nature genetics

Article

https://doi.org/10.1038/s41588-025-02326-8

Genetic variants affecting RNA stability influence complex traits and disease risk

Received: 26 July 2023

Accepted: 4 August 2025

Published online: 05 September 2025

Check for updates

Elaine Huang[1], Ting Fu[2], Ling Zhang[2], Guanao Yan[3,13], Ryo Yamamoto[1], Sari Terrazas[4], Thuy Linh Nguyen[2], Carlos Gonzalez-Figueroa[2], Armen Khanbabaei[5], Jae Hoon Bahn[2], Rajagopal Varada[2], Kofi Amoah[1], Jonatan Hervoso[1], Michelle T. Paulsen[6,7], Brian Magnuson[8], Mats Ljungman[6,7], Jingyi Jessica Li[1,3,9,10,11] & Xinshu Xiao[1,2,4,12]

Gene expression is modulated jointly by transcriptional regulation and messenger RNA stability, yet the latter is often overlooked in studies on genetic variants. Here, leveraging metabolic labeling data (Bru/BruChase-seq) and a new computational pipeline, RNAtracker, we categorize genes as allele-specific RNA stability (asRS) or allele-specific RNA transcription events. We identify more than 5,000 asRS variants among 665 genes across a panel of 11 human cell lines. These variants directly overlap conserved microRNA target regions and allele-specific RNA-binding protein sites, illuminating mechanisms through which stability is mediated. Furthermore, we identified causal asRS variants using a massively parallel screen (MapUTR) for variants that affect post-transcriptional mRNA abundance, as well as through CRISPR prime editing approaches. Notably, asRS genes were enriched significantly among a multitude of immune-related pathways and contribute to the risk of several immune system diseases. This work highlights RNA stability as a critical, yet understudied mechanism linking genetic variation and disease.

Identifying genetic variants that regulate gene abundance is a common strategy to decipher the mechanisms that underlie traits and diseases. It is well established that transcriptional regulation and variable stability of transcripts jointly determine steady-state messenger RNA abundance. However, the former has received far greater attention than the latter. As a result, known functional genetic variants associated with gene abundance are linked primarily to transcriptional regulation (for example, by disruption of transcription factor binding sites[1,2] or core promoter motifs[3]) rather than mRNA stability regulation.

Despite the limited attention, the role of mRNA stability in determining gene abundance has long been established[4,5]. Genome-wide characterizations of mRNA stability have revealed large variabilities in decay rates across genes[6,7]. Factors such as sequence composition[8], presence of AU-rich elements (AREs)[9], expression of RNA-binding proteins (RBPs), microRNA target sites[10] and translational efficiency[11] have all been implicated in modulating mRNA stability. Genetic variants, as mutations in the DNA template for mRNA transcription, have the propensity to alter stability-modulating sequences. Thus, genetic

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA. [2]Department of Integrative Biology and Physiology, University of California, Los Angeles, CA, USA. [3]Department of Statistics and Data Science, University of California, Los Angeles, CA, USA. [4]Molecular Biology Interdepartmental Program, University of California, Los Angeles, CA, USA. [5]Molecular, Cellular, and Integrative Physiology Interdepartmental Program, University of California, Los Angeles, CA, USA. [6]Center for RNA Biomedicine and Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA. [7]Departments of Radiation Oncology and Environmental Health Sciences, University of Michigan, Ann Arbor, MI, USA. [8]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [9]Department of Human Genetics, University of California, Los Angeles, CA, USA. [10]Department of Computational Medicine, University of California, Los Angeles, CA, USA. [11]Department of Biostatistics, University of California, Los Angeles, CA, USA. [12]Molecular Biology Institute, University of California, Los Angeles, CA, USA. [13]Present address: Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA. e-mail: gxxiao@ucla.edu

variants represent an important, yet understudied, class of features in mRNA stability regulation.

A handful of human studies have linked genetic variants to mRNA stability. Work by Pai et al. estimated that 19% of the expression quantitative trait loci (eQTLs) that they identified in lymphoblastoid cells might be regulated, at least in part, by differences in decay rates[12]. Model-based approaches[13] have been used to estimate mRNA decay rates in lung tissue, enabling identification of variants associated with RNA stability[14].

The above studies highlight the potential contribution of genetically regulated RNA stability towards gene regulation. However, a systematic characterization of stability-regulating genetic variants across different cellular contexts is still lacking. To fill these gaps, we aimed to provide a comprehensive account detailing the effects of genetic variants on RNA stability and their potential contributions to disease phenotypes. Using metabolic labeling data (Bru-seq/BruChase-seq) of 11 cell lines and a new computational workflow, RNAtracker, we examined transcriptome changes over time to identify allele-specific RNA stability (asRS) and allele-specific RNA transcription (asRT) events. We found >1,000 genes with asRS and/or asRT patterns across the cell lines with significant overlap. We showed that asRS variants can explain previously identified eQTL signals across a wide range of tissues. In addition, our data uncovered enrichment of asRS genes within immune-related pathways, many of which featured genes that help to functionally interpret genetic variants related to various immune-related diseases. Our study highlights the critical contributions of genetically mediated RNA stability—a previously underappreciated mechanism of regulation—towards human disease and biology.

Results

Overview of Bru-seq/BruChase-seq and RNAtracker

Bru-seq/BruChase-seq are a set of complementary experimental techniques for tracking the same population of RNA over time[15]. In this protocol, RNA is incubated with bromouridine nucleotides, which are subsequently incorporated into nascent transcripts. These bromouridine-labeled RNA molecules are then either isolated immediately for sequencing (Bru-seq) or 'chased' with uridine nucleotides for *n* hours so that any newly synthesized transcripts will incorporate uridine rather than bromouridine before sequencing (BruChase-seq). After *n* hours have passed, the bromouridine-labeled RNA is isolated for sequencing (leaving unlabeled transcripts behind). Thus, comparing transcript expression differences between Bru-seq and BruChase-seq samples enables inferences about degradation that may have occurred over the *n* hours (Fig. 1a). Typically, BruChase-seq data are collected at several timepoints to track changes in RNA abundance and Bru-seq data are considered time 0. We note that bromouridine labeling has minimal impact on gene expression (Extended Data Fig. 1a) and splicing[16], and thus is unlikely to confound our identification of asRS/asRT events.

To analyze and interpret Bru-seq/BruChase-seq in an allele-specific manner, we developed a computational workflow named RNAtracker (Fig. 1b and Methods). Briefly, in this workflow, data from several timepoints assayed by Bru-seq/BruChase-seq (Supplementary Note 1) are considered together to identify genes with allele-specific expression (ASE) patterns. Specifically, RNAtracker employs a beta-binomial mixture model to categorize genes probabilistically into those associated with asRS or asRT regulation (Supplementary Notes 2 and 3 and Methods). These categorizations are based on the principle that allele-specific transcriptional regulation affects all timepoints (starting at time 0) and allele-specific regulation of RNA stability induces ASE at later timepoints (no allelic bias at time 0).

RNAtracker categorizes genes by their mechanisms of genetic regulation

We obtained Bru-seq/BruChase-seq data from 16 different cell lines as part of the ENCODE project (Supplementary Table 1). For each cell line, data was collected at three timepoints with two replicates per timepoint: 0 h (Bru-seq), 2 h (BruChase-seq with 2-h uridine chase) and 6 h (BruChase-seq with 6-h uridine chase). Allelic counts were obtained at nonintronic heterozygous single nucleotide variant (SNV) positions in genes that did not overlap copy number variant (CNV) regions (Extended Data Fig. 1b). Five cell lines (K562, Panc1, PC-3, PC-9 and Caco-2) were excluded from downstream analysis as they each had fewer than 100 genes eligible for categorization (Extended Data Fig. 1c). Across the remaining 11 cell lines, we identified a total of 665 asRS genes (corresponding to 5,051 unique variants), and 491 asRT genes (corresponding to 3,397 unique variants) (Extended Data Fig. 1d and Supplementary Table 2). Genes exhibiting ASE patterns reflecting complex cases where both asRS and asRT may coexist were categorized separately (Methods). A total of 434 genes were assigned to this 'mixed' category across all cell lines (Extended Data Fig. 1d and Supplementary Table 2). An example asRS gene, *TJP2*, is shown in Fig. 1c, where allelic imbalance was not observed until the 6 h timepoint. In contrast, an asRT gene, *FN1* (Fig. 1c), exhibited allelic imbalance at times 0 h, 2 h and 6 h, supporting allele-specific transcriptional regulation.

We did not observe substantial differences in coverage or decay rate (Supplementary Note 4) across different groups of genes categorized as above, suggesting that these factors are unlikely to have skewed the categorization (Extended Data Fig. 2a,b). Removal of SNVs overlapping alternatively spliced regions had a minor effect on gene categorizations (Supplementary Note 5 and Extended Data Fig. 2c) as well, suggesting that alternative splicing is not likely to impact these gene categorizations in most cases. Assessing our workflow on simulated data (Supplementary Note 6) revealed an average precision of 0.97 and recall of 0.89 across all gene states (Extended Data Fig. 3a,b). When considering only genes that passed our confidence cut-offs (Methods), the average recall is 0.99 (Extended Data Fig. 3c).

Although the causal variant underlying asRT does not need to lie in the mRNA itself, RNAtracker cannot detect asRT genes without any heterozygous SNVs in the mRNA (Extended Data Fig. 4a). Such genes may carry heterozygous variants in the promoter/enhancer regions that regulate transcription but, without heterozygous SNVs in the mRNA to observe, they are untestable by RNAtracker. To address this limitation, we applied RNAtracker to identify ASE genes using testable intronic SNVs alone at timepoint 0 h. Since introns captured at 0 h Bru-seq most likely have not been spliced out, allelic imbalance at this timepoint implies that the gene is under transcriptional regulation. We call this class of genes 'intron-based asRT' (Extended Data Fig. 4b and Supplementary Table 2) and include them in the calculation of asRT prevalence (Fig. 1d), as well as all analyses hereafter. Notably, including the 'intron-based asRT' genes resulted in only minor shifts in asRT prevalence across most cell lines (Extended Data Fig. 4c). This approach does not apply to asRS genes, which are most likely regulated by SNVs in the mRNA.

The prevalence of genes under stability regulation (asRS plus mixed) was variable across cell lines, ranging from 6.2% in HUVEC to 26.5% in Calu3 (Fig. 1d). Prevalence was calculated by dividing the number of asRS plus mixed genes over the total number of genes categorized by RNAtracker in each cell line. We observed that most asRS genes were unique to a single cell line (Extended Data Fig. 5a). However, this observation may be due partially to differences in genetic background among the cell lines or limited sequencing depth in each sample to detect asRS events. As a result, common testable variants and genes are limited across cell lines (Extended Data Fig. 5b–d). Alternatively, it may also reflect cell-type-specificity of asRS. To further examine this latter possibility, we asked whether the overlap of asRS genes between a pair of cell lines was higher than expected by chance (Methods). A total of nine pairwise comparisons exhibited significant difference ($P < 0.05$). Notably, all of them showed that the shared asRS prevalence was greater than expected (Fig. 1e). We observed similar results on the variant level, in which most asRS variants were identified in a single

https://doi.org/10.1038/s41588-025-02326-8

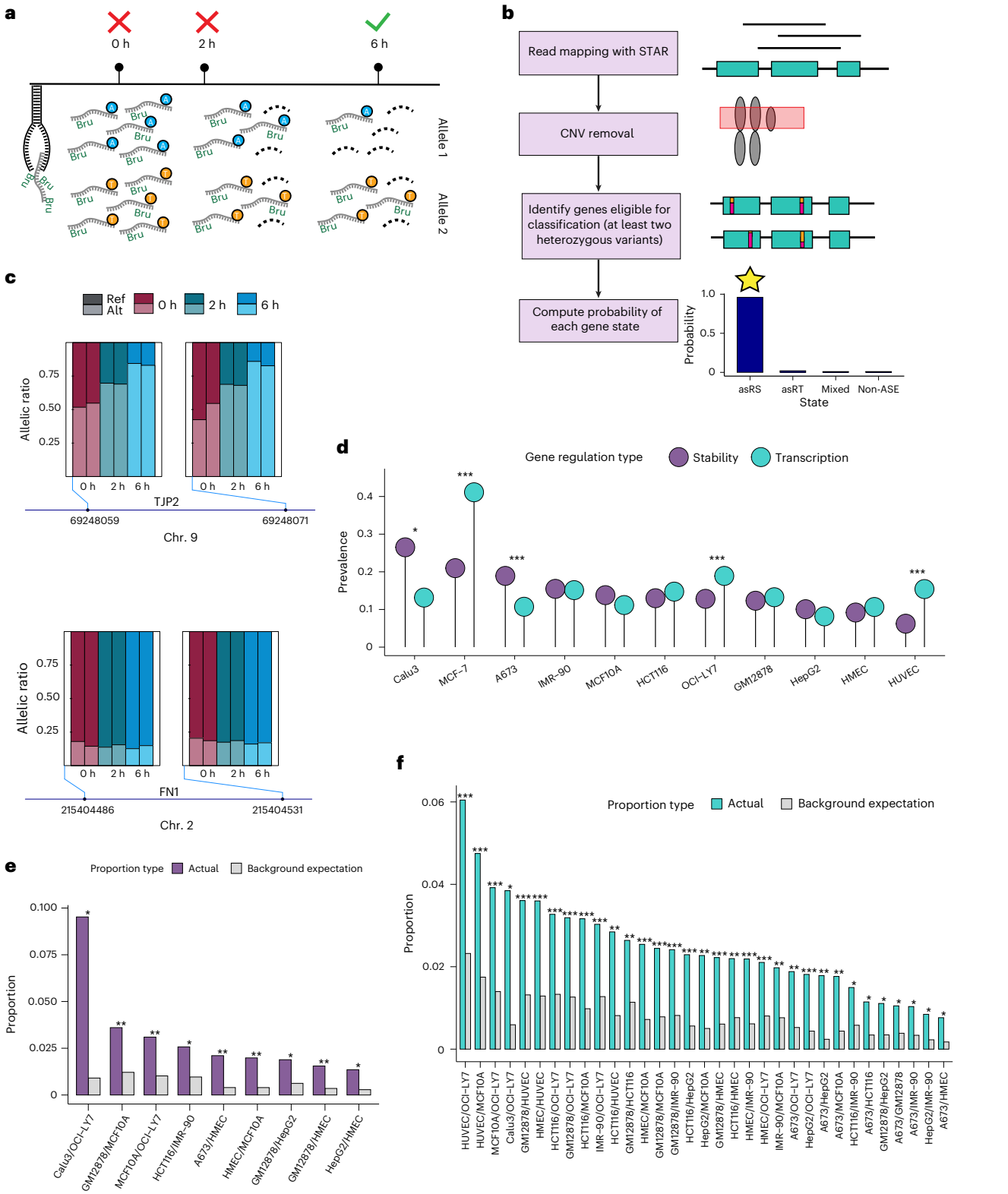


Fig. 1 | RNAtracker categorizes genes by the underlying mechanisms of allelic imbalance. a, Bru/BruChase-seq reads for a gene exhibiting asRS pattern (allele 1 exhibits greater degradation than allele 2, which becomes apparent at the 6-h timepoint; schematic illustration only). **b**, RNAtracker categorizes genes as asRS, asRT, 'mixed' or non-ASE by using a beta-binomial mixture model to calculate the posterior probability of each state. **c**, Example asRS (*TJP2*) and asRT (*FN1*) genes. Variants in *TJP2* exhibit balanced allelic ratios at 0 h, but unbalanced allelic ratios at 2 h and 6 h in the HCT116 cell line. Variants in *FN1* exhibit unbalanced allelic ratios at all three timepoints in the A673 cell line. Alt, alternative; ref, reference. **d**, Comparison of the prevalence of stability-regulated genes versus transcriptionally regulated genes. Stability-regulated genes include asRS and mixed genes. Transcriptionally regulated genes include asRT, intronic asRT and mixed genes. To calculate prevalence, the number of genes falling under each of these categories is summed and divided by the total number of categorized genes in the cell line. For each cell line, the prevalence of stability-regulated versus transcriptionally regulated genes was compared through two-sided Fisher's exact test (*$P \leq 0.05$, ***$P \leq 0.001$). **e**,**f**, Pairs of cell lines that exhibited a significant difference between the expected and actual proportion of overlapping asRS (**e**) or asRT (**f**) genes (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, two-sided binomial test).

https://doi.org/10.1038/s41588-025-02326-8

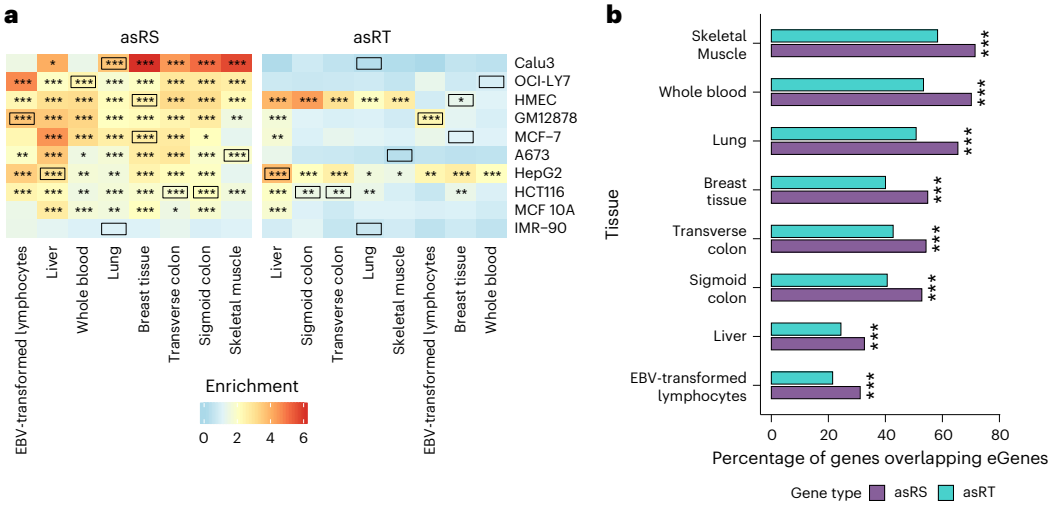


Fig. 2 | asRS and asRT events overlapping GTEx eQTL and their target genes. **a**, Enrichment (two-sided Fisher's exact test odds ratio compared against background variants) of asRS and asRT variants among GTEx eQTLs. Black boxes represent cell lines matched with their most biologically similar GTEx tissue; $^{*}P \le 0.05$, $^{**}P \le 0.01$, $^{***}P \le 0.001$. **b**, Percentage of asRS or asRT genes that overlapped genes with significant eQTLs (eGenes) in GTEx tissues; $^{***}P \le 0.00$) (two-sided Fisher's exact test comparing asRS and asRT enrichment). EBV, Epstein–Barr virus.

cell line (Extended Data Fig. 5e). All 24 pairwise comparisons in which the actual proportion of overlapping asRS variants was significantly different from the background expectation exhibited greater actual asRS prevalence than background (Extended Data Fig. 5f). Thus, the low fraction of overlap between cell lines is due largely to having different SNPs present in their genomes. In other words, given shared genetic background, the variant effects tend to be independent of cell type.

Similarly, genes under transcriptional regulation (asRT plus intron-based asRT plus mixed) also exhibited variability in prevalence across cell lines, ranging from 8.17% in HepG2 to 41.1% in MCF-7 (Fig. 1d and Supplementary Note 7). Like the asRS genes, most asRT genes were unique to a single cell line (Extended Data Fig. 6a). Nonetheless, 33 pairs of cell lines showed significant differences between the background expectation versus the actual proportion of overlapping asRT genes, all in which the shared asRT prevalence was greater than expected (Fig. 1f). Again, similar results were observed on the variant level (Extended Data Fig. 6b,c). Together, these results suggest that genetic variants often affect RNA stability or transcriptional regulation in a cell-type-independent manner, consistent with the genetically driven nature of such events.

asRS and asRT contribute to gene expression regulation

We next assessed the prevalence of asRS and asRT events among variants that have been associated previously with gene expression changes on a population-wide scale (GTEx eQTL data)[17]. We overlapped asRS and asRT variants with significant eQTLs, which are genetic loci associated with gene expression variation, among tissue types that most closely matched the cell lines in our dataset (Supplementary Table 1). Since regulatory variants of RNA stability are expected to be intragenic, for this analysis we required not only the asRS/asRT variant to match the eQTL, but also the eQTL target (eGene) to match the asRS/asRT gene. We observed that both asRS and asRT variants were enriched among significant eQTLs (Fig. 2a and Extended Data Fig. 7a). Although the lower enrichment of asRT variants is expected since only intragenic variants were considered, the high enrichment of asRS variants supports our prediction that these intragenic variants are associated with stability regulation. In addition, the magnitude of enrichment appeared to be unrelated to the biological similarity between the cell line/GTEx tissue, further supporting the cell-type-independent effects of asRS and asRT variants (Fig. 2a).

Moving from a variant-level to gene-level analysis, we found that the proportion of asRS genes that overlapped eGenes was greater than that of asRT genes (Fig. 2b and Extended Data Fig. 7b). This was the case for combined asRS or asRT genes across all cell lines (Fig. 2b), as well as for each cell line individually (Extended Data Fig. 7b). As with the eQTL variant overlap, the magnitude of enrichment was unrelated to the biological similarity between the cell line/GTEx tissue. Overall, we found that, among the subset of eGenes that overlapped the total set of genes categorized by RNAtracker, 15.6–19.2% overlapped asRS genes, whereas 20.6–23.7% overlapped asRT genes. The slightly higher percentage of overlap with asRT genes is to be expected since there are more asRT genes (when intron-based asRT genes are included) than asRS genes (Supplementary Table 2) in our data. Together, these analyses revealed that asRS and asRT both contribute towards shaping gene expression profiles that have been observed on a population-wide scale.

Delineating functional mechanisms and effects of asRS variants

As genetically mediated RNA stability has been underexplored previously despite its essential contributions to gene regulation, we focused on further analysis of asRS events for the remainder of the study. In an asRS gene, several genetic variants demonstrate ASE patterns. However, not all variants are necessarily functional with regards to their effect on mRNA stability. Nonetheless, the observation of asRS reflects the existence of one or more functional variants as *cis*-acting regulators of RNA stability. To hone in on the functional variants as well as their mechanisms of action, we first considered enrichment of asRS variants in binding regions of RBPs as determined through enhanced cross-linking immunoprecipitation (eCLIP) experiments. Relative to random controls (Methods), asRS variants were enriched significantly in the binding sites of known stability-regulating RBPs, such as MATR3 (ref. 18), FMR1 (ref. 19), TIA1 (ref. 20) and UPF1 (ref. 21) (Fig. 3a and Supplementary Table 3).

Showcasing a more granular view of how asRS variants may impact RBP binding, we also observed significant enrichment of asRS variants among allele-specific binding (ASB) sites in eCLIP data that were identified using our previously developed method BEAPR[22] (Fig. 3b). ASB reflects the functional role of an asRS variant in altering protein–RNA interactions. Notably, the RBP with the highest proportion of asRS variants among its ASB sites is SUB1 (Fig. 3c, Extended Data Fig. 8a and

https://doi.org/10.1038/s41588-025-02326-8

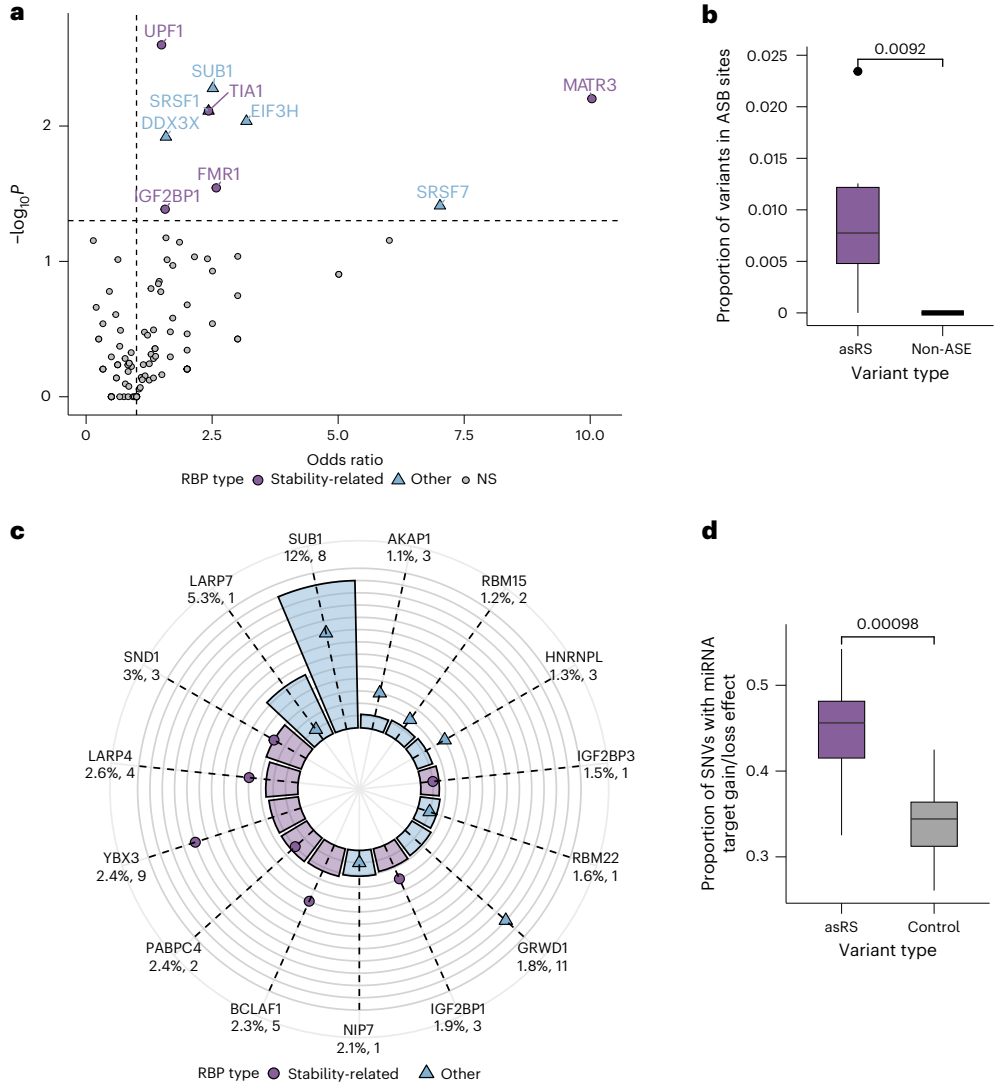


Fig. 3 | asRS variants are enriched within functional regions. a, Enrichment of asRS variants within the eCLIP peaks of each RBP. *P* values were calculated by two-sided Fisher's exact test. NS, nonsignificant. **b**, Proportion of asRS and control variants in each cell line ($n = 11$) that overlapped ASB sites. *P* value was calculated by two-sided Wilcoxon's signed-rank test. RBPs with known roles in RNA stability and decay are shown in purple. **c**, RBPs with the highest proportion of ASB sites that overlapped asRS variants. The height of the columns represents the proportion of ASB sites that overlapped asRS variants; the distance of the circular dot/triangle from the center reflects the number of ASB sites that overlapped asRS variants. Values of these two metrics are also shown below each RBP name. **d**, Proportion of asRS and control variants in each cell line ($n = 11$) that overlap miRNA target loss/gain SNPs. *P* value was calculated by Wilcoxon's signed-rank test. In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5× the interquartile range.

Supplementary Table 3), which also exhibited significant enrichment of asRS variants within its eCLIP peaks (Fig. 3a and Supplementary Table 3) and has been shown to stabilize its target RNAs[23].

In addition to RBPs, miRNAs are well-known regulators of RNA stability[24]. Thus, we asked whether asRS variants may alter miRNA targeting. SNPs in miRNA seed regions that create or disrupt miRNA binding sites (that is, target gain/loss effects, respectively) have previously been identified[25]. In total, 2,243 and 2,198 asRS variants overlapped these gain and loss sites, respectively. Using miRNAs expressed in each cell line (Methods), we observed that the proportion of asRS variants that overlapped miRNA target sites was significantly higher than that of control SNVs (Fig. 3d and Supplementary Table 3). Analogously, SNPs in miRNA seed regions are enriched significantly with asRS variants (Extended Data Fig. 8b).

Experimental support for asRS events

To provide orthogonal experimental support for asRS genes, we performed deep transcriptomic sequencing in GM12878, HCT116 and MCF-7 cells at various timepoints after treatment with the transcriptional inhibitor actinomycin D (ActD; 0 h untreated; 2 h, 8 h and 24 h post-treatment) (Fig. 4a). We then identified asRS genes with SNVs that demonstrated ASE at timepoints after 0 h, or that exhibited increased allelic imbalance compared to the 0 h timepoint (Methods). Furthermore, the direction of the imbalance (that is, whether the reference or alternative allele degrades faster over time) was required to be consistent with the observation in the Bru/BruChase-seq data. With these requirements, we obtained experimental support for 159 (74.3%) asRS genes out of a total of 214 testable in the ActD RNA sequencing (RNA-seq) data (Supplementary Table 4). asRS genes *GEN1*, *CDC137* and *C2CD2* are shown as examples in the MCF-7, GM12878 and HCT116 cell lines, respectively (Fig. 4b–d). We note that ActD treatment functions as a major disruptor of cellular physiology and may affect post-transcriptional processes such as RNA localization[7]. As a result, one should not view ActD-based experiments as a gold standard to evaluate the performance of asRS prediction. Nonetheless, it functions as an orthogonal support for a select number of asRS events.

https://doi.org/10.1038/s41588-025-02326-8

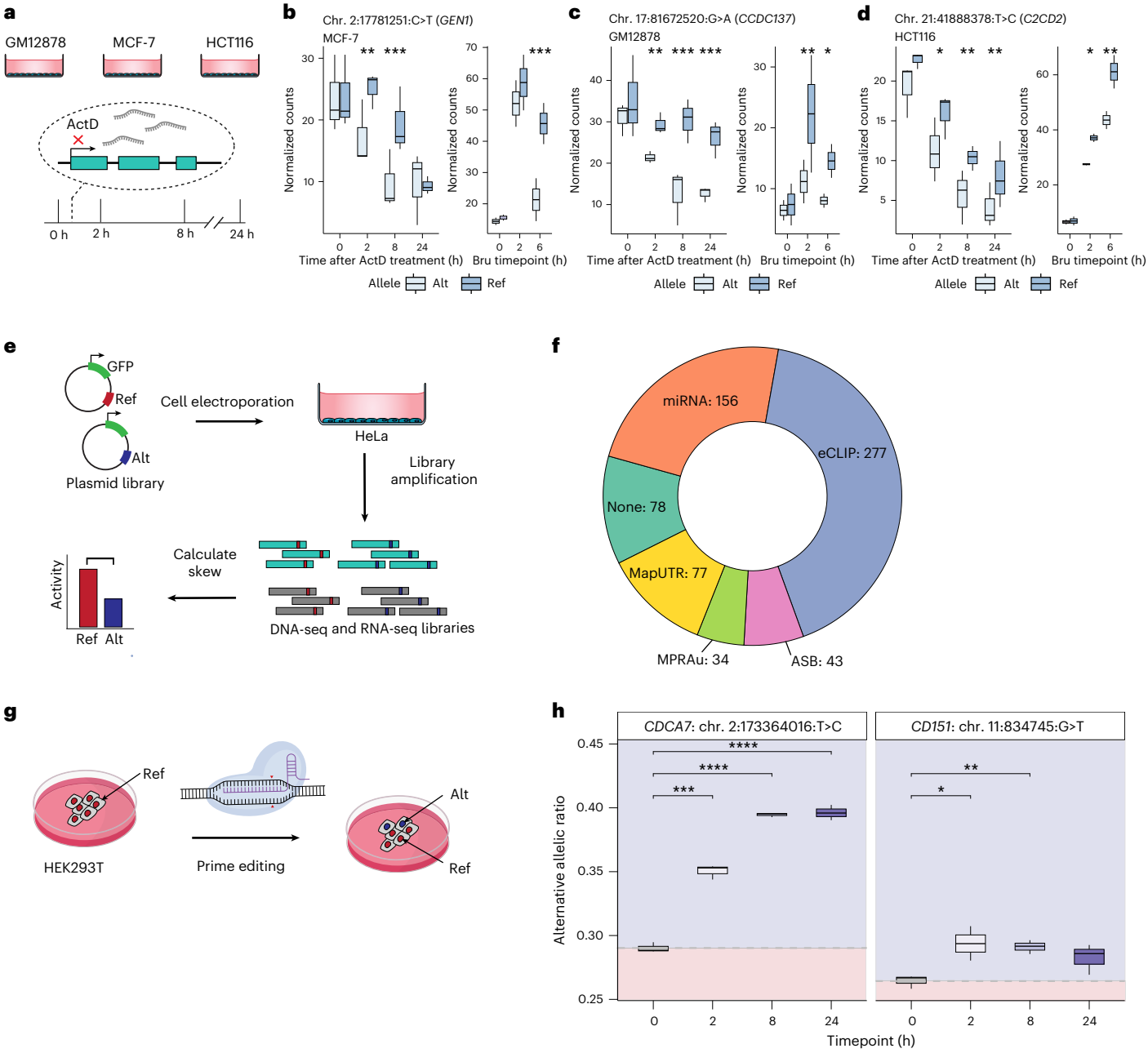


Fig. 4 | Experimental support for asRS variants. a, Schematic illustration of ActD RNA-seq experiments. **b**–**d**, Examples of asRS SNVs in MCF-7 (**b**), GM12878 (**c**) and HCT116 (**d**) that exhibit stability-mediated regulation in ActD RNA-seq. Left: comparison of normalized counts for each SNV in ActD RNA-seq at pretreatment (0 h) and post-treatment (2 h, 8 h, 24 h) timepoints (three replicates per timepoint). Right: comparison of normalized counts for each SNV in Bru/BruChase-seq data (two replicates per timepoint). $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$ (SNV allele-specific expression test *P* value; Supplementary Note 1). **e**, Schematic illustration of MapUTR design. **f**, Validation method for asRS genes. If a gene had variants validated by several methods, the following priority order was used: MapUTR > MPRAu > ActD > ASB > eCLIP > miRNA. **g**, Schematic illustration of prime editing. **h**, Alternative allelic ratio (number of alternative allele reads/(number of reference allele reads + number of alternative allele reads)) observed in CRISPR-edited HEK293T cells RNA-seq before (0 h) and after (2 h, 8 h, 24 h) treatment with ActD (three replicates per timepoint). Regions in blue indicates higher alternative allelic stability compared to the 0 h timepoint; regions in red indicate higher reference allelic stability. Dashed gray line represents the mean alternative allelic ratio observed in the 0 h samples. *P* value calculated by a two-sided *t* test. From left to right: $P = 0.00023$; $P = 7 \times 10^{-5}$; $P = 4.6 \times 10^{-5}$ (*CDCA7* variant); $P = 0.049$; $P = 0.0035$; $P = 0.1$ (*CD151* variant). In boxplots, minima/maxima represent least/greatest SNV counts or alternative allelic ratios, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5× the interquartile range.

Whereas the ActD experiments showcased that asRS genes are undergoing stability-mediated regulation, they do not reveal which specific variants are functional with regards to their effects on allele-specific degradation. Variants with ASE may simply be tag variants reflecting the existence of a functional SNV. To hone in on these functional variants, we leveraged data from massively parallel reporter assays (MPRA) that were designed to identify functional variants affecting post-transcriptional regulation (Fig. 4e). These experiments involve cloning oligonucleotides containing the variant of interest and their genomic context into the 3′ UTR of a reporter gene. After cellular transfection of the plasmid reporters, sequencing data of the plasmid DNA and mRNA are compared to identify sites associated with significant allelic expression

differences. Sites that exhibit differences above a specified threshold can then be nominated as candidate functional variants.

We first overlapped asRS variants with SNVs tested by a screening method from our laboratory called MapUTR[26]. Among the asRS variants that overlapped those evaluated by MapUTR, 106 (29.04%) were identified as functional, defined as exhibiting significantly different activity scores between their alternative and reference alleles ($\log_{\text{fold change (FC)}} \geq 0.1$, false discovery rate (FDR) ≤ 0.1; Methods). On the gene level, 55.07% (76 out of 138) of the MapUTR tested asRS genes had at least one functional variant (Supplementary Table 5). We also found that 61 out of the 365 (16.71%) asRS variants tested by a separate massively parallel assay (MPRAu[27]) were identified as functional transcript abundance-modulating variants (tamVars). On a gene level, 40 out of 107 (37.38%) MPRAu tested asRS genes had at least one tamVar (Supplementary Table 5). Combined with the RBP and miRNA analyses in the last section, we were able to nominate at least one functional variant for 88.3% (587 out of 665) of all asRS genes (Fig. 4f).

Prime editors can be used to introduce variants at specific genomic positions (Fig. 4g). To further hone in on causal asRS variants, we introduced select asRS variants into the genome of HEK293T cells, which have proven editing efficiency with prime editing[28]. We prioritized testing variants within genes that were supported by our MPRA (MapUTR) (Supplementary Table 5). We were able to successfully perform genome editing for variant chr. 2:173364016:T>C in *CDCA7* and variants chr. 11:838672:C>T and chr. 11:834745:G>T in *CD151*. After confirming successful genome editing (with an average editing efficiency of 25.24% across the three variants) (Extended Data Fig. 9a–c), we performed gene-targeted sequencing of cells at various timepoints after treatment with ActD (0 h untreated; 2 h, 8 h and 24 h post-treatment). To assess the variant effect on stability, we compared the variant allelic ratio at each post-ActD treatment timepoint with that of the untreated timepoint. A significant difference in allelic ratio at a post-treatment timepoint compared to the 0 h timepoint points to a difference in the stability of the two alleles. Under this evaluation, chr. 11:834745:G>T in *CD151* and chr. 2:173364016:T>C in *CDCA7* were both identified as causal stability-regulating variants (Fig. 4h). We note that the allelic ratio at 0 h is similar to the DNA allelic ratio (0.24 for *CDCA7*:chr. 2:173364016:T>C and 0.27 for *CD151*:chr. 11:834745:G>T).

In *CDCA7*, we found that an increased proportion of reads were assigned to the alternative allele (C) of chr. 2:173364016:T>C at all post-treatment timepoints compared to the 0 h timepoint. This suggests that the alternative allele confers higher RNA stability to the gene compared to the reference allele. Indeed, in the Bru/BruChase-seq data (although in a different cell line, OCI-LY7, from HEK293T), we also observe that the alternative allele exhibits higher expression at both 2 h and 6 h, suggesting greater stability compared to the reference allele (Extended Data Fig. 9d). Similarly in *CD151*, we found that an increased proportion of reads were assigned to the alternative allele (T) of chr. 11:834745:G>T at all post-treatment timepoints compared to the 0 h timepoint. This suggests that the alternative allele confers higher RNA stability than the reference allele. Although we are unable to compare this prime editing result with the Bru/BruChase-seq data due to insufficient BruChase-seq coverage at 2 h and 6 h for this variant, our findings suggest functional role for what would otherwise be an understudied variant. Overall, the above results demonstrate causality for the two variants on RNA stability.

asRS genes are enriched in immune-related pathways

To elucidate the functional importance of stability-regulated genes, we first performed Gene Ontology (GO) enrichment analysis to identify biological processes that featured a significant number of asRS genes (Methods). Out of all enriched GO terms, 'positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay' exhibited by far the most significant enrichment (Fig. 5a and Supplementary Table 6). This group includes the genes *CNOT1* and *PABPC1*, which have been studied closely in tandem for their contributions towards generating cycles of mRNA deadenylation[29]. Notably, another mRNA stability-related term, 'miRNA metabolic process,' was also one of the most significant hits. Other significant terms were related to cell adhesion and junction organization or immune response—all of which are closely related functionalities. Although proper cell adhesion functioning is important across all cells, regulation of this process is especially relevant to immune cells[30] and inflammatory processes[31].

To explore the complete list of enriched GO terms more thoroughly, we clustered terms by semantic similarity (Methods). This allowed us to ascertain whether there exist groups of related biological pathways that were consistently enriched among asRS genes. From this analysis, we again observed that 'cell adhesion' and 'cell–cell junction organization' as two clusters with the highest average enrichment scores (Fig. 5b). We also observed several clusters of terms highlighting immune-related processes (such as 'innate immune response' and 'defense response to Gram-positive bacterium'), as well as catabolic processes such as 'proteolysis' and 'positive regulation of autophagy'—a process in which cytosolic material, including proteins, is delivered to lysosomes for degradation[32] (Fig. 5c). Notably, the extent of enrichment for these processes appears specific to asRS genes, as we did not observe the same level of significance among asRT genes (Extended Data Fig. 10a).

asRS variants are enriched among genome-wide association studies hits

Genome-wide association studies (GWAS) can relate genetic variants to various traits by identifying significant associations between specific variants and traits on a population-wide scale[33]. Modulating mRNA stability is one mechanism through which variants may contribute towards specific phenotypes. Indeed, we found that asRS variants were enriched significantly among hits with genome-wide significance ($P < 5 \times 10^{-8}$) reported in the GWAS catalog[34] (Fig. 6a).

Whereas the RNAtracker workflow does not pinpoint the exact causal variant in asRS genes, our MPRA assay and ASB analysis allowed us to identify likely functional asRS variants. Of the 239 total variants examined, 24 directly overlapped significant GWAS hits (Supplementary Table 7). Several asRS genes (for example, *CCND1*, *CDK6*, *EPHA3*, *IL7R*, *WDFY4*) harbored one or more variants associated with several traits, including autoimmune disorders such as multiple sclerosis, primary biliary cirrhosis, rheumatoid arthritis and systemic lupus erythematosus. In most cases, the mechanism through which these asRS-overlapping GWAS hits contribute to disease risk has not been explored.

Next, we performed stratified linkage disequilibrium (LD) score (S-LDSC) regression to assess disease heritability enrichment of asRS variants. Given their apparent relevance to immune-related processes, we focused on immune-related diseases with GWAS summary statistics available through the GWAS catalog (Methods). Notably, for several autoimmune diseases (such as rheumatoid arthritis and systemic lupus erythematosus), asRS variants demonstrated heritability enrichment across several independent studies (Fig. 6b and Supplementary Table 7). Together, our data suggest that stability-regulating variants in these genes may contribute to disease susceptibility.

asRS genes are associated with immune-related diseases

To reinforce the relevance of asRS genes to the immune-related diseases of interest, we took an approach that is analogous to that of summary-based transcriptome-wide association studies (TWAS[35]). TWAS is similar in principle to GWAS; however, rather than associating specific genetic variants with traits, it can identify associations between the expression of genes and various traits. Specifically, we built genetics-based gene expression prediction models using genic SNPs (Methods) to infer the expression of asRS genes in GTEx participants

https://doi.org/10.1038/s41588-025-02326-8

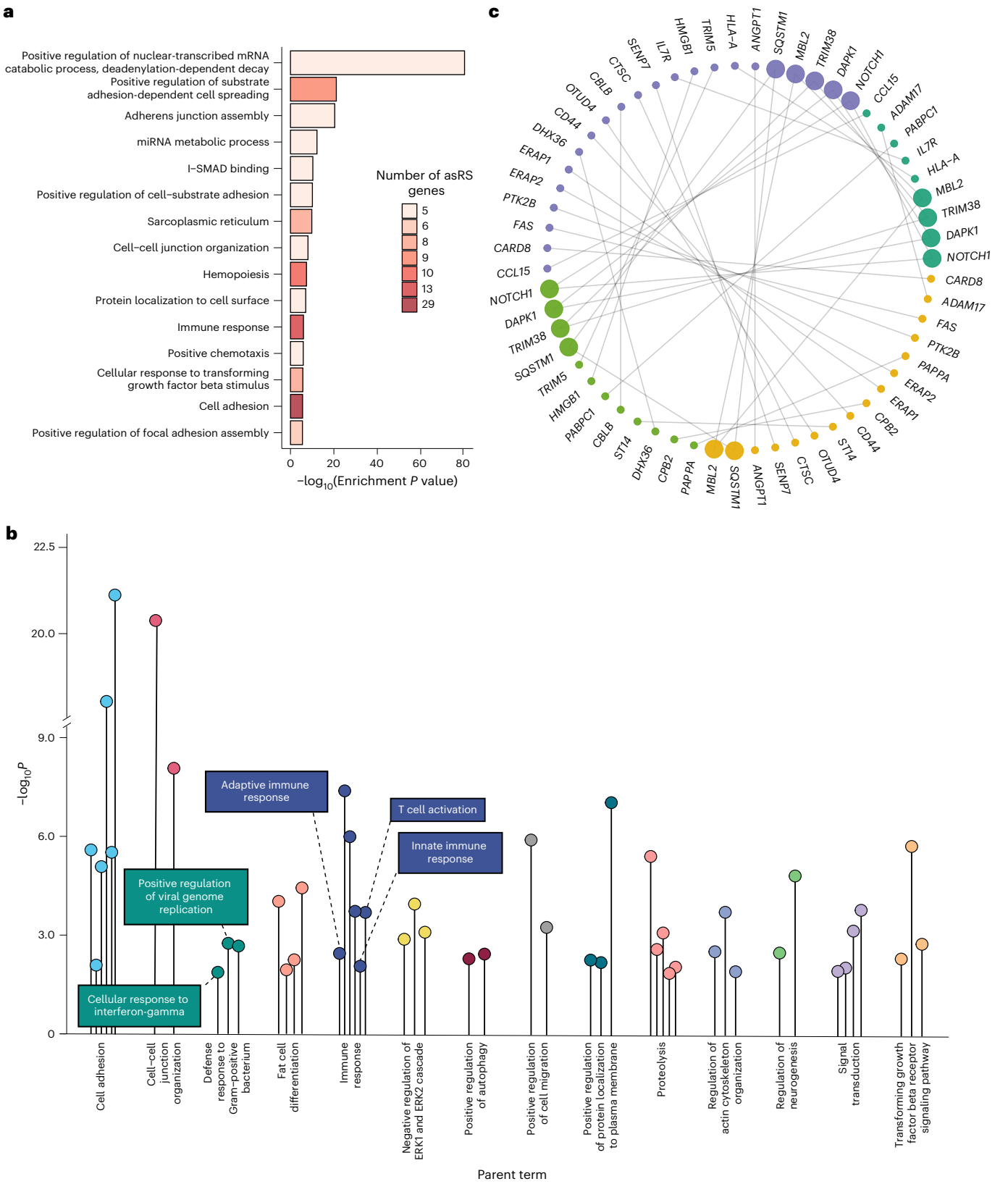


Fig. 5 | asRS genes are enriched in immune-related processes. a, Top 15 enriched GO terms for asRS genes. **b**, rrvgo plot showing GO terms after clustering. A selection of immune-related terms is highlighted. 'Positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay' is omitted for plotting purposes because its *P* value is an outlier. **c**, asRS genes in the rrvgo groups 'innate immune response,' 'proteolysis,' 'defense response to Gram-positive bacterium' and 'positive regulation of autophagy.' For **a** and **b**, *P* values were derived from an empirical Gaussian distribution of number of control genes containing each GO term (Methods).

https://doi.org/10.1038/s41588-025-02326-8

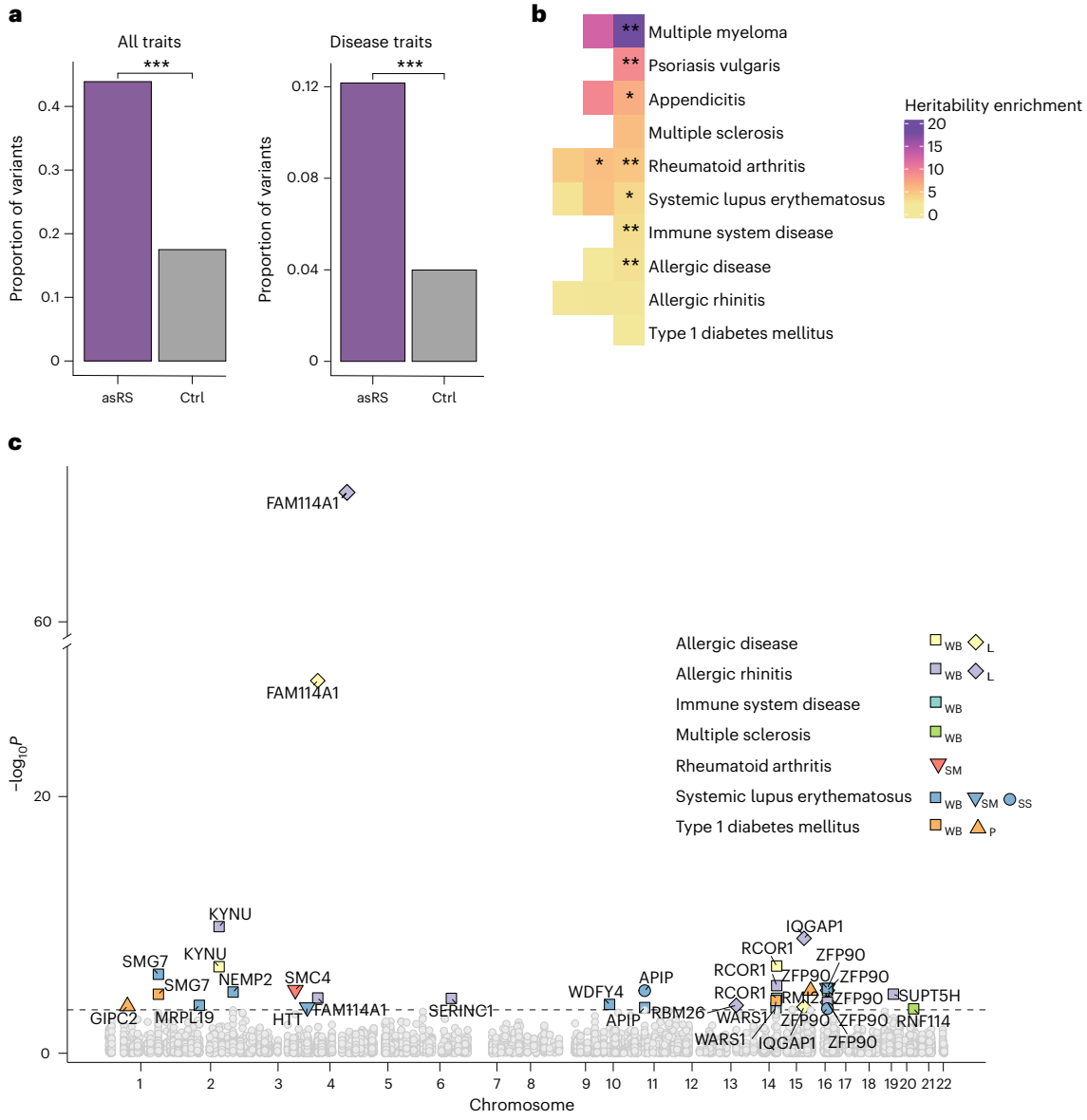


Fig. 6 | asRS events illuminate mechanisms underlying significant GWAS hits for various diseases. a, Proportion of asRS and control (Ctrl) variants matched with at least one significant GWAS hit based on shared tag SNPs. Top: all GWAS traits ($P = 1.31 \times 10^{-77}$); bottom: disease-related traits only ($P = 9.21 \times 10^{-23}$). **b**, Disease heritability enrichment estimates with $P < 0.1$ for asRS variants. Each cell represents the heritability enrichment from a different GWAS summary statistics file (*$P \leq 0.05$, **$P \leq 0.01$). **c**, Gene–trait prediction results by TWAS (all significant associations are labeled). *P* values are based on the estimated association between predicted asRS gene expression and disease as a weighted linear combination of SNP-trait standardized effect sizes and have been FDR-adjusted. Legend indicates GTEx tissues tested for each significant trait. L, lung; P, pancreas; SM, skeletal muscle; SS, suprapubic skin; WB, whole blood.

from disease-relevant tissues (Supplementary Table 8). These prediction models were then used to identify associations between asRS genes and traits using GWAS summary statistics of immune-related diseases. Restricting the models to genic SNPs helps minimize the influence of transcriptional regulatory variants, thus enriching for stability-mediated mechanisms. Across all evaluated asRS gene–trait pairs, 17 unique genes (out of 414 tested) were associated significantly with disease (Fig. 6c and Supplementary Table 8).

This TWAS-like analysis uncovered additional disease-related asRS genes that were not apparent from the direct overlap of GWAS hits with functional asRS variants. The strongest observed association was between *FAM114A1* and allergic rhinitis in lung tissue ($P = 5.41 \times 10^{-71}$). This gene encodes the nervous system overexpressed protein NOXP20 and has been implicated in regulating apoptosis in melanocytes[36] and angiotensin II signaling in cardiac cells[37], yet its role in inflammatory processes in the lung remains largely unexplored. Other notable genes include the nonsense-mediated decay factor *SMG7*, significantly associated with both systemic lupus erythematosus ($P = 7.01 \times 10^{-7}$ in whole blood) and type 1 diabetes mellitus ($P = 2.53 \times 10^{-5}$ in whole blood), as well as the E3 ubiquitin ligase protein-encoding gene *RNF114*, which displayed a significant association with multiple sclerosis ($P = 3.45 \times 10^{-4}$ in whole blood). Collectively, these findings strengthen the notion that genetically mediated mRNA stability represents a key mechanism contributing to the pathogenesis of immune-related diseases.

Discussion

In this study, we present a systematic analysis of allele-specific RNA stability, independent of transcriptional regulation, in human cells. Employing metabolic labeling data (Bru-seq/BruChase-seq) with the RNAtracker computational workflow, our approach sheds light on

the role of mRNA stability, distinguishing between gene abundance changes that result from transcriptional regulation versus decay rate variability. Because stability-regulating variants often reside within the mRNA, the allele-specific approach of RNAtracker ensures that the causal SNP is probably among variants we evaluate, making our workflow particularly effective for revealing stability-mediated regulatory events. On the other hand, variants that regulate transcription, such as promoters or enhancers, may reside outside of the genes. Nonetheless, RNAtracker can still capture transcriptionally regulated genes based on the read counts of its mRNA heterozygous variants—even though they may not necessarily be causal.

Future applications of RNAtracker can continue to extend our paradigm for understanding stability-mediated regulation of mRNA abundance. Indeed, the workflow can be adapted readily for use with any data that tracks the same population of RNA across different timepoints. These include other forms of uridine labeling[38] as well as collecting RNA at several timepoints after transcriptional inactivation. Whereas our analyses show that asRS and asRT events are largely cell-type-independent, variants regulated by cell-type-specific *trans*-acting factors (such as RBPs and miRNAs) may present a class of exceptions. We accounted for cellular context by limiting our analysis to cell line-specific miRNAs and using relevant GTEx tissue types when possible. Nonetheless, the generation of datasets from samples of more relevant cellular contexts will facilitate a more precise understanding of stability regulators.

In summary, we present a workflow for identifying stability-mediating variants and provide a comprehensive characterization of their biological roles. Our results highlight their contributions to disease and nominate functional explanations for poorly understood variant–trait associations, demonstrating RNA stability as a key link between genetic variants and disease.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-025-02326-8.

References

1. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
2. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
3. Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
4. Liebhaber, S. A. mRNA stability and the control of gene expression. *Nucleic Acids Symp. Ser.* **36**, 29–32 (1997).
5. Hollams, E. M., Giles, K. M., Thomson, A. M. & Leedman, P. J. MRNA stability and the control of gene expression: implications for human disease. *Neurochem. Res.* **27**, 957–980 (2002).
6. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
7. Tani, H. et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).
8. Courel, M. et al. GC content shapes mRNA storage and decay in human cells. *eLife* **8**, e49708 (2019).
9. LaMarre, J., Gingerich, T. J., Feige, J.-J. & LaMarre, J. AU-rich elements and the control of gene expression through regulated mRNA stability. *Anim. Health Res. Rev.* **5**, 49–63 (2004).
10. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
11. Wu, Q. et al. Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife* **8**, e45396 (2019).
12. Pai, A. A. et al. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet.* **8**, e1003000 (2012).
13. Alkallas, R., Fish, L., Goodarzi, H. & Najafabadi, H. S. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease. *Nat. Commun.* **8**, 909 (2017).
14. Li, J.-R., Tang, M., Li, Y., Amos, C. I. & Cheng, C. Genetic variants associated mRNA stability in lung. *BMC Genomics* **23**, 196 (2022).
15. Paulsen, M. T. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc. Natl Acad. Sci. USA* **110**, 2240–2245 (2013).
16. Bedi, K. et al. Co-transcriptional splicing efficiencies differ within genes and between cell types. *RNA* **27**, 829–840 (2021).
17. The GTEx Consortium et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
18. Salton, M. et al. Matrin 3 binds and stabilizes mRNA. *PLoS ONE* **6**, e23882 (2011).
19. Zhang, G. et al. Dynamic FMR1 granule phase switch instructed by m6A modification contributes to maternal RNA decay. *Nat. Commun.* **13**, 859 (2022).
20. Meyer, C. et al. The TIA1 RNA-binding protein family regulates EIF2AK2-mediated stress response and cell cycle progression. *Mol. Cell* **69**, 622–635 (2018).
21. Kim, Y. K. & Maquat, L. E. UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. *RNA* **25**, 407–422 (2019).
22. Yang, E.-W. et al. Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nat. Commun.* **10**, 1338 (2019).
23. Zhang, J. et al. An integrative ENCODE resource for cancer genomics. *Nat. Commun.* **11**, 3696 (2020).
24. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79**, 351–379 (2010).
25. Liu, C.-J. et al. miRNASNP-v3: a comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. *Nucleic Acids Res.* **49**, D1276–D1281 (2021).
26. Fu, T. et al. Massively parallel screen uncovers many rare 3′ UTR variants regulating mRNA abundance of cancer driver genes. *Nat. Commun.* **15**, 3335 (2024).
27. Griesemer, D. et al. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260 (2021).
28. Chen, P. J. et al. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **184**, 5635–5652 (2021).
29. Bresson, S. & Tollervey, D. Tailing off: PABP and CNOT generate cycles of mRNA deadenylation. *Mol. Cell* **70**, 987–988 (2018).
30. Springer, T. A. Adhesion receptors of the immune system. *Nature* **346**, 425–434 (1990).
31. González-Amaro, R., Diaz-González, F. & Sánchez-Madrid, F. Adhesion molecules in inflammatory diseases. *Drugs* **56**, 977–988 (1998).
32. Ryter, S. W., Cloonan, S. M. & Choi, A. M. K. Autophagy: a critical regulator of cellular metabolism and homeostasis. *Mol. Cells* **36**, 7–16 (2013).
33. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 37–49 (2021).
34. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).

35. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
36. Zhou, M. et al. Inhibition of Fam114A1 protects melanocytes from apoptosis through higher RACK1 expression. *Aging* **13**, 24740–24752 (2021).
37. Subbaiah, K. C. V., Wu, J., Tang, W. H. W. & Yao, P. FAM114A1 influences cardiac pathological remodeling by regulating angiotensin II signaling. *JCI Insight* **7**, e152783 (2022).
38. Imamachi, N. et al. BRIC-seq: a genome-wide approach for determining RNA stability in mammalian cells. *Methods* **67**, 55–63 (2014).

https://doi.org/10.1038/s41588-025-02326-8

Methods

Ethics

This research study did not require approval from any specific ethics board/committee.

CNV removal

We obtained absolute copy numbers generated by the Cancer Cell Line Encyclopedia (CCLE)[39] using the ABSOLUTE algorithm[40]. These copy number calls were overlapped with the GENCODE v.36 gene annotation. Only genes that overlapped copy number regions in which the minor ABSOLUTE copy number call and the major ABSOLUTE copy number call were equal to 1 were retained in downstream analysis.

If ABSOLUTE calls were not available, we used the CNVpytor[41] software with bin size set to 10 kb to analyze cell lines that had publicly available whole-genome sequencing data (Supplementary Table 1 for data sources). We then filtered the CNV calls by requiring P value < 0.0001, CNV size ≥50,000 and at least half of the reads to be uniquely mapped. We used the default mean-shift caller for cell lines that are diploid or near diploid and the joint-caller for cell lines that are known to be polyploid (Supplementary Table 1).

For the remaining cell lines, we downloaded Hi-C data in pairs format (standard text format for pairs of genomic loci given at 1 bp point positions) and used 'cload' from the cooler[42] software to convert these files into *.cool matrices at 20 kb resolution. We used the 'calculate-cnv' and 'segment-cnv' modules from NeoloopFinder[43] to identify CNV regions in each cell line using the Hi-C cool files as input. 'calculate-cnv' was run with the '–enzyme' parameter set to uniform and 'segment-cnv' was run with bin size 1,000 and ploidy set to two (default) for all cell lines except for Caco-2, in which ploidy was set to three. All genomic segments with copy number!= 2 were considered CNV regions. We filtered out genes if they overlapped any predicted CNV region.

Identification of asRS, asRT and mixed genes with RNAtracker

To categorize a gene as asRS, asRT or mixed, RNAtracker fits a beta-binomial mixture model (Supplementary Notes 2 and 3) for the reference allelic counts of the gene's testable SNVs (total read counts ≥10 and minor read count ≥2) at each timepoint. Combining three timepoints (0 h, 2 h, 6 h), RNAtracker categorizes genes into one of seven possible states (listed in the table below), where each state is a triplet corresponding to three timepoints. At each timepoint, a gene is encoded as 1 for ASE and 0 for non-ASE. We denote the total count of the ith SNV of a gene g at timepoint t by n_{gi}^t, among which we assume the reference allelic count follows beta-binomial($n_{gi}^t, \alpha_0^t, \beta_0^t,$) if gene g is non-ASE or beta-binomial($n_{gi}^t, \alpha_1^t,\ _1^t$) if gene g is ASE, $t = 0, 2, 6$. Since we do not want to assume that the reference allelic count is always greater than the alternative allelic count, we ensure the beta distributions are symmetrical by setting the two beta distribution parameters equal, that is, $\alpha_0^t = \beta_0^t, \alpha_1^t = \beta_1^t$. Specifically, in our implementation, we assume the reference allelic counts at 2 h and 6 h share the same parameters. To summarize, we have distributions at timepoints $t = 0$ following beta-binomial($n_{gi}^0, \alpha_0^0, \beta_0^0,$) if gene g is non-ASE or beta-binomial($n_{gi}^0, \alpha_1^1, \beta_1^1$) if gene g is ASE; at $t = 2$ or 6 following beta-binomial($n_{gi}^t, \alpha_0^{2,6}, \beta_0^{2,6}$) if gene g is non-ASE or beta-binomial($n_{gi}^t, \alpha_1^{2,6}, \beta_1^{2,6}$) if gene g is ASE.

	0h	2h	6h
State 0 (non-ASE)	0	0	0
State 1 (asRS)	0	1	1
State 2 (asRS)	0	0	1
State 3 (asRT)	1	1	1
State 4 (mixed)	1	0	1
State 5 (mixed)	1	0	0
State 6 (mixed)	1	1	0

First, in a preprocessing step, we focused on the 0 h data only and assumed that the reference allelic counts of each gene either follow the ASE beta-binomial distribution or the non-ASE beta-binomial distribution. Only genes with at least two testable SNVs are evaluated. We used π_1^0 to represent the probability of a gene being ASE (or $\pi_0^0 = 1 - \pi_1^0$ for a gene being non-ASE) at 0 h, where π refers to a fixed (nonrandom) parameter (or unknown constant) to be estimated. The expectation–maximization (EM) algorithm is then used to estimate the parameters $(\pi_1^0, \alpha_0^0, \alpha_1^0)$.

Upon convergence of the EM algorithm, we labeled each gene as non-ASE (0) or ASE (1) at 0 h based on the gene's posterior probabilities for the two states. Our assignment of genes to the two states is a two-step procedure: first, we assigned every gene to the state at which its posterior probability is greater than 0.5; second, based on the initially assigned genes, we retained a gene in a state only if its posterior probability at that state is at least (1) 0.95 or (2) the first tercile of the posterior probabilities of all genes initially assigned to that state.

Second, after determining whether the gene exhibits ASE at 0 h in the preprocessing step, RNAtracker jointly considers data from the 2 h and 6 h timepoints to categorize genes into one of the seven triplet states. For genes that are labeled non-ASE in the previous step, the EM algorithm is used to estimate parameters $\alpha_0^{2,6}, \alpha_1^{2,6}, \pi_0^{2,6}, \pi_1^{2,6}$, the first two of which are defined as the symmetric hyperparameters for the beta-binomial representing non-ASE and ASE genes, respectively, at each of the latter two timepoints (2 h and 6 h) and the last two of which are defined as the probability of a gene being in state 0 or state 1, respectively. The probability of a gene being in state 2 is $\pi_2^{2,6} = 1 - \pi_0^{2,6} - \pi_1^{2,6}$. Similarly, for genes that are labeled ASE in the previous step, the EM algorithm is used to estimate parameters $\alpha_0^{2,6}, \alpha_1^{2,6}, \pi_3^{2,6}, \pi_4^{2,6}, \pi_5^{2,6}$. Again, $\alpha_0^{2,6}$ and $\alpha_1^{2,6}$ are defined as the symmetric hyperparameters for the beta-binomial representing non-ASE and ASE genes respectively at each of the latter two timepoints (2 h and 6 h), and $\pi_3^{2,6}, \pi_4^{2,6}, \pi_5^{2,6}$ are defined as the probability of a gene being in state 3, state 4 or state 5, respectively. The probability of a gene being in state 6 is $\pi_6^{2,6} = 1 - \pi_3^{2,6} - \pi_4^{2,6} - \pi_5^{2,6}$.

We then used a procedure similar to the two-step procedure utilized in the preprocessing step, with an additional refinement step to assign genes to the seven states. In step 1, we assigned every gene to the state with the highest posterior probability. In step 2, based on the initially assigned genes, we retained a gene in a state only if its posterior probability at that state is at least (1) 0.95 or (2) the first tercile of the posterior probabilities of all genes initially assigned to that state. Finally, in step 3, for each gene, we require at least half of its SNVs at each ASE timepoint (encoded as 1 in the table above) to exhibit allelic imbalance in the same direction across the two replicates. In other words, at least half of the SNVs in the gene must have the same sign (reference allelic ratio − 0.5) for the two replicates. Genes that fail this threshold are considered uncategorized.

To summarize, for the three timepoints, the RNAtracker model has a total of ten independent parameters $(\alpha_0^0, \alpha_1^0, \pi_1^0, \alpha_0^{2,6}, \alpha_1^{2,6}, \pi_0^{2,6}, \pi_1^{2,6}, \pi_3^{2,6}, \pi_4^{2,6}, \pi_5^{2,6})$ and three parameters that are constrained by $\pi_0^0 = 1 - \pi_1^0$, $\pi_2^{2,6} = 1 - (\pi_0^{2,6} + \pi_1^{2,6})$ and $\pi_6^{2,6} = 1 - (\pi_3^{2,6} + \pi_4^{2,6} + \pi_5^{2,6})$. Our justification for the model parameterization (in which parameters for 0 h are estimated separately from the 2 h and 6 h data) is that we want to implement stringent thresholds for calling genes ASE or non-ASE at 0 h, as this timepoint is crucial to distinguishing asRS from asRT genes. Moreover, the distribution of reference allelic counts at 0 h was found to be distinct from that of 2 h and 6 h; hence, the beta-binomial parameters are the same for 2 h and 6 h, but different from 0 h.

For details on the criteria for testable SNVs, identification of ASE SNVs, estimation of initial hyperparameters, and an extended explanation of RNAtracker, please refer to Supplementary Note 1.

Cell-type comparisons

To calculate the background expectation of shared asRS genes between each pair of cell lines, we obtained the number of asRS genes in each cell line and divided each value by the number of common testable asRS genes. The product of these two values was used as the background expectation, that is, the expected proportion of overlap. A binomial test was used to evaluate whether the background expectation differed from the actual proportion of shared asRS genes between the two cell lines. The same analysis was applied to asRT genes.

GTex analysis

Significant GTEx *cis*-eQTLs were downloaded from the GTex portal (v.8 release) at https://www.gtexportal.org/home/datasets (filename: GTEx_Analysis_v8_eQTL.tar). Fisher's exact test was used to compute the odds ratio (that is, enrichment) of asRS or asRT variants overlapping significant GTEx *cis*-eQTLs compared to background variants. Background variants were those found in genes categorized as non-ASE by RNAtracker. For each overlap, we required the eQTL-associated gene to match the asRS, asRT or background gene. We also compared the enrichment of asRS and asRT genes among eGenes using Fisher's exact test. This analysis was performed per tissue using asRS or asRT events combined across all cell lines (Fig. 2b), as well as in each cell line separately (Extended Data Fig. 7b).

Deep transcriptomic profiling of ActD-treated cells

RNA-seq (NovaSeq X Plus 150 PE) was performed for GM12878, MCF-7 and HCT116 cells before treatment with 10 μg ml^{-1} (GM12878, HCT116) or 5 μg ml^{-1} (MCF-7) of ActD, as well as 2 h, 8 h and 24 h post-treatment (three replicates per timepoint). These reads were processed using the same procedure as the Bru/BruChase-seq data (that is, STAR mapping with WASP filtering, followed by obtaining read counts at heterozygous SNV positions). To confirm the genotypes for these three cell lines, we sequenced their genomic DNA and called variants using the GATK germline short-variant discovery pipeline (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels). To be considered validated, we required asRS genes to have at least one SNV that is non-ASE at 0 h but ASE at a latter timepoint. The allelic imbalance at the latter timepoint must be more extreme than the 0 h timepoint and at least 0.1 in at least one replicate. Alternatively, if the SNV is ASE at 0 h, its allelic imbalance must be more extreme at a latter timepoint and be at least 0.1 in at least one replicate. In either case, the direction of the allelic imbalance must be consistent with what is observed in the Bru/BruChase-seq data.

We used our previous approach in which we derive an empirical Gaussian distribution for the read coverage of each SNV to evaluate the probability that the average read count of the minor allele was generated from the same distribution as that of the major allele[22]. A SNV was deemed ASE if its Benjamini–Hochberg adjusted *P* value was less than 0.05. Allelic imbalance is calculated based on the delta allelic ratio at each SNV position: delta allelic ratio = abs(SNV allelic ratio − 0.5).

Prime editing

Prime editing was performed using the PE7 approach, which features a prime editor protein (PE7) fused to the RNA-binding, N-terminal domain of the small RNA-binding exonuclease protection factor La[44]. For each asRS variant that we evaluated with prime editing, we designed spacer and extension sequences for engineered prime editing guide RNAs (epegRNAs) using pegFinder[45]. pegLIT[46] was used to design linker patterns for each epegRNA. Golden Gate assembly was used to clone the spacer, extension and epegRNA scaffold sequences (Supplementary Table 9) into the pU6-tevopreq1-GG-acceptor vector (Addgene, catalog number 174038) for epegRNA constructs. We then transfected pCMV-PE7 (Addgene, catalog number 214812) and the plasmid expressing each epegRNA into HEK293T cells, respectively. gDNA was extracted 72 h post-transfection to confirm genome editing events. Total RNA was then harvested from cells 0 h (pretreatment) and 2 h, 8 h and 24 h after treatment with 10 μg ml^{-1} of ActD (three replicates per timepoint). gDNA was also harvested from cells before ActD treatment (0 h). After reverse transcription using the SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific, catalog number 18090010), the cDNA was amplified using gene-specific primers (Supplementary Table 9) to generate amplicons containing the variant of interest. Amplicons containing different variants from the same timepoint were pooled together before a second round of PCR to add Illumina adapters for sequencing. The PCR reactions were stopped before the plateau of the amplification curves. The libraries were purified using 2% agarose gel and sequenced with NovaSeq X Plus 150 PE.

Adapters were trimmed with bbduk (https://sourceforge.net/projects/bbmap/) before reads were mapped to GRCh38 with STAR (v.2.7.8a)[47]. To focus on reads from mature mRNA sequences, we filtered out unspliced reads before quantifying variant allelic counts with perbase (https://github.com/sstadick/perbase). Variants with a significantly different (Student's *t*-test; $P < 0.05$) allelic ratio at post-ActD treatment timepoints (2 h, 8 h or 24 h) compared to the 0 h pretreatment timepoint were identified as causal variants.

Massively parallel reporter assays

A total of 365 asRS variants (Supplementary Table 5) were included in the MPRA experiment (following the MapUTR[26] screening method) in HeLa cells. In brief, synthetic DNA oligonucleotides containing the variants of interest and their flanking sequences (164 nucleotides total) were cloned into the 3′ UTR of the *eGFP* gene. The expression of this reporter gene was driven by the cytomegalovirus early enhancer/chicken beta actin (CAG) promoter. These oligos were then introduced into HeLa cells by electroporation. Following electroporation (24 h), total RNA was extracted for sequencing targeting the tested variant regions. Specifically, the test sequences were amplified from both the plasmid library and mRNA to generate DNA sequencing and RNA-seq libraries. Three biological replicates were collected for each experiment and a high correlation was observed between replicates ($R = 0.84$). Sequencing data of the plasmid DNA and mRNA were compared to identify sites associated with significant expression differences between the two alleles using MPRAnalyze[48]. FDR ≤ 0.1 and $|\ln(FC)| \geq 0.1$ were required to call significance.

tamVars identified by MPRAu[27] were obtained from Supplementary Table 1 of the corresponding study. Variants identified as a tamVar in at least one of the tested cell lines were considered functional variants.

Functional enrichment analysis

Allele-specific binding sites were obtained from our previous work (Supplementary Data 2 from ref. 22. After removing coordinate-unstable positions[49], we converted ASB sites from hg19 to hg38 coordinates to be consistent with the asRS variants. eCLIP data for reproducible peaks (as determined from the irreproducible discovery rate approach[50]) were downloaded from the ENCODE portal. Annotations for RBP functions were obtained from Supplementary Data 1 from ref. 50.

rsids for SNPs overlapping miRNA seed regions that create or disrupt miRNA binding sites were downloaded from miRNASNPv3 (ref. 25). To be included in the enrichment analysis, these SNPs were required to be in the seed regions of miRNAs that were expressed in the cell line under consideration (nonzero read counts in miRNA-seq; Supplementary Table 1).

Each asRS variant was matched with a control variant that was sampled randomly from the same chromosome and type of genomic region (that is 3′ UTR, 5′ UTR, coding exon or exon in noncoding transcripts). asRS variants that appeared in more than one genomic context were assigned one control variant per genomic context. We overlapped all asRS and control variants with each set of functional annotations. For ASB and miRNA targeted sites, the proportion of asRS and control variants that overlapped each set of sites was calculated per cell line. Two-sided Wilcoxon's signed-rank test was then used to assess whether

the asRS variant proportion was significantly greater than the control variant proportion. For the eCLIP annotations, we used two-sided Fisher's exact test to calculate enrichment of asRS variants that overlapped each set of functional annotations compared to control variants. The enrichment test was performed using the combined list of asRS and control variants across all cell lines. A pseudocount of 1 was added to avoid division by zero errors.

GO enrichment analysis

GO terms were downloaded from Ensembl using biomaRt[51]. The enrichment analysis was performed using all asRS genes as the set of query genes. For each asRS gene, a random control gene with gene length and average gene expression (across all samples) within 10% relative to that of the asRS gene was chosen. A total of 10,000 sets of control genes were obtained and a Gaussian distribution was fit to the number of control genes containing each GO term. This distribution was used to calculate the enrichment *P* value of the GO term among all asRS genes. Focusing on significant (FDR < 0.05) GO terms with at least five asRS (or asRT) genes, we then used rrvgo[52] to group terms by semantic similarity (threshold = 0.7). rrvgo assigns parent terms to each group based on the GO term that has the most significant enrichment *P* value. Groups with two or more GO terms are shown in Fig. 4b. The 'innate immune response' cluster was renamed 'immune response' to more accurately describe the range of GO terms within this group.

GWAS catalog analysis

All reported associations were downloaded from the GWAS catalog (17 April 2023) and filtered to include variants that passed genome-wide significance at $P < 5 \times 10^{-8}$. We obtained GRCh38 genotype reference files from the 1000 Genomes project (subsampled for the EUR and CEU populations) (https://www.internationalgenome.org/data-portal/data-collection/grch38). Tag SNPs (required to be within 250 kb and exhibit $r^2 \geq 0.8$ with the target variant) were generated using plink (v.1.90)[53] for all 2,242 asRS variants that were present in the genotype reference files. The overall enrichment of asRS variants that shared tag SNPs (across all traits) with significant GWAS associations compared to a random set of control variants was computed using two-sided Fisher's exact test.

S-LDSC regression[54] was used to estimate disease heritability. This analysis was run on all available harmonized summary statistics (31 January 2025) from GWAS catalog that were categorized under the EFO term EFO0000540 (immune system disease). Variant sets were defined by all genic variants inside asRS genes as determined by RNAtracker. LD scores for the regression were calculated using genotype reference files from 1000 Genomes project EUR samples within the variant sets for each chromosome. Disease heritability was then calculated using summary statistics for each disease of interest (Supplementary Table 7b), and s.e. values for heritability estimates were computed using the jackknifing approach. We required the enrichment s.e. to be less than the estimated enrichment for the result to be reported in Supplementary Table 7b.

Gene prediction of disease

Gene prediction models were built using the FUSION.compute_weights.R script from http://gusevlab.org/projects/fusion/. We matched each trait of interest to disease-relevant GTEx tissues (Supplementary Table 8). Subsequently, the genotypes and gene expression from each GTEx tissue of interest was used to compute gene prediction models for matched traits using variants that resided within asRS genes. The FUSION.assoc_test.R script was then used to estimate gene-disease associations.

Statistics and reproducibility

No statistical method was used to determine sample size. Sample size was set based on the number of Bru/BruChase-seq samples available through the ENCODE portal. We excluded data from cell lines that had an insufficient (<100) number of testable genes after CNV filtering. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Randomization and blinding were not relevant for our study given that samples were not allocated into experimental groups. The software (including specific version) and statistical tests used in the data analysis have been reported in Methods to facilitate reproducibility of the results.

Reporting summary

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

Data availability

Bru-seq/BruChase-seq from 16 human cell lines (GM12878, HCT116, HepG2, IMR-90, K562, MCF-7, PC-3, Panc1, PC-9, A673, MCF10A, Calu3, Caco-2, OCI-LY7, endothelial cell of umbilical vein (HUVEC) and mammary epithelial cell (HMEC)) were downloaded from the ENCODE data portal (https://www.encodeproject.org/). Accession IDs can be found in Supplementary Table 1b. The GRCh38 reference genome and gene annotation can be found at https://www.gencodegenes.org/human/release_36.html (filenames: GRCh38.primary_assembly.genome.fa.gz; gencode.v36.primary_assembly.annotation.gtf.gz). Significant GTEx *cis*-eQTLs were downloaded from the GTex portal (v.8 release) at https://www.gtexportal.org/home/datasets (GTEx_Analysis_v8_eQTL.tar). ABSOLUTE CNVs from CCLE can be obtained from https://depmap.org/portal/data_page/?release=CCLE+2019&file=CCLE_ABSOLUTE_combined_20181227.xlsx&tab=allData. Allele-specific binding sites were obtained from our previous work (Supplementary Data 2 from ref. 22). SNPs overlapping miRNA seed regions that create or disrupt miRNA binding sites were downloaded from miRNASNPv3 (ref. 25). eCLIP data for reproducible peaks (as determined from the irreproducible discovery rate approach[50]) were downloaded from the ENCODE portal. ActD RNA-seq data can be accessed on GEO (Series record GSE276016). MapUTR sequencing data can be accessed on GEO (Series record GSE298114). CRISPR editing results can be accessed on GEO (Series record GSE298112). All GWAS summary statistics used in this paper can be downloaded from the GWAS catalog (https://www.ebi.ac.uk/gwas/; accession codes in Supplementary Table 8a). GRCh38 genotype reference files from the 1000 Genomes project can be found at https://www.internationalgenome.org/data-portal/data-collection/grch38. Source data are provided with this paper.

Code availability

Code for reproducing the RNAtracker gene categorization results and other data analysis scripts is available via GitHub at https://github.com/gxiaolab/RNAtracker and via Zenodo at https://doi.org/10.5281/zenodo.15528784 (ref. 55). We used bbduk from the BBmap package (v.38.91) (https://sourceforge.net/projects/bbmap/) for read adapter trimming, STAR[47] (v.2.7.8a) for read mapping, Picard Tools (https://broadinstitute.github.io/picard/) (v.1.94) to remove PCR duplicates and extract uniquely mapped reads, NeoloopFinder[43] (v.0.3.0) for CNV predictions, rrvgo[52] (v.1.6.0) for GO enrichment analysis, PLINK[56] (v.1.9) to obtain tag SNPs and bedtools (v.2.30.0)[57] to overlap genomic regions. Perbase (v.0.10.0) (https://github.com/sstadick/perbase) was used to obtain variant allelic counts in the CRISPR prime editing sequencing data. MPRAnalyze[48] was used to identify functional variants in the MapUTR data. S-LDSC[54] was used to estimate disease heritability. FUSION.compute_weights.R from http://gusevlab.org/projects/fusion/ was used to build gene prediction models. For analyzing whole-genome sequencing data, we used bwa mem[58] (v.0.7.17) for read mapping and CNVpytor[41] (v.1.3.1) for identifying CNV regions. The VGAM[59] (v.1.1) R package was used to compute probability density values and simulate allelic counts. Genome coordinate conversions were performed using liftOver (https://www.bioconductor.org/packages/release/workflows/html/liftOver.html). Other R packages used for plotting include ComplexUpSet[60], ComplexHeatmap[61] and AllelicImbalance[62].

References

39. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
40. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
41. Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *Gigascience* **10**, giab074 (2021).
42. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
43. Wang, X. et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
44. Yan, J. et al. Improving prime editing with an endogenous small RNA-binding protein. *Nature* **628**, 639–647 (2024).
45. Chow, R. D., Chen, J. S., Shen, J. & Chen, S. A web tool for the design of prime-editing guide RNAs. *Nat. Biomed. Eng.* **5**, 190–194 (2021).
46. Nelson, J. W. et al. Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol.* **40**, 402–410 (2022).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Ashuach, T. et al. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
49. Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22**, bbab069 (2021).
50. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
51. Smedley, D. et al. BioMart–biological queries made easy. *BMC Genomics* **10**, 22 (2009).
52. Sayols, S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. *microPubl. Biol.* https://doi.org/10.17912/micropub.biology.000811 (2023).
53. 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).
54. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
55. gxiaolab. Gxiaolab/RNAtracker: for publication. *Zenodo* https://doi.org/10.5281/zenodo.15528784 (2025).
56. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Human Genet.* **81**, 559–575 (2007).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Yee, T. W. The VGAM package for categorical data analysis. *J. Stat. Softw.* **32**, 1–34 (2010).
60. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
61. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
62. Gådin, J. R., van't Hooft, F. M., Eriksson, P. & Folkersen, L. AllelicImbalance: an R/bioconductor package for detecting, managing, and visualizing allele expression imbalance data from RNA sequencing. *BMC Bioinformatics* **16**, 194 (2015).

Acknowledgements

We thank members of the Xiao laboratory for helpful discussions and comments on this work. This work was supported in part by grants from the National Institutes of Health (U01HG009417 and R01AG075206 to X.X.). E.H. was supported by the Graduate Research Fellowship of the NSF under Grant No. DGE-2034835. T.F. was supported by the UCLA Hyde Fellowship and Dissertation Year Fellowship. K.A. was supported by the University of California-Historically Black Colleges and Universities (UC-HBCU) Fellowship. S.T. was supported by the NIH T32GM145388. M.L. was supported by NHGRI grant UM1 HG009382. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

E.H., L.Z. and X.X. designed the study with inputs from all other authors. E.H., L.Z., K.A., R.Y. and J.H. conducted the bioinformatics works. E.H., G.Y. and J.J.L. worked on the statistical modeling. T.F., S.T., T.L.N., C.G.-F., A.K., J.H.B. and R.V. conducted the molecular, cellular and biochemical experiments. M.T.P. and B.M. generated the Bru-seq/BruChase-seq data. X.X., J.J.L. and M.L. provided supervisory inputs. All authors contributed to the writing of the paper. All authors approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

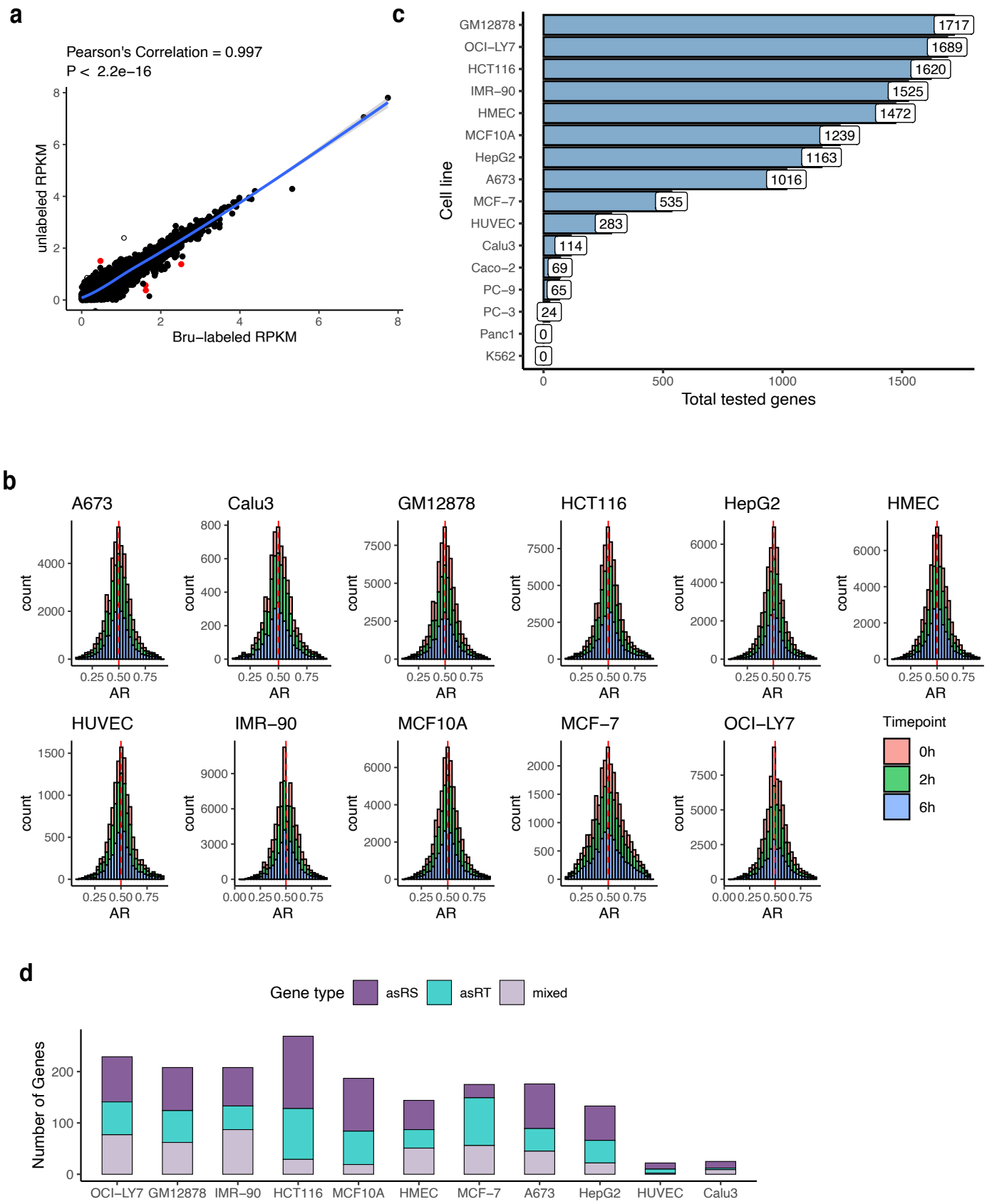
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02326-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02326-8.

Correspondence and requests for materials should be addressed to Xinshu Xiao.

Peer review information *Nature Genetics* thanks Michael Hagemann-Jensen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02326-8

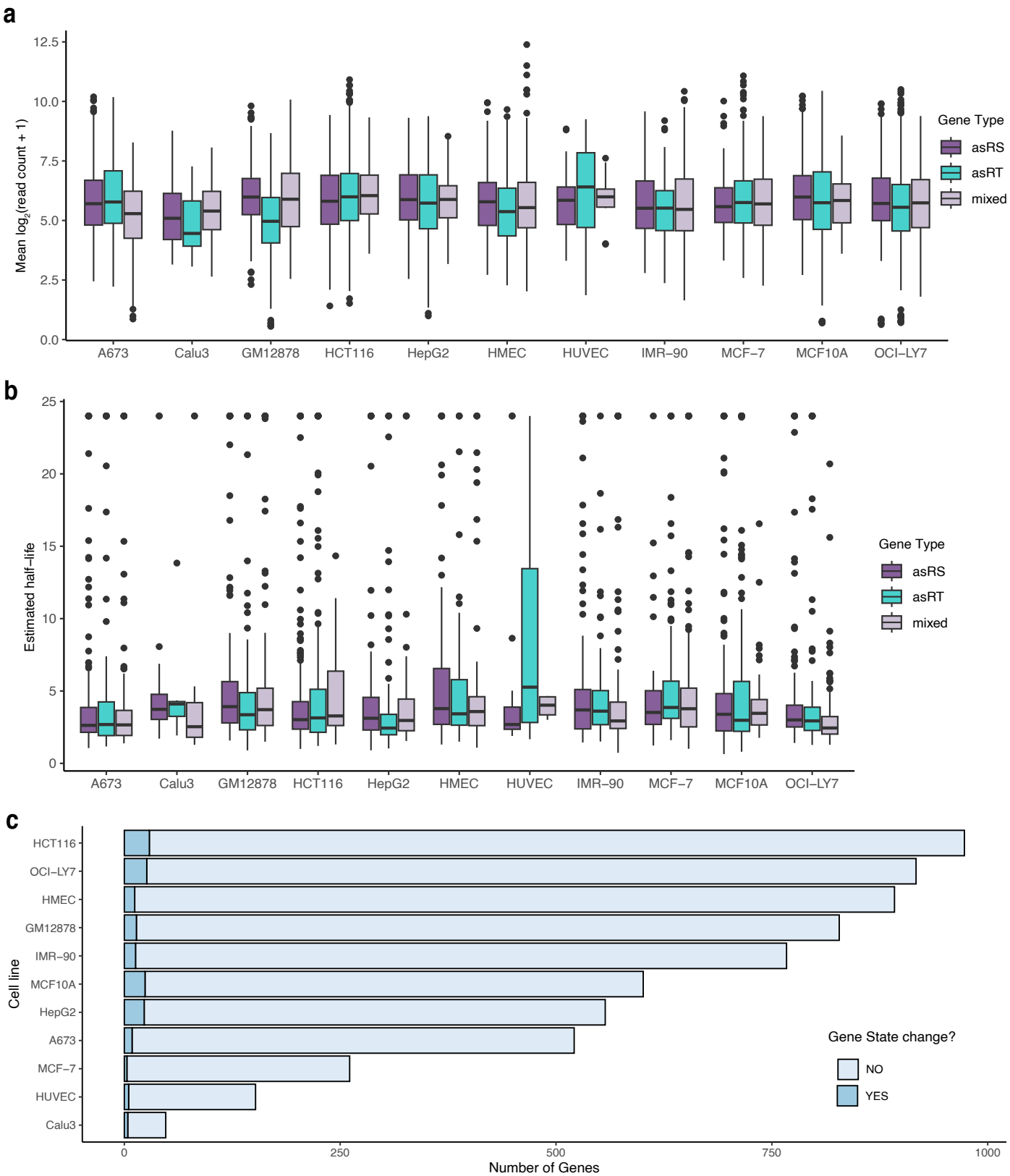


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-025-02326-8

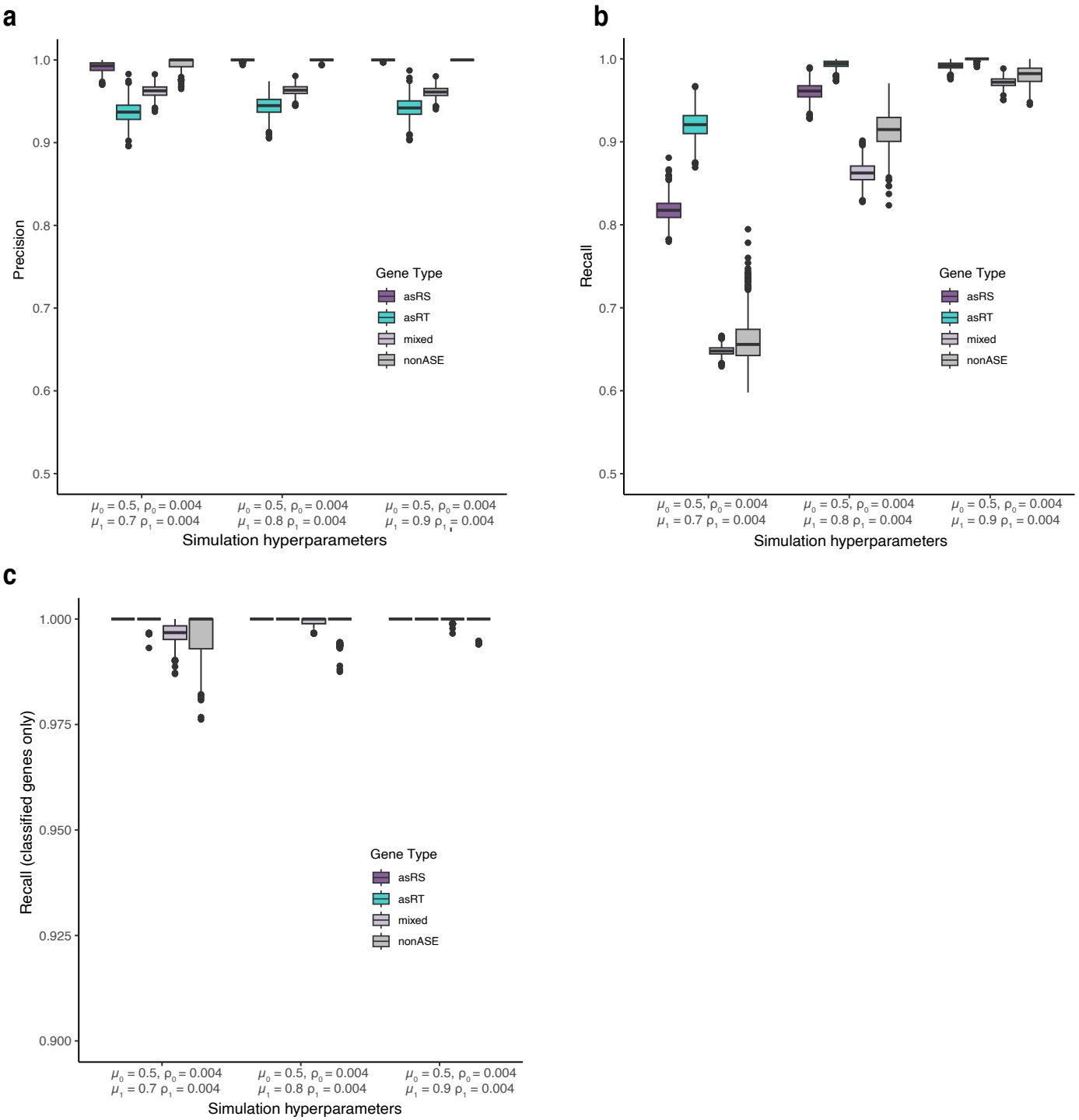
Extended Data Fig. 1 | RNAtracker facilities the classification of ASE genes. **a**, Transcriptomic comparison of Bru-labeled vs. unlabeled K562 cells based on two-sided Pearson's correlation test ($p < 2.2$ e-16). **b**, Allelic ratio distribution after copy-number variant (CNV) removal for the 11 cell lines with >100 genes eligible for classification. Allelic ratio (AR) is calculated by dividing the number of reference allelic counts by total counts per variant. **c**, Number of genes eligible for classification by RNAtracker in each cell line. **d**, Number of genes identified as asRS, asRT, or mixed.

https://doi.org/10.1038/s41588-025-02326-8



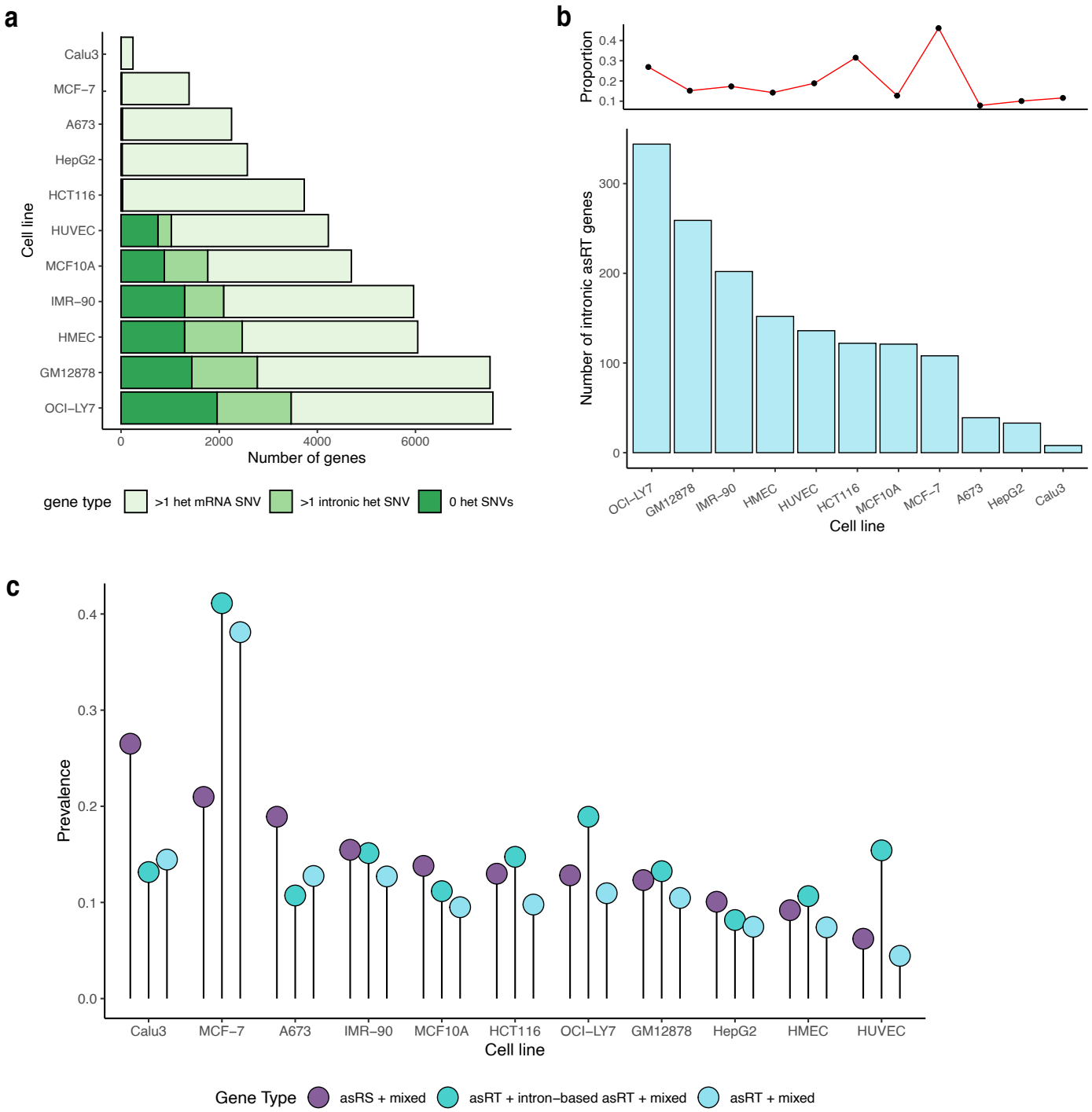
Extended Data Fig. 2 | Read coverage, gene half-life, and alternative-splicing contribute minimally to RNAtracker performance. a, Coverage of asRS, asRT, and mixed genes. For each gene, we take the average coverage across all genic regions. **b**, Estimated half-lives of asRS, asRT, and mixed genes. In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5 * the interquartile range. **c**, Number of genes with and without gene classification changes after removing SNVs in alternatively spliced regions.

https://doi.org/10.1038/s41588-025-02326-8



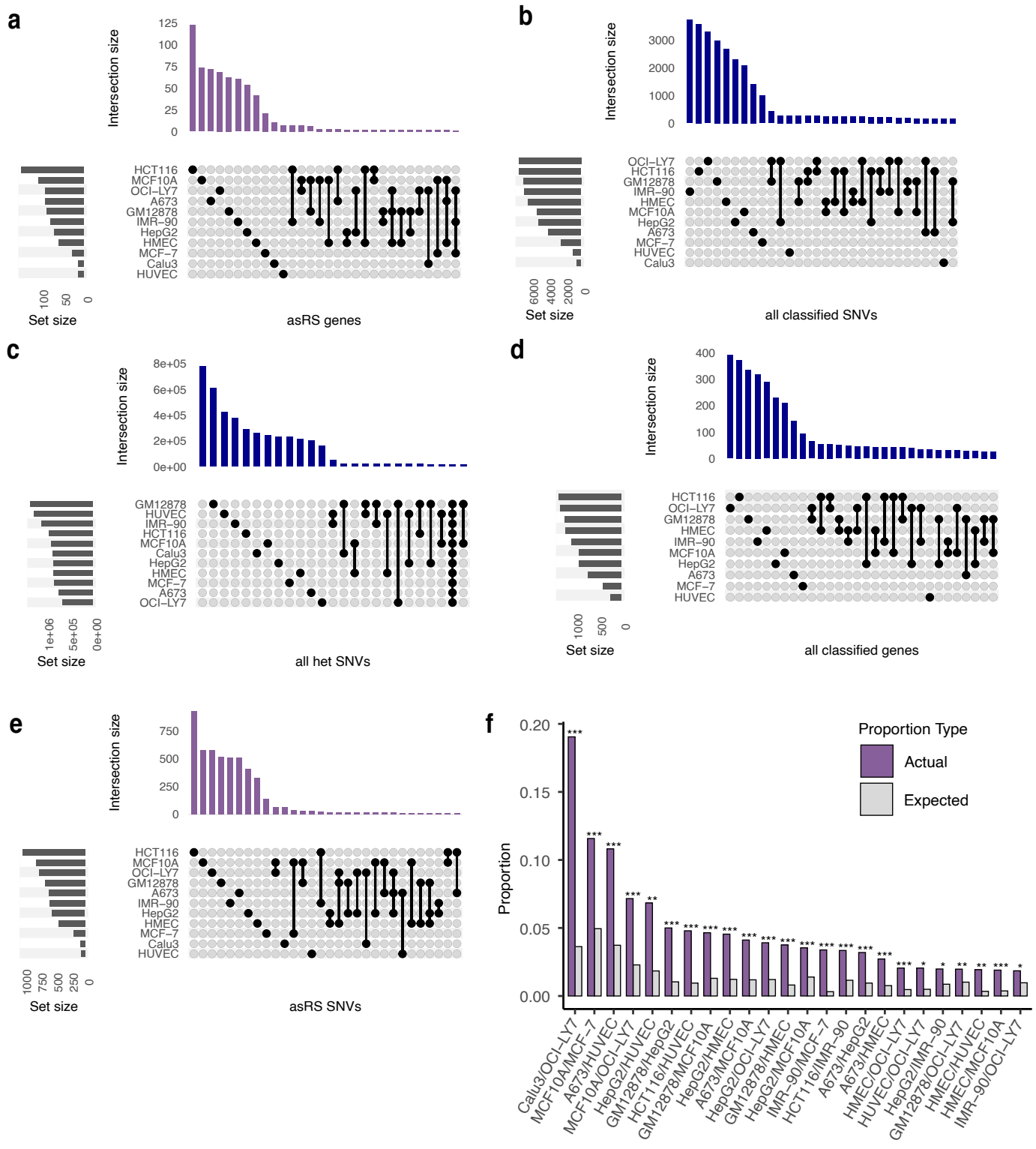
Extended Data Fig. 3 | RNAtracker exhibits high Precision and Recall in simulations. a,**b**,**c** 1,000 allelic count datasets were simulated by sampling reads from beta-binomial models using hyperparameters representing 3 different allelic imbalance conditions to assess Precision (**a**), Recall (**b**), and Recall among genes passing classification thresholds only (**c**). In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5 * the interquartile range.

https://doi.org/10.1038/s41588-025-02326-8



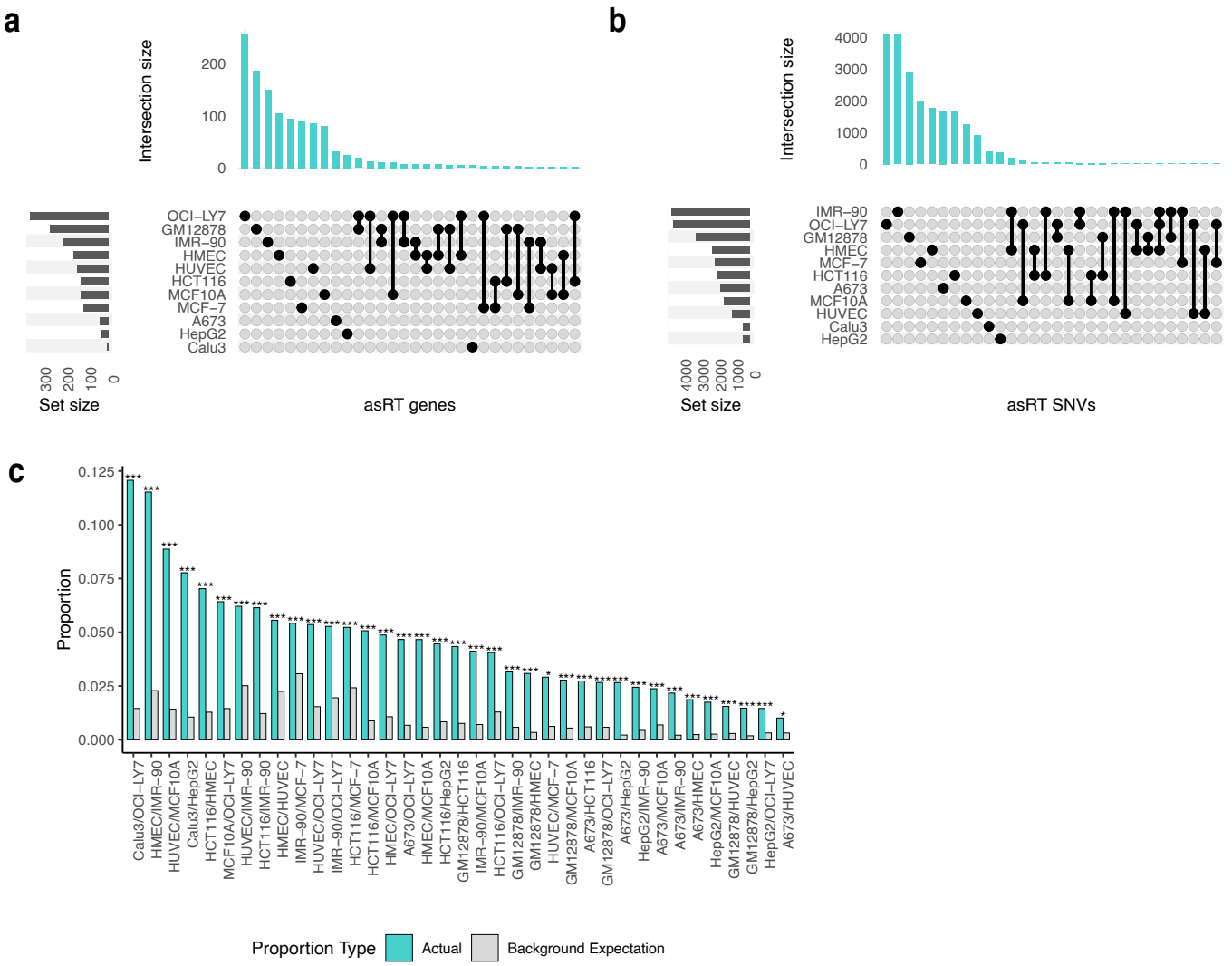
Extended Data Fig. 4 | Bias against asRT identification varies across cell lines. a, Number of expressed genes with and without heterozygous genic single-nucleotide variants (SNVs). Genes without heterozygous SNVs are further categorized into whether they have intronic heterozygous SNVs or 0 heterozygous SNVs (even when introns are considered). To be considered expressed, a gene must have average base coverage ≥10 across all genic regions, across all 6 timepoint samples. **b**, Prevalence (top panel) and number (bottom panel) of intron-based asRT genes. Prevalence is calculated using the total number of genes that are testable based on having SNVs in intronic regions. **c**, Comparison of the prevalence of stability-regulated genes versus transcriptionally regulated genes (including and excluding intron-based asRT genes). Stability regulated genes include asRS and mixed genes. Transcriptionally regulated genes include asRT, intronic asRT, and mixed genes. To calculate prevalence, the number of genes falling under each of these categories is summed and divided by the total number of genes that were tested using RNAtracker in the cell line.

https://doi.org/10.1038/s41588-025-02326-8



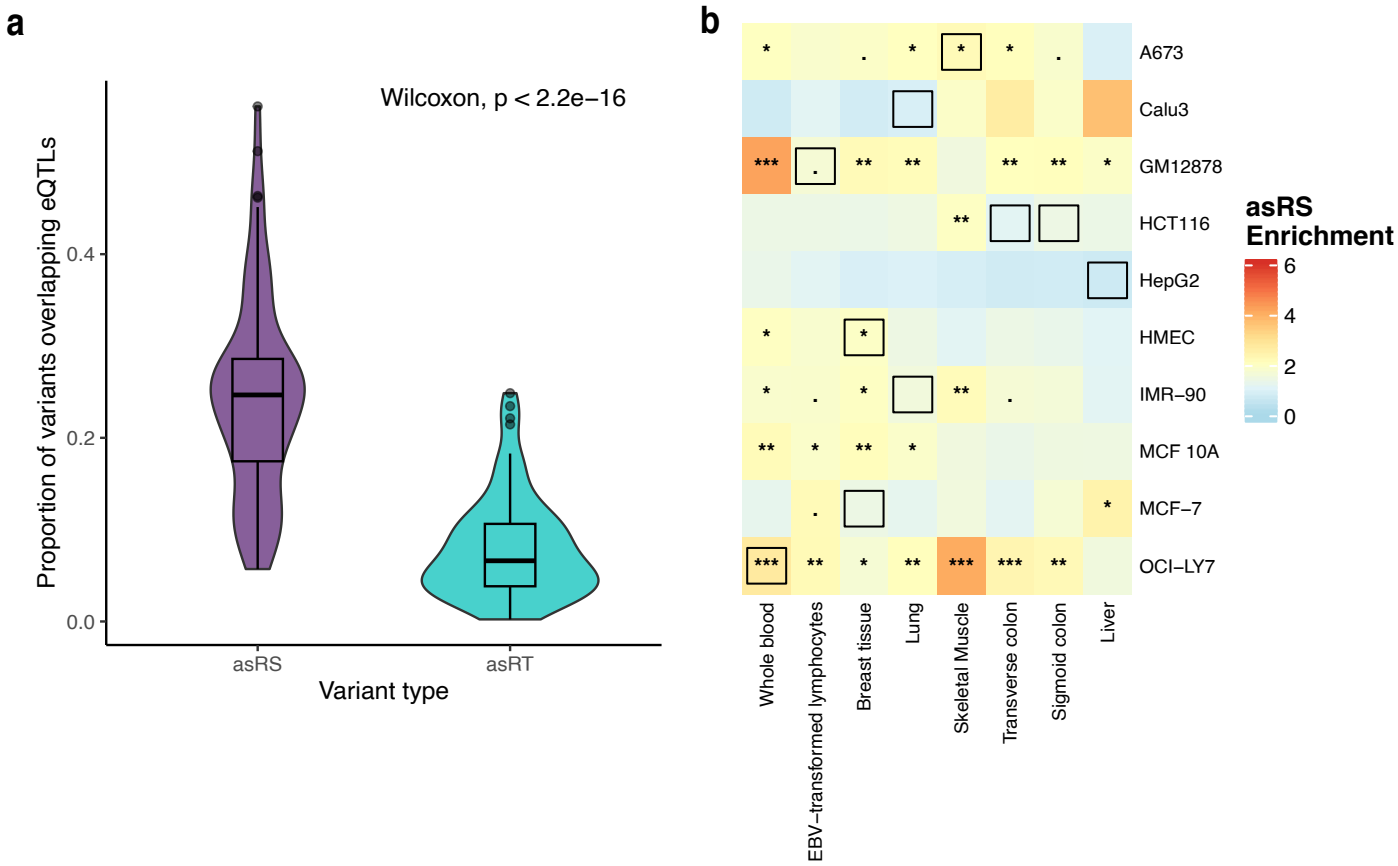
Extended Data Fig. 5 | Low proportion of asRS sharing across cell lines can be attributed to their unique genetic backgrounds. a, UpSet plot of asRS genes that are unique to or shared across cell lines. **b**,**c** UpSet plot of heterozygous single-nucleotide variants (SNVs) within genes classified by RNAtracker (**b**), as well as all (including intronic) heterozygous SNVs (**c**) that are unique to or shared across cell lines. Note that Calu3 is not shown because only the top 30 largest intersections are plotted. **d**, UpSet plot of all genes categorized by RNAtracker across cell lines. **e**, UpSet plot of asRS variants that are unique to or shared across cell lines. **f**, Pairs of cell lines that exhibited a significant difference between the expected and actual proportion of overlapping asRS variants. All 24 comparisons had significantly greater actual proportion than expected. *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001 (two-sided binomial test).

https://doi.org/10.1038/s41588-025-02326-8



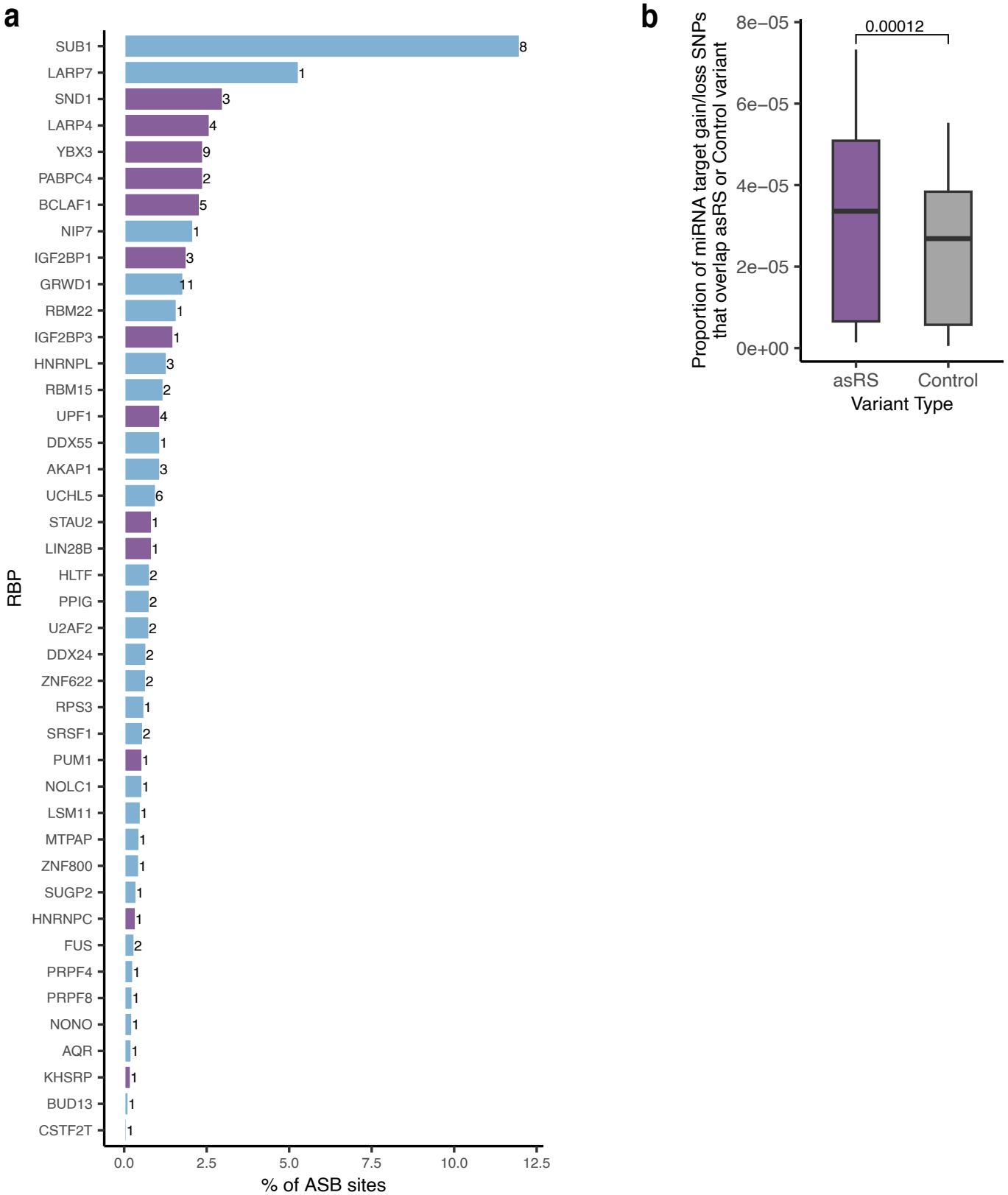
Extended Data Fig. 6 | Low proportion of asRT sharing across cell lines can be attributed to their unique genetic backgrounds. a, UpSet plot of asRT genes shared across cell lines. **b**, UpSet plot of asRT variants shared across cell lines. **c**, Pairs of cell lines that exhibited a significant difference between the expected and actual proportion of overlapping asRT variants. 33 out of 33 of these comparisons had significantly greater actual proportion than expected. *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001 (two-sided binomial test).

https://doi.org/10.1038/s41588-025-02326-8



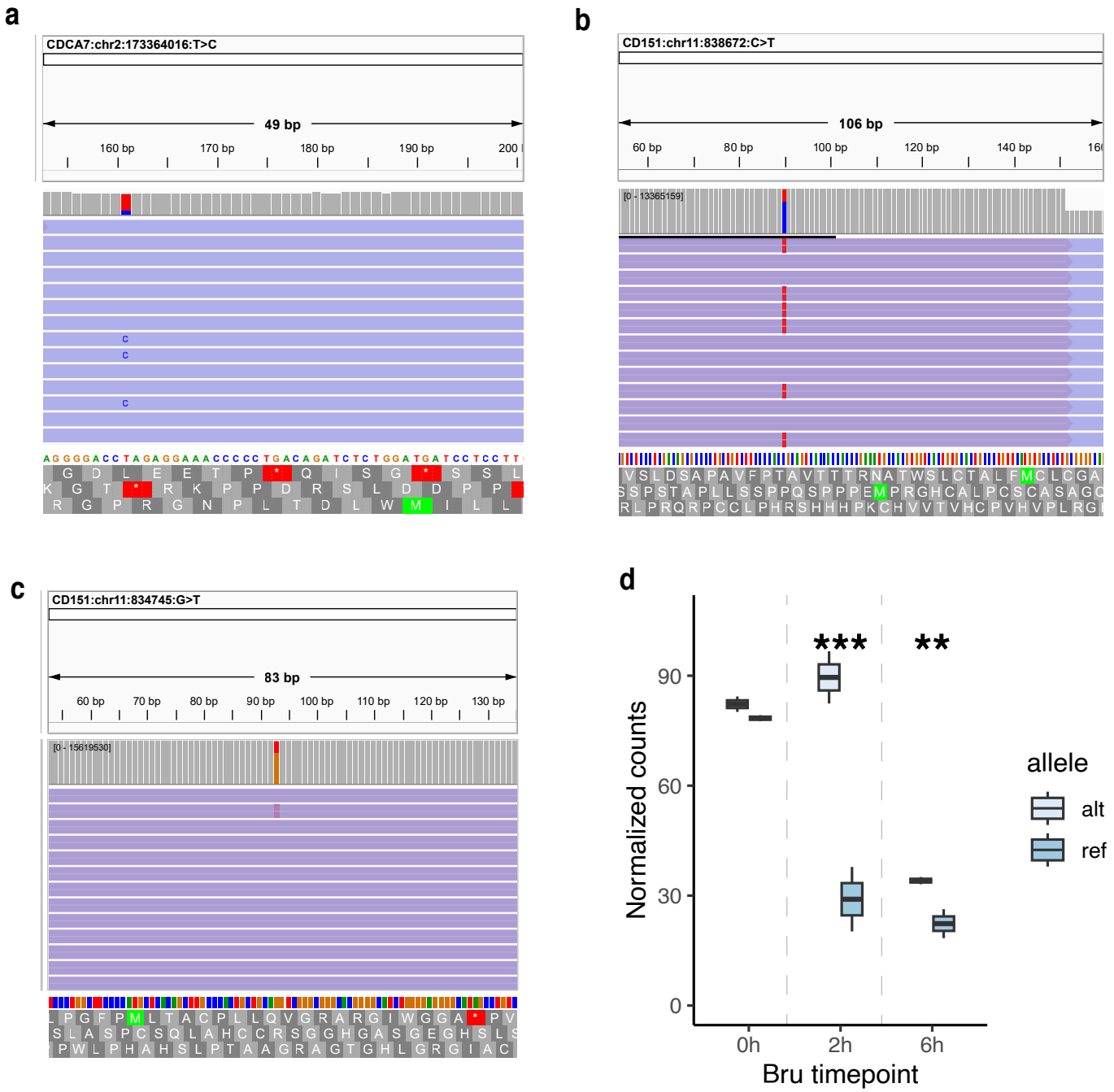
Extended Data Fig. 7 | asRS and asRT events are both important contributors to gene expression. a, Proportion of asRS and asRT variants in each cell line ($n = 11$) that overlapped expression quantitative trait loci (eQTLs) ($p = 3.71e\text{-}20$). P value was calculated via a two-sided Wilcoxon's signed rank-test. In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5 * the interquartile range. **b**, Enrichment (that is Fisher's exact test odds ratio) of asRS genes that overlap eGenes (compared to asRT genes).*p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001. None of the comparisons in which the asRT overlap proportion was higher than the asRS overlap proportion were significant.

https://doi.org/10.1038/s41588-025-02326-8



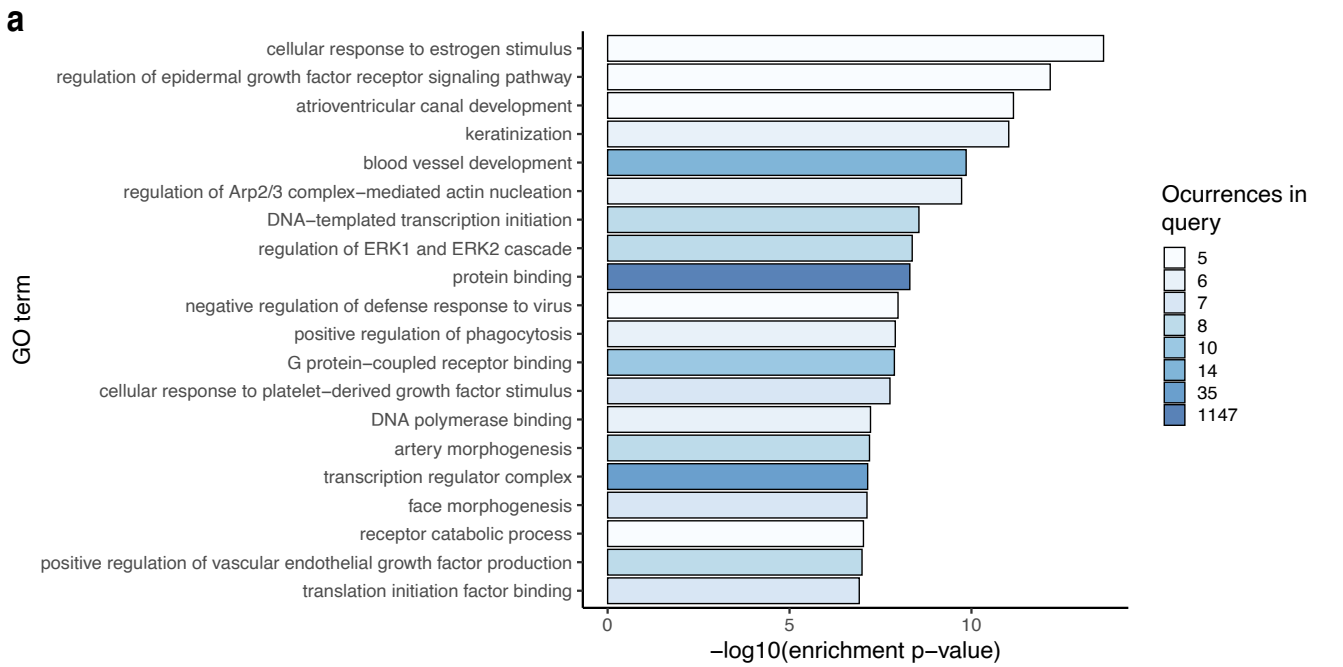
Extended Data Fig. 8 | asRS variants may function by disrupting interactions with trans-regulatory factors. a, RNA binding proteins (RBPs) with allele-specific binding (ASB) sites that overlap asRS variants. X axis shows percentage of each RBP's ASB sites that overlap asRS variants. Number of ASB sites that overlap asRS variants is shown to the right of each bar. Purple: RBPs involved with RNA stability and decay according to previous manual literature curation. **b**, Proportion of miRNA target gain/loss SNPs that overlap asRS or control variants in each cell line (n = 11). In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5 * the interquartile range.

https://doi.org/10.1038/s41588-025-02326-8



Extended Data Fig. 9 | Prime editing supports the causality of asRS variants. **a**-**c**, Genomic DNA sequencing supports the successful genome editing of chr2:173364016:T > C (**a**), chr11:838672:C > T (**b**), and chr11:834745:G > T (**c**). **d**, Comparison of SNV normalized counts in Bru/BruChase-seq data (2 biological replicates per timepoint) for chr2:173364016:T > C in *CDCA7*. In boxplots, minima/maxima represent least/greatest proportion values, bounds show 25th and 75th percentiles, and whiskers indicate values within 1.5 * the interquartile range.

https://doi.org/10.1038/s41588-025-02326-8



Extended Data Fig. 10 | asRS and asRT genes are involved in various pathways. a, Top 20 enriched Gene Ontology (GO) terms for asRT genes. P values were derived from an empirical Gaussian distribution of number of control genes containing each GO term (Methods).

nature portfolio

Corresponding author(s): Xinshu Xiao

Last updated by author(s): May 27, 2025

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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 for data collection.
Data analysis	Code for reproducing the RNAtracker gene categorization results is available at https://github.com/gxiaolab/RNAtracker. We used bbduk from the BBmap package (v. 38.91) (https://sourceforge.net/projects/bbmap/) for read adaptor trimming, STAR (v2.7.8a) for read mapping, Picard Tools (https://broadinstitute.github.io/picard/) (v1.94) to remove PCR duplicates and extract uniquely mapped reads, NeoloopFinder (v0.3.0) for CNV predictions, rrvgo (v1.6.0) for GO enrichment analysis, PLINK (v1.9) to obtain tag SNPs, and bedtools (v2.30.0) to overlap genomic regions. Perbase (v0.10.0) (https://github.com/sstadick/perbase) was used to obtain variant allelic counts in the CRISPR prime editing sequencing data. MPRAnalyze was used to identify functional variants in the MapUTR data. S-LDSC was used to estimate disease heritability. FUSION.compute_weights.R from http://gusevlab.org/projects/fusion/ was used to build gene prediction models. For analyzing WGS data, we used bwa mem (v0.7.17) for read mapping and CNVpytor (v.1.3.1) for identifying copy-number variant regions. The VGAM (v1.1) R package was used to compute probability density values and simulate allelic counts. Genome coordinate conversions were performed using liftOver (https://www.bioconductor.org/packages/release/workflows/html/liftOver.html). Other R packages used for plotting include ComplexUpSet, ComplexHeatmap, and AllelicImbalance.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Bru-seq/BruChase-seq from 16 human cell lines (GM12878, HCT116, HepG2, IMR-90, K562, MCF-7, PC-3, Panc1, PC-9, A673, MCF10A, Calu3, Caco-2, OCI-LY7, endothelial cell of umbilical vein (HUVEC) and mammary epithelial cell (HMEC)) were downloaded from the ENCODE data portal (https://www.encodeproject.org/). Accession IDs can be found in Supplementary Table 1. The GRCh38 reference genome and gene annotation can be found at https://www.gencodegenes.org/human/release_36.html (filenames: GRCh38.primary_assembly.genome.fa.gz; gencode.v36.primary_assembly.annotation.gtf.gz). Significant GTEx cis-eQTLs were downloaded from the GTex portal (V8 release) at https://www.gtexportal.org/home/datasets (GTEx_Analysis_v8_eQTL.tar). ABSOLUTE copy-number variants from CCLE can be obtained from https://depmap.org/portal/data_page/?release=CCLE+2019&file=CCLE_ABSOLUTE_combined_20181227.xlsx&tab=allData. Allele-specific binding sites were obtained from our previous work rsids for SNPs overlapping miRNA seed regions that create or disrupt miRNA binding sites were downloaded from miRNASNPv3. eCLIP data for reproducible peaks (as determined from the irreproducible discovery rate, or IDR, approach) were downloaded from the ENCODE portal. ActD RNA-seq data can accessed on GEO (Series record GSE276016). MapUTR sequencing data can be accessed on GEO (Series record GSE298114). CRISPR editing results can be accessed on GEO (Series record GSE298112). All GWAS summary statistics used in this paper can be downloaded from the GWAS catalogue (https://www.ebi.ac.uk/gwas/; accession codes in Supplementary Table 8). GRCh38 genotype reference files from the 1000 Genomes project can be found at: https://www.internationalgenome.org/data-portal/data-collection/grch38.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	No statistical method was used to determine sample size. Sample size was set based on the number of Bru-Seq/BruChase-seq replicates that were available per cell line through the ENCODE portal. Sample size was sufficient to identify asRS and asRT events.
Data exclusions	We included Bru-Seq/BruChase-seq samples from 11 out of the 16 deeply profiled ENCODE cell lines. The excluded cell lines had insufficient (<100) number of testable genes.
Replication	We required genes to have at least two testable SNVs in order to be evaluated by RNAtracker to increase the chances that we were giving RNAtracker enough information to make a reliable categorization. While we did not verify reproducibility using an independent Bru/BruChase-seq dataset, ActD was used as an alternative method for identifying stability-regulated genes.
Randomization	Randomization was not relevant for our study of genes regulated by RNA stability and transcriptional regulation as samples were not allocated into experimental groups.
Blinding	Blinding was not relevant for our study of genes regulated by RNA stability and transcriptional regulation as samples were not allocated nto experimental groups.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Source: ATCC
Authentication	None
Mycoplasma contamination	Tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	None

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA