Promoter architecture determines co-translational regulation of mRNA

María Lorena Espinar Calvo[1,†], Miquel Àngel Schikora Tamarit[2,†], Júlia Domingo Espinós[2,3,†], Lucas B. Carey[2,*]

Affiliations:

[1] *Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. 10 Aiguader, 08003 Barcelona, Spain.*

[2] *Universitat Pompeu Fabra (UPF), Barcelona, Spain.*

[3] *EMBL-CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.*

[*]Corresponding author. Email: lucas.carey@upf.edu (L.B.C.)

[†]Equal contribution

keywords: systems biology ; transcription

Abstract

Information that regulates gene expression is encoded throughout each gene, but the interaction between regulatory regions is poorly understood. Here we measure mRNA levels for 10,000 open reading frames (ORFs) transcribed from either an inducible or constitutive promoter. We find that the strength of co-translational regulation on mRNA levels is determined by promoter architecture. Using a novel computational-genetic screen of 6402 RNA-seq experiments we identify the RNA helicase Dbp2 as the mechanism by which co-translational regulation is reduced specifically for inducible promoters. Finally, we find that for constitutive genes, but not inducible genes, most of the information encoding regulation of mRNA levels in response to changes in growth rate is encoded in the ORF and not in the promoter. Thus the ORF sequence is a major regulator of gene expression, and a non-linear interaction between promoters and ORFs determines mRNA levels.

Introduction

Precise control of gene expression is essential(Skotheim et al. 2008), and the way this regulatory information is encoded in enhancers, promoters, 5' untranslated regions (UTRs) and other regulatory regions throughout throughout the gene sequence has been well characterized(Shalem et al. 2013; Keren et al. 2013; Dvir et al. 2013; Kudla et al. 2009; Shalem et al. 2015). The contribution of sequences features within the open reading frame (ORF) to both steady-state mRNA expression levels and to changes in expression, however, is controversial, as ORF-encoded sequence features may co-evolve with, but not directly influence expression(Plotkin and Kudla 2011). In addition, recent work has shown that some transcription factors (TFs) can influence mRNA stability and

localization(Braun et al. 2015; Bregman et al. 2011; Zid and O'Shea 2014). However, two major questions remain unclear: by which mechanisms do elements in the promoter influence the lifecycle of the mRNA, and where in the gene is the regulatory information encoded.

To determine if promoters influence the contribution of coding sequences to steady-state mRNA levels we generated a pooled library of over 10,000 yeast genomic DNA fragments cloned into plasmids containing either inducible TATA box containing GALL promoter or the TATA-less constitutive RPL4A promoter. This library exhibits over four orders of magnitude of expression, most of which can be predicted using a sequence-feature based mathematical model. Intriguingly, in both the library and in native transcripts, co-translational features such as codon usage more strongly influence mRNA levels when expression is driven by a constitutive promoter. To identify the molecular mechanism for this difference in co-translational regulation between promoter architectures we performed a computational genetic screen across 6402 RNA-seq experiments, and identified the RNA helicase Dbp2 as specifically insulating transcripts of TATA+, but not TATA-less, promoters from the effects of co-translational regulation on steady-state mRNA levels. Finally, we used RNA-seq and promoter-YFP expression data to show that, specifically for TATA-less promoters, the ORF contains more regulatory information than the promoter.

Results

ORF-encoded sequence features play an active role in regulating gene expression in native transcripts

To determine the ability of coding sequences to regulate gene expression we developed a method to decouple native genomic sequences from their 5' and 3' regulatory context **(Fig. 1A, fig. S1)**. We digested the yeast genome with four restriction enzymes and inserted the resulting random

genomic DNA fragments into a plasmid containing the inducible GALL promoter, a start codon, the ClaI restriction site and three staggered stop codons **(fig. S1A)**. This generated a pool of over 10,000 plasmids, each of which contains a single random fragment of the yeast genome. The cloned random gDNA fragments derive from the entire yeast genome and therefore contain parts of native coding sequences as well as untranslated regions (UTRs), promoters, transcription terminators, and other chromosomal features **(fig. S1B)**. We transformed the plasmids into yeast and measured expression of each transcript as the ratio between RNA and DNA abundance **(Fig. 1B).** In spite of having the same promoter, 5' untranslated region and transcription terminator, the expression of the gDNA fragments varies of over four orders of magnitude **(Fig. 1C)**. To determine the sources of this regulation and to quantify the contribution of individual sequence features to changes in expression we computed over 4000 sequence features for each insert, such as dimer and trimer nucleotide counts, GC content, codon bias and the presence of premature stop codons **(see methods)**. We fit a linear model with a minimal number of predictive features to the experimental data and found that a model with seven features can accurately predict expression (R^2=0.64 for all inserts, and R^2=0.61 for inserts with no premature stop codon) **(Fig. 1D, Fig. 2B)**, suggesting that a small number of ORF-encoded sequence features generate a large amount of the variation in gene expression. To determine if ORF-encoded sequence features regulate expression in native genes we used the same model, without taking into account 3'UTR length, to predict expression of native genes from their coding sequence **(Fig. 1E)**.

The ability of the ORF-feature model to predict expression could be because coding sequences play an active regulatory role or due to co-evolution and selection for certain sequence properties in genes whose expression levels are determined by more stereotypical regulatory regions such as the promoter. To differentiate correlation from causation we took advantage of data in which the strength of 859 promoters driving YFP was measured(Keren et al. 2013). We find that ORF features are equally good as promoters in predicting steady-state mRNA levels, and that a combined model including both promoter-YFP data and ORF features performs significantly better **(Fig. 1F)**.

Thus, ORF-encoded sequence features play an active role in regulating gene expression in native transcripts.

Co-translation regulation is stronger for mRNAs regulated by TATA-less promoters

While the majority of experiments that investigated transcriptional, co-translational and post-translational regulation used inducible promoters(Shalem et al. 2013; Radhakrishnan et al.; Puchta et al. 2016; Meaux et al. 2008; Shah et al. 2013), most promoters are in fact not condition-specific. Promoters can be roughly divided into two broad categories, those with a TATA box (SAGA-dominated and inducible or regulated) and those lacking a conserved TATA box (TFIID-dominated and constitutive)(Basehoar et al. 2004; Tirosh and Barkai 2008; de Jonge et al. 2016; Struhl 1986; Huisinga and Pugh 2004). To determine if ORF-encoded sequence features are regulatory for constitutive promoters we built the random gDNA fragment library in a second plasmid in which the regulated TATA box containing GALL promoter was replaced with the equally strong constitutive ribosomal RPL4A promoter **(Fig 2A, fig. S1).** We found that codon bias and 3'UTR length, both of which require active translation to be interpreted by the cell(Gardin et al. 2014; Kervestin and Jacobson 2012), have more importance in a model trained on RPL4Apr data compared to GALLpr data **(Fig. 2B)**. Consistent with this, both 3'UTR length, which in yeast is the main determinant of nonsense mediated decay (NMD)(He and Jacobson 2015), and codon bias, have a stronger effect on expression when transcription is driven by the constitutive RPL4A promoter **(Fig. 2C,D)**.

To determine if this promoter-specific effect of co-translational regulation affects native genes we divided promoters into two categories, those with an evolutionarily conserved TATA box (inducible) and those without (TATA-less, or constitutive)(Basehoar et al. 2004). To measure the effect of NMD separately for each class of transcript we used RNA-seq measurements from a Δ*upf1* strain, which is defective in nonsense-mediated-decay (NMD)(Smith et al. 2014). Consistent with

results from the gDNA library, native transcripts from TATA-less promoters are affected by NMD in a manner that increases with 3'UTR length, while transcripts from promoters with an evolutionarily conserved TATA box are unaffected by NMD **(Fig. 2E)**. To determine if the effect of codon bias on native genes is stronger for TATA-less promoters we used a linear model to predict mRNA levels from either promoter-YFP expression data alone (Keren et al. 2013) or from a model that includes both promoter-YFP data and codon bias (tAI(dos Reis et al. 2003)), and measured the increase in R^2 when including codon bias as a model feature (**Fig. 2F**). Consistent with results from the gDNA library, codon bias is more strongly predictive of mRNA levels for transcripts driven by TATA-less promoters (**Fig. 2G**). These differences are not due to differences in expression, mRNA stability or codon bias, since the distribution of mRNA expression, degradation rates and tAI are similar between the two classes of genes **(fig. S2)**. To determine if ORF-encoded sequence features are more predictive of native expression levels for TATA-less genes, we predicted expression from sequence separately for the two classes of genes. ORF features are more predictive of expression in TATA-less genes **(Fig. 2H,I)**. In both, for the gDNA library and native genes, co-translational regulation affects mRNA levels more strongly when expression is driven by TATA-less promoters, suggesting the existence of some mechanism that carries information from the promoter to the ribosome.

A genetic screen of 6402 RNA-seq experiments identifies the RNA helicase Dbp2 as TATA containing promoter specific insulator damper of co-translational regulation

To identify the molecular mechanism underlying the difference in co-translational regulation between TATA and TATA-less promoters we performed a computational genetic screen for mutants that alter our ability to predict mRNA levels from promoter-YFP data. We screened 6402 RNA-seq experiments, all RNA-seq experiments that have been done in *S. cerevisiae(Ziemann et al. 2015)*, for mutants that affect the ability of codon bias to predict expression in a promoter-class-specific manner **(Fig. 3A)**. For each of the 6402 experiments we calculated the increase in R^2 upon including codon

bias as a predictive feature. Codon bias has more predictive power for TATA-less genes across thousands of RNA-seq experiments (**Fig. 3B)**. We identified a single mutant which significantly reduces this difference across multiple biological replicates (t-test $p < 10^{-6}$). Deletion of Dbp2 (Beck et al. 2014) increases the effect of codon bias on gene expression specifically for TATA containing genes **(Fig. 3C, fig. S3)**. This suggests that Dbp2, a co-transcriptionally loaded RNA helicase(Ma et al. 2016; Cloutier et al. 2012), specifically reduces the effect of co-translational regulation on mRNA levels in a promoter-specific manner. Interestingly, Dbp2 physically interacts with nucleosomes, Pol-II-associated GTFs , and the ribosome(Stark et al. 2006) **(fig. S3)**.

To confirm the results of the screen and to determine if Dbp2 plays a role in other forms of co-translational regulation of mRNA levels, such as NMD, we built a set of four synthetic NMD reporters **(Fig. 3D)**. Each plasmid contains two fluorescent proteins, mCherry and YFP, with or without a premature termination codon (PTC) in the intervening linker. In the presence of a stop codon in the linker, YFP becomes a long 3'UTR and this targets the transcript for NMD(Muhlrad and Parker 1999). Each of the two constructs is driven by either a constitutive (RPL4A) or a regulated (GALL) promoter. We transformed the plasmids into wild-type, *Δupf1* and *Δdbp2* yeast and measured mCherry and YFP expression **(fig. S2)**. The NMD effect is the ratio of the mCherry signal in the noPTC and PTC constructs. In wild-type cells the effect of NMD is stronger in the TATA-less RPL4Apr plasmids, consistent with the gDNA library and native genes. Consistent with the computational RNA-seq screen, transcripts from the inducible GALL promoter are more strongly affected by NMD in a *Δdbp2* strain **(Fig. 3E)**. This suggests a model in which Dbp2 mutes the effect of codon bias and NMD on mRNA levels specifically for TATA containing promoters.

For TATA-less promoters, growth-rate mediated regulation of expression is implemented in the coding sequence, not the promoter

Why implement regulation in coding sequences? In response to changes in growth rate, cells alter the expression of thousands of genes(Brauer et al. 2008; García-Martínez et al. 2016). In yeast, expression of translation-related genes is positively correlated with growth rate, while genes associated with environmental stress response, respiration and oxidative phosphorylation are negatively correlated with growth rate **(Fig 4A, Table S1)**(Brauer et al. 2008). TATA-less genes tend to be positively correlated with growth rate, while TATA containing genes are negatively correlated (Fisher's exact test $p=10^{-27}$, odds ratio = 4.0). To determine if the information encoding regulation of expression in response to changes in growth rate is in the promoter, we took advantage of a dataset in which promoter-YFP expression data for 859 genes were measured in ten different environmental conditions(Keren et al. 2013). For each gene we calculated the slope, how promoter-YFP expression changes with growth rate, and compared the slopes of promoter-YFP data to the change of steady-state mRNA levels upon changes in growth. We find that changes in the mRNA levels of constitutive genes (TATA-less) are worse predicted by promoter-YFP data **(Fig. 4B)**. Therefore, for TATA-less genes, information encoding growth-rate mediated regulation of expression lies outside of the promoter. To determine if the information that regulates expression with growth rate is encoded in the ORF, we predicted expression-growth rate slopes from the coding sequence of each gene. We find that ORF-features fail to predict growth-rate-mediated changes in expression for TATA containing genes, but do so quite well for TATA-less genes **(Fig. 4C)**. For constitutive genes, there is more regulatory information in the ORF than in the promoter **(Fig. 4D,E)**.

Discussion

Genetic screens involving targeted and random mutation in both *cis* and *trans* regulatory elements have played a major role in defining the machinery that controls gene expression(Struhl 1995; Sharon et al. 2012). However, the global output of all biomedical research is exponentially greater than the ability of any single lab. In the genomic era, naturally existing genetic variation has been utilized to understand gene regulation(Rockman and Kruglyak 2006; Lindeboom et al. 2016). In this era of publically available data we are rapidly approaching a point at which functional genomic data will be available for almost all mutants in a wide variety of conditions. Here we show that publically available expression data can be used to discover molecular mechanisms underlying entirely novel types of gene regulation.

How gene expression is regulated was recognized as a problem prior to the identification of DNA as the genetic material(Monod 1966). Understanding how expression levels are encoded in the genome and how genetic variation affects expression has been one of the biggest challenges in molecular biology for the past fifty years(Beer and Tavazoie 2004; Rockman and Kruglyak 2006). Much of this work has focused on non-coding regions, specifically on the ability of transcription factors to bind distinct regulatory motifs in promoters and enhancers. The limited work that has been done understanding the regulatory potential of coding sequences has been done in an isolated manner, without considering the other regulatory regions of the gene(Kudla et al. 2009; Puchta et al. 2016; Radhakrishnan et al.). Our results here show that coding sequences contain a large amount of regulatory information that affects gene expression, and that this information interacts with information in the promoter architecture. Experimental systems and computational analyses that treat regulatory units in isolation miss molecular mechanisms that affect the expression of thousands of genes.

Why do organisms implement regulation at the level of the ORF? Measurements of steady-state mRNA levels from cells proliferating at different rates suggested that thousands of genes are regulated in response to changes in growth rate in both yeast(Brauer et al. 2008) and mammalian cells [Badia et al. in progress]. However, promoter activity, as measured by promoter-YFP signal, was found to scale with growth rate identically for almost all genes, suggesting little or no gene-specific regulation with growth rate(Keren et al. 2013). Here we show that information encoded in the ORF and interpreted in a manner that depends on the promoter architecture and on the RNA helicase Dbp2, resolves this conflict. Constitutive genes, which are the majority of genes in yeast, are regulated mostly by sequences encoded in the ORF. Why? Translation-mediated regulation is less noisy than that mediated by transcription factors (Carey et al. 2013). Therefore implementing global regulation at the level of the ORF may be a more reliable and efficient means to coordinate changes in expression across thousands of genes.

Methods

Library plasmids construction

All plasmids are named and described in **Table S2**. To generate the random gDNA fragment library we used as a vector backbone the plasmid pRS416 carrying the URA3 selectable marker. We cloned into pRS416 the GALL inducible promoter and CYC1 transcriptional terminator, generating P47. We next used Gibson assembly (NEB, E2611S) to insert the ATG-ClaI-Stop fragment (**Supplementary Figure 1**). This fragment was comprised of two annealed oligonucleotides, 174 and 175 (see **Table S3**) in 10X Oligo Annealing Buffer (400 µl 1M Tris pH 8, 80ul 0.5M EDTA pH8, 800ul 2.5M NaCl, 2720 µl DNase/RNase-Free Water) in a reaction containing: top and bottom strand DNA oligos (200

µM each), 2 µl of 10X Oligo Annealing Buffer and DNase/RNase-Free Water for a total volume of 20 µl. After incubation at 94ºC for 4 minutes, we allowed the reaction to cool to room temperature for 10 minutes. To obtain the RPL4Apr plasmid (P86) we amplified using Phusion polymerase (Thermo Fisher Scientific, F530S) in GC buffer the RPL4A promoter from genomic DNA using primers 222 and 223 (**Table S3**) and cloned the amplicon via Gibson assembly (NEB, E2611S) into the SacI and XbaI restriction sites of plasmid P47.

Library construction

The random yeast genomic sequences were obtained by cutting 40 µg of isolated genomic DNA (MasterPure™ Yeast DNA Purification Kit. Epicentre, MPY80200) from yeast strain FY4, a prototrophic S288C strain, for 4 hours at 37 ºC with a combination of 4 restriction enzymes (TaqαI, HpaII, MspI and AciI). All restriction enzymes used in this study are from New Englands Biolabs. Size selection was performed using a 2.5% agarose gel and purified DNA was ligated into ClaI cut and dephosphorylated (rSAP, New Englands Biolabs, M0371S) vector (P47). Ligated products were desalted by drop dialysis using 13 mm diameter, Type-VS Millipore membrane (Merck Millipore, VSWP01300) and 3 µl were transformed by electroporation using the UltraClone Kit with 10G Elite DUO (Lucigen, 60117-1). *E. coli* transformants were selected on LB medium (0.5% yeast extract, 1% NaCl, 15% bactotryptone) supplemented with 100 µg/ml ampicillin as a selection marker.

To obtain the RPL4Apr plasmid library with the same inserts that are present in the GALLpr library we amplified with Q5 DNA polymerase reaction mix (NEB, M0491S) the library previously cloned into P47 using primers 345 and 198 (see **Table S3**). After cutting pRS416-RPL4Apr (P86 in **Table S2**) with XbaI and XhoI restriction enzymes we cloned the PCR library fragments into pRS416-RPL4Apr via Gibson assembly (NEB, E2611S), obtaining the new promoter library that was transformed in *E. coli* electrocompetent cells (UltraClone Kit with 10G Elite DUO. Lucigen, 60117-1) and plated on LB medium (0.5% yeast extract, 1% NaCl, 15% bactotryptone) supplemented with 100

µg/ml ampicillin as a selection marker. The constitutive and inducible library plasmids were purified using the Nucleobond XtraMidi Plus EF (Machery-Nagel, 740422.50).

Yeast transformation

The gDNA libraries cloned into P47 (GALL) and P86 (RPL4A) were transformed into BY4741 (Y49) (**Table S2**) via the lithium acetate method(Gietz and Woods 2006) and plated in glucose synthetic complete dropout plates lacking uracil for plasmid selection. In the same way, we transformed the GALL inducible library of the yeast native ORF in the mutant strains Y194 and Y195 (**Table S2**). After 2 days of growth at 30C we collected all the transformants and we resuspended them to OD of 0.5 in synthetic complete dropout plates lacking uracil supplemented with 2% galactose as a carbon source in order to get the GALL promoter induction. After growing the cells overnight we scraped off the plates and inoculated four 500 ml flasks of the same medium to OD of 0.025 as biological replicates. The cultures were grown about 10 hours at 30°C to reach an exponential phase with an OD of 0.5, where we took 1.5 ml samples in order to extract plasmid DNA (MasterPure™ Yeast DNA Purification Kit, Epicentre) and transcribed RNA (MasterPure™ Yeast RNA Purification Kit, Epicentre) from all the biological replicas. We pelleted at 4 ºC and froze every sample at -80 ºC until processing.

NGS sample preparation

We used a gene specific primer to generate the cDNA from each RNA preparation (**Supplementary Figure 1**, primer 198 depicted in yellow and see **Table S3**) using the ThermoScript RT-PCR System (Life Technologies, 11146-024). With barcoded primers we amplified the second DNA strand and the library plasmids DNA in order to differentiate between replicas and sample origin, in this case, DNA and RNA (for details of barcoded primers combination for the libraries see **Table S3** and **Table S4**). The PCR sample was set up as 50 µl reaction of Q5 DNA polymerase (NEB, M0491S) in the following cycling conditions: 98°C for 30s; 98°C for 10s, 55-64°C (depending on

primer combination for demultiplexing on technical variability and biological replicates) for 30s and 72°C for 30s (20 cycles); and 72°C for 2 min. PCR products were purified using MinElute PCR Purification Kit (Qiagen, 28004). NGS libraries were prepared from 100 ng of the purified DNA amplicons using Ovation Rapid DR System (Nugen, 0319-32) according to manufacturer's instructions. Each library was visualized on a Bioanalyzer (Agilent Technologies) and quantified by qPCR with a Kapa Library Quantification Kit (Kapa Biosystems, KK4835). The libraries were sequenced through Illumina HiSeq 2500 125bp paired-end reads method.

NMD reporter plasmids construction and measurement

The four NMD reporter constructs GALLpr-mCherry-PTC-YFP, GALLpr-mCherry-linker-YFP, RPL4Apr-mCherry-PTC-YFP and RPL4Apr-mCherry-linker-YFP (P71, 68, 74 and P78, respectively. For details see **Table S2**) were generated by fusion PCR reaction with overlapping oligos amplifying mCherry with primers 538 and 539 (in case of stop codon presence in the linker between mCherry and YFP) or primers 538 and 540 (in case of stop codon absence in the linker between mCherry and YFP). YFP was amplified with primers 541 and 542 (see **Table S2** for primers details). For all PCR reactions Phusion polymerase (Thermo Fisher Scientific, F530S) in GC buffer was used. After fusion of both fragments, the final product was purified using MinElute PCR Purification Kit (Qiagen, 28004) and cloned in the ClaI cut and dephosphorylated with rSAP (New Englands Biolabs, M0371S) pRS416-GaLLpr and pRS416-RPL4Apr vectors (P47 and P86 in **Table S2**). The four NMD reporter plasmids were transformed into Y49 and multiple transformants of each construct were inoculated into 96 well plates containing SCGal-URA media. Cells were grown overnight, diluted 1:50, and measured by flow cytometry after 15 hours of growth, when the OD600 was around 0.5. Wells in which the cell density was too high or too low were discarded. All flow cytometry was performed on a BD LSRFortessa (BD Biosciences) with 488nm and 561nm lasers with 530/28 and 610/20 filters for YFP or mCherry, respectively. Analysis of flow-cytometry data was performed using a previously described custom MATLAB pipeline (Carey et

al. 2013). These two promoters have approximately equal expression levels **(Table S2).** The same procedure was done for the constructs in Y194 and Y401 strains.

Sequencing data processing

Samples and replicas for the library were all compact in fastq files with 125bp paired-end reads. We used the primers to amplify the library as barcodes to demultiplex samples and replicates. We first removed 3' Illumina adapters and 3' possible PCR amplification primer barcodes if necessary (in case reads exceeded the length of the inserted fragment). Next, by detecting the PCR primer barcode we split reads into different samples of origin (RNA or DNA, replicate 1 or 2 and strain). After trimming the PCR primers of the reads sequence, reads were mapped to the yeast genome (bowtie2) to obtain the inserted sequence and the count of reads per fragment inserted. Once obtained the inserted fragment in the genome we counted the number of reads of each individual insert had at the level of DNA and RNA. To obtain the full sequence of the generated transcript of each insert we attached the constant 5' and 3' nucleotide sequence of the plasmid construct which contained the start and stop codons (**Supplementary Table 5**).

Expression values and data analysis

Once obtained a number of reads per unique sequence for each sample and replica, we first normalized read counts by the total amount of reads per sample and replica following the formula:

$$a_{ij} = a_{ij} * \left(\frac{\sum_{j=1}^{S} \sum_{i=1}^{R} a_{ij}}{S \sum_{i=1}^{R} a_{ij}} \right)$$

Where a_{ij} represents the read count for a unique sequence variant per replica and sample. S is the number of samples (RNA, DNA and technical replicas) and R the total number of unique transcripts. In order to normalize by technical and biological bias we used as values of expression the $\log_2$ ratio between the averaged replicas DNA reads and RNA reads:

$$log_2\left(\frac{RNA+p}{DNA+p}\right)$$

Where p represents a pseudocount (p = 0.5). In order to uniformize the measurements across all libraries we calculated expression as the z-score value of this last log_2 ratio. For the codon bias and NMD analysis we regressed out the contribution of insert GC and length. This allowed us to measure the GC and length-independent role of 3'utr and CAI to expression.

Model to predict expression in the ORF library using inserts features

To build a multiple linear regression model to predict expression values in the ORF library we scanned for ORF features that could significantly predict expression. Predictors included premature termination codons, 3'UTR length, codon bias (tRNA adaptation index, tAI) insert length and GC content. From the subset of fragments without PTCs that are fully contained within native yeast ORFs, the directionality of the inserted fragment had a significant impact on expression (fragments oriented in the same direction as the native gene had higher expression that reverse oriented ones, t-test $p<10^{-11}$). To select sequence features that determined this differences we generated a classifier model with more than 4000 features (dimers, trimers, hexamers, frequency of bases…) and subsequently select those features that better distinguished this two categories. We then selected the subset of features that are predictive of orientation, predictive of expression, and likely to be functionally interpreted by the translating ribosome inside the cell. For all model predictions we used a 10-fold cross-validation scheme to assess the accuracy of all models and to test for overfitting.

Predicting mRNA level of native genes based on ORF features.

To measure the contribution of the coding region to changes in expression we used a linear regression model that included exclusively ORF features selected previously using the gDNA library. 3'UTR length and premature termination codon features were not used for models for native genes. We used 10-fold cross validation to test the accuracy of the model. The same model was used to

predict the expression of genes for which the promoter contribution to expression was previously measured (Keren et al. 2013).

Yeast NMD strength analysis

To measure NMD strength for native yeast genes wild-type (SRR1258470,SRR1258471) and *upf1* (SRR1258533,SRR1258534) RNA-seq reads were downloaded from NCBI SRA (PRJNA245106) (Smith et al. 2014). Expression was quantified using kallisto (Bray et al. 2016) (--single -l 180 -s 20) and orf_coding_all.fasta from SGD (Cherry et al. 2012) (yeast genome R64-2-1). For each gene the NMD effect was defined as log2(mean(expression *upf1*) / mean(expression WT)). To work in a parameter regime comparable to that of the gDNA library, only genes with 3'UTR lengths (Nagalakshmi et al. 2008) from 200 - 400nt were analyzed.

Digital expression explorer analysis

We obtained the processed data of 6402 RNAseq dataset from the digital expression explorer database(Ziemann et al. 2015). Expression was calculated normalizing sequence read counts by transcript length and sum of reads in each experiment. The effect of codon bias on expression was inferred for each experiment as described in Results **(Fig. 2F)**. We next searched for experiments in which the promoter-dependent effect changes from the "wild-type" behavior. We grouped the expression datasets by experimental setup (each setup corresponds to a different work) and graphed each of them separately in the same space (TATA/TATA-less vs "effect of codon bias on expression"). We looked for experimental setups in which a subset of common samples (for example with the same altered genotype) differs consistently from another subset (for example wild-type strains). Any such clustering would have been informative of a mechanism.

Data access

DNA-seq and RNA-seq data generated in this study are deposited in the National Center for Biotechnology Information Gene Expression Omnibus under accession GSE100452. All code for figures and analysis are on GitHub (https://github.com/MikiSchikora/Promoter_architecture_Espinar17/).

Acknowledgments

L.B.C. was supported by Ministerio de Economía y Competitividad (MINECO) (BFU2015-68351-P) and AGAUR (2014SGR0974). We'd like to thank Wenfeng Qian, Shaohuan Wu, Aaron New, Francesc Posas, and members of the Barcelona Yeast Group and the CRG/UPF Computational Genomics and Systems Biology groups for helpful comments on the results and manuscript.

Disclosure declaration

The authors declare no competing financial interest.

References

References

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699–709.

Beck ZT, Cloutier SC, Schipma MJ, Petell CJ, Ma WK, Tran EJ. 2014. Regulation of glucose-dependent gene expression by the RNA helicase Dbp2 in Saccharomyces cerevisiae. *Genetics* **198**: 1001–1014.

Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**: 185–198.

Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D. 2008. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell* **19**: 352–367.

Braun KA, Dombek KM, Young ET. 2015. Snf1-Dependent Transcription Confers Glucose-Induced Decay upon the mRNA Product. *Mol Cell Biol* **36**: 628–644.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, Choder M. 2011. Promoter elements regulate cytoplasmic mRNA decay. *Cell* **147**: 1473–1483.

Carey LB, van Dijk D, Sloot PMA, Kaandorp JA, Segal E. 2013. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol* **11**: e1001528.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**: D700–5.

Cloutier SC, Ma WK, Nguyen LT, Tran EJ. 2012. The DEAD-box RNA helicase Dbp2 connects RNA quality control with repression of aberrant transcription. *J Biol Chem* **287**: 26155–26166.

de Jonge WJ, O'Duibhir E, Lijnzaad P, van Leenen D, Groot Koerkamp MJ, Kemmeren P, Holstege FC. 2016. Molecular mechanisms that distinguish TFIID housekeeping from regulatable SAGA promoters. *EMBO J*. http://dx.doi.org/10.15252/embj.201695621.

dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res* **31**: 6976–6985.

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A* **110**: E2792–801.

García-Martínez J, Troulé K, Chávez S, Pérez-Ortín JE. 2016. Growth rate controls mRNA turnover in steady and non-steady states. *RNA Biol* **13**: 1175–1181.

Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. 2014. Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* **3**. http://dx.doi.org/10.7554/eLife.03735.

Gietz RD, Woods RA. 2006. Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol* **313**: 107–120.

He F, Jacobson A. 2015. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annu Rev Genet* **49**: 339–366.

Huisinga KL, Pugh BF. 2004. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. *Mol Cell* **13**: 573–585.

Keren L, Zackay O, Lotan-Pompan M, Barenholz U, Dekel E, Sasson V, Aidelberg G, Bren A, Zeevi D, Weinberger A, et al. 2013. Promoters maintain their relative activity levels under different

growth conditions. *Mol Syst Biol* **9**: 701.

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

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**: 255–258.

Lindeboom RGH, Supek F, Lehner B. 2016. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet*. http://dx.doi.org/10.1038/ng.3664.

Ma WK, Paudel BP, Xing Z, Sabath IG, Rueda D, Tran EJ. 2016. Recruitment, Duplex Unwinding and Protein-Mediated Inhibition of the Dead-Box RNA Helicase Dbp2 at Actively Transcribed Chromatin. *J Mol Biol* **428**: 1091–1106.

Meaux S, van Hoof A, Baker KE. 2008. Nonsense-mediated mRNA decay in yeast does not require PAB1 or a poly(A) tail. *Mol Cell* **29**: 134–140.

Monod J. 1966. From enzymatic adaptation to allosteric transitions. *Science* **154**: 475–483.

Muhlrad D, Parker R. 1999. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA* **5**: 1299–1307.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Neymotin B, Athanasiadou R, Gresham D. 2014. Determination of in vivo RNA kinetics using RATE-seq. *RNA* **20**: 1645–1652.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32–42.

Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G. 2016. Network of epistatic interactions within a yeast snoRNA. *Science*. http://dx.doi.org/10.1126/science.aaf0965.

Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Coller J. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell*. http://www.sciencedirect.com/science/article/pii/S0092867416311503.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat Rev Genet* **7**: 862–872.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* **153**: 1589–1601.

Shalem O, Carey L, Zeevi D, Sharon E, Keren L, Weinberger A, Dahan O, Pilpel Y, Segal E. 2013. Measurements of the impact of 3' end sequences on gene expression reveal wide range and sequence dependent effects. *PLoS Comput Biol* **9**: e1002934.

Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, Segal E. 2015. Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet* **11**: e1005147.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands

of systematically designed promoters. *Nat Biotechnol* **30**: 521–530.

Skotheim JM, Di Talia S, Siggia ED, Cross FR. 2008. Positive feedback of G1 cyclins ensures coherent cell cycle entry. *Nature* **454**: 291–296.

Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, Coller J, Baker KE. 2014. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Rep* **7**: 1858–1866.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**: D535–9.

Struhl K. 1986. Constitutive and inducible Saccharomyces cerevisiae promoters: evidence for two distinct molecular mechanisms. *Mol Cell Biol* **6**: 3847–3853.

Struhl K. 1995. Yeast transcriptional regulatory mechanisms. *Annu Rev Genet* **29**: 651–674.

Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Res* **18**: 1084–1091.

Zid BM, O'Shea EK. 2014. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature* **514**: 117–121.

Ziemann M, Kaspi A, Lazarus R, El-Osta A. 2015. Digital Expression Explorer: A user-friendly repository of uniformly processed RNA-seq data. *ComBio2015, volume POS-TUE-099, Melbourne*. https://www.researchgate.net/profile/Mark_Ziemann/publication/282287193_Digital_Expression_Explorer_A_user-friendly_repository_of_uniformly_processed_RNA-seq_data/links/560a7ba008ae4d86bb13926d.pdf.

44. B. Neymotin, R. Athanasiadou, D. Gresham, Determination of in vivo RNA kinetics using RATE-seq. *RNA*. **20**, 1645–1652 (2014).

Figures

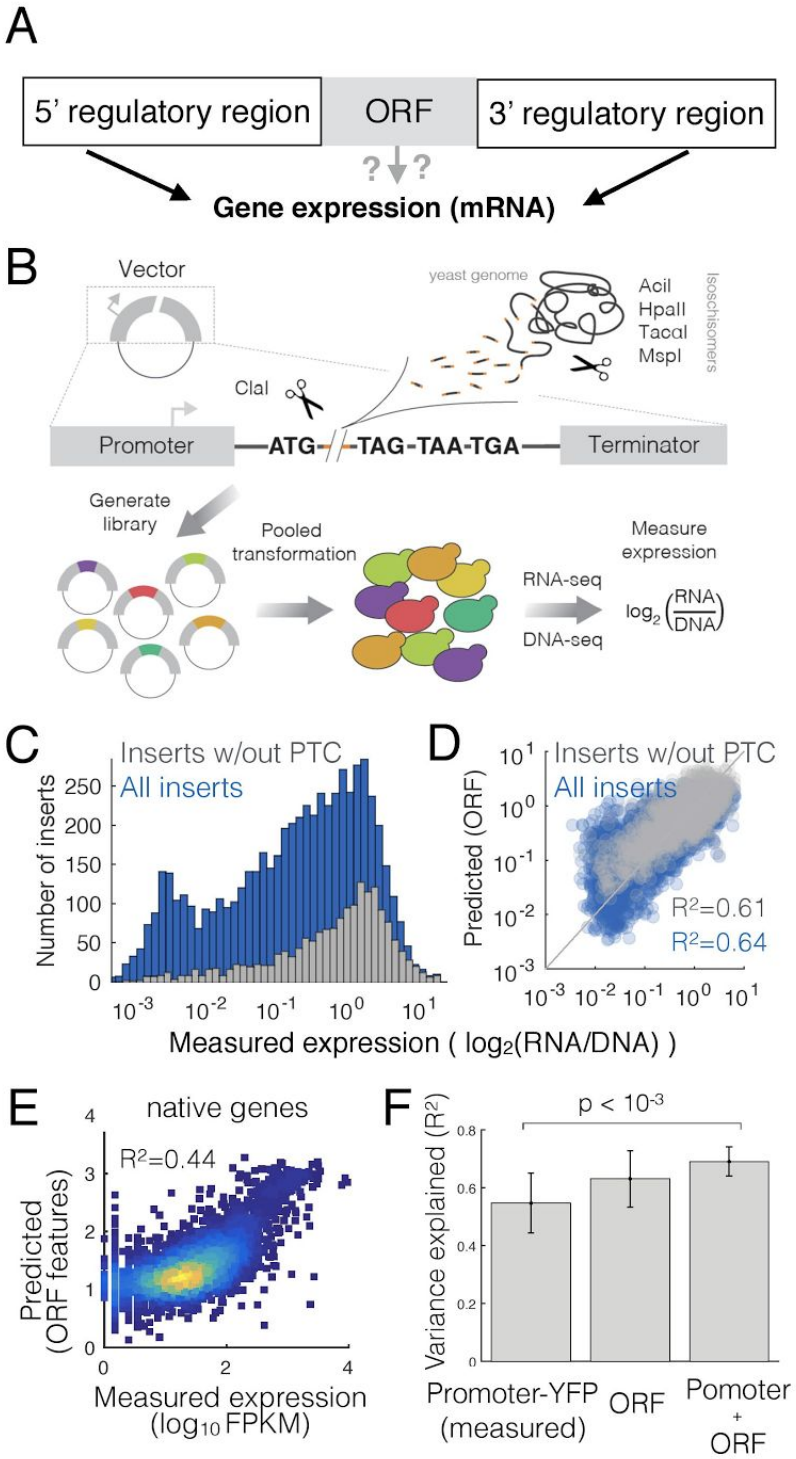

A
5' regulatory region
ORF
3' regulatory region
? ?
Gene expression (mRNA)
B
Vector
yeast genome
AcuI
HpaII
TaqαI
MspI
Isoschisomers
ClaI
Promoter
ATG
TAG-TAA-TGA
Terminator
Generate library
Pooled transformation
Measure expression
RNA-seq
DNA-seq
log2 (RNA/DNA)
C
Inserts w/out PTC
All inserts
Number of inserts
D
Inserts w/out PTC
All inserts
Predicted (ORF)
R2=0.61
R2=0.64
Measured expression (log2(RNA/DNA))
E
native genes
R2=0.44
Predicted (ORF features)
Measured expression (log10 FPKM)
F
p < 10-3
Variance explained (R2)
Promoter-YFP (measured)
ORF
Pomoter + ORF

Figure 1. An expression library of genomic fragments to quantify the ability of ORF-encoded sequence features to regulate gene expression. (A) A schematic view of the initial question: *Does the open reading frame determine gene expression?* (**B)** Scheme of the gDNA library preparation and expression measurements. **(C)** Measured expression distributions for all inserts (blue) and only those lacking a premature termination codon (grey). **(D)** Measured vs predicted expression levels in the gDNA library. Expression is predicted from the sequence of each gDNA insert using a 10-fold cross-validated linear model (R^2 is calculated across all test data from all cross-validations). (**E)** Expression predicted from the sequence of each native yeast ORF using the same features as for the gDNA library. **(F),** Including ORF-encoded features in a model of expression increases the ability of promoter-YFP data to predict steady-state mRNA levels. Error bars are standard deviation from 10-fold cross-validation.

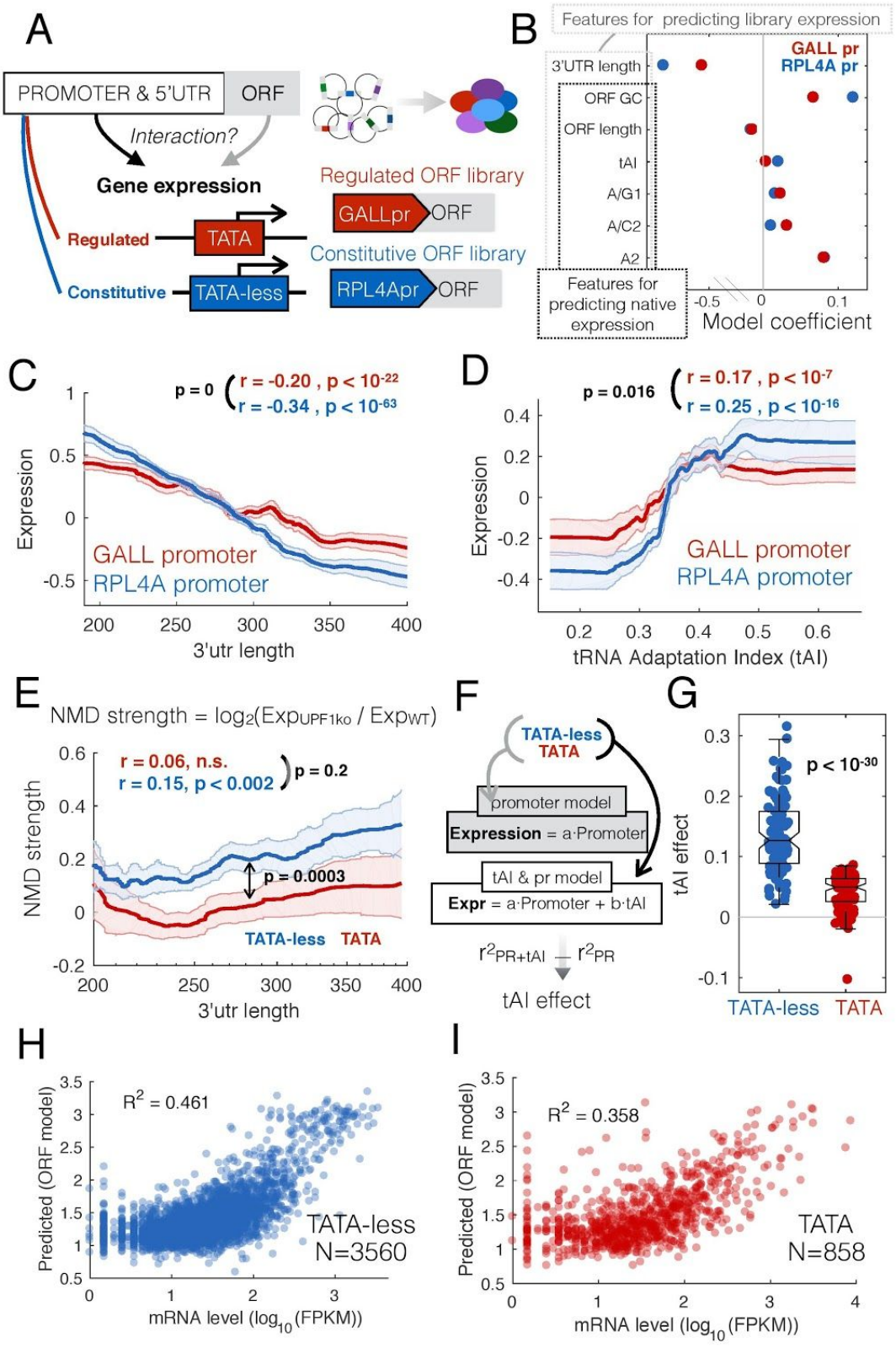


Figure 2. A gDNA library with different promoters identifies sequence features that interact with the promoter to determine gene expression. (A) The gDNA library was cloned under the

control of either the $TATA^+$ GALL promoter or the TATA-less ribosomal RPL4A promoter and the expression of both libraries measured in yeast growing on galactose as a carbon source. **(B)** Coefficients from the multiple linear model based on ORF sequence features from libraries with the two different promoters. Outlined are the features used for predicting expression in both the libraries or native genes. **(C,D)** Lines show the median expression for inserts binned by 3'UTR length (c) or codon bias (d). Correlation values are for unbinned data and the p-value is a test for a significant difference between the two correlation values using bootstrapping. **(E)** NMD effect, measured as the log2 ratio in mRNA (TPM) between *upf1* and wild-type cells for native transcripts. Lines show the median NMD effect across transcripts binned by 3'UTR length for TATA-containing (red) and TATA-less promoters (blue). The p-value shown between the red and blue lines is for a t-test on unbinned data between TATA and TATA-less promoters. **(F)** The makeup of two linear models, one that predicts mRNA levels from promoter-YFP data, and the other that includes codon bias (tRNA adaptation index, tAI) as an additional predictor. For both models, native genes are split into two classes, TATA and TATA-less, and tAI effect is the difference in R^2 between the two models. **(G)** Difference in tAI effect for random samplings of equal numbers of genes from each class. **(H,I)** An ORF-encoded sequence feature model was trained to predict mRNA levels for TATA and TATA-less promoters (R^2 = squared Pearson correlation coefficient).

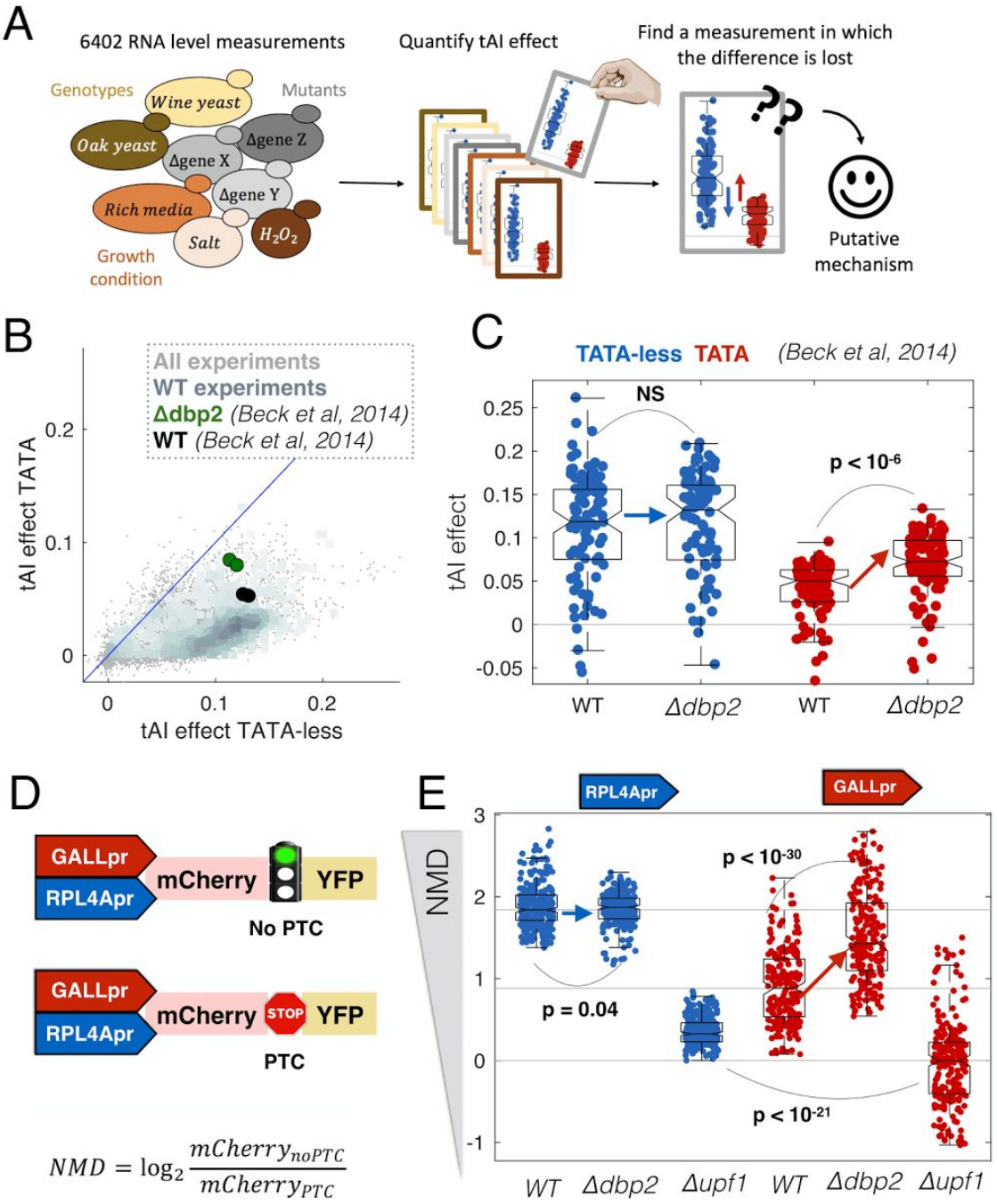

A
6402 RNA level measurements
Genotypes
Wine yeast
Mutants
Oak yeast
Δgene Z
Δgene X
Δgene Y
Rich media
Salt
H_2O_2
Growth condition
Quantify tAI effect
Find a measurement in which the difference is lost
Putative mechanism
B
All experiments
WT experiments
Δdbp2 (Beck et al, 2014)
WT (Beck et al, 2014)
tAI effect TATA
tAI effect TATA-less
C
TATA-less TATA (Beck et al, 2014)
NS
p < 10-6
tAI effect
WT
Δdbp2
WT
Δdbp2
D
GALLpr
RPL4Apr
mCherry
YFP
No PTC
STOP
PTC
$NMD = \log_2 \frac{mCherry_{noPTC}}{mCherry_{PTC}}$
E
RPL4Apr
GALLpr
NMD
p = 0.04
p < 10-30
p < 10-21
WT
Δdbp2
Δupf1
WT
Δdbp2
Δupf1

Figure 3. A computational genetic screen for mutants that alter promoter-specific co-translational regulation. (A) We analyzed expression data for 6402 RNA-seq experiments in *S. cerevisiae(Ziemann et al. 2015)*, and for each experiment predicted mRNA levels from promoter-YFP data alone, or promoter-YFP and codon bias (tAI). **(B)** For each for the 6402 experiments we calculated the ability of tAI to improve R^2 for TATA-less and TATA-containing genes. Highlighted is a single experiment (Beck et al. 2014) in which the R^2 changes in a promoter-class-specific manner between wild-type and mutant cells. **(C)** Shown are the changes in R^2 for wild-type and Δ*dbp2* cells for TATA containing and TATA-less transcripts. Each point corresponds to the effect of tAI on expression from equal-sized random samples of TATA (red) and TATA-less (blue) genes. Boxplots show the median and interquartile range. **(D)** Synthetic system for measuring the effect of NMD on gene expression. NMD is the ratio in mCherry expression between plasmids with and without a premature termination codon (PTC). **(E)** In each genotype we measured mCherry expression eight times (biological replicates) for each of the four plasmids. As NMD effect is the ratio between two plasmids, for each promoter and genotype we calculated 8^2 ratios. Boxplots show the median and interquartile range for each set of ratios.

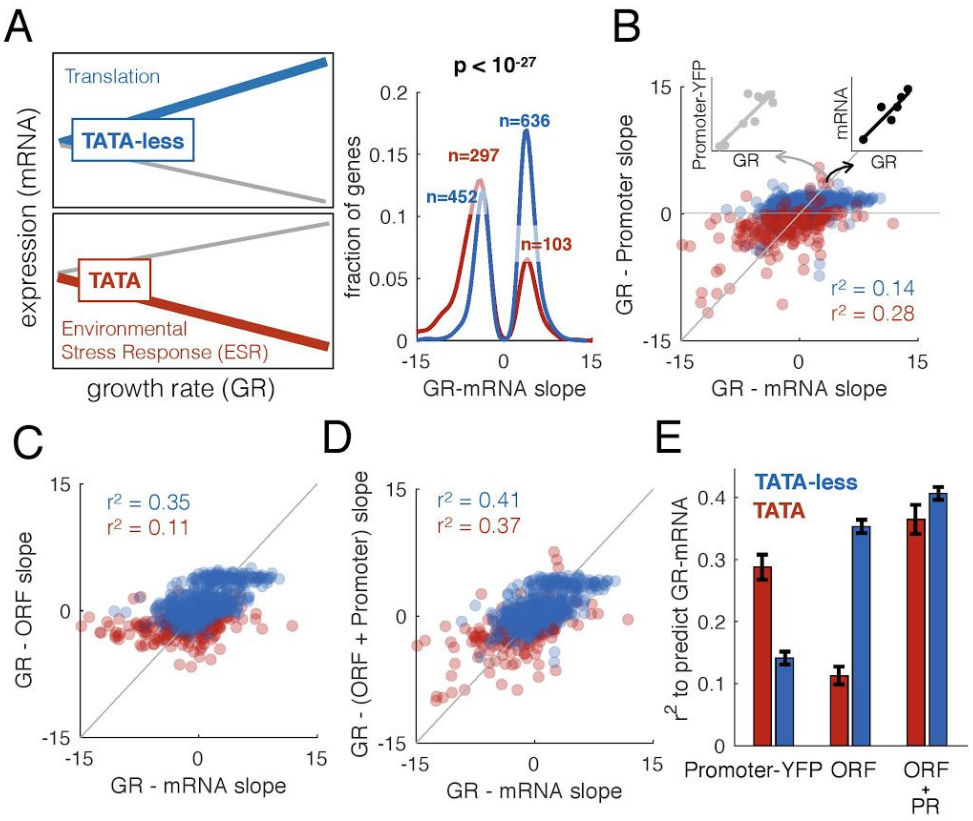
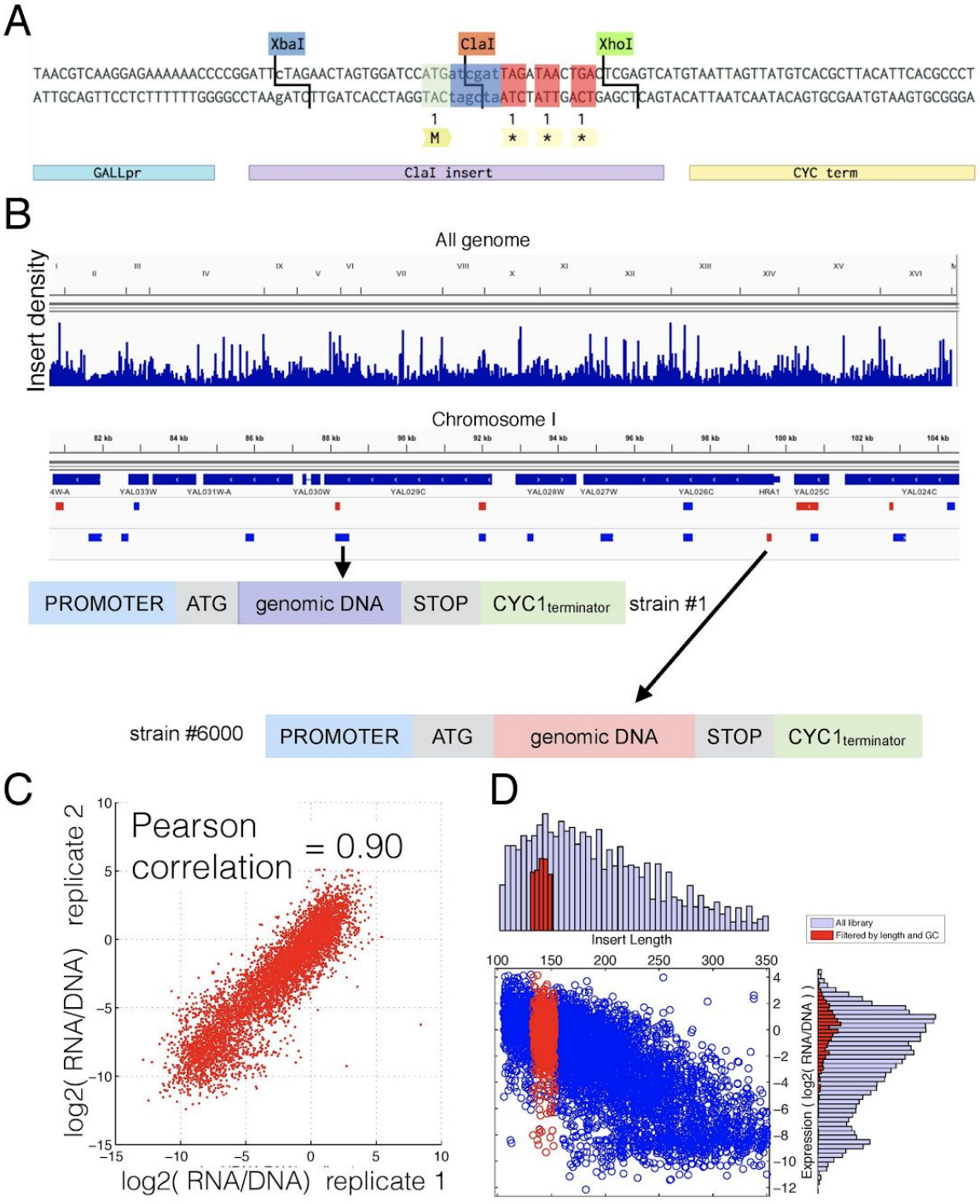
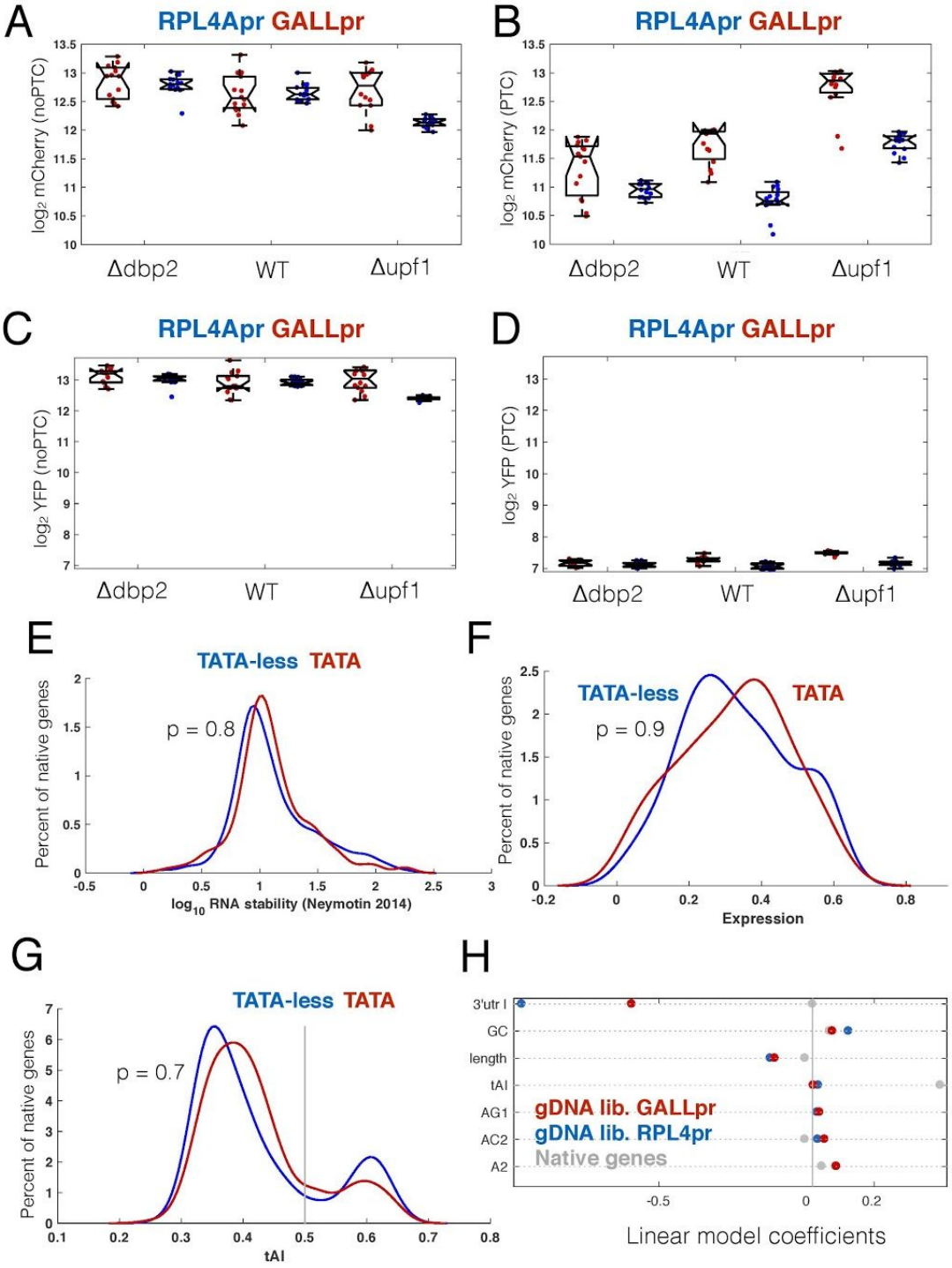


Figure 4. ORF features regulate expression in response to changes in growth rate. (A) Brauer et al(Brauer et al. 2008) grew yeast at six different growth rates in each of six different environments and measured how the expression of each gene changed as a function of growth rate (slope GR-mRNA). TATA-less genes (blue, N=1088) increase in expression, while those with TATA boxes (red, N=401) more often decrease. The p-value is from a Fisher's Exact test. **(B)** Keren et al(Keren et al. 2013) measured promoter-YFP expression at different growth rates. Each point shows the slope (expression change as a function of growth rate) for promoter-YFP (y-axis) and steady-state mRNA (x-axis) for TATA (red) or TATA-less (blue) genes. **(C)** GR-mRNA was predicted from ORF features allowing us to calculate the slope between growth rate and the contribution of ORF to expression (GR-ORF). Shown is the relationship between GR-ORF and GR-mRNA (see fig. A), for each type of genes. **(D)** GR-mRNA was predicted from both GR-Promoter and ORF features (GR- (ORF + Promoter))**.** Shown is the relationship between this and GR-mRNA, for each type of genes **(E)** R^2 values for a model that predicts GR-mRNA from GR-Promoter, ORF features or ORF + Promoter, for each type of genes. Error-bars correspond to the standard deviation, calculated with bootstrapping.

Supplementary Figures

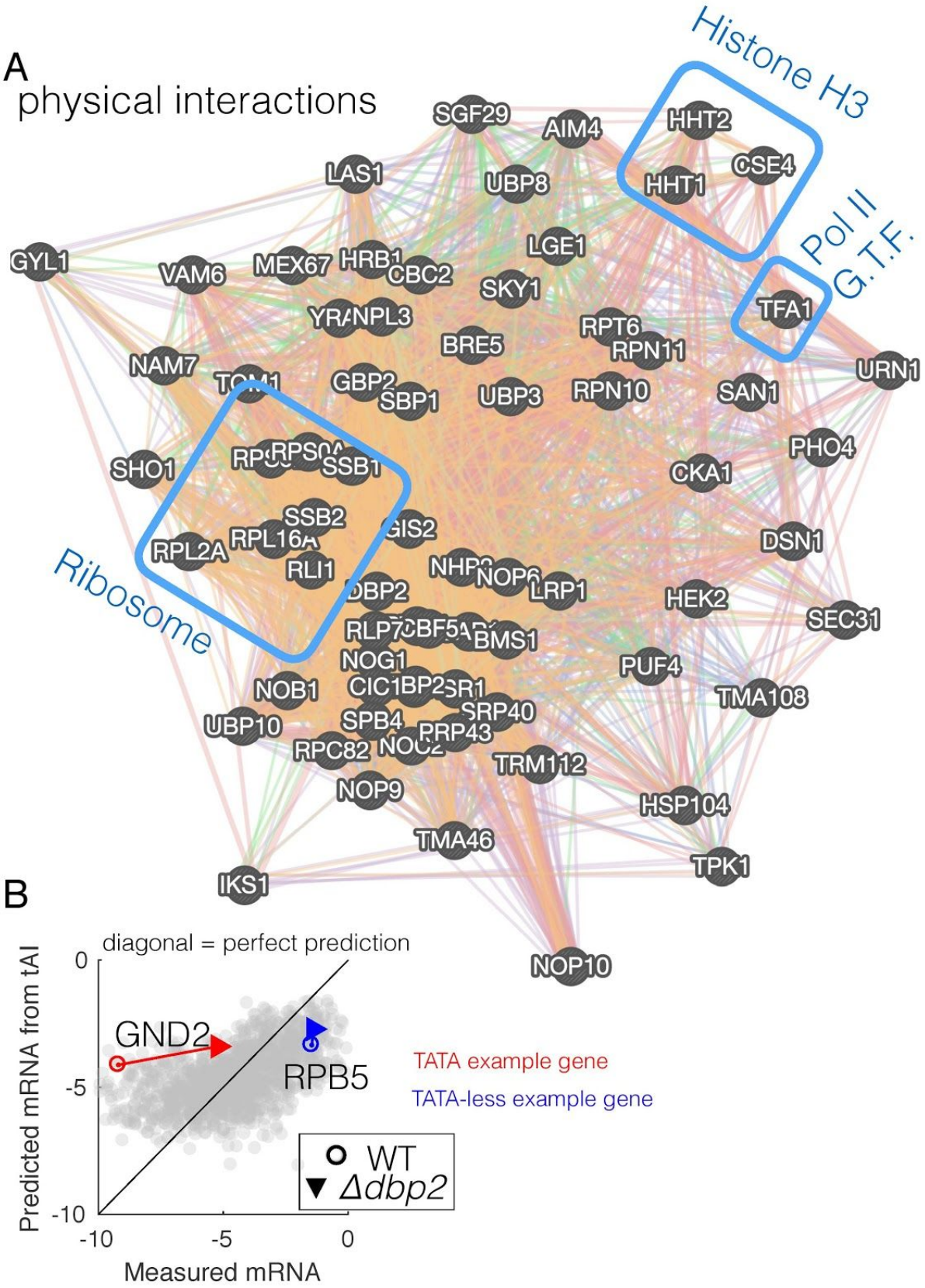


Supplementary Figure 1. gDNA library and construct. (A) Sequence details of the ClaI restriction site (dark blue) where fragments of the yeast genome are cloned. The construct contains a GALL promoter (left-bottom blue) a designed region where the ClaI site is located (bottom purple annotation) and the CYC1 terminator (bottom yellow annotation). The designed region contains a start codon ATG (green), followed by the ClaI site and finishing with three stop codons (red) separated by one nucleotide to cover all possible reading frames. **(B)** Top: distribution of the genomic fragments inserted in the GALL library. Bottom: two example resulting strains after insertion of two chromosome I fragments. **(C)** Expression measurements across biological replicates are reproducible. **(D)** Shown are the length and expression distribution for the entire GALL library (blue) and a small subset with similar insert length (133-153 nt) and GC content (0.35-0.45% GC). In this subset with expression varies by 4 orders of magnitude.



Supplementary Figure 2. Several parameters of constitutive and regulated genes. (A) mCherry expression of the NMD reporter constructs lacking a premature termination codon (PTC) between mCherry and YFP (see **Fig. 3D**), driven by the constitutive (blue) and regulated (red) promoter. Each construct was measured in wild-type, *upf1Δ* and *dbp2Δ* strains**.** Each point is from a different

biological replicate. **(B)** The same as in (A), but for constructs bearing a PTC in the intervening mCherry-YFP linker. **(C,D)** The same as (A,B), but showing YFP expression measurements. **(E)** Distribution of RNA half-life values reported Neymotin et al (Neymotin et al. 2014). Shown are these values for TATA (red) and TATA-less (blue) genes. The p value corresponds to a ttest. **(F)** Distribution of mRNA expression levels (in $\log_2$ TPM) for the same groups as in (E). **(G)** tRNA Adaptation Index (tAI) scores for for the same groups as in (E). The line refers to the threshold used in the analysis of how tAI affects expression in native genes (see Results). (**H**) Coefficients obtained with the linear model that predicts expression from sequence features.


A
physical interactions
Histone H3
Pol II
G.T.F.
Ribosome
SGF29
AIM4
HHT2
CSE4
HHT1
LAS1
UBP8
LGE1
GYL1
VAM6
MEX67
HRB1
CBC2
SKY1
TFA1
RPT6
RPN11
BRE5
NAM7
URN1
GBP2
SBP1
UBP3
RPN10
SAN1
PHO4
SHO1
SSB1
CKA1
SSB2
GIS2
DSN1
RPL2A
RPL16A
RLI1
LRP1
HEK2
SEC31
BMS1
NOG1
PUF4
NOB1
TMA108
SRP40
UBP10
SPB4
PRP43
RPC82
TRM112
NOP9
HSP104
TMA46
TPK1
IKS1
NOP10
B
diagram = perfect prediction
GND2
RPB5
TATA example gene
TATA-less example gene
WT
Δdbp2
Predicted mRNA from tAI
Measured mRNA
0
-5
-10
-10
-5
0

Supplementary Figure 3. DBP2 functional properties. (A) DBP2 physical interactome of protein-protein interactions (Cherry et al. 2012). It interacts with ribosomal, histone H3 and General Transcription Factors (GTFs) of Pol-II (TAF1 is the yeast ortholog of the mammalian TFIIE). **(B)** Shown is the ability of tAI to predict steady-state mRNA levels for an RNA-seq experiment performed in a WT genotype (Beck et al. 2014). Two example genes with different promoters (TATA, and TATA-less) show different changes regarding the effect of tAI in a *dbp2Δ* strain, performed under the same conditions.

Supplementary Tables

Supplementary Table 1: Results of Gene Onthology Analysis

Supplementary Table 2: Table of strains and plasmids used in this study

Supplementary Table 3: Table of used primers

Supplementary Table 4: PCR Primer barcode combination

Supplementary Table 5: Nucleotide sequences of 5' and 3' regions of random ORF library plasmid construct (5' to 3')