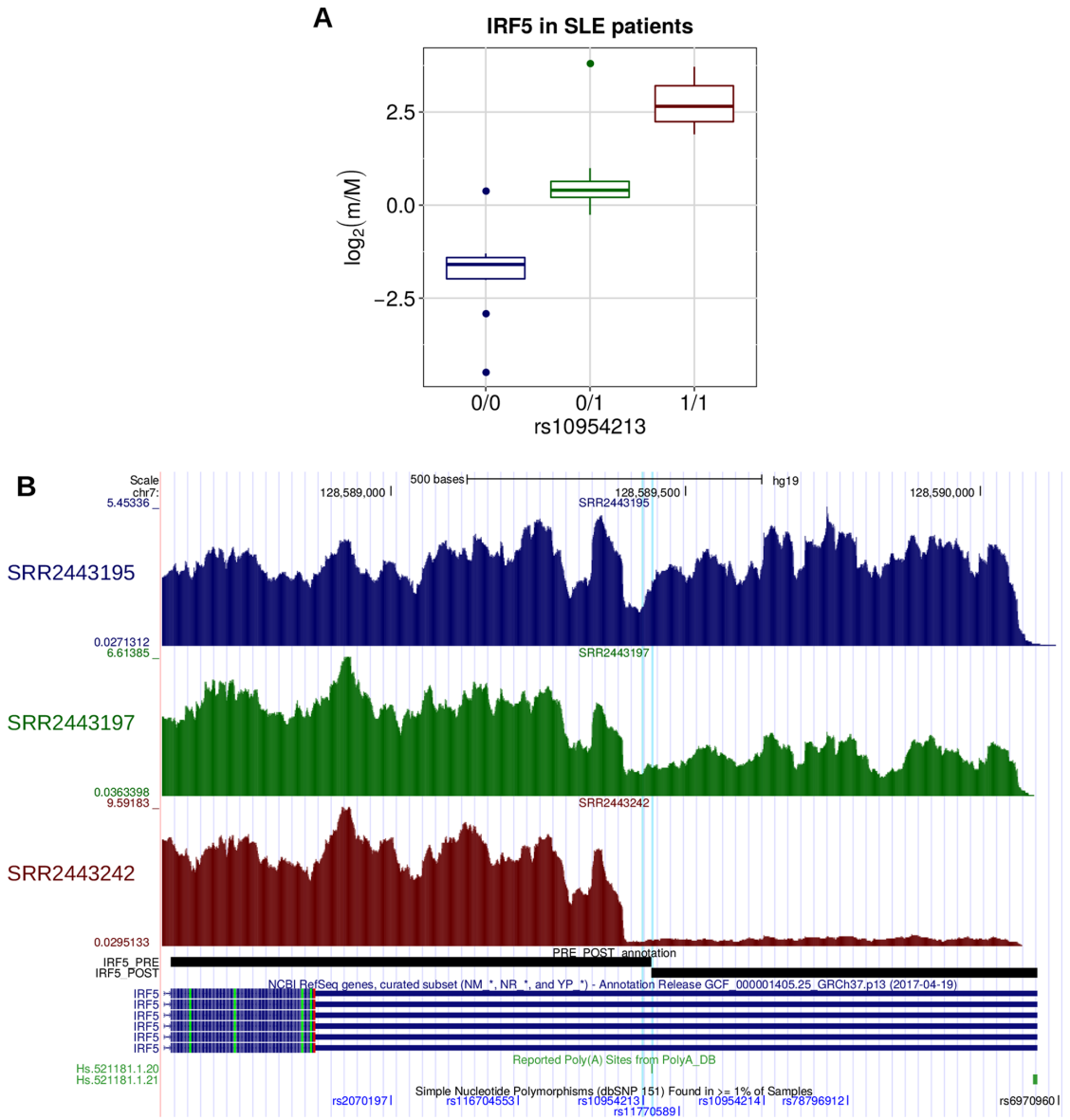

FIGURE 9 | (A) The effect of rs10954213 on the relative expression of the *IRF5* alternative isoforms was investigated also in a small cohort of systemic lupus erythematosus (SLE) patients. The boxplot shows the variation of the log2-transformed *m/M* values obtained for *IRF5* as a function of the genotype of the individuals. **(B)** Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per thousand mapped reads, are shown for three representative individuals: SRR2443195 (homozygous for the reference allele), SRR2443197 (heterozygous), and SRR2443242 (homozygous for the alternative allele). *IRF5* RefSeq, *IRF5* PRE/POST segments, poly(A) sites, and common SNPs are shown. The rs10954213 variant and the affected poly(A) site (Hs.521181.1.20) are highlighted.

Moreover, the significant overlap between our apaQTLs and the eQTLs identified in Lappalainen et al. (2013) 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 (Elkon et al., 2013; Berkovits and Mayr, 2015). Indeed, most of the apaQTLs we found were not identified as eQTLs in Lappalainen et al. (2013).

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 nontranscribed 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. In addition, 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 (Gruber et al., 2016), 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 (**Figure S4**), in agreement with previous evidence (Gieselmann et al., 1989). 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 a higher risk of both multiple sclerosis and SLE in the Sardinian population (Steri et al., 2017).

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 Lappalainen et al. (2013). 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 (Ji et al., 2011; Elkon et al., 2013). The mechanisms behind this coupling are thought to include the interaction between rates of Pol II elongation and alternative polyadenylation and the recruitment, by the transcription machinery, of trans-acting factors affecting PAS choice (Tian and Manley, 2017). Moreover, it has been shown that RBPs involved in APA regulation can interact with promoters (Oktaba et al., 2015).

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 (Võsa et al., 2015) 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 (Ha et al., 2018). 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 (Dominguez et al., 2018; Hentze et al., 2018).

Furthermore, it is reasonable to assume that also noncanonical 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 (Yue et al., 2018). 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, Wu and Bartel (2017) 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 (Tian and Manley, 2017). Therefore, it is not surprising that its perturbation has been associated with multiple pathological conditions (Chang et al., 2017b; Manning and Cooper, 2017). 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 posttranscriptional 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 (Takata et al., 2017) and Alzheimer's disease (Raj et al., 2018). 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 (Hoarau et al., 2004; Kulkarni et al., 2017), and genes encoding antigen-presenting molecules account for the highest fraction of genetic risk for many neurological diseases (Misra et al., 2018).

A gene-based alternative approach to the interpretation of GWAS has been recently proposed. In the original implementation of Transcriptome Wide Association Studies (TWAS) (Gamazon et al., 2015), 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, Gusev et al. (2016) 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 (Gusev et al., 2018) and Alzheimer's disease (Raj et al., 2018). 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 (Grassi et al., 2016). Second, our strategy depends on a preexisting 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 (Grassi et al., 2016; Ha et al., 2018). In addition, although the method is generally able to successfully discriminate APA events from alternative splicing events, it may give rise to spurious associations when intron retention is present within the 3′ UTRs, and therefore, such special cases should be inspected with particular attention. Finally, 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 (Aguet et al., 2017), 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 (Elkon et al., 2013).

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 mechanistic interpretation of genetic variants associated to phenotypic traits and diseases.

MATERIAL AND METHODS

Data Sources

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) (O'Leary et al., 2016; Casper et al., 2017). The corresponding transcript-gene map was downloaded from NCBI (version 69) and the Bioconductor R package org.Hs.eg.db v3.4.0 (Carlson, 2016) 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) (Lee et al., 2007).

ChromHMM annotations (Ernst et al., 2011) 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/ (Delaneau et al., 2019).

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 (Lappalainen et al., 2013) in the EBI ArrayExpress archive (https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/). We also downloaded genotypic data for the same individuals and the results of the eQTL/trQTL mapping analyses. 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) (The 1000 Genomes Project Consortium, 2015), but the GEUVADIS consortium additionally obtained genomic data for 41 of them through genotyping with single nucleotide polymorphism (SNP) array followed by genotype imputation (Lappalainen et al., 2013). Furthermore, whole-blood RNA-Seq data for 99 individuals affected by SLE were downloaded from the NCBI SRA database (SRP062966) (Leinonen et al., 2011; Hung et al., 2015).

Regulatory Motifs and Related Expression Data

Different collections of regulatory motifs were downloaded. A list of 18 PAS motifs was obtained from Gruber et al. (2016), microRNA seeds were downloaded from TargetScan 7.2 (Agarwal et al., 2015), and positional weight matrices (PWMs) describing the binding specificities of RNA-binding proteins were downloaded from the CISBP-RNA dataset (Ray et al., 2013), 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 and E-GEUV-1 datasets on the EBI ArrayExpress archive (https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-2/) (Lappalainen et al., 2013).

GWAS Catalog

A collection of genomic loci associated with human complex traits was obtained by downloading the NHGRI-EBI GWAS Catalog, v1.0.2 (MacArthur et al., 2017). This resource is continuously updated: the version we used was downloaded on October 10, 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 (Malone et al., 2010), including the parent category of each trait (the version of the downloaded file was r2018-09-30). In addition, the dbSNP Build 151 (Sherry et al., 1999) collection of human genetic variants was downloaded for hg19.

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 **File S5**).

The coordinates of the human poly(A) sites were converted from hg17 to hg19 using liftover (Hinrichs et al., 2006) 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 (Grassi et al., 2016). 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 contained into both the long and the short isoform, while the POST segment should be contained exclusively into the long isoform. The GTF file used for the computation of *m*/*M* values is available as **File S6**.

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.

Computation of *m*/*M* Values

Using the Bioconductor R package Roar (Grassi et al., 2016), 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 $m/M_{a,i}$ of gene *a* in the i_{th} individual is defined as

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

where l_{PRE_a} and l_{POST_a} are, respectively, the length of the PRE and POST segment of the gene *a*, and $\#r_{PRE_{a,i}}$ 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 <100 missing *m*/*M* values were selected for the following investigation, leaving us with 6,256 genes.

Genotypic Data Preprocessing

Starting from the downloaded VCF files, we extracted genotypic data for 373 EUR individuals for whom also RNA-Seq data are available using VCFtools (Danecek et al., 2011). In addition, only common genetic variants with minor allele frequency (MAF) > 5% were considered in all the following analyses. The MAF values were computed taking into account that the reference allele reported in the VCF file may not always be the most frequent one in the EUR population considered by itself, and we conservatively attributed the most frequent homozygous genotype to individuals for which the genotype was missing, thus being sure to exclude all the less frequent variants from the analysis. We are aware that these MAF values may be an underestimate of the real ones, and therefore, in all the enrichment analyses (see below for details), we instead used MAF values obtained ignoring individuals with missing data.

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 (Price et al., 2008).

We filtered out all the genetic variants falling within 24 long-range LD (LRLD) regions whose coordinates were derived from Price et al. (2008). In addition, following Novembre et al. (2008), we performed an LD pruning of the genetic variants using the –indep-pairwise function from PLINK v1.9 (Chang et al., 2015) to recursively exclude genetic variants with pairwise genotypic R^2 > 80% within sliding windows of 50 SNPs (with a 5-SNP increment between windows). Also in this case, VCFtools (Danecek et al., 2011) was used to apply all these filters to the VCF files, and finally, EIGENSTRAT v6.1.4 (Price et al., 2006) was used to run the PCA on the remaining genotypic data at the genome-wide level.

apaQTL Mapping

From a statistical point of view, we adopted the same strategy used in standard eQTL mapping analyses (Lappalainen et al., 2013) 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 TES. 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:

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

where $log_2(m/M_{a,i})$ is the log2 transformed m/M value computed for the a gene in the i_{th} individual, $g_{j,i}$ 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 performed using the CRAN R package MatrixEQTL (Shabalin, 2012). 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 (Danecek et al., 2011). Following Lappalainen et al. (2013), in all our models, we included both the imputation status of the individuals and the first three PCs obtained from genotypic data as covariates, to correct for possible biases due to population stratification (**Figure S7**) 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 (**Figure S8**). A permutation-based procedure was implemented (Churchill and Doerge, 1994): 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 performed. Finally, to control for multiple testing, the empirical P values were corrected with the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) and models with a corrected empirical $P < 0.05$ were considered statistically significant. Manhattan plots were drawn using the CRAN R package qqman (Turner, 2018).

Comparison With Other Molecular QTLs

To compare the genes for which we detected one or more apaQTLs with those for which eQTL/trQTL were reported (Lappalainen et al., 2013), we translated the Ensembl Gene IDs (ENSG) to NCBI Entrez Gene IDs using Ensembl v67 (Zerbino et al., 2018) retrieved using the Bioconductor R package biomaRt v2.30 (Durinck et al., 2005, Durinck et al., 2009). Two hundred twenty-nine ENSGs could not be translated with this procedure and were therefore excluded from this analysis.

Enrichment Analyses

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 to obtain a subset of independent candidate variants [the same strategy was implemented for example in Li et al. (2013) 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 $P < 0.05$ for at least one gene) and null variants (i.e., nominal $P > 0.1$ in all the fitted models). We excluded the "gray area" variants with nominal $P < 0.1$ but empirical corrected $P > 0.05$ 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:

$$t_j = \beta_0 + \beta_1 \times Feature_j + covariates + \epsilon_j \quad (3)$$

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

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).

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 Ernst et al. (2011), 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.

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.

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 Delaneau et al. (2019) 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).

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 polymorphisms (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 (Wu and Bartel, 2017). 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 Agarwal et al. (2015), 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 Lappalainen et al. (2013), each microRNA was considered expressed if its expression value was >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 (Grant et al., 2011) 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 >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.

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=GRCh37).

The rs10954213 Variant in SLE Patients

In the analysis of SLE patient RNA-seq data, we were interested in the *IRF5* gene only. Therefore, RNA-Seq reads were aligned to a reduced genome comprising the gene sequence and an additional 50 bp at its 3′ end using Bowtie v2.2.3 (Langmead et al., 2009) and TopHat v2.0.12 (Trapnell et al., 2009). 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. Initially, 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 a heterozygous individual. Then, genotype quality was assessed using VCFx version 1.2b (Castelli et al., 2015; Lima et al., 2016) with default parameters to filter out low-confidence genotypes. In this way, we obtained 11 homozygotes for the reference allele, 22 heterozygotes, and 14 homozygotes for the alternative allele (**Figure S9**), while 52 individuals with missing genotype information were excluded from the subsequent analysis. Notably, the genotypes are in Hardy–Weinberg equilibrium (chi-squared $P = 0.705$). A Kruskal–Wallis test was then used to evaluate the differences in m/M values between genotypes.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ebi.ac.uk/Tools/geuvadis-das/

AUTHOR CONTRIBUTIONS

EM: conceived and planned the work, performed most computational analyses, wrote the manuscript, approved the final version; FM: performed computational analysis, approved the final version; EG: performed computational analysis, approved the final version; SG: performed computational analysis, approved the final version; PP: conceived and planned the work, wrote the manuscript, approved the final version.

FUNDING

This work was supported by a grant from Compagnia di San Paolo, Turin, Italy [grant Torino_call_L2_252].

ACKNOWLEDGMENTS

This manuscript has been released as a Pre-Print at bioRxiv bioarxiv (Mariella et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00714/full#supplementary-material

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, e05005. doi: 10.7554/eLife.05005

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

An, J. J., Gharami, K., Liao, G.-Y., Woo, N. H., Lau, A. G., Vanevski, F., et al. (2008). Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* 134, 175–187. doi: 10.1016/j.cell.2008.05.045

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis:

multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodological)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berkovits, B. D., and Mayr, C. (2015). Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* 522, 363–367. doi: 10.1038/nature14321

Carlson, M. (2016). [Dataset] org.Hs.eg.db: Genome wide annotation for Human.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., et al. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* 46, D762–D769. doi: 10.1093/nar/gkx1020

Castelli, E. C., Mendes-Junior, C. T., Sabbagh, A., Porto, I. O., Garcia, A., Ramalho, J., et al. (2015). HLA-E coding and 3′ untranslated region variability determined by next-generation sequencing in two West-African population samples. *Hum. Immunol.* 76, 945–953. doi: 10.1016/j.humimm.2015.06.016

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Chang, D., Nalls, M. A., Hallgrímsdóttir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., et al. (2017a). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* 49, 1511–1516. doi: 10.1038/ng.3955

Chang, J. W., Yeh, H. S., and Yong, J. (2017b). Alternative polyadenylation in human diseases. *Endocrinol. Metab. (Seoul)* 32, 413–421. doi: 10.3803/EnM.2017.32.4.413

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194. doi: 10.1038/nrg2537

Cunninghame Graham, D. S., Manku, H., Wagner, S., Reid, J., Timms, K., Gutin, A., et al. (2007). Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.* 16, 579–591. doi: 10.1093/hmg/ddl469

Czyzyk, T. A., Ning, Y., Hsu, M.-S., Peng, B., Mains, R. E., Eipper, B. A., et al. (2005). Deletion of peptide amidation enzymatic activity leads to edema and embryonic lethality in the mouse. *Dev. Biol.* 287, 301–313. doi: 10.1016/j.ydbio.2005.09.001

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., et al. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* 364, eaat8266. doi: 10.1126/science.aat8266

Dominguez, D., Freese, P., Alexis, M. S., Yeo, G. W., Graveley, B. R., and Burge, C. B. (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* 70, 854–867.e9. doi: 10.1016/j.molcel.2018.05.001

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Eipper, B. A., Glembotski, C. C., and Mains, R. E. (1983). Bovine intermediate pituitary alpha-amidation enzyme: preliminary characterization. *Peptides* 4, 921–928. doi: 10.1016/0196-9781(83)90091-8

Elkon, R., Ugalde, A. P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* 14, 496–506. doi: 10.1038/nrg3482

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49. doi: 10.1038/nature09906

Feng, D., Stone, R. C., Eloranta, M.-L., Sangster-Guity, N., Nordmark, G., Sigurdsson, S., et al. (2010). Genetic variants and disease-associated factors contribute to enhanced IRF-5 expression in blood cells of systemic lupus erythematosus patients. *Arthritis Rheum.* 62, 562–573. doi: 10.1002/art.27223

Ferreira, P. G., Oti, M., Barann, M., Wieland, T., Ezquina, S., Friedländer, M. R., et al. (2016). Sequence variation between 462 human individuals fine-tunes functional sites of RNA processing. *Sci. Rep.* 6, 32406. doi: 10.1038/srep32406

Floor, S. N., and Doudna, J. A. (2016). Tunable protein synthesis by transcript isoforms in human cells. *Elife* 5, e10921. doi: 10.7554/eLife.10921

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., et al. (2011). Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 21, 741–747. doi: 10.1101/gr.115295.110

Gaier, E. D., Kleppinger, A., Ralle, M., Covault, J., Mains, R. E., Kenny, A. M., et al. (2014). Genetic determinants of amidating enzyme activity and its relationship with metal cofactors in human serum. *BMC Endocr. Disord.* 14, 58. doi: 10.1186/1472-6823-14-58

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Gieselmann, V., Polten, A., Kreysing, J., and von Figura, K. (1989). Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site. *Proc. Natl. Acad. Sci. U. S. A.* 86, 9436–9440. doi: 10.1073/pnas.86.23.9436

Graham, R. R., Kyogoku, C., Sigurdsson, S., Vlasova, I. A., Davies, L. R. L., Baechler, E. C., et al. (2007). Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U. S. A.* 104, 6758–6763. doi: 10.1073/pnas.0701266104

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Grassi, E., Mariella, E., Lembo, A., Molineris, I., and Provero, P. (2016). Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinform.* 17, 423. doi: 10.1186/s12859-016-1254-8

Gruber, A. J., Schmidt, R., Gruber, A. R., Martin, G., Ghosh, S., Belmadani, M., et al. (2016). 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. *Genome Res.* 26, 1145–1159. doi: 10.1101/gr.202432.115

Gunderson, S. I., Beyer, K., Martin, G., Keller, W., Boelens, W. C., and Mattaj, L. W. (1994). The human U1A snRNP protein regulates polyadenylation *via a* direct interaction with poly(A) polymerase. *Cell* 76, 531–541. doi: 10.1016/0092-8674(94)90116-3

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548. doi: 10.1038/s41588-018-0092-1

Ha, K. C. H., Blencowe, B. J., and Morris, Q. (2018). QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.* 19, 45. doi: 10.1186/s13059-018-1414-4

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* 19, 327–341. doi: 10.1038/nrm.2017.130

Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. (2016). Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* 352, 1413–1416. doi: 10.1126/science.aad9868

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., et al. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34, D590–D598. doi: 10.1093/nar/gkj144

Hoarau, J.-J., Cesari, M., Caillens, H., Cadet, F., and Pabion, M. (2004). HLA DQA1 genes generate multiple transcripts by alternative splicing and polyadenylation of the 3′ untranslated region. *Tissue Antigens* 63, 58–71. doi: 10.1111/j.1399-0039.2004.00140.x

Hung, T., Pratt, G. A., Sundararaman, B., Townsend, M. J., Chaivorapol, C., Bhangale, T., et al. (2015). The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science* 350, 455–459. doi: 10.1126/science.aac7442

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* 106, 7028–7033. doi: 10.1073/pnas.0900028106

Ji, Z., Luo, W., Li, W., Hoque, M., Pan, Z., Zhao, Y., et al. (2011). Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* 7, 534. doi: 10.1038/msb.2011.6

Kozyrev, S. V., Lewén, S., Reddy, P. M. V. L., Pons-Estel, B., Witte, T., Junker, P., et al. (2007). Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus. *Arthritis Rheum.* 56, 1234–1241. doi: 10.1002/art.22497

Krijger, P. H. L., and de Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782. doi: 10.1038/nrm.2016.138

Kulkarni, S., Ramsuran, V., Rucevic, M., Singh, S., Lied, A., Kulkarni, V., et al. (2017). Posttranscriptional regulation of HLA-A protein expression by alternative polyadenylation signals involving the RNA-binding protein syncrip. *J. Immunol.* 199, 3892–3899. doi: 10.4049/jimmunol.1700697

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., et al. (2008). Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* 40, 225–231. doi: 10.1038/ng.2007.57

La Rosa, P., Bielli, P., Compagnucci, C., Cesari, E., Volpe, E., Farioli Vecchioli, S., et al. (2016). Sam68 promotes self-renewal and glycolytic metabolism in mouse neural progenitor cells by modulating Aldh1a3 pre-mRNA 3′-end processing. *Elife* 5, e20750. doi: 10.7554/eLife.20750

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi: 10.1186/gb-2009-10-3-r25

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Lazzari, E., and Jefferies, C. A. (2014). IRF5-mediated signaling and implications for SLE. *Clin. Immunol.* 153, 343–352. doi: 10.1016/j.clim.2014.06.001

Lee, J. Y., Yeh, I., Park, J. Y., and Tian, B. (2007). PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.* 35, D165–D168. doi: 10.1093/nar/gkl870

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Lembo, A., Di Cunto, F., and Provero, P. (2012). Shortening of 3′ UTRs correlates with poor prognosis in breast and lung cancer. *PLoS ONE* 7, e31129. doi: 10.1371/journal.pone.0031129

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., et al. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet.* 4, 103. doi: 10.3389/fgene.2013.00103

Liang, S., and Lutz, C. (2006). p54nrb is a component of the snRNP-free U1A (SF- A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA* 12, 111–121. doi: 10.1261/rna.2213506

Lima, T. H. A., Buttura, R. V., Donadi, E. A., Veiga-Castelli, L. C., Mendes-Junior, C. T., and Castelli, E. C. (2016). HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. *Hum. Immunol.* 77, 841–853. doi: 10.1016/j.humimm.2016.07.231

Lutz, C. S., Murthy, K. G., Schek, N., O'Connor, J. P., Manley, J. L., and Alwine, J. C. (1996). Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency *in vitro*. *Genes Dev.* 10, 325–337. doi: 10.1101/gad.10.3.325

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., et al. (2010). Modeling sample variables with an experimental factor ontology. *Bioinformatics* 26, 1112–1118. doi: 10.1093/bioinformatics/btq099

Manning, K. S., and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nat. Rev. Mol. Cell Biol.* 18, 102–114. doi: 10.1038/nrm.2016.139

Mariella, E., Marotta, F., Grassi, E., Gilotto, S., and Provero, P. (2019). The length of the expressed 3′ UTR is an intermediate molecular phenotype linking genetic variants to complex diseases. *bioRxiv* 540088. doi: 10.1101/540088

Masamha, C. P., Xia, Z., Yang, J., Albrecht, T. R., Li, M., Shyu, A.-B., et al. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* 510, 412–416. doi: 10.1038/nature13261

Mayr, C. (2018). What are 3′ UTRs doing? *Cold Spring Harb. Perspect. Biol.* a034728. doi: 10.1101/cshperspect.a034728

Mayr, C., and Bartel, D. P. (2009). Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673–684. doi: 10.1016/j.cell.2009.06.016

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* 49, 1126–1132. doi: 10.1038/ng.3892

Miller, J. E., and Reese, J. C. (2012). Ccr4-Not complex: the control freak of eukaryotic cells. *Crit. Rev. Biochem. Mol. Biol.* 47, 315–333. doi: 10.3109/10409238.2012.667214

Millevoi, S., Decorsiere, A., Loulergue, C., Iacovoni, J., Bernat, S., Antoniou, M., et al. (2009). A physical and functional link between splicing factors promotes pre-mRNA 3′ end processing. *Nucleic Acids Res.* 37, 4672–4683. doi: 10.1093/nar/gkp470

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S. Z., Keller, W., Antoniou, M., et al. (2006). An interaction between U2AF 65 and CF Im links the splicing and 3′ end processing machineries. *EMBO J.* 25, 4854–4864. doi: 10.1038/sj.emboj.7601331

Misra, M. K., Damotte, V., and Hollenbach, J. A. (2018). The immunogenetics of neurological disease. *Immunology* 153, 399–414. doi: 10.1111/imm.12869

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* 5, 4698. doi: 10.1038/ncomms5698

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., et al. (2008). Genes mirror geography within Europe. *Nature* 456, 98–101. doi: 10.1038/nature07331

Oktaba, K., Zhang, W., Lotz, T. S., Jun, D. J., Lemke, S. B., Ng, S. P., et al. (2015). ELAV links paused Pol II to alternative polyadenylation in the *Drosophila* nervous system. *Mol. Cell* 57, 341–348. doi: 10.1016/j.molcel.2014.11.024

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., et al. (2008). Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* 83, 132–135. doi: 10.1016/j.ajhg.2008.06.005

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337. doi: 10.1093/bioinformatics/btq419

Raj, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* 50, 1584–1592. doi: 10.1038/s41588-018-0238-1

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177. doi: 10.1038/nature12311

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647. doi: 10.1126/science.1155390

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis *via* large matrix operations. *Bioinformatics* 28, 1353–1358. doi: 10.1093/bioinformatics/bts163

Shaw, G. M., Lu, W., Zhu, H., Yang, W., Briggs, F. B., Carmichael, S. L., et al. (2009). 118 SNPs of folate-related genes and risks of spina bifida and conotruncal heart defects. *BMC Med. Genet.* 10, 49. doi: 10.1186/1471-2350-10-49

Sherry, S. T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* 9, 677–679. doi: 10.1101/GR.9.8.677

Shi, Y., and Manley, J. L. (2015). The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. *Genes Dev.* 29, 889–897. doi: 10.1101/gad.261974.115

Spies, N., Burge, C. B., and Bartel, D. P. (2013). 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 23, 2078–2090. doi: 10.1101/gr.156919.113

Steri, M., Orrù, V., Idda, M. L., Pitzalis, M., Pala, M., Zara, I., et al. (2017). Overexpression of the cytokine BAFF and autoimmunity risk. *N. Engl. J. Med.* 376, 1615–1626. doi: 10.1056/NEJMoa1610528

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* 8, 14519. doi: 10.1038/ncomms14519

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Thomas, L. F., and Sætrom, P. (2012). Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. *PLoS Comput. Biol.* 8, e1002621. doi: 10.1371/journal.pcbi.1002621

Tian, B., and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* 18, 18–30. doi: 10.1038/nrm.2016.116

Tian, B., Pan, Z., and Lee, J. Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 17, 156–165. doi: 10.1101/gr.5532707

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q–Q and manhattan plots. *J. Open Source Softw.* 3, 731. doi: 10.21105/joss.00731

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi: 10.1126/science.1058040

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Võsa, U., Esko, T., Kasela, S., and Annilo, T. (2015). Altered gene expression associated with microRNA binding site polymorphisms. *PLOS ONE* 10, e0141351. doi: 10.1371/journal.pone.0141351

Wu, X., and Bartel, D. P. (2017). Widespread Influence of 3'-end structures on mammalian mRNA processing and stability. *Cell* 169, 905–917.e11. doi: 10.1016/j.cell.2017.04.036

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 1254806. doi: 10.1126/science.1254806

Yoon, O. K., Hsu, T. Y., Im, J. H., and Brem, R. B. (2012). Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* 8, e1002882. doi: 10.1371/journal.pgen.1002882

You, L., Wu, J., Feng, Y., Fu, Y., Guo, Y., Long, L., et al. (2015). APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* 43, D59–D67. doi: 10.1093/nar/gku1076

Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., et al. (2018). VIRMA mediates preferential m6A mRNA methylation in 3′ UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov.* 4, 10. doi: 10.1038/s41421-018-0019-0

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098

Zhernakova, D. V., de Klerk, E., Westra, H.-J., Mastrokolias, A., Amini, S., Ariyurek, Y., et al. (2013). DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* 9, e1003594. doi: 10.1371/journal.pgen.1003594

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.