Impact of Rare Non-coding Variants on Human Diseases through Alternative Polyadenylation Outliers

Lei Li

lei.li@szbl.ac.cn

Shenzhen Bay Laboratory https://orcid.org/0000-0003-3924-2544

Xudong Zou

Institute of Systems and Physical Biology, Shenzhen Bay Laboratory https://orcid.org/0000-0002-2958-0438

Zhaozhao Zhao

Fudan university

Yu Chen

Fudan university

Kewei Xiong

Shenzhen Bay Laboratory

Zeyang Wang

Shenzhen Bay Laboratory https://orcid.org/0000-0001-5735-0675

Shuxin Chen

Shenzhen Bay Laboratory

Hui Chen

Institute of Systems and Physical Biology, Shenzhen Bay Laboratory

Gong-Hong Wei

Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School Basic Medical Sciences, Shanghai Medi https://orcid.org/0000-0001-6546-9334

Shuhua Xu

School of Life Sciences, Fudan University

Wei Li

University of California, Irvine https://orcid.org/0000-0001-9931-5990

Ting Ni

Collaborative Innovation Center of Genetics and Development, Human Phenome Institute, School of Life Sciences, Fudan University https://orcid.org/0000-0001-7007-1072

Article

Keywords:

Posted Date: March 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3907149/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Although rare non-coding variants (RVs) play crucial roles in human complex traits and diseases, understanding their functional mechanisms and identifying those most closely associated with diseases continue to be major challenges. Here, we constructed the first comprehensive atlas of alternative polyadenylation (APA) outliers (aOutliers) from 15,201 samples across 49 human tissues. Strikingly, these aOutliers exhibit unique characteristics markedly distinct from those of outliers based on transcriptional abundance or splicing. This is evidenced by a pronounced enrichment of RVs specifically within aOutliers. Mechanistically, aOutlier RVs frequently alter poly(A) signals and splicing sites, and experimental perturbation of these RVs indeed triggers APA events. Furthermore, we developed a Bayesian-based APA RV prediction model, which successfully pinpointed a specific set of RVs with significantly large effect sizes on complex traits or diseases. A particularly intriguing discovery was the observed convergence effect on APA between rare and common cancer variants, exemplified by the combinatorial regulation of APA in the *DDX18* gene. Together, this study introduces a novel APA-enhanced framework for individual genome annotation and underscores the importance of APA in uncovering previously unrecognized functional non-coding RVs linked to human complex traits and diseases.

Introduction

The human genome harbors numerous rare genetic variants[1], each with a minor allele frequency (MAF) of less than 1%. Many of these rare variants strongly contribute to human diseases[2–5]. While exome sequencing of large population cohorts has identified numerous rare protein-coding variants associated with both common and rare diseases[6], the vast majority of rare variants (RVs) are located in non-coding regions. These non-coding RVs do not function through altering the protein sequences, thereby posing a significant challenge in interpreting their functions. To address this challenge, analysis of population-scale transcriptomic data has been used to uncover functional rare non-coding variants affecting gene expression or splicing outliers[7–10]. Despite these efforts, a significant portion of disease-associated RVs remain uncharacterized.

Alternative polyadenylation (APA) of mRNA is a widespread post-transcriptional regulatory mechanism observed across various species. By employing different polyadenylation sites within 3'untranslated regions (3' UTRs), genes can produce various mRNA isoforms with either shortened or extended 3' UTRs. These 3' UTRs contain many regulatory elements that modulate the abundance or localization of the mRNA and protein[11–15]. Moreover, APA can also occur in intronic regions, leading to truncated mRNA or proteins [16,17]. Accordingly, disruptions in APA events have been increasingly implicated in many human diseases[17–19]. For example, altered APA leading to 3' UTR shortening of competing-endogenous RNAs for tumor suppressor genes can result in the release of microRNAs, inhibiting tumor suppressor genes and potentially leading to tumorigenesis[20]. Moreover, recent studies have reported the ubiquitous genetic regulation of APA, highlighting its importance in the functional interpretation of disease-associated non-coding variants[21–23]. A notable example is a single-nucleotide polymorphism (SNP; rs10954213) within

the 3' UTR of interferon regulatory factor 5 (*IRF5)*, which can alter the length and stability of its 3' UTR, thereby contributing to systemic lupus erythematosus susceptibility[24]. In our previous study, we built an atlas of human 3' UTR APA quantitative trait loci (3'aQTLs) across human tissues, identifying approximately 0.4 million common SNPs associated with interindividual APA changes, which colocalize with 16.1% of trait-associated genetic variants[25]. Yet, these studies mainly focus on assessing the APA regulation of common variants. To our knowledge, the effect of RVs on APA has not been explored.

Here, to better understand the impact of RVs on APA, we systematically analyzed aberrant APA events across 49 human tissues from the Genotype-Tissue Expression Project (GTEx). We identified 1,534 multi-tissue APA outliers (aOutliers) from European individuals. Intriguingly, 74.2% of these aOutliers are associated with genes not previously identified in outlier analysis of other molecular phenotypes (e.g., expression or splicing). These aOutliers exhibit distinct characteristics, such as unique 3' UTR length and GC-contents, setting them apart from other types of molecular outliers. Moreover, a significant enrichment of deleterious RVs was observed in regions proximal to these aOutliers. To prioritize functional RVs impacting APA, we developed a Bayesian hierarchical model and identified a distinct set of RVs with large effect sizes on human complex traits and disease phenotypes. Intriguingly, we observed and demonstrated strong convergence effects between prioritized RVs and common variants in regulating 3' UTR APA, exemplified by the combinatorial regulation of APA in *DDX18*. Lastly, to facilitate broad access to aOutliers-associated RVs, we have constructed a user-friendly portal at http://bioinfo.szbl.ac.cn/rareAPA/index.php. Collectively, our findings indicate that APA highlights a specific set of RVs with significant impacts on human traits and diseases, providing a new avenue for interpreting rare human non-coding genetic variants.

Results

The landscape of APA outliers across 49 human tissues

We first conducted a comprehensive identification of 3' UTR and intronic APA events in 15,201 GTEx RNA-seq samples from 49 human tissues of 838 individuals (Fig. 1a) using our Dapars2[25,26] and IPAFinder[18] algorithms, respectively (see Materials and Methods) (Supplementary Fig. 1). Considering the potential influence of many known and unknown technical confounders on APA usage among samples, we regressed out these confounders, such as age, sex, sequencing platform, and other hidden confounders inferred by using probabilistic estimation of expression residuals (PEER) factors (Supplementary Fig. 2). We then calculated Z-scores for the PEER-adjusted 3' UTR and intronic APA usage in each tissue to identify individuals with aberrant APA usage for a specific gene, which we refer to as APA outliers (aOutliers) with an absolute Z-score > 3. The individuals and genes were designated as “aOutlier individuals” and “aOutlier genes”, respectively. Importantly, a single gene could be associated with multiple outlier individuals, and conversely, one individual could be an aOutlier individual for multiple genes. Our analysis of these aOutliers revealed that, on average, 68.5% of all transcripts per tissue were present in at least one outlier individual (Supplementary Fig. 3a). The number of aOutlier genes strongly

correlated (Spearman's correlation rho = 0.91, $P < 2.2 \times 10^{-16}$) with sample size across tissues (Supplementary Fig. 3b), suggesting that additional aOutlier genes might be discovered as more RNA-seq samples become available. This strong sample size correlation was further confirmed by down-sampling analyses in representative tissues (Supplementary Fig. 3c). Moreover, we noticed that the incidence of an aOutlier identified in one tissue being replicated in another was as low as 14.3% (Supplementary Fig. 4), indicating a significant degree of tissue-specificity among these single-tissue aOutliers.

We further defined multi-tissue aOutliers based on aberrant APA usage across five or more tissues (see Materials and Methods). From this analysis, we identified a total of 2,147 multi-tissue aOutliers, comprising 1,930 3' UTR aOutliers and 217 intronic aOutliers based on the genomic location of the APA event. Focusing specifically on the 715 European individuals, in whom we detected 1,534 multi-tissue aOutliers, including 1,334 3' UTR and 200 intronic aOutliers (Fig. 1b and Supplementary Figs. 5 and 6). In our further investigation into the distribution of multi-tissue aOutliers across different tissues, we found that intronic aOutliers exhibited a broader replication pattern than 3' UTR aOutliers (one-sided Wilcoxon rank–sum test $P = 3.35 \times 10^{-14}$; Fig. 1c, d). Notably, among these aOutliers, several significant genes were identified (Fig. 1e–g and Supplementary Fig. 7a–f), including *SUGP1*, known for its crucial role in mRNA splicing regulation in cancer[27,28]. In certain outlier individual(s), *EIF2A*, *FLYWCH*, *TP53RK*, and *SUGP1* exhibited increased usage of distal poly(A) sites, whereas genes such as *UNC5A, RAB31*, and *LSS* preferentially use proximal poly(A) sites. Additionally, genes like *COL4A2* (Fig. 1g), *ADCY4*, and *HMGCL* (Supplementary Figs. 7g, h) were found to utilize intronic poly(A) sites in outlier individuals. Altogether, the single and multi-tissue aOutliers we identified represent the first comprehensive atlas of aberrant APA events across 49 human tissues.

aOutliers represent a unique gene set with characteristics distinct from other molecular outliers

To determine the extent of sharing between aOutliers genes and those identified as expression outlier or splicing outlier genes (i.e., eOutliers and sOutliers, respectively), we conducted a comparative analysis using the same datasets. Remarkably, we found that 74.2% of multi-tissue aOutlier genes were not detected by analysis of multi-tissue eOutliers or sOutliers (Fig. 2a and Supplementary Fig. 8a). For example, *TRIT1*, a human tRNA isopentenyl transferase 1 gene, is an aOutlier-only gene that preferentially utilizes a distal poly(A) site in outlier individuals across multiple tissues (median Z-score > 11) (Fig. 2b). This finding suggests that multi-tissue aOutliers represent a novel set of aberrant genes not detectable by traditional eOutlier and sOutlier analyses.

Further comparisons between the genomic lengths of multi-tissue aOutliers and eOutliers disclosed that aOutlier genes have significantly longer 3' UTRs than eOutlier genes (one-sided Wilcoxon rank–sum test, $P = 1.4 \times 10^{-16}$) (Fig. 2c and Supplementary Fig. 8b). In contrast, aOutlier genes have only slightly longer 5' UTRs than eOutliers (one-sided Wilcoxon rank–sum test, $P = 0.004$; Supplementary Fig. 8c), and no significant difference was observed in coding sequence length (two-sided Wilcoxon rank–sum test, $P = 0.19$). Furthermore, aOutlier genes have a lower GC-content (Fig. 2d) in their 3' UTR regions (one-sided

Wilcoxon rank–sum test, $P = 6.8 \times 10^{-6}$) than eOutlier genes. Gene ontology enrichment analysis[29] on multi-tissue aOutliers further highlighted specific biological processes and signaling pathways unique to these genes (Supplementary Fig. 9). Collectively, these data indicate that aOutliers comprise a distinct gene set with unique molecular and functional characteristics, thereby significantly distinguishing them from other types of molecular outliers.

RVs are significantly enriched among APA outliers

To assess the impact of RVs (MAF < 0.01) on aberrant APA usage, we computed odds ratios (ORs) for RVs located within varying proximity of the gene body (window size: 1 kb, 2 kb, or 10 kb) to multi-tissue aOutlier genes in outlier individuals compared to those in nonoutlier individuals. Our analysis revealed strong enrichment of nearby RVs in multi-tissue aOutliers (Supplementary Fig. 10a). Interestingly, we observed higher ORs for the enrichment of insertion and deletions (indels) than for single-nucleotide variants (SNVs) (Supplementary Fig. 10a, b). Furthermore, the degree of enrichment became more pronounced when we considered RVs located in closer proximity to the aOutlier genes or employed increased Z-score thresholds (Supplementary Fig. 10b, c).

To gain further functional insights into aOutliers-associated RVs, we first determined the proportions of these RVs with functional category using Variant Effect Predictor (VEP)[30]. A higher proportion of aOutliers-associated RVs had function annotation than nonoutliers, increasing with higher Z-score thresholds (Fig. 2e). The functional categories of aOutliers-associated RVs were largely distinct from those associated with eOutliers and sOutliers. For example, aOutliers-associated RVs are strongly enriched in the 3' UTR region (OR = 4.6 and 10.1, respectively; Fig. 2f and Supplementary Fig. 10d).

To examine whether aOutliers-associated RVs are more likely to be deleterious and potentially pathogenic, we further employed Combined Annotation-Dependent Depletion (CADD) scores[31] to stratify RVs into three groups: (1) lowly deleterious, CADD score 0–15; (2) moderately deleterious, CADD score ≥ 15 but < 25; and (3) highly deleterious, CADD score ≥ 25. Highly deleterious RVs showed significantly higher enrichment (20-fold increase for singletons and 11-fold increase for RVs with MAF < 1%; Fig. 2g) in aOutliers compared to moderately deleterious RVs (10-fold increase for singletons and 6-fold increase for RVs with MAF < 1%) and lowly deleterious RVs (2-fold increase for singletons and RVs with MAF < 1%). In total, we identified 179 rare SNVs with CADD scores ≥ 15 near 155 aOutlier genes (two-sided Fisher's exact test, $P = 5.2 \times 10^{-107}$; Supplementary Table 1). In two examples, the rare SNV rs557639120 in *SUGP1* (CADD score = 18.4, MAF in GTEx = 0.0056, and gnomAD = 0.0033) leads to an increase in distal poly(A) site usage in its 3' UTR. Similarly, the rare SNV rs759305120 in *COL4A2* (CADD score = 34, MAF in GTEx = 0.0007 and gnomAD = 0.000031) leads to preferential use of its intronic poly(A) site (Supplementary Table 1). We also identified 211 indels near 186 aOutlier genes (two-sided Fisher's exact test, $P = 1.9 \times 10^{-16}$; Supplementary Table 2), including 49 located in 3' UTR. For example, an indel variant (C > CAAAT, rs112906978) at the 3' UTR of *ACSF3* introduces a canonical "AAUAAA" motif near a poly(A) site, leading to three aOutliers (Supplementary Fig. 10e, f). Enrichment of RVs was also observed in single-tissue aOutliers across nearly all individual tissues (including SNVs and Indels) (Fig. 2h and

Supplementary Fig. 11). Considered collectively, our analyses reveal that a distinct class of RVs is significantly associated with aOutlier genes.

Rare APA variants frequently alter the 3' UTR PAS, 5' splice sites, and RNA binding proteins (RBPs) binding sites

We next investigated the potential regulatory mechanisms of aOutliers-associated RVs on aberrant APA usage. We first focused on 3' UTR aOutliers-associated RVs and performed motif enrichment analysis to determine the prevalence of RVs altering 3'end processing. Our results show that 3' UTR aOutliers-associated RVs frequently alter polyadenylation signals (PAS) and AU-rich motifs, such as "AWUAAA" and "AAUAAA" (Fig. 3a). Additionally, by using saturation mutagenesis data[32], we found that RVs associated with aOutliers have a more significant impact on poly(A) site usage than RVs associated with nonoutliers (one-sided Wilcoxon rank–sum test $P = 1.32 \times 10^{-23}$; Supplementary Fig. 12a). Notably, we observed a significant proportion of large-effect RVs (fold change, LFC > 1) associated with aOutliers compared to nonoutliers (50.3% *vs.* 6.6%; one-sided Wilcoxon rank–sum test $P = 6.1 \times 10^{-44}$; Supplementary Fig. 12b), indicating their pronounced effects on 3' UTR APA. To further experimentally validate these findings, we selected four top-ranked 3' UTR aOutlier genes by median Z-score and utilized a minigene reporter system containing reference allele and alternative allele of four rare variants in selected genes, including *MKKS* (Fig. 3b), *SUGP1*, *TP53RK*, and *ATP5F1E*. In all four cases, we could detect significant changes in the poly(A) site usage, which agreed well with the predicted effects of these RVs (Figs. 3c, d and Supplementary Fig. 13a, b).

Further investigation into multi-tissue intronic aOutliers revealed a higher incidence of RVs at 5' splice donor sites than at acceptor sites (Fig. 3e). Compared to nonoutlier RVs, aOutlier RVs are 19 to 441 times more prevalent at donor sites, and up to 47 times more prevalent at acceptor sites. Specifically, aOutlier RVs are 441 times more prevalent in the "D + 1" site and “D + 4” site and 302 times more prevalent in the "D + 2" site relative to the nonoutlier RVs. For example, RVs that alter the first nucleotide of the "GT" sequence in the intron of *COL4A2* (Fig. 1h) and the intron of *TXNRD2* lead to significant intronic APA events in these genes (Fig. 3f and Supplementary Fig. 13c). We also found that RVs altering the last base of exon 11 in *ADCY4* and exon 4 in *HMGCL* resulted in intronic APA events (Fig. 3g and Supplementary Fig. 7e, f). Based on these findings, we hypothesized that RVs affecting canonical donor sites drive intronic aOutliers. This hypothesis is also supported by our recent finding that mutations near the donor sites can promote IPA usage, potentially by blocking U1 small-nuclear RNP binding[33]. Predicting the strength of donor sites with MAXENT[34] showed a reduced strength of mutant donor sites compared to wild type (Fig. 3h, i). We then performed intronic APA minigene reporter assays for *TXNRD2* and *COL4A2* with RVs at the conserved donor sites, as well as *HMGCL* and *ADCY4* with RVs at the last base of the exons. For these assays, we cloned fragments containing full-length intronic sequences, including the donor sites, and upstream and downstream exons into the pcDNA3.1 vector. Results from 3' Rapid Amplification of cDNA Ends (3' RACE) assays indicate that all four RVs significantly increase alter IPA regulation relative to the wild-type sequence (Fig. 3j, k and Supplementary Fig. 13d, e).

Lastly, we investigated whether aOutlier-associated RVs impact other transcriptional and posttranscriptional regulation of target genes. DeepBind[35] analysis of 927 binding motifs revealed 11 significantly enriched motifs in aOutlier-associated RVs (Supplementary Fig. 14a) using randomly shuffled RVs as control, including known APA regulator *PABPN1*[36]. Furthermore, we analyzed 166 publicly accessible RBPs cross-linking immunoprecipitation sequencing (CLIP-seq) datasets from the Encyclopedia of DNA Elements (ENCODE) project[37]. We found seven RBPs's CLIP-seq data are strongly enriched with multi-tissue aOutlier RVs compared to nonoutlier RVs (Fig. 3l and Supplementary Fig. 14b), including *LARP4*, an APA regulator identified in our previous study[25], and a known APA regulator *CSTF2T*. Knockdown of the two RBPs resulted in widespread APA dysregulation (Supplementary Fig. 14c, d), affecting two aOutlier genes, *SREBF2* (Supplementary Fig. 14e) and *TOLLIP* (Fig. 3m), in which the associated RVs were inside binding peaks of *LARP4* (Supplementary Fig. 14f) and *CSTF2T* (Fig. 3n), respectively. Beyond these known APA regulators, other RBPs such as *TIA1*, *UPF1*, and *SAFB2* were also identified as potential new APA regulators (Supplementary Figs. 14g-i). Collectively, these results suggested that aOutlier-associated RVs trigger aberrant APA usage through altering PAS, splice sites, or RBP binding sites.

Inclusion of APA significantly improves functional RV effect prediction

To prioritize potentially impactful RVs for the interpretation of individual genomes, we repurposed the traditional Watershed[7] method into an APA-included version (aWatershed). This revised aWatershed model is an unsupervised probabilistic Bayesian hierarchical graphical model incorporating three RNA outlier signals, including aOutliers, eOutliers, and sOutliers, and annotations of a matched individual genome (Supplementary Table 3). The aWatershed model can allow us to quantify the posterior probability of an RV leading to a functional effect on APA usage (Supplementary Figs. 15a, b; Materials and Methods). To evaluate the aWatershed performance on the GTEx v8 data, we used held-out individual pairs with the same RVs as the evaluation dataset. By applying aWatershed prediction on the first individual of each pair and evaluating this prediction using the outlier status of the second individual as a label, we observed that our model significantly outperforms both the RIVER (RNA-informed variant effect on regulation) model[8], a simplification of the Watershed model which integrates genomic features with aOutlier signals alone, and the GAM (genomic annotation model), a generalized logistic regression model based on genomic features alone (Fig. 4a and Supplementary Fig. 15c). 93% of aWatershed prioritized RVs have low posterior probabilities in the GAM (Fig. 4b), highlighting the importance of transcriptomic aOutlier signals in functional RVs prioritization. Moreover, aWatershed successfully captures the regulatory mechanisms underlying the effect of RVs on aOutlier signal (Fig. 4c). Strikingly, the integrated aWatershed model can prioritize RVs associated with 73.8% of aOutliers, in contrast to only 12.4% when relying on the genomic features alone (Fig. 4d).

Next, we used the saturation mutagenesis data[32] to further evaluate the efficacy of aWatershed in prioritizing RVs with significant effects on APA regulation. In this analysis, we stratified RVs into two

groups based on aWatershed APA posterior probabilities and compared poly(A) usage between them. We found that RVs in the group with high posterior probability had significantly larger effects on APA than those in the low posterior probabilities group (Fig. 4e, f), suggesting our aWatershed model is effective in identifying RVs with substantial APA effects. Furthermore, our analysis revealed that aWatershed successfully identified many functional RVs overlooked by the previous variant prediction model[38], as exemplified by two RVs in *RPL13A* and *PAAF1*, respectively (Supplementary Fig. 15d). Overall, aWatershed prioritized 1,799 RVs predicted to impact 278 APA genes (Supplementary Table 4). Interestingly, there was minimal overlap between RVs impacting APA and those affecting gene expression or splicing, as only 60 of these 1,799 RVs were common to those categories. For example, the RV rs191575428 within the 3' UTR of *MTHFD2*, which exhibited a high aWatershed APA posterior probability of 0.997 based on aOutliers, showed considerably lower posterior probabilities for expression and splicing (0.055 and 0.008, respectively). This variant is associated with 3' UTR lengthening in outlier individuals without changing gene expression levels (Supplementary Fig. 15e, f). Further extending the aWatershed model to prioritize tissue-specific functional RVs by integrating genomic features with single-tissue aOutliers signals, we observed that the tissue-aWatershed model outperforms both the tissue-RIVER model and tissue-GAM model (Supplementary Figs. 16–17). In summary, by leveraging these aOutliers, we have implemented a robust Bayesian hierarchical variant effect prediction model aWatershed that effectively prioritizes rare functional variants with significant effects on APA regulation.

Analysis of aOutliers prioritizes RVs impacting complex traits and diseases

To test the hypothesize that aWatershed RVs could be used to interpret the complex traits and diseases, we first examined the 278 genes prioritized by aWatershed and cross-referenced with genes annotated in the Online Mendelian Inheritance in Man (OMIM) database[39]. We identified 21.2% of the prioritized genes were well-known disease genes (Supplementary Fig. 18a). For example, we identified a prioritized RV, rs79940214, associated with *MKKS* (Supplementary Fig. 18b), which encoded a centrosome-shuttling protein and was associated with many genetic diseases, including McKusick-Kaufman syndrome (OMIM id: 236770)[40,41] and Bardet-Biedl syndrome 6 (OMIM id: 605231)[42,43]. Another example is one prioritized intronic RV, rs76984877, that is associated with gene *EXT2* (Supplementary Fig. 18b), which was associated with hereditary multiple exostosis, type 2[44,45]. We also identified five prioritized RVs associated with gene *BCR* (Supplementary Fig. 18b), which has been frequently reported to be associated with chronic myeloid leukemia[46,47].

We further cross-referenced aWatershed-prioritized RVs with trait variants from 1,234 well-powered GWAS summary statistics from UK Biobank (UKBB) and literature (Supplementary Table 5), resulting in 1,385 RVs associated with 171 aOutlier genes in 1,186 traits (Supplementary Table 6). We focused on the subset of 623 traits, which also have evidence of colocalization with 3'aQTLs (Supplementary Table 7). Notably, aOutlier prioritized RVs fell in or nearby genes had evidence of colocalization with 3'aQTLs having larger trait effect size than the non-colocalized RVs ($P = 0.0014$, one-sided Wilcoxon rank–sum test; Fig. 5a). We also conducted a permutation test to determine whether these prioritized RVs exhibit

larger effect sizes on these complex traits and diseases. We found that the mean odds of aOutlier-prioritized RVs had a more significant effect size than non-prioritized RVs ($P = 2.5 \times 10^{-15}$, one-sided and paired Wilcoxon rank–sum test; Supplementary Fig. 19a). To exemplify the larger effect size in aOutlier-prioritized RVs, we focused on two traits: height related traits (UKBB trait ID: 50_irnt and 20015_irnt) and high blood pressure (UKBB trait ID: 6150_4). This analysis revealed a significant shift in the odds favoring RVs with higher aWatershed posterior probabilities over those with lower ones ($P = 1.6 \times 10^{-9}$ and $P = 2.3 \times 10^{-54}$, respectively; one-sided Wilcoxon rank–sum test; Fig. 5b, c; Supplementary Fig. 19b). In the case of height related traits and high blood pressure, these aOutlier prioritized RVs had larger effect sizes on the trait than other variants within a 1Mb of the RV, including RVs prioritized by eOutliers or sOutliers. Notably, for the height related traits, the RV (rs112567314), located in the intron of *CUL3*, had a greater effect size than other variants within 1 Mb and RVs prioritized by eOutlier or sOutlier (Fig. 5d and Supplementary Table 6). Similarly, for high blood pressure, the RV (rs893929), located in the intron of *USP38*, also had a greater effect size than 99.6% of variants within 1 Mb, including the nearest trait-associated significant variants as well as eOutlier or sOutlier RVs (Fig. 5e). Collectively, our results demonstrate the capability of aWatershed in prioritizing RVs with large effect sizes on APA, significantly impacting complex traits and diseases.

Strong convergence between rare and common variants on DDX18 links APA regulation with cancer susceptibility

Emerging evidence suggests potential interactions between rare and common variants in affecting the same disease genes[48–50]. As expected, we also observed the strong convergence effect on 3' UTR APA regulation between RVs and common variants (Supplementary Fig. 20). To further mechanistically examine their convergence effects on disease, we focused on aWatershed prioritized RVs and their associated genes. We found 126 out of the 278 aOutlier RV associated genes were also identified as susceptibility to disease risks, including cancer risks through 3'aQTLs in our gene-based association studies[51,52] (Fig. 6a). Among the top-ranked APA genes that were prioritized by both RV and 3'aQTLs analyses (Fig. 6b), we noticed several highly constrained genes (pLI score > 0.9), and we particularly focused on the gene *DDX18*, a member of the DEAD-box RNA helicase family, that was identified as an APA-mediated susceptibility gene across many cancer types[53,54]. Moreover, CRISPR-Cas9 based gene essentiality screens also demonstrated that *DDX18* has an essential role in cancer cell proliferation[55,56] (Fig. 6c). Examining our 3'aQTLs data revealed significant associations between common variants and 3' UTR APA of *DDX18* across tissues, with the most significant one was found near the 3' end (Fig. 6d and Supplementary Fig. 21a-c). Intriguingly, an aWatershed prioritized RV, rs1680042046, located near the distal poly(A) site of *DDX18*, was identified in the outlier individual (Fig. 6e, f). This RV alters the hexamer motif "AUUAAA" to "AUUAAG" (Supplementary Fig. 21b) and has a highly deleterious effect (CADD = 17.5) (Supplementary Table 1).

To further experimentally validate the convergence effect of RVs and common variants on *DDX18*, we designed minigenes introducing the APA variants by PCR-based site-directed mutagenesis in HEK293T

and MCF7 cells (Fig. 6g). We then performed 3' RACE to quantitatively evaluate the effect of the common variant (rs1052628; A > G) alone, the RV (rs1680042046) alone, or their joint effect on APA. We first mutate the reference A allele to the alternative G allele for either RV or the common variants. In HEK293T cells, this mutation decreased the use of the distal poly(A) site (dPAS) for both the common variant and RV (two-sided Student's t-test $P = 1.5 \times 10^{-6}$ and 2.3×10^{-7}; Figs. 6h, i), indicating that both variants indeed trigger *DDX18* APA regulation. A similar APA effect was also observed in MCF7 cells (Fig. 6j). To further assess the functional roles of *DDX18* APA regulation in breast cancer cells, we measured *DDX18* protein level using luciferase reporter assays and assessed the effect of gene silencing on the proliferation of MCF7 cells proliferation. We observed lower luciferase activities in the short 3' UTR isoform of *DDX18* and the reporter containing RV or both RV and common variant (Supplementary Fig. 21d, e). Knockdown of *DDX18* in MCF7 results in inhibition of cell proliferation (Supplementary Fig. 21f, g). Collectively, these findings highlight the critical role of rare variants in understanding the risk of common diseases and offer a novel approach to linking functional rare variants to complex diseases.

Discussion

The human genome contains a plethora of rare genetic variants whose functional effects and underlying molecular mechanisms are challenging to interpret. In this study, we introduce the aOutlier as an emerging molecular phenotype reflecting aberrant 3' UTR or intronic APA usage across multiple samples. aOutlier can be used to identify functional rare APA variants. By analyzing population-scale transcriptomics data using our DaPars2[25,26] and IPAfinder algorithms[18], we identified 1,534 multi-tissue aOutliers based on European individuals. These aOutlier genes exhibit unique molecular features, such as genomic lengths and GC-content, setting them apart from other molecular outliers, such as eOutliers and sOutliers. Importantly, aOutliers can aid in identifying a distinct class of rare functional variants. We observed that aOutliers-associated RVs are more likely to be deleterious and are highly enriched in outlier individuals. Mechanistically, these aOutlier-associated RVs can modulate APA usage by either altering PAS, AU-rich elements, or splice donor sites, as confirmed by saturation mutagenesis data and 3' RACE experiments.

To further enhance the utility of our aOutlier atlas, we adapted a Bayesian hierarchical prediction model (aWatershed) by incorporating genomic features with multiple functional signals, including aOutliers, eOutliers, and sOutliers. This integration aims to predict the probability of RV leading to aberrant APA usage. Notably, our aWatershed model outperformed models trained only on genomic features or those combined with aOutlier signals alone. Moreover, aWatershed-prioritized RVs exhibited more significant effects on APA regulation than non-prioritized RVs. The predictive power of aWatershed was validated using GWAS summary data from the UKBB, showing that aWatershed-prioritized RVs had larger trait effect sizes than non-prioritized RVs, as exemplified by RVs near *POLR2L* and *ATP5F1D* associated with height and BMI, respectively.

Interestingly, we observed a significant proportion of intersection between aOutlier transcripts and 3'aQTL associated transcripts in matched tissue, suggesting the potential interplay of common variants and RV

in APA regulation, similar to previous findings in gene expression studies[48,49,57,58]. Additionally, a rare deletion 16p11.2 and common variants in chromosome 16p modulate downstream gene expression and affect the risk for autism[48]. Moreover, using minigene reporters and 3' RACE assays, we demonstrated the potential additive effect of rare and common APA variants on *DDX18* 3' UTR regulation. We further demonstrated that the regulation of *DDX18* 3' UTR contributes to *DDX18* protein expression level, which is tightly linked to breast cancer cell proliferation. In summary, our study identifies a novel set of rare functional variants that influence APA and connects these RVs to human trait phenotypes, providing valuable information for the identification of novel genes associated with increased disease risk.

Materials and Methods

GTEx data collection and processing

We downloaded both RNA-seq raw sequencing data and whole-genome genotype data of the v8 release of the GTEx project from dbGAP (accession: phs000424.v7.p2). Expression outlier (eOutlier) and splicing outlier (sOutlier) data, and the metadata of samples (filename: GTEx_Analysis_v8_Annotations_SampleAttributesDD.xlsx) and subjects (filename: GTEx_Analysis_v8_Annotations_SubjectPhenotypesDD.xlsx) were downloaded from GTEx Portal (https://gtexportal.org/home/). The GTEx RNA-seq dataset contains 17,832 samples representing 54 biological tissues collected from 838 donors. For this study, we included data from 49 tissues, each with at least 70 samples. Original GTEx RNA-seq reads were aligned with the human genome (hg38/GRCh38) using STAR v.2.7.3a[59], with the following alignment parameters: outSAMtype, BAM; SortedByCoordinate; outSAMstrandField, intronMotif; outFilterMultimapNmax, 10; outFilterMultimapScoreRange, 1; alignSJDBoverhangMin, 1; sjdbScore, 2; alignIntronMin, 20; and alignSJoverhangMin, 8. The aligned BAM files were sorted and further converted to bedGraph format using BEDTools v.2.27.1[60]. The genotype data in VCF format (filename: GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_838Indiv_Analysis_Freeze.vcf.gz) was processed with vcftools v.0.1.13 to calculate MAF across all subjects and extract allele information for each variant.

3' UTR APA and intronic APA quantification

To quantify the 3' UTR APA, we analyzed alignment files in BAM format using the DaPars2 algorithm. We followed the workflow implemented in our 3'aQTL analysis[25,61]. Briefly, the BAM files were firstly transformed to bedGraph format with a bin size of 1, which records the read coverage of each position in the genome. Before analyzing APA, we downloaded the gene annotation file containing all transcripts of genome build hg38 in RefSeq database through the UCSC Genome Browser, from which we extracted 3' UTR region of each transcript using script "DaPars_Extract_Anno.py". The DaPars2 algorithm then detects the proximal poly(A) site in the 3' UTR region of each transcript and calculates the relative usage of the distal poly(A) site by the script “Dapars2_Multi_Sample.py” for all samples in each of the 49 tissues. This is indicated as the Percent of Distal Poly (A) site Usage Index (PDUI) only if a proximal poly(A) site is detected. For intronic APA detection and quantification, we used IPAfinder[18], which is a python-based tool

that enables *de novo* identification and quantification of intronic APA (IPA) events using RNA-seq data. IPAfinder can identify potential IPA sites and calculate the Intronic poly(A) site Usage Index (IPUI), which represents the proportion of total transcripts that are intronic-polyadenylated for each intronic APA event[18,33,62]. BAM files were analyzed together by IPAfinder and separated by tissues.

Covariate correction and normalization

To avoid batch effects and unobserved confounders in each tissue, we adjusted the sample genotype and APA usages with known covariates, such as population structure, sex, and sequencing platform. Briefly, for genotype data, we first removed sites marked as "wasSplit" from the GTEx analysis freeze variant call format (VCF) using BCFtools v.1.10.2. We further applied the PEER model[63] with sex, age, sequencing platform, and the top five genotype principal components as known covariates to estimate a set of latent covariates for PDUI or IPUI values in each tissue. The number of PEER factors was optimized based on tissue sample size, as suggested by the GTEx Consortium; 15 PEER factors were chosen for sample sizes < 150, 30 PEER factors were selected for sample sizes ranging from 150 to 250, and 35 peer factors were chosen for sample sizes > 250. Before running the PEER model for inferring hidden covariates, PDUI/IPUI values in each tissue were quantile normalized to remove batch effects.

APA outlier calling

After inferring the hidden covariates for each tissue, we calculated PDUI/IPUI residuals by regressing out inferred PEER factors and known covariates, including population structure, sex, and sequencing platform, using the function "PEER_getResiduals". In each individual tissue, we obtained normal-distributed $Z\left(g,t\right)$ score for each gene (g) in the tissue (t) by scaling the PDUI/IPUI residuals across samples with the following equation, $X_r\left(g,t\right)$ denotes the residuals of PDUI/IPUI values, $\bar{X_r}\left(g,t\right)$ and $sd\left(X_r\left(g,t\right)\right)$ represent the mean and standard deviation of the residuals across samples, respectively:

$$Z\left(g,t\right)=\frac{X_r\left(g,t\right)-\bar{X_r}\left(g,t\right)}{sd\left(X_r\left(g,t\right)\right)}$$

We defined two types of aOutliers in the current study. One is single-tissue aOutlier, which is called from a single tissue based on the Z-score of each gene in that tissue. When the absolute Z-score of an individual exceeds a threshold of three for a gene, then the individual is called a single-tissue aOutlier for that gene. The other is multi-tissue aOutlier, for which we calculated a median Z-score[7,8] for each APA event across all tissues when data were available and restricted our analysis to individuals with APA measurements in at least five tissues. Multi-tissue aOutliers were defined as those with an absolute median Z-score > 3. The same threshold was used for eOutlier and sOutlier calling[7,8]. Our method allowed that one gene could have multiple aOutlier individuals, and one individual could also be aOutliers of multiple genes. To account for situations in which widespread aberrant APA might occur in an individual due to non-genetic influences, we removed 11 individuals in which the proportion of tested genes identified as multi-tissue

aOutliers exceeded 1.5 times the interquartile range of the distribution for aOutlier gene proportion across all individuals. These 11 individuals were marked as global outliers.

Estimation of replication rates of aOutliers

To estimate the replication rate of aOutliers between different tissues, we selected one of the 49 GTEx tissues each as discovery tissue, and compared aOutliers detected in it with those of the other tissues. For replication rate calculation, we only considered the shared aOutlier genes in both tissues and an aOutlier to be replicated only when the gene and individual of the aOutlier matched between the compared tissue pairs. For multi-tissue aOutliers replication, we used the cross-validation method described in a previous study[8] to estimate their replication rate. In brief, the 49 human tissues were separated into two groups, one group with 39 tissues as the discovery group, the other group has the remaining ten tissues as the replication group. Each time we randomly sampled t (t = 10, 15, 20, 25, 30) tissues from the discovery group and called multi-tissue aOutliers in the discovery group using a Z-score threshold of 3 in at least five tissues as described above. Then we estimated the replication rate as the proportion of multi-tissue aOutliers in the discovery group with an absolute median Z-score 2 or 3 in the replication group. We also computed the expected replication rate by randomly selecting individuals in the discovery group with at least five tissues that have APA usage for the gene and determined the replication status in the replication group. For each discovery group size (t), we repeated this process 10 times.

RV annotation

We defined RVs as those with MAF < 1% within the GTEx European individuals and with MAF < 1% in non-Finnish Europeans within gnomAD[64]. Singletons were defined as RVs with minor allele only presents once in GTEx European individuals and were extracted using vcftools. The annotation of RVs was performed by Ensembl VEP (release 104), which assigned 36 different annotation terms to each RV, including protein-coding gene position (e.g., "splice_donor" "splice_acceptor," "frameshift”) and regulatory regions (e.g., "TFBS_ablation", "TF_binding_site"). Annotation terms were grouped into one of the four classes based on predicted impact: "High", "MODERATE", "MODIFIER" and "LOW". The high-impact one was used for variants assigned with two or more annotations. In addition to 36 VEP annotations, we added two other annotations to each RV; "PAS region" describes variants located within 50 bp upstream of the annotated PAS, and "PAS signal" refers to variants located at the PAS motif "AAUAAA" and its additional 14 variants ("AUUAAA", "UAUAAA", "AGUAAA", "AAAAAA", "AACAAA", "AAGAAA", "AAUAUA", "AAUACA", "CAUAAA", "UUUAAA", "ACUAAA", "AAUAGA" and "GAUAAA"). We also used genomic annotations of variants extracted from CADD v.1.5 release (http://cadd.gs.washington.edu/download).

RV enrichment analysis

We examine the enrichment of Rare Variants (RVs), including single-nucleotide variants (SNV) and small insertion and deletion (indel) near aOutlier genes. Only genes with at least one aOutlier individual were considered, and the remaining individuals for the same genes were treated as nonoutlier controls. We first counted the RVs present within 1kb, 2kb, or 10kb of the outlier genes in both outlier and control

individuals and built a 2 × 2 contingency table for each of the flanking region, containing the number of aOutliers with RVs, the number of nonoutlier controls with RVs, the number of aOutliers without RVs, and the number of controls without RVs. We then calculated Odds Ratios (ORs), *P* value, and 95% confidence interval (CI) using Fisher's exact test in R base package. We grouped variants into four groups based on their MAF (0–1%, 1–5%, 5–10%, and 10–25%), and performed enrichment analysis for each group. We also conducted enrichment for RVs that stratified by VEP annotations and CADD scores as described above.

Enrichment analysis for RVs that influence PAS and AU-rich motifs

To identify potential regulatory variants associated with aberrant APA events, we defined RVs located within the gene body or in the 10-kb region surrounding outlier genes in outlier individuals as aOutlier RVs. Those in nonoutlier individuals in the same region were classified as nonoutlier RVs. For each aOutlier RV and nonoutlier RV located in the 50-bp region upstream (PAS region) of the poly(A) sites annotated in PolyA_DB V.3.2[65,66], we extracted its upstream and downstream 5 base pairs sequences and examined whether it matched with one of the 15 known PAS motifs by using script "dna-pattern" in RSAT (https://github.com/rsa-tools). We then summarized all tested RVs and conducted PAS motif enrichment analysis using Fisher's exact test, which determined the odds ratios (ORs) and 95% confidence intervals (CIs) for each PAS motif. To perform enrichment analysis at the 12 known AU-rich motifs, we restricted RVs to those within the 100 bp flanking the annotated poly(A) sites. We then counted RV enrichment analysis for each of the AU-rich motifs using the same method as for PAS motifs.

Identification of aOutlier RVs enriched RNA motifs

We focused on multi-tissue aOutlier associated RVs located within the gene body region, which spans from 3 kb downstream of the transcription start site (TSS) to the end of the gene. We extracted the 3 base pairs of sequences flanking each RV from both sides. Next, we used DeepBind v0.11[35] to score these 7-mer sequences using 617 pre-built models, including 515 transcription factors and 102 RNA-binding proteins from Homo sapiens. For each 7-mer sequence, we selected the top three motifs with a DeepBind score of at least 0.1. To validate the enrichment of RVs in predicted binding motifs, we created a control set of RVs by randomly shuffling the genomic locations of multi-tissue aOutlier associated RVs within the same gene body regions. We used Fisher's exact test to estimate the level of enrichment.

Identification of aOutlier RVs enriched RBPs

We obtained CLIP-seq data for 166 RNA-binding proteins (RBPs) from the Encyclopedia of DNA Elements (ENCODE) data portal for HepG2 and K562 cells. We only considered significant binding peaks with *P-values* < 0.01, shared by two biological replicates for each RBP. To assess the enrichment of aOutlier RVs in RBP binding peaks, we selected RVs associated with multi-tissue aOutliers within gene body regions representing the region of 3 kb downstream of the transcription start site (TSS) to the end of the gene. We created a control RV set by randomly shuffling the genomic locations of multi-tissue aOutlier associated

RV set within the same gene body regions. We then counted the RVs in binding peaks of each RBP using bedtools. Finally, we compared the RVs between the two sets using Fisher's exact test to determine the enrichment.

Development of a Bayesian prediction model that integrates APA outlier signals

To prioritize rare functional variants with significant impact, we improved the Watershed Bayesian hierarchical model by incorporating APA outlier signals with other layers of transcriptomic outlier signals and genomic annotations. The improved model called aWatershed, includes a layer of genomic annotation features (*G*) which denotes the 40 observed genomic features aggregated over all RVs in the outlier individual that are within 10 kb region of the gene, a fully connected conditional random field (CRF) layer (*Z*) represent the unobserved regulatory variables for each of the three transcriptomic outlier signals (APA, mRNA expression, and splicing), and a layer of variables (*E*) representing the observed outlier status of each transcriptomic data type. The three layers were linked by the following conditional distributions:

$$Z|G \sim CRF\left(\alpha, \beta_{APA}, \beta_{RNA}, \beta_{Splice}, \theta\right),$$

$$E_k|Z_k \sim Categorical(\varphi_k) \forall k \in K,$$

$$\varphi_k \sim Dirichlet(C, \ldots, C),$$

$$\beta_k \sim Normal\left(\frac{1}{\lambda}\right), (k = APA, Expression, Splice),$$

Where K represents the three outlier signals (APA, Expression, and Splice), β_k are parameters defining the contribution of the 40 genomic features to the CRF of the three outlier signals, α defines the intercept of the CRF for each outlier signal, θ represent parameters defining the edge weights between pairs of the three outlier signals, ϕ_k are the paramters denoting the categorical distributions of each of the three outlier signal, and C and λ are hyper-parameters.

To train and evaluate aWatershed, we utilized all gene-individual pairs that have at least one of the three multi-tissue outlier signals, which are defined as the absolute value of Z-score greater than 3 or *P-value* less than 0.0027 for splicing outliers, measured in GTEx v8 data. We also used a set of 38 binary and continuous genomic annotation features aggregated across all rare variants within the 10-kb region, flanking each gene. We then trained aWatershed to learn edge weights connecting the three transcriptomic outlier signals, weights representing the contribution of each genomic annotation for each type of outlier signal, and other parameters, as described previously[7,8].

To evaluate aWatershed, we selected pairs of individuals with the same set of rare variants associated with the same gene (known as "N2pair") from the training dataset. We estimated the posterior probability

of a functional rare variant in the first individual of the pair and used the outlier status of the second individual as a label for evaluation. We also trained and evaluated the genomic annotation model on each layer of the three transcriptomic signals to determine whether the integration of transcriptomic outlier signals contributes to the prediction of rare functional variants. We compared the results to those obtained from the aWatershed model. After evaluation, we utilized the aWatershed prediction model to calculate posterior probabilities.

3'aQTL mapping across 49 GTEx tissues

Genetic associations between GTEx common variants within 1 Mb of each gene and PEER-corrected APA usage were mapped by Matrix eQTL[67], as described in our previous study[25]. Known covariates, including sex, RNA integrity number, platform, top five genotype principal components, and unobserved covariates inferred from PEER, were used during 3'aQTL mapping with Matrix eQTL. The number of PEER covariates for each tissue was used as suggested by the GTEx Consortium. We performed 1,000 rounds of permutation to obtain empirical *P*-values for each gene, which were then adjusted using the R package qvalue.

Colocalization analysis between GWAS summary statistics and 3'aQTL

We conducted colocalization analysis comparing GWAS summary statistics from the UKBB and literature and 3'aQTL summary data from 49 human tissues using the coloc v.5.1.0.1 package in R[68]. Only GWAS summary data for traits with at least 10,000 cases (binary traits) or 50,000 participates (continuous trait) and with at least 10 SNPs overlapped with aOutlier-associated RVs were kept, which resulted in 1,186 well-powered traits. We extracted the sentinel SNPs for each GWAS trait, defined as GWAS SNPs with $P < 5 \times 10^{-8}$, located at least 1 Mb away from more significant variants. We then searched for colocalizing signals within the 1-Mb region surrounding each sentinel SNP. The coordinates from 3'aQTL summary data were converted from human genome build 38 (hg38) to build 37 (hg19) by CrossMap software[69] to match the version used in all GWAS summary statistics. As defined by the coloc method, five posterior probabilities under five different null hypotheses were calculated. In detail, PP0 represents the null model of no association. PP1 and PP2 represent the probability that causal genetic variants are associated with disease signals or 3'aQTL only. PP3 represents the probability that the genetic effects of trait signals and 3'aQTL are independent, and PP4 represents the probability that trait signals and 3'aQTL data share causal SNPs. The current study classified colocalized events as those with PP4 > 0.75.

3' UTR APA transcriptome-wide association study (3'aTWAS) analysis

We used APA quantitative data that was previously used for 3'aQTL mapping[25,52,61] and genotype data of individual genomes from whole genome sequencing (WGS) of GTEx consortia to construct 3'aTWAS model using FUSION[70] for each of the 49 human tissues. To avoid the effects of confounders, well-established factors used in 3'aQTL mapping, including gender, sequencing platform, and other covariates, were incorporated to adjust APA usages. To build the TWAS model, four different models embedded in

FUSION were used for weight calculation, including best linear unbiased predictor (blup), elastic-net regression (enet), lasso regression (lasso), and single best eQTL (top1). Subsequently, the cross-validation approach was employed to choose the optimal 3'aTWAS model for each gene. Of note, only genes exhibiting significant heritability estimates (*cis-*h^2) (Bonferroni-corrected P < 0.05) were retained for subsequent analysis. The built models were then applied to GWAS summary statistics for gene-based association analysis, and a significant association was defined by the FDR < 0.05. The disease risk genes identified by 3'aTWAS in two or more tissues were used for further analysis.

Prioritization of trait-associated RVs

To determine the frequency with which randomly selected aWatershed-prioritized RVs exhibit larger GWAS effect sizes than matched non-prioritized RVs, we conducted a random sampling test (n = 1000) on all RVs using posterior probabilities obtained from the aWatershed prediction model and effect sizes from UKBB GWAS summary statistics. We used aWatershed-prioritized RVs based on aOutlier signals, matched non-prioritized RVs, as well as GWAS effect sizes, gene IDs, and prioritized scores as input data. We defined matched non-prioritized RVs as those with a posterior probability of < 0.1 and MAF within ± 0.001 of the selected prioritized RVs in the UKBB cohort.

For each gene in each trait, we randomly selected one prioritized RV and one matched non-prioritized RV and then identified the one with the largest absolute GWAS effect size in the pair. By summarizing all genes in the trait, we computed the odds of observing a prioritized RV with a larger absolute effect size than a non-prioritized RV across all genes. To generate a null distribution of odds, we repeated this process for matched non-prioritized variants only and randomly selected and compared two non-prioritized RVs for each gene.

Cell culture

HEK293T and MCF7 cells were purchased from the Cell Bank of the Type Culture Collection at the Shanghai Institute of Biochemistry & Cell Biology, Chinese Academy of Science. Cells were maintained in Dulbecco's modified Eagle medium (DMEM; Invitrogen, #11960044) supplemented with 10% fetal bovine serum (Gibco), 100-μg/ml streptomycin, and 100-units/ml penicillin at 37°C in a humidified incubator with 5% CO_2.

Plasmid construction

All primers used in this study are listed in Supplementary Table 8. For intronic APA (IPA) minigenes, the candidate intron and its flanking exons were amplified from genomic DNA as wild-type fragments. For 3' UTR APA minigenes, the 3' UTR of each gene was amplified from genomic DNA as wild-type fragments, and mutations were introduced by PCR-based site-directed mutagenesis. In short, genomic DNA from HEK293T and MCF7 cells was amplified by PCR using primers to generate two 20–25 bp overlapping fragments containing a mutant site. The IPA wild-type and mutant fragments were subcloned into the EcoRI and BamHI sites of the pcDNA3.1 vector, while 3' UTR APA wild-type and mutant fragments were

subcloned into the XhoI and PmeI sites of the mpCHECK2 vector by the One Step Cloning Kit (Vazyme). Two sets of predesigned shRNAs from Sigma against *DDX18* were used to clone into pLKO.1-puro vector.

Transient transfection

For transient transfection, HEK293T and MCF7 cells were plated in a 2-ml culture medium at 6×10^5 cells/well in six-well plates. After 24 h of culture, cells were transfected with 2 μg of wild-type or mutant minigene plasmid using Lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. The culture medium was replaced at 6 h post-transfection, and cells were harvested for RNA extraction at 48 h post-transfection. Total RNA was extracted using TRIzol reagent (Invitrogen), according to the manufacturer's instructions, and cDNA was synthesized using the FastKing RT Kit (Tiangen, KR116) with the S-CDS primer. All cDNA was diluted 4-fold in nuclease-free double-distilled H_2O for further use.

3' RACE

The total length of 3' UTR was identified and amplified from the total RNA of NCI-H1299 cells by 3' RACE using the HiScript-TS 5'/3' RACE Kit (Vazyme, RA101) following the manufacturer’s protocol. 3' RACE was performed using the S-PCR primer and pcDNA3.1-F or mpCHECK2-F primer to distinguish minigene RNA from endogenous RNA, respectively. The 3' RACE PCR products were separated by gel electrophoresis, and excised bands were purified for Sanger sequencing using the Zymoclean Gel DNA Extraction kit. Cleaned DNA fragments were cloned into the PCE2 vector using the 5 min TA/Blunt-Zero Cloning Kit (Vazyme, C601) and bidirectionally sequenced with M13 forward and reverse primers. At least five colonies were sequenced for every gel product that was purified. Primer sequences are listed in Supplementary Table 8.

Dual-luciferase reporter assay

MCF-7 cells were seeded 1 day prior to transfection. The Renilla luciferase in the mpCHECK-2 vector was transfected into cells using Lipofectamine 3000 Transfection Reagent (Invitrogen, cat#: L3000015) according to the manufacturer's instructions. Forty-eight hours post-transfection, firefly, and renilla luciferase activities were measured by Dual-Luciferase Assay System (Promega, #E1980) on a BioTek Synergy H1 plate reader with full waveband. Each assay was measured in three independent replicates.

Cell viability and proliferation assays for shRNA-mediated knockdown

shRNA-expressing lentivirus was produced with the third-generation packaging system in human embryonic kidney (HEK) 293T cells. For lentivirus infection, target cells (MCF7) were seeded in a 6-well plate 16–18 h before infection and were grown to 70–80% confluency upon transduction. The culture medium was removed, and cells were incubated with virus supernatant along with 8 μg/ml polybrene. Puromycin was applied to kill non-infected cells 2days after infection. After two days of selection, when non-infected control cells were all dead, surviving cells were split and maintained with the same concentration of puromycin. Cells were trypsinized, resuspended at 1×10^4 cells/ml, and seeded in 96-

well plates, with each well containing 100ul medium of 1×10^3 cells. Cell viability and proliferation were determined using CCK8 assays (Yeasen, cat#: 40203ES76) at designated time points (day 1, day 3, day 5, and day 7) by measuring the absorbance at 450 nm, following the manufacturer's instructions. Values were obtained from three replicate wells for each treatment and time point. Results were representative of three independent experiments.

The comprehensive data portal for aOutliers

We have established a database along with a web interface called rareAPA (http://bioinfo.szbl.ac.cn/rareAPA/index.php) on a standard LAMP (Linux + Apache + MySQL + PHP) system, which serves as a comprehensive resource presenting detailed and comprehensive information on rare APA events and their associated RVs. All these data in the rareAPA were stored in MySQL (www.mysql.com). The interactive web pages were implemented using HTML, CSS, JavaScript, and PHP languages (www.php.net), with several JavaScript libraries (JQuery.js, DataTable.js, and IGV.js) and Bootstrap framework (a popular framework for developing interactive websites) on Red Hat Linux powered by an Apache server (www.apache.org). This data portal is valuable for exploring aOutliers and their associated functional rare variants. With rareAPA, users can search, browse, and visualize important information on aOutliers in 49 human tissues. Users can search by gene or tissue name and scrutinize rare APA events among individuals in each tissue. Additionally, users can also visualize aOutliers using a scatter plot or explore them through a genome browser. Furthermore, rareAPA provides a curated list of prioritized RVs using the aWatershed algorithm, allowing users to examine rare variants and their aWatershed posterior scores. Additionally, rareAPA offers batch downloading of all single-tissue aOutliers and multi-tissue aOutliers. The rareAPA is freely available online without registration or login requirements.

Declarations

Code availability

DaPars2 is available at https://github.com/3UTR/DaPars2, and IPAFinder can be accessed through https://github.com/ZhaozzReal/IPAFinder. The codes for mapping 3'aQTL are available at https://github.com/3UTR/3aQTL-pipe. The custom scripts and source codes for data analysis relevant to this study are available, under the MIT license, at Github repository: https://github.com/Xu-Dong/rareAPA and Zenodo: https://doi.org/10.5281/zenodo.10576656.

Data availability

The raw data of whole transcriptome and genome sequencing data from the GTEx project V8 are available at the database of Genotypes and Phenotypes (dbGaP) under the accession number: phs000424.v7.p2 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2][71]. All processed GTEx data, including gene expression outlier (eOutlier) and splicing outlier (sOutlier), are available via the GTEx portal (http://gtexportal.org). GWAS summary

statistics used in this study were obtained from UK Biobank GWAS (https://www.nealelab.is/uk-biobank), Finn Gen (https://www.finngen.fi/en), and JENGER (http://jenger.riken.jp). The details about the GWAS summary statistics are listed in Supplementary Table 5. Genomic and functional annotations of rare variants are available via the Combined Annotation Dependent Depletion (CADD v1.5, https://cadd.gs.washington.edu/), and gnomAD v3.1(https://gnomad.broadinstitute.org/). The crosslinking and immunoprecipitation (CLIP) assay data for RNA binding proteins used in this study are available at The Encyclopedia of DNA Elements (ENCODE, https://www.encodeproject.org/). The data described in this study are freely available for querying, visualizing, and downloading at http://bioinfo.szbl.ac.cn/rareAPA/index.php, a website portal dedicated to rare APA.

Author Contributions

L.L. T.N., and W.L. conceived and supervised the project. X.Z., and Z.Z. performed the bioinformatics analysis with the help from K.X., and H.C. X.Z. constructed the website. Y.C. performed the experiments with the help from Z.W. and S.C., X.Z., T.N., W.L., and L.L. interpreted the data and wrote the manuscript. G.W. and S.X. reviewed and revised the manuscript.

Competing Interests

The authors declare no competing interests.

Acknowledgments

We thank Dr. Jian Yang from Westlake University for providing feedback on the manuscript. We also thank members of the Li laboratory for helpful discussions. This work was supported by the National Natural Science Foundation of China (no. 32100533, 32370721, 32288101, 32030020) and startup funds from Shenzhen Bay Laboratory to L.L. We also thank Qin Wang at the Shenzhen Bay Laboratory Supercomputing Center for high-level computing support and the Medical Science Data Center of Fudan University.

References

1. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
2. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-3 (2012).
3. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
4. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-4 (2012).

5. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
6. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527-532 (2021).
7. Ferraro, N.M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**(2020).
8. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239-243 (2017).
9. Hernandez, R.D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nat Genet* **51**, 1349-1355 (2019).
10. Fresard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25**, 911-919 (2019).
11. Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol* **11**(2019).
12. Tian, B. & Manley, J.L. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18-30 (2017).
13. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171-194 (2017).
14. Berkovits, B.D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363-7 (2015).
15. Di Giammartino, D.C., Nishida, K. & Manley, J.L. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**, 853-66 (2011).
16. Mitschka, S. & Mayr, C. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol* **23**, 779-796 (2022).
17. Singh, I. *et al.* Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat Commun* **9**, 1716 (2018).
18. Zhao, Z. *et al.* Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res* **31**, 2095-2106 (2021).
19. Masamha, C.P. *et al.* CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412-6 (2014).
20. Park, H.J. *et al.* 3' UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat Genet* **50**, 783-789 (2018).
21. Mittleman, B.E. *et al.* Alternative polyadenylation mediates genetic regulation of gene expression. *Elife* **9**(2020).
22. 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).
23. Li, L., Li, Y., Zou, X., Peng, F., Cui, Y., Wagner, E.J., Li, W. Population-scale genetic control of alternative polyadenylation and its association with human diseases. *Quantitative Biology* **10**, 44-54 (2022).

24. 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 U S A* **104**, 6758-63 (2007).
25. Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53**, 994-1005 (2021).
26. Feng, X., Li, L., Wagner, E.J. & Li, W. TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res* **46**, D1027-D1030 (2018).
27. Liu, Z. *et al.* Pan-cancer analysis identifies mutations in SUGP1 that recapitulate mutant SF3B1 splicing dysregulation. *Proc Natl Acad Sci U S A* **117**, 10305-10312 (2020).
28. Alsafadi, S. *et al.* Genetic alterations of SUGP1 mimic mutant-SF3B1 splice pattern in lung adenocarcinoma and other cancers. *Oncogene* **40**, 85-96 (2021).
29. Chen, E.Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
30. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
31. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
32. 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).
33. Zhao, Z. *et al.* Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers. *Nucleic Acids Res* **49**, 10369-10381 (2021).
34. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).
35. 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**, 831-8 (2015).
36. Jenal, M. *et al.* The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* **149**, 538-53 (2012).
37. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854-867 e9 (2018).
38. Linder, J., Koplik, S.E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23**, 232 (2022).
39. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-7 (2005).
40. Slavotinek, A.M. *et al.* Mutation analysis of the MKKS gene in McKusick-Kaufman syndrome and selected Bardet-Biedl syndrome patients. *Hum Genet* **110**, 561-7 (2002).
41. Stone, D.L. *et al.* Mutation of a gene encoding a putative chaperonin causes McKusick-Kaufman syndrome. *Nat Genet* **25**, 79-82 (2000).
42. Slavotinek, A.M. *et al.* Mutations in MKKS cause Bardet-Biedl syndrome. *Nat Genet* **26**, 15-6 (2000).

43. Katsanis, N. *et al.* Mutations in MKKS cause obesity, retinal dystrophy and renal malformations associated with Bardet-Biedl syndrome. *Nat Genet* **26**, 67-70 (2000).
44. Wuyts, W. *et al.* Mutations in the EXT1 and EXT2 genes in hereditary multiple exostoses. *Am J Hum Genet* **62**, 346-54 (1998).
45. Stickens, D. *et al.* The EXT2 multiple exostoses gene defines a family of putative tumour suppressor genes. *Nat Genet* **14**, 25-32 (1996).
46. Quintas-Cardama, A. & Cortes, J. Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood* **113**, 1619-30 (2009).
47. Salesse, S. & Verfaillie, C.M. BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. *Oncogene* **21**, 8547-59 (2002).
48. Weiner, D.J. *et al.* Statistical and functional convergence of common and rare genetic influences on autism at chromosome 16p. *Nat Genet* **54**, 1630-1639 (2022).
49. Schrode, N. *et al.* Synergistic effects of common schizophrenia risk variants. *Nat Genet* **51**, 1475-1485 (2019).
50. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509-516 (2022).
51. Cui, Y. *et al.* Alternative polyadenylation transcriptome-wide association study identifies APA-linked susceptibility genes in brain disorders. *Nat Commun* **14**, 583 (2023).
52. Chen, H. *et al.* A distinct class of pan-cancer susceptibility genes revealed by alternative polyadenylation transcriptome-wide association study. *medRxiv*, 2023.02.28.23286554 (2023).
53. Dong, G. *et al.* DDX18 drives tumor immune escape through transcription-activated STAT1 expression in pancreatic cancer. *Oncogene* **42**, 3000-3014 (2023).
54. Redmond, A.M. *et al.* Genomic interaction between ER and HMGB2 identifies DDX18 as a novel driver of endocrine resistance in breast cancer cells. *Oncogene* **34**, 3871-80 (2015).
55. McFarland, J.M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat Commun* **9**, 4610 (2018).
56. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576 e16 (2017).
57. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet* **55**, 198-208 (2023).
58. Wu, N. *et al.* TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *N Engl J Med* **372**, 341-50 (2015).
59. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
60. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
61. Zou, X. *et al.* Using population-scale transcriptomic and genomic data to map 3' UTR alternative polyadenylation quantitative trait loci. *STAR Protoc* **3**, 101566 (2022).

62. Ma, X. *et al.* ipaQTL-atlas: an atlas of intronic polyadenylation quantitative trait loci across human tissues. *Nucleic Acids Res* **51**, D1046-D1052 (2023).
63. 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**, 500-7 (2012).
64. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**, 1012-1030 (2022).
65. Wang, R., Zheng, D., Yehia, G. & Tian, B. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Res* **28**, 1427-1441 (2018).
66. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**, D315-D319 (2018).
67. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
68. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
69. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-7 (2014).
70. Grishin, D. & Gusev, A. Allelic imbalance of chromatin accessibility in cancer identifies candidate causal risk variants and their mechanisms. *Nat Genet* **54**, 837-849 (2022).
71. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

Figures

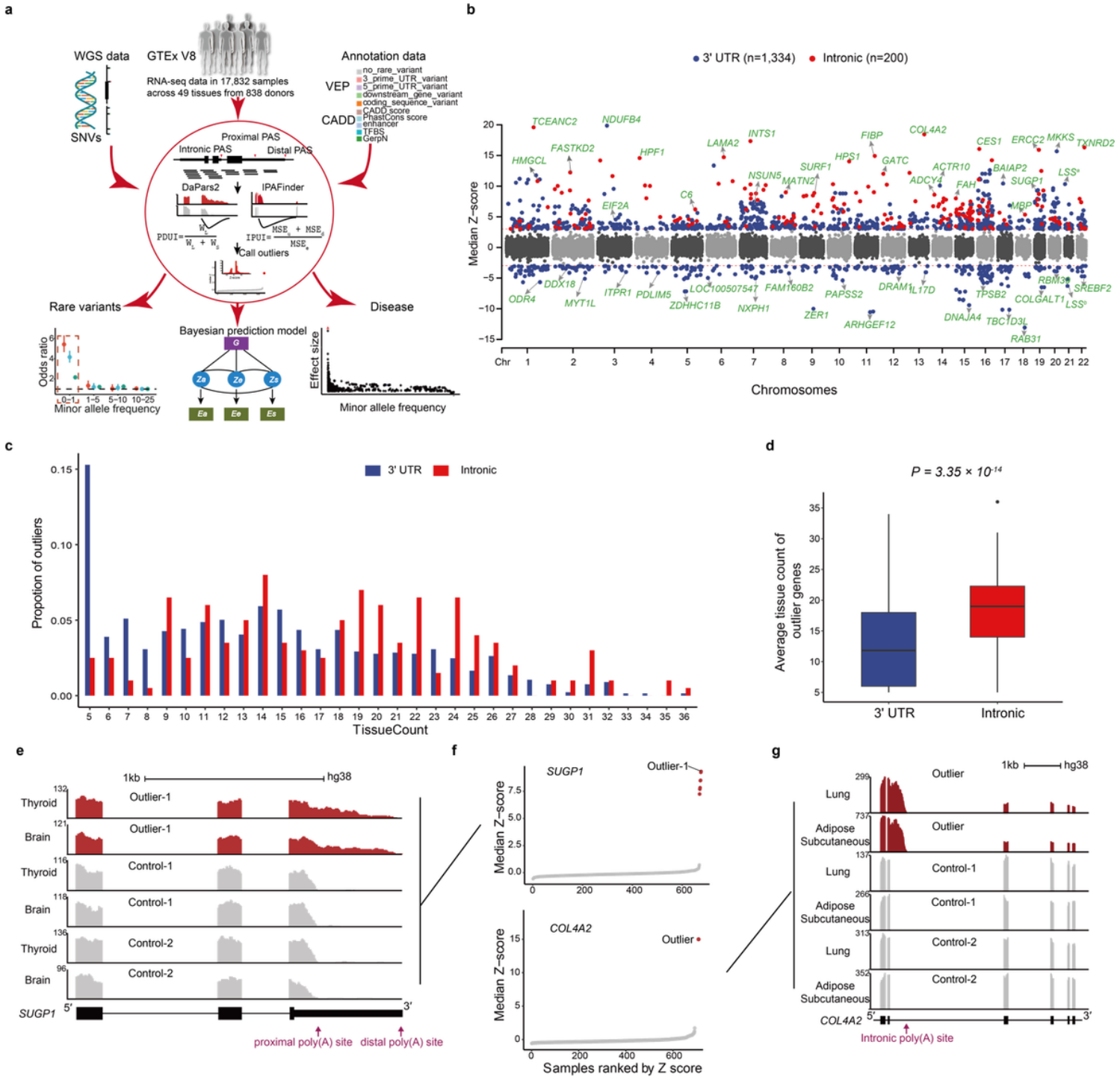

Figure 1

Atlas of human alternative polyadenylation (APA) outliers (aOutliers). a. Schematic illustrating the overall design of this study. **b.** Distribution of 3' untranslated region (UTR; blue) and intronic (red) aOutliers across the human genome. Genes with the highest (for positive median Z-scores) or lowest (for negative median Z-scores) Z-score at each chromosome region were labeled. **c.**Distribution of the number of tissues in which 3' UTR aOutliers (deep blue) and intronic aOutliers (red) were detected. **d.** Comparison of average tissue counts of 3' UTR aOutliers (deep blue; n=603 genes) and intronic aOutliers (red; n=100 genes). Box plots show the median and first and third quartiles, and whiskers extend up to 1.5 times the

interquartile range. **e.** RNA sequencing (RNA-seq) read coverage of the *SUGP1* gene 3' UTR in outlier individuals (red) and nonoutlier individuals (gray) in the Lung and Brain hippocampus. **f.** Median Z-score distribution of the *SUGP1* and *COL4A2* genes across individuals. Outliers are highlighted with red dots. **g.** RNA-seq read coverage of the *COL4A2* gene at the region of “exon5-intron-exon6” in outlier individuals (red) and nonoutlier individuals (gray). For the data shown in this figure, significance was calculated using the single-tailed Wilcoxon rank–sum test.

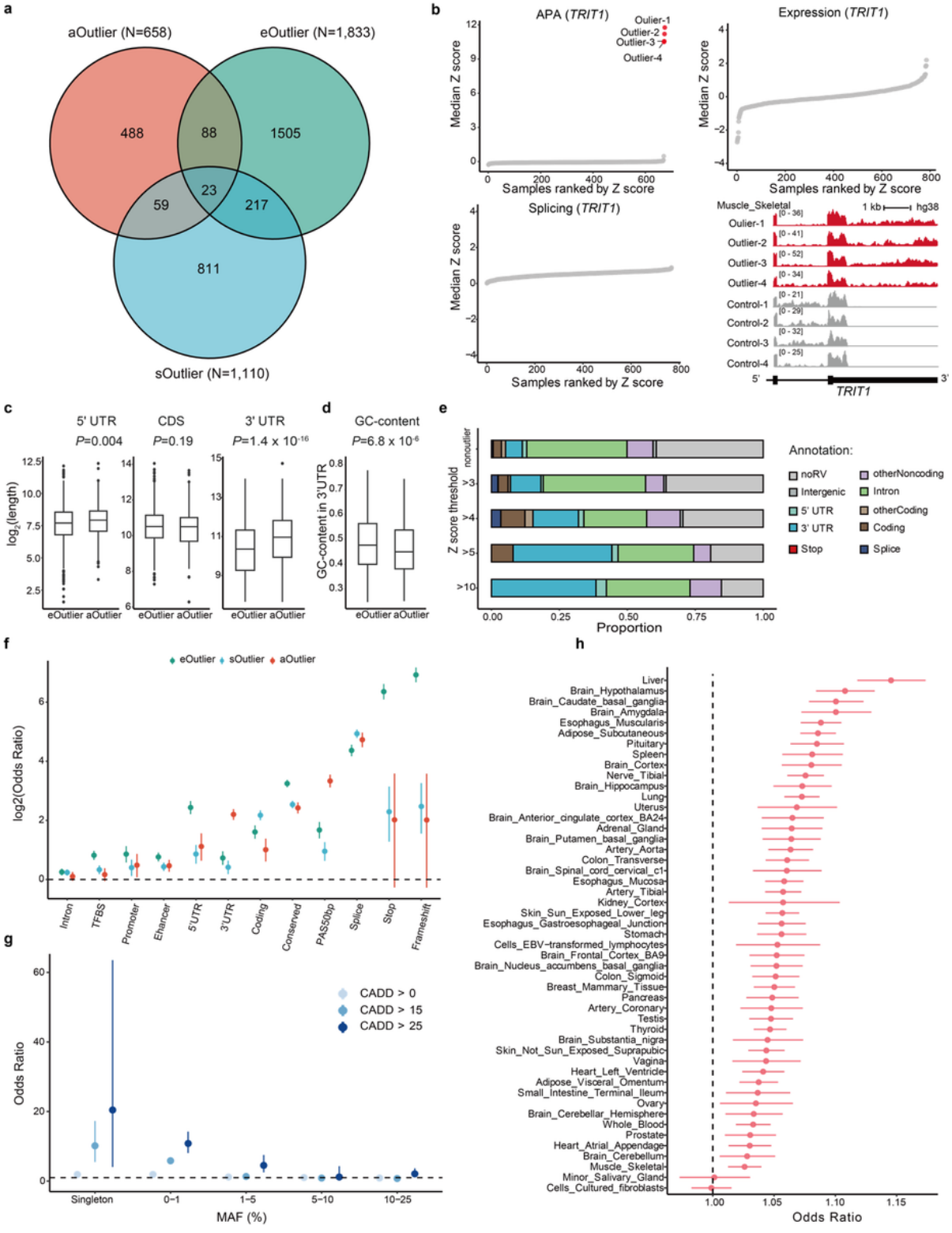

Figure 2

Outliers are distinct from other molecular outliers. a. Number of aOutlier genes also detected by analysis of expression outliers (eOutliers) and splicing outliers (sOutlier) in the same dataset. **b.** Example of an aOutlier-only gene, not detected by eOutlier and sOutlier analysis. **c.** Analysis of 5¢ UTR length, coding sequence (CDS) length, and 3¢ UTR length in aOutlier genes (n=562 genes) compared to eOutlier genes (n=1,833 genes). *P*-values were calculated using the one-sided Wilcoxon rank–sum test. Box plots show the median and first and third quartiles, and whiskers extend up to 1.5 times the interquartile range. **d**. Analysis of GC-content in 3¢ UTR regions of aOutlier (n=562) and eOutlier genes (n=1,833). *P*-values were calculated using the one-sided Wilcoxon rank–sum test. Box plots show the median and first and third quartiles, and whiskers extend up to 1.5 times the interquartile range. **e.** The proportion of aOutliers with nearby RVs of different categories. aOutliers were stratified by absolute median Z-score thresholds: $Z < 1$ (nonoutlier), $Z > 3$, $Z > 4$, $Z > 5$, and $Z > 10$. RV categories were assigned by VEP (v.104), and some were manually merged. Terms are defined as follows: "Splice" includes RVs at the splice donor site, splice acceptor site, and splice region; "Stop" includes RVs resulting in stop gained, stop lost, and start lost; "Coding" includes missense variant, stop retained variant; "otherCoding" includes CDS and synonymous variant; and "other noncoding" includes downstream gene variant, upstream gene variant, and non-coding transcript exon variant. **f.** Enrichment of RVs of different categories in aOutliers (red), eOutliers (green), and sOutliers (blue). "PAS50bp" represents the 50 base pairs upstream of the annotated poly(A) site. "Conserved" RVs are defined by mammalian phaseCons score > 0.9. Data are presented as log2 odds ratios (OR) and 95% confidence intervals (CI). **g.** Enrichment of deleterious single-nucleotide variants (SNVs) in aOutliers; variants within 1 kb of aOutlier genes were counted. Data are presented as ORs and 95% CIs. **h.** Enrichment of RVs in single-tissue aOutliers. Data are presented as odds ratio and 95% CI (y-axis) for each tissue (x-axis).

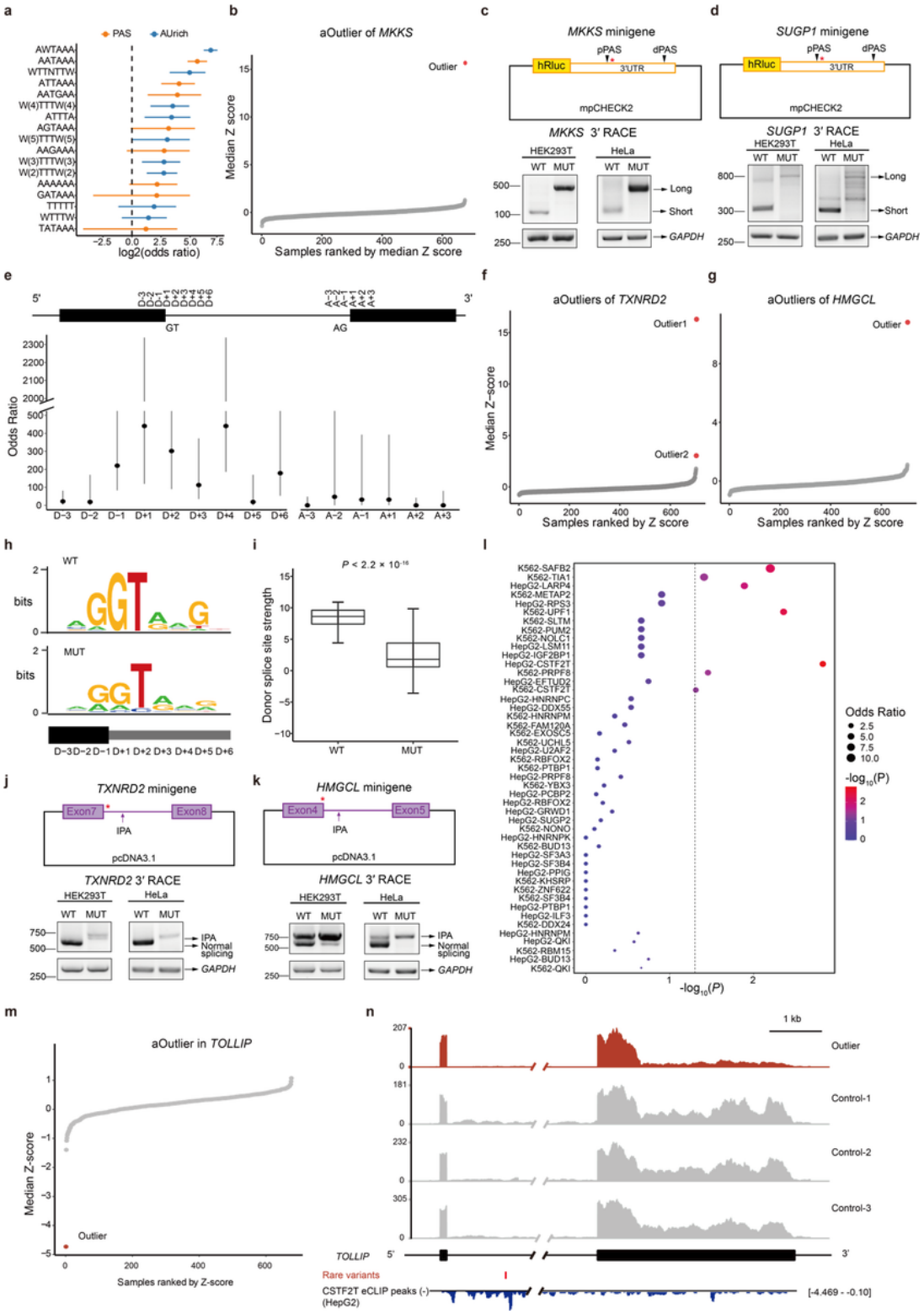

Figure 3

Functional rare variants (RVs) and RBPs associated with aOutliers. a. Enrichment of aOutlier-associated RVs in poly(A) signal (PAS) and AU-rich motifs. Central dots show the log transformed odds ratio, and lines show 95% confidence intervals**. b.** The aOutlier of gene *MKKS*. Data are presented as median Z score (y-axis) and samples (x-axis) ranked by median Z score. The outlier individual was represented as a red dot. **c-d.** The minigenes and 3' RACE assays for the 3' UTRs of *MKKS* (c) and *SUGP1* (d) in HEK293

and HeLa cells. The structures of each minigene reporter are shown at the top, and PCR priming data for both long and short isoforms are presented below. Tested RVs that alter PAS motifs are indicated with red asterisks. *GAPDH* was used as a reference in all assays. **e.** Enrichment of intronic aOutlier-associated RVs that disrupt splice sites compared with RVs associated with nonoutliers. The splice site was defined as nine bp (indicated as "D-3" to "D+6") for the donor site and six bp (indicated as "A-3" to "A+3") for the acceptor site. Enrichments are presented as ORs and 95% CIs. **f-g.** aOutliers in gene *TXNRD2* (f) and *HMGCL* (g). **h.** Consensus donor site sequences in outlier and nonoutlier individuals. **i.** Strength of donor splice sites in intronic aOutlier individuals (MUT) and controls (WT). The center horizontal lines represent the median values; boxes span from the 25th to 75th percentile, and whiskers extend to 1.5 × interquartile range. Significance was calculated using the single-tailed Wilcoxon rank–sum test. **j-k.** Minigenes and 3' RACE assays for the intronic APA of *TXNRD2* (j) and *HMGCL* (k) in HEK293 and HeLa cells. The structures of each minigene reporter are shown at the top, and PCR priming data for both long and short isoforms are presented below. Tested RVs that alter PAS motifs are indicated with red asterisks. *GAPDH* was used as a reference in all assays. **l**. Enrichment of RNA-binding protein (RBP) binding regions in aOutlier-associated RVs compared to nonoutlier RVs. Data are presented as $-\log_{10}(P)$ (y-axis) and odds ratio (dot size). **m.** The scatter plot shows aOutlier in gene TOLLIP. **n.** One rare variant in aOutlier individual of gene TOLLIP was identified located at the binding regions of CSTF2T. The RNA-seq reads coverage of the aOutlier and three nonoutlier individuals in the 3' UTR region were presented, and the binding peaks of the CSTF2T regulator by eCLIP assay were presented below. The rare variant was presented in red.

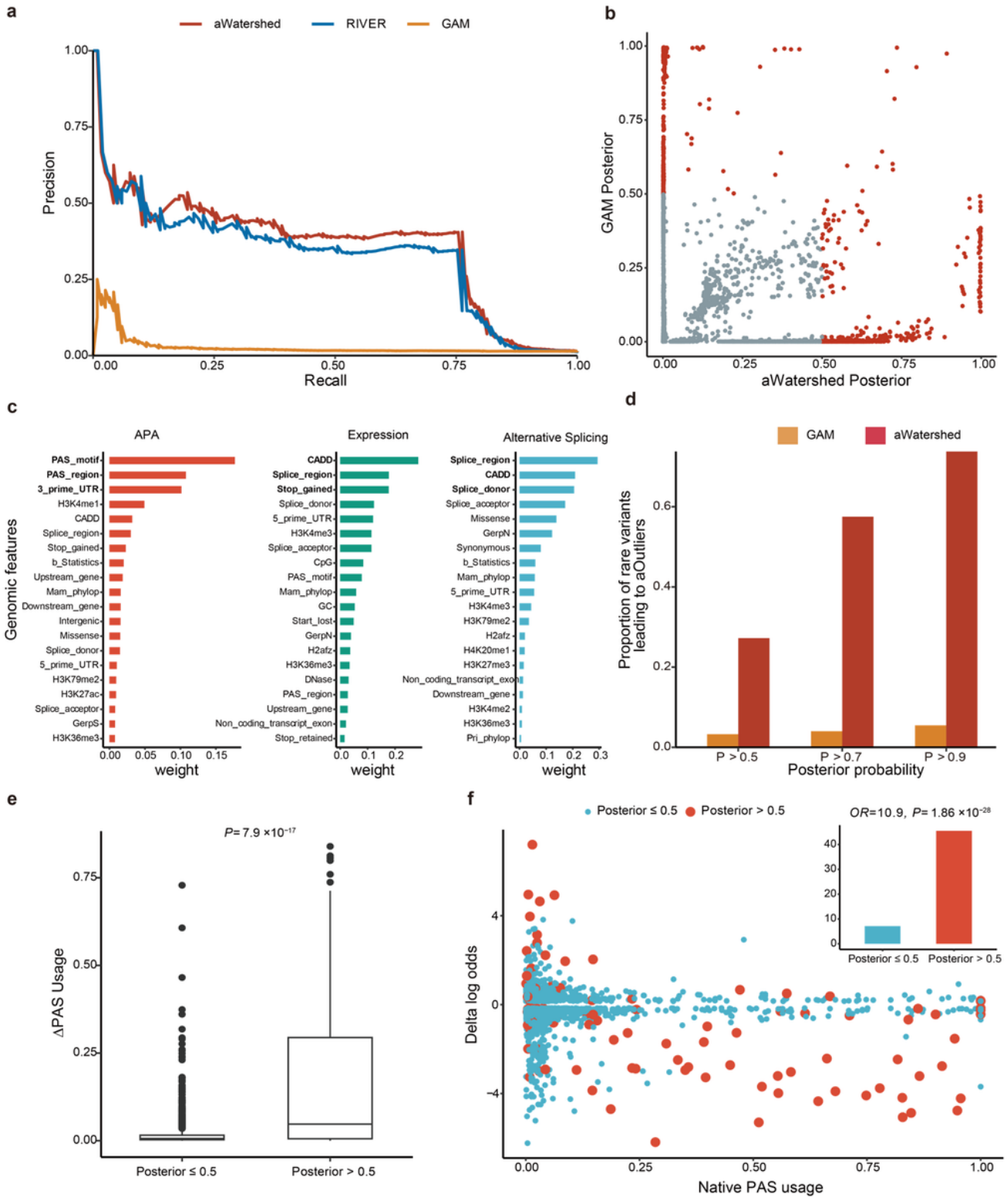

Figure 4

Development and evaluation of the APA-based Watershed (aWatershed) model. a. Performance of the aWatershed model (red) compared to the RIVER model (blue) and the genomic annotation model (orange). Data are presented as the area under the precision–recall curve (AUC-PR). **b.** The correlation between aWatershed predicted posteriors and GAM predicted posteriors. RVs with posteriors > 0.5 in either group were filled with red. **c.** Edge weights connecting top genomic annotation features to latent

regulatory variables in aOutlier signal (red), eOutlier signal (green), and sOutlier signal (blue), ranked by weight in decreasing order. The top three most influential genomic features are highlighted in bold font. **d.** The proportion of RVs leading to aOutliers. RVs were stratified based on aWatershed (red) and GAM (orange) posterior probability for APA signal. **e, f.** Evaluation of aWatershed-prioritized RVs using the data estimated from a published massively parallel reporter assay. RVs were stratified into two groups based on aWatershed APA posterior probabilities (i.e., probabilities > 0.5 (red) and £ 0.5 (blue)), and poly(A) site usage change was compared for reference and alternative alleles in each group (e), and proportion of large-effect (absolute log Fold change > 1) was compared between the two groups (f). Box plots show the median and first and third quartiles, and whiskers extend up to 1.5 times the interquartile range. *P*-values were calculated using the one-sided Wilcoxon rank–sum test.

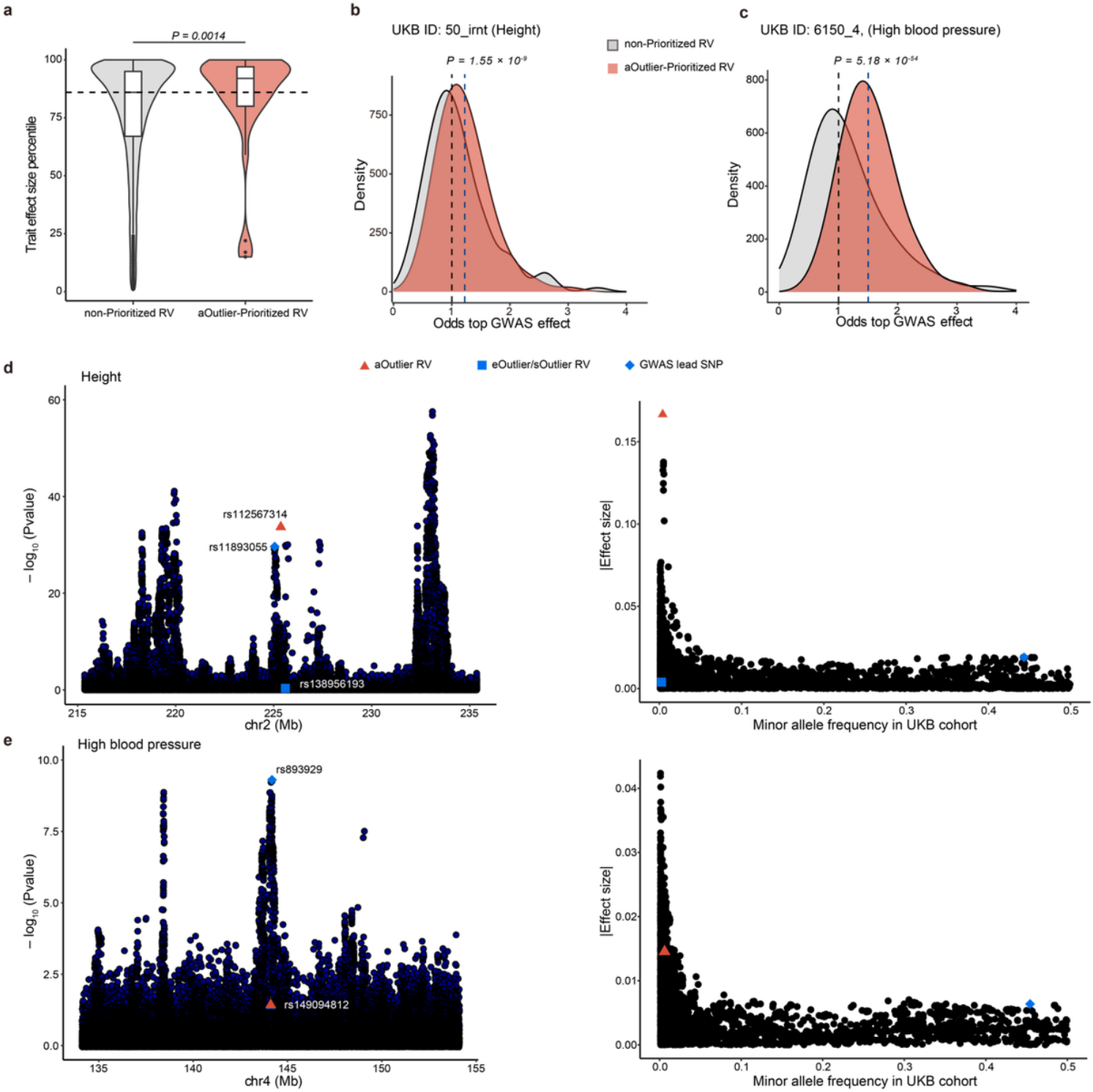

Figure 5

Trait effect sizes for aOutlier RVs prioritized by aWatershed. a. Comparison of the trait effect size of aOutlier prioritized RV (red) nearby genes with evidence of colocalization to non-prioritized RVs (gray) nearby the same genes. Box plots show the median and first and third quartiles, and whiskers extend up to 1.5 times the interquartile range. *P*value was calculated using the one-sided Wilcoxon rank–sum test (n = 77,388). **b,c.**Distribution (red) of odds estimated from permutation test assessing how often randomly drawn aWatershed-prioritized RVs have larger effect sizes in GWAS of height (b) or high blood pressure

(c) than matched non-prioritized RVs across genes. The null distribution (in gray) of odds was obtained from a permutation test by randomly drawning two RVs from a non-prioritized RV set only. *P* value was calculated from the one-sided Wilcoxon rank–sum test. **d.** Manhattan plot (left) across 20 Mb in chromosome 2 for GWAS signals of height (50_irnt) in the UKBB. The aOutlier prioritized RV rs112567314 in the colocalized region was highlighted by the red triangle, and the GWAS lead SNP is indicated by a blue diamond. The blue square denotes RVs prioritized by eOutliers or sOutliers in the same region. UKBB MAF *vs.* effect size for all variants within 1Mb of the aOutlier prioritized RV was shown on the right. **e.** Manhattan plot (left) across 20 Mb in chromosome 4 for GWAS signals of high blood pressure (6150_4) in the UKBB, and the scatter plot (right) shows the UKBB MAF *vs.* effect size for all variants in a 2Mb region cross the aOutleir prioritized RV (rs149094812). The red triangle highlights the aOutlier prioritized RV, and the blue diamond highlights the GWAS lead SNP.

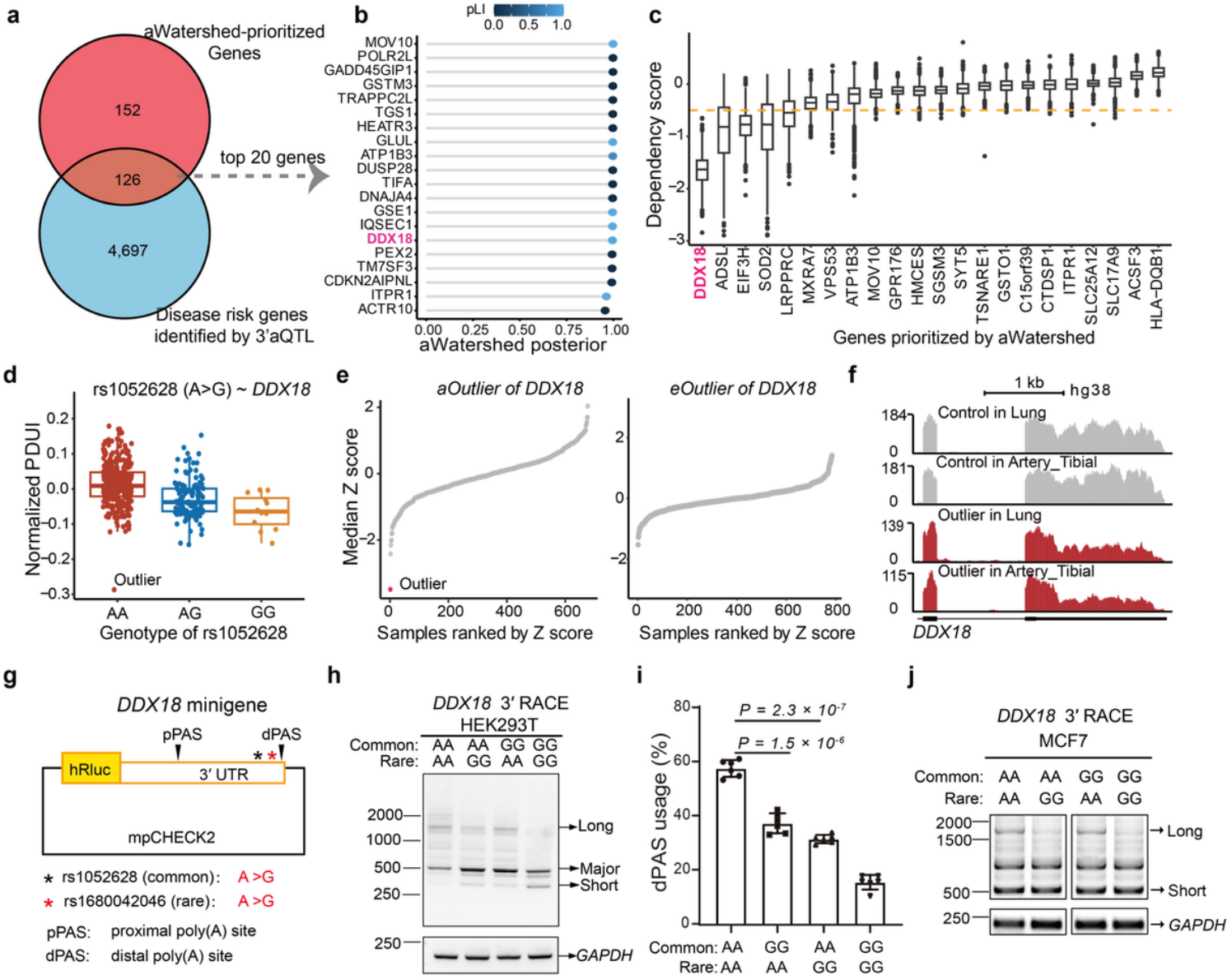

Figure 6

The convergence effect of rare and common variants links APA to human diseases. a. Intersection of disease risk genes identified by 3'aQTLs using gene-based methods, including colocalization and transcriptome-wide association study and genes associated with aWatershed-prioritized functional APA RVs. 278 genes associated with RVs having aWatershed posterior score > 0.5 were involved. **b.** The top 20 intersected genes are ranked by aWatershed posterior score. **c.** Distribution of dependency scores estimated from CRISPR-Cas9 essentiality screening assays in cancer cells for genes associated with RVs prioritizing by aWatershed and also identified as cancer risk genes by 3'aQTLs analysis. **d.** Boxplot shows the association between the common variant (rs1052628) and 3' UTR APA of *DDX18*. The outlier individual with the rare variant was labeled. **e.** aOutliers (left) and eOutliers (right) of gene *DDX18*. Data are presented as median Z-score (y-axis), and individuals (x-axis) are ranked by median Z-score. Outliers are highlighted with red dots. **f.** RNA-seq reads coverage of *DDX18* 3' UTR region and the last second exon in the outlier individual (red) and non-outlier (control) individual (gray). **g.** Minigenes of *DDX18* 3' UTR containing the common APA variant and the rare APA variant. **h.** 3' RACE assays with *DDX18* minigenes containing only the RV or only the common variant, or both the RV and the common variant. Assays were performed in HEK293T cells. **i.** The bar plot shows the usage of dPAS in each minigene measured through image J. Error bar represents the standard deviation and a two-sided student t-test was used to test the difference (n=6 independent experimental replicates). **j.** The 3' RACE assays of *DDX18* minigenes in MCF7 cells. *GAPDH* was used as the loading control.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx
- SupplementalTables.xlsx