Integrated multiplexed assays of variant effect reveal *cis*-regulatory determinants of catechol-*O*-methyltransferase gene expression

Ian Hoskins[a], Shilpa Rao[a], Charisma Tante[a], Can Cenik[a,*]

Affiliations:

[a] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA

* Corresponding author: ccenik@austin.utexas.edu

Keywords
MAVE, COMT, mutagenesis, ribosome load, RNA secondary structure, variant effects

Abstract:
Multiplexed assays of variant effect are powerful methods to profile the consequences of rare variants on gene expression and organismal fitness. Yet, few studies have integrated several multiplexed assays to map variant effects on gene expression in coding sequences. Here, we pioneered a multiplexed assay based on polysome profiling to measure variant effects on translation at scale, uncovering single-nucleotide variants that increase and decrease ribosome load. By combining high-throughput ribosome load data with multiplexed mRNA and protein abundance readouts, we mapped the *cis*-regulatory landscape of thousands of catechol-*O*-methyltransferase (*COMT*) variants from RNA to protein and found numerous coding variants that alter *COMT* expression. Finally, we trained machine learning models to map signatures of variant effects on *COMT* gene expression and uncovered both directional and divergent impacts across expression layers. Our analyses reveal expression phenotypes for thousands of variants in *COMT* and highlight variant effects on both single and multiple layers of expression. Our findings prompt future studies that integrate several multiplexed assays for the readout of gene expression.

Introduction:
Genome-wide association studies have identified regulatory variants that impact RNA expression, transcript isoform usage, and translation (Battle et al., 2015; Cenik et al., 2015; GTEx Consortium, 2020; Lappalainen et al., 2013; Li et al., 2013); however, these methods have been limited to surveying common population variants. Consequently, interpreting the impact of rare genetic variation on gene expression remains a major challenge. Multiplexed assays of variant effect (MAVEs) employ deep sequencing to profile the phenotypic effects of thousands of genetic variants in a single target gene at once (Esposito et al., 2019). MAVEs have been developed to assay specific phenotypic readouts such as protein activity (Chiasson et al., 2020; Majithia et al., 2016; Starita et al., 2015; Suiter et al., 2020), protein-protein interactions (Faure et al., 2022; Fowler et al., 2010; Starr et al., 2020), mRNA abundance (Findlay et al., 2018; Kircher et al., 2019), protein abundance (Chiasson et al., 2020; Matreyek et al., 2018; Suiter et al., 2020), and organismal fitness (Sun et al., 2016; Weile et al., 2017).

Even though variant effects on several layers of gene expression have been assayed, no MAVE has been developed to measure coding variant effects on translation at scale in native coding sequences. Past efforts have focused on quantifying the ribosome occupancy of reporters with randomized 5′ UTRs (Sample et al., 2019) or unnatural sequence elements such as poly-charged peptides (Burke et al., 2022; Tesina et al.,

2020). In this study, we developed a high-throughput assay for translation based on polysome profiling, which measures the number of ribosomes on mRNAs via sucrose gradient fractionation and RNA-sequencing. Utilizing our established software for MAVE analysis (Hoskins et al., 2023), we profiled variant effects on translation at single-nucleotide resolution, and integrated these translation measurements with MAVEs measuring mRNA and protein abundance for catechol-*O*-methyltransferase (*COMT*), an enzyme of major biomedical significance.

COMT catabolizes dopamine, epinephrine, norepinephrine, and catechol estrogens. Several common population SNPs of *COMT* were significant in genome-wide association studies of addictive behavior, cognition, and mood and neurological disorders (Bray et al., 2003; Egan et al., 2001; Guillot et al., 2015; Hendershot et al., 2012; Lin et al., 2017; Meyer-Lindenberg et al., 2006; Shifman et al., 2002). Drugs targeting COMT are used in the treatment of neurological disorders such as Parkinson's disease (Bialecka et al., 2008; Ruottinen and Rinne, 1998). Further, COMT breaks down catechol estrogens which can be genotoxic and may contribute to cancer (Dawling et al., 2001). In prostate and breast cancer cell lines, overexpression of *COMT* inhibited cancer progression by limiting cell invasion and enhancing apoptosis (Hashimoto et al., 2021; Janacova et al., 2023). In addition, 2-methoxyestradiol, a product of COMT catabolism of 2-hydroxyestradiol, inhibits cancer progression (Fotsis et al., 1994; Klauber et al., 1997; Mukhopadhyay and Roth, 1998), suggesting *COMT* plays a suppressive role in cancer.

While variant effects on COMT protein function have been described for the common population variants rs4680, rs4818, and rs4633 (Chen et al., 2004; Nackley et al., 2006), and a noncoding variant in exon 1 (NM_001135162.2) (Lim et al., 2021), there is little annotation of rare variant effects on *COMT* gene expression. For example, expression phenotypes are not mapped for most *COMT* variants annotated with "drug response" in the ClinVar database (Landrum et al., 2016). COMT is produced as both a membrane (MB-) and soluble (S-COMT) protein isoform, with alternative translation initiation producing the S-COMT isoform. It was previously suggested that common coding variants near the translation initiation site (TIS) of S-COMT influence protein abundance and enzyme activity, possibly through alteration of mRNA secondary structure (Nackley et al., 2006; Tsao et al., 2011). Thus, we sought to annotate rare *COMT* variants in the same coding region and test the possibility that coding variants influence expression through alteration of RNA secondary structure.

Results

Evidence of RNA structure in the *COMT* coding sequence

Prior studies used computational predictions to nominate a regulatory role of mRNA secondary structure in the early coding region of *COMT* (Nackley et al., 2006; Tsao et al., 2011). In line with previous studies, we found that the 5′ UTR of COMT was folded with a modest probability, and the first coding exon, which coincides with clinically-relevant variants (Landrum et al., 2016), was predicted to be highly structured (exons 2 and 3 of NM_000754.4, **Figure 1A, Methods**).

Since an increasing number of experimental methods have been deployed to assess RNA secondary structure, we next analyzed RNA probing data from the RASP database (Li et al., 2021) to assess experimental evidence for structure. Dimethyl sulfate (DMS) profiling methods yielded signal in *COMT* coding and untranslated regions. DMS-seq (Rouskin et al., 2014), which is based on RT termination and mapping of truncated 3′ ends, indicated most regions were relatively unstructured in K562 cells; however, DMS-MaPseq (Zubradt et al., 2017), a mutational profiling method that relies on DMS scar readthrough and mismatch enumeration, indicated strong secondary structure across most regions of exons 2 and 3 in HEK293T cells (**Figure 1B**). Both methods indicated higher structure just downstream of the TIS for the MB-COMT isoform, which overlaps the predicted stem structures from mFold. In addition, DIM-2P-seq (Wu and Bartel, 2017), which enriches structures near the 3’ end, indicated long-range structure in the 3′ UTR (**Supplementary Figure 1A**).

Our analysis of predicted and experimental secondary structure data indicates structure in the 5′ UTR (exon 2), coding exons 3 and 4, and the 3′ UTR (exon 6). Based on these results, we hypothesized disruption of mRNA structure and possibly alternative translation initiation in exon 3 (S-COMT) may cause expression changes identifiable at multiple levels of gene expression (mRNA abundance, ribosome load, and protein abundance) (Lim et al., 2021; Nackley et al., 2006; Tsao et al., 2011).

Characterization of *COMT* uORFs and common population SNPs

To dissect regulatory elements impacting *COMT* expression, we generated a HEK293T cell line that stably expressed the conserved 5′ UTR exon and MB-COMT coding sequence with a N-terminal Flag tag and a C-terminal fluorescent tag from a landing pad locus (Matreyek et al., 2020) (**Methods**). To better characterize the translation pattern of our transgene, we conducted ribosome profiling experiments and found evidence for translation initiation, not only at the MB-COMT start codon, but also at

Figure 1

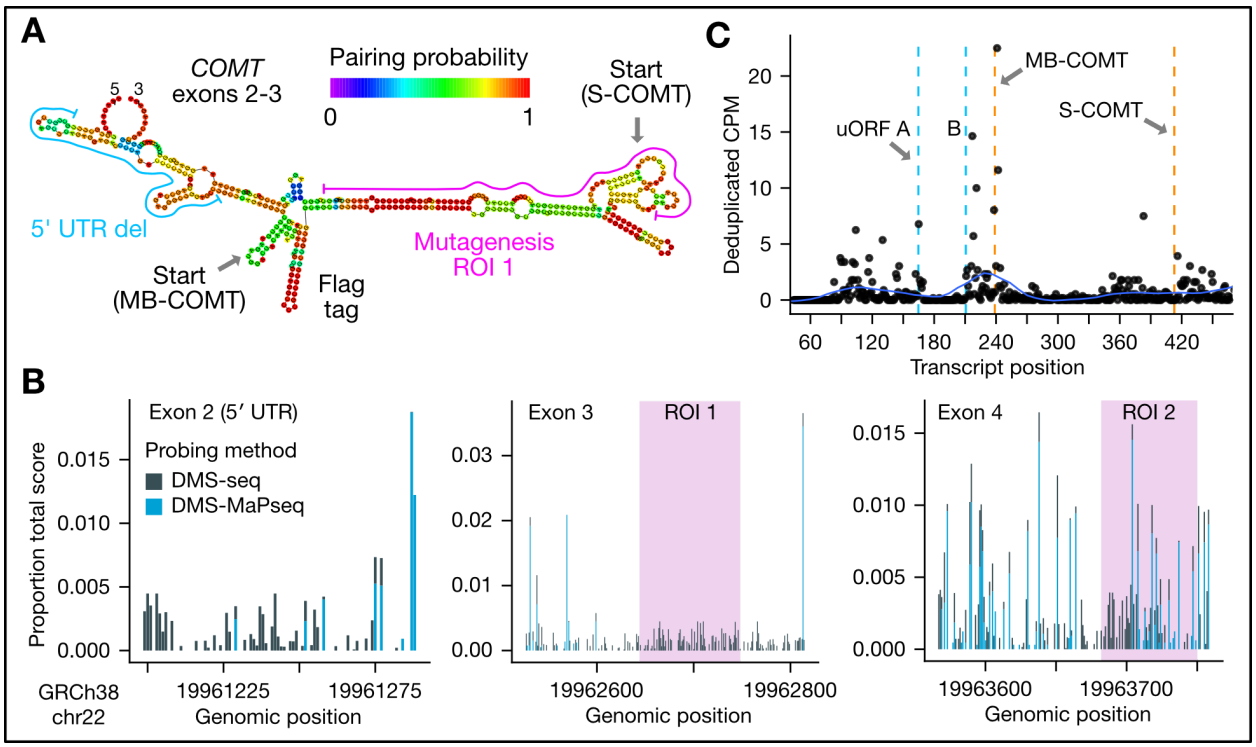


Figure 1: *COMT* RNA secondary structure and uORFs

A) Predicted *COMT* mRNA structure. RNAfold was used to predict RNA secondary structure in *COMT* exons 2 and 3 (NM_000754.4, including a Flag tag). Highlighted are the locations of the membrane and soluble isoform TIS, an isolated deletion (see Supplementary Figure 1B), and the first mutagenesis region. B) *COMT* RNA probing data. Dimethylsulfate (DMS) probing data available from the RASP database (Li et al., 2021) is shown. DMS-seq was carried out in H562 cells (Rouskin et al., 2014), and DMS-MaP-seq data in HEK293T cells (Zubradt et al., 2017). DMS-MaP-seq produced showed little to no signal in exons 2 and 3 relative to DMS-seq. C) Translation initiation in the 5′ UTR of COMT. Ribosome protected fragment (RPF) 5′ ends are plotted as counts per million reads (CPM) following unique molecular identifier based PCR deduplication. Vertical dotted lines mark the expected 5′ end of RPFs when the P-site is positioned at the start codons of the candidate uORFs. Orange lines mark canonical TIS 5′ ends, and blue lines the uORF 5′ ends. ROI=region of interest; UTR: untranslated region; uORF=upstream open reading frame; TIS=translation initiation site.

two upstream open reading frames (uORFs) in the 5′ UTR (**Figure 1C**). Both uORFs (hereafter uORF A and B), showed high ribosome footprint signal at CUG start codons in previous translation initiation site (TIS)-sequencing data in multiple cell types (Michel et al., 2014) **(Supplementary Figure 1B**).

We characterized the potential effect of uORF translation on gene expression using several single-nucleotide variants and deletions (**Supplementary Figure 1C, Supplementary Figure 2, Supplementary Note**). Our analysis revealed potential regulation of MB-COMT expression by uORF B. We also characterized the effect of common population variants in the coding region on *COMT* expression (**Supplementary Figure 3, Supplementary Note**). Altogether, we noted possible alteration of *COMT* expression by noncoding variants but not common coding variants, although these effects may be due to our transgene design (see **Limitations**). To rigorously annotate the large space of possible variant effects in the coding region of *COMT*, we proceeded by characterizing variant effects on RNA abundance, ribosome load, and protein abundance after mutagenizing two *COMT* regions spanning the region of the common population SNPs.

Design of multiplexed assays for *COMT* variant effect analysis

To scale our measurements of variant effect, we had mutagenesis libraries synthesized to generate all alternate silent and missense codons, and one nonsense codon, at each of 58 amino acid positions in *COMT* (**Figure 2A**, **Methods**). We focused our mapping efforts on two regions spanning *COMT* variants of clinical significance (Landrum et al., 2016) (**Supplementary Figure 2A**). The first mutagenesis region overlaps the predicted mRNA stem structure in exons 2-3, and the second region covers the SNPs in exon 4 that were reported to modulate protein abundance and enzyme activity (rs4680, rs4818) (Nackley et al., 2009, 2006). We opted to target these regions instead of catalytic regions of *COMT* to interrogate variant effects tied to RNA folding and codon usage in the early coding sequence of *COMT*. Each library exhibited high uniformity of representation of intended alterations (**Supplementary Figure 4A**), and multi-nucleotide polymorphisms (MNPs) were at globally lower frequencies compared to SNPs (**Supplementary Figure 4B**).

After stable expression of each *COMT* library in four replicate cell lines, we sequenced the mutagenized regions using genomic DNA, total RNA, and polysomal RNA from four pooled fractions after sucrose gradient fractionation (**Methods**). To measure mRNA and protein abundance by an orthogonal method, we used flow cytometry to bin cells into

Figure 2

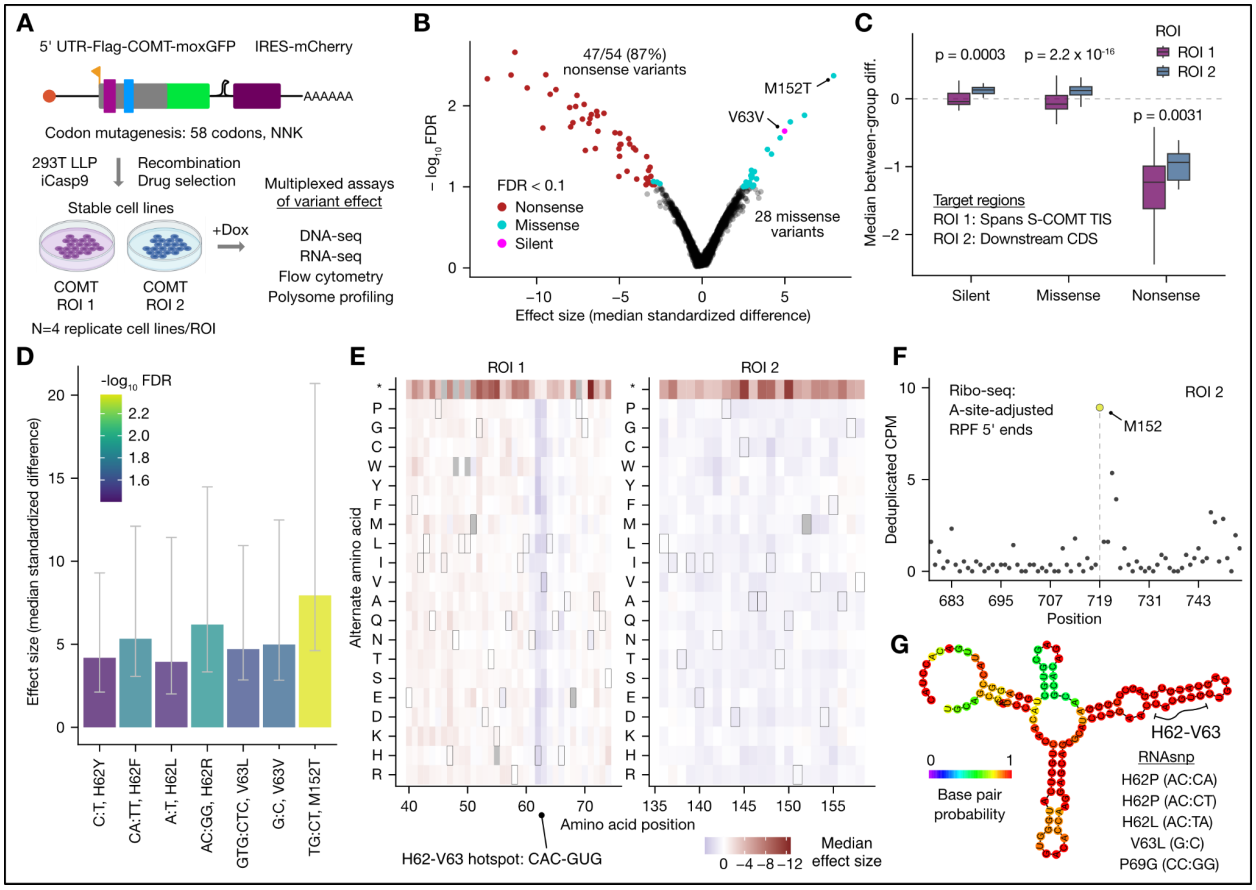


Figure 2: Design of integrated multiplexed assays and *COMT* variant effects on RNA abundance

In panels B-E, the effect size is determined by ALDEx2 (Fernandes et al., 2014, 2013; Gloor et al., 2016) (Methods). A) *COMT* mutagenesis of two regions of interest (ROIs). Two regions spanning common population and clinically relevant SNPs were mutagenized to generate 62/64 of alternate codons at each of 58 codon positions. B) *COMT* variant effects on RNA abundance. C) RNA abundance effects differ by *COMT* region. Outliers > 1.5 * interquartile range are not shown. Significance between regions was determined by unpaired two-sided Welch t-tests. D) Effects by variant position. The reference and alternate nucleotides and amino acid position are denoted for variants at a FDR < 0.1. Error bars indicate 75% confidence intervals. E) Position-specific effects on RNA abundance. Gray tiles indicate incomplete observations. Tiles with a border are the wild-type amino acid. F) Potential ribosome pause with M152 in the A-site. Plotted are counts per million deduplicated reads (CPM) of ribosome-protected fragment 5′ ends within adjusted by codon position in the A-site (Methods). G) Predicted local secondary structure at H62-V63. Minimum free energy structure (Methods) of a 120 nt sequence spanning the H62-V63 hotspot.

one of four populations based on moxGFP (protein) and mCherry (mRNA proxy) fluorescence. In total, we generated over 120 sequencing libraries for all gene expression readouts.

After excluding the first two polysome fractions and one flow cytometry population with low library yields and a high degree of bottlenecking (**Methods**), 96% of variants (3396/3527) were detected in all biological replicates and readouts, enabling analysis of multi-layer effects of gene expression. While replicates of DNA and RNA readouts were highly reproducible (>0.9 Pearson correlation), we observed modest reproducibility for polysome and flow cytometry readouts (~0.5-0.8) (**Supplementary Figure 4C, Limitations**). For example, in region 1 and 2 respectively, the mean Pearson correlation was 0.61 and 0.72 for polysome fraction 3 and 0.51 and 0.80 for fraction 4. Poorer reproducibility for these readouts is possibly due to more limited starting material. After summarizing biological replicates (N=4) for each readout, correlations between gDNA, total RNA, and polysomal RNA readouts were satisfactory (Pearson's $r > 0.7$). Flow cytometry readouts were the least correlated with other readouts as expected due to the different modality of measurement (**Supplementary Figure 4D**).

<u>*COMT* variant effects on RNA abundance are highly position- and domain-specific</u>
We first identified *COMT* coding variants with differential abundance at the RNA level relative to gDNA (**Figure 2B**, **Methods**, FDR < 0.1). Of variants with altered RNA abundance, we detected predominantly nonsense and missense variants, including a single silent variant (V63V). Specifically, 87% of nonsense variants (47/54) were significantly down in total RNA relative to gDNA, and our point estimate for nonsense variant effects on *COMT* RNA abundance was approximately a two-fold decrease relative to the gDNA frequency (fold change of 0.43 +/- 0.13; mean +/- standard deviation; **Methods**). These observations suggest that our sequencing-based estimate of variant effects on mRNA abundance has both sufficient power to detect changes and low systematic bias.

Interestingly, although in region 1 the majority of missense variants decreased RNA abundance, in region 2, most silent and missense variants increased in RNA abundance (**Figure 2C**). Yet, nonsense variant effects were highly similar between the two regions, indicating a potential biological origin of the observation. Other than M152T, which exhibited the largest increase in RNA abundance, variants with the strong effects on RNA abundance were overwhelmingly located in two adjacent codons: H62 and V63. At these codons, both silent and missense variants uniformly

led to increased RNA abundance (**Figure 2D**). Moreover, when aggregated at the level of amino acid, increased RNA abundance was observed for nearly all alternate amino acids (**Figure 2E**). Notably, the rs4633 silent variant is in the wobble position of H62, and this codon is located in the stable stem structure of *COMT* exon 3 (**Figure 1A**). In HEK293 cells, rs4633 increased protein abundance for the low pain sensitivity haplotype (Tsao et al., 2011). While Tsao et al. suggested that the rs4633 T allele increases protein abundance by relaxing secondary structure near the TIS to facilitate translation initiation, our data supports a mechanism of increased RNA abundance.

We then turned to investigating potential mechanisms causing altered RNA abundance, including differences in codon stability (Narula et al., 2019; Wu et al., 2019), ribosome dwell time (Narula et al., 2019), dicodons (Gamble et al., 2016), dipeptides (Burke et al., 2022), tRNA abundance (Behrens et al., 2021; Gogakos et al., 2017), RNA binding protein motifs (Benoit Bouvrette et al., 2020), RNA editing sites (Picardi et al., 2017), and predicted variant effects on folding (Sabarinathan et al., 2013). Our analysis indicated no significant global differences between variants that increased and decreased in RNA with respect to most of these potential mechanisms (codon stability, dwell time, dicodons, dipeptides, tRNA abundance, RBP motifs, RNA editing; **Methods**).

While there were no global differences, specific variants highlight potential mechanisms altering mRNA abundance. For example, we examined our ribosome profiling data and noted a potential ribosome stall when M152 is situated in the A-site (**Figure 2F**). We hypothesize the M152T variant (ATG:ACT) might perturb this putative stall and stabilize the RNA. This effect was specific to the ACT codon, as all other variants to threonine at this position showed no effect. Unlike M152, we did not find any strong footprint signals at the H62 or V63 codons. While the H62 codon (CAC) has a low codon stability coefficient, V63 (GUG) has an optimal coefficient (Narula et al., 2019; Wu et al., 2019), suggesting the effect at this region is not solely due to codon stability. We noted the H62-V63 hotspot is positioned in a high-probability stem structure (**Figure 2G, Figure 1A**). Using RNAsnp (Sabarinathan et al., 2013), we found three variants at H62 that were predicted to alter RNA structure (alternate codons CCA, CCT, CTA; RNAsnp $p < 0.1$, **Methods**). Taken together with prior computational evidence for a role of RNA structure (Nackley et al., 2006; Tsao et al., 2011), we hypothesize variants at H62 and V63 alter RNA folding, which may have a secondary impact on RNA abundance.

<u>Flow cytometry nominates *COMT* variants with specific effects on protein abundance</u>

Next, we determined which *COMT* variants alter RNA and protein abundance by flow cytometry, and attempted to separate these effects by identifying variants that alter protein fluorescence conditioned on RNA fluorescence (uncorrelated fluorescence readouts). Representative data for one biological replicate of the wild-type control and the *COMT* variant libraries is shown in **Figure 3A**. We observed high correlation between moxGFP (protein) and mCherry (RNA proxy) fluorescence (Pearson's r 0.97, 0.91, and 0.85 for the wild-type template, region 1, and region 2, respectively).

To test the reliability of our flow cytometry data, we first determined the enrichment or depletion of silent and nonsense variants in specific populations of cells (**Figure 3A**). We found nonsense variants were enriched in the uncorrelated population (P3; uncorrelated RNA, low protein) but depleted in the correlated population (P1; correlated RNA, high protein) (**Figure 3B**). Conversely, silent variants were depleted from P3 but were well-represented in P1. We defined deleterious variants as those enriched in P3 (**Figure 3C-E**). Aside from variants at H62 and H66, wild-type reference positions with deleterious variants overwhelmingly encoded a nonpolar, uncharged amino acid (G, A, V, L, I), suggesting hydrophobicity in the early coding region (S-COMT 1st and 5th helix, 2nd beta-strand) may be important for COMT folding (**Figure 3E**). We examined variants at these positions to determine if specific chemical properties of the alternate amino acid may correlate with the direction and magnitude of effect. While changes from nonpolar amino acids to charged or polar amino acids were the predominant deleterious variants, some deleterious variants were changes from a smaller nonpolar amino acid to a bulkier nonpolar amino acid (A136I, G157I) (**Figure 3F**).

We then investigated the relationship of RNA and protein abundance. We found that variants down at the RNA level were depleted from P1 and enriched in P3, suggesting that some variant effects on protein abundance stem from underlying changes to RNA abundance (**Figure 3G**). We finally asked whether variants with deleterious effects on protein abundance overlapped evolutionarily constrained residues. We used DeMask (Munro and Singh, 2020), a linear model for variant effect prediction, to test if there was a difference between variants enriched and depleted in P3. We found deleterious variants enriched in P3 had lower DeMask scores compared to variants depleted in P3, reinforcing the importance of conserved residues in COMT (**Figure 3F**).

Altogether, we mapped the effect of 3070 variants on COMT protein abundance and identified twelve constrained amino acids. Notably, our flow cytometry strategy may be more sensitive to variants that specifically alter protein abundance independent of

Figure 3

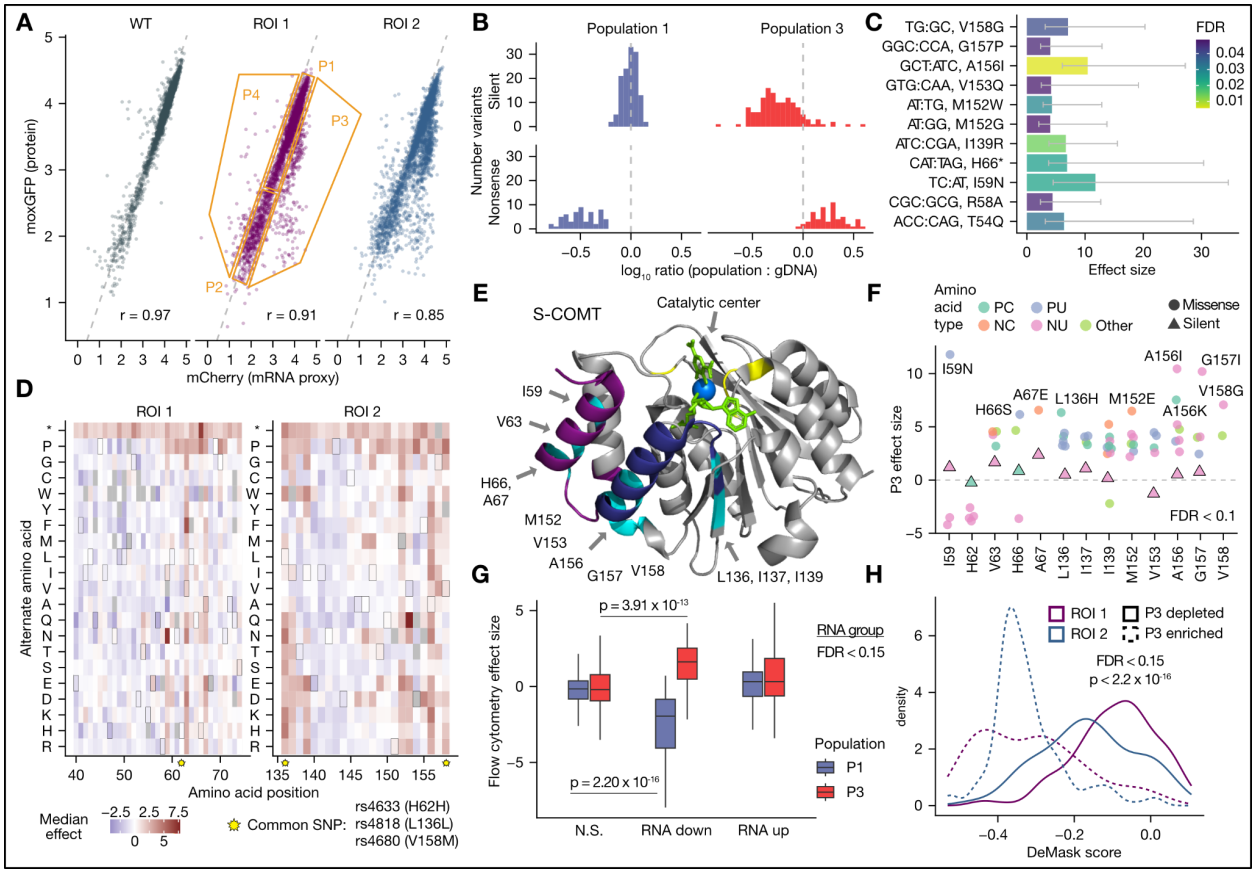


Figure 3: *COMT* variant effects on protein abundance

A) Flow cytometry gating method. A representative biological replicate is shown. Orange lines represent the gating strategy. B) Silent and nonsense variants are enriched and depleted in expected populations. C) Top deleterious variants on protein abundance. Variants enriched in P3 compared to gDNA are shown. Error bars indicate 75% confidence intervals. D) COMT protein variant effect map. Gray tiles indicate incomplete observations. Tiles with a border are the wild-type amino acid. Yellow stars indicate locations of common population SNPs. E) Positions with deleterious variants in the S-COMT structure, with annotations relative to MB-COMT. Regions 1 and 2 are indicated by dark purple and dark blue, respectively. Deleterious positions are highlighted in cyan. Substrates (3-,5-dinitrocatechol, S-adenosylmethionine) are shown in lime green, the catalytic Mg^{2+} ion in blue, and catalytic residues in yellow. Note mutational tolerance in the loop near the catalytic center. F) Missense variant effects at constrained positions. Variants at positions in panel E are colored according to the property of the alternate amino acid. PC: positive charged (R, H, K); NC: negative charged (D, E); PU: polar uncharged (S, T, N, Q); Other (C, P); NU: nonpolar uncharged (G, A, V, L, I, M, F, Y, W). G) Comparison with RNA abundance. Variants in P1 and P3 at

are stratified by their effect on RNA abundance at FDR < 0.15. Outliers > 1.5 * interquartile range are omitted for clarity. H) Comparison with linear model of protein abundance effect. Kernel density estimate of the DeMask model score is shown for variants in P3 at a FDR < 0.15. In panels G and H, significance was determined with two-sided Wilcoxon tests.

effects on RNA abundance. The positions with deleterious effects largely encoded nonpolar amino acids and made contacts with adjacent secondary motifs, whereas the loop nearest the SAM binding site was highly tolerant to mutation except for a single amino acid substitution (N142P).

A MAVE using polysome profiling identifies *COMT* variant effects on ribosome load

We reasoned that some effects on RNA or protein abundance could be attributed to impacts on translation, via alteration of elongation or co-translational protein folding (Kimchi-Sarfaty et al., 2007; Simhadri et al., 2017). To our knowledge, no prior study measured ribosome load for codon mutagenesis libraries in native coding sequences, although previous studies have profiled ribosome load on transcript isoforms (Floor and Doudna, 2016) and randomized 5′ UTR reporters (Sample et al., 2019).

We sequenced RNA from several polysome *metafractions* (pooled raw fractions) for the *COMT* variant libraries and compared abundance in each metafraction to the total RNA abundance (**Supplementary Figure 5A**, **Methods**). Biological replicates were modestly correlated for heavy-polysomal metafractions (Pearson correlation mean +/- S.D. for metafraction 3 was 0.61 +/- 0.09 and 0.72 +/- 0.23 for region 1 and region 2, respectively; for metafraction 4 correlation was 0.51 +/- 0.11 and 0.80 +/- 0.16). Poor correlation was observed for the first and second metafractions, as these libraries were very low yield (**Supplementary Figure 5B**). Thus, we dropped these metafractions from subsequent analyses. Consistent with previous measurements in HEK293T cells (Floor and Doudna, 2016), *COMT* associates predominantly with heavy polysomes, leading to low abundance of messages with 1-3 loaded ribosomes (**Supplementary Figure 5C**). High polysome load likely explains low yield for the first two metafractions.

We hypothesized nonsense variants should be depleted in mid (F3) to heavy (F4) polysome metafractions relative to silent variants, as premature termination should trigger nonsense-mediated decay (NMD) and removal of these mRNAs from the translating pool. Indeed, we found nonsense variants showed lower enrichment in F3 and F4 compared to silent variants (**Figure 4A**, two-sided Wilcoxon rank sum test $p=0.07$ and $p=5.26 \times 10^{-9}$). We propose a modest effect for nonsense variants in the mid-polysome fraction (F3) could be due to lack of exon-junction complex deposition on mRNAs from a lack of introns in our transgene, limiting the efficiency of NMD (Lloyd et al., 2020).

Figure 4

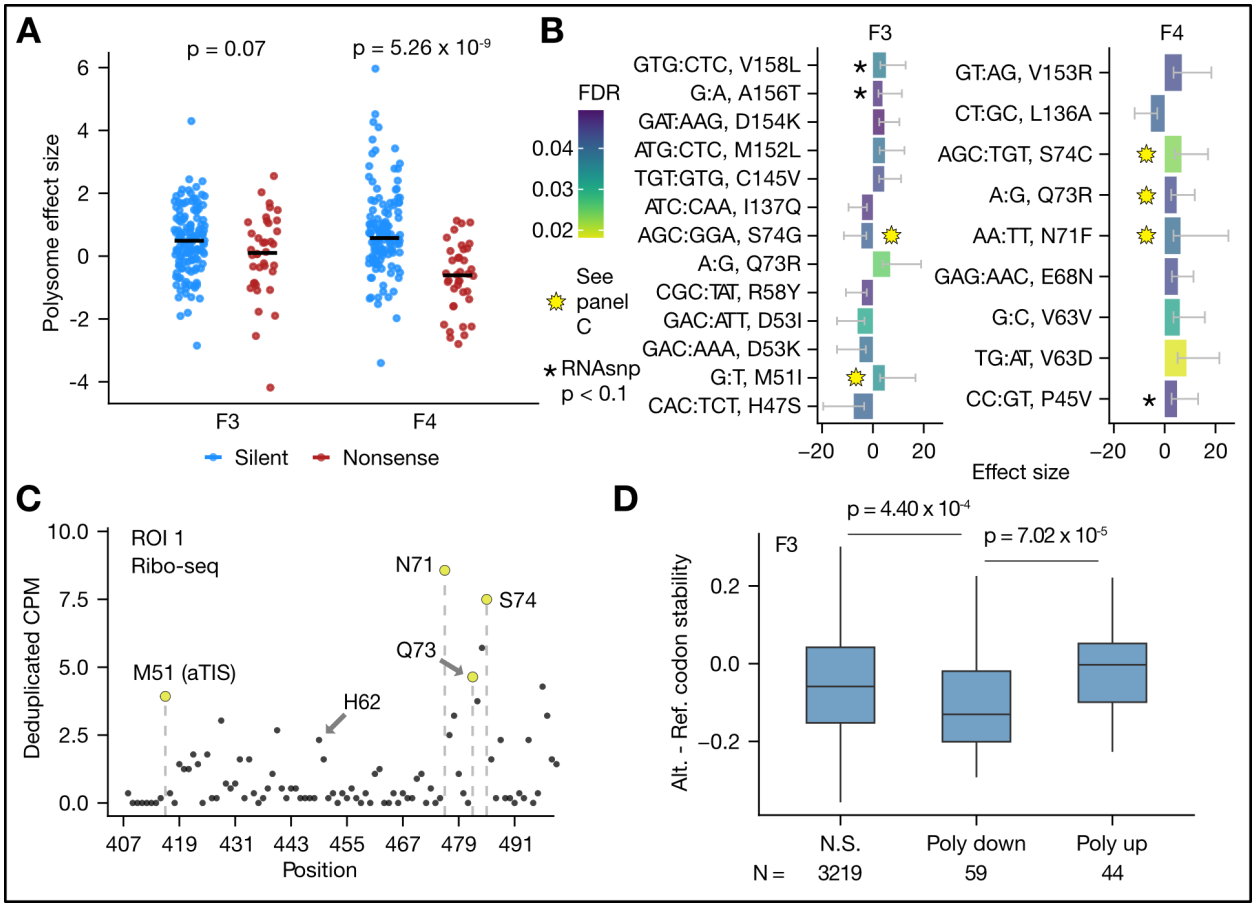


Figure 4: *COMT* variant effects on ribosome load

A) Nonsense variants have decreased ribosome load relative to silent variants. F3 and F4 refer to mid to heavy polysome fractions, respectively. B) Variants with differential ribosome load. Variants at putative pause sites are demarcated with a yellow star. Error bars indicate 75% confidence intervals. C) Putative pause sites in *COMT*. The y-axis reports PCR-deduplicated counts per million (CPM) of ribosome-protected fragments (RPFs) in the *COMT* template. RPF 5′ ends were A-site adjusted (Methods), and vertical dashed lines highlight positions of differential variants. M51 is the alternative TIS (aTIS) for S-COMT. D) Variant effects on ribosome load are consistent with codon stability. The difference between alternate and reference codon stability coefficients (Methods) (Narula et al., 2019) was computed for F3 variants at a FDR < 0.1. Significance was determined by a one-sided Wilcoxon test.

We proceeded by identifying variants with altered mid- (F3) and heavy- (F4) polysome loading. We identified 22 variants with differential abundance in F3 and F4 (**Figure 4B**, FDR < 0.05). We uncovered variants at positions previously identified as significant in our RNA and protein abundance assays, including V63V (increased RNA, increased ribosome load). We cross-referenced our ribosome profiling data and noted potential ribosome A-site pauses at M51, N71, Q73, and S74 of MB-COMT, the location of 23% (5/22) of variants with altered ribosome load (**Figure 4C**). The majority of these variants (4 of 5) showed increased ribosome load, but our observational result is not sufficiently powered to establish a statistical relationship between ribosome pausing and load.

We finally asked whether codon stability or prior ribosome pausing data (Narula et al., 2019; Wu et al., 2019) provide a mechanistic interpretation for variants with altered ribosome load. In line with codon stability as a predictor of translational efficiency (Presnyak et al., 2015), variants with low codon optimality were depleted from polysomes compared to variants with optimal codons (**Figure 4D**, F3, one-tailed Wilcoxon rank sum test $p = 7.02 \times 10^{-5}$ and 8.73×10^{-5} for independent datasets reporting codon stability scores (Narula et al., 2019; Wu et al., 2019)). We did not find an association with polysome effect size and tRNA abundance, or scores for ribosome dwelling at the A-site or P-site (data not shown).

In summary, we tested the application of polysome profiling for multiplexed assay readout for thousands of *COMT* variants. We find less than 1% of queried variants (22/3341) dysregulate mid- to heavy-polysome load (FDR < 0.05), and few of these variants were detected at the RNA- and protein-level. Our results highlight the utility of polysome profiling readouts to complement RNA and protein abundance measurements for multiplexed assays of variant effect.

<u>Unsupervised machine learning of *COMT* post-transcriptional gene expression</u>
Our analyses identified variants with effects on specific molecular phenotypes, generating a rich dataset enabling us to investigate feedforward relationships between RNA abundance, ribosome load, and protein abundance. We next proceeded by utilizing machine learning methods to map the relationship between gene expression signatures for thousands of variants across layers.

We first conducted an analysis using our single-layer effect sizes (ALDEx2, (Fernandes et al., 2014, 2013; Gloor et al., 2016), **Methods**) to map sequential expression conditioned on the prior layer (RNA abundance given gDNA abundance, ribosome load

given RNA, and protein given RNA). Nonsense variants primarily had an effect on RNA abundance, with modest additional effects on protein abundance and polysome load (**Figure 5A**, **Methods**). We then investigated variants that were differential in any gene expression layer and hierarchically clustered these variants (**Figure 5B, Methods)**. Notably, several variants led to divergent expression signatures (effect in different directions between layers) including at H62, V63, M152, L136, and A156.

Then, we used compositional data analysis (interquartile-range normalized log ratios) (Greenacre, 2021) and an unsupervised machine learning method (self organizing map) (Kohonen, 1982; Wehrens and Buydens, 2007) to group *COMT* variants by their expression signature (**Methods; Figure 5C-D**). Of note, three clusters (1, 2, and 9) contained variants deleterious on protein abundance; cluster 9 contained variants deleterious on both RNA and protein abundance; and variants in cluster 8 exhibited decreased ribosome load with little to no changes on RNA or protein abundance.

We characterized clusters in terms of their composition of silent, missense, and nonsense variants. Strikingly, cluster 9 was highly enriched for nonsense variants and contained 19 missense variants and no silent variants (**Figure 5E**). Clusters further away from this cluster showed few nonsense variants but an increase in the proportion of silent variants. A large proportion (14/19) of the missense variants in cluster 9 were at position I59 and A67. Although I59 is thought to be relatively tolerant to mutation, A67 is constrained (DeMask (Munro and Singh, 2020) entropy < 20% percentile). These findings implicate that these specific missense variants lead to gene expression profiles highly similar to nonsense changes.

We finally tested whether the variants which show signatures of deleterious protein abundance might alter RNA folding, as previously suggested (Nackley et al., 2006). A predicted alteration of RNA folding was tested for 19 missense variants from cluster 9 and 11 silent variants from adjacent clusters 7 and 8 (**Figure 5F**). Of these 30 variants, four (13%) were predicted to alter RNA folding (RNAsnp $p < 0.1$). For example, the silent variant E56E (G>A) was among this set, but V63V (G>C) was not predicted to significantly alter folding. We also found that variants with deleterious effects on protein abundance were indeed associated with greater predicted RNA folding effects (**Figure 5G**, two-sided Wilcoxon rank sum test $p = 5.61 \times 10^{-10}$). Thus, perturbed RNA folding may alter translation initiation or elongation rates with consequent impacts on protein abundance.

Figure 5

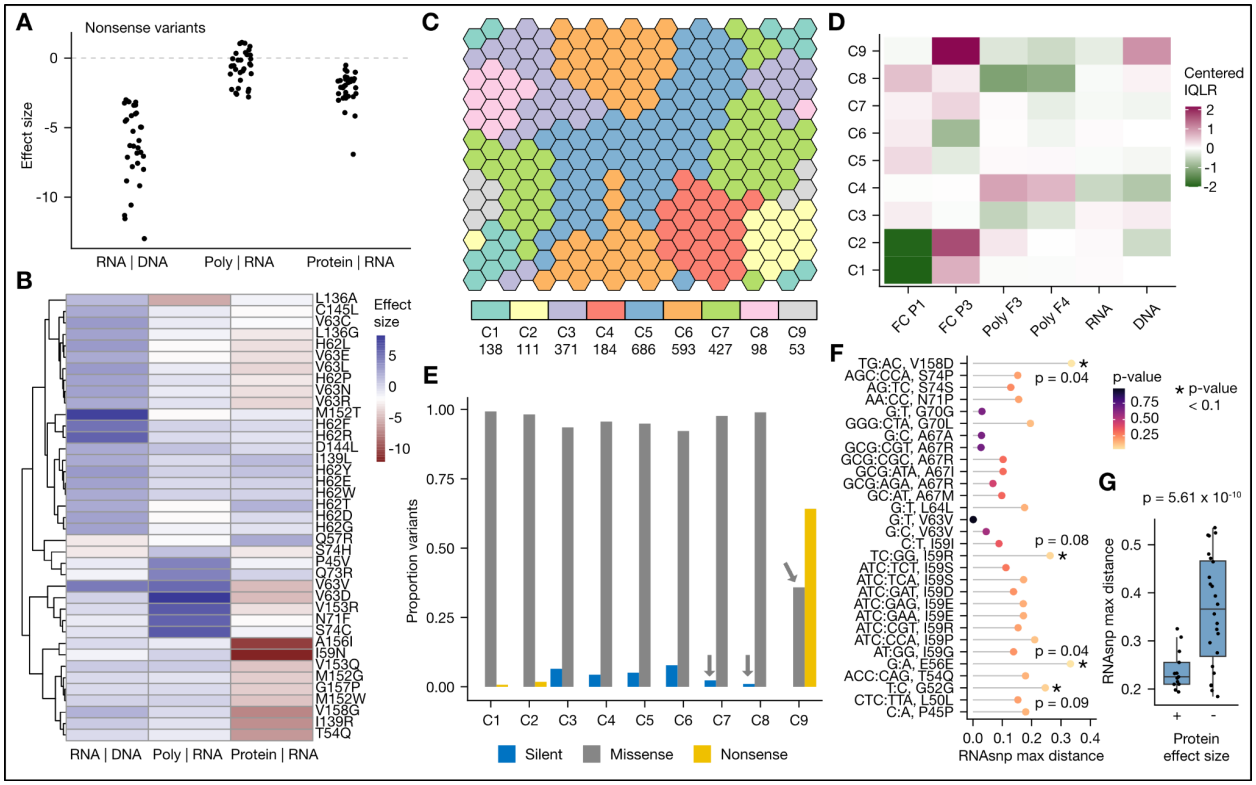


Figure 5: Integrated analysis of *COMT* multi-layer gene expression

Panels A, B, and G show ALDEx2 effect sizes, i.e. effect sizes conditioned on the former expression layer. The 'protein | RNA' effect (enrichment in flow cytometry P3) was multiplied by -1 to signify a deleterious impact. For 'Poly | RNA', metafraction 4 was used. A) Nonsense variants predominantly impact RNA abundance. Effect sizes are shown for 32 nonsense variants tested in all expression layers. B) Hierarchical clustering of variants. Nucleotide changes were omitted for clarity. C) Clustering of variant expression topology. A self-organizing map (SOM) was trained and each of 260 neurons (hexagons) were subsequently grouped by affinity propagation clustering (Methods). Nearby SOM neurons share a similar expression topology. D) Exemplars for each affinity propagation cluster. The interquartile log ratios (IQLR) are shown, indicating the representative expression. In contrast to panels A and B, a positive value for flow cytometry (FC) P3 indicates enrichment in P3 (deleterious effect on protein abundance), while for all other comparators, a negative value indicates a deleterious effect. E) Variant composition of affinity propagation clusters. The proportion of variants that are silent, missense, and nonsense in each cluster are plotted. Arrows indicate variants tested in panel F. F) Predicted effects on RNA structure for missense variants in cluster 9 and silent variants in clusters 7 and 8. Variants with a potential impact on RNA structure were determined with RNAsnp (Methods). Variants impacting structure

(measured by max Euclidean distance from the wild-type base pairing probabilities) have greater distances. G) Alteration of RNA structure may underlie protein abundance effects. Variants affecting protein abundance at a FDR < 0.1 and RNAsnp p-value < 0.15 are plotted. Significance was determined with a two-sided Wilcoxon rank sum test.

Discussion

Our results for variant effects on multiple gene expression phenotypes are collated and provided for each multiplexed assay, along with annotations of codon stability, tRNA abundance, ribosome dwell time, and predicted effects on RNA structure (**Supplementary Table 1, 2**). To our knowledge, we are the first to integrate three multiplexed assays of variant effect to measure RNA abundance, ribosome load, and protein abundance at single-nucleotide resolution. We coalesced multiplexed data from several gene expression layers and identified diverse *cis*-regulatory elements controlling expression in both a directional and divergent manner. Variants impacting multiple expression layers were rare, underscoring the benefit of harnessing several multiplexed assays to annotate variant phenotypes. Our data suggest that variant effects may be explained by diverse mechanisms including changes in codon stability and perturbation of RNA structure or ribosome pausing.

Prior studies reported a strong correlation between ribosome load and protein abundance (Floor and Doudna, 2016; Sample et al., 2019), which prompted us to utilize polysome profiling as the basis of a novel MAVE to measure ribosome load. However, the reliability of polysome profiling in reporting active translation is debated (Richter and Coller, 2015). Some mRNAs are polysome-associated but translationally repressed (Andrews et al., 2011; Braat et al., 2004; Clark et al., 2000; Richter and Coller, 2015). Although we identified several variants that alter ribosome load, some coinciding with putative ribosome pause sites, we found no strong associations between ribosome load and protein abundance for single-nucleotide variants. This may be due to our analysis strategy, which conditioned effects of ribosome load and protein abundance on underlying RNA abundance.

To fully realize the power of integrative multiplexed assays, readouts of protein function are needed in combination with gene expression measurements. We attempted to establish a high-throughput assay of COMT enzyme activity using a COMT substrate (7,8-dihydroxy-4-methylcoumarin, DHMC), which was suggested to fluoresce upon methyl transfer (Qian et al., 2016). However, we failed to detect activity of endogenous COMT after treating HEK293T cells with DHMC, including after expression of our transgene in multiple experiments (data not shown). One possibility is that because our COMT transgene is predominantly membrane-localized, and since the catalytic face is extracellular (Schott et al., 2010), the catabolic product may be released into the supernatant and therefore preclude sensitive detection *in cellula*. Alternatively, the

reporter may be nonfunctional, highlighting the need to develop more rigorously tested reporters of enzyme activity.

Characterization of *COMT* variants nominated mutations at hydrophobic residues that decrease protein abundance (MB-COMT residues I59, V63, A67, A156, G157, and V158). By pairing our data with predicted effects on RNA folding, we found variants at P45, G52, E56, I59, H62, V63, P69, M152, A156, G157, and V158 that may alter gene expression through perturbation of RNA structure. Our discovery of the H62-V63 RNA abundance hotspot is consistent with prior reports of a compensatory effect of rs4633 (H62 silent mutation) on decreased protein abundance (Tsao et al., 2011). Future work might use RNA probing methods to characterize RNA folding of single mutants to confirm association of variant effects with altered secondary structure. One intriguing possibility is to develop multiplexed assays of variant effect on RNA folding, using mutational profiling RNA probing methods (Weng et al., 2020; Zubradt et al., 2017).

Finally, we cross-referenced the variants targeted in our multiplexed assays with expression quantitative trait loci (eQTL) data (Kwong et al., 2022) and population variant data from gnomAD (Karczewski et al., 2020) to determine if any common or rare variants might impact *COMT* gene expression. While none of the variants we targeted were previously identified as eQTLs, our analyses nominate specific variants (FDR < 0.15) observed in human populations that might impact expression (**Supplementary Table 3**). For example, we identified rs4633 (H62H, C>T); rs201641100 (P143P, C>T); rs8192488 (A146A, C>T); and a variant H62Y (C>T) that increased RNA abundance, as well as L49V (C>G) which decreased RNA abundance.

At the level of ribosome load, we identified rs758271838 (L60L, C>T), rs149909767 (G70R, G>A); and rs759340648 (D144N, G>A) that increased ribosome load, and rs750684179 (G52G, T>C) which decreased ribosome load and was predicted to alter RNA secondary structure. Finally, of variants that may be deleterious on protein abundance, we identified rs1342715506 (I59I, C>T); rs757163626 (A67V, C>T); rs375558228 (A67A, G>A); and rs567862500 (A147T, G>A). Our study underscores the advantages of employing comprehensive gene expression profiling for rare variants, a strategy that revealed several variants in human populations potentially influencing *COMT* gene expression. Further research will be required to explore the functional implications of these variants on expression phenotypes for *COMT* and other clinically relevant genes.

Limitations

One important caveat is that our *COMT* transgene is expressed without introns, with partial vector-derived UTRs, and with N- and C-terminal affinity tags, which could impact its expression relative to the endogenous *COMT* gene (**Supplementary Note**). These and other latent variables might explain our finding of insignificant effects of common population SNPs. This includes the rs4818 variant, which generates the high-pain sensitivity haplotype and was previously reported to lead to low protein abundance (Nackley et al., 2006). Nonetheless, we identified the encompassing residues of rs4818 (I136) and rs4680 (V158) as positions intolerant to mutation by our multiplexed protein abundance assay.

Second, high ribosome load on our transgene, combined with our polysome fraction pooling strategy (in which the first two metafractions did not produce sufficient yield), led to a limited dynamic range. This may lead to lower reproducibility of polysome profiling readouts compared to readouts from gDNA and total RNA, which could be a limitation for scaling multiplexed translation assays to large targets. Potential reasons for modest reproducibility of variant abundance (**Supplementary Figure 4**) include biological variability in translation and a technically challenging experimental workflow (three-step RNA purification with >30 RNA precipitations *per metafraction*, **Methods**). Robotic automation of RNA purification steps and use of scalable library preparation methods (Zheng et al., 2014) might be leveraged for multiplexed polysome profiling of large target genes. For future high-throughput polysome profiling assays, use of stable-expression tools, a carefully chosen polysome fraction pooling design, and a relatively high number of biological replicates (Matreyek et al., 2020, 2017) may be required.

Materials and Methods

Cell culture

HEK293T LLP iCasp9 Blast cells (Matreyek et al., 2020) were periodically tested and confirmed free of Mycoplasma, and were cultured in Dulbecco's Modified Eagle Medium (Thermo Fisher Scientific, 11995065) with 10% fetal bovine serum (Gibco), and 1% penicillin-streptomycin (Gibco, 15140122).

Cloning of dual-tagged *COMT* transgene

C-terminal fluorophore tagging of COMT has been previously reported and is compatible with enzyme activity assays (Schott et al., 2010; Sei et al., 2010). A plasmid containing the *COMT* noncoding exon 2 (NM_000754.4) and coding sequence was

obtained from DNASU (clone HsCD00617865), and gBlocks were ordered for 1) the same *COMT* noncoding exon 2 with a downstream Flag tag; and 2) moxGFP (Costantini et al., 2015), which facilitates membrane localization. The 5′ UTR and coding sequence were amplified in Q5 polymerase reactions (NEB, M0491) with 10 ng HsCD00617865 template and 500 nM forward and reverse primers for 25 cycles using manufacturer's recommendations with 55 ˚C annealing temperature. See Supplementary Table 4 for primer sequences. Products were purified with the Nucleospin PCR and Gel Cleanup Kit (Takara, 740609), and the 5′ UTR, CDS, and moxGFP gBlock were joined using a Gibson assembly reaction (NEB) following manufacturer recommendations.

To generate the 5′ UTR deletion clone, 2 µL of the assembly reaction was used as template for a final Q5 PCR using Gateway-compatible primers with attB1/attB2 tails and the aforementioned PCR conditions. The product was subsequently transferred into a pDONR223 entry clone via a Gateway BP reaction (Thermo Fisher Scientific), and transferred to a destination vector (pDEST_HC_Rec_Bxb_v2) compatible with recombination into the 293T LLP iCasp9 Blast cell line with a 1 h LR reaction (Thermo Fisher Scientific, 11-791-020). To generate a clone containing the full 5′ UTR exon 2, first a CDS-moxGFP product was amplified from the validated 5′ UTR deletion clone in a 50 µL NEB Q5 reaction for 25 cycles with 30 s annealing at 55 ˚C. The product was extracted from a 1% agarose-TBE gel (Zymo Gel Extraction kit, 10 µL elution), then assembled with the 5′ UTR gBlock in a Gibson reaction. A final 50 µL Q5 PCR reaction was performed with 2 µL assembly reaction template for 30 cycles with 30 s annealing at 50 ˚C. The product was extracted from a 1% agarose-TBE gel as previously detailed. The transgene was transferred to the pDONR223 and pDEST_HC_Rec_Bxb_v2 with 1 h Gateway BP and LR reactions, respectively.

The isolated transgene clones contained either the full 91 nt of exon 2 of the MB-COMT 5′ UTR (Endo. 5′ UTR), or a 53 nt exon 2 deletion (5′ UTR del). Both transgenes shared a common Kozak sequence between -21-1, a N-terminal Flag tag, a C-terminal moxGFP tag upstream of a bicistronic IRES-mCherry element, and partial vector-derived 5′ and 3′ UTRs from the landing pad vector pDEST_HC_Rec_Bxb_v2. Constructs were verified using Sanger sequencing.

<u>*COMT* variant library generation</u>

The 5′ UTR-Flag-COMT-moxGFP clone in pDEST_HC_Rec_Bxb_v2 was sent to Twist Biosciences for mutagenesis of membrane-COMT codons 40-74 and 136-158. All

silent and missense variants and the amber nonsense variant were generated at each codon with attB sites. Inserts for each target region (region 1, region 2) were received and pooled separately, with codons in each region pooled at equimolar ratios. A Gateway BP reaction was performed with 70 ng of pooled inserts and 150 ng pDONR223 at 25 ˚C overnight (19 h). 1.5 µL BP reaction was electroporated into 25 µL Endura Electrocompetent cells (Lucigen, 60242) with the following conditions: 2 mm cuvette, 2500 V, 200 Ohms, 25 µF. Cells were outgrown for 1 h at 37 ˚C in 1 mL Lucigen outgrowth media, then 600 µL was plated on Nunc Square Bioassay dishes, scraped in 8 mL LB Miller broth, and plasmid library purified with the ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, D4203) using 200 µL 50 ˚C elution buffer, and including endotoxin removal. Library sizes were estimated at 942,000 and 510,000 clones for region 1 and region 2 libraries, respectively (or 434-fold and 357-fold coverage of each variant).

1 µg of each library in pDONR223 was digested with Blp1 (NEB) for 1 h 15 min at 37 ˚C. Digests were run on a 0.8% agarose-TAE gel and full-length plasmids were extracted with the Nucleospin PCR and Gel Cleanup Kit (Takara, 740609). A LR reaction was performed with 150 ng entry library and 150 ng pDEST_HC_Rec_Bxb_v2, incubated at 25 ˚C for 21 h. 25 µL Endura Electrocompetent cells were transformed with 1.5 µL LR reaction using the same conditions as for the BP reaction. After 14.5 h, colonies were collected with 7 mL LB Miller broth, and 3 mL was pelleted and processed with the ZymoPURE II Plasmid Maxiprep Kit as above.

Generation of single *COMT* mutants

The common population variants (rs4633, rs4818, rs4680, rs6267, rs74745580) were generated in the 5′ UTR-Flag-COMT-moxGFP clone in pDEST_HC_Rec_Bxb_v2 by Genscript. The 5′ UTR mutants were generated using the NEB Q5 Site-Directed Mutagenesis Kit (Fisher Scientific, E0554S) according to manufacturer recommendations but extending the KLD reaction to 10 min. The cycling parameters were: initial denaturation at 98 ˚C for 30 s; 25 cycles of denaturation at 98 ˚C for 10 s, anneal at 62 ˚C for 30 s, extension at 72 ˚C for 3 min 30 s; final extension at 72 ˚C for 2 min. Mutagenesis primers are listed in Supplementary Table 4. All mutants were bidirectionally validated by Sanger sequencing.

Stable expression of *COMT* variant libraries and single mutants

For each region 1 and region 2 biological replicate cell line, 20 µg of the library along with an equal mass of Bxb1 recombinase (pCAG-NLS-HA-Bxb1) was transfected into a

15 cm dish of HEK293T LLP iCasp9 Blast cells using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008), with volumes scaled based on 3.75 µL reagent per 6-well. For single mutants, 1.5 µg was transfected for each of the donor and Bxb1 recombinase plasmids. After 48 h, at near full confluency, 2 µg/mL Doxycycline and 10 nM AP1903 (MedChemExpress, HY-16046), both solubilized in DMSO, were added for negative selection of non-recombined cells. The next day, dead cells were removed and recombined cells were grown out for an additional two days with fresh media containing Doxycycline and AP1903. Cells were recovered for two days by growth in media without Doxycycline and AP1903. Before functional assays, transcription was induced with 2 µg/mL Doxycycline for 24 h (total RNA, polysome RNA readouts) or 21 h (flow cytometry readouts), and cells were stimulated with fresh media for 1 h 15 min prior to harvest. Importantly, because our *COMT* transgene is membrane localized and faces the extracellular space, cells are collected by pipetting in cold PBS, and not by trypsinization.

RT-qPCR

1-2 µg of extracted total RNA and polysomal RNA was treated with DNaseI (NEB, M0303) in a 100 µL reaction at 37 ˚C for 15 min, then re-purified by the RNA Clean and Concentrator Kit (Zymo Research, R1015) and eluted in 10 µL water. 100-200 ng of treated RNA was denatured at 65 ˚C for 5 min followed by RT primer annealing at 4 ˚C for 2 min, using 1 µL of 20 µM oligodT(18)6N and 50 ng random hexamers. Primed total RNA was included in a 20 µL SuperScript IV cDNA synthesis reaction (Thermo Fisher, 18090010) with SUPERase-In RNase inhibitor (Thermo Fisher, AM2696), and first-strand cDNA was synthesized by incubating at 23 ˚C for 10 min, 55 ˚C for 1 h, followed by RT inactivation at 80 ˚C for 10 min. RNA was digested with addition of 5 U RNaseH (NEB, M0297) to the first strand cDNA synthesis reaction and incubation at 37 ˚C for 20 min. 2 µL of a 1:10 dilution of 1st strand cDNA was assayed by qPCR on Applied Biosystems Viia-7 instruments in 10 µL reactions with PowerUp SYBR Green master mix (Applied Biosystems) according to manufacturer recommendations, with 500 nM primers and annealing at 55 ˚C for 15 s. See Supplementary Table 4 for primer sequences to the target (*COMT, mCherry*) and calibrator (*EEF2*).

Immunoblotting

No more than 10 µg total protein, quantified by Pierce BCA Protein Assay Kit, was loaded on Bolt 8% Bis-Tris Gels (Thermo Fisher Scientific) in 1x LDS sample buffer with reducing conditions and addition of Bolt Antioxidant, then resolved at 200 V for 25 min. Proteins were transferred to 0.45 µm PVDF membranes at 20 V for 1 h 15 min at 4

˚C. Membranes were dried for 10 min, then reactivated in methanol, and blocked for 1 h in 1x Pierce Blocking Buffer or 3% non-fat dry milk in TBST (0.1% Tween-20). Membranes were rotated with 1 µg/mL mouse anti-FLAG M2 antibody (Millipore Sigma) blocking buffer overnight at 4 ˚C, then washed four times for 5 min in TBST. Blots were finally rotated with 0.2 µg/mL AF647-conjugated goat anti-mouse IgG secondary antibody (Thermo Fisher Scientific) in TBST for 1 h at RT before washing 4 x 5 min in TBST, and imaging on a Odyssey CLx instrument (700 nm detection).

Polysome fractionation and polysomal RNA purification

For *COMT* library stable cell lines, cells at 80-95% confluency (one 15 cm dish per gradient) were incubated with 100 µg/mL cycloheximide (Sigma Aldrich, C4859) in media for 10 min at 37 ˚C, then resuspended in ice cold PBS with 100 µg/mL cycloheximide (PBS-CHX), collected at 4 ˚C by pipetting, and pelleted at 300 x g for 7 min at 4 ˚C. Pellets were flash frozen in liquid nitrogen, then lysed on ice for 10 min with 300-400 µL lysis buffer (20 mM Tris-HCl pH 7.5; 150 mM NaCl; 5 mM $MgCl_2$, 1 mM fresh DTT, 20 mM ribonucleoside vanadyl complex (Fisher Scientific, 50-812-650), 1X protease inhibitor cocktail EDTA-free (Fisher Scientific, 539196), and 100 µg/ml cycloheximide. Lysates were clarified by centrifugation at 1,300 x g for 10 min at 4 ˚C, and 1/10th of the lysate was saved for total RNA extraction. 20-50% sucrose gradients were made with the BioComp Gradient Maker in 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, and 100 µg/ml cycloheximide using Beckman Coulter UC tubes 9/16 x 3-1/2 (Fisher Scientific, NC9194790), and equilibrating overnight at 4 ˚C. Clarified lysates were loaded onto sucrose gradients, and fractionated by ultracentrifugation at 38,000 rpm for 2.5 h at 4 ˚C in a Beckman Coulter SW-41 Ti rotor.

Approximately 500 µL polysome fractions were collected using the Biocomp Piston Gradient Fractionator (v8.04) with a Triax flow cell (BioComp model FC-1) and Gilson Fraction Collector using a piston speed of 0.2 mm/s. RNA was precipitated from raw polysome fractions by addition of sodium acetate to 300 mM pH 5.2, 2 µL Glycoblue, and 2x volumes cold ethanol, stored overnight at -20 ˚C. RNA was resuspended in 50-80 µL ultrapure water, and these raw fractions were pooled into metafractions and extracted by phenol-chloroform (5:1) and ethanol precipitation overnight at -20 ˚C. After a single 70% ethanol wash, polysomal RNA was resuspended in 30 µL ultrapure water. Lastly, DNaseI treatment and a column purification was performed prior to library preparation (see *RNA extraction and cDNA synthesis*).

Ribosome profiling

A 15 cm dish of cells with stable expression of the 5′ UTR-containing COMT transgene at 95% confluency was washed twice with cold PBS containing 100 µg/mL cycloheximide, then lysed on the plate with 1.5 mL lysis buffer (20 mM Tris HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 100 µg/ml cycloheximide, 100 µg/ml anisomycin, 1% Triton-X, and 1x protease inhibitor cocktail EDTA-free). Lysate was partitioned into two tubes and 7 µL RNaseI (Invitrogen, AM2249) was added and incubated for 1 h at 4 ˚C. Digestion was stopped by addition of ribonucleoside-vanadyl complex to 20 mM final. Monosomes were pelleted through a sucrose cushion (1 M sucrose, 20 mM Tris HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1x protease inhibitor cocktail EDTA-free) by spinning at 38,000 rpm at 4 ˚C in a SW-41 Ti rotor (Beckman Coulter). Monosome pellets were resuspended in 400 µL lysis buffer and rocked overnight. RNA was extracted with 1.2 mL phenol/chloroform/isoamyl alcohol and precipitated with 300 mM sodium acetate pH 5.2, 1.5 µL Glycoblue, and 2.5 volumes of cold absolute ethanol overnight at -20 ˚C. RNA was pelleted at 21,000 x g for 1 h, washed once with cold 70% ethanol, and resuspended with 30 µL water.

15 µL of RNA was run on a 15% TBE-urea polyacrylamide gel at 150 V for 1.5 h. The gel was stained with 1x SYBR Gold and footprints between 21 nt and 34 nt were cut and isolated by crush and soak in RNA extraction buffer (300 mM sodium acetate pH 5.5, 1 mM EDTA, 0.25% v/v SDS) with overnight rotation. The eluate was passed through a Spin X Centrifuge Tube Filter with a cellulose acetate membrane (Thermo Fisher Scientific, 07-200-385), and the RNA was precipitated with 300 mM sodium acetate pH 5.2, 1.5 µL Glycoblue, and 2.5 volumes of cold absolute ethanol overnight at -20 ˚C. RNA was pelleted at 21,000 x g for 1 h, washed once with cold 70% ethanol, and resuspended in 12 µL water. 25 ng of RNA in 8 µL water was input to the D-Plex Small RNA kit (Diagenode, C0503000) and libraries were prepared according to the manufacturer recommendations with the exception that 9 cycles of PCR were carried out. The library was purified with 1.25x AMPure beads (Beckman Coulter, A63880) and quantified by the High Sensitivity DNA Kit (Agilent, 5067-4626).

<u>Flow cytometry analysis and flow cytometry sorting</u>

Following induction with doxycycline and media stimulation (see *Stable expression of COMT variant libraries*), cells were washed once with cold PBS, resuspended with PBS, and assayed in a 5 mL Falcon polystyrene round bottom tube with a cell-strainer cap (Fisher Scientific, 08-771-23). Analysis of point mutants and deletion cell lines was conducted on a LSRFortessa SORP instrument (BD) with FACSDiva v6.1.3, and flow cytometry was conducted on a FACSAria Fusion SORP instrument (BD) at the Center

for Biomedical Research Support Microscopy and Imaging Facility at UT Austin (RRID# SCR_021756). Gating was performed with FlowJo. A gate for cells versus debris was defined based on FSC-H versus SSC-H, and another gate for single cells versus doublets was defined based on FSC-A versus FSC-H. Fluorescence gating was based on autofluorescence of 239T LLP iCasp9 Blast cells without recombination or of a stable cell line expressing the template transgene but without doxycycline induction. After flow cytometry sorting, each population was expanded up to a confluent T-25 flask, washed, trypsinized, pelleted, and flash frozen prior to gDNA extraction.

COMT library preparation with amplicon method

gDNA extraction

gDNA was extracted from approximately 3-4 million cells with the Cell and Tissue DNA Isolation Kit (Norgen Biotek Corp, 24700), including RNaseA treatment at 37 ˚C for 15 min, and eluting in 200 µL warm elution buffer.

RNA extraction and cDNA synthesis

Approximately 3-4 million cells were solubilized with 1 mL QIAzol (Qiagen, 79306) and 0.2 mL chloroform in 5PRIME Phase-Lock Gel heavy tubes (QuantaBio, 2302830), according to the manufacturer's recommendations. RNA was precipitated at -20 ˚C following the addition of 2 µL GlycoBlue (Thermo Fisher, AM9515) and 2.5 volumes of cold absolute ethanol. RNA was washed once with cold 70% ethanol then resuspended in 30 µL water. 10 µg total RNA or 15 µL of polysomal RNA (half of 30 µL eluate, volumetric inputs) was treated with DNaseI (NEB, M0303) in a 100 µL reaction at 37 ˚C for 15 min, then re-purified by the RNA Clean and Concentrator Kit (Zymo Research, R1015) and eluted in 20 µL water.

10 µL DNaseI-treated total RNA (<10 µg) or DNaseI-treated polysomal RNA (200 ng - 5 µg) was denatured at 65 ˚C for 5 min followed by RT primer annealing at 4 ˚C for 2 min, using 2 pmol pDEST_HC_Rec_Bxb_v2_R primer specific for the landing pad (see Supplementary Table 4**)** and 2.5 mM random hexamers. Primed total RNA was included in 20 µL SuperScript IV cDNA synthesis reaction (Thermo Fisher, 18090010) with SUPERase-In RNase inhibitor (Thermo Fisher, AM2696), and first-strand cDNA was synthesized by incubating at 23 ˚C for 10 min, 55 ˚C for 1 h, followed by RT inactivation at 80 ˚C for 10 min. RNA was digested with addition of 5 U RNaseH (NEB, M0297) to the first strand cDNA synthesis reaction and incubation at 37 ˚C for 20 min.

Landing pad amplification (PCR1)

Nucleic acids (gDNA, 1st strand cDNA) were amplified with Q5 polymerase (NEB, M0491) in 50 µL PCR reactions for 16 cycles with 500 nM landing-pad-specific primers (pDEST_HC_Rec_Bxb_v2_F/R) flanking the entire *COMT* insert (~1.7 kb). For negative control (plasmid) samples, 10 ng was used as template. For total RNA and polysome RNA samples, 15 µL of unpurified 1st strand cDNA synthesis reaction was used as template. For total gDNA samples, 4 µg template was used in each of three PCR reactions. For flow-sorted gDNA samples, either 1.75 µg template was used in three PCR reactions or 2.5 µg was input into two reactions, with 14 cycles instead of 16. The cycling parameters were: initial denaturation at 98 ˚C for 30 s; 3-step cycling with denaturation at 98 ˚C for 10 s, anneal at 65 ˚C for 30 s, extension at 72 ˚C for 1 min; final extension at 72 ˚C for 2 min.

Products from total RNA and polysome RNA samples were purified with the QIAquick PCR Purification Kit (Qiagen, 28106) and eluted in 30 µL elution buffer. Products for total gDNA samples were pooled, resolved on a 0.8% agarose/TAE gel, stained with 1x SYBR Gold, and extracted using the Nucleospin Gel and PCR Cleanup Kit (Takara, 740609) or Zymoclean Gel DNA Recovery Kit (Zymo Research, D4007) with 20 µL 70 ˚C elution buffer. At this point, PCR1 products for gDNA samples cannot typically be seen in the gel, and the expected band size is cut using comparison to a high-molecular weight ladder.

Coding sequence amplification (PCR2)

Purified PCR1 products (landing pad insert) were amplified for each of COMT target regions in a 50 µL NEB Q5 reaction (NEB, M0491) for 8 cycles, following the same cycling parameters as for PCR1 and 500 nM tailed primers (see Supplementary Table 4). 15 µL PCR1 product was used as template (50% of eluate for total RNA and polysome RNA samples; 75% of eluate for gel-extracted gDNA samples). Products were purified with the Nucleospin Gel and PCR Cleanup Kit (Takara, 740609) with 1:5 buffer NTI dilution and 25 µL 70 ˚C elution buffer, or with the QIAquick PCR Purification Kit (Qiagen, 28106) with 30 µL elution buffer.

Illumina adapter addition (PCR3)

A final Q5 PCR reaction was carried out for 8 cycles with the same formulation as PCR2 but using NEBNext Multiplex Oligos for Illumina Dual Index Primers Set 1 (NEB, E7600S) according to manufacturer's recommendations (65 ˚C annealing). 25 µL/30 µL eluate was used as template for total gDNA, total RNA, and polysome RNA samples. For flow cytometry gDNA samples, 15 µL/25 µL eluate was used as template. PCR3

products (final library) were purified with the Nucleospin Gel and PCR Cleanup kit (Takara, 740609) and eluted in 25 µL 70 ˚C buffer.

Quantification and size-selection of final libraries

Final libraries were quantified using the KAPA Library Quantification Kit for Illumina (Roche, KK4873) according to the manufacturer's recommendations, using 1:10,000 dilution of libraries and size-correction with an average fragment size of 300 nt (region 1) and 285 nt (region 2). qPCR was performed on the Applied Biosystems ViiA 7 instrument and values were used to pool individual libraries at equimolar ratios to target ~5 M read pairs per library. Final library pools were size-selected use PAGE purification on a 15% TBE gel followed by a crush-and-soak method (Sambrook and Russell, 2006), or with BluePippin 2% gel (Sage Science, BDF2010) to select fragments from 250-340 bp. Following size-selection, final pools were quantified by High Sensitivity DNA Kit (Agilent, 5067-4626) prior to sequencing.

Next-Generation Sequencing

All libraries were sequenced 2 x 150 bp (paired-end) on Illumina platforms with minimum 5% PhiX. Libraries were sequenced at MedGenome, Inc. on the HiSeq X and at Novogene Corporation, Inc. on the NovaSeq 6000. Raw FASTQs were obtained directly from the vendor.

Ribosome profiling analysis

Analysis was conducted with RiboFlow (umi_devel branch) and RiboR (Ozadam et al., 2020). The reference files were modified to include a Flag tag upstream of the *COMT* coding sequence to enable mapping of footprints arising from the stably-expressed *COMT* transgene. Footprints with lengths between 28 and 35 were used for analysis. A- and P-site offsets from the 5′ end of the footprint were determined by examining the metagene plots centered at the stop codon. The A-site offset was determined to be 16 nt. This value was used to nominate positions with ribosomal pausing at the A-site in Figures 2F and 4D.

Variant calling and preprocessing

Variants were called with satmut_utils v1.0.3-dev001 (Hoskins, 2023). Analysis of *COMT* libraries prepared with the amplicon method utilized the satmut_utils *call* parameters `-n 2 -m 1 --r1_fiveprime_adapters TACACGACGCTCTTCCGATCT --r1_threeprime_adapters AGATCGGAAGAGCACACGTCT --r2_fiveprime_adapters

AGACGTGTGCTCTTCCGATCT --r2_threeprime_adapters AGATCGGAAGAGCGTCGTGTA`.

Random forest models were trained on simulated data as previously described for libraries prepared from the amplicon method (Hoskins et al., 2023). Variants were post-processed by the following criteria: variants are within the mutagenesis region; 2) nonsense variants must match the amber codon UAG; 3) $\log_{10}$ variant frequency is > -5.8; 4) SNP variants with false positive random forest predictions ($p > 0.5$) in more than half of the gDNA libraries were filtered out. Amino acid positions were adjusted to account for the Flag tag such that positions align with the endogenous membrane-COMT protein isoform.

Frequencies were normalized to the wild-type count at each position (variant count + 0.5 / wild-type count + 0.5) (Rubin et al., 2017). For each variant, normalized frequencies in the negative control were subtracted from the frequency observed in total gDNA, total RNA, polysomal RNA, and flow cytometry gDNA samples, and variants with resultant negative values were filtered out. Normalized frequencies were batch-corrected to account for the library preparation experiment using ComBat (Leek et al., 2022) with default parameters. Total gDNA and total RNA frequencies were adjusted with a model matrix, whereas no model matrix was used for polysomal RNA and flow cytometry gDNA libraries, as readouts to be compared were in different batches. The median among technical replicates was computed for each biological replicate (independent stable cell line). For multi-layer analysis, the median among biological replicates was first computed before unsupervised learning.

<u>Analysis of *COMT* variant differential abundance between gene expression layers</u>
The wild-type normalized, batch-corrected frequencies were converted to counts per million and input to ALDEx2 (Fernandes et al., 2014, 2013; Gloor et al., 2016) with parameters `paired=FALSE, denom="iqlr", mc.samples=128`. ALDEx2 effect size is an estimate of the median standardized difference between groups. The comparisons made were 1) total RNA to total gDNA; 2) polysome metafraction 3 or 4 to total RNA; 3) flow cytometry population 1 and 3 to total gDNA. Polysome metafraction 1 and 2 were not used for analysis due to low yield and sequencing depth. To estimate the fold change for nonsense variant effects on RNA abundance, limma-trend (Ritchie et al., 2015) was used on variant log frequencies using empirical Bayes moderation with an estimated proportion of differential genes of 0.03.

RNA secondary structure analysis

The RNAfold web server (Gruber et al., 2008) was used to predict RNA structure. In Figure 1A, exons 2-3, including the Flag tag, were folded using default parameters. In Figure 2G, a 120 nt sequence spanning the H62-V63 codons was folded. Minimum free energy structures were plotted. To predict variant impacts on RNA structure, command-line RNAsnp (Sabarinathan et al., 2013) was used with default parameters (global folding mode 1, 200 nt window flanking the variant). RNA probing data from the RASP database (Li et al., 2021) was examined with the Integrative Genome Viewer.

Analysis of *COMT* variant effects on protein abundance

We identified variants enriched in flow cytometry population 3 (P3) compared to total gDNA using ALDEx2. Enrichment in P3 (FDR < 0.05) is interpreted as a deleterious effect. In Figure 5, effect sizes were multiplied by -1 to indicate a deleterious effect. PyMol Open-Source v2.5.0 (DeLano, 2002) was used to visualize the S-COMT structure (PDB 3BWM). To keep variant annotations consistent in the text, we mapped annotations relative to MB-COMT onto the S-COMT crystal structure.

Codon stability and ribosome pausing analysis

Codon stability coefficients and ribosome dwell time scores were obtained (Narula et al., 2019; Wu et al., 2019), and the difference in these scores between the alternate and reference codon was computed and compared to the direction of effect. For codon stability data, one-sided Wilcoxon rank sum tests were used with the expectation that variants that decrease ribosome load have lower codon stability coefficient differences. All other data was tested with two-sided Wilcoxon rank sum tests.

Unsupervised machine learning of multi-layer gene expression

Median frequencies across biological replicates for each gene expression layer were transformed to interquartile log ratios by dividing each variant frequency by the interquartile range. IQLRs were scaled (centered and standardized within a layer), then for each variant, robust centering to the median was performed across layers. This ensures the nominal variant abundance in the library is not learned. Using the Kohonen R package (Wehrens and Buydens, 2007), a self-organizing map (SOM) (Kohonen, 1982) with 260 neurons was trained using a 13 x 20 toroidal grid with hexagonal topology and bubble neighborhood. The optimal number of neurons was determined by training 500 SOMs with different random initialization seeds, and choosing the seed that minimizes the distance between the input data and the matching codebook vectors. The dataset was presented to the network 100 times with a linearly decreasing

learning rate of 0.05. Subsequently, neurons were clustered with affinity propagation clustering (Frey and Dueck, 2007) using the R apcluster package (Bodenhofer et al., 2011). The pairwise similarity matrix used negative squared distances, and exemplar preferences were initialized with a q parameter value of 0.1.

Declarations

Supplementary files

Supplementary Table 1: ALDEx2 effects and functional data

Supplementary Table 2: Interquartile log ratios

Supplementary Table 3: Variants in gnomAD v3.1.2

Supplementary Table 4: Primers used in cloning, mutagenesis, and library preparation

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Supplementary data and code used for analysis is located on GitHub:
https://github.com/ijhoskins/integrated_MAVEs
Raw sequencing data and processed results will be made available on GEO.

Competing interests

We declare no competing interests.

Funding

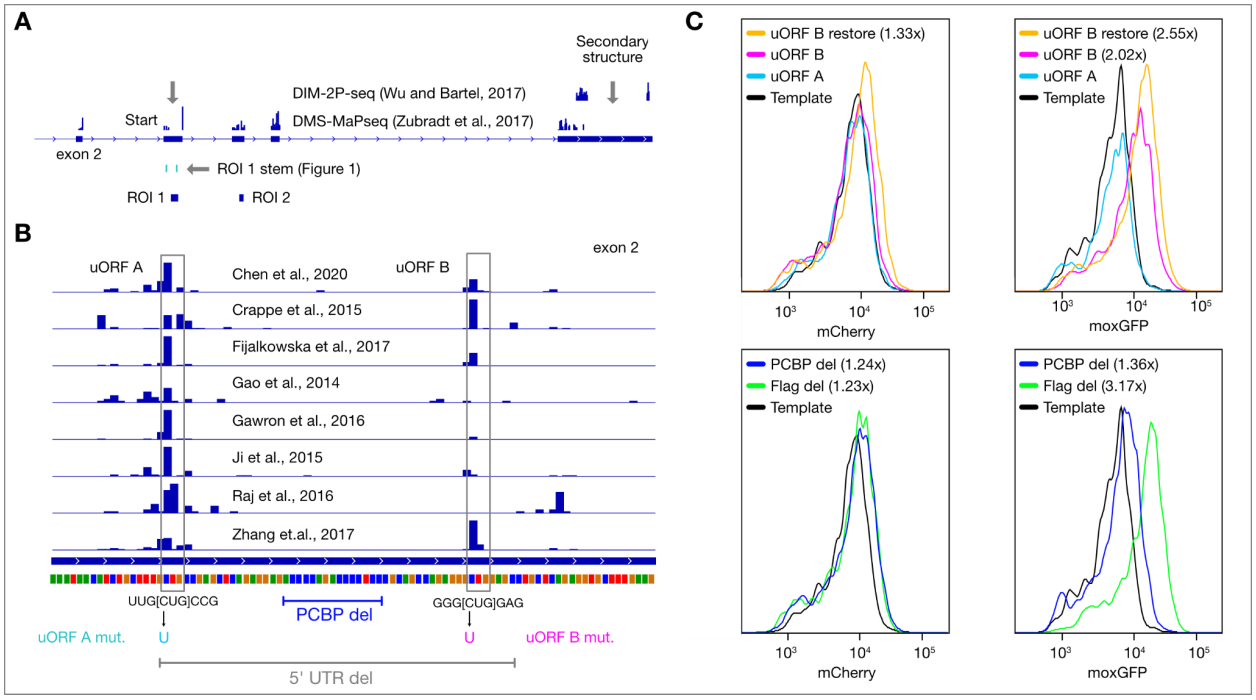
This work was supported in part by the Welch Foundation grant F-2027-20230405 to C.C. C.C. is a CPRIT Scholar in Cancer Research supported by CPRIT grant RR180042.

Authors' contributions

IH- Investigation, Methodology, Formal analysis, Validation, Visualization, Writing-Original Draft. **SR**- Investigation. **CT**- Investigation. **CC**- Conceptualization, Writing-Review and Editing, Supervision, Project administration, Funding acquisition.

Acknowledgements
We would like to thank the Doug Fowler lab for providing the HEK293T landing pad cell line and the Fritz Roth lab for providing the plasmid backbone used for integration into the landing pad.

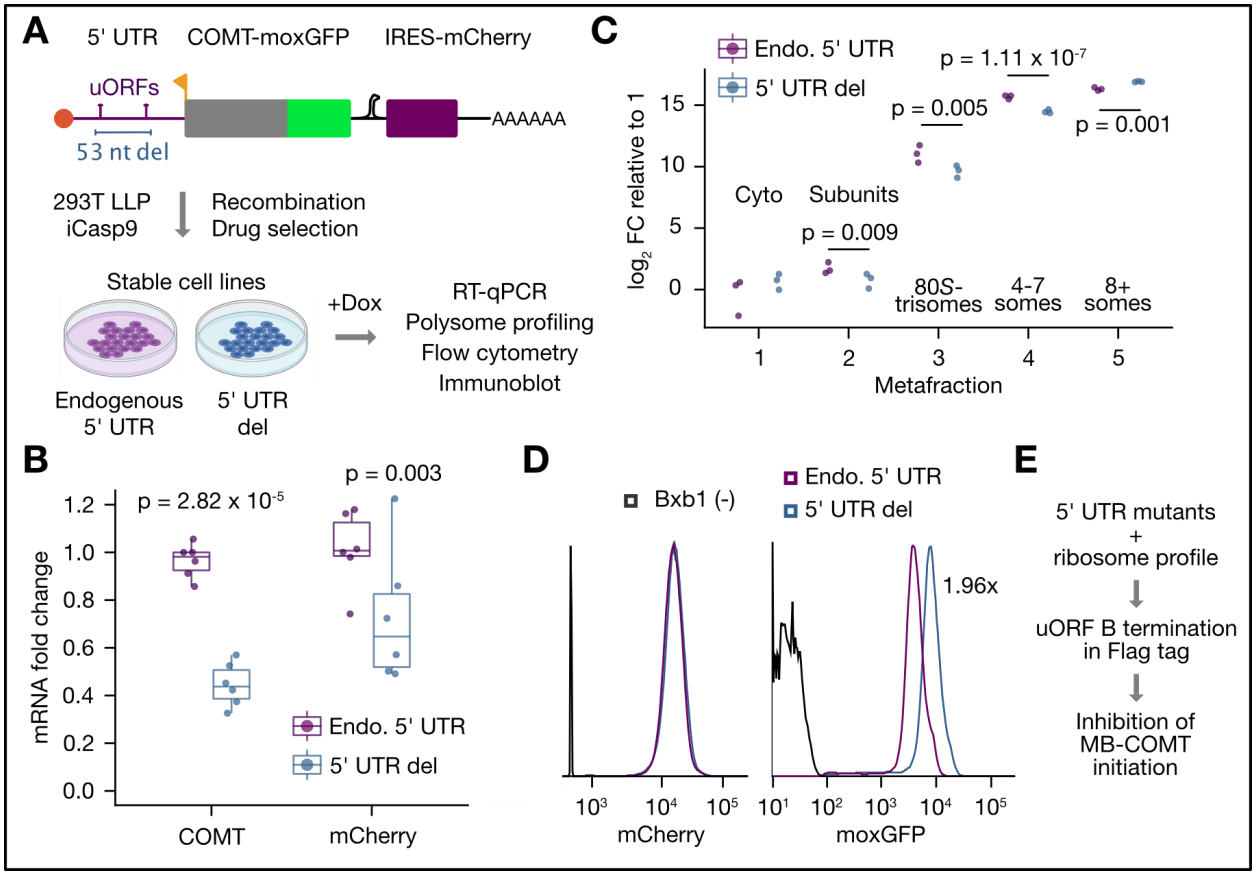
Supplementary Figure 1



Supplementary Figure 1: *COMT* 3′ UTR RNA secondary structure, uncharacterized uORFs, and noncoding variant effects on mRNA and protein abundance

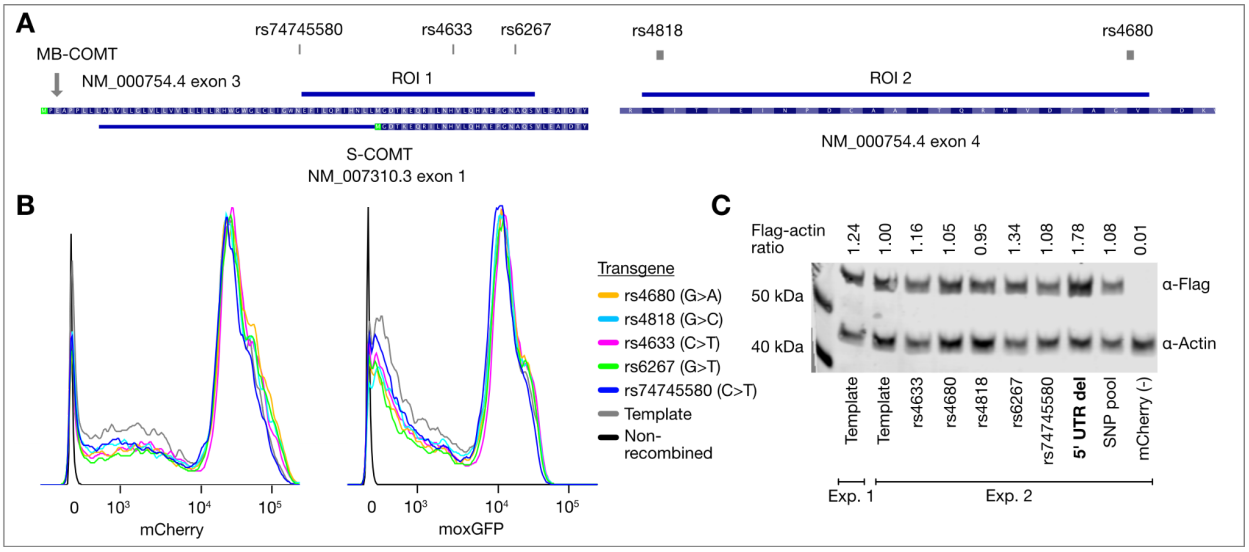
A) *COMT* RNA structure in the 3′ UTR. RNA probing data for HEK293T cells, available from the RASP database (Li et al., 2021), is shown. DIM-2P-seq employs polyA capture to enrich for structures at the 3′ end (Wu and Bartel, 2017). B) *COMT* uORF translation initiation sites. TIS-ribosome profiling data from GWIPS-viz (Michel et al., 2014) was visualized. Two uORFs starting at CUG were identified in the 5′ untranslated exon 2 (NM_000754.4) in multiple cell lines including HEK293T. Exon 2 is conserved among RefSeq transcript isoforms for MB-COMT. At bottom, isolated mutations and deletions are annotated, which are assayed in panel C (uORF start codon mutants, PCBP del) and Supplementary Figure 2 (5′ UTR del). C) Flow cytometry data of noncoding and Flag tag single-nucleotide variants. “uORF A” and "uORF B" are the variants indicated in panel B. "uORF B restore" is a silent variant in the Flag tag that restores the endogenous uORF B frame. "Flag del" is a deletion of the Flag tag. In parentheses is the fold change of median fluorescence between each mutant and the wild-type template. All variants with a reported fold change were significant by two-sided Wilcoxon rank sum tests at $p < 1 \times 10^{-9}$. TIS=translation initiation site; DMS=dimethyl sulfate mutation profiling methods; uORF=upstream open reading frame; UTR=untranslated region; MB-COMT=membrane protein isoform.

Supplementary Figure 2



Supplementary Figure 2: The *COMT* 5′ UTR regulates mRNA and protein abundance
A) Schematic of *COMT* transgene and strategy for stable expression. A *COMT* transgene with a partial 5′ UTR (exon 2 only) and a 53 nt deletion was cloned with a N-terminal Flag, C-terminal moxGFP, and bicistronic mCherry ORF. Transgenes were stably expressed in a HEK293T landing pad line and assayed following induction of transcription with Doxycycline (Dox). B) The 5′ UTR del decreases *COMT* mRNA abundance. The mRNA fold change was determined by RT-qPCR and plotted relative to the median of Endo. 5′ UTR replicates. C) The 5′ UTR del mRNA is decreased in low polysomes and enriched in high polysomes. RNA from raw polysome fractions were pooled into "metafractions" and RNA was extracted and assayed by RT-qPCR. D) The 5′ UTR del leads to higher COMT protein abundance. Flow cytometry results of the transgenic cell lines and a non-recombined (negative control) cell line are shown. E) Proposed model for regulation by the 5′ UTR. 5′ UTR mutants indicated initiation at uORF B and termination within the Flag tag blocks initiation at the canonical TIS. TIS=translation initiation site; moxGFP=monomeric, oxidation-resistant GFP (Costantini et al., 2015); IRES=internal ribosome entry site; 293T LLP iCasp9=lentiviral landing pad (Matreyek et al., 2020).

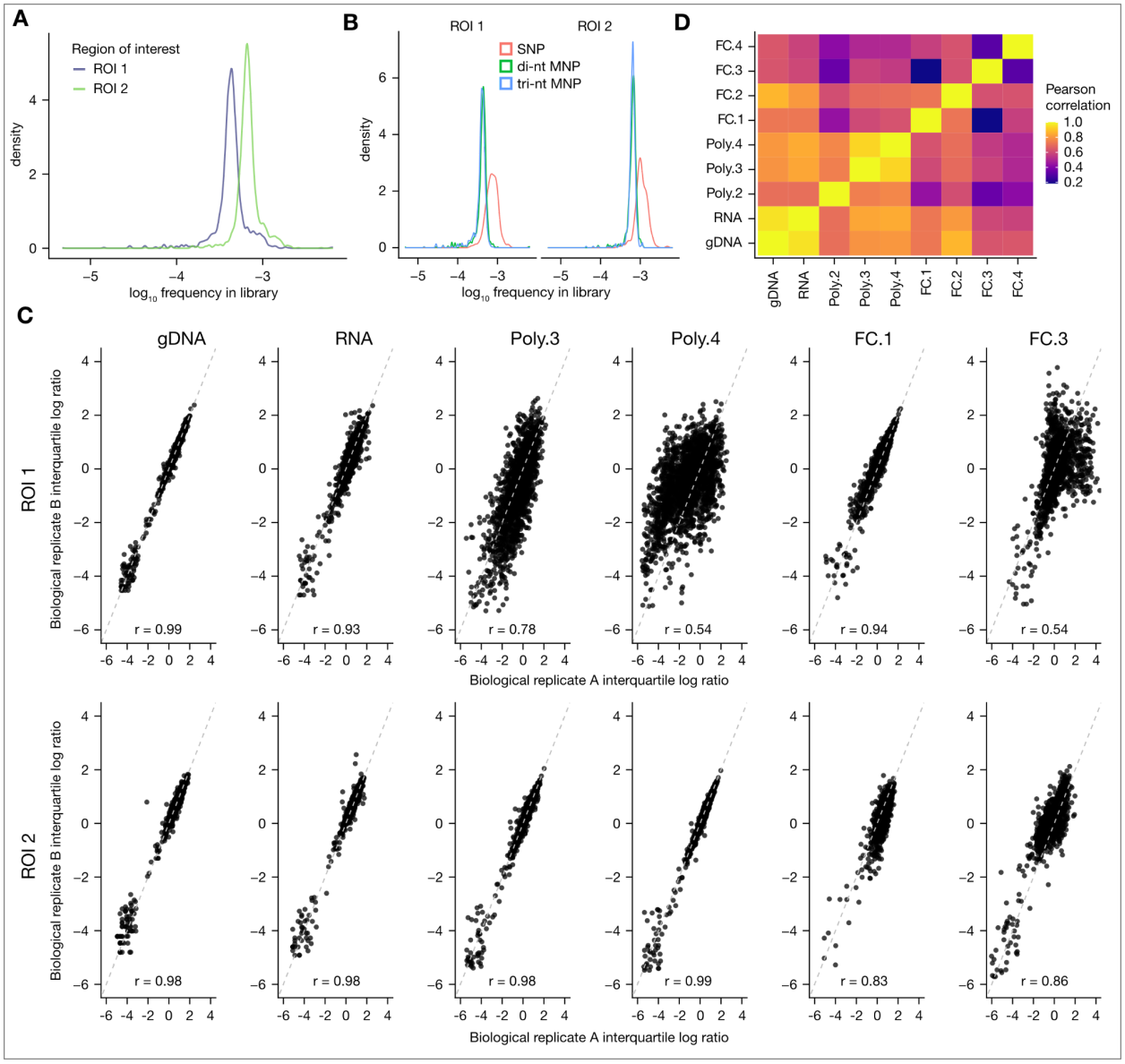
Supplementary Figure 3



Supplementary Figure 3: *COMT* common population SNPs have no impact on mRNA and protein abundance in the absence of complete UTRs

A) *COMT* regions of interest and location of common population SNPs. B) No impact of common population variants on *COMT* mRNA or protein abundance. Fluorescence of mCherry (mRNA proxy) and moxGFP (protein) is plotted for each common population variant (indicated with dbSNP identifiers), along with a non-recombined negative control and the non-mutagenized template, which consists of the LPS haplotype (Nackley et al., 2006) without the intronic SNP rs6269. B) Immunoblotting of common population SNPs. Background-corrected Flag signal to Actin signal from densitometry analysis is indicated at the top of each lane. For the template transgene, lysates from two separate experiments were analyzed. "SNP pool" refers to a stable cell line with all variants expressed. "mCherry (-)" refers to a negative control for Flag expression, an untagged mCherry payload.

Supplementary Figure 4

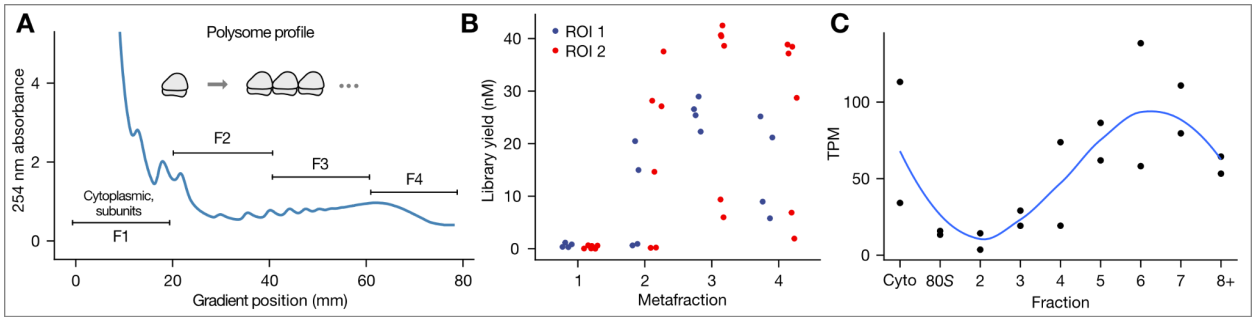


Supplementary Figure 4: *COMT* library quality control and reproducibility of variant abundance estimates

In panels A and B, plotted is kernel density estimation of $\log_{10}$ transformed frequencies reported in the Twist Biosciences quality control report. In panels C and D, and FC.1-4 refer to the four sorted populations, and Poly.2-4 refer to the 2nd-4th metafractions (see Supplementary Figure 5A for schematic). A) Variant uniformity for *COMT* target regions. B) Frequency distributions by variant type. SNPs are at higher frequencies than di-nt and tri-nt MNPs, which have two and three mismatches, respectively. C) Biological replicate correlation. Two of the four biological replicates are shown as examples. Metafractions and flow cytometry populations that were not subsequently analyzed

(Poly.1, Poly.2, FC.2, FC.4) are omitted. D) Correlation between gene expression readouts. Plotted is correlation of the median $\log_{10}$ frequency of biological replicates between readouts.

Supplementary Figure 5



Supplementary Figure 5: *COMT* low versus high polysome fractions

A) Representative polysome trace indicating the fraction pooling strategy. F1 and F2 were not subsequently analyzed due to inadequate library yield or reproducibility. B) Yields for polysome metafraction libraries. Yield was quantified by qPCR using Illumina adapter specific primers (Methods). C) Polysome profile of endogenous COMT in HEK293T cells. Transcripts per million mapped reads (TPM) is plotted using data from TrIP-seq (Floor and Doudna, 2016). Blue line is a loess fit with confidence intervals omitted for clarity. Cyto: cytoplasmic fraction.

References

Andrews S, Snowflack DR, Clark IE, Gavis ER. 2011. Multiple mechanisms collaborate to repress nanos translation in the Drosophila ovary and embryo. *RNA* **17**:967–977.

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**:664–667.

Behrens A, Rodschinka G, Nedialkova DD. 2021. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol Cell* **81**:1802–1815.e7.

Benoit Bouvrette LP, Bovaird S, Blanchette M, Lécuyer E. 2020. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res* **48**:D166–D173.

Bialecka M, Kurzawski M, Klodowska-Duda G, Opala G, Tan E-K, Drozdzik M. 2008. The association of functional catechol-O-methyltransferase haplotypes with risk of Parkinson's disease, levodopa treatment response, and complications. *Pharmacogenet Genomics* **18**:815–821.

Bodenhofer U, Kothmeier A, Hochreiter S. 2011. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**:2463–2464.

Braat AK, Yan N, Arn E, Harrison D, Macdonald PM. 2004. Localization-dependent oskar protein accumulation; control after the initiation of translation. *Dev Cell* **7**:125–131.

Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, Owen MJ, O'Donovan MC. 2003. A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. *Am J Hum Genet* **73**:152–161.

Burke PC, Park H, Subramaniam AR. 2022. A nascent peptide code for translational control of mRNA stability in human cells. *Nat Commun* **13**:6829.

Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, Ricci E, Snyder MP. 2015. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res* **25**:1610–1621.

Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S, Kolachana BS, Hyde TM, Herman MM, Apud J, Egan MF, Kleinman JE, Weinberger DR. 2004. Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. *Am J Hum Genet* **75**:807–821.

Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, Sun S, Roth FP, DeSloover D, Marks DS, Rettie AE, Fowler DM. 2020. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**. doi:10.7554/eLife.58026

Clark IE, Wyckoff D, Gavis ER. 2000. Synthesis of the posterior determinant Nanos is spatially restricted by a novel cotranslational regulatory mechanism. *Curr Biol*

10:1311–1314.
Costantini LM, Baloban M, Markwardt ML, Rizzo M, Guo F, Verkhusha VV, Snapp EL. 2015. A palette of fluorescent proteins optimized for diverse cellular environments. *Nat Commun* **6**:7670.
Dawling S, Roodi N, Mernaugh RL, Wang X, Parl FF. 2001. Catechol-O-methyltransferase (COMT)-mediated metabolism of catechol estrogens: comparison of wild-type and variant COMT isoforms. *Cancer Res* **61**:6716–6722.
DeLano WL. 2002. Pymol: An open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr*.
Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE, Goldman D, Weinberger DR. 2001. Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc Natl Acad Sci U S A* **98**:6917–6922.
Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF. 2019. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*. doi:10.1186/s13059-019-1845-6
Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. 2022. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**:175–183.
Fernandes AD, Macklaim JM, Linn TG, Reid G. 2013. ANOVA-like differential gene expression analysis of single-organism and meta-RNA-seq. *PLoS One*.
Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**:15.
Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM, Shendure J. 2018. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**:217–222.
Floor SN, Doudna JA. 2016. Tunable protein synthesis by transcript isoforms in human cells. *Elife* **5**. doi:10.7554/eLife.10921
Fotsis T, Zhang Y, Pepper MS, Adlercreutz H, Montesano R, Nawroth PP, Schweigerer L. 1994. The endogenous oestrogen metabolite 2-methoxyoestradiol inhibits angiogenesis and suppresses tumour growth. *Nature* **368**:237–239.
Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. 2010. High-resolution mapping of protein sequence-function relationships. *Nat Methods* **7**:741–746.
Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* **315**:972–976.
Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. 2016. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell* **166**:679–690.
Gloor GB, Macklaim JM, Fernandes AD. 2016. Displaying Variation in Large Datasets:

Plotting a Visual Summary of Effect Sizes. *J Comput Graph Stat* **25**:971–979.
Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T. 2017. Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell Reports*. doi:10.1016/j.celrep.2017.07.029
Gothelf D, Law AJ, Frisch A, Chen J, Zarchi O, Michaelovsky E, Ren-Patterson R, Lipska BK, Carmel M, Kolachana B, Weizman A, Weinberger DR. 2014. Biological Effects of COMT Haplotypes and Psychosis Risk in 22q11.2 Deletion Syndrome. *Biological Psychiatry*. doi:10.1016/j.biopsych.2013.07.021
Greenacre M. 2021. Compositional Data Analysis. *Annu Rev Stat Appl* **8**:271–299.
Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res* **36**:W70–4.
GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**:1318–1330.
Guillot CR, Fanning JR, Liang T, Berman ME. 2015. COMT Associations with Disordered Gambling and Drinking Measures. *J Gambl Stud* **31**:513–524.
Hashimoto Y, Shiina M, Maekawa S, Kato T, Shahryari V, Kulkarni P, Dasgupta P, Yamamura S, Saini S, Tabatabai ZL, Dahiya R, Tanaka Y. 2021. Suppressor effect of catechol-O-methyltransferase gene in prostate cancer. *PLoS One* **16**:e0253877.
Hendershot CS, Lindgren KP, Liang T, Hutchison KE. 2012. COMT and ALDH2 polymorphisms moderate associations of implicit drinking motives with alcohol use. *Addiction Biology*. doi:10.1111/j.1369-1600.2010.00286.x
Hoskins I. 2023. satmut_utils: v1.0.3-dev001. *Zenodo*. doi:10.5281/zenodo.7499446
Hoskins I, Sun S, Cote A, Roth FP, Cenik C. 2023. satmut_utils: a simulation and variant calling package for multiplexed assays of variant effect. *Genome Biol* **24**:1–27.
Janacova L, Stenckova M, Lapcik P, Hrachovinova S, Bouchalova P, Potesil D, Hrstka R, Müller P, Bouchal P. 2023. Catechol-O-methyl transferase suppresses cell invasion and interplays with MET signaling in estrogen dependent breast cancer. *Sci Rep* **13**:1285.
Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**:434–443.
Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman

MM. 2007. A “silent” polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**:525–528.

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**:3583.

Klauber N, Parangi S, Flynn E, Hamel E, D’Amato RJ. 1997. Inhibition of angiogenesis and breast cancer in mice by the microtubule inhibitors 2-methoxyestradiol and taxol. *Cancer Res* **57**:81–86.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biol Cybern* **43**:59–69.

Kwong A, Boughton AP, Wang M, VandeHaar P, Boehnke M, Abecasis G, Kang HM. 2022. FIVEx: an interactive eQTL browser across public datasets. *Bioinformatics* **38**:559–561.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**:D862–8.

Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511.

Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC. 2022. sva: Surrogate Variable Analysis.

Lim Y, Arora S, Schuster SL, Corey L, Fitzgibbon M, Wladyka CL, Wu X, Coleman IM, Delrow JJ, Corey E, Others. 2021. Multiplexed functional genomic analysis of 5’untranslated region mutations across the spectrum of prostate cancer. *Nat Commun* **12**:1–18.

Lin C-H, Ray Chaudhuri K, Fan J-Y, Ko C-I, Rizos A, Chang C-W, Lin H-I, Wu Y-R. 2017. Depression and Catechol-O-methyltransferase (COMT) genetic variants are associated with pain in Parkinson’s disease. *Scientific Reports*. doi:10.1038/s41598-017-06782-z

Li P, Zhou X, Xu K, Zhang QC. 2021. RASP: an atlas of transcriptome-wide RNA secondary structure probing data. *Nucleic Acids Res* **49**:D183–D191.

Li Q, Makri A, Lu Y, Marchand L, Grabs R, Rousseau M, Ounissi-Benkalha H, Pelletier J, Robert F, Harmsen E, Hudson TJ, Pastinen T, Polychronakos C, Qu H-Q. 2013.

Genome-wide search for exonic variants affecting translational efficiency. *Nat Commun* **4**:2260.
Li W, Chen Y, Yin B, Zhang L. 2014. Pain in Parkinson's disease associated with COMT gene polymorphisms. *Behav Neurol* **2014**:304203.
Lloyd JPB, French CE, Brenner SE. 2020. Polysome fractionation analysis reveals features important for human nonsense-mediated mRNA decay. *Cold Spring Harbor Laboratory*. doi:10.1101/2020.03.08.981811
Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, Duby M, Sharpe T, Kalkhoven E, Rosen ED, Barroso I, Ellard S, UK Monogenic Diabetes Consortium, Kathiresan S, Myocardial Infarction Genetics Consortium, O'Rahilly S, UK Congenital Lipodystrophy Consortium, Chatterjee K, Florez JC, Mikkelsen T, Savage DB, Altshuler D. 2016. Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**:1570–1575.
Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, Bhatia S, Evans WE, Relling MV, Yang W, Shendure J, Fowler DM. 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**:874–882.
Matreyek KA, Stephany JJ, Chiasson MA, Hasle N, Fowler DM. 2020. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res* **48**:e1.
Matreyek KA, Stephany JJ, Fowler DM. 2017. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res* **45**:e102.
Meyer-Lindenberg A, Nichols T, Callicott JH, Ding J, Kolachana B, Buckholtz J, Mattay VS, Egan M, Weinberger DR. 2006. Impact of complex genetic variation in COMT on human brain function. *Mol Psychiatry* **11**:867–77, 797.
Michel AM, Fox G, M Kiran A, De Bo C, O'Connor PBF, Heaphy SM, Mullan JPA, Donohue CA, Higgins DG, Baranov PV. 2014. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res* **42**:D859–64.
Mukhopadhyay T, Roth JA. 1998. Superinduction of wild-type p53 protein after 2-methoxyestradiol treatment of Ad5p53-transduced cells induces tumor cell apoptosis. *Oncogene* **17**:241–246.
Munro D, Singh M. 2020. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. *Bioinformatics*. doi:10.1093/bioinformatics/btaa1030
Nackley AG, Shabalina SA, Lambert JE, Conrad MS, Gibson DG, Spiridonov AN, Satterfield SK, Diatchenko L. 2009. Low enzymatic activity haplotypes of the human catechol-O-methyltransferase gene: enrichment for marker SNPs. *PLoS One* **4**:e5237.
Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*

314:1930–1933.
Narula A, Ellis J, Taliaferro JM, Rissland OS. 2019. Coding regions affect mRNA stability in human cells. *RNA* **25**:1751–1764.
Ozadam H, Geng M, Cenik C. 2020. RiboFlow, RiboR and RiboPy: an ecosystem for analyzing ribosome profiling data at read length resolution. *Bioinformatics* **36**:2929–2931.
Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. 2017. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* **45**:D750–D757.
Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. 2015. Codon optimality is a major determinant of mRNA stability. *Cell* **160**:1111–1124.
Qian X-K, Wang P, Xia Y-L, Dou T-Y, Jin Q, Wang D-D, Hao D-C, Bi X-L, Ge G-B, Yang L. 2016. A highly selective fluorescent probe for sensing activities of catechol- O -methyltransferase in complex biological samples. *Sensors and Actuators B: Chemical*. doi:10.1016/j.snb.2016.03.074
Richter JD, Coller J. 2015. Pausing on Polyribosomes: Make Way for Elongation in Translational Control. *Cell* **163**:292–300.
Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**:e47.
Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**:701–705.
Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. 2017. A statistical framework for analyzing deep mutational scanning data. *Genome Biol* **18**:150.
Ruottinen HM, Rinne UK. 1998. COMT inhibition in the treatment of Parkinson's disease. *Journal of Neurology*. doi:10.1007/pl00007743
Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. 2013. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat* **34**:546–556.
Sambrook J, Russell DW. 2006. Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc* **2006**. doi:10.1101/pdb.prot2936
Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G. 2019. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37**:803–809.
Schott BH, Frischknecht R, Debska-Vielhaber G, John N, Behnisch G, Düzel E, Gundelfinger ED, Seidenbecher CI. 2010. Membrane-Bound Catechol-O-Methyl Transferase in Cortical Neurons and Glial Cells is Intracellularly Oriented. *Front Psychiatry* **1**:142.
Sei Y, Li Z, Song J, Ren-Patterson R, Tunbridge EM, Iizuka Y, Inoue M, Alfonso BT,

Beltaifa S, Nakai Y, Kolachana BS, Chen J, Weinberger DR. 2010. Epistatic and functional interactions of catechol-o-methyltransferase (COMT) and AKT1 on neuregulin1-ErbB signaling in cell models. *PLoS One* **5**:e10789.

Shifman S, Bronstein M, Sternfeld M, Pisanté-Shalom A, Lev-Lehman E, Weizman A, Reznik I, Spivak B, Grisaru N, Karp L, Schiffer R, Kotler M, Strous RD, Swartz-Vanetik M, Knobler HY, Shinar E, Beckmann JS, Yakir B, Risch N, Zak NB, Darvasi A. 2002. A highly significant association between a COMT haplotype and schizophrenia. *Am J Hum Genet* **71**:1296–1302.

Simhadri VL, Hamasaki-Katagiri N, Lin BC, Hunt R, Jha S, Tseng SC, Wu A, Bentley AA, Zichel R, Lu Q, Zhu L, Freedberg DI, Monroe DM, Sauna ZE, Peters R, Komar AA, Kimchi-Sarfaty C. 2017. Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J Med Genet* **54**:338–345.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. 2015. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**:413–422.

Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, King NP, Veesler D, Bloom JD. 2020. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**:1295–1310.e20.

Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, Yang W, Hoshitsuki K, Singh M, Trehan A, Parish C, Smith C, Li L, Bhojwani D, Yuen LYP, Li C-K, Li C-H, Yang Y-L, Walker GJ, Goodhand JR, Kennedy NA, Klussmann FA, Bhatia S, Relling MV, Kato M, Hori H, Bhatia P, Ahmad T, Yeoh AEJ, Stenmark P, Fowler DM, Yang JJ. 2020. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proc Natl Acad Sci U S A* **117**:5394–5401.

Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, Theesfeld CL, Bansal P, Sahni N, Yi S, Yu A, Tyagi T, Tie C, Hill DE, Vidal M, Andrews BJ, Boone C, Dolinski K, Roth FP. 2016. An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*. doi:10.1101/gr.192526.115

Tesina P, Lessen LN, Buschauer R, Cheng J, Wu CC-C, Berninghausen O, Buskirk AR, Becker T, Beckmann R, Green R. 2020. Molecular mechanism of translational stalling by inhibitory codon combinations and poly(A) tracts. *EMBO J* **39**:e103365.

Tsao D, Shabalina SA, Gauthier J, Dokholyan NV, Diatchenko L. 2011. Disruptive mRNA folding increases translational efficiency of catechol-O-methyltransferase variant. *Nucleic Acids Res* **39**:6201–6212.

Wehrens R, Buydens LMC. 2007. Self- and Super-organizing Maps in R: The kohonen Package. *J Stat Softw* **21**:1–19.

Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, Wu Y, Pons C, Wong C, van Lieshout N, Yang F, Tasan M, Tan G, Yang S, Fowler DM, Nussbaum R, Bloom JD, Vidal M, Hill DE, Aloy P, Roth FP. 2017. A framework for exhaustively mapping functional missense variants. *Mol Syst Biol* **13**:957.

Weng X, Gong J, Chen Y, Wu T, Wang F, Yang S, Yuan Y, Luo G, Chen K, Hu L, Ma H, Wang P, Zhang QC, Zhou X, He C. 2020. Keth-seq for transcriptome-wide RNA structure mapping. *Nat Chem Biol* **16**:489–492.

Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, Wright M, Bazzini AA. 2019. Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife* **8**. doi:10.7554/eLife.45396

Wu X, Bartel DP. 2017. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell* **169**:905–917.e11.

Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, Chen J, Robinson HE, Shim HS, Chmielecki J, Pao W, Engelman JA, Iafrate AJ, Le LP. 2014. Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med* **20**:1479–1484.

Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS, Rouskin S. 2017. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods* **14**:75–82.