PNAS RESEARCH ARTICLE | PLANT BIOLOGY OPEN ACCESS

Variation in upstream open reading frames contributes to allelic diversity in maize protein abundance

Joseph L. Gage[a,b,1], Sujina Mali[c], Fionn McLoughlin[c], Merritt Khaipho-Burch[d], Brandon Monier[a], Julia Bailey-Serres[e], Richard D. Vierstra[c], and Edward S. Buckler[a,d,f,1]

Contributed by Edward S. Buckler; received July 14, 2021; accepted February 22, 2022; reviewed by Jian Lu, Magnus Nordborg, and Nicky Whiffin

The 5′ untranslated region (UTR) sequence of eukaryotic mRNAs may contain upstream open reading frames (uORFs), which can regulate translation of the main ORF (mORF). The current model of translational regulation by uORFs posits that when a ribosome scans a mRNA and encounters an uORF, translation of that uORF can prevent ribosomes from reaching the mORF and cause decreased mORF translation. In this study, we first observed that rare variants in the 5′ UTR dysregulate maize (*Zea mays L.*) protein abundance. Upon further investigation, we found that rare variants near the start codon of uORFs can repress or derepress mORF translation, causing allelic changes in protein abundance. This finding holds for common variants as well, and common variants that modify uORF start codons also contribute disproportionately to metabolic and whole-plant phenotypes, suggesting that translational regulation by uORFs serves an adaptive function. These results provide evidence for the mechanisms by which natural sequence variation modulates gene expression, and ultimately, phenotype.

uORF | proteome | rare alleles | maize | gene expression

Significance

Proteins are the machinery which execute essential cellular functions. However, measuring their abundance within an organism can be difficult and resource-intensive. Cells use a variety of mechanisms to control protein synthesis from mRNA, including short open reading frames (uORFs) that lie upstream of the main coding sequence. Ribosomes can preferentially translate uORFs instead of the main coding sequence, leading to reduced translation of the main protein. In this study, we show that uORF sequence variation between individuals can lead to different rates of protein translation and thus variable protein abundances. We also demonstrate that natural variation in uORFs occurs frequently and can be linked to whole-plant phenotypes, indicating that uORF sequence variation likely contributes to plant adaptation.

Understanding the mechanisms by which genetic variation produces a phenotype will depend on learning how gene expression is regulated differently between alleles. Although moderate correlations between messenger RNA (mRNA) and protein have been observed across genes (r = 0.3 to 0.8; reviewed in ref. 1), the correlation across individuals tends to be quite low ($r < 0.25$ in maize [*Zea mays* L.] and humans) (2, 3). Additionally, there is only modest overlap between variants associated with mRNA levels and variants associated with protein levels in maize, mice, and humans (3–6). Together, these findings suggest an extensive role of posttranscriptional regulation in determining gene expression across eukaryotes. Protein synthesis is energetically expensive, which implies high selective pressure on individuals to produce the appropriate quantity of each protein (7, 8). Despite this strong selective pressure, protein levels are heritable and vary between individuals (6, 9), which means that there must be adaptive advantages to allelic differences in protein abundance. Motivated by these observations, an outstanding question is, How does genetic variation contribute to variation in protein abundance?

Secondary structure and high GC content of the 5′ untranslated region (UTR) can affect translation efficiency, presumably by the formation of secondary structure that obstructs scanning of the 43S preinitiation complex and completion of initiation (10–13), a mechanism which can regulate gene expression response to environmental conditions (14). However, single base pair substitutions (the focus of this study) do not cause appreciable changes in secondary structure or GC content of the 5′ UTR. Upstream open reading frames (uORFs) located in the 5′ UTR and preceding the main ORF (mORF) also contribute to posttranscriptional gene regulation, generally by reducing translation of the mORF (11, 13, 15). Studies in humans have shown that variants which create or disrupt uORFs are under negative selection and associated with disease phenotypes (16, 17). Across eukaryotes, the strength of the uORF Kozak sequence is predictive of translation initiation efficiency (18). While the effects of uORF translation on protein abundance have been demonstrated in several genes in plants and humans (15, 16, 19), across genes in *Arabidopsis thaliana* (20), and across alleles in massively parallel reporter assays (11, 13), less is known about the genome-wide effects that natural sequence variation in uORFs has on allelic protein abundance in crop plants.

Previous studies of genetic variability for protein abundance between individuals (3–5, 21, 22) have primarily used genome-wide association studies (GWAS), an approach that is most effective for alleles that are at high frequency in the population being studied. To be sufficiently powerful, GWAS require large population sizes and are still generally best suited to identifying common alleles, which tend to have smaller effect sizes but can be important for locally adapted phenotypes. In contrast, rare variants tend to have larger, deleterious effects (23–25) but are difficult to identify by GWAS.

Author contributions: J.L.G., J.B.-S., R.D.V., and E.S.B. designed research; J.L.G., S.M., F.M., M.K.-B., and B.M. performed research; J.L.G, S.M., F.M., M.K.-B., and B.M. analyzed data; and J.L.G., S.M., B.M., and E.S.B. wrote the paper.

Reviewers: J.L., Peking University; M.N., Gregor Mendel Institute; and N.W., Wellcome Trust Centre for Human Genetics.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jlgage@ncsu.edu or esb33@cornell.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2112516119/-/DCSupplemental.

Published March 29, 2022.

Because rare alleles 1) are more likely to have large effects and 2) have been implicated in dysregulation of mRNA abundance (26, 27), we reasoned that, in aggregate, rare alleles would be ideal candidates for identifying specific sets of variants that contribute most to posttranscriptional regulation. Our findings for rare alleles could then be validated in the pool of common alleles.

In this study, we answer the preceding questions with maize, which has large homogeneous tissues that facilitate high-quality proteome extraction, extremely high levels of phenotypic and genetic diversity (28, 29), and rapid linkage disequilibrium decay that allows high genetic resolution with fewer individuals (30, 31). We use two distinct sets of diverse inbred maize lines: a technically and biologically replicated set of four inbreds and a mostly unreplicated set of 95 inbreds previously described (3). We first use rare alleles, which generally have larger and more disruptive effects, to learn that variants in the 5′ UTR, specifically those which modify uORFs, have the most influence on protein abundance. We then use those findings to identify and test common variants which likely have weaker but adaptive effects, and demonstrate their effects on protein abundance as well as their contribution to metabolic and whole-plant phenotypes.

Results and Discussion

Regulation of gene expression is largely genetically determined; mRNA abundance is heritable in organisms across kingdoms (32), as are proteomic (6, 9) and metabolite levels (33). Quantification of 7,524 peptides in a pair of leaves in developmentally matched juvenile plants (four replicated maize inbreds; Dataset S1) confirmed that protein abundance is heritable in maize, with median heritability of 0.79, similar to mRNA and metabolite abundances (33) (*SI Appendix*, Figs. S1 and S2). These high heritabilities not only demonstrate low measurement error but also reveal extant adaptive genetic variation in protein abundances.

We identified rare variants based on a minor allele frequency (MAF) < 0.02 in the maize HapMap3.2.1 population, which contains >1,200 varieties of maize and its wild relatives from around the world (34). Rare variants were classified based on their location within five genic features: promoter, 5′ UTR, coding sequence (CDS), intron, or 3′ UTR (Fig. 1*B*). In each genic feature, we tested for differing protein abundance among 95 diverse inbred lines (Dataset S2), classified at each single-nucleotide polymorphism (SNP) by whether they have the common or rare allele. Introns were predicted to show little effect on protein abundance, due to their typical absence from the mRNA; we observe a corresponding lack of association between intronic rare alleles and protein abundance (Fig. 1*A*). Rare alleles in the promoter and 5′ UTR were associated with more variable protein abundance, with rare alleles in the 5′ UTR showing the strongest effect (Fig. 1 *A* and *C* and *SI Appendix*, Fig. S3). This finding reinforces previous implication of the 5′ UTR in posttranscriptional regulation (11–13, 16–18, 35, 36), and the lack of any significant effect from variants in the 3′ UTR contrasts with recent work focused on the role of the 3′ UTR in translational regulation in maize (37).

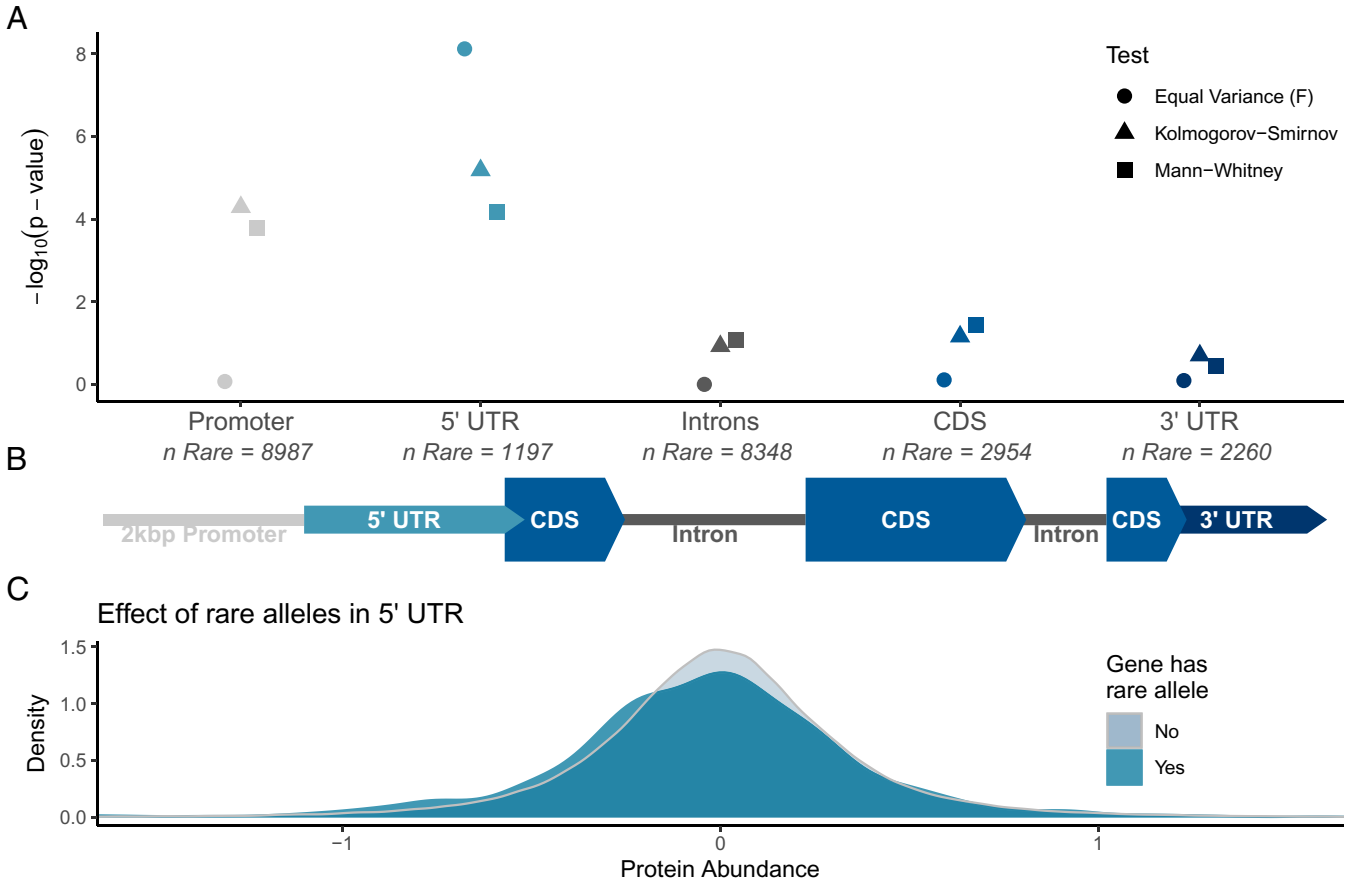

Fig. 1. Individuals with rare alleles have altered protein abundance. (*A*) In each genic region shown in *B* (generic gene), individuals with at least one rare allele (MAF < 0.02) were compared to individuals with no rare alleles in any of the genic regions. The two-sided Kolmogorov–Smirnov and Mann–Whitney tests test for differences in the distribution of protein abundance between individuals with and without rare alleles; the one-sided *F* test for equality of variance tests for greater variability among individuals with rare alleles. Rare alleles in the 5′ UTR are significantly associated with dysregulated protein abundance by all three tests. Number of observations without rare allele = 184,668. (*C*) Distribution of protein abundance (log2 ratio against B73) for individuals with rare alleles in the 5′ UTR (blue), compared to individuals without rare alleles (light blue).

The 5′ UTR can contribute to posttranscriptional regulation of gene expression, and machine learning models have been successful at predicting protein abundance from sequence. However, application of previously developed machine learning models (11, 38) proved unable to predict allelic differences in protein abundance from 5′ UTR and transcript sequence in four maize inbreds (full results are available in Bitbucket repository https://bitbucket.org/bucklerlab/p_protein_diverse_maize/src/master/src/1_four_inbreds/). Instead, we turned to evaluating the mechanisms by which these observed rare variants might be disrupting translation.

We hypothesized that rare alleles can disrupt existing uORF start codons or create start codons that generate novel uORFs, thereby decreasing or increasing (respectively) uORF translation and ultimately altering translation of the mORF (Fig. 2 *A* and *B*). Based on this hypothesis, and the inverse relationship between uORFs and mORF translation efficiency (*SI Appendix*, Fig. S4), we predicted that 1) rare alleles that weaken existing uORF start codons will be associated with greater mORF protein abundance (Fig. 2*A*) and 2) rare alleles that cause new or strengthen existing uORF start codons will be associated with lower mORF protein abundance (Fig. 2*B*).

To test the first prediction, we identified uORFs that show evidence of translation based on ribosome profiling data (39), then searched for rare alleles within or near the start codons of those uORFs. A score was assigned to the common and rare alleles based on similarity of their surrounding sequence to the maize Kozak sequence (*SI Appendix*, Fig. S5) (40), and variants with rare alleles that weaken the Kozak sequence were labeled as "derepressive variants," based on the hypothesized effect of the rare allele on mORF protein abundance (Dataset S3). Derepressive variants were compared to rare alleles anywhere else in the 5′ UTRs of genes with translated uORFs, including within uORFs themselves but excluding the start codons of translated uORFs. This comparison revealed that derepressive variants are associated with 26% greater mORF protein abundance than rare alleles elsewhere in the 5′ UTR of genes with translated uORFs (Fig. 2*C*; P = 1.4e-02, two-sided Mann–Whitney test). We also studied the difference in protein abundance between individuals with the common and rare derepressive alleles. In 9 out of 14 cases, the rare derepressive alleles were associated with increased mORF protein abundance relative to the common allele (Fig. 2*D*).

To test the second prediction, we performed analyses similar to those described above but focused on genes with no annotated uORFs, or with annotated uORFs that did not show any evidence of translation (39). We searched for rare alleles that increased their surrounding sequence's similarity to the maize Kozak sequence (40), and labeled them as "repressive variants," based on their hypothesized effect on mORF protein abundance (Dataset S3). We compared repressive variants to rare alleles which are located in genes with no annotated or translated uORFs but decrease the surrounding sequence's similarity to the Kozak motif. Repressive variants are associated with a 45% decrease in protein abundance of the corresponding mORF (Fig. 2*E*; P = 3.3e-03, two-sided Mann–Whitney test). Comparing individuals with rare alleles at repressive variants to individuals with common alleles revealed that 15 out of 25 rare alleles were associated with decreased mORF protein abundance (Fig. 2*F*). For both repressive and derepressive variants, we observed at least one variant with large effects in the direction opposite to what we anticipated. These provide evidence that, although our annotations seem to be identifying SNPs with the predicted effects, other factors also influence final protein abundance.

Because variants in the 5′ UTR can also influence mRNA abundance by affecting mRNA stability (41) and transcription initiation (42, 43), we performed the same tests described above, but with mRNA abundance instead of protein abundance (Dataset S4), to see whether repressive and derepressive variants are impacting protein abundance through mRNA levels. We found no evidence that repressive or derepressive variants impact mRNA abundance, which implies that their hypothesized effects are taking place during translation (*SI Appendix*, Figs. S6–S8).

Given that rare, putatively deleterious variants appear to dysregulate mORF protein abundance by altering start codons of uORFs, we wondered whether historical mutations with similar mechanisms may have conferred adaptive advantage and risen to a higher allele frequency via positive selection. We identified common (MAF > 0.1) repressive and derepressive variants using the same criteria described above (Dataset S5), and tested their association with mORF protein abundance. We found an enrichment of significant associations between derepressive variants and mORF protein abundance (Fig. 3*A*; one-sided Mann–Whitney test).

We performed an equivalent test with mRNA abundance to check whether derepressive variants are influencing protein abundance indirectly by altering mRNA levels, and found that, while some derepressive variants are associated with altered mRNA levels, they are generally not the same variants as are associated with altered protein levels (*SI Appendix*, Figs. S9 and S10). The derepressive variants that are associated with changes in mRNA abundance could be acting directly on mRNA stability or transcription factor binding sites (41–43), or they could be in linkage disequilibrium with other expression quantitative trait loci, for example, in promoter regions (26).

We reasoned that, if these common alleles are adaptive, they may also affect metabolic or physical phenotypes. Indeed, derepressive variants show greater than 18% enrichment for GWAS hits over all common variants in 5′ UTRs, with >80% of derepressive variants associated with at least one phenotype. Derepressive variants being enriched for GWAS hits may explain previous observations that the 5′ UTR has an outsized contribution to quantitative traits in maize (44). The enrichment of GWAS hits for metabolic phenotypes is particularly notable because uORFs have been implicated in metabolite-based regulation involving the translating ribosome (45, 46). On the other hand, repressive variants show a statistically insignificant 2% enrichment (Fig. 3*B*; two-sided binomial test).

Derepressive variants show stronger association with protein abundance and greater enrichment for GWAS hits than repressive variants. Although we are studying common SNPs, which we assume have not been under negative selection, it is possible that reduced translation of mORFs is under strong negative selection (47) and that more repressive variants than derepressive variants are false positives. While derepressive variants disrupt the Kozak sequence of uORFs that already show evidence of translation, repressive variants strengthen the Kozak sequence but have not been shown to result in translation of the putative uORFs they create. SNPs may be more likely to disrupt existing uORFs than to create new translated uORFs (48).

None of the derepressive or repressive variants we identified overlap with genes that have conserved peptide uORFs (CPuORFs) across angiosperms (49). This is possibly because the strong conservation of CPuORFs means any new mutations are subject to strong negative selection. Studies identifying CPuORFs have also been largely focused on dicots. However, many of the uORFs containing derepressive and repressive variants that we identified are conserved between members of the Andropogoneae tribe (50), although they are not significantly more or less conserved than all other uORFs (*SI Appendix*, Fig. S11).

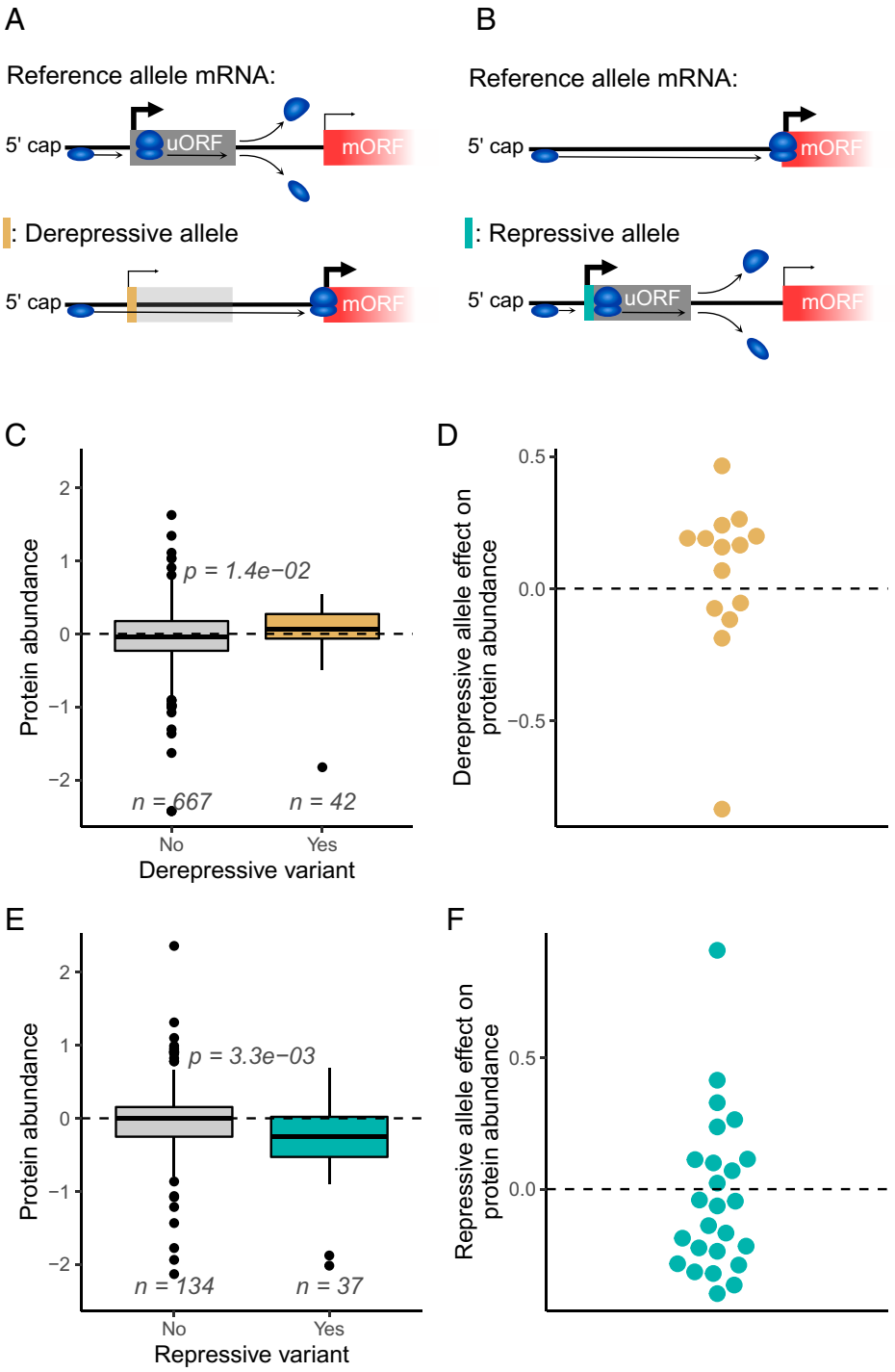

Fig. 2. (*A*) Rare SNP alleles which disrupt or weaken the start codon of an existing translated uORF are associated with derepression of the mORF, whereas (*B*) rare alleles which cause a new start codon or strengthen an existing start codon in the 5′ UTR are associated with repression of the mORF. *C* and *E* show the effects of derepressive and repressive rare alleles, compared to the effects of other rare alleles in similar contexts (two-sided Mann–Whitney test; box plots show median and first and third quartiles, and whiskers extend no farther than 1.5 times the interquartile range; outliers are displayed as points). *D* and *F* show the effect that derepressive and repressive alleles have on protein abundance, on a per-gene basis. Individuals with derepressive alleles often show an increase in protein abundance over individuals with the common allele. Individuals with repressive alleles often show a decrease in protein abundance. Each point represents, for a single gene, the difference in median protein abundance of between individuals with the rare (derepressive or repressive) allele and individuals with the common allele. Note that the differences between individuals with rare alleles and individuals with common alleles are often of low confidence due to the fact that the rare allele group frequently contains only one or two observations. In *C*–*F*, protein abundance is represented as a log2 ratio against B73.

All results up to this point have been based on genes that have protein abundance data. We expanded our scope to all annotated genes in B73 and identified variants that may have repressive or derepressive function (Dataset S6). Genes with derepressive variants were enriched for biological process gene ontology (GO) terms related to metabolic and biosynthetic processes (Dataset S7), consistent with evidence that uORFs contribute to regulation of metabolic pathways in plants (51, 52). We identified a potential derepressive variant in the 5′ UTR of an adenosylmethionine decarboxylase, Zm00001eb184470,

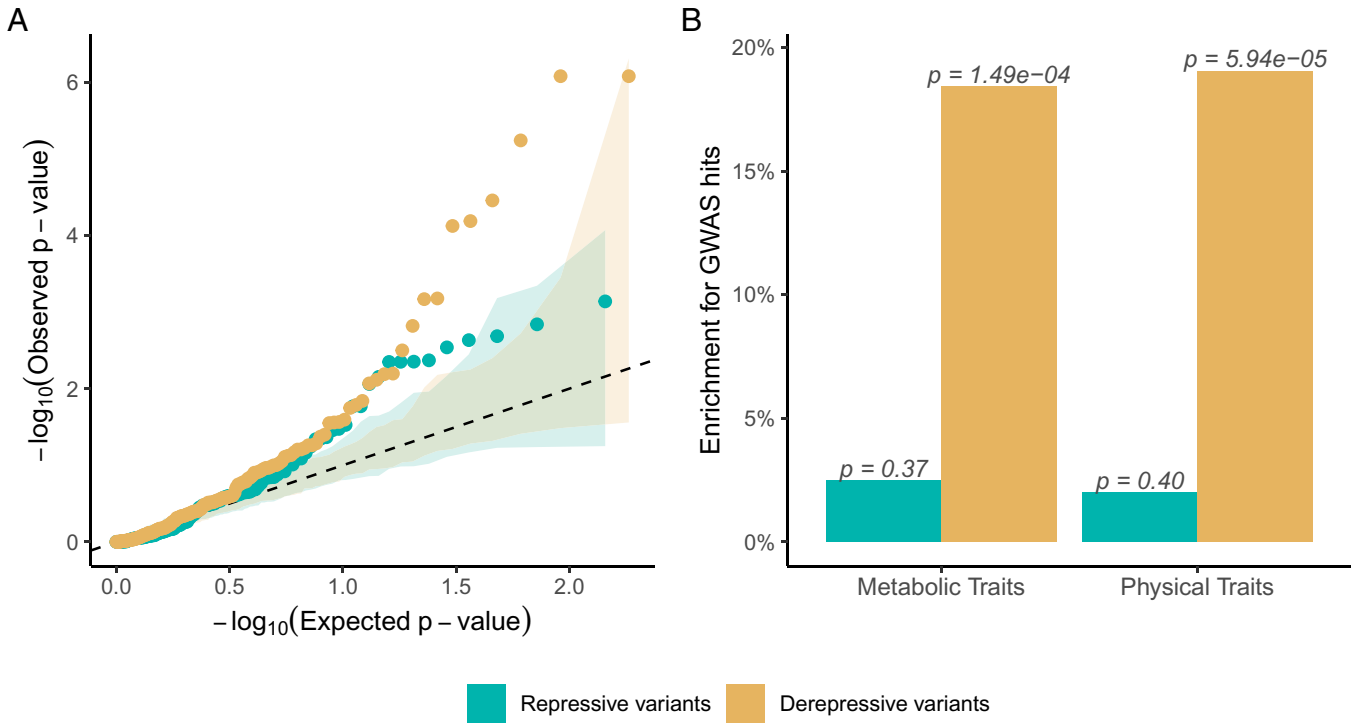

Fig. 3. (*A*) Common variants (MAF > 0.1) show significant association with mORF protein abundance based on one-sided *t* test for increased (derepressive variants; *n* = 183) or decreased (repressive variants; *n* = 144) mORF protein abundance. A random effect to control for kinship was included in the model, and SNPs with less than three individuals having either allele were excluded. Shaded areas show distribution of *P* values over 100 permutation tests; dashed line marks a one-to-one relationship. (*B*) Common derepressive variants show a significantly increased number of GWAS hits relative to all common SNPs in 5′ UTRs (two-sided binomial test; derepressive *n* = 191, repressive *n* = 153).

the ortholog of which has been shown to be regulated by uORFs in *A. thaliana* (53).

Beyond plants, it may be the case that certain protein families are regulated by uORFs across eukaryotes. We found derepressive variants in the uORFs of Zm00001eb252410 and Zm00001eb136490, which both encode proteins from the heat shock protein 70 family, a member of which has been shown to be regulated by uORFs in humans (54). Another derepressive variant is in the uORF of Zm00001eb070400, which is a multidrug resistance–associated protein (ABC transporter C family member 2), also shown to be regulated by uORFs in humans (55).

These results demonstrate that rare alleles in the 5′ UTR have dysregulatory effects on the proteome. Investigating that pattern further, we have also shown that variation in uORF start codons contributes to allelic differences in mORF protein abundance. Not only do these effects show up at the protein abundance level, but variants in uORFs are enriched for GWAS hits of metabolic and physiological traits. These findings could be used to prioritize derepressive variants, particularly ones at common allele frequencies, in genomic prediction models. They are also a step toward building models of allelic variation between individuals, based on mechanistic effects of sequence variation. Although rare alleles are infrequent within a given site, hundreds of rare variants in each individual contribute to genetic burden, which can be better modeled by categorizing the effects of particular rare alleles. Because these effects relate to fundamental aspects of the translational machinery, we anticipate that these findings will extend beyond maize and be applicable in most eukaryotes. These findings can be used to engineer or predict variation in protein regulation and can potentially be applied to address problems in synthetic biology, genome editing, crop improvement, and human disease.

Materials and Methods

Proteomic Datasets. Two distinct proteomic datasets were used in this study. The first consisted of four inbred lines (B73, Mo17, CML103, and P39) for which genome assemblies are publicly available (56). These were used for analyses in which accurate genomic sequence or biological replication was needed. The second dataset was generated by Jiang et al. (3), which consists of 98 diverse inbred lines. Protein abundance data for these lines were determined by using B73 as the reference for aligning peptides to the proteome. These data were used for analyses requiring a larger sample size and SNP sets called against a common reference genome.

Replicated Proteomics on Four Diverse Inbreds. The plants were grown using 3:1 Metromix 900 (SunGro)/Turface MVP (Profile Products) in the greenhouse under a 16-h light/8-h dark photoperiod and 27 °C/22 °C day/night temperatures. Third and fourth leaves of 2-wk-old plants were collected separately, flash frozen, and stored at −80 °C. Sample preparation and mass spectrometry analysis were performed as described by McLoughlin et al. (57). Each genotype was analyzed with five biological and three technical replicates. Raw mass spectrometry data were analyzed against the B73 v5 proteome (58) with Proteome Discoverer (version 2.0.0.802; Thermo Fisher Scientific). Peptides were assigned by SEQUEST HT, allowing one missed tryptic cleavage, a precursor mass tolerance of 10 ppm, and a fragment mass tolerance of 0.02 Da. Carbamidomethylation of cysteines and oxidation of methionine were used as static and dynamic modifications, respectively. Only peptides with false discovery rates of 0.01 (high confidence) were used for data analysis.

Heritability Calculations. We fit each peptide with the model $y_{ij} = g_i + e_{ij}$, where y_{ij} is the abundance of the peptide in replicate j of inbred i, g_i is the genetic effect of inbred i, and e_{ij} is the residual for replicate j of inbred i. Both terms g and e were random effects, with variances σ_g^2 and σ_e^2, respectively. Heritability was calculated as $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$.

Protein Abundance Prediction. We attempted to predict peptide abundance, using several sequence-based models that predict protein or mRNA abundance.

The first model we tested was described by Washburn et al. (38), who developed it to predict which of two orthologous genes in *Z. mays* and *Sorghum bicolor* had higher mRNA expression levels. We used the same inputs, namely, 1,500 base pairs (bp) of DNA sequence from each of the promoter and terminator regions of each gene, and attempted to predict which of two inbred lines had higher peptide abundance.

The second model we tested was published by Cuperus et al. (11), who trained their model on a library of random *Saccharomyces cerevisiae* 5′ UTRs and corresponding protein abundances. We used their pretrained model to try to predict peptide abundance in maize. Because the peptide abundances in our dataset cannot be compared between genes, we instead compared the log2 ratio of observed peptide abundances to the log2 ratio of predicted peptide abundances for each pairwise comparison between inbreds at each peptide in our dataset.

The third model we tested was based on the observation that codon bias can be predictive of protein abundance (59–61). We trained a multiple linear regression model to predict the log ratio of peptide abundance between two inbreds, using, as explanatory variables, the differences in codon counts between alleles of genes encoding the same peptide.

Uplifting Jiang Proteomic Data. The proteomic data generated by Jiang et al. (3) was uplifted to version 5 of the B73 proteome (58). The Uniprot IDs in the published data were used to obtain protein sequences, which were searched in v5 using blastp (62). Proteins with a match in the v5 proteome that had >90% identity and >90% coverage (2,523 out of 2,750) were kept. Similar to the four-inbred experiment described above, the protein abundances reported by Jiang et al. were normalized by calculating the log2 ratio against B73. For example, if, for a given gene, inbred X had a protein abundance of 20 and B73 had a protein abundance of 10, the transformed protein abundance for inbred X would be represented as $\log_2(20/10) = 1$. This log ratio transformation was done to enable comparisons between genes, which otherwise cannot be directly compared, due to the nature of the proteomic data.

Calling HapMap SNPs on Jiang Inbreds. The maize HapMap3.2.1 (34) SNP data were uplifted to B73 v5 coordinates using CrossMap (63). Whole-genome sequencing was available for 95 of the 98 maize inbreds (64), and was used to call SNPs at the same v5 positions and using the same methods as HapMap3.2.1 (34, 65).

Effects of Rare Variants on Protein Abundance. To study the relationship between rare, putatively deleterious variants and protein abundance, we classified SNPs as rare if they had a MAF of <0.02 in the maize HapMap 3.2.1 panel, which contains over 1,200 diverse maize varieties and wild relatives, and represents an independent dataset from the Jiang inbred lines for defining MAF (34).

Variants were categorized based on overlap with five annotated features: promoters (2,000 bp upstream of the transcription start site), 5′ UTRs, 3′ UTRs, CDS, and introns. SNPs were categorized using the GenomicFeatures (66) package in R (67), based on the Zm00001eb.1 annotation of B73 (58). Each combination of gene, feature, and inbred was given a binary classification for whether or not it contained a rare (MAF < 0.02) SNP allele. Within each feature, the relative protein abundance of all gene–inbred combinations that did contain a rare allele was compared to a null consisting of protein abundances in gene–inbred combinations that did not have rare variants in any of the five features. The number of individuals per gene in the null group ranged from 14 to 95. Distributions of protein abundance between the rare variant and null categories were statistically compared by the Kolmogorov–Smirnov test (68, 69), Mann-Whitney test, and F-test for equality of variance.

Identifying Translated uORFs. Two biological samples of riboseq data generated by Lei et al. (39) on nonstressed B73 seedlings were obtained from National Center for Biotechnology Information Sequence Read Archive accessions SRX845439 and SRX845455. Riboseq data consist of sequenced fragments of mRNA which are bound by ribosomes and therefore assumed to be undergoing translation. Cutadapt (70) version 1.18 was run using the parameters "-a CTGTAGGCACCATCAAT -m 20" to remove adapters and discard any reads shorter than 20 bp. Reads were mapped against version 5 of the B73 genome using hisat2 (71) version 2.2.1 with default parameters except for "–trim5 1" to remove the first base pair from each read, which the original authors describe as frequently representing an untemplated addition during reverse transcription (39). The alignments were sorted and indexed with samtools (72) version 1.11, and reads that mapped to more than one location in the genome were discarded.

The uORFs were computationally identified using the R (67) package ORFik (73). The uORFs were identified using the pattern (ATG|TTG|CTG-3n-TAA|TAG|TGA), since translated uORFs can initiate on noncanonical start codons (74–76). Any uORFs that overlapped with annotated CDSs were discarded, and the remaining uORFs were used as targets for read counting. Reads overlapping computationally identified uORFs were counted by htseq (77) version 0.11.3 using htseq-count with the argument "–nonunique all" so that reads were not discarded if they mapped to multiple overlapping uORFs or uORFs on different annotated transcripts. Read positions were shifted to reflect the location of the ribosome P site by using the functions detectRibosomeShifts() and shiftFootprints() in ORFik (73). Only reads with lengths between 26 and 34 bp (89% of reads) were used.

The log(fragments per kilobase million [FPKM] + 1) values for the two biological samples were correlated, $r = 0.99$ for CDS, $r = 0.78$ for uORFs. The lower correlation for uORFs was primarily due to uORFs with reads in one sample but not the other; discounting those, the correlation of log(FPKM) between samples was $r = 0.87$. These correlations were high enough that read counts from both samples were pooled, and FPKM was calculated on the pooled counts.

Translated uORFs were defined as computationally identified uORFs that had FPKM > 20, FPKM < 1,759, total length > 15 bp, and total length < 333 bp. These cutoffs were chosen based on the fifth and 95th percentiles of the distributions of FPKM and length across all uORFs, in order to exclude uORFs that may be in misannotated 5′ UTRs or have an unusually high number of reads mapping to them. We were relatively lenient with calling translated uORFs because, as described below, we were more interested in identifying uORFs that may be translated than in obtaining a highly accurate quantification of translation.

Identifying Variants That Strengthen or Weaken uORF Start Sequence. An empirical maize Kozak sequence was determined by creating a position weight matrix of the sequence spanning –3 to +4 nucleotides relative to the translation start site of all annotated maize gene model CDSs. The range of –3 to +4 was chosen to reflect the portion of the larger Kozak sequence (–6 to +4) that is highly conserved in green plants (78). Variants that weaken the Kozak of existing uORFs were identified by searching for variants that fell within the –3 to +4 range around the start of uORFs that show evidence of translation (described above). Two versions of the 7-bp sequence around the start codon were created, one with each SNP allele, and they were scored for their similarity to the Kozak sequence using the R package Biostrings (79). Variants that decreased the similarity to the Kozak sequence by 0.5 or more were classified as "derepressive."

Identification of "repressive" variants was performed similarly, but with rare variants that were in the 5′ UTR of genes with no annotated uORFs, or of genes with all annotated uORFs having FPKM < 5. Because an annotated uORF was not present in all instances, unlike in the preceding paragraph, it was not obvious where to position the variants within the Kozak sequence. We calculated the similarity score against the Kozak with the variant site in all positions from –3 to +4, and chose the position with the highest score, reasoning that it represented the context that was closest to making a uORF start. We then substituted the alternate allele at the SNP site and compared the Kozak score, classifying variants that increased the score by >0.25 as "repressive."

Identifying and Testing Adaptive Variants. Repressive and derepressive variants with MAF > 0.1 were identified by the same criteria as described above for rare alleles. Of 115,836 common variants in the 5′ UTR of all genes, 9,117 were in the 5′ UTR of genes with protein abundance information. Of those, we identified 191 derepressive and 153 repressive variants. One-sided *t* tests were performed to test for increase or decrease of protein abundance associated with derepressive or repressive alleles, respectively. Models included a random effect to account for kinship between individuals, using the R package sommer (80). The results from these tests were compared to *P* values from shuffling the genotypes and performing the same test 100 times for each variant.

GWAS Hit Enrichment. To determine the number of physiological and metabolic traits associated with SNPs within 5′ UTR regions, a graph database was used to store and query GWAS results with additional biological information

in maize. Neo4j (v4.2.0) and Cypher (v4.2.0) were used as the graph database management system and querying language, respectively (81). Results were obtained by interfacing with the database using the Neo4j diver, neo4r (67, 82). In total, 3,874 metabolite traits collected from the Goodman diversity panel (83) and 333 physiological traits collected from both the Goodman diversity and Nested Association Mapping panels (84, 85) were used for association testing and SNP–trait relation queries. Significant counts were obtained by filtering associations between SNP and trait with P value $<$ 10e-5. Enrichment for GWAS hits among repressive and derepressive variants was performed using a one-sided binomial test of the alternative hypothesis that the proportion of repressive or derepressive GWAS hits was greater than the proportion of all 5′ UTR GWAS hits.

Conservation among Andropogoneae. Alignments between maize and five other grasses from the Andropogoneae tribe (50) were filtered to regions that overlap the annotated B73 5′ UTRs. The uORFs, identified above, were classified based on whether at least one other species had alignments covering 90% or more of the uORF. The uORFs with rare or common repressive or derepressive variants were tested against all uORFs by two-sided χ^2 test to see whether repressive and derepressive variants are associated with greater or lesser conservation among the Andropogoneae.

GO Enrichment. All repressive or derepressive variants and their cognate genes were identified genome-wide using the criteria previously described. GO enrichment was performed using the R package topGO (86). Fisher's test was used to compare Biological Process GO categories for genes with repressive or derepressive variants against all genes, using the GO annotations available at https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0_Zm00001eb.1.interproscan.tsv.gz.

Data Availability. SNP data can be found at http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Gage_uORF_allelic_variation_2021. The mass spectrometry proteomics data for four maize inbreds have been deposited to the ProteomeXchange Consortium via the PRIDE (79) partner repository with the dataset identifier PXD026378 (https://www.ebi.ac.uk/pride/archive/projects/PXD026378). All other data can be found in *SI Appendix*.

Code Availability. Code for all analyses can be found at https://bitbucket.org/bucklerlab/p_protein_diverse_maize/.

ACKNOWLEDGMENTS. This material is based upon work supported by the NSF Postdoctoral Research Fellowship in Biology under Grant IOS-1906619, NSF Grant IOS-1840687, NSF Grant IOS-1822330, and the US Department of Agriculture Agricultural Research Service.

Author affiliations: [a]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853; [b]Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695; [c]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130; [d]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853; [e]Department of Botany and Plant Sciences, Center for Plant Cell Biology, University of California, Riverside, CA 92521; and [f]Agricultural Research Service, US Department of Agriculture, Ithaca, NY 14853

1. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020).
2. C. Cenik *et al.*, Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25**, 1610–1621 (2015).
3. L. G. Jiang *et al.*, Characterization of proteome variation during modern maize breeding. *Mol. Cell. Proteomics* **18**, 263–276 (2019).
4. A. Battle *et al.*, Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
5. J. M. Chick *et al.*, Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**, 500–505 (2016).
6. L. Wu *et al.*, Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013).
7. J. A. Birchler, R. A. Veitia, Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14746–14753 (2012).
8. M. Lynch, G. K. Marinov, The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15690–15695 (2015).
9. L. Parts *et al.*, Heritability and genetic basis of protein level variation in an outbred population. *Genome Res.* **24**, 1363–1370 (2014).
10. J. R. Babendure, J. L. Babendure, J. H. Ding, R. Y. Tsien, Control of mammalian translation by mRNA structure near caps. *RNA* **12**, 851–861 (2006).
11. J. T. Cuperus *et al.*, Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).
12. G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin, Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
13. P. J. Sample *et al.*, Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2019).
14. B. Y. W. Chung *et al.*, An RNA thermoswitch regulates daytime growth in *Arabidopsis*. *Nat. Plants* **6**, 522–532 (2020).
15. R. Kawaguchi, J. Bailey-Serres, Regulation of translational initiation in plants. *Curr. Opin. Plant Biol.* **5**, 460–465 (2002).
16. D. S. M. Lee *et al.*, Disrupting upstream translation in mRNAs is associated with human disease. *Nat. Commun.* **12**, 1515 (2021).
17. N. Whiffin *et al.*; Genome Aggregation Database Production Team; Genome Aggregation Database Consortium, Characterising the loss-of-function impact of 5′ untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 2523 (2020).
18. H. Zhang *et al.*, Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nat. Commun.* **12**, 1076 (2021).
19. D. A. Halterman, R. P. Wise, Upstream open reading frames of the barley Mla13 powdery mildew resistance gene function co-operatively to down-regulate translation. *Mol. Plant Pathol.* **7**, 167–176 (2006).
20. H. Y. L. Wu, P. Y. Hsu, Actively translated uORFs reduce translation and mRNA stability independent of NMD in Arabidopsis. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.09.16.460672. Accessed 10 November 2021.
21. J. Fu *et al.*, System-wide molecular evidence for phenotypic buffering in *Arabidopsis*. *Nat. Genet.* **41**, 166–167 (2009).
22. A. Ghazalpour *et al.*, Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* **7**, e1001393 (2011).
23. M. Kimura, Rare variant alleles in the light of the neutral theory. *Mol. Biol. Evol.* **1**, 84–93 (1983).
24. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
25. E. Marouli *et al.*; EPIC-InterAct Consortium; CHD Exome+ Consortium; ExomeBP Consortium; T2D-Genes Consortium; GoT2D Genes Consortium; Global Lipids Genetics Consortium; ReproGen Consortium; MAGIC Investigators, Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
26. K. A. G. Kremling *et al.*, Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523 (2018).
27. J. Zhao *et al.*, A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
28. E. S. Buckler, B. S. Gaut, M. D. McMullen, Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* **9**, 172–176 (2006).
29. M. I. Tenaillon *et al.*, Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9161–9166 (2001).
30. M. A. Gore *et al.*, A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
31. D. L. Remington *et al.*, Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11479–11484 (2001).
32. E. E. Schadt *et al.*, Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
33. J. J. B. Keurentjes *et al.*, The genetics of plant metabolism. *Nat. Genet.* **38**, 842–849 (2006).
34. R. Bukowski *et al.*, Construction of the third-generation *Zea mays* haplotype map. *Gigascience* **7**, gix134 (2018).
35. M. Ringnér, M. Krogh, Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLOS Comput. Biol.* **1**, e72 (2005).
36. D. R. Morris, A. P. Geballe, Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20**, 8635–8642 (2000).
37. W Zhu *et al.*, Large-scale translatome profiling annotates functional genome and reveals the key role of genic 3′ untranslated regions in translatomic variation in plants. *Plant Commun.* **2**, 100181 (2021).
38. J. D. Washburn *et al.*, Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5542–5549 (2019).
39. L. Lei *et al.*, Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J.* **84**, 1206–1218 (2015).
40. M. Kozak, The scanning model for translation: An update. *J. Cell Biol.* **108**, 229–241 (1989).
41. X. J. Hua, B. Van de Cotte, M. Van Montagu, N. Verbruggen, The 5′ untranslated region of the At-P5R gene is involved in both transcriptional and post-transcriptional regulation. *Plant J.* **26**, 157–169 (2001).
42. P. Sega, K. Kruszka, Ł. Szewc, Z. Szweykowska-Kulińska, A. Pacak, Identification of transcription factors that bind to the 5′-UTR of the barley PHO2 gene. *Plant Mol. Biol.* **102**, 73–88 (2020).
43. C. P. Yu, J. J. Lin, W. H. Li, Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. *Sci. Rep.* **6**, 25164 (2016).
44. E. Rodgers-Melnick, D. L. Vera, H. W. Bass, E. S. Buckler, Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3177–E3184 (2016).
45. C. Hanfrey, M. Franceschetti, M. J. Mayer, C. Illingworth, A. J. Michael, Abrogation of upstream open reading frame-mediated translational control of a plant S-adenosylmethionine decarboxylase results in polyamine disruption and growth perturbations. *J. Biol. Chem.* **277**, 44131–44139 (2002).
46. S. van der Horst, T. Filipovska, J. Hanson, S. Smeekens, Metabolite control of translation by conserved peptide uORFs: The ribosome as a metabolite multisensor. *Plant Physiol.* **182**, 110–122 (2020).
47. D. G. MacArthur *et al.*; 1000 Genomes Project Consortium, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
48. J. G. Monroe, J. K. McKay, D. Weigel, P. J. Flood, The population genomics of adaptive loss of function. *Heredity* **126**, 383–395 (2021).
49. S. van der Horst, B. Snel, J. Hanson, S. Smeekens, Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid sequences in the 5′ leader of mRNAs in *Arabidopsis thaliana*. *RNA* **25**, 292–304 (2019).

50. B. Song *et al*., *Constrained non-coding sequence provides insights into regulatory elements and loss of gene expression in maize*. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.07.11.192575. Accessed 15 April 2021.
51. J. Carbonell, M. A. Blázquez, Regulatory mechanisms of polyamine biosynthesis in plants. *Genes Genomics* **31**, 107–118 (2009).
52. A. G. von Arnim, Q. Jia, J. N. Vaughn, Regulation of plant translation by upstream open reading frames. *Plant Sci.* **214**, 1–12 (2014).
53. C. Hanfrey *et al*., A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. *J. Biol. Chem.* **280**, 39229–39237 (2005).
54. S. R. Starck *et al*., Translation from the 5′ untranslated region shapes the integrated stress response. *Science* **351**, aad3867 (2016).
55. Y. Zhang, T. Zhao, W. Li, M. Vore, The 5′-untranslated region of multidrug resistance associated protein 2 (MRP2; ABCC2) regulates downstream open reading frame expression through translational regulation. *Mol. Pharmacol.* **77**, 237–246 (2010).
56. S. Mali, R. D. Vierstra, Variation in upstream open reading frames contributes to allelic diversity in protein abundance. PRIDE (PRoteomics IDEntifications Database). https://www.ebi.ac.uk/pride/archive/projects/PXD026378. Deposited 31 May 2021.
57. F. McLoughlin *et al*., Maize multi-omics reveal roles for autophagic recycling in proteome remodelling and lipid turnover. *Nat. Plants* **4**, 1056–1070 (2018).
58. M. B. Hufford *et al*., De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
59. L. Jeacock, J. Faria, D. Horn, Codon usage bias controls mRNA and protein abundance in trypanosomatids. *eLife* **7**, e32496 (2018).
60. T. Tuller, Y. Y. Waldman, M. Kupiec, E. Ruppin, Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3645–3650 (2010).
61. S. Klumpp, J. Dong, T. Hwa, On ribosome load, codon bias and protein abundance. *PLoS One* **7**, e48542 (2012).
62. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
63. H. Zhao *et al*., CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
64. N. Yang *et al*., Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* **51**, 1052–1059 (2019).
65. J. L. Gage, Gage_uORF_allelic_variation_2021. Cyverse Data Commons. https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/Gage_uORF_allelic_variation_2021. Deposited 8 November 2021.
66. M. Lawrence *et al*., Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
67. R Core Team, *R: A Language and Environment for Statistical Computing* (R Core Team, 2018).
68. A. Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione [in Italian]. *Inst. Ital. Attuari. Giorn.* **4**, 83–91 (1933).
69. N. Smirnov, Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **19**, 279–281 (1948).
70. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
71. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
72. H. Li *et al*.; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. H. Tjeldnes *et al*., ORFik: A comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics* **22**, 336 (2021).
74. H. Zhang, Y. Wang, J. Lu, Function and evolution of upstream ORFs in eukaryotes. *Trends Biochem. Sci.* **44**, 782–794 (2019).
75. M. G. Kearse, J. E. Wilusz, Non-AUG translation: A new start for protein synthesis in eukaryotes. *Genes Dev.* **31**, 1717–1731 (2017).
76. Y. R. Li, M. J. Liu, Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. *Genome Res.* **30**, 1418–1433 (2020).
77. S. Anders, P. T. Pyl, W. Huber, HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
78. G. Hernández, V. G. Osnaya, X. Pérez-Martínez, Conservation and variability of the aug initiation codon context in eukaryotes. *Trends Biochem. Sci.* **44**, 1009–1021 (2019).
79. H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy, *Biostrings : Efficient manipulation of biological strings*. Biostrings version 2.58.0. https://rdrr.io/bioc/Biostrings/. Accessed 8 November 2021.
80. G. Covarrubias-Pazaran, Genome assisted prediction of quantitative traits using the R package sommer. *PLoS One* **11**, e0156744 (2016).
81. Neo4j Graph Data Platform, The leader in graph databases. *Neo4j Graph Database Platf*. Neo4j version 4.1.0. https://neo4j.com/product/neo4j-graph-database/. Accessed 15 June 2020.
82. C. Fay, neo4r: A 'Neo4J driver'. neo4r version 0.1.3. https://github.com/neo4j-rstats/neo4r. Accessed 23 June 2021.
83. S. A. Flint-Garcia *et al*., Maize association population: A high-resolution platform for quantitative trait locus dissection. *Plant J.* **44**, 1054–1064 (2005).
84. J. Yu, J. B. Holland, M. D. McMullen, E. S. Buckler, Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539–551 (2008).
85. J. L. Gage, B. Monier, A. Giri, E. S. Buckler, Ten years of the maize nested association mapping population: Impact, limitations, and future directions. *Plant Cell* **32**, 2083–2093 (2020).
86. A. Alexa, J. Rahnenfuhrer, topGO: Enrichment analysis for gene ontology. topGO version 2.42.0. https://bioconductor.org/packages/release/bioc/html/topGO.html. Accessed 3 May 2021.