nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00864-5

An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability

Lei Li[1,7], Kai-Lieh Huang[2,7], Yipeng Gao[3], Ya Cui[1], Gao Wang[4], Nathan D. Elrod[2], Yumei Li[1], Yiling Elaine Chen[5], Ping Ji[2], Fanglue Peng[6], William K. Russell[2], Eric J. Wagner[2] ✉ and Wei Li[1] ✉

Genome-wide association studies have identified thousands of noncoding variants associated with human traits and diseases. However, the functional interpretation of these variants is a major challenge. Here, we constructed a multi-tissue atlas of human 3′UTR alternative polyadenylation (APA) quantitative trait loci (3′aQTLs), containing approximately 0.4 million common genetic variants associated with the APA of target genes, identified in 46 tissues isolated from 467 individuals (Genotype-Tissue Expression Project). Mechanistically, 3′aQTLs can alter poly(A) motifs, RNA secondary structure and RNA-binding protein-binding sites, leading to thousands of APA changes. Our CRISPR-based experiments indicate that such 3′aQTLs can alter APA regulation. Furthermore, we demonstrate that mapping 3′aQTLs can identify APA regulators, such as La-related protein 4. Finally, 3′aQTLs are colocalized with approximately 16.1% of trait-associated variants and are largely distinct from other QTLs, such as expression QTLs. Together, our findings show that 3′aQTLs contribute substantially to the molecular mechanisms underlying human complex traits and diseases.

Genome-wide association studies (GWAS) have identified thousands of genetic variants that have been associated with quantitative traits and common diseases. However, the vast majority of variants occur in noncoding regions, resulting in significant challenges when attempting to elucidate the molecular mechanisms through which these variants contribute to diseases and phenotypes. To provide functional interpretations of GWAS loci, researchers have suggested employing several molecular QTL analyses, including expression QTLs (eQTLs)[1], which are genetic variants associated with the expression of one or more genes. Although these genetic variants can be informative and, in many cases, are thought to impact the transcription of nearby genes, the roles played by a large fraction of trait-associated noncoding variants is unexplained[2].

APA plays an important role during the posttranscriptional regulation of most human genes. By employing different polyadenylation (poly(A)) sites, genes can either shorten or extend 3′UTRs that contain *cis*-regulatory elements, such as microRNAs (miRNA) or RNA-binding protein (RBP) binding sites[3]. Therefore, APA can affect the stability and translation efficiency of target messenger RNA and the cellular localization of proteins[4]. The diverse landscape of poly(A) sites can substantially impact both normal development and the progression of diseases, such as cancer[5]. The broad importance of alternative polyadenylation is well exemplified by the altered expression of NUDT21, a key APA regulator, in diseases such as glioblastoma[6] and idiopathic pulmonary fibrosis[7]. More recently, our work revealed a more nuanced interpretation of APA since 3′UTR shortening in breast cancer represses tumor suppressor genes in *trans* by disrupting competing endogenous RNA crosstalk[8].

In addition to being associated with gene expression, genetic variations have been identified as critical regulatory factors for the APA of individual genes in certain cell lines[9,10]. Moreover, APA-associated genetic changes have been linked to the development of multiple disease states, including cancer[11], α-thalassemia[12], facioscapulohumeral muscular dystrophy[13], bone fragility[14], neonatal diabetes[15] and systemic lupus erythematosus[16,17]. As a prime example of these studies, one SNP (rs10954213) within the 3′UTR of *IRF5* can alter the 3′UTR length and affect mRNA stability[17], which can further contribute to systemic lupus erythematosus susceptibility. Aside from these few isolated examples, the broad implications of genetic determinants impacting APA in various human tissues and their association with phenotypic traits and diseases have not been systematically examined.

Previous studies identified APA-associated SNPs using 3′-end profiling methods, which have not been widely adopted; thus, these methods have only been applied to small sample sizes[9,18]. In contrast, RNA sequencing (RNA-seq) has been extensively used during eQTL studies; however, only a few RNA-seq data have been analyzed in a manner that would systematically identify and quantify APA events[19]. To obtain an insight into the genetic basis of APA regulation in human tissues, we used our dynamic analyses of APA from RNA-seq (DaPars) algorithm[20] to construct an atlas of tissue-specific, human APA events, using 8,277 RNA-seq datasets coupled with whole-genome sequencing genotype data derived from 46 tissues and isolated from 467 individuals by the Genotype-Tissue Expression Project (GTEx)[1]. In total, we identified 403,215 common *cis*-acting genetic variants associated with APA (3′aVariants), which were colocalized with 16.1% of trait-associated

[1]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA, USA. [2]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [3]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA. [4]The Gertrude H. Sergievsky Center and Department of Neurology, Columbia University, New York, NY, USA. [5]Department of Statistics, University of California, Los Angeles, CA, USA. [6]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [7]These authors contributed equally: Lei Li, Kai-Lieh Huang. ✉e-mail: ejwagner@utmb.edu; wei.li@uci.edu

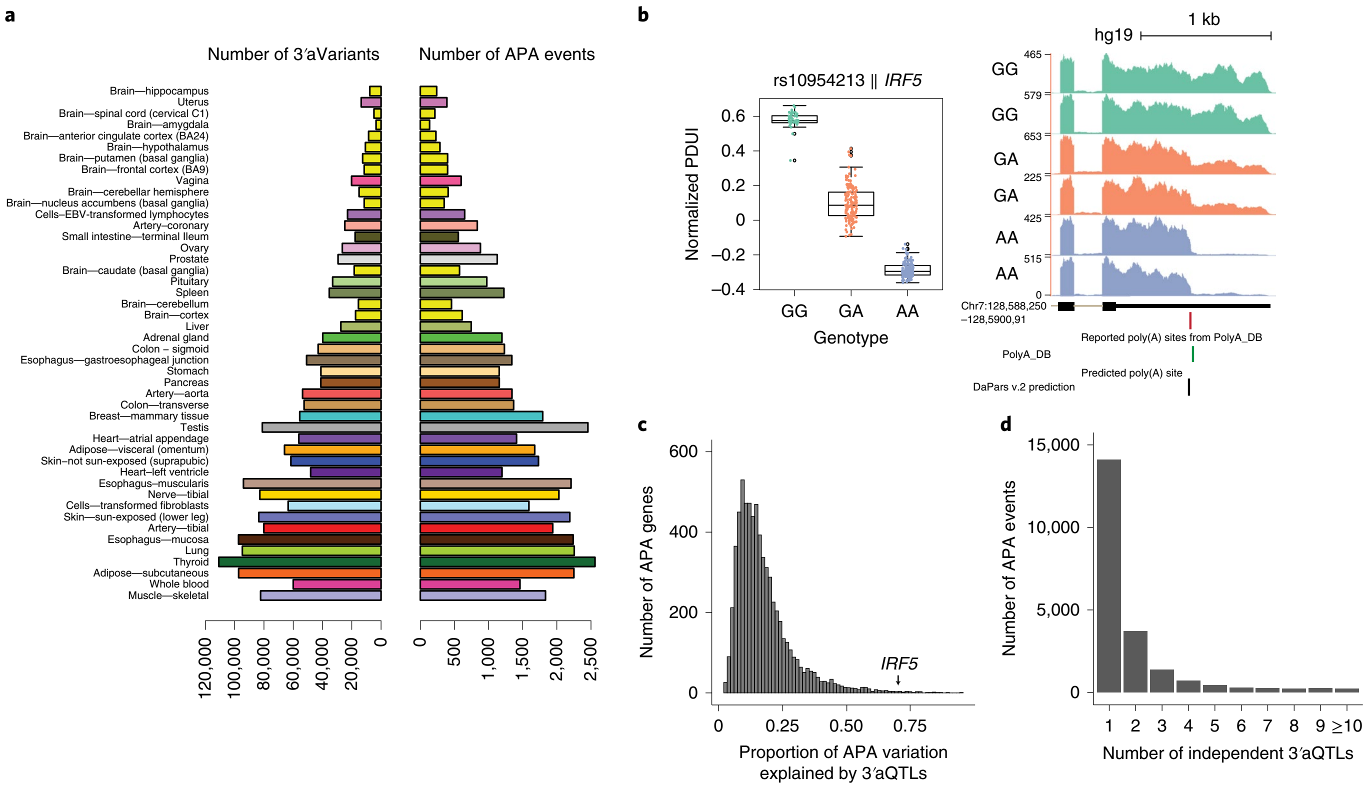


Fig. 1 | Atlas of genetic variations associated with 3′UTR usage across 46 human tissues. a, Distribution of the number of APA events and significant 3′aVariants (FDR ≤ 0.05) for each tissue, sorted by the tissue sample sizes. Each color code indicates a tissue of origin. **b**, Example of a 3′aVariant (rs10954213) that is strongly associated with the *IRF5* 3′UTR usage in whole blood. Left: Distribution of the normalized PDUI for each genotype. Each dot in the box plot represents the normalized PDUI value for one particular sample ($n = 396$). The center horizontal lines within the plot represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× interquartile range (IQR) (bottom). Right: RNA-seq coverage track for the *IRF5* 3′UTR. The bottom four tracks show the RefSeq gene structure, 3′aVariant location, reported poly(A) site location and DaPars v.2 prediction. **c**, Average fraction of APA variations that can be explained by 3′aQTLs for each transcript. The *y* axis represents the total transcripts across all human tissues studied. **d**, The distribution of independent 3′aQTLs across tissues.

variants in at least 1 tissue. Collectively, the results of our study indicated that 3′aQTLs reveal the genetic architecture of an emerging molecular phenotype and can be used to interpret a significant portion of the human genetic variants found outside of coding regions.

Results

An atlas of human 3′aQTLs. To detect global APA events in primary human tissues, we used our DaPars v.2.0 algorithm to identify APA events retrospectively and directly using 8,277 standard RNA-seq samples in 46 tissue types from the GTEx v.7 project. The multi-sample DaPars v.2 regression framework calculates a percentage of distal poly(A) site usage index (PDUI) value for each gene in each sample (Supplementary Fig. 1). The PDUI values can then be normalized further after corrections for known covariates including sex, sequencing platform, population structure, RNA integrity number and inferred technical covariates using probabilistic estimation of expression residual (PEER) factors[21]. The inferred PEER factors were strongly associated with several known covariates for each sample and donor (Extended Data Figs. 1–3). We then used Matrix eQTL to identify common genetic variations associated with differential 3′UTR usage (3′aQTLs) in each tissue[22] (Methods). Genes with a 3′aQTL are called 3′aGenes and the corresponding significant variants are called 3′aVariants. Using a false discovery rate (FDR) threshold of 5%, we identified 403,215 3′aVariants associated with 11,613 3′aGenes across 46 tissues, representing approximately 51% of annotated genes (Fig. 1a). Across all tissues, we discovered 56.7% of protein-coding and 26.1% of long noncoding RNA genes detected in at least 1 tissue (Supplementary Fig. 2). The tissues with the highest numbers of 3′aQTLs tended to have larger sample sizes (Supplementary Table 1). This strong association between 3′aQTL number and sample size suggests that additional APA events and 3′aQTLs will continue to be discovered as additional RNA-seq datasets become available. In addition, our global analysis of recent saturation mutagenesis data[23] showed that 3′aQTLs are more enriched in the variants that lead to more notable APA changes (Extended Data Fig. 4).

To evaluate the performance of our 3′aQTL detection method using the current sample size, we compared the detected 3′aQTLs with previously reported SNPs that have been associated with variations in 3′UTR usage. Although previous studies of APA events have been limited to a few cell types, such as lymphoblastoid cells, our approach recaptured many of these 'experimentally validated' 3′aQTLs. For example, the strong association between the SNP rs10954213 and the alternative 3′UTR of *IRF5* (ref. [17]), which encodes a transcription factor involved in multiple immune processes, was replicated in a whole-blood 3′aQTL analysis (Fig. 1b). Interestingly, we also found that this genetic effect on *IRF5* was shared in 22 other tissues, suggesting that the multi-tissue context analysis of this locus could aid further investigations into how *IRF5* variants contribute to autoimmune diseases (Supplementary Fig. 3). Of the 15 previously reported SNP-associated APA genes that were identified in lymphoblastoid cell lines[9,10,24–26], our 3′aQTL analysis was able to recapture 13 (87%) in Epstein–Barr virus (EBV)-transformed lymphocytes (Supplementary Fig. 4). This observation indicated that

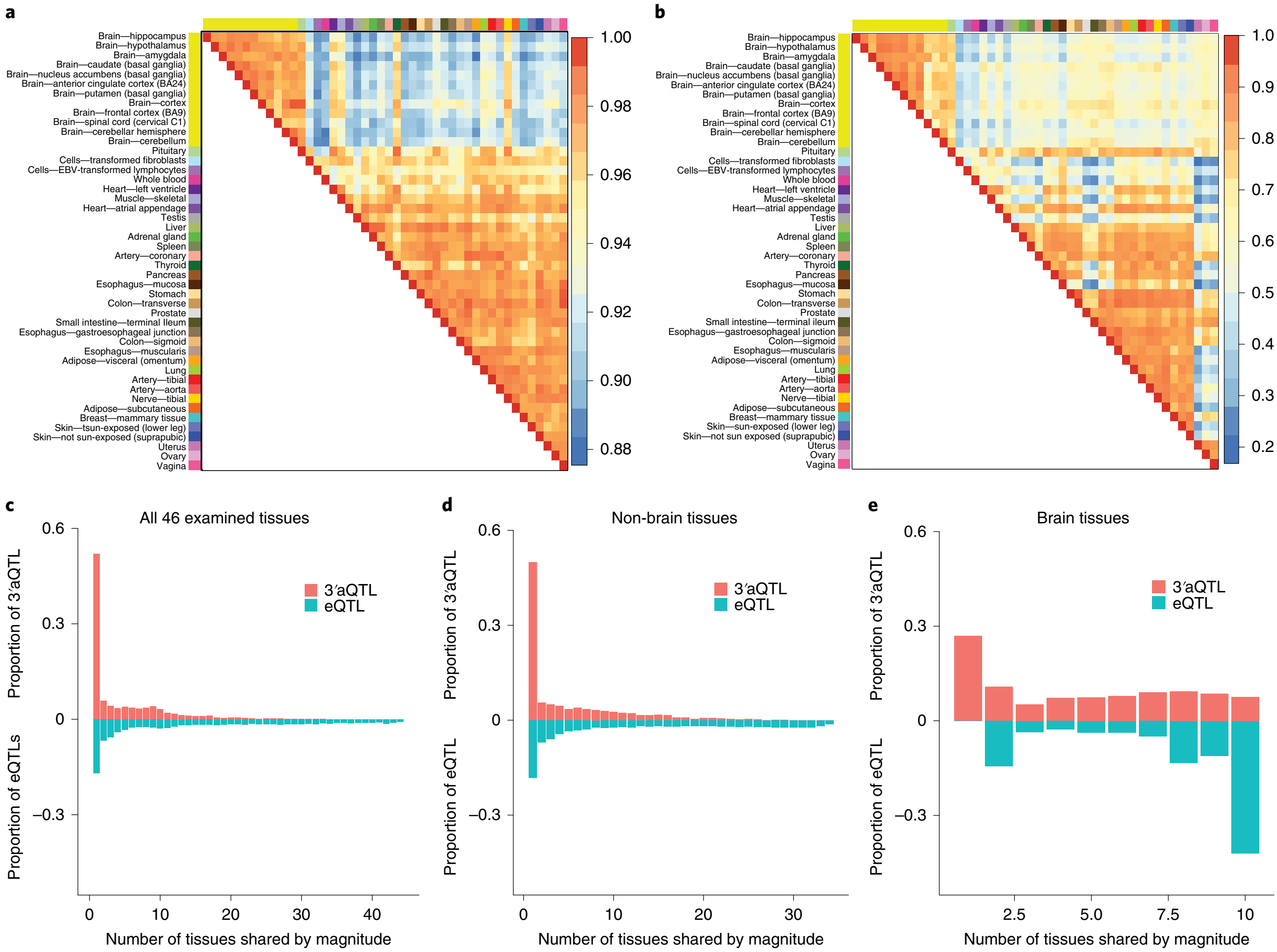


Fig. 2 | Tissue-specific 3′aQTLs. a, Pairwise 3′aQTL sharing by sign among tissues: for each pair of tissues, the proportion of shared lead 3′aQTLs, with the same direction of effect, was calculated. **b**, Pairwise 3′aQTL sharing by magnitude among tissues: for each pair of tissues, the proportion of shared lead 3′aQTLs, with the same direction of effect and within a twofold effect size, was calculated. **c–e**, Histograms showing the estimated proportion of tissues that shared lead 3′aQTLs/eQTLs, by magnitude, with other tissues, among all 46 examined tissues (**c**), among non-brain tissues only (**d**) and among brain tissues only (**e**).

the currently available datasets can be used to capture most of the known APA-associated SNPs in human tissues.

To investigate the global distribution of 3′aQTLs across the human genome, we used Manhattan plots to visualize the locations of 3′aQTLs, with their associated *P* values (Supplementary Fig. 5a). Significant 3′aQTLs were distributed across each chromosome. Importantly, previously reported APA genes were readily detected, including *IRF5* (ref. [17]), *ERAP1* (ref. [10]), *THEM4* (ref. [10]), *EIF2A* and *DIP2B*[9]; however, most of the detected 3′aQTL genes represented, to the best of our knowledge, new events. Several of these new 3′aQTL genes are particularly noteworthy, including *CHURC1* (Supplementary Fig. 5b), which encodes a zinc-finger transcriptional activator that is important during neuronal development[27], and *TPSAB1* (Supplementary Fig. 5c), which encodes α-tryptase and reportedly plays a role in multisystem disorders, such as irritable bowel syndrome, caused by elevated basal serum tryptase levels[28].

We applied heritability estimation and genetic fine-mapping to elucidate the genetic architecture of APA gene variations caused by 3′aQTLs. Specifically, we used a linear mixed model in the genome-wide complex trait analysis genome-based restricted maximum likelihood program[29] to estimate the heritability of the APA variations contributed by all 3′aVariants in each gene, within the 1-megabase (Mb) *cis* region. We observed that 3′aQTLs can explain, on average, 25.2% of APA variations (Fig. 1c). At the individual tissue level, 3′aQTLs can explain between 15.5 and 51.2% of APA variations (Supplementary Table 2). Furthermore, 3′aQTLs could explain >50% of APA variations in 2.2% of APA genes, which are enriched in antigen processing and response to interferon-γ (IFN-γ)-mediated signal pathways (Supplementary Fig. 6). For example, 72.7% of the *IRF5* APA variations can be explained by 3′aQTLs. We also found that 3′aQTLs can explain, on average, 16.2% of APA gene expression changes (Supplementary Fig. 7). To account for correlations among the identified 3′aQTLs, due to linkage disequilibrium (LD), we used sum of single effects (SuSiE) regression[30] to fine-map independent associations (summarized as 95% single-effect credible sets) for each APA transcript in each tissue. SuSiE produces clusters of association signals and each signal is designed to capture exactly one causal SNP independent from those captured by other clusters. SNPs within each signal cluster are highly correlated due to LD. *ALDH16A1* is an APA example where SuSiE revealed two independent 3′aQTL signal clusters (the lead 3′aQTLs are rs1006938 and rs73582462). The maximum R^2 between any 2 SNPs taken separately from the 2 clusters is very small (0.03), suggesting that there are indeed 2 independent signals

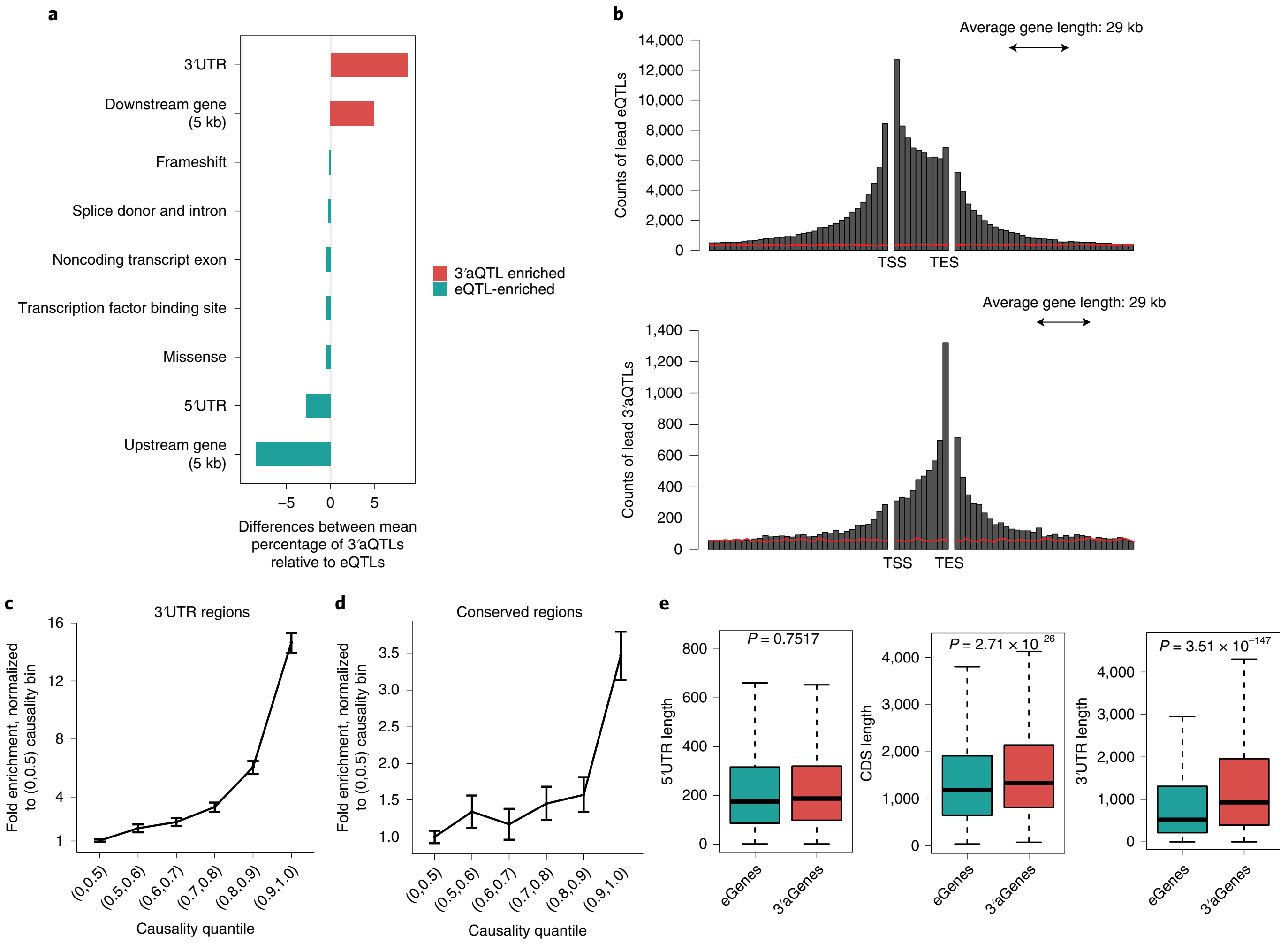


Fig. 3 | 3′aQTL represent a new type of molecular QTL. a, Differences between the mean percentages of lead 3′aQTLs and eQTLs for different annotations. The colors indicate the statistical significance of the differences, with red indicating 3′aQTL-enriched annotations, with an FDR ≤ 0.01, and green indicating eQTL-enriched annotations, with an FDR ≤ 0.01. **b**, Relative distances between eQTLs or 3′aQTLs and their associated genes. TES, transcription end site. The red line represents randomly selected positions within the ±1-Mb window for each gene. **c**, Fold enrichment and 95% confidence intervals (CIs) for 3′aQTLs in each causality bin for the intersection with 3′UTR regions. **d**, Fold enrichment and 95% CIs of 3′aQTLs that intersect with conserved regions, which were defined as regions with UCSC phastCons conservation scores >0.8. **e**, Genomic length was compared between 3′aGenes and eGenes. *P* values were calculated using a two-sided *t*-test. The center horizontal lines of the box plot show the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). *n* = 46 tissues examined.

detected. *IRF5* is another APA example where SuSiE detected only one signal cluster (the lead 3′aQTL is rs10954213). In total, 35% of tissue-transcript pairs were associated with more than 1 independent 3′aQTL, which indicated the widespread, allelic heterogeneity of 3′aQTL effects (Fig. 1d). Altogether, the approximately 0.4 million 3′aQTLs we identified provide an extensive display of how common genetic variants are associated with 3′UTR usage across multiple human tissues and expand the number of known 3′aQTLs by several orders of magnitude compared with all previously reported APA-associated SNPs.

Patterns of tissue specificity for 3′aQTLs. To examine how *cis* regulatory elements contribute to APA events in tissue-specific or shared manners (Supplementary Fig. 8), we used multivariate adaptive shrinkage (MASH)[31] to estimate the effect sizes of 3′aQTLs shared across all 46 tissues, while controlling for nongenetic correlations, such as sample overlap. The heterogeneity of cross-tissue effects was evaluated based on the sharing of signs (effects in the same direction) and magnitudes (effects in the same direction and within a twofold effect size change) among 3′aQTLs. This analysis revealed that human tissues cluster into two major groups—brain tissues and non-brain tissues—using hierarchical clustering with complete linkage (Fig. 2a). We also noted that some biologically related tissues grouped within 'non-brain' tissues, such as the uterus/vagina/ovary and colon/stomach groups (Fig. 2b). These patterns revealed developmental and functional similarities between different tissues due to APA regulation. In addition, we found that, although 78.4% of tissues had 3′aQTLs with the same sign, only 13.9% of shared 3′aQTLs displayed similar magnitudes. Compared with eQTLs shared among tissues (85% shared among tissues by sign and 36% shared among tissues by magnitude)[31], 3′aQTLs exhibited similar sign effects (Supplementary Figs. 9 and 10) but a much lower degree of shared-magnitude effects (Fig. 2c–e, Supplementary Fig. 11 and Extended Data Fig. 5). One possible explanation is that APA events are more tissue-specific than gene expression (Supplementary Fig. 12). Considered collectively, these observations suggested that 3′aQTL effect sizes exhibit greater tissue specificity than that of eQTLs.

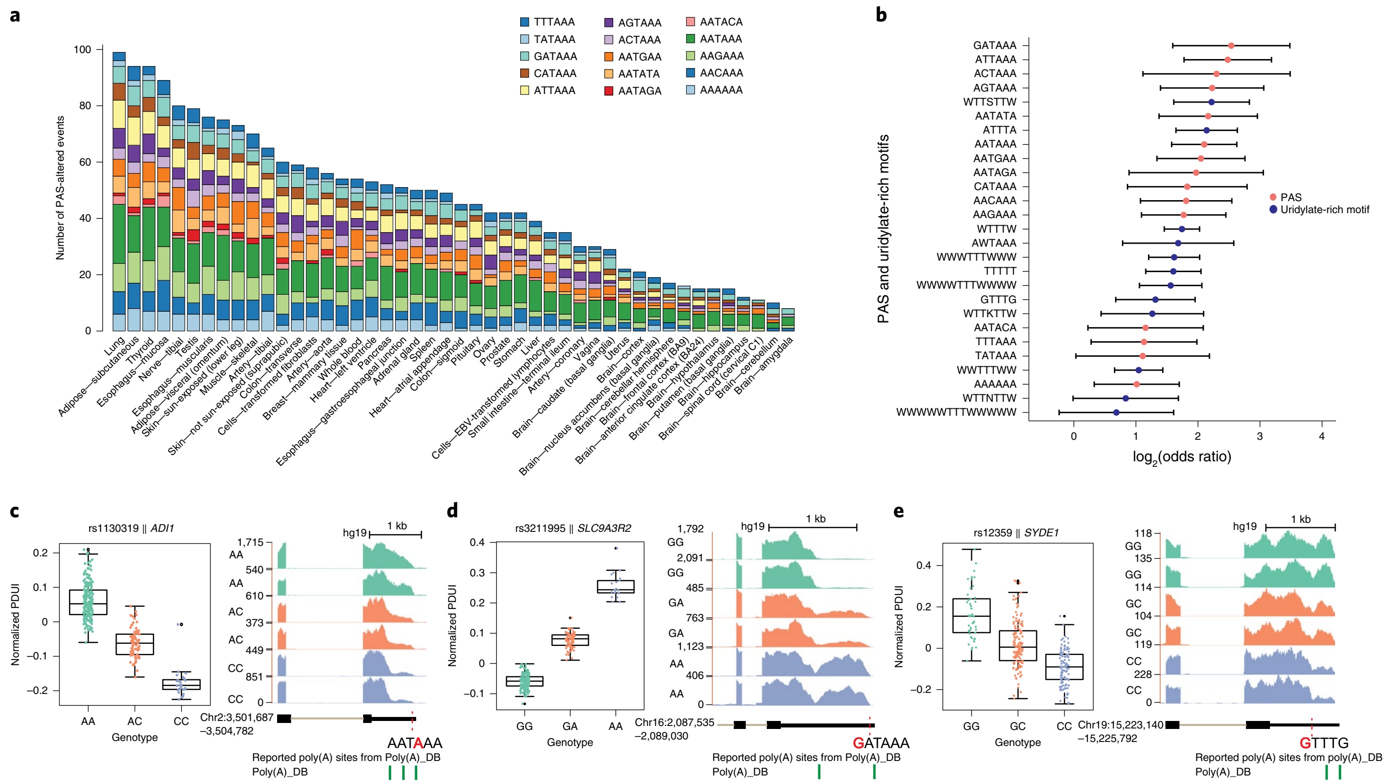


Fig. 4 | 3′aQTLs can alter PAS and uridylate-rich motifs in human tissues. a, Summary of the PAS altered by 3′aVariants across human tissues. The *x* axis shows the tissue names and the *y* axis lists the number of 3′aQTLs that alter the PAS. **b**, Enrichment of 3′aVariants that alter PAS and uridylate-rich motifs and are proximal to poly(A) sites, compared with the rest of the genome. Data are presented as odds ratio and 95% CI. **c**, Box plot showing the significant correlation between the 3′aQTL rs1130319 and *ADI1* APA events for each genotype. Each dot represents a normalized PDUI value from a single sample. The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). The coverage plot illustrates that this SNP could disrupt the canonical PAS. The red dotted line in the RefSeq gene structure indicates the location of the 3′aVariant. The PAS is shown, with the 3′aQTL highlighted in red. **d**, Box plot showing that the 3′aQTL rs3211995 is strongly correlated with the *SLC9A3R2* 3′UTR change for each genotype. The coverage plot illustrates that this SNP could 'create' a canonical PAS. **e**, Box plot showing that the 3′aQTL rs12359, which alters the uridylate-rich motif, is strongly associated with *SYDE1* 3′UTR usage for each genotype.

3′aQTLs have distinct molecular features. To characterize the relationships between different QTLs, we classified lead 3′aQTLs and lead eQTLs across 46 tissue types according to the functional categories defined in SnpEff v.5.0 (ref. [32]). As expected, we found that 3′aQTLs were significantly enriched in 3′UTRs ($P=2.68\times10^{-30}$) or located within 5 kilobases (kb) downstream of genes ($P=9.43\times10^{-08}$), whereas eQTLs were significantly enriched within gene promoters/upstream regions ($P=1.11\times10^{-34}$) or within 5′UTRs ($P=1.42\times10^{-32}$) (Fig. 3a). This observation is consistent with the metagene analysis encompassing the relative position distributions of 3′aQTLs and eQTLs over their associated genes (Fig. 3b). 3′aQTLs are distributed approximately symmetrically around the 3′UTR region and 34% of 3′aQTLs are located in downstream gene regions, likely due to the LD effect[1,33] (Methods). 3′aQTLs also differ markedly from splicing QTLs (sQTLs)[33], which are enriched primarily within gene bodies and splice regions (Extended Data Fig. 6). We also cross-referenced the recent 549 protein QTLs[34] (pQTLs) with lead 3′aQTLs and lead eQTLs. We found that 154 multi-tissue 3′aQTLs are pQTLs for the same gene in 1 or more tissues and 78.5% of pQTL-overlapped 3′aQTLs are not eQTLs. These data suggest that some 3′aQTLs can affect protein expression levels independent of gene expression.

To further determine the genomic context of 3′aQTLs, while also accounting for LD effects, we examined the enrichment of 3′aQTLs according to their posterior causal probabilities. Fine-mapped 3′aQTLs were allocated into six bins based on causality quantiles. We found that 27.4% of 3′aQTLs in the most causal bin (larger than the 90th quantile) were associated with a 14-fold enrichment in 3′UTR regions compared with 3′aQTLs in the least causal bins (less than the 50th quantile) (Fig. 3c). Interestingly, 3′aQTLs are also highly enriched in conserved regions (University of California Santa Cruz (UCSC) phastCons conservation score >0.8) (Fig. 3d) but not in transcription factor binding sites (Supplementary Fig. 13).

Moreover, the structures of 3′aGenes and eQTL-associated genes (eGenes) differed considerably. Compared with eGenes, 3′aGenes harbored comparable 5′UTRs but much longer coding sequences (CDS) ($P=2.71\times10^{-26}$) and 3′UTR lengths ($P=3.51\times10^{-147}$) (Fig. 3e). Furthermore, a significantly higher number of adenylate-uridylate-rich elements proximal to poly(A) sites were observed in 3′aGenes than in eGenes ($P=7.61\times10^{-198}$), suggesting that 3′aGenes harbor more potentially regulatory elements that control APA events (Supplementary Fig. 14). 3′aGenes are also enriched in ontologies related to immune and environmental responses, such as the IFN-γ-mediated signaling pathway. This is in contrast with eGenes, which were underrepresented in genes related to the environmental response[1]. Considered collectively, the results of these analyses suggested that 3′aQTLs and the genes affected by them have different molecular features than other previously defined QTLs and their modulated genes.

Alterations of poly(A) motifs are associated with APA. Next, we investigated the potential mechanisms through which genetic variations contribute to APA events. We hypothesized that some

3′aQTLs alter the motifs important for the 3′-end processing of transcripts. Alterations to the polyadenylation signal (PAS) can produce distinct mRNA isoforms, with 3′UTRs of differing lengths. However, only a few cases have been reported from a limited number of cell lines[9,35]. To systematically examine the prevalence of PAS-altering 3′aQTLs among human populations, we extracted significant 3′aVariants located within 50 base pairs (bp) upstream of annotated poly(A) sites from the Poly(A) database (PolyA_DB)[36], UCSC, Ensembl and RefSeq gene annotations, and performed motif searches based on 15 common PAS motif variants. In total, we identified 2,135 3′aVariants that alter the PAS and generate alternative 3′UTR lengths in their associated genes across 46 human tissues (Fig. 4a and Supplementary Table 3). A total of 991 3′aVariants either disrupted the canonical PAS (AATAAA) or changed other PAS variants to the canonical PAS ($P = 2.827 \times 10^{-10}$) (Fig. 4b). For example, a change in the rs1130319 SNP from the reference A allele to the C allele, which impairs the canonical PAS, AATAAA, correlated with the preferred use of a cryptic poly(A) site in the *ADI1* 3′UTR (Fig. 4c). We validated our finding using recent saturation mutagenesis data[23], where the same 3′aVariant disruption of the *ADI1* canonical poly(A) motif resulted in a 20-fold decrease in the abundance of the long isoform (Extended Data Fig. 7a). In another case, a G>A change in rs3211995 resulted in a strong PAS (AATAAA), instead of the weak noncanonical GATAA motif, at the 3′-end of *SLC9A3R2*, which correlated with a shift to an mRNA isoform with a longer 3′UTR (Fig. 4d). Again, saturation mutagenesis confirmed that this 3′aVariant resulted in a 42.52-fold increase in the abundance of the long isoform (Extended Data Fig. 7b). We also found that 3′aVariants are prone to alter those PAS variants that are proximal to annotated poly(A) sites (Fig. 4b). In addition to the PAS, we also investigated whether 3′aVariants could alter uridylate-rich elements, which are also important for 3′-end processing[4]. Interestingly, adenylate-uridylate, guanylate-uridylate and uridylate-rich motifs were also frequently altered by 3′aQTLs (Fig. 4b and Supplementary Fig. 15). For example, a 3′aVariant at the guanylate-uridylate-rich motif, GTTTG, located near the proximal poly(A) site of the gene *SYDE1*, could lead to significant 3′UTR shortening (Fig. 4e). The uridylate-rich motif variations on APA have been described before[37]. Collectively, these results suggested that a small fraction of detectable APA events are the result of 3′aVariants alterations of PAS or uridylate-rich motifs.

APA-associated RBP binding sites and RNA secondary structure. Alterations in polyadenylation signals can explain only a small percentage of 3′aQTLs, suggesting that most 3′aQTLs affect APA via other mechanisms. To test this hypothesis, we analyzed the extent to which 3′aQTLs interfere with either the transcriptional or posttranscriptional regulation of target genes. First, we used DeepBind v.0.11 (ref. [38]) to evaluate the enrichment of 3′aVariants in 927 binding motifs of 538 DNA-binding proteins and RBPs, in each tissue, using randomly shuffled 3′aVariants as a control group. We identified 125 motifs that were significantly enriched in 3′aVariants, 17 of which were common among at least 20% of the tissues examined (Supplementary Fig. 16). Proteins associated with these 17 common motifs were significantly enriched ($P = 1.06 \times 10^{-5}$; hypergeometric test) with known poly(A) factors, such as PABP[39], CPEB4 (refs. [39,40]), SRSF7 (ref. [41]), RBFOX1 (ref. [42]) and HNRNPC, which was recently described as an APA regulator[43].

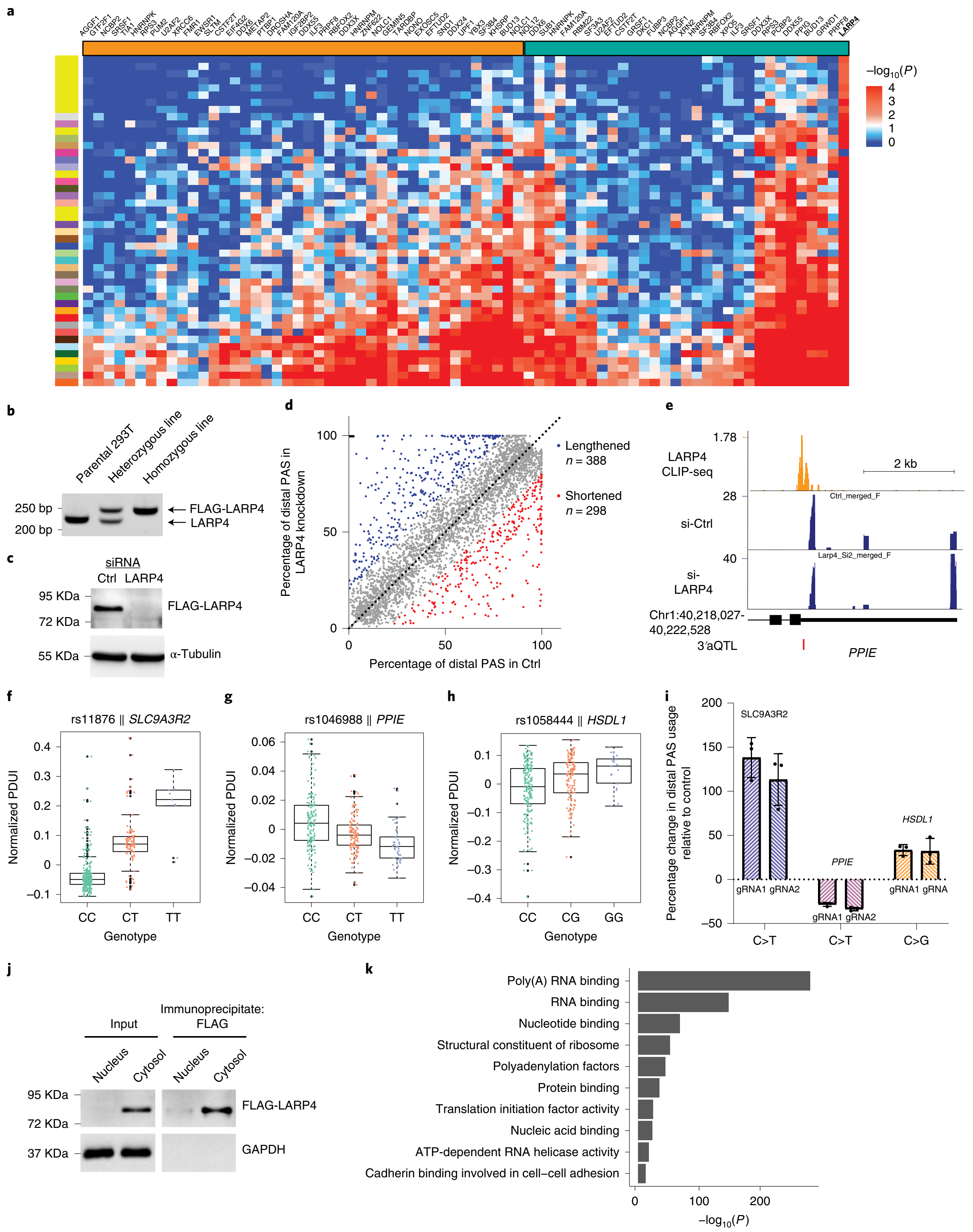
We then analyzed 166 RBP cross-linking immunoprecipitation sequencing (CLIP-seq) datasets, which were available from the Encyclopedia of DNA Elements (ENCODE) project[44]. These datasets are particularly useful because 81.2% of RBPs are not included in the DeepBind resource. We examined whether 3′aQTLs were significantly enriched within the CLIP-seq binding peaks of each RBP compared with a random sequence dataset. We further integrated a new computational strategy to predict the *trans*-regulator of APA (Methods and Extended Data Fig. 8) and identified 73 RBPs that preferentially bound to regions containing 3′aQTLs, including several poly(A) factors, such as CSTF, in addition to many splicing factors (Fig. 5a and Supplementary Table 4). Consistent with a potential functional significance, these splicing factors have previously been linked to alternative 3′UTR usage[40,41].

To evaluate the association between 3′aQTL and RNA structural features, we decided to use the riboSNitch data[45], which are defined as DNA variants affecting RNA secondary structure changes by parallel analysis of RNA structure experiments. We cross-referenced these riboSNitch data with our lead 3′aQTLs. The overlap event was defined as high LD ($R^2 \geq 0.8$) between lead 3′aQTL and riboSNitch for the same transcript. We found that 10.6% of riboSNitch data overlapped with 3′aQTLs (Supplementary Fig. 17), suggesting a strong correlation between 3′aQTLs and RNA secondary structure.

3′aQTL analysis facilitates the identification of APA regulators such as LARP4. Among the 73 3′aQTL-enriched RBPs (Fig. 5a), we found that 1 tumor suppressor, La-related protein 4 (LARP4), with binding sites primarily within 3′UTR regions (Supplementary Fig. 18), was selectively bound to 3′aQTL-containing regions across most tissues. LARP4 is an RBP that binds to the poly(A) tail of mRNA molecules[46] and regulates mRNA translation; however, to our knowledge, its role in APA regulation has not yet been reported. Our observation that LARP4 binding involves regions enriched with 3′aQTLs suggests that LARP4 might be an APA regulator. Importantly, our approach cannot distinguish whether LARP4 APA regulation is mediated through impacting poly(A) site choice in the

Fig. 5 | LARP4 is an APA regulator. a, Heatmap showing the 3′aVariant significance for RBPs identified by ENCODE in each tissue. The left bar shows the color code for each tissue; the top color bar represents the K562 and HepG2 cell lines, separately. Values in the heatmap represent the degree of enrichment for 3′aQTLs in RBP binding peaks compared with the control. **b**, PCR screening gel of clonal 293T lines with homozygous FLAG-LARP4. The primers flanking the integration site of a FLAG epitope tag at the N terminus of the *LARP4* TSS were used. The representing gel of parental 293T cells, a heterozygous targeted line and a homozygous line ($n = 3$ from 12 clonal lines screened) are shown. **c**, Western blot analysis of 293T cells transfected with either control or *LARP4* siRNA to knock down endogenous LARP4 protein. Protein lysates were extracted from whole cells after 72 h of knockdown. The gel represents one of two effective siRNAs tested, as shown in the source data files. **d**, Scatterplot analysis of PAC-seq data comparing distal poly(A) site usage between control and *LARP4* knockdown cells. **e**, Representative genome browser images of the *PPIE* gene, whose poly(A) is regulated by LARP4 and binds with LARP4, as assessed by LARP4 CLIP-seq. **f**–**h**, Predicted effects of three 3′aQTLs located within the LARP4 binding sites. Each box plot represents the PDUI differences in relation to the SNP genotypes ($n = 431$ for *SLC9A3R2* (**f**), $n = 396$ for *PPIE* (**g**) and $n = 431$ for *HSDL1* (**h**)). The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). **i**, Quantitative RT–qPCR analysis showing the altered APA regulation of three genes in response to CRISPR genome editing to introduce the 3′aQTL that was predicted to alter LARP4 binding. The 3′aQTL of each gene was targeted by two independent gRNAs and each gRNA editing was repeated ($n = 3$, shown by each dot) biologically. Data are presented as the mean ± s.d. **j**, Western blot analysis of nuclear and cytosolic extraction from the homozygous FLAG-LARP4 293T cell line. LARP4 subcellular localization was examined by anti-FLAG M2 antibody. The FLAG immunoprecipitates from each fractionation were subjected to mass spectrometry for orthogonal analysis, which confirmed the results of the western blot through definitive peptide identification. **k**, Functional annotation of the enrichment analysis for LARP4-associated proteins, based on the mass spectrometry results.

nucleus or through regulating differential stability of short/long mRNA isoforms in the cytoplasm. To test the hypothesis that LARP4 regulates APA, we first CRISPR-engineered 293T cells to harbor a single FLAG epitope tag within both copies of the endogenous *LARP4* gene (Fig. 5b). We then transfected these cells with either control small interfering RNA (siRNA) or *LARP4*-targeting siRNA



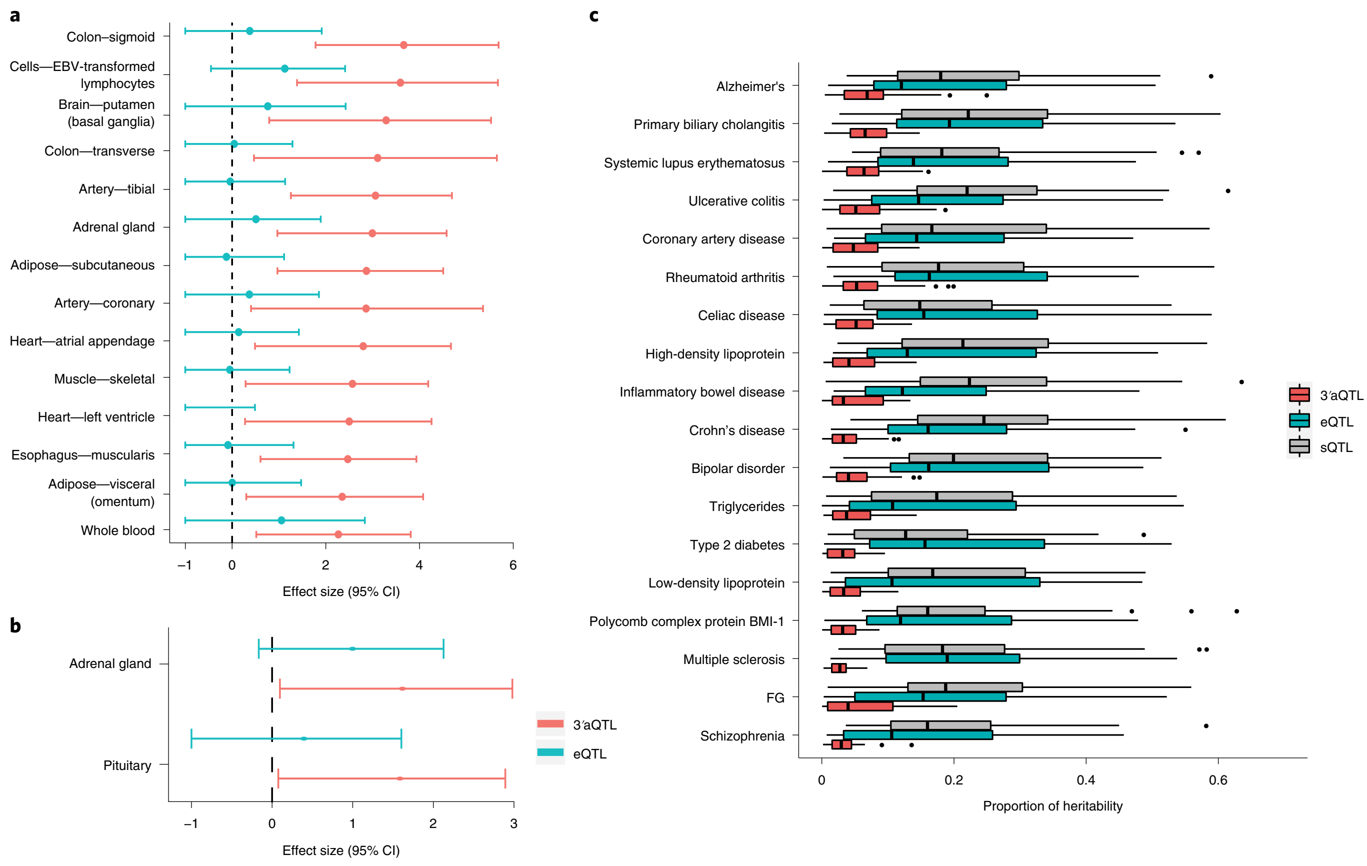


Fig. 6 | Association between 3′aQTLs and human GWAS diseases/traits. a,**b**, Tissues with 3′aVariant enrichment but no eQTL enrichment for Alzheimer's disease (**a**) and rheumatoid arthritis (**b**). The enrichment values (effect size) were calculated using functional genome-wide association analysis, which quantifies the relationships between trait-associated variants and 3′aQTLs/eQTLs. For example, a positive value indicates that variants with stronger association evidence in GWAS are more likely to be 3′aQTLs/eQTLs. The estimated lower and upper bound 95% CIs for the enrichment value are also shown. **c**, Partitioned heritability plot for the percentage of phenotypic variance (*x* axis) that can be explained for 28 traits (*y* axis) by eQTLs, 3′aQTLs and sQTLs in aggregate. FG, fasting glucose. The center lines within the box plot represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). *n* = 46 tissues examined.

and observed the robust depletion of FLAG-*LARP4* (Fig. 5c). RNA was isolated from both control and knockdown cells and analyzed using 3′-end sequencing (poly(A)-ClickSeq (PAC-seq))[47]. Using PAC-seq, we observed broad changes in poly(A) site usage after knockdown of *LARP4*, which is consistent with a role for *LARP4* in APA regulation (Fig. 5d). Importantly, several of the genes that contain 3′aQTLs that are predicted to alter LARP4 binding were also found to exhibit robust APA in response to *LARP4* knockdown (Fig. 5e and Extended Data Fig. 9). To further test the model that LARP4 can regulate APA, we focused on three genes that exhibit changes in APA after *LARP4* knockdown and contain 3′aQTLs within their LARP4 binding sites, as assessed using the LARP4 CLIP-seq data (Fig. 5f–h). We designed CRISPR-based homologous recombination templates that would allow the introduction of the *LARP4* 3′aQTL into 293T cells (Supplementary Table 5). Cells transfected with Cas9, the homologous recombination template and either of two independent single-guide RNAs (sgRNAs) were selected and APA was assessed using quantitative reverse-transcription PCR (RT–qPCR). In all three cases, we could detect notable changes in the distal poly(A) site selection, which agreed with the predicted effects of 3′aQTLs (Fig. 5f–i), suggesting that the 3′aQTL is sufficient to alter APA regulation. Finally, we generated nuclear and cytoplasmic extracts from FLAG-LARP4 cells, purified LARP4 (using FLAG affinity resin) and analyzed the purified complexes using mass spectrometry (Fig. 5j and Supplementary Table 6). Consistent with previous reports, LARP4 was primarily, but not exclusively, cytoplasmic, and we could robustly detect associated proteins involved in poly(A)-binding. Surprisingly, we also detected numerous components of the cleavage and polyadenylation machinery associated with LARP4, suggesting a potential direct role in APA regulation (Fig. 5k). Altogether, these results support a function of LARP4 in APA regulation and further validate the use of 3′aQTLs as a discovery tool for APA regulators.

3′aQTLs can explain a significant proportion of disease heritability. The GWAS approach has commonly been used to associate genetic variants with complex human traits and diseases. However, explaining how these genetic variations, particularly noncoding variations, contribute to specific phenotypes can be difficult. We hypothesized that 3′aQTLs could be used to interpret GWAS noncoding variants, particularly those located near 3′UTRs (Supplementary Figs. 19 and 20). In this study, we compiled GWAS summary statistics for 23 common human diseases and traits from previously published studies (Supplementary Table 7) and evaluated the enrichment of 3′aVariants within trait-associated GWAS SNPs for each tissue using functional genome-wide association analysis[48]. We identified the enrichment of 3′aVariants within 11.5% of tissue-trait pairs. When further compared with known eVariants that are enriched for these traits, we observed that, overall, eQTLs had larger effects than 3′aQTLs for 26.5% of the tissue-trait pairs

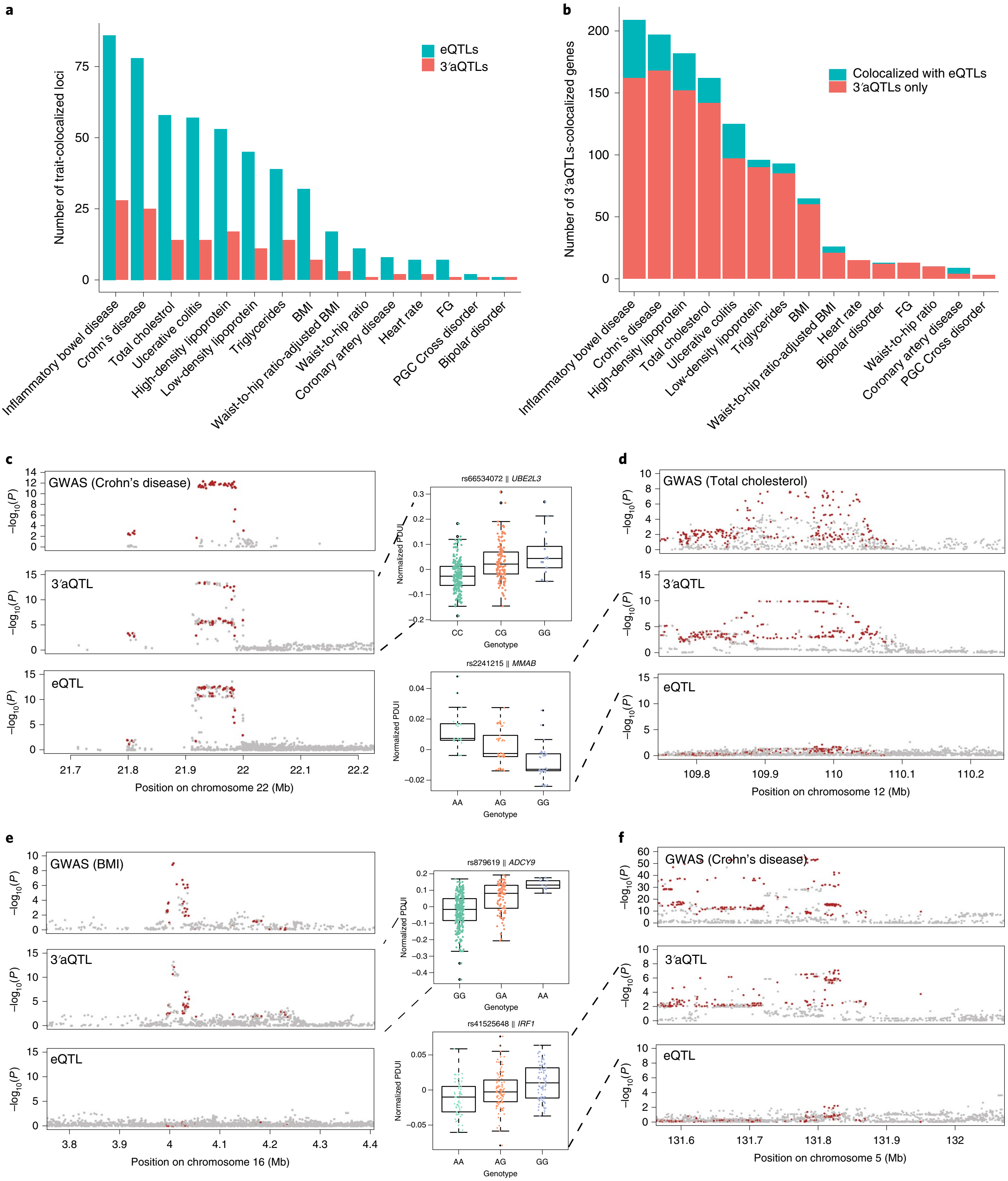


Fig. 7 | Colocalization of 3′aQTLs with complex trait-associated loci. a, Total number of colocalized GWAS signals (*y* axis) for each of 15 traits (*x* axis) across 46 human tissues. FG, fasting glucose. **b**, Number of 3′aQTL-colocalizing genes, colored based on whether the gene also colocalized with eQTLs. Blue represents 3′aQTL-colocalizing genes that are also eQTL colocalizing genes and red represents 3′aQTL-colocalizing genes are not eQTL colocalizing genes. **c**, Colocalization map of Crohn's disease-associated genes with 3′aQTLs in whole blood. The box plot shows that 3′aQTLs are strongly associated with 3′UTR usage in *UBE2L3* (*n* = 396). The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). Red indicates 3′aQTLs shared with GWAS SNPs. **d**, Colocalization map of the total cholesterol level trait with 3′aQTLs and eQTLs in liver tissue (*n* = 135). **e**, Colocalization map of BMI trait with 3′aQTLs and eQTLs in skeletal muscle tissue (*n* = 431). **f**, Colocalization map of Crohn's disease with 3′aQTLs and eQTLs in transformed fibroblasts (*n* = 291).

examined. However, in 9.8% of pairs, we found that 3′aQTLs exhibited the increased enrichment of GWAS SNPs compared with eQTLs (Supplementary Table 8), including those associated with Alzheimer's disease and rheumatoid arthritis. Notably, many of the 3′aVariants were enriched in tissues relevant to their respective diseased states, such as the brain putamen (basal ganglia) for Alzheimer's disease and the pituitary gland for rheumatoid arthritis (Fig. 6a,b). Of note, 3′aVariants were also enriched in less biologically relevant tissues, which may represent common 3′aVariants across many tissues or new trait-associated tissues[2].

To quantify the proportion of regulatory variations associated with heritability for each trait, we conducted a partitioned heritability analysis, using LD score regression[49]. Of the traits examined, the median range of SNP heritability that could be explained by 3′aQTLs, sQTLs and eQTLs was 3–7, 13–25 and 10–19% per trait, respectively. Notably, 3′aQTLs were particularly effective for explaining a large proportion of heritability associated with several autoimmune diseases, such as ulcerative colitis, primary biliary cholangitis and Alzheimer's disease. For some diseases, such as multiple sclerosis, 3′aQTLs contributed little to heritability (Fig. 6c and Extended Data Fig. 10). Taken together, although the role of APA in the modulation of these diseases has been studied at the single-gene level, such as for *tau* in Alzheimer's disease[50] and *TCF7L2* in type 2 diabetes[51], our results suggested that 3′aQTLs can explain a significant proportion of disease-associated variants.

Many trait-colocalizing 3′aQTLs are independent of gene expression. The enrichment of 3′aQTLs within disease-associated loci provide disease-specific knowledge about the overall impact of 3′aQTLs but does not necessarily imply a causal relationship. Therefore, we investigated the extent to which 3′aQTLs may function as causal variants for human phenotypes. We used colocalization analysis[52], which identifies 3′aQTLs that share the same putative causal variants with trait-associated signals, to examine 15 complex diseases and traits with known minor allele frequencies (MAFs). Of note, the colocalization model has limited power for the identification of multiple causal variants per gene. In total, 801 trait-associated variants colocalized with either eQTL or 3′aQTL signals. Consistent with previous results[1], 57% of trait-associated variants colocalized with eQTLs in 1 or more tissues. Interestingly, 16.1% of trait-associated variants colocalized with 3′aQTLs in at least 1 tissue (Fig. 7a). Of note, this 3′aQTL colocalization may still be driven by eQTLs or sQTLs (Supplementary Fig. 21). We found that 14 colocalizing 3′aQTLs were overlapped with pQTLs[34]. For example, rs503366 is not only a pQTL for *MTRF1L*, but also the lead 3′aVariant that colocalized with bipolar disorder GWAS variants (the posterior probability of a model with one shared causal variant (PP4) = 0.922). We also found that 83.7% (1,019 out of 1,218) of 3′aQTL-colocalizing genes were not eQTL colocalizing genes (Fig. 7b and Supplementary Table 9). We separated all 3′aGenes into two groups based on whether they overlapped with eQTLs. Within each group, we analyzed the differences of APA usage and gene expression with different 3′aQTL alleles. We observed no APA usage differences between eQTL-overlapped 3′aGenes and non-eQTL-overlapped 3′aGenes ($P = 0.06$; Supplementary Fig. 22a). We further found that eQTL-overlapped 3′aGenes tended to have notable gene expression changes ($P < 2.2 \times 10^{-16}$) (Supplementary Fig. 22b), whereas non-eQTL-overlapped 3′aGenes had almost no associated gene expression changes. To explore the potential regulatory mechanisms, we cross-referenced the 3′UTR regions of 3′aGenes with the TargetScan human v.6.2 (ref. [53]) miRNA binding sites and ENCODE RBP CLIP-seq peaks. We found that eQTL-overlapped 3′aGenes have overall greater miRNA binding site density within the 3′UTR region than non-eQTL-overlapped 3′aGenes ($P = 5.695 \times 10^{-5}$; Supplementary Fig. 22c). We did not find any enrichment of RBP binding sites. These results suggest that eQTL-overlapped 3′aGenes tend to affect gene expression through miRNA-mediated regulation but not through RBP regulation.

UBE2L3 is a representative example of the 16.3% of genes that colocalized with both 3′aQTLs and eQTLs. *UBE2L3* is an E2 ubiquitin-conjugating enzyme that promotes the activation of nuclear factor kappa B signaling during immune responses[54]. The rs66534072 locus in *UBE2L3* has been associated with gene expression levels and confers risk for autoimmune diseases[55]. However, the mechanisms through which these genetic variants affect gene expression are unknown. We determined that *UBE2L3* can be subject to APA and can exhibit dynamic 3′UTR use among different individuals. Moreover, the lead 3′aQTL SNP, rs66534072, was significantly correlated with 3′UTR use in *UBE2L3* (Fig. 7c). Specifically, the C allele was associated with the shortening of the *UBE2L3* mRNA 3′UTR, whereas the G allele was associated with the lengthening of the 3′UTR. We examined the tissues where rs66534072 serves as a 3′aQTL for *UBE2L3* and found that most are known to be affected by autoimmune diseases.

Most 3′aQTL trait-colocalized gene pairs are specific to 3′aQTLs and not eQTLs. For instance, *MMAB* encodes an enzyme involved in adenosylcobalamin formation, which is crucial for cholesterol degradation[56]. A total of 288 3′aQTLs were found to associate with *MMAB* 3′UTR use and were directly correlated with total cholesterol level GWAS loci on chromosome 12 (Fig. 7d). Similarly, variants on chromosome 16 that were associated with body mass index (BMI) also colocalized with 3′aQTLs that regulate 3′UTR length changes in *ADCY9* (Fig. 7e). We also observed a strong colocalization pattern between 3′aQTLs in *IRF1* and the significant GWAS loci for multiple autoimmune diseases, including ulcerative colitis, Crohn's disease and inflammatory bowel disease (Fig. 7f). *IRF1* is induced by IFN-γ signaling and promotes innate and acquired immune responses[57]. In contrast, except in musculoskeletal tissue, no strong association between eQTL and *IRF1* expression was observed. Colocalization analyses of musculoskeletal tissue revealed no colocalization patterns between disease-associated loci and *IRF1* eQTLs. In contrast, colocalization patterns for *IRF1* 3′aQTLs and autoimmune diseases were identified in multiple tissues, including transformed fibroblasts (PP4 = 0.97). These results suggested that *IRF1*-associated 3′aQTLs, more than *IRF1*-associated eQTLs, can explain most of the effects of the *IRF1* variations associated with these diseases. Collectively, our data suggest that many 3′aQTLs contribute to human diseases and traits, independent of their roles in the regulation of gene expression.

Discussion

We defined 3′aQTLs as the genetic basis for an emerging human molecular phenotype that is responsible for alternative 3′UTR usage. By reanalyzing large-scale GTEx data, using our DaPars v.2 algorithm, we identified 11,613 APA genes and approximately 0.4 million 3′aQTLs across 46 human tissues. 3′aQTLs were found to be sufficient to alter APA regulation, as demonstrated by CRISPR-based experiments and saturation mutagenesis data. In contrast with other molecular QTLs, such as eQTLs, 3′aQTLs are highly enriched within 3′UTRs. Mechanistically, 3′aQTLs likely induce changes in 3′UTR usage by either modulating the strength of poly(A) signal motifs, RNA secondary structure or RBP binding sites. 3′aQTLs that reside outside of gene-transcribed regions are likely to involve a more complex mechanistic basis as evidenced by recent work revealing connections between DNA methylation, gene looping and APA regulation[58,59]. eQTLs are important molecular features associated with human phenotypic variations. In this study, we demonstrated that 3′aQTLs represent molecular features that contribute to phenotypic variation in human populations at an unexpectedly similar level as eQTLs. Furthermore, we also validated the use of 3′aQTLs as a discovery tool for identifying APA regulators, such as LARP4.

We found that 3′aQTLs can explain a substantial proportion of trait heritability. Colocalization analyses found that 16.1% of trait-associated loci colocalized with 1 or more 3′aQTLs in human tissues. Furthermore, very few of the 3′aQTL-colocalizing trait-associated loci overlapped with eQTLs, indicating that 3′aQTLs and eQTLs are largely independent. We speculate that eQTL-independent 3′aQTLs regulate the stability, translation or cellular localization of target genes independently of the regulation of gene expression. Collectively, the results of our in-depth analyses of the genetic influence of APA events in 46 human tissues increase the fraction of common noncoding variations that can be associated with molecular phenotypes and suggest interpretations that explain how natural variations can shape human phenotypic diversity and tissue-specific diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00864-5.

Received: 30 January 2020; Accepted: 5 April 2021;
[redacted]

References

1. GTEx Consortium.Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
2. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
3. Mayr, C. Regulation by 3′-untranslated regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
4. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
5. Mayr, C. What are 3′ UTRs dDoing? *Cold Spring Harb. Perspect. Biol.* **11**, a034728 (2018).
6. Masamha, C. P. et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412–416 (2014).
7. Weng, T. et al. Cleavage factor 25 deregulation contributes to pulmonary fibrosis through alternative polyadenylation. *J. Clin. Invest.* **129**, 1984–1999 (2019).
8. Park, H. J. et al. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat. Genet.* **50**, 783–789 (2018).
9. Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* **8**, e1002882 (2012).
10. Zhernakova, D. V. et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
11. Stacey, S. N. et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43**, 1098–1103 (2011).
12. Higgs, D. R. et al. α-Thalassaemia caused by a polyadenylation signal mutation. *Nature* **306**, 398–400 (1983).
13. van der Maarel, S. M., Tawil, R. & Tapscott, S. J. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. *Trends Mol. Med.* **17**, 252–258 (2011).
14. Fahiminiya, S. et al. A polyadenylation site variant causes transcript-specific BMP1 deficiency and frequent fractures in children. *Hum. Mol. Genet.* **24**, 516–524 (2015).
15. Garin, I. et al. Recessive mutations in the *INS* gene result in neonatal diabetes through reduced insulin biosynthesis. *Proc. Natl Acad. Sci. USA* **107**, 3105–3110 (2010).
16. Hellquist, A. et al. The human *GIMAP5* gene has a common polyadenylation polymorphism increasing risk to systemic lupus erythematosus. *J. Med. Genet.* **44**, 314–321 (2007).
17. Graham, R. R. et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl Acad. Sci. USA* **104**, 6758–6763 (2007).
18. Cannavò, E. et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541**, 402–406 (2017).
19. Mariella, E., Marotta, F., Grassi, E., Gilotto, S. & Provero, P. The length of the expressed 3′ UTR is an intermediate molecular phenotype linking genetic variants to complex diseases. *Front. Genet.* **10**, 714 (2019).
20. Xia, Z. et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.* **5**, 5274 (2014).
21. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
22. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
23. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
24. Kwan, T. et al. Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225–231 (2008).
25. Hoarau, J.-J., Cesari, M., Caillens, H., Cadet, F. & Pabion, M. *HLA DQA1* genes generate multiple transcripts by alternative splicing and polyadenylation of the 3′ untranslated region. *Tissue Antigens* **63**, 58–71 (2004).
26. Cunninghame Graham, D. S. et al. Association of *IRF5* in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.* **16**, 579–591 (2007).
27. Sheng, G., dos Reis, M. & Stern, C. D. Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* **115**, 603–613 (2003).
28. Lyons, J. J. et al. Elevated basal serum tryptase identifies a multisystem disorder associated with increased *TPSAB1* copy number. *Nat. Genet.* **48**, 1564–1569 (2016).
29. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
31. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
32. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
33. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
34. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
35. Thomas, L. F. & Sætrom, P. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. *PLoS Comput. Biol.* **8**, e1002621 (2012).
36. Lee, J. Y., Yeh, I., Park, J. Y. & Tian, B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.* **35**, D165–D168 (2007).
37. Sun, H. S. et al. A polymorphic 3′UTR element in ATP1B1 regulates alternative polyadenylation and is associated with blood pressure. *PLoS ONE* **8**, e76290 (2013).
38. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
39. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.* **9**, 563–576 (2012).
40. Bava, F.-A. et al. CPEB1 coordinates alternative 3′-UTR formation with translational regulation. *Nature* **495**, 121–125 (2013).
41. Müller-McNicoll, M. et al. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev.* **30**, 553–566 (2016).
42. Chen, P.-F., Hsiao, J. S., Sirois, C. L. & Chamberlain, S. J. RBFOX1 and RBFOX2 are dispensable in iPSCs and iPSC-derived neurons and do not contribute to neural-specific paternal *UBE3A* silencing. *Sci. Rep.* **6**, 25368 (2016).
43. Gruber, A. J. et al. A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**, 1145–1159 (2016).
44. Dominguez, D. et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **70**, 854–867.e9 (2018).
45. Wan, Y. et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
46. Yang, R. et al. La-related protein 4 binds poly(A), interacts with the poly(A)-binding protein MLLE domain via a variant PAM2w motif, and can promote mRNA stability. *Mol. Cell. Biol.* **31**, 542–556 (2011).
47. Routh, A. et al. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res.* **45**, e112 (2017).
48. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

49. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
50. Dickson, J. R., Kruse, C., Montagna, D. R., Finsen, B. & Wolfe, M. S. Alternative polyadenylation and miR-34 family members regulate tau expression. *J. Neurochem.* **127**, 739–749 (2013).
51. Locke, J. M., Da Silva Xavier, G., Rutter, G. A. & Harries, L. W. An alternative polyadenylation signal in *TCF7L2* generates isoforms that inhibit T cell factor/lymphoid-enhancer factor (TCF/LEF)-dependent target genes. *Diabetologia* **54**, 3078–3082 (2011).
52. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
53. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
54. Lewis, M. J. et al. *UBE2L3* polymorphism amplifies NF-κB activation and promotes plasma cell development, linking linear ubiquitination to multiple autoimmune diseases. *Am. J. Hum. Genet.* **96**, 221–234 (2015).
55. Wang, S. et al. A functional haplotype of *UBE2L3* confers risk for systemic lupus erythematosus. *Genes Immun.* **13**, 380–387 (2012).
56. Holleboom, A. G., Vergeer, M., Hovingh, G. K., Kastelein, J. J. & Kuivenhoven, J. A. The value of HDL genetics. *Curr. Opin. Lipidol.* **19**, 385–394 (2008).
57. Kano, S. et al. The contribution of transcription factor IRF1 to the interferon-γ-interleukin 12 signaling axis and T_H1 versus T_H-17 differentiation of $CD4^+$ T cells. *Nat. Immunol.* **9**, 34–41 (2008).
58. Nanavaty, V. et al. DNA methylation regulates alternative polyadenylation via CTCF and the cohesin complex. *Mol. Cell* **78**, 752–764.e6 (2020).
59. Mittleman, B. E. et al. Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**, e57492 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Mapping of GTEx RNA-seq data. Original RNA-seq reads were aligned with the human genome (hg19/GRCh37) using STAR v.2.5.2b[60], with the following alignment parameters: outSAMtype, BAM; SortedByCoordinate; outSAMstrandField, intronMotif; outFilterMultimapNmax, 10; outFilterMultimapScoreRange, 1; alignSJDBoverhangMin, 1; sjdbScore, 2; alignIntronMin, 20; and alignSJoverhangMin, 8. The resulting sorted BAM files were converted into bedGraph formats using BEDTools version 2.17.0 (ref. [61]).

Covariate correction. To account for hidden batch effects and other unobserved covariates in each tissue, we first corrected the sample genotype for population structure. Briefly, we first removed sites marked as 'wasSplit' from the GTEx analysis freeze variant call format (VCF) using BCFtools v.1.3, leaving 39,741,769 biallelic sites. The variants were further filtered with a call rate of >99% and MAF >5%; LD pruning was performed using PLINK v.2.0. The top three principal components from the principal component analysis were consistent with the known three main subpopulations, including White, Black or African American and Asian, in the GTEx samples. We further used PEER[21] with sex, RNA integrity number, top 5 genotype principal components and genotyping platforms as the known covariates to estimate a set of latent covariates for the PDUI values in each tissue. The number of PEER factors was optimized based on suggestions from the GTEx Consortium[1]; for tissue sample sizes <150, 15 PEER factors were chosen. Thirty PEER factors were chosen if the sample size ranged from 150 to 250 and 35 peer factors were chosen for >250 samples. We analyzed the correlation between PEER factors and covariates reported for the GTEx samples and noticed that many of these covariates were strongly associated with PEER factors, such as nucleic acid isolation batch and total ischemic time, which were associated across tissues (Extended Data Fig. 1). We also included three measurements for 3′Bias statistics: (1) 3′ 50-base normalization, which is the ratio between the coverage at the 3′-end and the average coverage of the full transcript, averaged over all transcripts; (2) 5′ 50-base normalization, which is the ratio between the coverage at the 5′-end and the average coverage of the full transcript, averaged over all transcripts; and (3) the number of transcripts that have at least one read at their 5′-end. The inferred PEER factors were highly correlated with the 3′Bias statistics (Extended Data Fig. 1), indicating that most of the 3′Bias effects have been corrected by our PEER analysis.

Furthermore, to comprehensively evaluate the other genotypic covariates, we correlated the PEER factors with donor covariates in each tissue. We observed that our PEER factors were consistently correlated with several donor covariates such as donor death, ischemic time, Hardy scale, EBV immunoglobulin M antibody and age (Extended Data Fig. 2).

3′aQTL mapping for each tissue. A whole-genome sequencing variant file for 635 individuals was obtained from the GTEx database of Genotypes and Phenotypes (dbGaP) website (phs000424.v7.p2), under the name 'GTEx_Analysis_2016-01-15_v7_WholeGenomeSeq_635Ind_PASS_AB02_GQ20_HETX_MISS15_PLINKQC.vcf.gz', from which 17 samples and all the variants that failed to pass the quality control step initially defined by the GTEx Consortium[1] were removed. Any individuals with no RNA-seq data were also removed. 3′aQTL mapping was performed separately for each tissue. Subset VCF data for each tissue were extracted, using BCFtools. VCF files were transformed into an SNP matrix file, including genotyping information, using BioAlcidae v.2.27.1 (ref. [62]). SNPs with a MAF of <0.01 were filtered and at least 10 counts per allele were required. We then tested associations for SNPs within an interval of 1 Mb from the 3′UTR region, with normalized PDUI values, in each tissue, using Matrix eQTL[22], in a linear regression framework.

Permutation analysis was conducted to identify significant 3′aQTL-associated gene pairs. Individual labels were randomly sampled 1,000 times and the minimum P value for each SNP and gene was recorded after each 3′aQTL mapping. These empirical P values were adjusted using the qvalue v.2.0.0 R package[63]. Genes with a $q < 0.05$ were considered to be significant APA genes. All APA gene-associated 3′aQTLs were subsequently identified with the FDR set to 5%.

Fine-mapping of causal variants to 3′aQTLs. We used SuSiE[30] to fine-map 3′aQTL. SuSiE can operate on individual-level data (genotypes and APA phenotypes) and can efficiently analyze loci containing many independent effect variables. We allowed a maximum of 10 independent effects in our analysis. Additionally, we verified our SuSiE results with causal variant identification in associated regions analysis[64], which uses summary statistics (z-scores derived from 3′aQTL association P values) and LD matrices but is limited to the detection of a small number of independent effects per region due to its computational capability constraints.

3′aQTL sharing and specificity analyses among tissues. 3′aQTL sharing and specificity among tissues were analyzed using MASH[31]. Briefly, we converted 3′aQTL association statistics to MASH formats. Lead 3′aQTLs and random SNP sets for each APA gene were extracted from each tissue to calculate MASH priors. A total of 4,470 genes, with no data missing from any tissue, were retained to train the MASH model. Prior covariance matrices were inferred via Empirical Bayes matrix factorization, implemented in factors and loadings by adaptive shrinkage; the multivariate 3′aQTL model was constructed using MASH. Posterior effect sizes were computed by applying the trained model to the lead 3′aQTLs sets. MASH aims to elucidate the heterogeneity of 3′aQTL effect sizes across tissues (Fig. 2). With MASH, we can learn about which 3′aQTLs have tissue-specific effect sizes and which have effect sizes consistent across tissues. This provides interesting insights into the genetic architecture of APA in different tissues. The MASH model was trained on a large random subset of SNPs[31], not the lead SNPs. The trained model was then applied to one lead SNP per gene for posterior inferences, to avoid dealing with LD between SNPs when more than one SNP in a gene was involved. Such 'one effect per region' simplification is widely accepted in a similar context to circumvent LD complications when it comes to evaluating association signals in a small region[48,52,65]. This essentially limits the scope of the investigation to a subset of 3′aQTLs but it is sufficient for our purpose to learn patterns of 3′aQTL sharing across tissues. If a lead SNP is only significant in one tissue and not the others, it will be considered a tissue-specific 3′aQTL; however, if the lead SNP is also significantly associated with APA in other tissues, even though the associations in these tissues are not as strong as the tissue based on which it is selected, it will be considered a shared 3′aQTL among tissues.

To examine whether MASH-estimated magnitudes were affected by read depth, we first downsampled 80% of the raw reads in each sample for the 5 representative tissues and reran the whole analysis. The correlations between the same tissues with different sequencing depths (100 versus 80%) were much stronger than the correlations between different tissues with the same sequencing depths (Supplementary Fig. 9a). We also downsampled the samples in each tissue to match the lowest coverage level, 15 million reads, among the included tissue samples. Still, we observed much stronger correlations between the same tissues with different sequencing depths than between different tissues at the same sequencing depth (Supplementary Fig. 9b).

Prediction of *trans* regulator of APA. For a gene G in a tissue type, all samples were ranked based on the expression levels of gene G. The top 10 most highly expressed samples and bottom 10 least expressed samples were chosen as the two groups. If the mean gene expression fold change between the two groups was >2 with $P < 0.05$, these two groups were treated as control and knockdown groups. Then, the PDUI values between the groups could be compared to identify significant dynamic APA genes between the high and low expression groups of gene G. Using this strategy, we calculated the number of 3′UTR shortening or lengthening effect of each gene, which regulates significant dynamic APA events between the high and low expression groups. The gene will be predicted as a *trans* regulator of APA if $P < 0.05$. We have validated our method in a few known APA regulators, such as CSTF2, which was described as an APA regulator promoting 3′UTR shortening. We observed that there was a marked shift of 3′UTR shortening in individuals with highly expressed CSTF2 (Extended Data Fig. 8a). We also investigated our newly detected APA regulator, LARP4. We often observed many APA events when comparing LARP4$^{\mathrm{high}}$ and LARP4$^{\mathrm{low}}$ individuals (Extended Data Fig. 8b).

Colocalization analyses. We utilized a Bayesian colocalization approach to identify GWAS signals that could exhibit the same genetic effects between eQTLs and 3′aQTLs, using the coloc v.3.2-1 R package[52]. The full summary statistics for 15 GWAS were used when the MAF was available. For each GWAS trait, we extracted the sentinel SNPs, which were defined as GWAS SNPs with $P < 5\times10^{-8}$ and located at least 1 Mb away from more significant variants. The colocalized signals were searched for within the 100-kb surrounding region of sentinel SNPs. As defined by the coloc method, five posterior probabilities (PPs) were calculated. PP0 represents the null model of no association. PP1 and PP2 represent the probability that causal genetic variants are either associated with disease signals only or with 3′aQTLs only, respectively. PP3 represents the probability that the genetic effects of disease signals and 3′aQTLs are independent and PP4 represents the probability that disease signals and 3′aQTLs share causal SNPs. The genes were defined as colocalization events if PP4 ≥ 0.75 and PP4/(PP4 + PP3) ≥ 0.9. Region visualization plots were constructed using LocusZoom v.1.4 (ref. [66]). LDs between reference SNPs and 3′aQTLs were calculated using PLINK[67].

Cell culture and cloning. The HEK 293T cell line (catalog no. CRL-3216; ATCC) was grown in high-glucose DMEM supplemented with 10% FCS and 50 U ml^{-1} penicillin-streptomycin (Thermo Fisher Scientific). The oligonucleotides used for cloning are listed in Supplementary Table 5. pST1374-NLS-flag-linker-Cas9 and pGL3-U6-sgRNA-PGK-puromycin plasmids for CRISPR targeting were a gift from X. Huang (plasmid nos. 44758 and 51133, respectively; Addgene). Each pair of oligonucleotides of sgRNAs was annealed and cloned into a pGL3-U6-sgRNA plasmid. The identities were confirmed by Sanger sequencing. *LARP4* RNA interference experiments were performed using a two-hit strategy, as described previously[68]. Briefly, 60 pmol of *LARP4* siRNA (SASI_Hs01-00187288; Sigma-Aldrich) was diluted in 100 µl of Opti-MEM. For each siRNA, 3 µl of RNAiMAX (Thermo Fisher Scientific) was diluted in 100 µl of Opti-MEM and incubated for 5 min at room temperature. Diluted siRNA and RNAiMAX were mixed and incubated for another 20 min at room temperature. Cells were seeded in 12-well plates at a density of 4×10^5, in 1 ml of regular growth medium, immediately before adding the complexes. Transfected cells were incubated at 37 °C and 5% CO_2 for 24 h. For the second forward transfection, 90 pmol of siRNA and 4.5 µl of RNAiMAX were used to form transfection complexes, as described for the first transfection. The medium was replaced with fresh medium before adding the complexes. After another 24 h, cells were expanded to 6-well plates and grown for a total of 72 h before being collected. To check protein expression, anti-FLAG-HRP M2 (catalog no. A8592; Sigma-Aldrich) in 1:5,000 dilution, anti-alpha Tubulin (catalog no. ab15246; Abcam) in 1:2,000 dilution and anti-GAPDH (catalog no. AM4300; Thermo Fisher Scientific) in 1:4,000 dilution was used for western blotting.

CRISPR genomic editing. CRISPR was used to precisely incorporate the FLAG sequence (gat tac aag gat gac gac gat aag), as described previously[69], into all endogenously expressed LARP4 proteins, at the N terminus. Briefly, a 100-bp genomic sequence surrounding the translation start site (TSS) was input into the CRISPOR program (http://crispor.tefor.net/crispor.py) for guide RNA (gRNA) prediction.

Two gRNAs were selected based on the following: (1) the shortest distance between the Cas9 cutting site (NGG is the protospacer adjacent motif) and the FLAG insertion site; and (2) the specificity score, based on the number of off-target effects. To design the single-strand DNA donor template, a 200-bp genomic sequence (including the 24 bases of the FLAG sequence in the middle) surrounding the TSS was synthesized by Integrated DNA Technologies. The protospacer adjacent motif on the donor template was mutated silently to avoid being attacked by transfected gRNA/Cas9. Equal amounts of gRNA and Cas9 plasmids (720 ng in total) were mixed with 10 pM (approximately 660 ng) of donor template and transfected into 4×10^5 HEK 293T cells in 24-well plates with Lipofectamine 2000. Cells were moved to 6-well plates after overnight incubation; selection (10 µg ml^{-1} of blasticidin and 1 µg ml^{-1} puromycin) was started 24 h after transfection for a total of 48 h. Cells were expanded in regular growth medium, without selection antibiotics. FLAG western blots were performed to determine the signal from pools of cells and confirm the signal from clonal lines. Genomic DNA was extracted from those clonal lines with FLAG signals on western blots. PCR was performed to amplify the approximately 200-bp fragment containing the FLAG sequence; the product was resolved on agarose gels to determine the homogeneity of FLAG insertion on all alleles of the target gene.

For 3′aQTL alterations, double-stranded DNA donor templates (approximately 500 bp from each of three genes) were amplified from HEK 293T genomic DNA. 3′aQTLs were designed to be located approximately two-thirds downstream from the 5′-end for higher CRISPR efficiency[70]. PCR-based mutagenesis was performed to alter the 3′aQTLs. Transfection and selection were performed as described above. RNA and genomic DNA were extracted from a pool of cells for distal PAS usage measurements and Sanger sequencing, respectively.

PAC-seq. To identify alternative polyadenylation sites, PAC-seq[47,71] was adopted to sequence *LARP4* knockdown samples. Briefly, poly(A) mRNA was enriched from 5 µg of total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), as described by the manufacturer's protocol. All enriched mRNA was reverse-transcribed into complementary DNA. First, 2 µl of a 5-mM mixture containing 3′-azido-2′,3′-dideoxyadenosine-5′-triphosphate (N-4007, N-4008 and N-4014; TriLink Biotechnologies) and deoxynucleoside triphosphate, at a ratio of 1 to 5, was added to the RNA sample together with 1 µl of 100 µM 3′Illumina_4N_21T primer. Regular RT–qPCR steps, using SuperScript III, were performed. The sample was treated with 1 µl of ribonuclease H (Thermo Fisher Scientific) for 20 min at 37°C, followed by 10 min at 80°C for inactivation. cDNA was purified using AMPure XP beads, as described by the manufacturer's instructions, and eluted in 12 µl of 50 mM of HEPES, pH 7.4. The click reaction was performed by first adding 23 µl of premixed Click-Adaptor (20 µl of dimethylsulfoxide and 3 µl of 5 µM of Click-Adaptor) to 10 µl of cDNA and then adding 2.4 µl of premixed catalyzer (0.4 µl of 50 mM of vitamin C and 2 µl of 10 mM of Copper(II)-TBTA (Lumiprobe)). After a 30-min incubation at room temperature, 2.4 µl of catalyzer was added to the reaction to boost reaction efficiency. 5′ Clicked cDNA was purified using AMPure XP beads.

PCR amplification was performed using 5′ short universal primer and 3′ indexing primer, which has a unique index for each sample. One*Taq* 2X Master Mix (New England Biolabs) was used to amplify the library under the following conditions: 1 min at 94 °C, 30 s at 55 °C, 10 min at 68 °C and 16 cycles of 30 s at 94 °C, 30 s at 55 °C, 2 min at 68 °C. Finally, the PCR extension was performed at 68 °C for 5 min, followed by 4 °C, indefinitely. The library was purified using AMPure XP beads; size selection was performed on 2% E-Gel EX Agarose Gels (Thermo Fisher Scientific), targeting fragments between 200 and 400 bp. The library was extracted from the gel using ZYMO DNA Clean & Concentrator 5 and quantified by a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) before being sequenced on an Illumina next-generation sequencer. PAC-seq data were analyzed with the differential poly(A) clustering DPAC[72] pipeline using the exon-centric approach, with the --P --M --C --A --B and --D options. The results were filtered such that genes or exons required a minimum of 10 mean reads in each sample, a 1.5-fold change and an adjusted $P < 0.01$ to be considered significantly differentially expressed. Genes with more than one PAS also required a percentage distal PAS usage change of 20% to be considered a change in the length of the 3′UTR.

Nuclear and cytosolic protein extraction. Cells were washed and collected in cold PBS and resuspended in a fivefold cell pellet volume of Buffer A (10 mM of Tris, pH 8, 1.5 mM of $MgCl_2$, 10 mM of KCl, 0.5 mM of dithiothreitol (DTT) and 0.2 mM of phenylmethylsulfonyl fluoride). Cells were allowed to swell during a 15-min rotation at 4 °C, then pelleted at 1,000*g* for 10 min, after which cells were homogenized in twofold the original cell pellet volume Buffer A with a Dounce pestle B for 20 strokes on ice. Nuclear and cytosolic fractions were separated by centrifugation at 2,000*g* for 10 min. For the cytosolic fraction, 10× Buffer B (300 mM of Tris, pH 8, 1.4 M of KCl and 30 mM of $MgCl_2$) was added to the supernatant to a final concentration of 1× Buffer B. Debris was removed by centrifugation at 15,000*g* for 30 min at 4°C. For the nuclear fraction, the pellet was washed once with Buffer A before resuspending the original cell pellet volume of Buffer C (20 mM of Tris, pH 8, 420 mM of NaCl, 1.5 mM of $MgCl_2$, 25% glycerol, 0.2 mM of EDTA, 0.5 mM of phenylmethylsulfonyl fluoride and 0.5 mM of DTT). The sample was homogenized with a Dounce pestle B for 20 strokes on ice and rotated for 30 min at 4 °C before centrifugation at 15,000*g* for 30 min at 4 °C. Supernatants were collected from both fractions and subjected to dialysis in Buffer D (20 mM of HEPES, 100 mM of KCl, 0.2 mM of EDTA, 0.5 mM of DTT and 20% glycerol) overnight at 4°C. Lysates were centrifuged again at 15,000*g* for 3 min at 4 °C to remove any precipitates before downstream applications.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw GTEx RNA-seq and genotype files are available to authorized users through dbGaP release, under accession no. phs000424.v7.p2. A list of 3′aQTLs, lead 3′aQTLs and their associated *APA* genes, isoform usage-controlled 3′aQTLs and expression-controlled 3′aQTLs are freely available at Synapse (accession no. syn22236281; https://doi.org/10.7303/syn22236281). Raw and processed PAC-seq data for the LARP4-depletion experiment have been deposited with the Gene Expression Omnibus under accession no. GSE139548. The proteomics data have been deposited with the MassIVE database under accession no. MSV000087000. A website portal dedicated to trait- and disease-associated 3′aQTLs can be accessed at https://wlcb.oit.uci.edu/3aQTL/index.php. Source data are provided with this paper.

Code availability

The open-source DaPars v.2.0 program is freely available at https://github.com/3UTR/DaPars2.

References

60. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Lindenbaum, P. & Redon, R. bioalcidae, samjs and vcffilterjs: object-oriented formatters and filters for bioinformatics files. *Bioinformatics* **34**, 1224–1225 (2018).
63. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
64. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
65. Aerts, J. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
66. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
67. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
68. Wagner, E. J. & Garcia-Blanco, M. A. RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol. Cell* **10**, 943–949 (2002).
69. Baillat, D., Russell, W. K. & Wagner, E. J. CRISPR–Cas9 mediated genetic engineering for the purification of the endogenous integrator complex from mammalian cells. *Protein Expr. Purif.* **128**, 101–108 (2016).
70. Song, F. & Stieger, K. Optimizing the DNA donor template for homology-directed repair of double-strand breaks. *Mol. Ther. Nucleic Acids* **7**, 53–60 (2017).
71. Elrod, N. D., Jaworski, E. A., Ji, P., Wagner, E. J. & Routh, A. Development of Poly(A)-ClickSeq as a tool enabling simultaneous genome-wide poly(A)-site identification and differential expression analysis. *Methods* **155**, 20–29 (2019).
72. Routh, A. DPAC: a tool for differential poly(A)–cluster usage from poly(A)-targeted RNAseq data. *G3 (Bethesda)* **9**, 1825–1830 (2019).

Acknowledgements

We thank L. Hou, Z. Cui and members of the Li laboratory for helpful discussions. This work was supported by a Computational Cancer Biology Training Program fellowship Cancer Prevention Research Institute of Texas (CPRIT) grant no. RP170593 (K.-L.H.), National Institutes of Health grants to W.L. (nos. R01HG007538, R01CA193466, R01CA228140), P.J. (no. R03CA223893), E.J.W. (no. R01GM134539) and a University of California, Irvine (UCI) Cancer Center support grant (no. P30CA062203). E.J.W. acknowledges support from University of Texas Medical Branch (UTMB) Department of Biochemistry and Molecular Biology startup funds and the Welch Foundation (no. H-1889). The UTMB Mass Spectrometry Facility is supported in part by CPRIT grant no. RP190682 (W.K.R.). We also thank I. Toufique and P. Papadopoulos at the UCI Research Cyber Infrastructure Center for high-performance computing support.

Author contributions

L.L. and W.L. conceived and supervised the project. L.L., Y.G., Y.C., Y.E.C. and G.W. performed the data analyses. K.-L.H., N.D.E., W.K.R. and P.J. performed the experiments. L.L., Y.L., Y.C., F.P., E.J.W. and W.L. interpreted the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

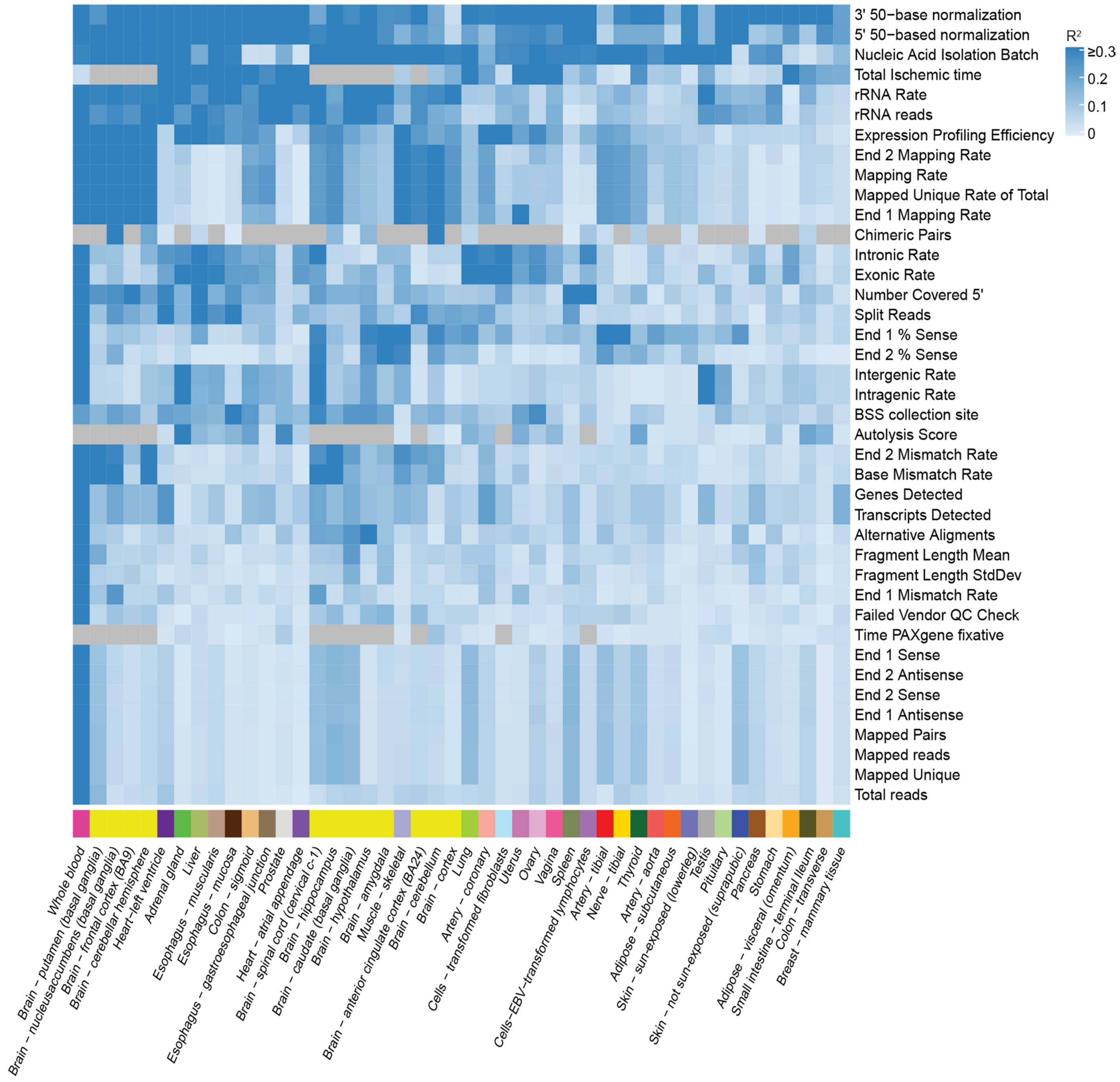
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00864-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00864-5.

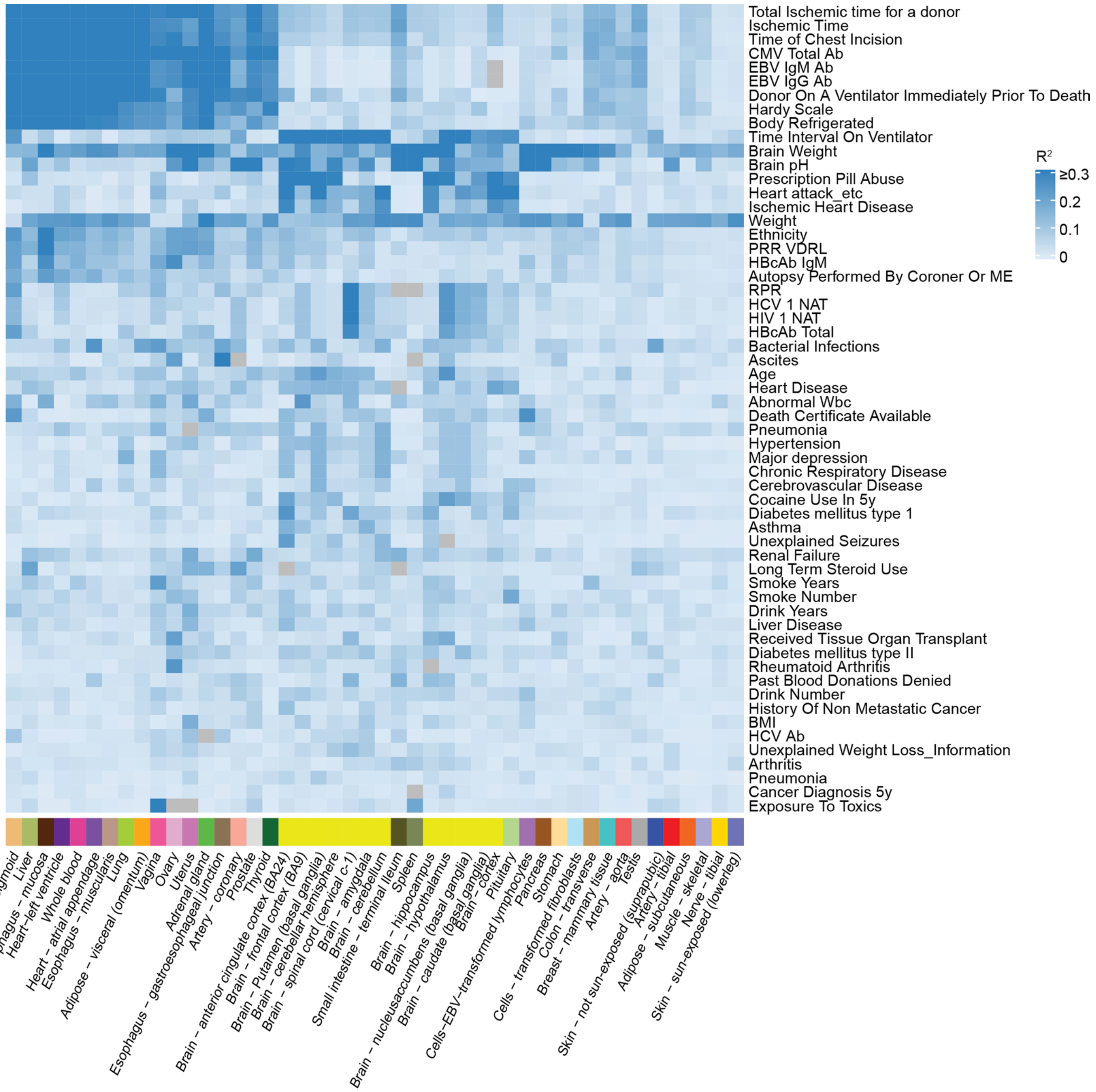
Correspondence and requests for materials should be addressed to E.J.W. or W.L.

Peer review information *Nature Genetics* thanks Stephen Montgomery, Bin Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

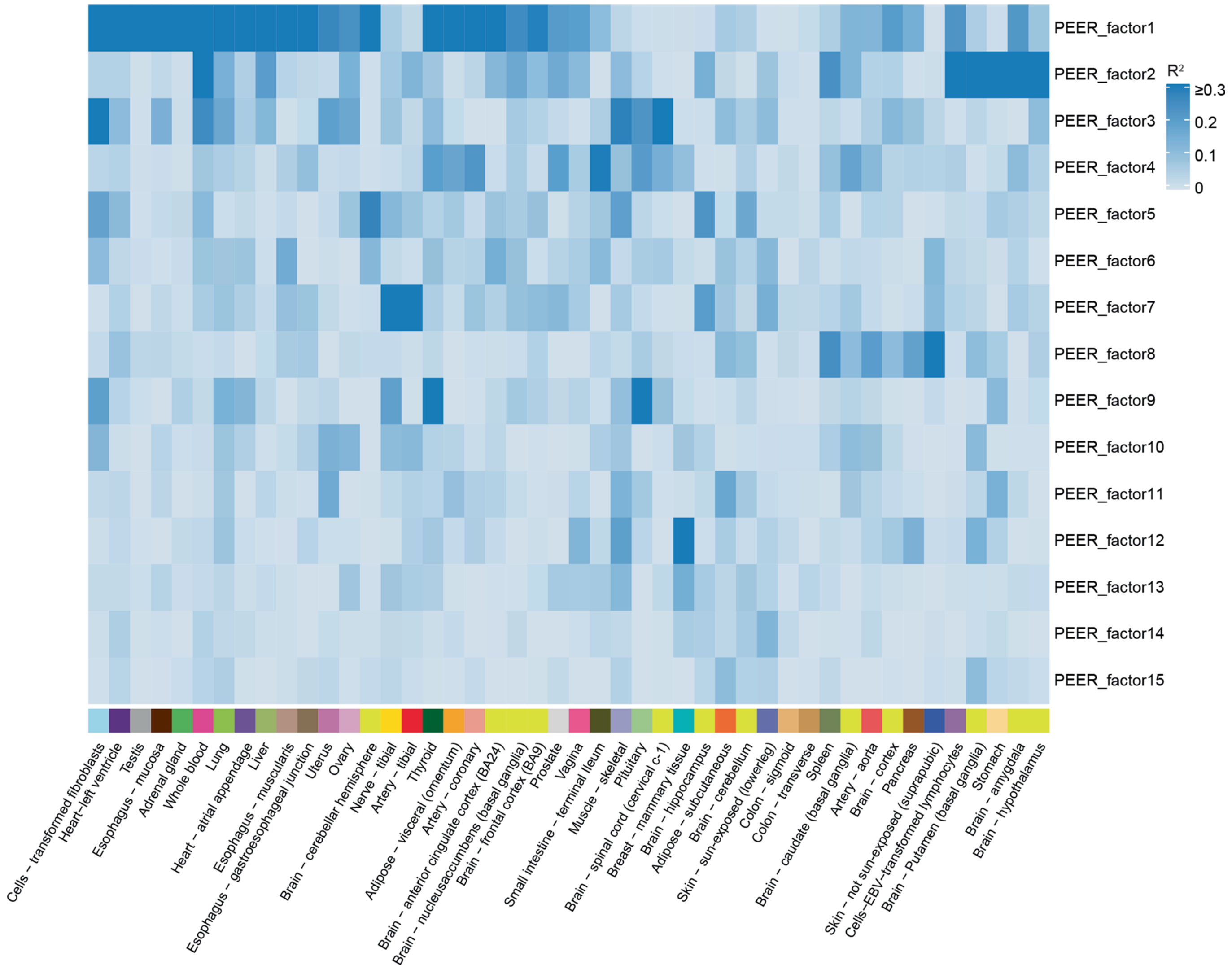
Reprints and permissions information is available at www.nature.com/reprints.



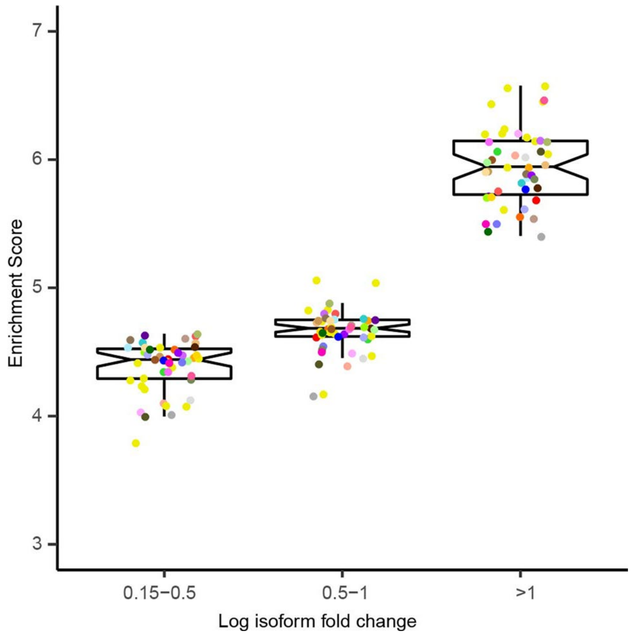
Extended Data Fig. 1 | Known technical covariates associated with inferred PEER factors in each tissue. The R^2 value in each cell represents the percentage of variance explained for each tissue/covariates pair. Only the most relevant sample-specific covariates were used. Gray color represents insufficient data to predict correlations. Each color code below indicates a tissue of origin.



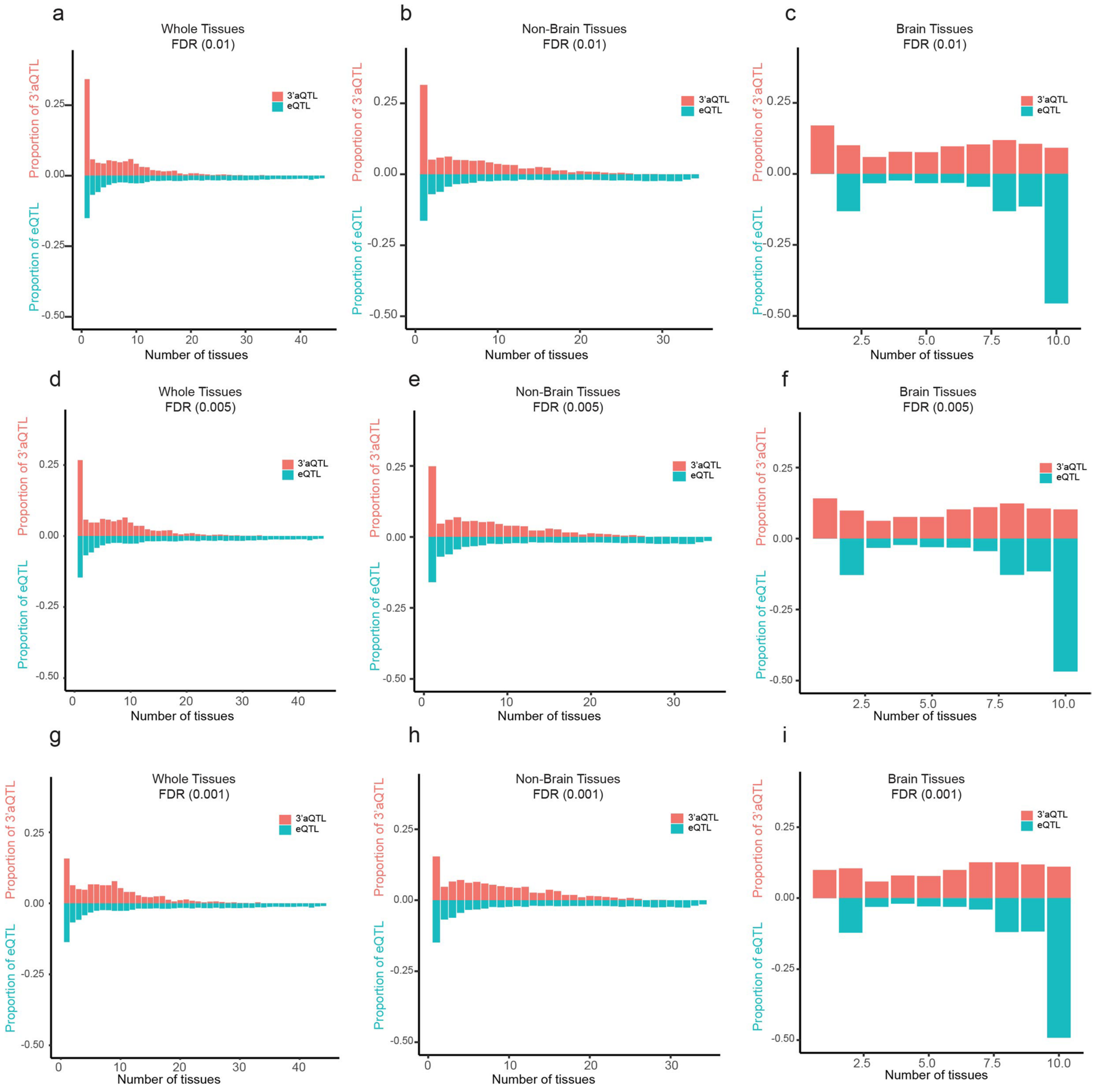
Extended Data Fig. 2 | Known donor covariates associated with inferred PEER factors in each tissue. The R^2 value in each cell represents the percentage of variance explained for each tissue/covariate pair. Only the most relevant donor-specific covariates were used. Gray color represents insufficient data to predict correlations. Each color code below indicates a tissue of origin.



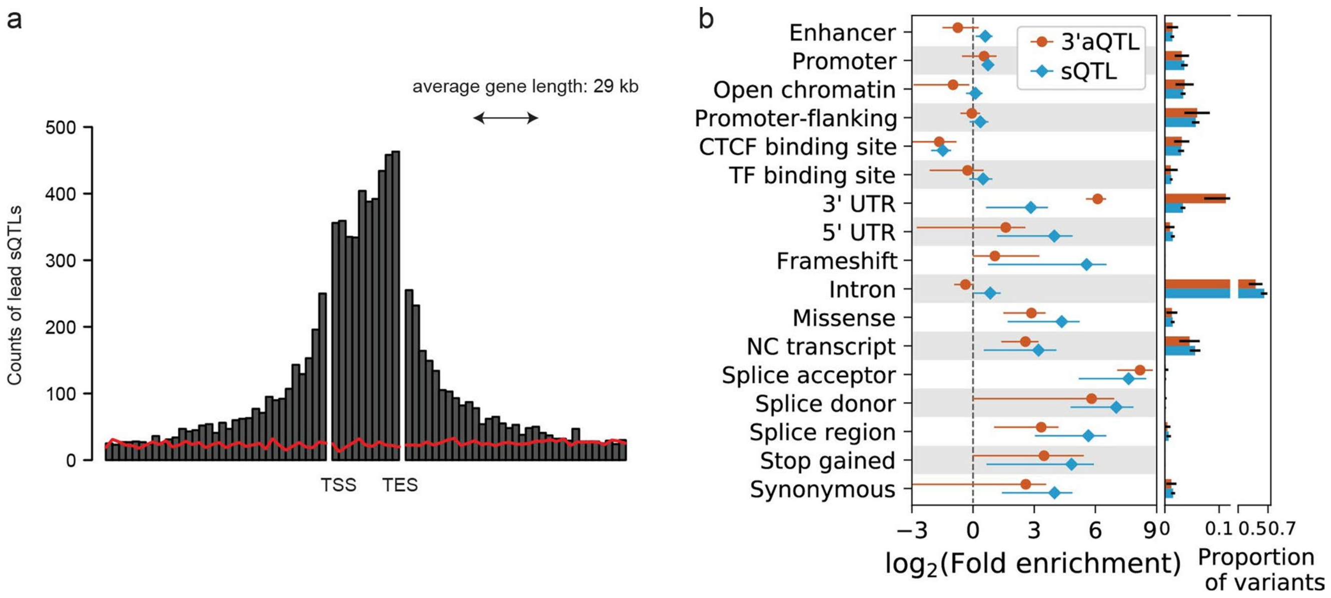
Extended Data Fig. 3 | PEER factors for gene expression associated with PEER factors for PDUI in each tissue. The R^2 value in each cell represents the correlation between the top PEER factors for gene expression (rows) and the most relevant PEER factors for PDUI for each tissue (columns). Each color code below indicates a tissue of origin.



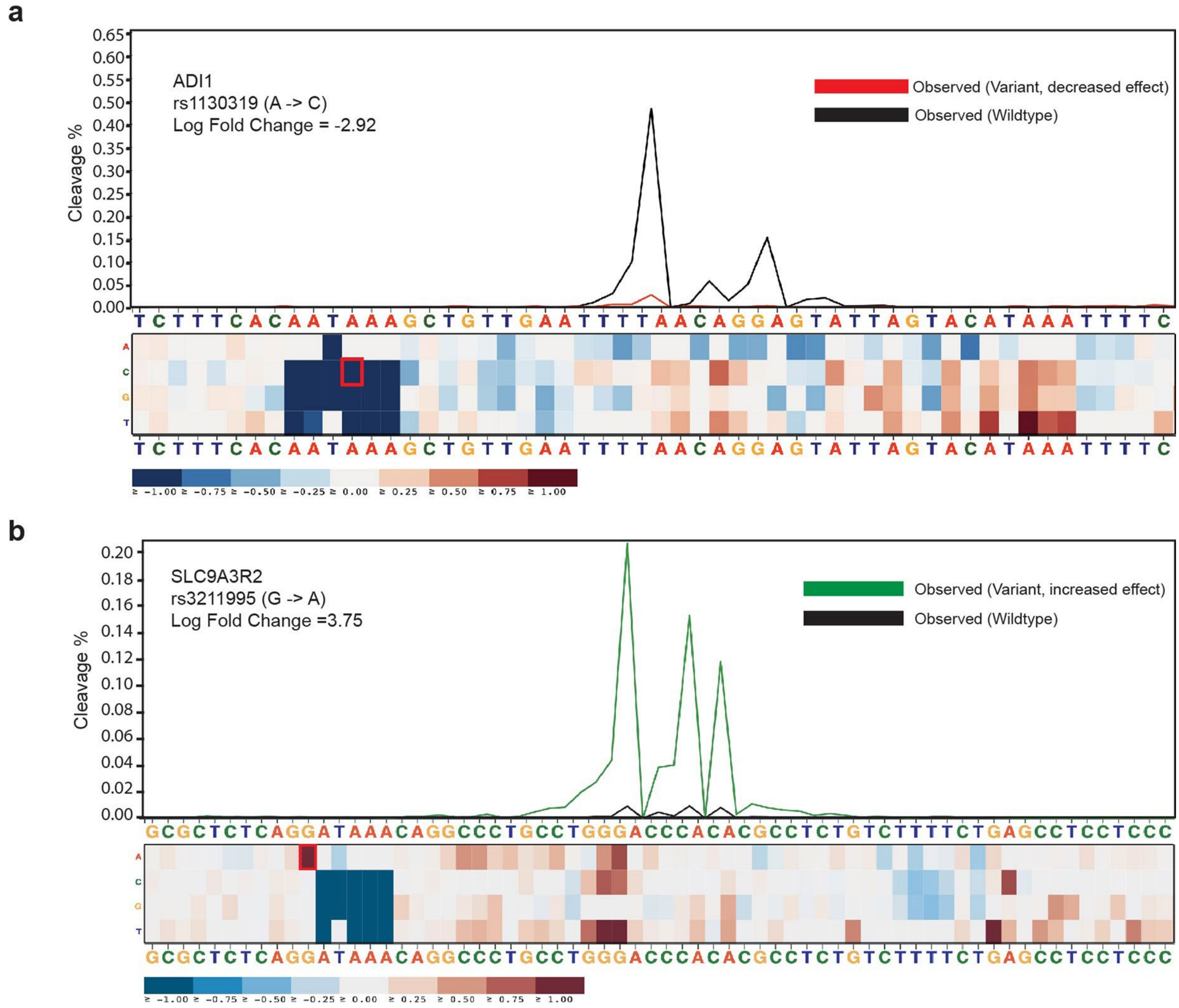
Extended Data Fig. 4 | Enrichment of 3′aQTL in different categories of mutagenesis variants annotations. The enrichment score represents the log odd ratio and accessed by the program Torus. The x-axis represents three categories of variants with different effects in predicting APA isoform log fold change due to the variant. Each color code indicates a tissue of origin. The saturation mutagenesis data with log isoform fold change < 0.15 are not available from Bogard *et al*.



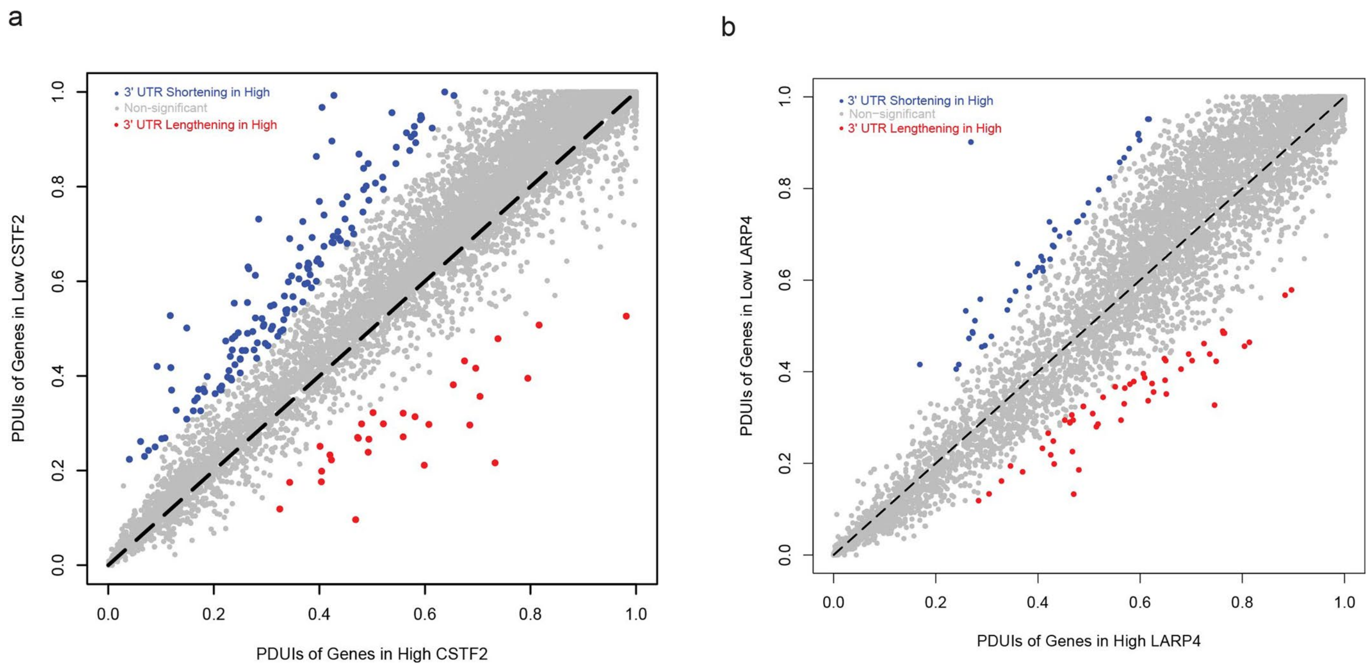
Extended Data Fig. 5 | The sharing magnitude of 3′aQTLs using different FDRs at 0.01, 0.005, 0.001. Histograms showing the estimated proportion of tissues that share lead 3′aQTLs /eQTLs, by magnitude, with other tissues, among all 46 examined tissues, among non-brain tissues only, and among brain tissues only.



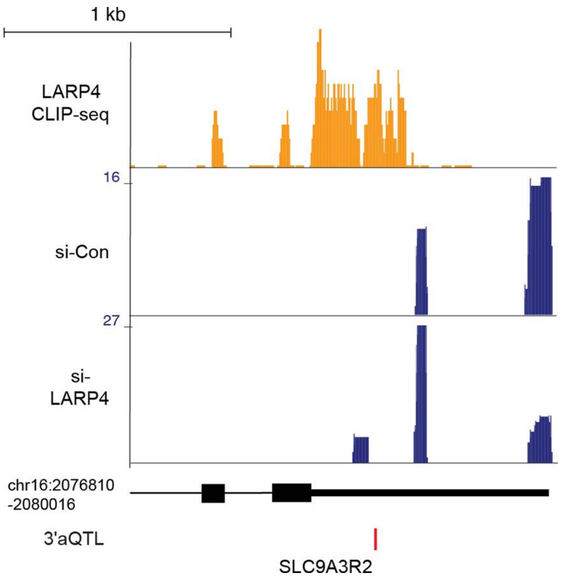
Extended Data Fig. 6 | sQTL have a distinct genomic distribution and functional enrichment compared with 3′aQTL. a, Relative position distance between sQTL and their associated genes. TSS represents the transcription start site; TES represents the transcription end site. Red line represents randomly selected positions within the +/− 1Mb window for each gene. **b**, 3′aQTL and sQTL enrichment in functional annotations. The enrichment is shown as mean with SD across tissues. The proportion of variants was also included for 3′aQTL and sQTL. Data are presented as mean value +/− Standard deviation. n = 46 tissues examined.



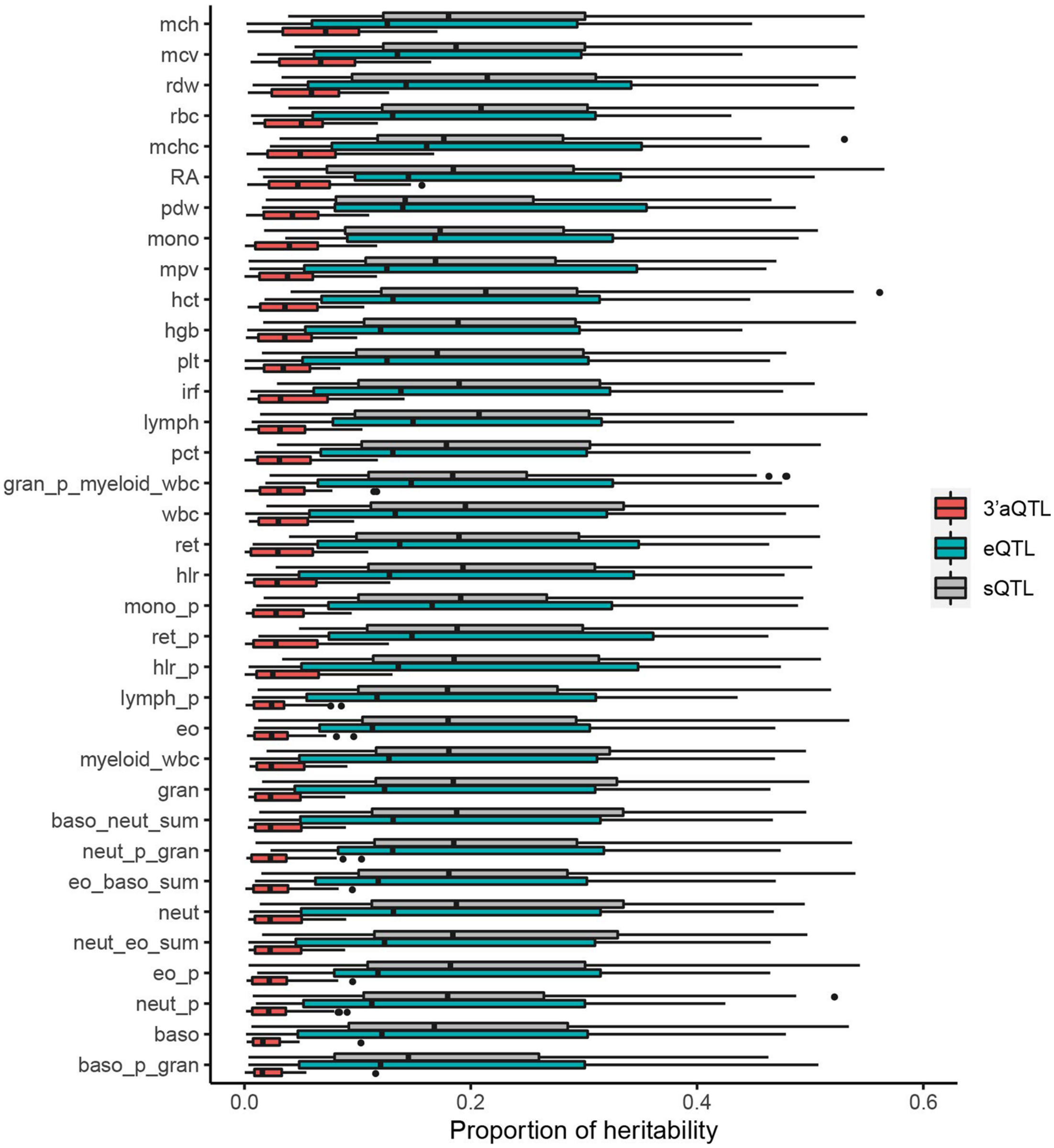
Extended Data Fig. 7 | 3′aQTLs are validated by saturation mutagenesis data. a, Saturation mutagenesis of the *ADI1* PAS. Shown above is the measured wild-type (black) and variant cleavage distribution (red) for the SNP rs1130319. The heatmap below shows the measured isoform fold changes as a result of each SNP. The red box color indicates the SNP rs1130319.



Extended Data Fig. 8 | Trans-regulator APA prediction. **a**, Scatterplot of the percentage of distal polyA site usage index (PDUI) in CSTF2 over-expressed and low-expressed samples where mRNA significantly shortened (blue) or lengthened (red) are colored. **b**, Scatterplot of PDUI changes for LARP4 over-expressed and low-expressed samples were shown.



Extended Data Fig. 9 | Representative genome browser images of the *SLC9A3R2* gene. *SLC9A3R2* APA is regulated by *LARP4* and binds *LARP4*, as assessed by *LARP4* CLIP-seq.



Extended Data Fig. 10 | A partitioned heritability plot for the percentage of phenotypic variance can be explained, for 35 traits, by 3′aQTLs, eQTLs, and sQTLs in aggregate. The trait/tissue pairs with heritability not significantly greater than 0 are removed. Centre horizontal lines show median values, boxes span from the 25th percentile to the 75th percentile. Whiskers extend to 1.5 × IQR (bottom), where IQR is the interquartile range. n = 46 tissues examined.

nature research

Corresponding author(s): Eric J Wagner and Wei Li

Last updated by author(s): 3/30/2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	STAR v2.5.2b, APAtrap v1.0, GETUTR v2.0.0, Cufflinks v2.2.1,PEER v1.3, GCTA v1.93, bcftools v1.3, Matrix eQTL v2.1.0, SuSiE (https://github.com/stephenslab/susieR), MEME v5.0.5, SMR v1.0.3, DeepBind v0.11, Fgwas v0.3.6, R 3.4.0, Ldsr v1.0.1, Coloc v3.2-1, bedtools v2.17.0, plink v2.0, The open-source DaPars v2 program is freely available at https://github.com/3UTR/DaPars2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the databases/datasets used in the study along with appropriately accessible links in the manuscript under the “Data availability” section as well as in this reporting summary. Raw GTEx RNA-seq and genotype files are available to authorized users through dbGaP release, under accession number phs000424.v7.p2. A list of 3'aQTLs, lead 3'aQTLs, and their associated APA genes, isoform usage-controlled 3'aQTLs, expression-controlled 3'aQTLs are freely available in Synapse (accession number: syn22236281, doi: 10.7303/syn22236281). Raw and processed PAC-seq data for the LARP4-depletion experiment have been deposited to GEO, under the accession number GSE139548. The proteomics data have been deposited to MassIVE database with accession number MSV000087000. A website portal dedicated to trait and disease associated 3'aQTLs can be accessed at https://wlcb.oit.uci.edu/3aQTL/index.php. AREsite2 database can be accessed at http://rna.tbi.univie.ac.at/AREsite2.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the availability of existing GTEx data.
Data exclusions	We removed diseased tissue cells and the leukemia cell line, and seven other tissues, including the cervix endocervix, cervix ectocervix, fallopian tube, bladder, kidney cortex, minor salivary gland, and brain substantia nigra due to small sample sizes.
Replication	The experiments has been performed independently with biological triplicates.
Randomization	The samples have been assigned randomly at the beginning of experiments.
Blinding	The bioinformatics analyses have been corroborated with blinded wet lab experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	1. Anti-Flag-HRP, clone M2 (Sigma, #A8592); 2. Anti-alpha-tubulin (Abcam, #ab15246); 3. anti-GAPDH, clone 6C5 (ThermoFisher, #AM4300).
Validation	1. ANTI-FLAG M2 monoclonal antibody is useful for detection, identification, and capture of fusion proteins containing a FLAG peptide sequence by common immunological procedures, such as Western blotting and Co-immunoprecipitation. 2. Anti-alpha-tubulin polyclonal antibody is used to detect human microtubule marker by Western blotting. 3. Anti-GAPDH monoclonal antibody is used to detect human GAPDH by Western Blotting. Anti-Flag-HRP M2 monoclonal antibody is registered with ID: AB_439702. It is used for detection of Flag fusion proteins (N-terminal and C-terminal) on Western blots application. The minimum detection range of Flag-fusion protein tested by company is around 8ng shown on Dot blot. Anti-alpha-Tubulin polyclonal antibody is registered with ID: AB_301787. It has been validated by company on Western blot application. This antibody gives a predominant band at expected molecular weight around 55KD after blotting whole cell extracts from mammalian cell lines. Anti-GAPDH monoclonal antibody is registered with ID: AB_437392. It has been validated by company on western blot application. This antibody gives a single band at expected molecular weight around 37KD after blotting whole cell extracts from mammalian cell lines. The citations of each antibody can be bound in the website, The Antibody Registry, by searching its ID.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	293T cell line is purchased from ATCC
Authentication	The 293T cell line was authenticated by STR profiling by ATCC
Mycoplasma contamination	The 293T cell line was tested negative of mycoplasma in our lab using MycoSensor qPCR Assay Kits (#302107, Agilent)
Commonly misidentified lines (See ICLAC register)	There is no commonly misidentified cell line used in the study