Integrative analysis of RNA, translation and protein levels reveals distinct regulatory variation across humans

Authors: Can Cenik[1], Elif Sarinay Cenik[1], Gun W. Byeon[1], Fabian Grubert[1], Sophie I Candille[1], Damek Spacek[1], Bilal Alsallakh[2], Hagen Tilgner[1], Carlos L. Araya[1], Hua Tang[1], Emiliano Ricci[3,4], Michael P. Snyder[1,*]

Affiliations:

[1] Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA 94305.

[2] Institute of Software Technology & Interactive Systems, Vienna University of Technology, Karlsplatz 13, Vienna, Austria.

[3] RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, Massachusetts, USA.

[4] *CIRI, International center for Infectiology Research, eukaryotic and viral translation team, Université de Lyon, INSERM U1111, Lyon, France*.

*Correspondence to: mpsnyder@stanford.edu

Keywords: Translation, human variation

Abstract: Elucidating the consequences of genetic differences between humans is essential for understanding phenotypic diversity and personalized medicine. Although variation in RNA levels, transcription factor binding and chromatin have been explored, little is known about global variation in translation and its genetic determinants. We used ribosome profiling, RNA sequencing, and mass spectrometry to perform an integrated analysis in lymphoblastoid cell lines from a diverse group of individuals. We find significant differences in RNA, translation, and protein levels suggesting diverse mechanisms of personalized gene expression control. Combined analysis of RNA expression and ribosome occupancy improves the identification of individual protein level differences. Finally, we identify genetic differences that specifically modulate ribosome occupancy - many of these differences lie close to start codons and upstream ORFs. Our results reveal a new level of gene expression variation among humans and indicate that genetic variants can cause changes in protein levels through effects on translation.

Introduction

Deciphering the molecular mechanisms that underlie human variation is essential for understanding human diversity and personalized medicine. To date, genetic variants that affect protein function in humans have been well studied, but those that control protein levels are less well characterized. Yet, misregulation of protein levels can have profound consequences for human health. For example, transcriptional regulatory mutations that increase telomerase gene expression have been identified in ~70% of melanoma patients (Horn et al., 2013; Huang et al., 2013) and are frequent in several other cancers (Huang et al., 2013). Similarly, changes in protein levels of SHANK3, neuroligins and neurexins have been linked to autism spectrum disorder, schizophrenia and learning disorders (Darnell, 2011). Therefore, understanding how RNA levels and translation efficiency control protein levels on an individual basis is required not only for understanding human phenotypic diversity, but also for personalized medicine as thousands of human genome sequences become available.

Protein expression is determined at many levels including (1) RNA expression, (2) translation efficiency and (3) protein stability. Recent studies have begun to unravel the extent of human variation at RNA levels and its control through transcription factor binding sites, and chromatin (Kasowski et al., 2010, 2013; McDaniell et al., 2010; Montgomery et al., 2010; Pai et al., 2012; Pickrell et al., 2010; Stranger et al., 2007; Westra et al., 2013). However, protein levels often correlate poorly with RNA expression (Ly et al., 2014; Vogel and Marcotte, 2012). Translation efficiency, i.e. the number of proteins synthesized per mRNA, has been suggested to account for a large component of the unexplained variation in protein levels (Marguerat et al., 2012; Schwanhäusser et al., 2011). While recent studies in yeast have begun to address the genetic control of translation efficiency (Albert et al., 2014; Artieri and Fraser, 2014; McManus et al., 2014; Muzzey et al., 2014), little is currently known about variation in translation efficiency and its genetic determinants in humans. Further, an integrated view of how expression is

controlled at many different levels is lacking in humans.

Here, we utilized RNA-seq and ribosome profiling to identify ribosome occupancy profiles. Ribosome profiling involves RNase digestion of unprotected RNA and isolation of ribosome-bound mRNA segments. The sequences of ribosome protected mRNA fragments can then be used to deduce the number of ribosomes per message in conjunction with RNA-seq data. We integrated these measurements with quantitative proteomics to reveal a comprehensive view of the variation in gene expression programs across a diverse set of humans.

Results

Measuring Ribosome Occupancy across Individuals at a Global Scale

To measure genome-wide ribosome occupancy of mRNAs, we first improved and adapted the ribosome profiling protocol (Ingolia et al., 2009, 2011, 2012) for lymphoblastoid cell lines (LCLs) (Figure 1A). A critical step in ribosome profiling method involves RNase digestion of unprotected RNAs before isolating and sequencing ribosome-associated mRNAs. While optimizing the protocol, we observed that RNase I digestion caused extensive degradation of ribosome integrity (Figure 1B-C, Supplemental Figure S1A). The loss of polyribosome signal was not accompanied by a corresponding increase in monosome signal (Figure 1C; Supplemental Figure S1A) but rather a shift towards lighter fractions indicative of free and degraded RNAs (Figure 1C; Supplemental Figure S1A). Hence, RNase I treatment can lead to a loss in ribosome integrity in addition to producing the expected 80S ribosome footprints (i.e. monosomes).

We tested whether other RNases could alleviate this problem and found that treatment with RNase A and RNase T1 (which collectively cut after C, U and G) resulted in complete digestion of polyribosomes into monosomes (Figure 1D; Supplemental Figure S1A). Recent

work in *Drosophila* and other systems also reported the importance of optimizing nuclease digestion to generate robust ribosome profiling data (Dunn et al., 2013; Ricci et al., 2014). Using our optimized ribosome profiling protocol, we generated ribosome occupancy maps for LCLs obtained from thirty individuals of diverse ethnic backgrounds: five Europeans, two Asian and twenty-three Yorubans with significant genetic diversity (Figure 1E). These lines were chosen because a) their genomes have been sequenced (The 1000 Genomes Project Consortium 2012; The International HapMap 3 Consortium 2010), b) their relative protein and RNA levels have been previously measured (Khan et al., 2013; Wu et al., 2013), and c) they can be grown in large quantities. Importantly, the ribosome occupancy maps were based on at least two replicate samples for the majority of individuals. In parallel, we generated 44 deep RNA-seq libraries (with a median of ~12M uniquely transcriptome mapped reads) from the same cells and combined these with those from previous work ('t Hoen et al., 2013; Lappalainen et al., 2013; Pickrell et al., 2010) thereby providing multiple RNA-seq replicates for most individuals.

We leveraged replicate measurements to assess data quality and its dependence on several parameters including alignment strategy, mRNA enrichment method, PCR artifacts, gene length normalization and batch effects (Supplemental Figure S2A-E, Supplemental Methods). In addition to verifying the high quality of the data, replicate measurements also enabled modeling of gene-specific variance in RNA expression and ribosome occupancy per individual, allowing robust derivation of individual-specific translation efficiency estimates. Specifically, we developed a linear modeling based approach to regress out the effects of RNA expression from ribosome occupancy measurements to calculate translation efficiency (Supplemental Methods).

Finally, for 28 individuals studied here, we previously measured relative protein abundances via isobaric tag-based quantitative proteomics using the same cell lines (Wu et al., 2013). In total, we present a combined analysis of 133 high-throughput sequencing libraries (83 RNA-seq and 50 ribosome profiling libraries) and extensive protein expression measurements

(Figure 1E).

Integrative Analysis of RNA Expression, Ribosome Occupancy and Protein Levels

We first considered the relationship between RNA expression, ribosome occupancy, and protein expression across genes. As expected, RNA expression and ribosome occupancy were highly correlated (Supplemental Figure S2F; Spearman ρ= 0.87, p-value < 2.2 x 10^{-16}; outlier robust correlation 0.88 using Donoho-Stahel estimator), albeit still lower than biological replicates of RNA expression data (Spearman ρ ~ 0.98, p-value < 2.2 x 10^{-16}) indicating that control of ribosome occupancy levels is distinct from RNA levels. Importantly, ribosome occupancy correlated better with protein levels than RNA expression correlated with protein levels (Supplemental Figure S2F-G; Spearman ρ of 0.54 and 0.43, Donoho-Stahel estimator based correlation coefficient 0.56, and 0.42, respectively; permutation test for difference in correlation coefficient p-value < 10^{-4}). Consistent with previous results (Ingolia et al. 2009), these results suggest that ribosome occupancy is a better predictor of protein level differences between genes.

While the correlation analysis reveals pairwise relationships, the interdependencies between RNA expression, translational efficiency, and protein levels are not captured. For example, some genes with high protein levels can have low RNA expression but very high translation efficiency, yielding a decreased correspondence between RNA expression and protein levels. To reveal such interdependencies, we utilized self organizing maps (SOM), an integrative machine learning method that is robust to noise and allows assessment of all relationships simultaneously (Figure 2A) (Kohonen, 1990; Wehrens and Buydens, 2007). Since SOMs are sensitive to differences in mean and variance of the input variables, we first converted each measurement into their relative rank order expressed as percentiles ensuring equal weighting of the input variables for the SOM training (Supplemental Methods). After

training, each neuron within the SOM contains genes that share a similar pattern of expression and protein level (Figure 2A; Supplemental Figure S2H).

The emerging map recapitulated the pairwise relationships between RNA expression, ribosome occupancy and protein levels across neurons (Figure 2B). We further grouped neurons in the SOM using a clustering approach (affinity propagation clustering (Frey and Dueck, 2007). This approach uncovered nine clusters in the SOM, revealing the distinct relationships between RNA expression and translation efficiency in determining protein levels. For example, genes in Cluster 6 have relatively high RNA expression but do not reach high protein levels as they are translationally repressed (Figure 2C).

We then examined functional (GO term) enrichments (Berriz et al., 2003, 2009) across the different clusters within the SOM and found specific functional enrichments for four of the nine clusters (Figure 2C; Supplemental Table S1). Genes with high translation efficiency and high protein levels were enriched for diverse functional categories such as the proteasome complex, glycolysis, mRNA splicing, and DNA damage checkpoint (Supplemental Table S1; selected examples are shown in Figure 2D; p-value < 0.05 for all categories using permutation based multiple hypothesis correction). Conversely, genes associated with translation and cytosolic ribosome constituents were enriched among those that exhibited very high RNA and protein levels despite having lower translation efficiencies (Supplemental Table S1; Figure 2D; p-value < 0.05 for all categories using permutation based multiple hypothesis correction), raising the possibility that higher protein stability or feedback mechanisms on translation efficiency modulate the levels of translation components. These findings indicate that some sets of functionally coherent genes adopt alternative strategies to achieve their respective steady-state protein levels.

Gene Expression Variability between Individuals

We next focused on how ribosome occupancy and RNA expression differ between individuals. We leveraged replicate measurements and identified genes with *significant* inter-individual variance in RNA expression or ribosome occupancy, exceeding technical noise. We found that ~27% of genes had statistically significant inter-individual variation in RNA expression compared to only ~7% of genes that had detectable variation in ribosome occupancy (Figure 3A; 3B; Holm's method adjusted p-value < 0.05 based on a simulation based likelihood ratio test). Consequently, ~20% of all genes exhibit inter-individual RNA expression variation which is not reflected in ribosome occupancy. These results were not explained by different sensitivities of the measurements (Supplemental Figure S3A). These results were also consistent when restricting the analysis to only the Yoruban individuals or when excluding RNA expression data not generated by our laboratory, indicating the robustness of the results.

Genes that exhibited significant inter-individual variation in both RNA expression and ribosome occupancy were highly enriched for gene ontology terms including: "chemokine receptor activity", "complement activation", " leukocyte migration", "antigen binding" (Figure 3C; Supplemental Table S2; p < 0.05 permutation based multiple hypothesis correction), indicating a role in immune functions. Consistently, protein levels that exhibit the most variation between individuals were previously shown to be enriched for "immune system process" (Wu et al., 2013). These functional categories are highly specific to the function of the studied cell type, LCLs (Figure 3C; Supplemental Table S2). Given that genes with significant inter-individual variation were directly pertinent to the function of the cell line studied here, it is likely that carrying out similar studies in other cell types will expand the set of genes whose expression levels differ significantly between individuals.

Within genes that exhibited significant inter-individual variation in both RNA expression and ribosome occupancy, we identified three subsets. Within the first subset the variability in RNA expression was comparable to variability in ribosome occupancy (Supplemental Figure S3E). This first subset contained 54% of all genes exhibiting inter-individual variation in both

RNA expression and ribosome occupancy. The second subset consisted of genes that had higher RNA level variability compared to ribosome occupancy variability. This subset encompassed nearly twice as many genes as the third subset where ribosome occupancy variability was higher than that of RNA expression (Supplemental Figure S3E). These results were consistent with the observation that for many genes inter-individual RNA expression variation is not reflected in ribosome occupancy. Taken together, our results are consistent with yeast studies that reported translational buffering of divergent RNA expression (Artieri and Fraser, 2014; McManus et al., 2014). However, we note that an alternative explanation of our findings is the presence of a pool of untranslated pool of mRNA (e.g., nuclear-retained or sequestered cytoplasmically in P bodies) that is variable between individuals.

A small, but interesting fraction of genes (0.7%) exhibited differential ribosome occupancy between individuals with no apparent differences at the RNA-level (Supplemental Table S3). These were enriched in genes coding for proteins involved in “cellular response to chemical stimulus” and the “Golgi apparatus” (Supplemental Table S2; $p < 0.05$ permutation based multiple hypothesis correction). These results suggest that translational control may play important roles in cellular signaling, whereby rapid cellular responses are often required.

Relationship between Individual Differences in Protein Levels, Ribosome Occupancy, and RNA Expression

An outstanding question in understanding phenotypic variation is how individual-specific protein levels relate to corresponding differences in gene expression. We previously measured relative protein levels for ~6000 proteins using the same cell lines (Wu et al., 2013). As expected, the protein level measurements were skewed towards genes that are more highly expressed and translated (Supplemental Figure S3B, Wilcoxon rank sum test $p < 2.2 \times 10^{-16}$). We first calculated the correlation between RNA expression and the corresponding protein level

across individuals. Consistent with previous results (Wu et al., 2013), the median correlation coefficient was 0.22, with 11% of genes showing a statistically significant correlation (Figure 3D; Spearman correlation coefficient, 5% FDR using Benjamini-Hochberg method; Supplemental Figure S3C).

We next repeated this analysis for the set of genes that we identified as having significant RNA expression variability between individuals. Among this subset, relative protein levels and RNA expression had a median correlation coefficient of 0.43 (Spearman correlation coefficient; Supplemental Figure S3D), indicating a partial correlation between RNA and protein variability.

Finally, we tested whether joint measurement of RNA expression and ribosome occupancy improved this correspondence. Specifically, we considered genes that exhibit significant inter-individual variation in both ribosome occupancy and RNA expression. Strikingly, 83% of these genes had statistically significant correlation between differences in protein levels and RNA expression (Figure 3D; Spearman correlation coefficient, 5% FDR using Benjamini Hochberg method) with a median correlation coefficient of 0.67; and Supplemental Figure S3C). The large difference between the correlation coefficients indicates that measuring both ribosome occupancy and RNA levels simultaneously greatly improves the ability to identify gene expression variability between individuals that will eventually result in personal differences in protein levels.

Genetic Determinants of Variability in Ribosome Occupancy

We next investigated whether genetic differences between individuals were associated with the observed variation in gene expression, specifically at the ribosome occupancy level. We used two complementary approaches. First, we used the 21 unrelated individuals from the Yoruban population and conducted a cis-quantitative trait loci (cis-QTL) mapping approach.

Using the cis-QTL mapping strategy, we identified significant association between single nucleotide polymorphisms (SNPs) and ribosome occupancy for 67 genes (Supplemental Figure S4A-D; Supplemental Table S4; 30% FDR). While 34 out of the 67 roQTLs were not associated with significant differences in RNA expression (nominal association p-value >0.05), this analysis cannot conclusively show that these roQTLs are not associated with RNA expression. Overall roQTLs had consistent effects on RNA expression and protein levels (Supplemental Figure S4A-D; Spearman ρ=0.86, $p < 2.2 \times 10^{-16}$). Consistent with recent work comparing two yeast strains (Albert et al., 2014), these results suggest that genetic effects that were propagated through RNA expression to ribosome occupancy caused consistent changes in protein levels for this set of genes.

The Role of uORFs in Modulating Ribosome Occupancy

We next adopted a targeted approach that was both better powered and enabled detection of combined effects of multiple genetic variants on ribosome occupancy. We first analyzed variants modifying upstream open reading frames (uORFs), which can alter protein expression by regulating translation (Wethmar et al., 2014). In humans, approximately half of annotated transcripts contain uORFs, and presence of uORFs is widely polymorphic across individuals (Barbosa et al., 2013; Calvo et al., 2009; Waern and Snyder, 2013; Supplemental Table S5). Disruption of a uORF in the *HR* gene has been previously shown to lead to Marie Unna hereditary hypotrichosis by modulating the translation of the main ORF suggesting that human disease can be associated changes in uORFs (Wen et al. 2009).

We correlated genetic alterations to uORF presence with ribosome occupancy of the main coding region and found 33 transcripts with significant association (Figure 4A-D, Supplemental Figure S4E-F; 5% FDR). One particular advantage of this targeted approach was the ability to detect changes to uORFs caused by multiple genetic variants. For example, two

different SNPs in the ZNF215 gene result in merging of two uORFs by removing a stop codon (Figure 4D). Merging of the uORFs significantly increased ribosome occupancy of the main coding region (Figure 4D; $p<0.001$).

In addition to impacting translational efficiency, nucleotide variants changing uORFs may alter RNA abundance. For example, they may change transcript stability by triggering nonsense-mediated decay (Kervestin and Jacobson, 2012). Alternatively, the variant or variants in linkage disequilibrium may alter transcriptional output as a proximal element downstream of the transcription start site. Of the 33 significant associations between changes in uORFs and ribosome occupancy, ~52% (17 out of 33) also had a significant effect on RNA levels (nominal p-value < 0.05), indicating the presence of uORFs and RNA levels are often coupled. However, for 16 other genes the observed effect was solely on ribosome occupancy suggesting direct modulation of the translation efficiency of the main reading frame (Figure 4A-B, Supplemental Table S5). We further observed that presence of uORFs could be associated with both increased and decreased ribosome occupancy of the main coding region (Supplemental Figure S4G). We verified the robustness of these results by limiting the analysis to data from Yoruban individuals and employing an alternative statistical framework based on linear mixed models (Supplemental Figure S4E-F and Supplemental Table S5). These results reveal that natural genetic variation within the human population can specifically cause personal differences in translation through changes to uORFs.

The Role of Sequences Surrounding the Start Codon in Modulating Ribosome Occupancy

We next analyzed the Kozak sequence, the region surrounding the start codon for its effect on translation efficiency (Figure 5A). Previous work has suggested that Kozak sequence is important for both start codon selection and translation efficiency of specific transcripts

(Kozak, 1987). However, the extent and impact of natural genetic variation affecting the Kozak sequence and the global effect of the Kozak region on translation efficiency have not been studied.

We first determined whether certain positions of the Kozak sequence have a global effect on translation efficiency. We found a highly significant and large effect of the nucleotides at position -3 and at position -2 on translation efficiency (Figure 5A; Supplemental Figure S5A; Bonferroni adjusted Kruskal-Wallis test $p=5.7 \times 10^{-20}$ for position -3; $p=1.2 \times 10^{-17}$ at position -2). Additionally, the two nucleotides immediately after the start codon had statistically significant effects on translation efficiency (Supplemental Figure S5A; Bonferroni adjusted $p < 2.8 \times 10^{-7}$). While previous work using reporter systems anticipated the significance of these features (for example, Kozak, 1987), our analyses highlight the role of sequence composition near the ATG in modulating translation efficiency of endogenous genes at a global scale.

The extent and potential role of natural variation that might alter the Kozak sequence across the genome remains largely unexplored in the human population (Xu et al., 2010; Supplemental Table S6). Among the set of individuals studied here, there were ~150 genetic variants altering the Kozak region in at least three individuals. ~65% of Kozak region variants reduced the PWM score of the reference sequence (Supplemental Figure S5B). This effect was even more pronounced for Kozak variants that were observed in a single individual. 77% of these reduced the PWM score of the reference sequence suggesting that selective pressure may be acting to optimize the Kozak sequence.

We next tested the effect of these variants on ribosome occupancy of the main coding region. We utilized the position weight matrix for the Kozak region to score the impact of each variant on the Kozak strength (Figure 5A). We found nine genes with Kozak variants that modified ribosome occupancy significantly with no significant effect on the RNA levels (Figure 5B-C; 10% FDR using Benjamini-Hochberg correction; RNA expression association p-value > 0.01; Supplemental Figure S5C; using a conservative linear mixed model, two of these genes

had $p < 0.01$), indicating the presence of variants specifically affecting translation efficiency.

Finally, to directly examine the role of genetic variation on translation efficiency, we used reporter assays (Jang et al., 1988) for six genes. These included four genes with Kozak region variants and two genes with uORF variants. We cloned the reference 5'UTR or the variant 5'UTR with a single base change at the Kozak region or the uORF upstream of a *Renilla* luciferase and transfected the resulting constructs into human HEK 293 cells. To normalize differences in RNA expression and transfection efficiency, an HCV internal ribosome entry site driven Firefly luciferase was cloned after the *Renilla* stop codon and the ratio of the *Renilla* to Firefly luciferase was quantified. Differences in this ratio between the reference and variant 5'UTR-containing reporters for four genes recapitulated the results from our ribosome profiling data i.e. sequences that were associated with reduced translational efficiency also gave low luciferase ratios. These results provide an independent validation of our conclusion that natural genetic variation can modify sequences surrounding the start codon leading to personal differences in translation (Figure 5D-F; Supplemental Figure S5D).

Discussion

This study demonstrates that translation efficiency varies among individuals and that nucleotides important for regulating translation efficiency can be identified. In several cases, we uncovered the mechanisms controlling translation efficiency variation in humans. These included uORFs and sequences near the translation initiation sites. Our study revealed that genetic differences between individuals could lead to gene expression differences at the level of translation.

We leveraged replicate measurements to identify genes with significant variability in RNA expression or ribosome occupancy between individuals. We found that genes that exhibit significant variability in both RNA expression and ribosome occupancy were highly enriched for

functions directly pertinent to LCLs such as immune response and leukocyte migration (Figure 3). Hence, extending this analysis framework to additional cell types or tissues will likely uncover more genes with variable expression between individuals.

We also investigated the relationship between protein levels differences and variability in RNA expression and translation. We found that joint analysis of RNA expression and translation improved our ability to identify the extent of gene expression variation that would be reflected in protein levels, indicating a tight coupling of translation efficiency and protein levels. These analyses were skewed towards genes with higher expression levels due to missing protein level measurements (Supplemental Figure S3B). Hence, further improvements in proteomics technology will be needed to test the generalizability of our results to lowly expressed proteins. Despite the significant improvements obtained by joint analysis of ribosome occupancy and RNA level measurements, there remains unexplained variability in protein levels. One potential contributor to this discrepancy is variability in protein degradation rates (Vogel and Marcotte, 2012). Another important future direction will be to investigate the contribution of RNA sequence features (such as in Vogel et al. 2010) to the relationship between RNA expression, ribosome occupancy, and protein levels.

Importantly, genes that have individual variability only in RNA expression are less likely to have corresponding differences at the protein level. Among this subset of genes, only 40% had statistically significant co-variation between RNA levels and protein levels (5% FDR). An important implication of this result concerns ongoing efforts that aim to identify genetic determinants of RNA expression (Battle et al., 2014; Lappalainen et al., 2013). These studies are in part motivated by the finding that most disease risk factors identified by genome-wide association studies lie in noncoding regions (Edwards et al., 2013). By linking genetic differences to RNA expression, these studies hope to uncover functional connections to disease states. Yet, our analyses suggest that the functional impact of RNA-level differences needs to be carefully considered to establish causal relationships to phenotype.

Recent consortium efforts measured RNA expression in large sets of genotyped samples (~900 in Battle et al., 2014; ~500 samples in Lappalainen et al., 2013) to identify trans-acting genetic effects on RNA expression. Interestingly, 85% of the trans-effects on RNA expression were mediated by the effects of the associated SNP on a nearby gene (Battle et al., 2014), indicating that changes in regulators of RNA expression lead to differences in RNA levels of distant transcripts. Similarly, genetic variation in translation regulators is likely to have trans-effects on ribosome occupancy of many transcripts. For example, levels of global regulators of translation such as MTOR, and translation initiation factors (e.g., EIF4E) have the potential to modulate the translation of a large number of targets (Mamane et al., 2007; Thoreen et al., 2012). In fact, a recent study that compared translation in two different strains of yeast suggested that the relative importance of trans-effects on translation is comparable to that for RNA levels (Albert et al., 2014). Future studies in the human population will likely uncover trans-acting and additional cis-acting genetic variants associated with translation, and reveal the contribution of population-level variation to translation variability.

A recent analysis (Battle et al. 2015) of RNA expression, ribosome occupancy and protein measurements from several human LCLs concluded that there is a scarcity of human genetic variants associated with translation-specific effects. However, we note critical limitations in their ribosome profiling data. As demonstrated in Figure 1 and recently by Miettinen and Björklund (2015), the nuclease digestion conditions employed in Battle et al. (2015) lead to severe ribosomal degradation and significantly lower monosome purity in ribosome profiling libraries. Moreover, the Battle et al. study design lacks replicate experiments, precluding proper assessment of the reproducibility of ribosome profiling measurements. In re-sequencing experiments, Battle et al. reported rank correlations lower than 0.9 (Spearman ρ) for the majority of their samples (Figure S2A in Battle et al. 2015). In contrast, we consistently achieved >0.98 rank correlations between biological replicates using independently grown cells and independently prepared ribosome profiling libraries. Here, by leveraging higher quality ribosome

profiling datasets with replicates and independent reporter experiments, we identify genetic variants associated with translation efficiency undetected in Battle et al.

Our study revealed several genetic variants that control translation efficiency variation in humans, including those affecting the Kozak region and upstream open reading frames (uORFs). A particularly interesting question is the molecular mechanisms of these sequence-function relationships. An intriguing feature of genetic variants modifying uORFs on translation was the observation that both gain and loss events could lead to increased translation of the downstream open reading frame (Supplemental Figure S4G) consistent with previous work that implicated both positive and negative regulation of translation efficiency by uORFs (Brar et al., 2012; Waern and Snyder, 2013). Whereas several mechanisms have been implicated in negative regulation of translation efficiency by uORFs, including nonsense mediated decay (Kervestin and Jacobson, 2012), less is known about the mechanisms of positive regulation by uORFs. Recent work identified a complex, DENR-MCT1, that catalyzes translation reinitiation downstream of certain uORFs (Schleich et al., 2014), suggesting that DENR-MCT1 or similar factors may act on subsets of uORFs to increase reinitiation frequency of the downstream ORF leading to higher translation efficiency.

Recent structural analysis of the yeast 48S translation initiation complex permitted an unprecedented view of the molecular environment of the start codon in eukaryotes (Hussain et al., 2014) revealing a potential mechanism by which Kozak region variants affect translation efficiency. Remarkably, this structural analysis revealed that eIF2alpha directly contacts nucleotides at positions -2 and -3, the same two positions that our global analysis of Kozak variants highlighted as being the most important for translation efficiency (Figure 5A). Thus, our results provide functional evidence that these residues are of general importance for translational efficiency.

Together, these results demonstrate that genetic alterations in the human population and disease-associated mutations may penetrate to phenotype through changes in translation. In

the era of personal genome sequencing, this information is crucial for understanding the role of genetic variants on gene expression, phenotypic traits and human disease susceptibility.

Methods:

RNA-seq experiments and Ribosome profiling experiments:

Human lymphoblastoid cell lines (LCLs) were obtained from Coriell Cell Repository. For replicate ribosome profiling replicate experiments, cells were grown separately to a density of 0.8-1.0 x 10^6 cells/mL. Approximately 10 million cells were pelleted at 250g at 4C and washed with PBS. The pellets were frozen in liquid nitrogen prior to cell lysis. 7 A260 Units of the cleared cell lysates were incubated with 300U Units of RNase T1 (Fermentas) and 500ng of RNase A (Ambion). A 34% (Weight/Volume) sucrose cushion was used to isolate ribosomes. Library preparation and sequencing was done as previously described with some modifications (Ingolia et al., 2012; Supplemental Methods).

For RNA-seq experiments, LCLs were grown to a density of $3x10^5$ – $6x10^5$ cells/ml. Total RNA was extracted using TRIzol reagent according to the manufacturer's instructions (Lifetechnologies), then purified using the Qiagen RNeasy kit (Qiagen, Valencia, CA) and treated with RNase-free DNase (Qiagen, Valencia, CA). RNA integrity was checked with a Bioanalyzer (Agilent, Santa Clara CA) and only samples with an RNA integrity number (RIN) of > 9.5 were subsequently subjected to either ribosomal depletion or poly-A-selection. For ribosomal RNA depletion, 5 µg of purified total RNA was depleted of rRNAs using the Ribo-Zero Magnetic Gold Kit (Human/Mouse/Rat) (Epicentre Biotechnologies, Madison, WI). For poly-A selection, 10 µg of purified total RNA were enriched by performing two cycles of selection using the Dynabeads mRNA Purification Kit (Life Technologies). Stranded libraries were prepared following the dUTP protocol (Parkhomchuk et al., 2009). For each cell line, we generated 2 x 101bp paired end RNA-seq data using two biological replicates of ribosomal RNA depleted and three biological replicates of poly-A-selected RNA.

Sequence alignment and processing:

To enable comparable analysis of high throughput sequencing datasets, we employed a uniform alignment and preprocessing pipeline. Reads were sequentially aligned using Bowtie 2 v.2.0.5 (Langmead and Salzberg, 2012). All reads mapping to human rRNA and tRNA sequences were filtered out. The remaining reads were aligned to APPRIS principal transcripts (release 12) (Rodriguez et al., 2013) from the GENCODE mRNA annotation v.15 (Harrow et al., 2012). For all transcript level analyses, reads that map only to coding regions were used. For details, see Supplemental Methods.

Ribosome profiling sample identity verification:

The cell line identity for all ribosomal profiling libraries were verified by comparing empirically generated genotype calls to the reference genotypes. Specifically, we utilized samtools mpileup utility in combination with bcftools (Li et al., 2009) to generate genotype calls from the ribosomal profiling read alignments. Finally, a custom perl script was used to compare the number of perfect matches between empirically called genotypes and the reference genotype that was available from the HapMap and the 1000 Genomes Project. For details, see Supplemental Methods.

Genotype data and processing:

Genome sequences were obtained from the 1000 Genomes Project pilot 2 trios and Phase1v3 (The 1000 Genomes Project Consortium 2012; The International HapMap 3 Consortium 2010) for 27 of the 30 individuals. The genome sequences of three cell lines (NA19139, NA19193, and NA19201) were imputed from HapMap release 28 data (The International HapMap 3 Consortium 2010; The International HapMap Consortium 2007) to the 1000 Genomes Phase1v3 reference panel (The 1000 Genomes Project Consortium 2012).

We included all variant calls provided by both release and pilot datasets without

additional score or source filtering. We subsetted all single nucleotide polymorphisms (SNPs) that overlap APPRIS transcripts andretained all phasing information from the VCF files. About ~8% of the variants in the pilot dataset were unphased, and for these variants, we randomly assigned the phase. For details, see Supplemental Methods.

Sequence data normalization and quality control:

After accounting for differences in mRNA enrichment method, ~9600 transcripts had a read count per million reads mapped (cpm) (as implemented in the edgeR package (McCarthy et al., 2012)) greater than one in at least 40 RNA-seq libraries and 36 ribosome profiling libraries. We used trimmed mean of M values to account for differences in library size (Robinson and Oshlack, 2010), and estimated the mean to variance relationship in the data using the voom method (Law et al., 2014). We explicitly specified the individual identifier to indicate which libraries were replicates from the same individual while applying the voom method. The inverse variance weights obtained from the voom method were used in all analyses where applicable. For details, see Supplemental Methods.

Calculation of translation efficiency:

When combined with RNA expression measurements, ribosome profiling enables the estimation of translation efficiency by capturing a snapshot of the transcriptome-wide ribosome occupancy. We treated ribosome profiling and RNA-seq as two experimental manipulations of the RNA pool of the cell. Translation efficiency was calculated using a linear model where the normalized expression values are dependent on the treatment (RNA-seq or Ribosome profiling) and the individual identifiers (limma R package Smyth, 2005). For details, see Supplemental Methods.

Self-Organizing maps for integrative gene expression analysis:

We used SOMs to explore the relationship between protein levels and the three expression measurements: RNA levels, ribosome occupancy and translation efficiency. SOMs rely on a suitable measure of distance between the transcripts for the clustering. To avoid skewing distance calculation due to difference in scale and variance of the expression measurements, expression levels and protein amounts were converted to percentiles using the empirical cumulative distribution function for each level. The kohonen R package (Wehrens and Buydens, 2007) was used for training the SOM with custom modifications to the plotting functions following (Xie et al., 2013). We then clustered the codebook vectors of the 140 units in the SOM using affinity propagation clustering (Frey and Dueck, 2007) as implemented in the apcluster R package (Bodenhofer et al., 2011). For details, see Supplemental Methods.

Gene set enrichment analysis:

FuncAssociate 2.0 was used for gene set enrichment analyses (Berriz et al., 2009). The background gene list was explicitly defined as the set of all genes that could potentially be included in the query set. We defined significant enrichments as GO terms with an odds ratio greater than 2 and adjusted p-value < 0.05. P-value adjustment was carried out using a permutation method to account for the overlap between the GO terms. We calculated the Kappa Similarity Score between all pairs of significantly enriched GO terms. We retained edges between all pairwise GO terms whose Kappa similarity score was greater than 0.1. Enriched GO terms were visualized with Cytoscape (Smoot et al., 2011) using the edge-weighted spring embedded layout. For details, see Supplemental Methods.

Analysis of between individual variation in RNA expression and ribosome occupancy:

Replicate measurements for RNA-seq and ribosome profiling were used to determine inter-individual variance while controlling for platform specific variance observed between

replicates from the same individual. To decompose these two variance components, we used a linear mixed effects model where we treated the individual as a random effect. As before, we utilized the inverse variance weights obtained from the voom approach and fitted the model using log-likelihood instead of a restricted maximum likelihood approach. We tested the null hypothesis that the variance of the random effect is zero. Rejection of the null hypothesis implied that there was significant inter-individual variance in the expression of the given transcript. We adopted a simulation-based approach using an exact likelihood ratio test implemented in RLRsim R package (Scheipl et al., 2008). Multiple-hypothesis correction was applied to RNA expression and ribosome occupancy p-values separately using Holm's method. For details, see Supplemental Methods.

Cis-QTL identification:

Association between gene expression and the genotype at each variant position located in the exons of the APPRIS transcripts was tested in the set of 21 unrelated Yoruban individuals using PLINK v1.07 (Purcell et al., 2007). For each transcript, replicate gene expression measurements were averaged for this analysis. The expression values were regressed on variant genotypes assuming an additive genetic model where genotype was coded as 0,1, or 2 copies of the alternate allele and restricting the testing to variants with a minor allele frequency >10% in the 21 unrelated Yoruban individuals.

Genetic Determinants of Variability in Ribosome Occupancy:

Testing the effect of uORF events on ribosome occupancy

We used AUG and CUG as potential start codons, and UAG, UAA and UGA as potential stop codons. CUG initiation has been reported in few well-documented cases such as *FGF2*, *VEGF*, *MYC* and *MHC class I* transcripts (Hann et al., 1988; Meiron et al., 2001; Schwab et al., 2003; Vagner et al., 1996). Additionally, recent studies mapping genome-wide translation

initiation sites have suggested that upstream translation initiates frequently from non-AUG codon, most prominently at CUG sites (Ingolia et al., 2011; Lee et al., 2012).

To group individuals by uORF differences on a given transcript, we first determined all possible combinations of uORF gain/loss events. We then tested whether the copy number of the uORF variants affects ribosome occupancy of the main coding region using two approaches. In the first approach, we used linear regression. In the second, more conservative approach, we fitted a linear mixed model assuming the difference in cell lines is an individual-specific random effect, i.e. treating the different cell lines of the same individual as "technical replicates". For details, see Supplemental Methods.

The effect of Kozak region sequence on translation efficiency:

We defined the Kozak region as the six nucleotides preceding the start codon and the two nucleotides following the start codon. We extracted the nucleotide sequence of this region from all annotated APPRIS transcripts and built a position weight matrix (PWM), which recapitulated the known Kozak sequence (Figure 5A). We tested whether the nucleotide content of the Kozak sequence affected translation efficiency using the Kruskal-Wallis test. Specifically, we tested whether transcripts split into four categories based on the nucleotide at a given position has the same translation efficiency. We corrected the p-value from this test using Bonferroni correction for the 8 tests (number of positions) that were performed.

Association between Kozak region genetic variants and ribosome occupancy:

Next, we collected all SNPs that intersect annotated Kozak regions. We scored the variant and the reference Kozak sequence using the PWM matrix obtained above. We coded each variant by the PWM score change and assumed an additive relationship between different positions in the Kozak region and copy number of the allele. We then tested whether the variants in the Kozak regions affect ribosome occupancy of the main coding region using a

linear model. For all Kozak variants affecting ribosome occupancy, we conducted the same association test using RNA expression level as the phenotype. As for the uORF analysis, we deemed RNA association to be not significant if the nominal p-value was greater than 0.05, or if the regression coefficient had the opposite sign. For details, see Supplemental Methods.

Luciferase reporter assays:

To assay translation efficiency, we used of a bicistronic luciferase reporter construct (Jang et al., 1988). This construct has an SV40 promoter that drives the expression of a bicistronic transcript that includes both the firefly and *Renilla* luciferase. While the *Renilla* luciferase translation is cap-dependent, firefly luciferase has an Hepatitis C virus (HCV) internal ribosome entry site (IRES) that enables cap-independent translation.

Gene segments were synthesized and cloned right in front of the start codon (ATG) of the *Renilla* luciferase using the CloneEZ system (GenScript). The bicistronic constructs were transfected into HEK293 cells. Cap dependent translation was calculated by taking the ratio of *Renilla* (cap) to firefly (HCV IRES) luciferase activity and derived from 5 replicate experiments. The HCV-IRES dependent translation of firefly luciferase accounted for differences in RNA expression and transfection efficiency. Outlier detection was carried out as described (Jacobs and Dinman, 2004). The difference between *Renilla* to firefly luciferase ratios was assessed using a Welch two sample two sided t-test. For details, see Supplemental Methods.

Data Access:

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE65912.

Acknowledgements:

CC is supported by Child Health Research Institute, Lucile Packard Foundation for Children's Health and the Stanford CTSA grant number UL1TR000093. GWB is supported by a Benchmark Stanford Graduate Fellowship. SC is supported by a Stanford Transformative and Innovation grant. DS was supported by NIH/NHGRI T32 HG000044 and the Genentech Graduate Fellowship. Hua Tang is supported by the NIH grant GM07305907. This research was supported by NIH grants 1U01HG00761101, HG007611, and HG006996. We would like to thank Basar Cenik, Douglas H Phanstiel, Robert J Nichols for comments. We would like to thank Ghia Euskirchen for sequencing support. Illumina sequencing services were performed by the Stanford Center for Genomics and Personalized Medicine.

Author contributions:

CC, MPS designed the study. CC carried out ribosome profiling experiments and coordinated the computational analyses. GWB carried out uORF analysis, SC contributed to ribosome occupancy QTL analysis, BA developed visualization tools, HT contributed to RNA-seq analysis. ESC performed luciferase reporter assays, FG generated RNA-seq data, DS carried out cell culture/maintenance. ER helped optimize the ribosome profiling protocol, CA, and Hua Tang provided statistical analysis guidance. CC, ESC, CA, and MPS wrote the manuscript with input from all authors.

Figure Legends:

Figure 1. Choice of RNase is critical for generating ribosome profiling data

(A) A schematic representation of ribosome profiling strategy was shown. A key step is the digestion of unprotected RNA segments with an RNase. The ribosome protected RNA segments are isolated using a sucrose cushion and prepared for high throughput sequencing **(B)** Human lymphoblastoid cells (GM12878) were lysed in the presence of cycloheximide. The samples were ultracentrifuged through a 10-50% sucrose gradient. Samples were fractionated while continuously monitoring absorbance at 254nm. A representative polysome profile is shown **(C)** Samples were prepared for ultracentrifugation as in panel (b) with the following exception. The cleared lysate was incubated with 100 Units of RNase I (Ambion) for 30 minutes at RT before the ultracentrifugation. (**D)** Samples were prepared as in panel (c), except 300 Units of RNase T1 (Fermentas) and 500ng of RNase A (Ambion) was used for the RNase digestion step. A complete digestion of polysomes into monosomes was observed **(E)** Schematic representation of the datasets used in the current study. Genotype, ribosome profiling, RNA-seq and mass spectrometry based proteomics data were collected from lymphoblastoid cells derived from a diverse group of 30 individuals.

Figure 2. Ribosome occupancy correlates better with absolute protein levels than RNA expression and protein levels

(A) A self-organizing map (SOM) was trained using ribosome occupancy, RNA expression, translation efficiency and protein levels. These measurements were converted into their relative rank order before training. After training, each neuron in the SOM contains several genes sharing similar expression patterns. **(B)** Four different colorings of the trained SOM depict the mean ribosome occupancy, RNA expression, translation efficiency or protein levels for each neuron. **(C)** Neurons of the SOM were grouped using affinity propagation clustering (Frey and Dueck, 2007). Shared coloring between nodes indicates membership to the same cluster. For

each cluster, the mean rank in ribosome occupancy (RO), RNA expression (RE), Translation efficiency (TE) and protein level (PL) was shown for the representative neuron of the cluster. The number of genes in each cluster (n) was shown. **(D)** For four out of nine clusters, significantly enriched gene ontology (GO) terms were identified (FuncAssociate (Berriz et al., 2009) permutation based corrected p-value < 0.05; Supplemental Table S1). For two clusters (5 and 8), selected GO categories were shown (log2 odds ratio; Supplemental Table S1 for the full list of enriched terms).

Figure 3. Identification of genes with significant inter-individual variability in RNA expression and ribosome occupancy improves the ability to identify personal differences in protein levels

(A) Ribosome occupancy and RNA expression was modeled using a linear mixed model treating individuals as a random effect and mean expression as the fixed effect. A simulation based exact likelihood ratio test (Scheipl et al., 2008) was used to compare the linear mixed model to a linear model that did not include the individual as a predictor. The number of genes that show significant inter-individual in RNA expression or inter-individual variation in ribosome occupancy was plotted (Holm's corrected p-val < 0.05). **(B)** The Venn diagram depicts the overlap between the two groups **(C)** Enriched gene ontology (GO) terms among genes with significant inter-individual variation in both RNA expression and ribosome occupancy was determined using FuncAssociate (Berriz et al., 2009). Cytoscape (Smoot et al., 2011) was used to visualize the enriched GO terms (permutation test corrected p-value <0.05, odds ratio > 3; Supplemental Table S2). Nodes correspond to GO terms and are colored by the corrected p-value. The size of the node is proportional to the logarithm of the odds ratio. The similarity between GO terms was quantified using Kappa similarity. The strength of the similarity was visualized using darker edge colors (Supplemental Methods). An edge-weighted spring embedded layout was shown. **(D)** For each gene, Spearman correlation was calculated between individual specific RNA expression and relative protein levels. The distribution of the

correlation coefficients was plotted as a density. Genes that showed significant variation in both RNA expression and ribosome occupancy between individuals were plotted with red bars and genes without detectable variation in RNA expression and ribosome occupancy were shown with white bars.

Figure 4. Nucleotide variants that modify upstream ORFs can alter ribosome occupancy of the main coding region

(A) We identified single nucleotide polymorphisms that generate, delete or otherwise modify an upstream open reading frame (uORF). We tested whether changes to uORFs affected ribosome occupancy of the main coding region using a linear regression framework. The absolute value of the effect size from the regression was plotted against the p-value of association. For 17 uORF changes shown with red circles, the association was solely with ribosome occupancy (nominal p-value > 0.05 or opposite signed regression coefficients for RNA expression, and Supplemental Table S5 for robustness to population stratification and linear mixed model). **(B)** A SNP in the 5'UTR of LENG8 gene introduces a premature in-frame stop codon that shortens an existing uORF. This event results in lower ribosome occupancy of the main coding region, as shown in the boxplot (p_{Ribo} = 0.002). The horizontal bar reflects the median of the distribution and the box depicts the inter-quartile range. The whiskers are drawn to 1.5 times the inter-quartile range. **(C)** In another example, SRRM1, a SNP completely eliminates an existing uORF by removing its start codon. The loss of this uORF is associated with reduced ribosome occupancy of the main coding region (p_{Ribo} = 0.0004; p_{RNA}= 0.19). **(D)** The reference sequence of ZNF215 gene has two short uORFs. Two different genetic variants eliminate the stop codon of the first uORF (UGA to UAC or UGA to CAA) resulting in merging of the two short uORFs into a single long uORF. The merging of the uORF significantly modulates both ribosome occupancy and RNA expression (p_{Ribo} = 0.0001; and $p_{RNA} = 10^{-9}$ respectively).

Figure 5. Nucleotide variants modulating the sequence around the translation initiation site alter translation efficiency

(A) Kozak region is defined as the six nucleotides preceding and two nucleotides following the start codon. The derived position weight matrix was visualized using WebLogo (Crooks et al., 2004). The upper panel shows the effects of each nucleotide at the -3 position on translation efficiency. The effect of nucleotides on translation efficiency was tested using Kruskal-Wallis test. **(B)** The effect of a Kozak region variant on the ribosome occupancy of NTPCR was assessed using a linear model (p-value = 1.1×10^{-6}). A boxplot was used to visualize the distribution of ribosome occupancy for individuals with given genotypes. The horizontal bar reflects the median of the distribution and the box is drawn to depict the inter-quartile range. **(C)** WDR11 had two naturally occurring SNPs in its Kozak region. An additive model was adopted to calculate the change in the position weight matrix score of the Kozak region. **(D)** 5'UTRs with or without Kozak variants were cloned into a translation efficiency reporter. The reporter expresses a biscistronic mRNA where the *Renilla* luciferase is translated under the control of the cloned 5'UTR and the Firefly luciferase is translated under the control of Hepatitis C virus (HCV) internal ribosome entry site (IRES). **(E)** and **(F)** The ratio of *Renilla* to Firefly luciferase activity was plotted for NTPCR (e) and WDR11 (f). Error bars represent s.e.m. The difference between the ratios was assessed using a two sided two sample t-test (* denotes p-value < 0.05).

References:

't Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. *31*, 1015–1022.

Albert, F. W., Muzzey, D., Weissman, J. S., & Kruglyak, L. (2014). Genetic Influences on Translation in Yeast. *PLoS Genetics*, *10*(10), e1004692. doi:10.1371/journal.pgen.1004692

Artieri, C. G., & Fraser, H. B. (2014). Evolution at two levels of gene expression in yeast. *Genome Research*, *24*(3), 411–421. doi:10.1101/gr.165522.113

Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. *9*, e1003529.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., Gilad, Y. (2015) Impact of regulatory variation from RNA to protein. Science, 347(6222), 664-667. doi: 10.1126/science.1260793.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*,*24*(1), 14–24.

Berriz, G.F., King, O.D., Bryant, B., Sander, C., and Roth, F.P. (2003). Characterizing gene sets with FuncAssociate. Bioinformatics *19*, 2502–2504.

Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M., and Roth, F.P. (2009). Next generation software for functional trend analysis. Bioinformatics *25*, 3043–3044.

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

Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2012). High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science (New York, N.Y.)*, *335*(6068), 552–557.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Darnell, J.C. (2011). Defects in translational regulation contributing to human cognitive and behavioral disease. Curr. Opin. Genet. Dev. *21*, 465–473.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife *2*, e01179.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the Dark Road from Association to Function. *American Journal of Human Genetics*, *93*(5), 779–797.

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111.014050.

Hann, S.R., King, M.W., Bentley, D.L., Anderson, C.W., and Eisenman, R.N. (1988). A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. Cell *52*, 185–195.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–8.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

Hussain, T., Llácer, J. L., Fernández, I. S., Munoz, A., Martin-Marcos, P., Savva, C. G., Lorsch, J.R., Hinnebusch, A.G., Ramakrishnan, V. (2014). Structural Changes Enable Start Codon Recognition by the Eukaryotic Translation Initiation Complex. *Cell*,*159*(3), 597–607.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Jacobs, J.L., and Dinman, J.D. (2004). Systematic analysis of bicistronic reporter assay data. Nucleic Acids Res. *32*, e160.

Jang, S. K., Kräusslich, H. G., Nicklin, M. J., Duke, G. M., Palmenberg, A. C., & Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *Journal of Virology*, *62*(8), 2636–2643.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription factor binding among humans. Science *328*, 232–235.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kervestin, S., and Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. Nat. Rev. Mol. Cell Biol. *13*, 700–712.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. Science *342*, 1100–1104.

Kohonen, T. (1990). The self-organizing map. Proc. IEEE *78*, 1464–1480.

Komili, S., Farny, N.G., Roth, F.P., and Silver, P.A. (2007). Functional specificity among ribosomal proteins regulates gene expression. Cell *131*, 557–571.

Kozak, M. (1987). At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J. Mol. Biol. *196*, 947–950.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Larsson, O., Sonenberg, N., and Nadon, R. (2010). Identification of differential translation in genome wide studies. Proc. Natl. Acad. Sci. U. S. A. *107*, 21487–21492.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M.J., Stratton, M.R., and Lamond, A.I. (2014). A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. Elife *3*, e01630.

Mamane, Y., Petroulakis, E., Martineau, Y., Sato, T.-A., Larsson, O., Rajasekhar, V. K., & Sonenberg, N. (2007). Epigenetic Activation of a Subset of mRNAs by eIF4E Explains Its Effects on Cell Proliferation. *PLoS ONE*, *2*(2), e242.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bähler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671–683.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.

McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235–239.

McManus, C. J., May, G. E., Spealman, P., & Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research*, *24*(3), 422–430. doi:10.1101/gr.164996.113

Meiron, M., Anunu, R., Scheinman, E.J., Hashmueli, S., and Levi, B.Z. (2001). New isoforms of VEGF are translated from alternative initiation CUG codons located in its 5'UTR. Biochem. Biophys. Res. Commun. *282*, 1053–1060.

Miettinen, T. P., & Björklund, M. (2015). Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3′ untranslated regions. *Nucleic Acids Research*, *43*(2), 1019–1034. doi:10.1093/nar/gku1310.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Muzzey, D., Sherlock, G., & Weissman, J. S. (2014). Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Research*, *24*(6), 963–973. doi:10.1101/gr.166322.113

Olshen, A.B., Hsieh, A.C., Stumpf, C.R., Olshen, R.A., Ruggero, D., and Taylor, B.S. (2013). Assessing gene-level translational control from ribosome profiling. Bioinformatics *29*, 2995–3002.

Pai, A.A., Cain, C.E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., Degner, J.F., Gaffney, D.J., Pickrell, J.K., Stephens, M., et al. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. *8*, e1003000.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. *37*, e123.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ricci, E.P., Kucukural, A., Cenik, C., Mercier, B.C., Singh, G., Heyer, E.E., Ashar-Patel, A., Peng, L., and Moore, M.J. (2014). Staufen1 senses overall transcript secondary structure to regulate translation. Nat. Struct. Mol. Biol. *21*, 26–35.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., and Tress, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. *41*, D110–7.

Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Comput. Stat. Data Anal. *52*, 3283–3299.

Schleich S, Strassburger K, Janiesch PC, Koledachkina T, Miller KK, Haneke K, Cheng Y-S, Küchler K, Stoecklin G, Duncan KE, et al. 2014. DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth. *Nature* **512**: 208–212.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431–432.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, (Springer New York), pp. 397–420.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

The International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Thoreen, C. C., Chantranupong, L., Keys, H. R., Wang, T., Gray, N. S., & Sabatini, D. M. (2012). A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*, *485*(7396), 109–113. doi:10.1038/nature11083

Vagner, S., Touriol, C., Galy, B., Audigier, S., Gensac, M.C., Amalric, F., Bayard, F., Prats, H., and Prats, A.C. (1996). Translation of CUG- but not AUG-initiated forms of human fibroblast growth factor 2 is activated in transformed and stressed cells. J. Cell Biol. *135*, 1391–1402.

Vogel C, Abreu R de S, Ko D, Le S-Y, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6**: 400.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Waern, K., and Snyder, M. (2013). Extensive transcript diversity and novel upstream open reading frame regulation in yeast. G3 *3*, 343–352.

Wehrens, R., and Buydens, L. (2007). Self-and super-organizing maps in R: the Kohonen package. J. Stat. Softw.

Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, Sun M, Li Y, Yang S, Zhang X-J, et al. 2009. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat Genet* **41**: 228–233.

Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A., and Leutz, A. (2014). uORFdb--a comprehensive literature database on eukaryotic uORF biology. Nucleic Acids Res. *42*, D60–7.

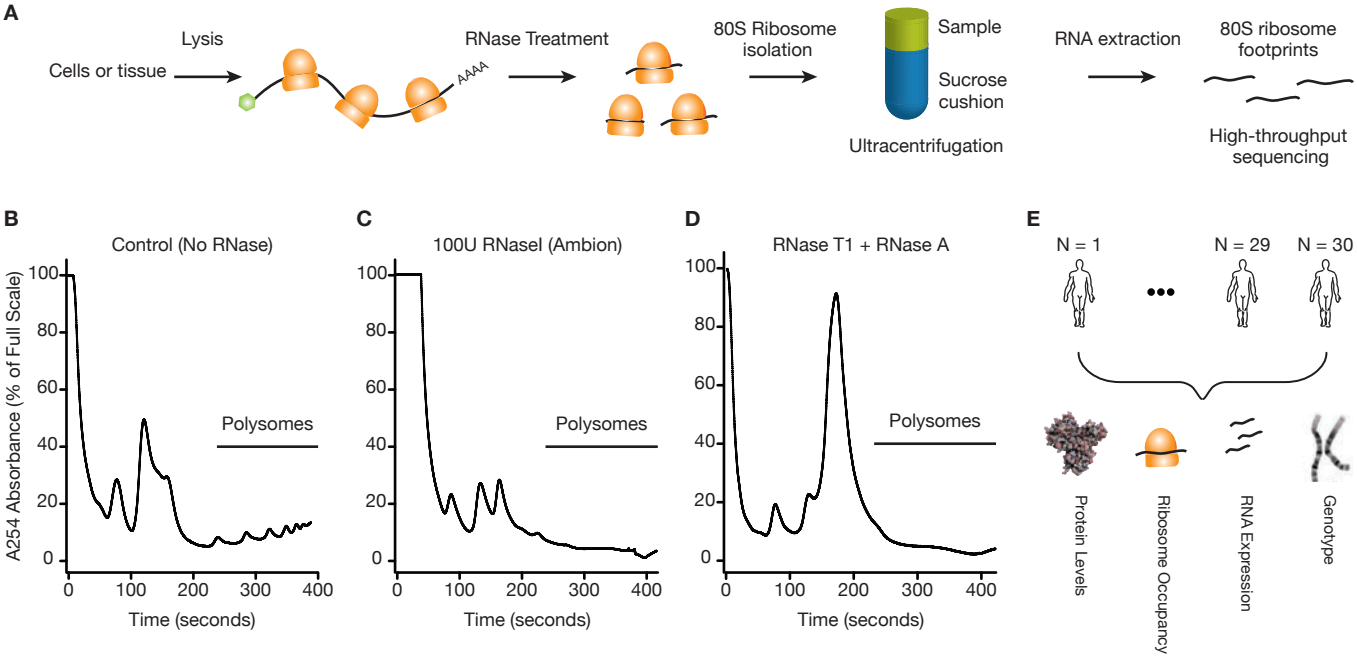
Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79–82.

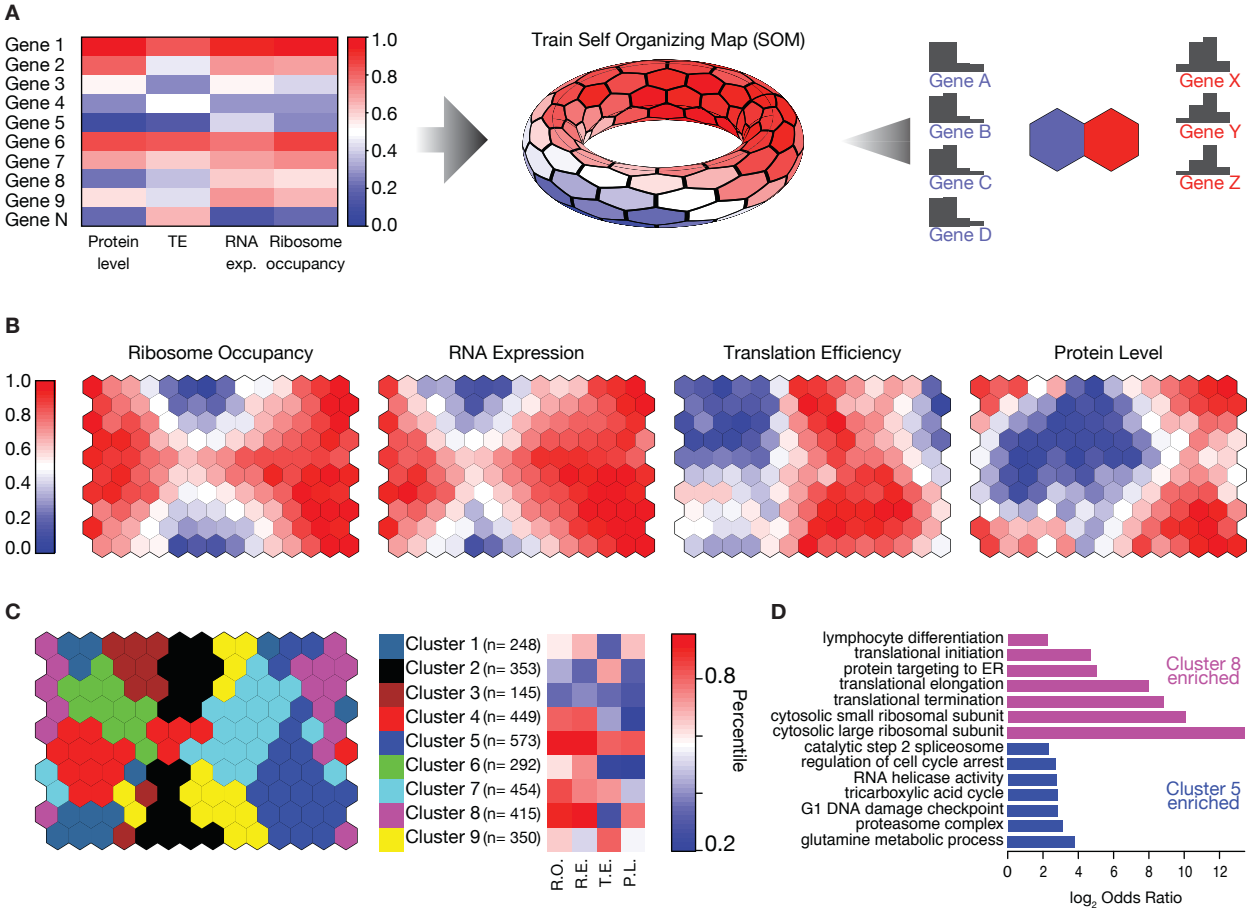
Xie, D., Boyle, A.P., Wu, L., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-acting factor colocalization in human cells. Cell *155*, 713–724.

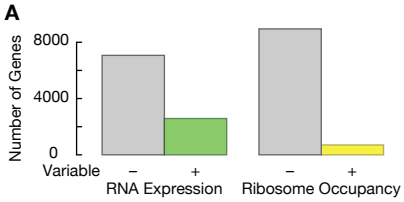
Xu, H., Wang, P., You, J., Zheng, Y., Fu, Y., Tang, Q., Zhou, L., Wei, Z., Lin, B., Shu, Y., et al. (2010). Screening of Kozak-motif-located SNPs and analysis of their association with human diseases. Biochem. Biophys. Res. Commun. *392*, 89–94.

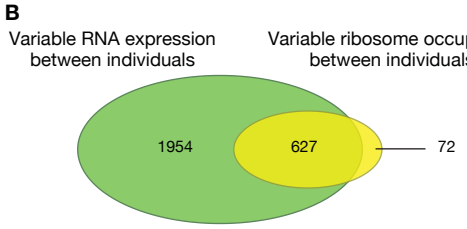
Yin, H. *The Self-Organizing Maps: Background, Theories, Extensions and Applications* 715-762 (Springer Berlin Heidelberg, 2008).

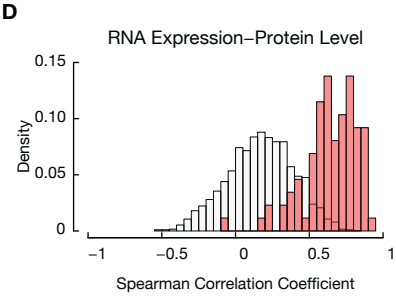
Figure 1

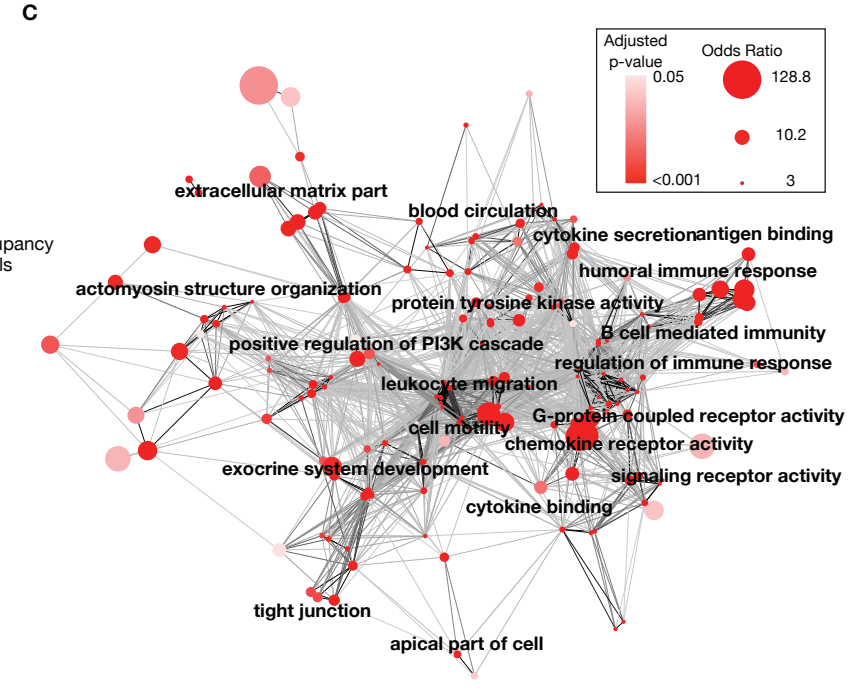

A
Lysis
Cells or tissue
AAAA
RNase Treatment
80S Ribosome isolation
Sample
Sucrose cushion
Ultracentrifugation
RNA extraction
80S ribosome footprints
High-throughput sequencing
B
Control (No RNase)
C
100U RNaseI (Ambion)
D
RNase T1 + RNase A
Polysomes
A254 Absorbance (% of Full Scale)
Time (seconds)
E
N = 1
N = 29
N = 30
Protein Levels
Ribosome Occupancy
RNA Expression
Genotype

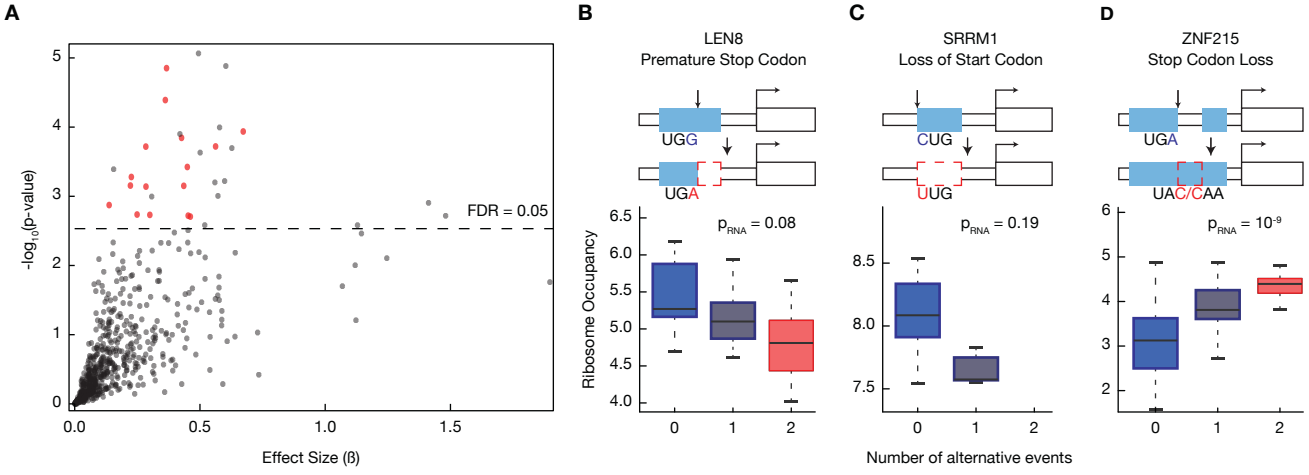

A
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
Gene N
Protein level
TE
RNA exp.
Ribosome occupancy
1.0
0.8
0.6
0.4
0.2
0.0
Train Self Organizing Map (SOM)
Gene A
Gene B
Gene C
Gene D
Gene X
Gene Y
Gene Z
B
Ribosome Occupancy
RNA Expression
Translation Efficiency
Protein Level
1.0
0.8
0.6
0.4
0.2
0.0
C
Cluster 1 (n= 248)
Cluster 2 (n= 353)
Cluster 3 (n= 145)
Cluster 4 (n= 449)
Cluster 5 (n= 573)
Cluster 6 (n= 292)
Cluster 7 (n= 454)
Cluster 8 (n= 415)
Cluster 9 (n= 350)
R.O.
R.E.
T.E.
P.L.
0.8
0.2
Percentile
D
lymphocyte differentiation
translational initiation
protein targeting to ER
translational elongation
translational termination
cytosolic small ribosomal subunit
cytosolic large ribosomal subunit
catalytic step 2 spliceosome
regulation of cell cycle arrest
RNA helicase activity
tricarboxylic acid cycle
G1 DNA damage checkpoint
proteasome complex
glutamine metabolic process
Cluster 8 enriched
Cluster 5 enriched
0
2
4
6
8
10
12
log2 Odds Ratio

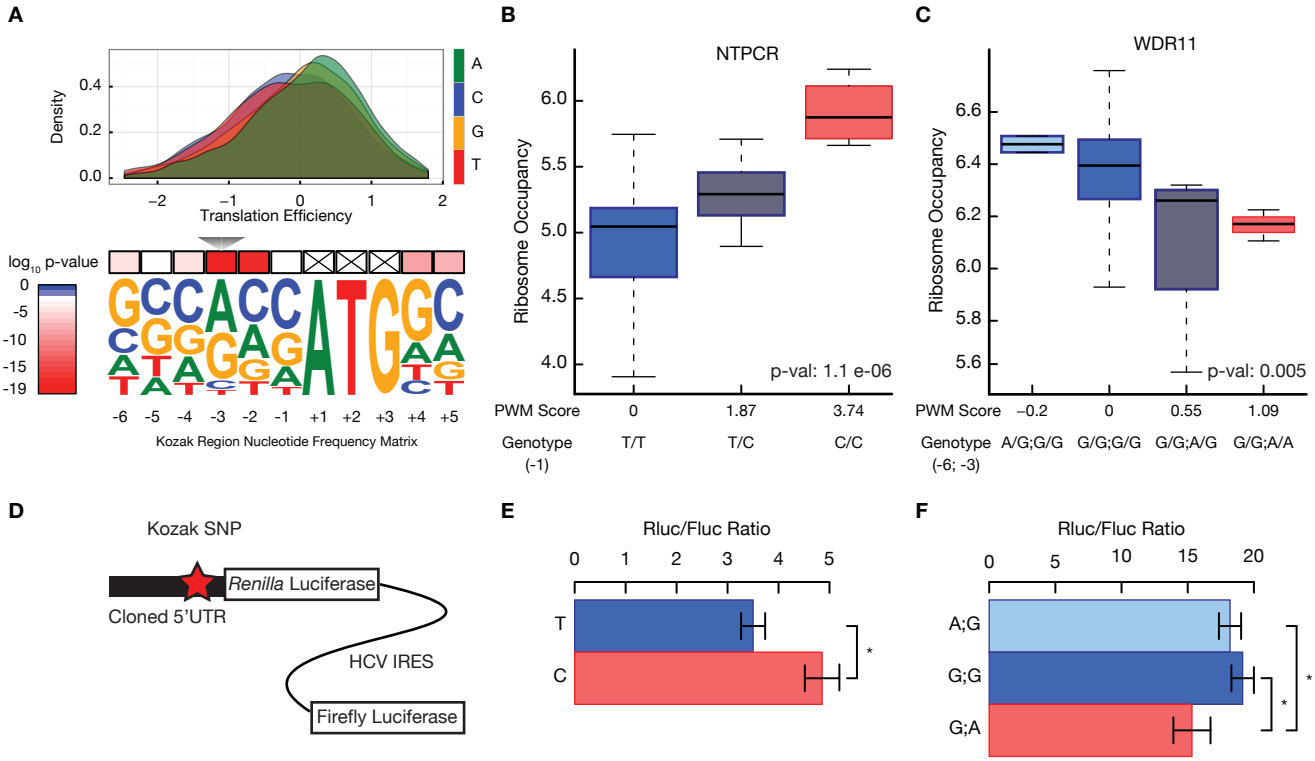

A
Number of Genes
8000
4000
0
Variable
–
+
–
+
RNA Expression
Ribosome Occupancy


B
Variable RNA expression between individuals
Variable ribosome occupancy between individuals
1954
627
72


D
RNA Expression–Protein Level
Density
0.15
0.10
0.05
0
–1
–0.5
0
0.5
1
Spearman Correlation Coefficient


C
Adjusted p-value
0.05
<0.001
Odds Ratio
128.8
10.2
3
extracellular matrix part
blood circulation
cytokine secretion
antigen binding
humoral immune response
actomyosin structure organization
protein tyrosine kinase activity
B cell mediated immunity
positive regulation of PI3K cascade
regulation of immune response
leukocyte migration
G-protein coupled receptor activity
cell motility
chemokine receptor activity
exocrine system development
signaling receptor activity
cytokine binding
tight junction
apical part of cell


A
-log10(p-value)
FDR = 0.05
Effect Size (ß)
B
LEN8
Premature Stop Codon
UGG
UGA
Ribosome Occupancy
p_RNA = 0.08
C
SRRM1
Loss of Start Codon
CUG
UUG
p_RNA = 0.19
Number of alternative events
D
ZNF215
Stop Codon Loss
UGA
UAC/CAA
p_RNA = 10-9


A
Density
Translation Efficiency
A
C
G
T
log10 p-value
Kozak Region Nucleotide Frequency Matrix
B
NTPCR
Ribosome Occupancy
p-val: 1.1 e-06
PWM Score
Genotype
(-1)
T/T
T/C
C/C
C
WDR11
Ribosome Occupancy
p-val: 0.005
PWM Score
Genotype
(-6; -3)
A/G;G/G
G/G;G/G
G/G;A/G
G/G;A/A
D
Kozak SNP
Renilla Luciferase
Cloned 5'UTR
HCV IRES
Firefly Luciferase
E
Rluc/Fluc Ratio
T
C
F
Rluc/Fluc Ratio
A;G
G;G
G;A

Integrative analysis of RNA, translation and protein levels reveals distinct regulatory variation across humans

Can Cenik, Elif Sarinay Cenik, Gun W Byeon, et al.

Genome Res. published online August 21, 2015
Access the most recent version at doi:10.1101/gr.193342.115