SUPPLEMENT

SEQUENCE SIGNATURES AND MRNA CONCENTRATION CAN EXPLAIN TWO-THIRDS OF PROTEIN ABUNDANCE VARIATION IN A HUMAN CELL LINE

Christine Vogel[1,$,*], Raquel de Sousa Abreu[2,$], Daijin Ko[3], Shu-Yun Le[4], Bruce A. Shapiro[4], Suzanne C. Burns[2], Daniel R. Boutz[1] , Devraj Sandhu[2], Edward M. Marcotte[1], Luiz O. Penalva[2,*]

[1] Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology,
University of Texas at Austin, TX, USA

[2] Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, TX, USA

[3] Department of Management Science and Statistics
University of Texas at San Antonio, TX, USA

[4] Center for Cancer Research Nanobiology Program, National Cancer Institute

[$] Equally contributing authors

* Corresponding authors: cvogel@mail.utexas.edu, penalva@uthscsa.edu

Content:

1 PROTEIN AND MRNA EXPRESSION DATA

1.1 Characteristics of mRNA expression data

Figure S1A. Accuracy (validity) of intensity based estimates of mRNA concentrations

mRNA concentrations based on signal intensity measured in single-channel microarrays (as we use in our analysis) correlate well with measurements from other methods that are assumed to be highly accurate: RNA-seq (A) and single-molecule sequencing (B)(Helicos)(Lipson *et al*, 2009; Nagalakshmi *et al*, 2008). Shown here is, for yeast, the RNA-seq (A) and Helicos (B) data measured for yeast (Lipson *et al*, 2009; Nagalakshmi *et al*, 2008) compared to affymetrix data (Holstege *et al*, 1998).

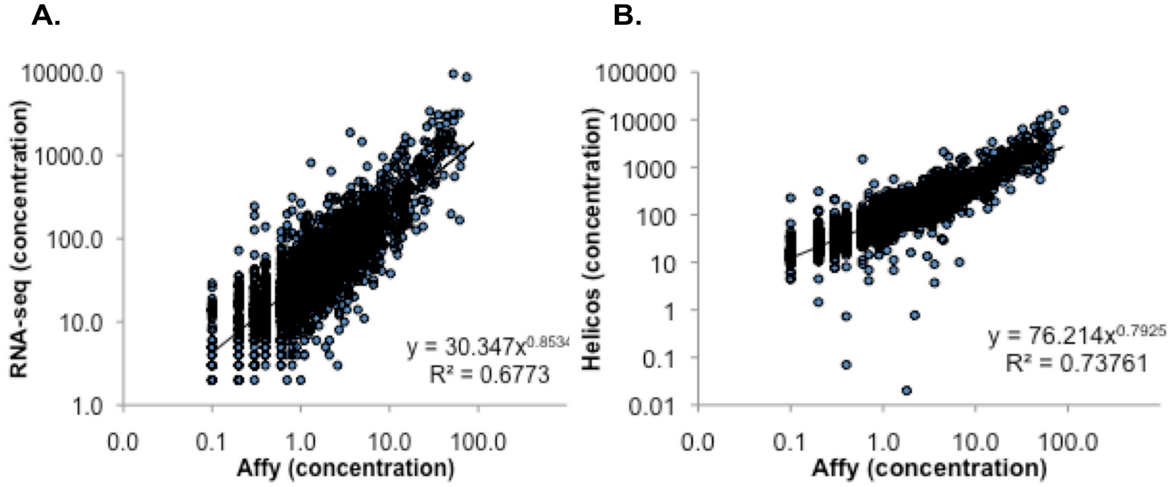


Figure S1B. Reproducibility (reliability) of mRNA concentration measurements

mRNA concentrations were measured in seven replicates from one biological sample (GSE20492, GEO database). The correlation between (log) intensities of replicate measurements lies between 0.96 and 0.99. An example is shown in the figure (GSM514930 vs GSM514931), with R^2=0.98 and R=0.99 (N=24,000).

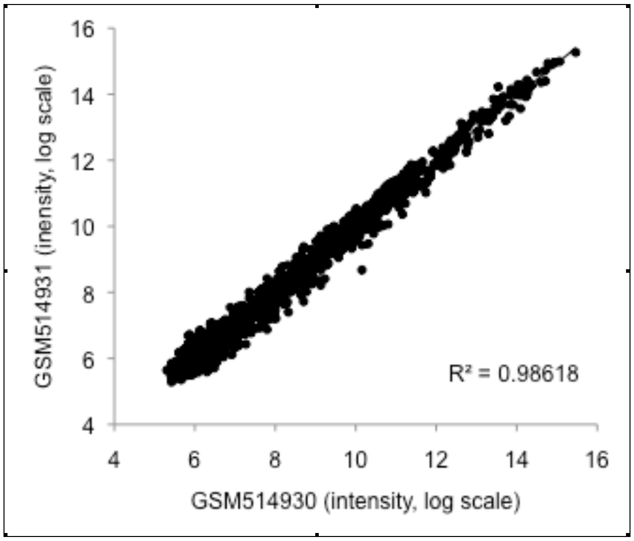
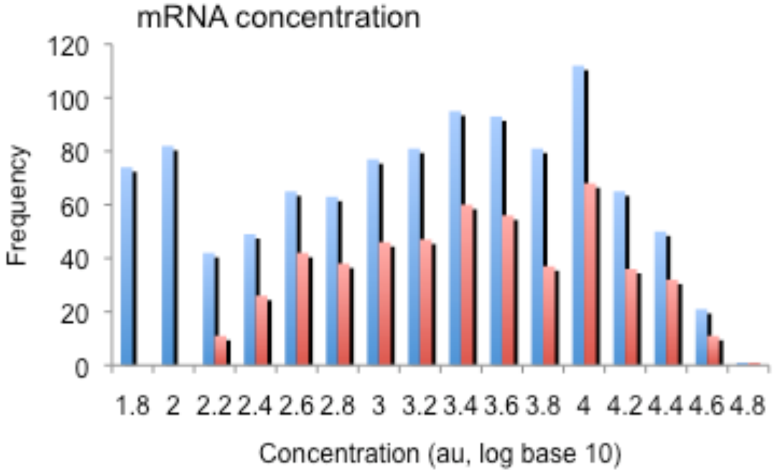


Figure S2. Frequency distribution of mRNA concentrations

To determine the gene expression profile of Daoy medulloblastoma cells, we used Nimblegen arrays. These arrays employ three probes per gene. Averaging over probe intensity accounts for probe-specific biases and improves accuracy of concentration estimates per gene. Moreover, Nimblegen uses TM (temperature of melting) balance to minimize differences between experiments and probes and to equalize hybridization conditions for each mRNA species in the sample. Thanks to these characteristics, we expect a reliable concentration estimates and consistent frequency distributions.

The mRNA data for 1025 proteins ('all', blue) showed an additional, smaller peak on the left attributed to technical artifacts. Thus for the high-confidence dataset (red, N=512), we removed genes with log(mRNA)<2 and proteins predicted to have a transmembrane helix (red). All data are supplied in the **Supplementary data** file (worksheet Data). au – arbitrary units



1.2 Characteristics of protein identification and spectral counting

Figure S3. Reproducibility

Spectral counting based methods of protein concentration measurements are highly reproducible: only 12 to 16% of the variation was accountable to noise in technical or biological replicates. Spectral counts are measured on a per-protein basis and vary by only 1-16% between injections (*not shown*). Shown here an example of two injections a, b correlating with two other injections c, d, with R^2=0.99 (linear) and R^2=0.84 (logarithmic), respectively (n>900).

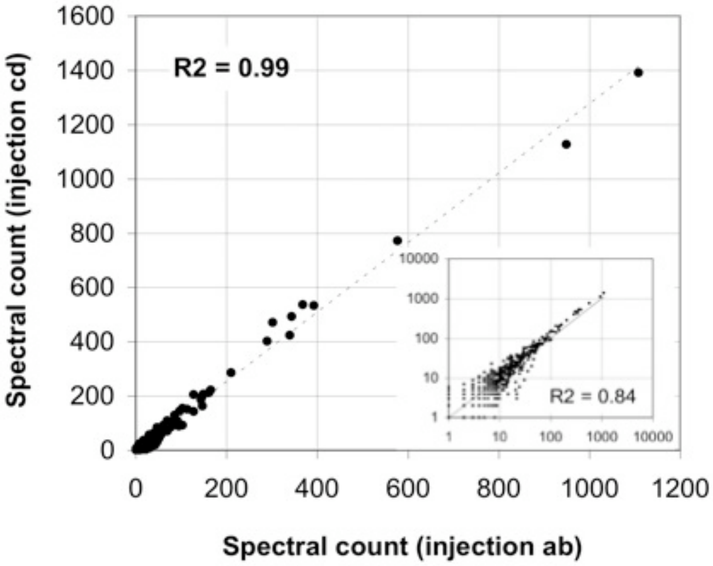
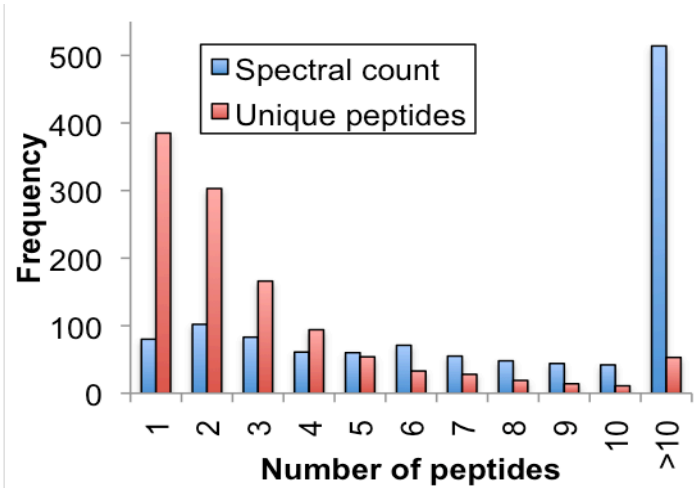


Figure S4. Single-peptide identifications

About one third of the 1,025 proteins in the Daoy dataset are identified by a single type of peptide (unique peptides = 1). However, in most of these cases, the peptide is observed multiple times (spectral count > 1), providing robust identification and quantitation of the corresponding protein. In <7% (80) of the cases, the spectral count equals 1, and the protein's identification relies on only a single hit. The large majority of proteins are identified and quantified by many peptides (72%, spectral count ≥ 5), demonstrating that the measurements of protein concentrations are based on robust data with multiple identifications per protein.



Protein quantitation

We developed a mass spectrometry based method, called APEX, to measure absolute protein concentrations in complex protein samples as well as to assess the statistical significance of differential protein expression (Lu *et al*, 2007; Vogel and Marcotte, 2008). While originally established in yeast, the following figures demonstrate APEX's accuracy in human cell systems. Two independent comparisons of the performance of different proteomics methods also supported the utility of APEX-based quantitation (Kuntumalla *et al*, 2009; von der Haar, 2008).

Figure S5. APEX can explain 84% of the variation in concentrations of a standardized mixture of 48 human proteins (UPS2, Sigma)

We tested a standard mixture of 48 human proteins in six different concentrations, spanning five orders of magnitude in concentration (*Sigma*, UPS2). The data was analyzed with our standard pipeline, using APEX and a decoy database of 6x48 shuffled protein sequences. We identified 43 of the 48 human proteins, detecting protein concentrations as low as 0.01 nM. Observed and expected concentrations correlated well (R^2=0.84, log scale). The log-average error, measured as fold-change between observed and expected concentration, is centered around 1.1 (10% error) with a standard deviation of 5.9-fold. APEX tends to slightly over-estimate protein concentrations in low concentrations.

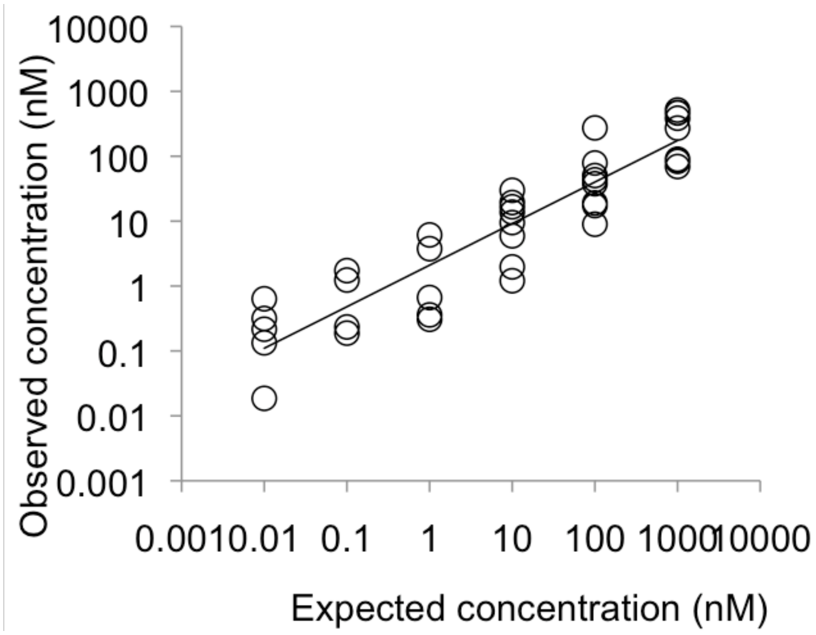


Table S1. APEX can explain 65% of the concentration variation of proteins spiked into human cell lysate background

To establish that concentration measurements are independent of the presence of a multitude of other molecules, we tested APEX for proteins spiked into a complex human cell lysate. We spiked nine non-human proteins into human K562 cellular lysate in three different mixtures, and retrieved concentration estimates for 26 of the 27 data points. The two measurements for GFP (Green Fluorescent Protein) may represent outliers, with >ten-fold error. Overall, we observe decent correlation of observed vs. expected concentrations (R^2=0.65 on log-scale; R^2=0.72 without GFP). The log-average fold-change error is 1.1±3.9 fold (1.1±2.8 fold without GFP), i.e., on average, the fold-error of logarithmic estimates is ~10%. Exp – expected, known concentration; Obs – observed concentration

		Mix 1		**Mix 2**		**Mix 3**	
Protein name	**Organism**	**Exp**	**Obs**	**Exp**	**Obs**	**Exp**	**Obs**
Ovalbumin	*Chicken*	**0.4**	2.0	**43.0**	29.5	**4.3**	20.6
Green fluorescent protein	*Aequorea victoria*	**0.3**	0.02	**0.3**	n/a	**0.3**	0.01
Streptavidin	*Streptomyces avidinii*	**3.6**	4.6	**3.6**	3.9	**3.6**	4.8
Ovotransferrin	*Chicken*	**99.6**	55.5	**99.6**	34.0	**99.6**	68.0
Lactalbumin	*Cow*	**4.4**	15.1	**174.8**	91.2	**8.7**	21.5
Lysozyme	*Chicken*	**12.2**	9.7	**0.6**	0.9	**60.8**	47.9
Lactoperoxidase	*Cow*	**4.2**	3.6	**4.2**	2.7	**4.2**	3.3
β-casein	*Cow*	**45.4**	19.1	**4.5**	5.3	**181.6**	36.8
β-lactoglobulin	*Cow*	**183.0**	105.8	**9.1**	23.9	**0.5**	12.4

Accounting for mass spectrometry biases

Electrospray ionization during mass spectrometry experiments is biased in terms of the amino acid composition of the observed peptides. During the APEX calculations we correct for these biases (Lu *et al*, 2007). To confirm that the amino acid biases observed in our analysis (see below, **Table 1** main text) are not due to the underlying mass spectrometry method, we conducted several tests.

A protein with long sequence produces more peptides than a short protein, increasing the probability of observing the spectra mapping to this protein. Thus, a pure spectral counting based protein quantitation overestimates the abundance of long proteins. Our quantitation method, APEX (Lu *et al*, 2007; Vogel *et al*, 2008), uses spectral counting, but adjusts for the number of peptides expected per protein using a correction factor called O_i. Further, hydrophobic peptides may have different propensities to ionize than peptides with charged amino acids, changing the spectral counts for proteins with hydrophobic peptides. APEX also corrects for the ability of peptides to ionize. The correction factor, O_i, consists of an estimate of the number of spectral counts based on protein length and based on peptide sequence properties.

Figure S6. The O_i value is independent of the protein-per-mRNA ratio given sequence length

The proteomics based correlation factor O_i accounts for both length and amino acid biases of the contributing peptides. If incorrect estimates of peptide ionization propensities were responsible for the amino acid biases observed in the main text (**Table 1**) one would suspect O_i to correlate with the protein-per-mRNA ratio (when correcting for the influence of sequence length). This is not the case (this figure). Thus, biases in the protein sampling introduced by mass spectrometry are accounted for in our quantitation method and do *not* cause the trends reported in **Table 1**.

Dataset: high-confidence data set (HCD) extracted measurements of protein concentrations in the Daoy medulloblastoma cell line (see methods, main text).

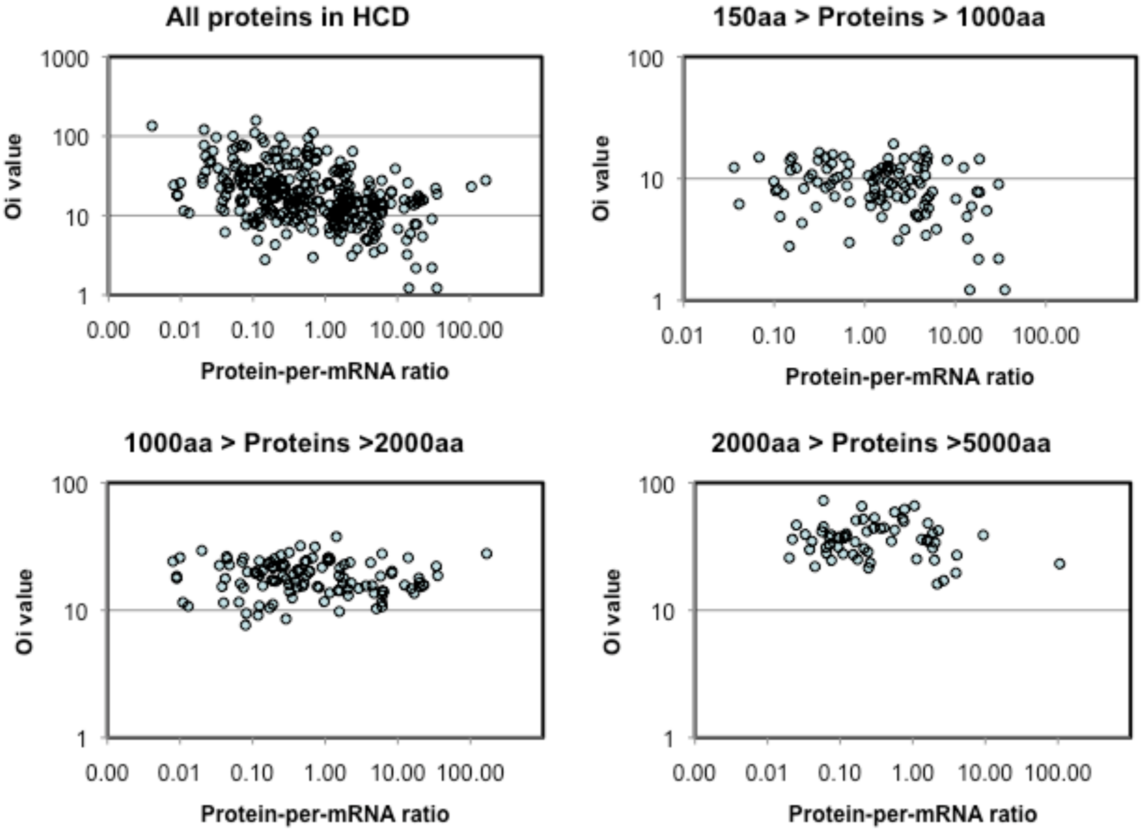


Figure S7. Differences in observed spectral counts are not only due to differences in sequence length

Long protein sequences produce more tryptic peptides (higher peptide count) than short sequences. Thus for proteins of equal concentrations we would expect a positive correlation between spectral or peptide counts and sequence length. APEX corrects for sequence length, and we observe a negative correlation between APEX-based protein concentrations and sequence length (see below). To test whether this negative correlation may be due to an over-compensation for length biases during the APEX process, we plot spectral counts vs. sequence length. Instead of a positive correlation, we observe no (if not a slightly negative) correlation between spectral counts and sequence length, implying that the positive correlation is counteracted by another opposing process, similar to what we observe for APEX-based concentrations and length. In other words, the mechanisms that result in short proteins being more highly abundant than long proteins are stronger than those that produce fewer peptide counts for short than for long proteins.

Dataset: high-confidence data set (HCD) extracted measurements of protein concentrations in the Daoy medulloblastoma cell line (see methods, main text).

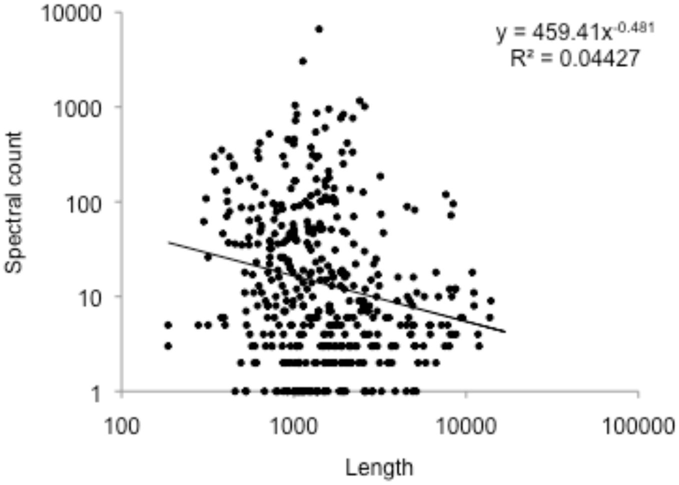
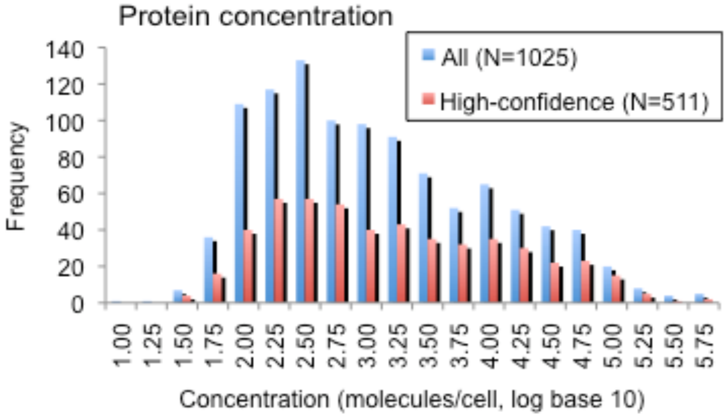


Figure S8. Frequency distribution of protein concentrations

Frequency distribution of protein concentrations measured for 1025 ('All') and 511 ('High-confidence dataset', HCD) proteins in the Daoy medulloblastoma cell line. Extraction of the HCD dataset is described in the methods, main text. All data are supplied in the **Supplementary data** file (worksheet Data).



2 SEQUENCE CHARACTERISTICS AND CORRELATION ANALYSIS

Table S2. Data and sources

The table lists all feature groups, their sources and a description of data preparation. All sequences stem from Ensembl v. 44.36f**.** The **Supplementary data file** reports correlations of all measures with protein and mRNA concentrations; **Table 1** (main text) reports significant correlations. To reduce redundancy between variables, we calculated the Spearman's rank correlation for each pair of variables (**Supplementary data file)**. We eliminated one variable from pairs if $|R_s|\geq0.9$ (high inter-correlation)**.** We kept variables whose correlation with *(log2) Protein abundance* was higher than the correlation with another variable. The **Supplementary data file** reports the inter-correlations that occurred between some of the measures.

Feature group / subgroup	**Comment**
Sequence lengths	R(seqinR) (Charif *et al*, 2005): we measure 3', 5', coding strand length, and the total length which is the sum of the former three lengths. We also report relative lengths (normalized by total length).
Nucleotide, dinucleotide, amino acid composition	R(seqinR) (Charif *et al*, 2005): residue frequencies are given as normalized by the total number of residues in the respective sequence part. As presented in (Karlin and Cardon, 1994), we searched for over/under-represented dinuclotides. We also searched for over-represented n-mers in the UTRs (Yoon *et al*, 2008), but the dataset was too small for meaningful statistical tests.
Amino acid properties	Relative frequencies of groups of amino acids and their physico-chemical properties calculated based on information from AAindex (http://www.genome.jp/aaindex/). Other calculations were done with R(seqinR) (Charif *et al*, 2005).
Codon Bias Index	Parsed using CodonW (CodonW): the higher the Codon Bias Index, the more biased is the codon usage in the sequence. Highly expressed proteins were taken from the protein expression data.
Codon usage	Perl scripts were used to assess codon usage. Number of tRNA genes per codon was obtained from (Lander *et al*, 2001).
G+C content	R(seqinR) (Charif *et al*, 2005): the higher the value, the higher the combined frequency of Guanine and Cytosine. Global G+C content = GC; G+C in the first codon position = GC1; G+C in the second codon position = GC2; and G+C in the third codon position = GC3. All normalized by length.
Upstream AUG; upstream Open Reading Frames in the 5' UTR	We used Perl scripts to parse the 5'UTR nucleotide sequence for a) AUG; b) AUG together with surrounding translation initiation site and in-frame STOP codon (uORF). The five to ten nucleotides surrounding the translation initiation site are thought to influence initiation efficiency (Kozak, 1987).
Degree of intrinsic unstructuredness of the protein	DisoPred (Ward *et al*, 2004): the larger the value for a given protein, the more intrinsically unstructured regions in its sequence (and the less stable it may be).
Secondary structures in 5' or 3' UTR	Using the Vienna RNA package (Gruber *et al*, 2008), we predicted the folding energy of the sequences. The smaller the energy, the more stable the secondary structures. Using the SEGFOLD/SIGSTB software (Le *et al*, 1990a; Le *et al*, 1990b), we predicted several measures of secondary structures in the 5' and 3' UTRs of the sequences locally (best score) and across the entire sequence part (mean), with respect to unusual folding regions (UFRs), significance score (*Sigscr*), thermodynamic stability, and the average folding energy. *Local*: we computed *Sigscr* for the first 100 (5'end) and last 150 nucleotides (3'end) in the sequence. The computation was completed by scanning the three fixed-length windows (20, 40, 60). For each sequence, we computed the local folding energy, its corresponding stability score (*Stbscr*) in the

	natural sequence and the corresponding *Sigscr* related to the randomly shuffled sample. Local unusual folding regions (UFRs) in RNA sequences closely correlate with RNA regulatory elements in gene expression. These UFRs often function as a binding target of cellular factors (Malim et al, 1989a, 1989b; Tiley et al., 1990; Wang et al., 1995; Bernstein et al., 1997; Akiri et al., 1998; Sella et al., 1999; Pozner et al., 2000; Chen et al., 2000; Yang et al., 2000; Kilav et al., 2004; Cencig et al., 2004; Yeh et al., 2008). The UFRs were assessed to have both highly statistical significance and thermodynamic stability in RNA sequence. *Mean*: the thermodynamic stability and statistical significance of RNA folding are assessed by *Stbscr* and *Sigscr*, respectively. The *Stbscr* is computed as the difference between the lowest free energy calculated for the segment sequence and the mean of the lowest free energies from all possible segments of the same size over the entire RNA sequence, divided by the standard deviation of the sample. The *Sigscr* is defined as the difference between the lowest free energy calculated for a segment of the real RNA sequence and the average of the lowest free energies of a large number of randomized segments with the same base composition and the same size divided by the standard deviation of the free energies from the random sample.
Poly-adenylation sites	Predicted using *polyadq* (Tabaska and Zhang, 1999) at http://rulai.cshl.org/tools/polyadq/polyadq_form.html.
miRNAs	First, we collected data on expression of miRNAs in the Daoy medulloblastoma cell line from www.microRNA.org. From the rank-ordered expression values, we selected the top 20 (50, 74, 90) most highly expressed miRNAs. Second, we used the *miRBase* (Griffiths-Jones *et al*, 2008) and *TargetScan* (Grimson *et al*, 2007) databases to obtain predicted targets for these miRNAs. All targets for the miRNAs of interest were selected if P-value<0.05. Target sequences were parsed for the frequency of all miRNAs predicted per 3'UTR, of miRNA families, and the frequencies of unique miRNAs and miRNA families, respectively. The higher any of the given numbers, the more miRNAs are predicted for this gene. Results presented here are for the 90 miRNAs of highest expression.
Protein stability index (PSI)	Extracted from reference (Yen *et al*, 2008). The higher the PSI, the more stable the protein.
mRNA decay rate	Extracted from reference (Yang *et al*, 2003) as molecules/hour. The larger the decay rate, the less stable the mRNA.
Phosphorylation sites	Extracted from PhosphoPep (Bodenmiller *et al*, 2008), a database of experimentally determined phosphorylated peptides. The higher the number of phosphorylated peptides per protein, the more likely is its multiple phosphorylation.
Ribosome attachment (as a measure of translation efficiency)	Extracted from reference (Mazan-Mamczarz *et al*, 2005), an analysis of nascent translation in human carcinoma cells exposed to short-wavelength UV light using cDNA microarrays. We used data for the untreated cells to analyze ribosome attachment in the following manner: the experiment was conducted in triplicate, with a collection and cDNA analysis of 10 to 12 fractions per replicate along the sucrose gradient. The more ribosomes attach to an mRNA, the later it elutes in the sucrose gradient; hence, later fractions contain translationally more active mRNAs than early fractions. For each mRNA within each replicate, we identified the fraction of its maximum elution using i) the raw microarray signal; ii) the array-normalized expression values or iii) rank-ordered expression values. Thus, the final data reports, for each mRNA, the (average) fraction in which it has its peak elution. While all three measures provided similar results, **Table 1** (main text) the correlation for rank-ordered expression (iii).
Protein function	Analyzed using the DAVID server (Huang da *et al*, 2007).
Kozak sequence	Using Perl scripts, we determined the nucleotide composition of the ten positions before and after the translation initiation codon AUG.

Number of alternative splice variants	Information on the number of different alternative splice variants were taken from two sources: the ASTD database at websites of the European Bioinformatics Institute (http://www.ebi.ac.uk/asd/), and the predicted splice information from the ENSEMBL gene prediction (which includes experimental data), using the human genome *Homo sapiens* v47.36.

Note on the use of protein-per-mRNA ratios

The ratio of protein-per-mRNA is an intuitive and often-used measure of translation efficiency and protein stability, e.g. (Nie *et al*, 2006; Wu *et al*, 2006). However, it is only correct to use (with linear regression methods) if protein and mRNA concentrations are linearly proportional: in this case, protein = a*mRNA, and protein/mRNA = a = constant.

Such linear relationship is not necessarily observed, as is shown in **Figure 2** (main text): the relationship between log(protein) and log(mRNA) can be described by a piece-wise linear function. Therefore, we conducted all individual correlation tests between features and protein expression (fixing mRNA) using a non-parametric correlation coefficient (Spearman rank). The combined analysis using a MARS model (see below) is piece-wise linear, thus approximating the non-linear case. We selected two populations (red, green in **Figure 2**, main text, and see Additional Results below) based on extreme protein-per-mRNA ratios. However, the selection procedure was also not based on the assumption of a linear relationship between protein and mRNA concentrations. Our analysis is thus independent of the exact mathematical relationship between protein and mRNA concentration.

Note on the use of partial correlation analysis

The use of partial correlation analysis has been criticized for its tendency to detect spurious correlations in noisy biological data, and principal component regression has been suggested as an alternative method, such as Principal Component Regression Analysis (PCRA)(Drummond *et al*, 2006). Though this suggestion is valid for investigation of linear relationships, we found that it is not useful for the non-linear relationships we observed in our data. A better way is to remove redundant features (highly correlated variables) using rank correlation and perform non-linear regression analysis such as MARS using selected variables. Since the principal components are linear combination of variables, PCRA models essentially the linear relationship between response and explanatory variables. Furthermore when the number of possible explanatory variables exceeds 130 (as in our case), the principal component itself may not be an easily interpretable quantity.

3 ADDITIONAL RESULTS

To obtain some of the results presented here, we compared characteristics (features) between the two extreme populations shown in **Figure 2** (main text).

P1: extremely low protein-per-mRNA ratio (red)

P3: extremely high protein-per-mRNA ratio (green)

Pt: all genes from the high-confidence dataset.

Table S3. Function enrichments

Functional analysis was performed using the DAVID's tool on functional annotation clustering (Huang da *et al*, 2007). Shown are over-represented functional clusters (Gene Ontology annotations only) in the two sets P1 (**A**) and P3 (**B**) compared to the high-confidence dataset. We applied a significance cutoff of 5% FDR, corresponding to approximately P-value<0.005.

A. Low number of proteins per mRNA (P1)

The genes include translation initiation factors EIF3C, D, F, M and EIF4B. The functional enrichment could not be reproduced when comparing P1 to the total set of genes (N=1025).

Term	Count	%	P-Value	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0003743~translation initiation factor activity	5	11	**0.003**	36	9	435	6.7	0.364	**0.36**	**3.84**

B. High number of proteins per mRNA (P3)

The set of enzymes in the first GO cluster includes: MDH1, PKM2, DLD, PGK1, TPI1, LDHB, LDHA, TXN, ETFA, MDH2, PDHB.

Term	Count	%	P-Value	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0006096~glycolysis	8	9	**0.001**	84	10	437	4.2	0.25	**0.25**	**0.76**
GO:0006006~glucose metabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0006007~glucose catabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0019318~hexose metabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0019320~hexose catabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0046365~monosaccharide catabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0046164~alcohol catabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0044275~cellular carbohydrate catabolic process	8	9	**0.002**	84	12	437	3.5	0.74	**0.49**	**3.54**
GO:0006091~generation of precursor metabolites and energy	11	12	**0.003**	84	22	437	2.6	0.83	**0.45**	**4.66**

Term	Count	%	P-Value	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0030530~heterogeneous nuclear ribonucleoprotein complex	7	8	**0.004**	75	10	400	3.7	0.40	**0.40**	**4.30**

Figure S9. Biases in translation initiation sites

We tested the two extreme sets of protein-per-mRNA ratios (**Figure 2**, main text, red and green) from the high-confidence dataset for biases in positional nucleotide composition around the translation initiation site. For each position, biases were assessed use the χ^2 test. For the high-confidence dataset (shown here), the composition in position -5 is significantly different between P1 and P3 (P-value<0.01). In 'all' data, the composition at position +4 is significantly different (P-value<0.01). The set of genes with low protein-per-mRNA ratios (P1) is enriched for Adenine at position +4, suggesting a sub-optimal translation initiation site.

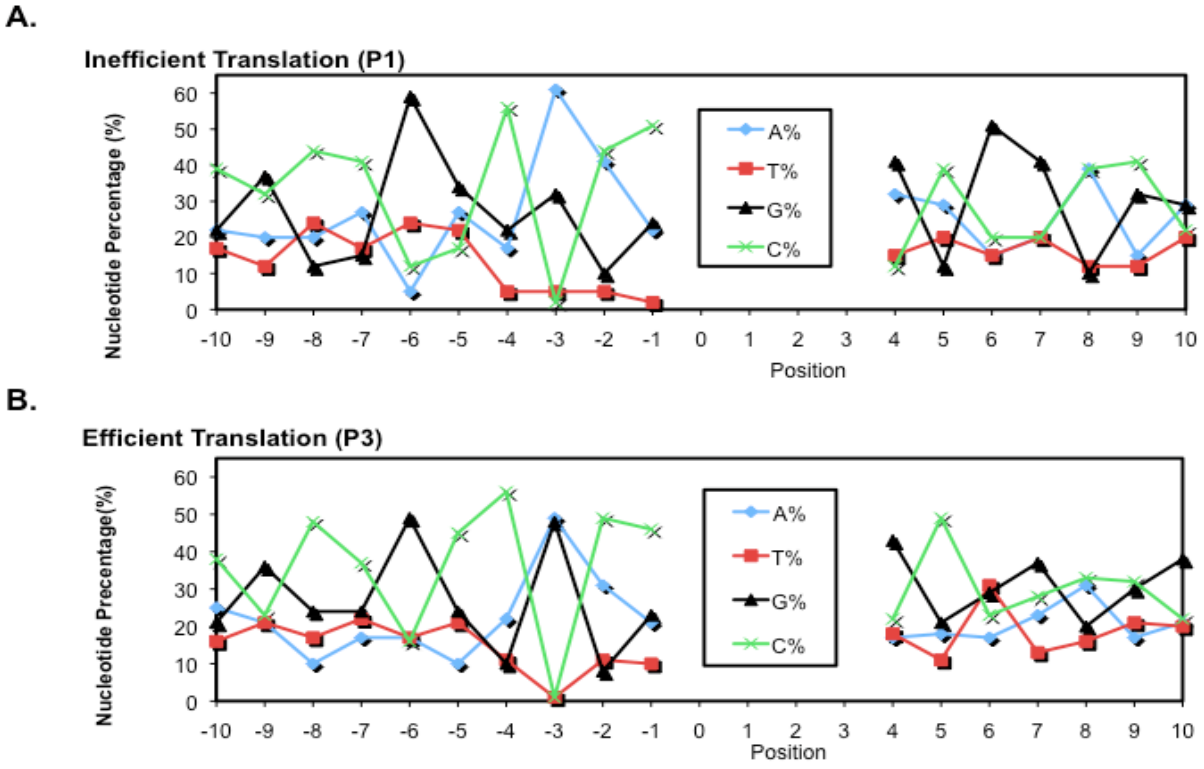


Figure S10. Codon usage: distributions

We assessed biases in several ways. We tested codon adaptation (CAI), codon bias (CBI), and the frequency of optimal codons (FOP) for their correlation with protein expression (**Table 1**). Second, we compared for all degenerate amino acids the distributions of codon usage in populations P1 and P3 using the χ^2 test (using absolute counts). Codon usage for Arginine, Serine, Proline, Leucine, and Glycine were significantly different between P1 and P3 (P-value<0.01). These biases were not explained by the number of tRNA genes for the respective codons (*not shown*).

P1	T			C			A			G		
T	**TTT**	0.50	PHE	**TCT**	0.22	SER	**TAT**	0.48	TYR	**TGT**	0.50	CYS
	TTC	0.50		**TCC**	0.20		**TAC**	0.52		**TGC**	0.50	
	TTA	0.09	LEU	**TCA**	0.16		**TAA**	0.38	TERM	**TGA**	0.39	TERM
	TTG	0.14		**TCG**	0.04		**TAG**	0.23		**TGG**	1.00	TRP
C	**CTT**	0.15	LEU	**CCT**	0.32	PRO	**CAT**	0.46	HIS	**CGT**	0.09	ARG
	CTC	0.17		**CCC**	0.29		**CAC**	0.54		**CGC**	0.17	
	CTA	0.08		**CCA**	0.31		**CAA**	0.29	GLN	**CGA**	0.11	
	CTG	0.38		**CCG**	0.08		**CAG**	0.71		**CGG**	0.20	
A	**ATT**	0.40		**ACT**	0.27	THR	**AAT**	0.52	ASN	**AGT**	0.16	SER
	ATC	0.44	LLE	**ACC**	0.33		**AAC**	0.48		**AGC**	0.21	
	ATA	0.16		**ACA**	0.30		**AAA**	0.46	LYS	**AGA**	0.24	ARG
	ATG	1.00	MET	**ACG**	0.10		**AAG**	0.55		**AGG**	0.19	
G	**GTT**	0.21	VAL	**GCT**	0.31	ALA	**GAT**	0.54	ASP	**GGT**	0.19	GLY
	GTC	0.21		**GCC**	0.38		**GAC**	0.46		**GGC**	0.32	
	GTA	0.13		**GCA**	0.24		**GAA**	0.46	GLU	**GGA**	0.26	
	GTG	0.45		**GCG**	0.08		**GAG**	0.54		**GGG**	0.23	

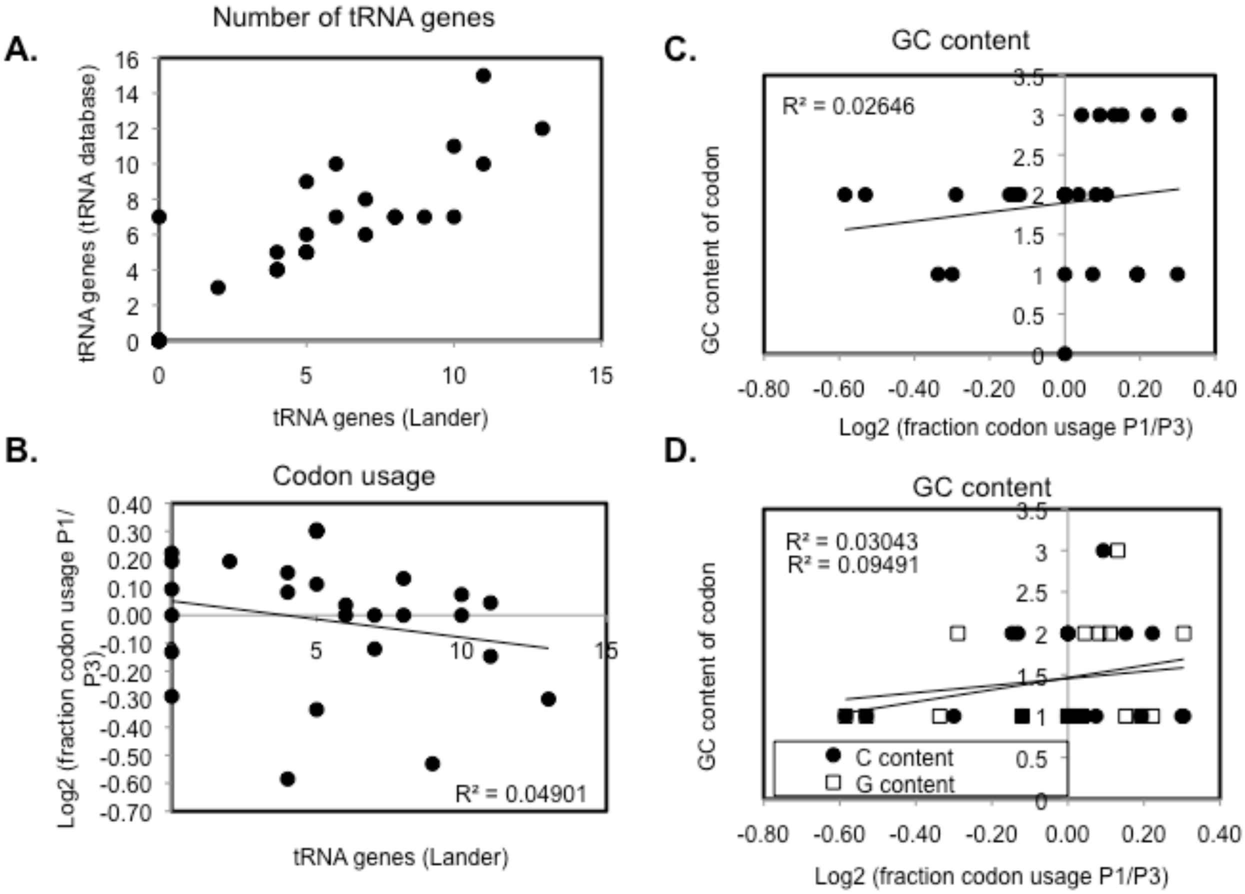
P2	T			C			A			G		
T	**TTT**	0.50	PHE	**TCT**	0.20	SER	**TAT**	0.48	TYR	**TGT**	0.48	CYS
	TTC	0.51		**TCC**	0.20		**TAC**	0.52		**TGC**	0.52	
	TTA	0.09	LEU	**TCA**	0.16		**TAA**	0.33	TERM	**TGA**	0.46	TERM
	TTG	0.14		**TCG**	0.05		**TAG**	0.21		**TGG**	1.00	TRP
C	**CTT**	0.14	LEU	**CCT**	0.31	PRO	**CAT**	0.45	HIS	**CGT**	0.10	ARG
	CTC	0.18		**CCC**	0.31		**CAC**	0.55		**CGC**	0.17	
	CTA	0.07		**CCA**	0.29		**CAA**	0.28	GLN	**CGA**	0.12	
	CTG	0.38		**CCG**	0.10		**CAG**	0.72		**CGG**	0.19	
A	**ATT**	0.39		**ACT**	0.27	THR	**AAT**	0.49	ASN	**AGT**	0.16	SER
	ATC	0.44	LLE	**ACC**	0.34		**AAC**	0.51		**AGC**	0.23	
	ATA	0.17		**ACA**	0.30		**AAA**	0.44	LYS	**AGA**	0.23	ARG
	ATG	1.00	MET	**ACG**	0.10		**AAG**	0.56		**AGG**	0.20	
G	**GTT**	0.20	VAL	**GCT**	0.29	ALA	**GAT**	0.50	ASP	**GGT**	0.18	GLY
	GTC	0.23		**GCC**	0.38		**GAC**	0.50		**GGC**	0.33	
	GTA	0.13		**GCA**	0.24		**GAA**	0.45	GLU	**GGA**	0.26	
	GTG	0.45		**GCG**	0.09		**GAG**	0.55		**GGG**	0.24	

P3	T			C			A			G		
T	**TTT**	0.50	PHE	**TCT**	0.22	SER	**TAT**	0.51	TYR	**TGT**	0.39	CYS
	TTC	0.50		**TCC**	0.23		**TAC**	0.50		**TGC**	0.61	
	TTA	0.09	LEU	**TCA**	0.13		**TAA**	0.48	TERM	**TGA**	0.39	TERM
	TTG	0.15		**TCG**	0.05		**TAG**	0.13		**TGG**	1.00	TRP
C	**CTT**	0.15	LEU	**CCT**	0.32	PRO	**CAT**	0.45	HIS	**CGT**	0.13	ARG
	CTC	0.17		**CCC**	0.29		**CAC**	0.55		**CGC**	0.17	
	CTA	0.08		**CCA**	0.29		**CAA**	0.27	GLN	**CGA**	0.12	
	CTG	0.37		**CCG**	0.10		**CAG**	0.73		**CGG**	0.16	
A	**ATT**	0.41		**ACT**	0.28	THR	**AAT**	0.48	ASN	**AGT**	0.15	SER
	ATC	0.44	LLE	**ACC**	0.34		**AAC**	0.53		**AGC**	0.22	
	ATA	0.14		**ACA**	0.30		**AAA**	0.44	LYS	**AGA**	0.26	ARG
	ATG	1.00	MET	**ACG**	0.08		**AAG**	0.56		**AGG**	0.17	
G	**GTT**	0.23	VAL	**GCT**	0.32	ALA	**GAT**	0.55	ASP	**GGT**	0.23	GLY
	GTC	0.22		**GCC**	0.36		**GAC**	0.45		**GGC**	0.31	
	GTA	0.13		**GCA**	0.23		**GAA**	0.48	GLU	**GGA**	0.27	
	GTG	0.43		**GCG**	0.09		**GAG**	0.52		**GGG**	0.20	

Pt	T			C			A			G		
T	**TTT**	0.50	PHE	**TCT**	0.21	SER	**TAT**	0.48	TYR	**TGT**	0.48	CYS
	TTC	0.50		**TCC**	0.20		**TAC**	0.52		**TGC**	0.52	
	TTA	0.09	LEU	**TCA**	0.16		**TAA**	0.36	TERM	**TGA**	0.44	TERM
	TTG	0.14		**TCG**	0.05		**TAG**	0.20		**TGG**	1.00	TRP
C	**CTT**	0.14	LEU	**CCT**	0.31	PRO	**CAT**	0.45	HIS	**CGT**	0.10	ARG
	CTC	0.18		**CCC**	0.30		**CAC**	0.55		**CGC**	0.17	
	CTA	0.08		**CCA**	0.29		**CAA**	0.28	GLN	**CGA**	0.12	
	CTG	0.38		**CCG**	0.09		**CAG**	0.72		**CGG**	0.19	
A	**ATT**	0.39		**ACT**	0.27	THR	**AAT**	0.50	ASN	**AGT**	0.16	SER
	ATC	0.44	LLE	**ACC**	0.34		**AAC**	0.50		**AGC**	0.23	
	ATA	0.16		**ACA**	0.30		**AAA**	0.45	LYS	**AGA**	0.23	ARG
	ATG	1.00	MET	**ACG**	0.10		**AAG**	0.56		**AGG**	0.19	
G	**GTT**	0.20	VAL	**GCT**	0.29	ALA	**GAT**	0.51	ASP	**GGT**	0.18	GLY
	GTC	0.22		**GCC**	0.38		**GAC**	0.49		**GGC**	0.33	
	GTA	0.13		**GCA**	0.24		**GAA**	0.45	GLU	**GGA**	0.26	
	GTG	0.45		**GCG**	0.09		**GAG**	0.55		**GGG**	0.23	

Figure S11. Codon usage: links to tRNA genes and GC content

The number of tRNA genes is a proxy of codon adaptation, and was taken from the human genome publication (Lander *et al*, 2001) and the tRNA database (http://gtrnadb.ucsc.edu/)(**A**). We mapped the log-base2 ratio of codon use in P1 vs. P2 to the number of tRNA genes for each codon (of significantly different amino acids, see **Figure S10**), but there was no obvious trend with respect to the preferred use of codons that have many tRNA genes in genes of large protein-per-mRNA ratios (P3)(**B**). We also could detect no bias in codon usage depending on GC content of the codons (**C, D**).



4 MARS ANALYSIS

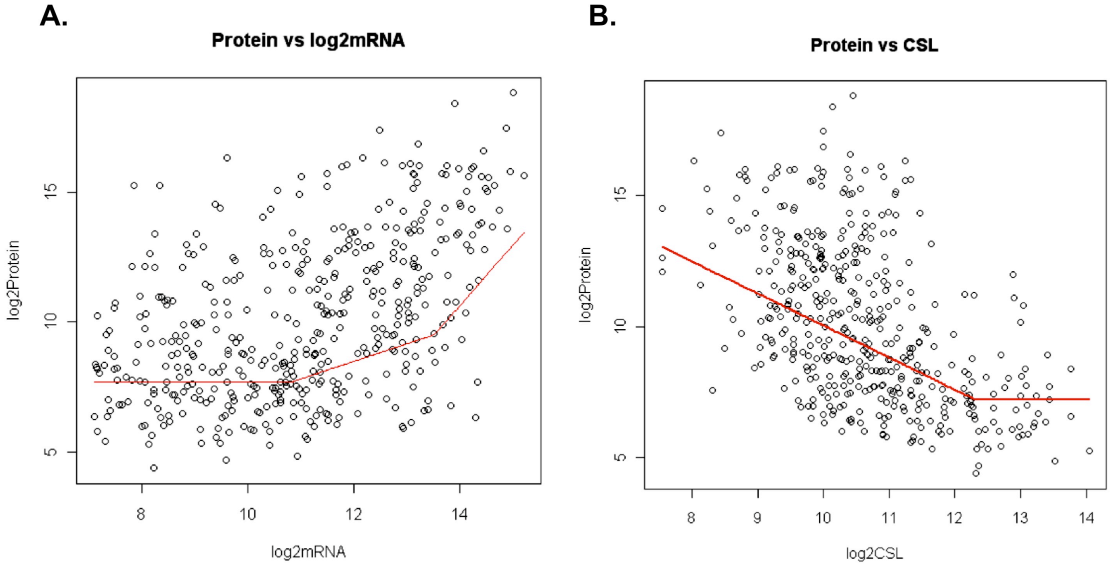
4.1 Data preparation

The data for the MARS analysis comprised an original set of 192 features which represent the explanatory variables (including “mRNA concentration” and sequence features), and “Protein abundance” as the response variable. To avoid extensive over-fitting, we applied stringent filters to the dataset before using it for the modeling. We removed all features that were redundant to other features with R_s>0.9 (see above). We log-transformed all features with extreme distributions ("*Total mRNA length", "Coding strand length", "5’UTR length", "3’UTR length"*). The MARS model with transformed variables resulted in a better fit than the model with untransformed variables and was henceforth used in the analysis (not shown). We removed features with >40 missing values, and all genes with >1 missing value. This procedure resulted in the dataset of 476 genes described by 133 features and no missing values (see **Supplementary data file**). Using different filters that resulted in more features (but fewer genes) resulted in strong over-fitting (*not shown*).

4.2 Model fitting

Multivariate Adaptive Regression Splines (MARS) is a method introduced by Friedman (Friedman, 1991) for solving regression-type problems, with the main purpose of predicting the values of a continuous response or outcome variable from a set of explanatory or predictor variables. Unlike linear regression models, MARS makes no assumption about the underlying functional relationship between the response and predictor variables and uncovers the functional relationships as well as the best predictor variables entirely from the data. MARS is particularly suited for our problem as the number of input dimensions is very high (133 features) and the functional relationships between the outcome and predictors are suspected to be non-linear. The non-linearity is exemplified in **Figure S12**, and consistent with the discrepancy between high Spearman’s rank correlation and low Pearson’s correlation coefficients.

Figure S12. Example of non-linear relationship between protein expression and mRNA.



In MARS, non-linear responses between protein expression and biological factors (variables, features) are described by a series of linear segments of differing slope, each of which is fitted using a basis function (Friedman, 1991; Hastie *et al*, 2001). In other words, in MARS the relationship between protein abundance *y* and selected feature variables $x_1, x_2, \ldots, x_k$ is modeled as

$$y= f_1(x_1) + f_2 (x_2) + \ldots +f_k (x_k) +error \qquad [1]$$

where $f_1, f_2, \ldots,$ and f_k are piecewise continuous linear functions. Each function f_j of the variable x_j is constructed by fitting basis functions to distinct intervals of the explanatory variables. In each function, piecewise linear segments, called splines, are smoothly connected together.

Fitting a spline basis function $B_i(x)$ both finds joining points of piecewise linear segments called knots and performs the 'smoothly' joining process. The fitting spline basis function is described by the following equations:

$$f_j(x_j) = \sum_{i=1}^{J} c_i B_i(x_j) . \qquad [2]$$

The basis function $B_i(x)$ has the form

$$B_i(\mathrm{x}) = \begin{cases} x - t_i & if\ x > t_i \\ 0 & otherwise \end{cases} \qquad [3A]$$

or

$$B_i(\mathrm{x}) = \begin{cases} t_i - x & if\ x < t_i \\ 0 & otherwise \end{cases} \qquad [3B]$$

where t_i is the knot placement.

The function $f(\mathbf{X})= f_1(x_1) + f_2 (x_2) + \ldots +f_k (x_k)$ can therefore be written as the weighted sum of basis functions

$$f_M(\mathbf{X}) = \sum_{j=1}^{k} \sum_{i=1}^{I_j} c_i B_i(x_j) \qquad [4]$$

where $M=I_1+\ldots+I_k$ is the total number of basis functions selected in the model. The MARS model selects the best set of basis functions (or equivalently knots t_i) from all possible candidate basis functions. The corresponding regression coefficients c_i will be estimated via regression. If a variable does not contribute to explanation of the variability in the response variable, the basis functions corresponding to the variable would be dropped from the model during the model building procedure.

4.3 Implementing MARS (model building)

Implementing MARS, i.e. the selection of knot positions and variables, involves a two-step procedure that is applied successively until a desired model is found. In the first step, we increase complexity by adding basis functions (variable and knot positions) until a maximum level of complexity has been reached. In the second step, we perform a backwards procedure to remove the least significant basis functions from the model with respect to the goodness of fit measure, the mean residuals of Generalized Cross-Validation (GCV, see below).

This algorithm is implemented as follows:

1. Start with the simplest model involving only the constant basis function.
2. Search the space of basis functions, for each variable and for all possible knots, and add those variables/knots which minimize prediction error.

3. Recursively apply step 2 until a model of pre-determined maximum complexity is achieved.
4. Finally, apply a pruning procedure where those basis functions are removed one by one that contribute least to the GCV mean residuals until the GCV error reaches its minimum.

Given a data set X containing n objects and p explanatory variables, note that if all of the input values are distinct, there would be $N=n\times p$ possible pairs of spline basis functions, including knot locations t_{ij} $(i=1, 2, \ldots, n; j=1, 2, \ldots, p)$. In such a model, steps 1- 3 lead to a very complex and over-fitted model. Although the model would fit the data well, it has poor predictive abilities for new objects (low generality).

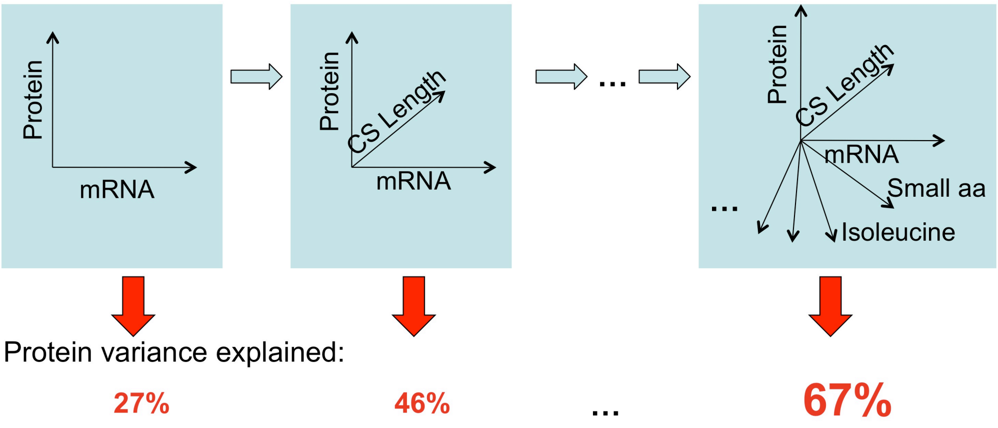
To improve the prediction (generality), the redundant basis functions are removed one at a time using a backward stepwise procedure in step 4. To determine which basis functions should be included in the model, MARS uses GCV (Craven and Wahba, 1979) as the overall goodness of fit. The Generalized Cross Validation criterion is a measure of the goodness of fit that takes into account not only the residual error but also model complexity. It is defined as the mean squared residual error divided by a penalty dependent on the model complexity. The GCV criterion for the model f_M is defined in the following way:

$$GCV(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}_M(\mathbf{X}_i))^2}{(1-\frac{C(M)}{n})^2}, \qquad [5]$$

where $C(M) =(M+1)+d*M$ is a complexity penalty that increases with the number of basis functions in the model. (M is the total number of basis functions selected in the model, as above.) The parameter d is a penalty for each basis function included into the model. Large values of d lead to fewer basis functions and therefore smoother function estimates. We adapted Hastie *et al.*'s (Hastie *et al*, 2001) recommendation that uss $d=2$ for the additive MARS model. The Generalized Cross Validation uses a formula to approximate the error that would be determined by leave-one-out validation. GCVs were introduced by Craven and Wahba (Craven *et al*, 1979) and extended by Friedman (Friedman, 1991) for MARS.

Figure S13. MARS model – simplified graphic representation

The graph illustrates how, by adding more and more variables, we explain variance in protein concentrations. The arrows represent 'dimensions' of the model, not linear relationships. Note that the R^2=0.27 reported in this diagram differs slightly from the R^2=0.29 reported in the main text, as the underlying datasets are different, using N=511 and N=476 (see **Section 4.1.**), respectively.



The importance of the variables in the MARS model

In the MARS model, the importance of each selected predictor variable is evaluated as its contribution to the goodness of fit of the model defined as $\sum_{i=1}^{n}(y_i - \hat{f}_M(\mathbf{X_i}))^2$, the residual sum of squares (RSS). The scoring of the importance of variables in the MARS model is similar to the leave-one-out cross-validation concept. To calculate scores of variable importance, MARS refits the model after deleting all terms involving the variable at issue and calculating the reduction in goodness of fit. The importance of the variables is a relative measure and scaled between 0 and 1. The most important variable is the one that, when dropped, decreases the model fit the most and it receives the highest score, i.e. 1. The less important variables receive the lower scores, which is the ratio of the reduction in goodness of fit of these variables to that of the most important variable.

Once the importances of variables are ranked we can evaluate the percentage of variance explained by each variable as follows. We define the "Cumulative variance explained (%) by the model" as

(Deviance of the null model – deviance of the model) / Deviance of the null model * 100,

where the deviance of the null model (model with no variable) is defined as $\sum_{i=1}^{n}(y_i - \bar{y})^2$ and the deviance of the model is defined as $\sum_{i=1}^{n}(y_i - \hat{f}_M(X_i))^2$, i.e.,

$$\text{Cumulative variance explained (\%)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{f}_M(\mathbf{X_i}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \text{*100 (\%)} \qquad [6]$$

Comparing the null model and the model with the most important variable (mRNA level), the cumulative variance explained (%) is the contribution of mRNA. The difference of the cumulative variance explained by the model with mRNA alone and the cumulative variance explained by the model with two variables, mRNA and the second most important predictor (Total length) is the variance explained by the total length, etc. We can add up each variable one-at-a-time to evaluate the additive variance explained by the added variable to the model.

The Generalized R-Square as a proxy of cross-validation

The term $\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{f}_M(\mathbf{X_i}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ in equation [6] is also called R-Square (R^2) of the model. It measures how well the model fits the training data. The predictive power of the model is estimated by Generalized R-Square ($GRSQ$) of the model defined as

$$GRSQ = 1 - \frac{GCV(M)}{GCV(0)} \qquad [7]$$

where $GCV(M)$ is defined in equation [5].

When testing a model, we generally want to test its performance with independent test data, not with the training data. However, such an independent test data is often unavailable, and so we resort to cross validation or leave-one-out methods which can be painfully slow. As an alternative, for models such as MARS or the Generalized Linear Model we can use a formula to approximate the error that would be determined by leave-one-out validation — and that approximation is the GCV. The formula adjusts the training residual sum of squares (RSS) to account for the flexibility of the model.

In sum, the GCV approximates the RSS that would be measured in independent data. Even when the approximation is not that good, it is usually good enough for comparing models during pruning. GCVs were introduced by Craven and Wahba (Craven *et al*, 1979), and extended by Friedman for MARS (Friedman and Roosen, 1995; Hastie *et al*, 2001). The GRSQ*100% measures the cumulative percentage of variance explained by the model to the test data in the same way that R-Squared*100% measures the cumulative percentage variance explained by the model to the training data.

All MARS models were fitted in R (Team, 2004) using functions contained in the 'earth' library (Milborrow, 2009).

Higher-order models

The MARS analysis offers a similar level of performance compared with other non-linear modeling techniques such as Generalized Additive Model (GAM, *not shown*). We chose MARS because the cross-validated selection procedure in MARS models selects a best set of variables and their functional forms while the GAM estimates unknown functional forms for a given, fixed set of variables. A more complex second order MARS model includes all the significant second order interaction terms. A second order MARS model was compared to our additive MARS model without any interaction term. The second order MARS model improved the first order model by about 5% in R^2 value (*not shown*), but the model was much harder to interpret so was not included in this study.

4.4 MARS model results

MARS selected 25 sequence variables (36 basis functions) including the variable *log₂(mRNA) abundance* with importance in explaining the variation of *log₂(Protein abundance)*. The RSq (R^2) = 0.67 and Generalized Cross Validated RSq (GRSq) = 0.55 (**Supplementary data file**).

Table S4A. MARS model – detailed results

Cumulative variance explained (%) by the MARS model includes, for each variable, the listed variable and all variables listed above the variable. For example for feature *AA Small* the model includes *log₂mRNA, log₂CSLength,* and *Phosphorylated*. It explains 48% of variability of the protein expression. The variance explained by a variable X_k is defined as the Cumulative variance explained (%) by the model with variables X_1, X_2, …,X_k minus the Cumulative variance explained (%) by the model with variables X_1, X_2, …,$X_{k-1.}$ In our analysis, mRNA is X_1 and the effects of other variables are estimated after adjusting the effect of mRNA first.

For the evaluation of the combined contributions of different groups of features (**Figure 3B, main text; Figure S14**), we predicted the cumulative variance explained using all features of the respective grouping. More detailed results are provided in the **Supplementary data** file. Abbreviations: AA – amino acid frequency/property. Di-nucleotides listed denote the sequential order of the nucleotides, i.e. CA does not equal AC. For further source and explanation of variables, see **Table S2**.

Feature	Cumulative variance explained	Grouping 1	Grouping 2
$\log_2$mRNA	26.9	mRNA	
$\log_2$ CSLength (length of coding region)	46.0	Coding sequence	Length
Phosphorylated (yes, no)	48.2	Coding sequence	Amino acids
AA Charge	50.5	Coding sequence	Amino acids
GC content in 3'UTR	52.9	3'UTR	Nucleotides
Local secondary structures 3'UTR_3end_window40_FoldingEnergy_best	54.4	3'UTR	Nucleotides
AA Small	55.5	Coding sequence	Amino acids
CA content in coding region	55.6	Coding sequence	Nucleotides
Targetscan90 miR families (3'UTR)	56.5	3'UTR	Nucleotides

Targetscan20 miR families (3’UTR)	58.0	3'UTR	Nucleotides
C concent in 3’UTR	58.8	3'UTR	Nucleotides
CAI (codon adaptation index)	59.6	Coding sequence	Nucleotides
GC content in coding region	60.0	Coding sequence	Nucleotides
GC content at codon position 1	60.7	Coding sequence	Nucleotides
AA relative R content	61.3	Coding sequence	Amino acids
AC content in 5’UTR	61.9	5'UTR	Nucleotides
AA Basic	62.7	Coding sequence	Amino acids
AG content in 3’UTR	63.4	3'UTR	Nucleotides
AA relative E content	63.7	Coding sequence	Amino acids
AA Isoelectric Point	64.7	Coding sequence	Amino acids
AA relative I content	65.3	Coding sequence	Amino acids
GC content in 5’UTR	65.7	5'UTR	Nucleotides
TC content in 5'	66.2	5'UTR	Nucleotides
GA content in coding region	66.8	Coding sequence	Nucleotides
AA content in 3’UTR	67.5	3'UTR	Nucleotides

Table S4B. MARS model – P-values

It is not possible to formally obtain P-values on the MARS R^2, but informal P-values can be obtained and should be interpreted differently from the traditional P-values of parameters of linear model. The P-value is the significance of the selected variable with respect to the estimated functional form. It is interpreted as ‘informal’ since we assume that the functional form is known rather than estimated. However, the functional forms of the features are the estimated forms from the MARS model.

Given the functional form of the feature, an F-statistic is calculated as FSTAT=(difference of deviance/df1)/(deviance of the full model/df2). The numerator degree of freedom df1 is the number of basis functions used in estimating the functional form of the feature in the MARS model. The denominator degree of freedom df2=439 is the residual degree of freedom of the full model. The null distribution of FSTAT is the F-distribution with the degrees of freedom df11 and df2, and the calculated P-value shows how individual features contribute to the MARS model. Thus the P-values of selected features in the MARS model are calculated as the difference of deviances between the model with all the features and the model with one feature eliminated.

For explanation of the feature name and source, please refer to **Table S1, Table S4A**, and the **Supplementary data file**.

	Deviance			**P-**
Feature	**Difference**	**df1**	**FSTAT**	**Value**
log2mRNA	469.28	1	142.1	1E-28
log2CSLength	592.73	2	89.74	2E-33
Phosphorylatedyes1	88.66	1	26.85	3E-07
AA_Charged	41.02	1	12.42	5E-04
gc_3UTR	68.15	3	6.88	2E-04
Local_3UTR_3end_window40_FoldingEnergy	67.79	2	10.26	4E-05
AA_Small	119.03	2	18.02	3E-08
CS_ca	77.41	1	23.44	2E-06
Targets90_TOTAL.MIRFAMS.3UTR	71.08	2	10.76	3E-05
Targets20_TOTAL.MIRFAMS.3UTR	65.81	1	19.93	1E-05
UTR3_C_nt	76.64	1	23.21	2E-06
CAI	164.05	2	24.84	6E-11

CS_GC	115.97	2	17.56	5E-08
CS_GC1	73.12	1	22.14	3E-06
AA_R_rel	58.86	2	8.91	2E-04
ac_5UTR	35.87	1	10.86	1E-03
AA_Basic	76.58	2	11.6	1E-05
ag_3UTR	64.63	1	19.57	1E-05
AA_E_rel	40.82	1	12.36	5E-04
AA_IsoelectricPoint	70.3	1	21.29	5E-06
AA_I_rel	23.26	1	7.04	8E-03
gc_5UTR	30.3	1	9.17	3E-03
tc_5UTR	20.89	1	6.33	1E-02
CS_ga	30.53	1	9.24	3E-03
aa_3UTR	29.53	2	4.47	1E-02

Summary of MARS model predictions and discussion of generality

Our MARS model results in an R^2=0.67 between observed and predicted protein concentrations (also see: **Figure 3A,** main text). Below we discuss the generality of this fit, i.e. the applicability to other datasets.

First, the modeling procedure also reports the Generalized R^2 estimating the correlation between observed and predicted protein concentrations penalizing for the number of variables (features) needed to get an optimal fit (see above). A large discrepancy between the R^2 (0.67) and the generalized R^2 implies over-fitting, as the high R^2 may only be due to selection of many variables. Our model reports a generalized R^2=0.55 which suggests a moderate generality of the model.

Second, we tested the model performance when reducing the number of selected features from 25 to a smaller number (setting the 'prune' option), assuming that the top-performing features are the most important ones and will also have predictive power in other datasets (see **Supplementary data file**). Restricting the model to 11 features, the resulting R^2=0.57 (and generalized R^2=0.52) suggest an even lower degree of over-fitting than for the whole model.

Third, we conducted ten-fold cross-validation, partitioning the dataset randomly into ten sets of equal sizes, analyzing model performance in each dataset independently, and then combining the results. We several types of cross-validated R^2 values in addition to the generalized (R^2=0.55). (i) A ten-fold cross-validated R^2 value based on the 25 variables selected in the model of the complete dataset and their estimated functional forms (61%). (iii) An R^2 based on the optimally selected terms fixing the number of terms to be selected in each of the ten model building steps to ~10 to 15 (using the nterm function), i.e. using the optimally pruned model. The R^2 obtained was 45% confirming that pruning reduces dataset-specific over-fitting. (iv) A complete cross-validation in which the MARS model newly selects variables and functional forms during each validation step (R^2=30%).

In sum, we can predict 67% of the protein abundance variation in our dataset, and achieve decent generality of the model, suggesting an ability to predict ~30-60% of the protein abundance variation in the generalized case, i.e. with other test data.

Figure S14. Explanation of variation in protein abundance by different feature groups

Using the complete MARS model, we assess the variance in protein abundance explained by different subsets of the features listed in **Table S4**. (Note that we do not simply add up percentages of the contribution, but used the glm() function to calculate combined results.) The percentages vary slightly depending on the groupings due to the nature of the calculations. Feature groupings are listed in **Table S4A. A**. Features grouped into length, amino acid or nucleotide characteristics, calculated in cumulative manner, we start with the strongest individual feature group and then successively add other feature groups, examining the difference in explained protein abundance variation. On their own, each individual feature group may explain an even higher proportion of variation, e.g. *Sequence Lengths* explain, on their own, ~30% of protein variance (see **B**). **B**. Individual contributions of different feature groups to explanation of protein abundance variation, calculated for each group separately.

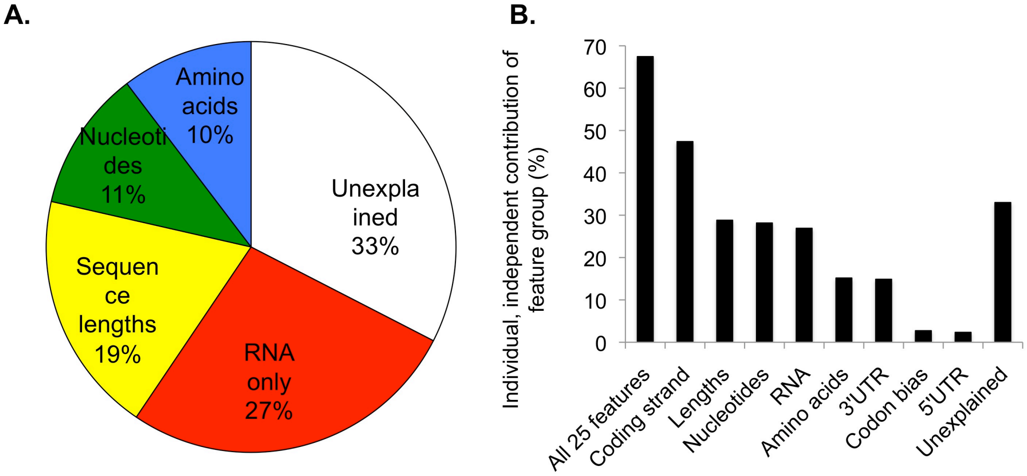
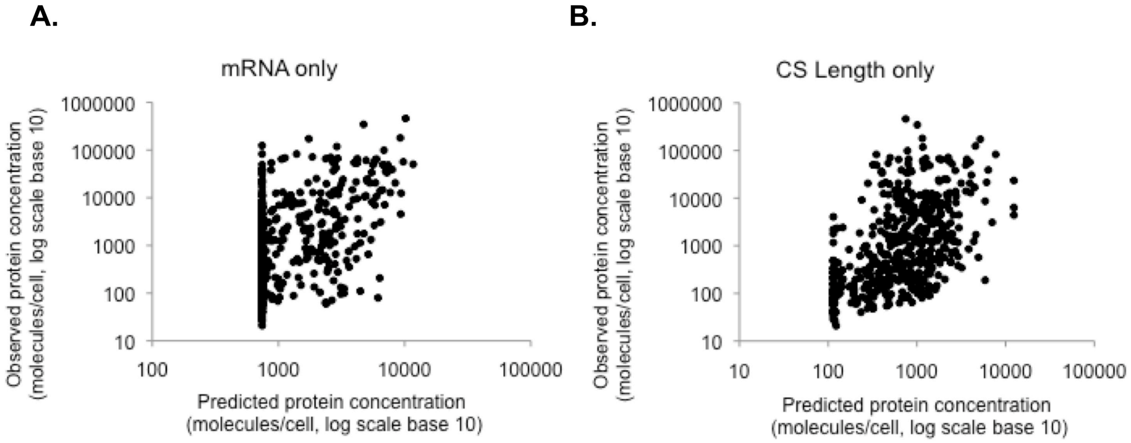


Figure S15. Prediction of protein concentrations

The graphs show the predicted protein concentrations using the MARS model and as only variables mRNA concentration (**A,** R^2=0.27) and Coding sequence length (**B**, R^2=0.27), respectively.



4.5 Alternative methods

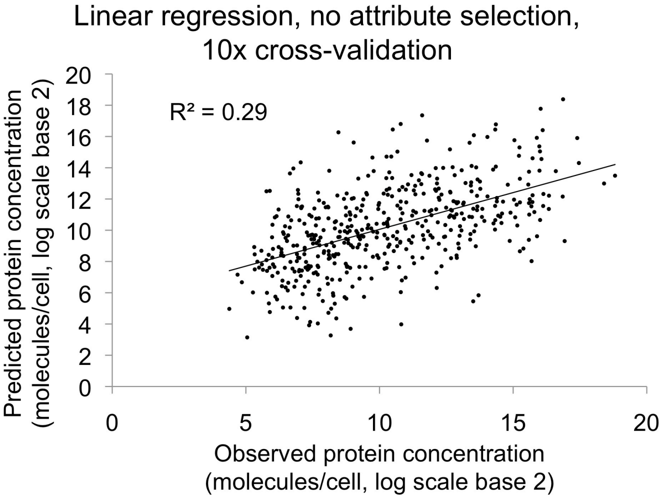
In addition to the MARS analysis, we tested the performance of linear regression, including the least absolute shrinkage and selection operator (LASSO), and principal component analysis combined with regression analysis. They are used to identify features that best explain variation in protein concentration. All analyses were conducted with the same dataset as used for the MARS model above, comprising 476 genes and 133 features.

4.5.1 Linear regression

Linear regression models linear relationships between the response variable (*log2 protein concentration*) and the *p* predictive variables in the form $y=\beta_0+\beta_1*x_1+\beta_2*x_2+\ldots+\beta_p*x_p+$ error. When p is large (i.e. >40), the model usually overfits the data and gives poor performance in prediction on other independent data. In such cases, we often would like to determine a smaller subset that exhibit the stronger relationship to the response variables. We use two methods, the subset selection (A) and the LASSO (B) (see below).

Figure S16. Linear regression (no variable selection)

The figure shows the result of linear regression without variable selection, i.e. this model is likely subject to over-fitting. The cross-validated prediction gives R^2=0.29. We conducted the analyses using the WEKA software version 3.4.13 (Hall *et al*, 2009)(http://www.cs.waikato.ac.nz/ml/weka/).



(A) AIC based subset selection method (linear regression)

The Akaike's information criterion (AIC) is a measure of the goodness of fit of the estimated linear model (Akaike, 1974). In the linear model defined above, it is defined as AIC = 2k+n[ln(RSS)] where RSS = $\sum_i (y_i-\beta_0+\beta_1*x_{1i}+\beta_2*x_{2i}+\ldots+\beta_k*x_{ki})^2$, the residual sum of squares.

The AIC based best subset selection method selects a subset $\{x_1,\ldots,x_k\}$ of size k from the p predictor variables that minimizes the AIC criterium in a stepwise manner. Increasing the number of predictors improves the goodness of fit, regardless of the number of free parameters in the data generating process. It also increases a penalty which is an increasing function of the number of estimated parameters. This penalty discourages over-fitting, and the preferred model is the one with the lowest AIC value. The AIC method attempts to find the model that best explains the data with a minimum number of predictors.

When the number p of predictors is large, i.e. >40, the number of subsets to be tested (2^p -1) is computationally infeasible. The difficulty may be overcome by starting with the intercept, and then sequentially adding to the model features that most improves the fit (*forward stepwise selection*). Alternatively, once can start with the full model, and sequentially delete the predictor that has the least impact on the fit (*backward*). We use the stepAIC function in R library MASS that implements hybrid stepwise selection strategies that consider both backward and forward moves at each step, and select the best out of the two methods. This stepwise procedure selected 42 features out of 133 as the best subset.

The observed R^2 =0.60 (R^2=0.51 in ten-fold cross-validation) for the 42 selected variables (linear model). For comparison, the MARS procedure selected 25 features with a Generalized R^2 = 0.67, and thus performed much better than the linear model which also used more variables.

Table S5. Linear regression with subset selection

		Std.	
Feature	**Estimate**	**Error**	**P-value**
log2CSLength *	-1.1938	0.1086	5.64E-25
log2mRNA *	0.3657	0.0584	9.37E-10
AA_G_rel	37.3118	8.0349	4.54E-06
CS_ta	-5.0533	1.1779	2.21E-05
CS_GC2	-18.6108	4.3637	2.46E-05
AA_L_rel	-27.7821	6.7241	4.32E-05
Phosphorylatedyes1 *	0.9172	0.2307	8.23E-05
ac_5UTR *	1.2774	0.3326	0.0001
CS_tc	-3.7125	0.9865	0.0002
Local_3UTR_3end_window40_Stability_best_Table10	-0.4098	0.1105	0.0002
CS_cc	-2.8951	0.8235	0.0005
CS_G_nt	-25.7544	7.4195	0.0006
CS_at	3.2288	0.9833	0.0011
Mean_3UTR_3end_window40_Energy_stdev_Table22	-0.5252	0.1634	0.0014
Targets90_TOTAL.MIRFAMS.3UTR *	61.9111	19.2621	0.0014
AA_solventwater_NOZY710101	13.6642	4.3067	0.0016
CS_tt	-2.2154	0.7609	0.0038
Targets20_TOTAL.MIRFAMS.3UTR *	-64.6502	22.7219	0.0046
AA_M_rel	-35.7507	12.8379	0.0056
AA_betasheet_CHOP780202	-19.4044	7.0039	0.0058
UTR3_C_nt *	-5.7528	2.2020	0.0093
CS_ag	2.7352	1.0632	0.0104
ag_5UTR	0.5885	0.2303	0.0110
AA_partitionenergies_GUYH850105	10.6742	4.5055	0.0183
AA_D_rel	-20.1999	8.6256	0.0196

gc_5UTR *	0.8745	0.3750	0.0201
AA_Small *	16.2866	7.1401	0.0230
Mean_3UTR_5end_window40_Stability_stdev_Table21	-1.2150	0.5444	0.0261
logAS_EBI	0.6721	0.3165	0.0343
Mean_3UTR_5end_window40_Energy_stdev_Table21	0.3717	0.1837	0.0436
ag_3UTR *	0.8215	0.4110	0.0463
CAI *	3.1794	1.7904	0.0765
log2UTR3Length	-0.1132	0.0642	0.0786
AA_Aromatic	-12.5713	7.3370	0.0874
logAS_SF	-0.6356	0.3932	0.1067
AA_betaturn_CHOP780203	-14.6160	9.0882	0.1085
AA_IsoelectricPoint *	0.1398	0.0892	0.1179
at_5UTR	0.2471	0.1636	0.1317
Potential_PEST	0.0818	0.0547	0.1352
ga_3UTR	-0.5338	0.3670	0.1465
Mean_3UTR_3end_window40_Significance_avg_Table22	-0.3161	0.2259	0.1624
tc_5UTR *	0.3519	0.2597	0.1762

(B) LASSO method (linear regression)

The least absolute shrinkage and selection operator (LASSO) is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Given a set of explanatory variables x_1, x_2 ...x_p and a response variable y, the LASSO method fits a linear regression model $y=\beta_0+\beta_1*x_1+\beta_2*x_2+\ldots+\beta_p*x_p$. The criterion it uses is:

$$\text{minimize RSS} = \sum_i (y_i-\beta_0+\beta_1*x_{1i}+\beta_2*x_{2i}+\ldots+\beta_k*x_{ki})^2 \text{ subject to } \sum_k |\beta_k| \leq t.$$

The first sum is taken over observations (cases) in the dataset. The bound t is a tuning parameter. When t is large, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on x_1, x_2, ..., x_p. However for sufficiently small t, some coefficients become exactly zero, resulting in LASSO performing a kind of continuous subset selection. Selecting t is similar to choosing the number of predictors to use in a regression model, and cross-validation is a good tool for estimating the best value for t.

In this study, we use $s = t / \sum_k |\beta_k|$ as a 'relative' tuning parameter. We use ten-fold cross-validation to estimate the best tuning parameter of s=0.05 which produces the R^2 value of 0.45 with 19 selected features. The LASSO fit for the whole data produces R^2=0.50 (no cross-validation). To evaluate the relative importance of selected variables, we standardized each feature variable by subtracting its mean and dividing it by its standard deviation first and fit the LASSO.

Table S6. LASSO

The table shows the 19 selected features including the intercept in the LASSO model and coefficients for the standardized feature variable. The MARS model selected 25 features out of 133. There were 6 common features between LASSO model fit and MARS fit, indicated with *.

Feature	Coefficient
(Intercept)	10.2120
log2CSLength *	-1.0639
log2mRNA *	0.8508
Phosphorylated *	0.2461
log2TotalLength	-0.2058
AA_G_rel	0.1639
AA_L_rel	-0.1425
AA_I_rel *	0.1321
AA_V_rel	0.1264

CS_ca *	0.1112
Local_3UTR_3end_window40_Stability_best_Table10	-0.0811
UTR5_G_nt	-0.0747
CS_at	0.0681
cc_5UTR	-0.0556
CS_tg	0.0490
tg_5UTR	-0.0303
CS_ga *	-0.0198
CS_gc	0.0137
ga_3UTR	-0.0122
logAS_EBI	0.0030

Comparisons between Linear models and MARS

Both Subset selection and LASSO adapt linear regression models. Despite many merits of these procedures, the non-linear relationship between feature variable and protein are not captured and show generally a worse fit than the MARS model. We chose to present the MARS model as it performs best with respect to feature selection, model fit (R^2), and incorporation of non-linearity.

Table S7. Comparison of model performance

GCV R^2 = Generalized Cross-Validated R^2; CV = ten-fold cross-validation

Model	Number of selected features	R^2	Features overlapping with MARS
MARS	25	0.61 (CV) 0.55 (GCV R^2)	-
Linear regression with AIC based subset selection	42	0.51 (CV)	13
LASSO	19	0.45 (CV)	6

4.5.2 Principal component regression

We employed singular value decomposition (SVD) to perform principal component analysis (PCA). The svd function in R (Team, 2009) deconvolutes the data (excluding *log2 protein concentration*) into a matrix of eigenvectors (V), their projections onto the genes (U) and the contributions of the eigenvectors to the variation in the matrix (D). Each eigenvector is a composite of the contributions of each of the features. By definition, the eigenvectors are orthogonal to each other, i.e. they each capture different parts of the variation in the data matrix. There is no theoretical reason for any of the eigenvectors to correlate better with *log2 protein concentration* than the individual features. However, we test for this correlation as it offers a means to reduce the complexity of the data matrix while isolating features that explain variation in protein concentration.

Table S8. Principal component regression (singular value decomposition)

We correlated the projections of each eigenvector (U) with the gene's corresponding *log2 protein concentration*, and vectors V4, V5, V6, and V11 display significant correlations. Vector V4 has the highest correlation, with $R=-0.48$ ($R^2=0.23$).

The four vectors including the top 20 contributing features are listed in the table. Vectors V4, V5, V6, and V11 explain a total of ~15% of the variation in the matrix, implying that only a small proportion of the information in the data can be used to explain variation in protein concentration (with this method). Feature names are explained in the **Supplementary data file.** The top contributing features to vectors V4, V5, V6, and V11 (largest eigen values) are similar to those identified in the MARS model: mRNA concentration, sequence lengths, PEST motifs, and structural features of the 3'UTR.

In sum, while the type of features identified via principal component regression is similar to those found in the MARS model, a much smaller proportion of variation in the data is explained in the former compared to the latter approach. In addition, the biological interpretation of eigenvectors is inherently difficult, which let us prefer the MARS model.

Eigenvector							
V4		**V5**		**V6**		**V11**	
Correlation of projections of eigenvector with log2 protein concentration							
Pearson (P-value)							
-0.48		-0.27		-0.21		0.20	
e-28		e-9		e-6		e-6	
Spearman (P-value							
-0.51		-0.26		-0.24		0.21	
e-33		e-8		e-7		e-6	
Fraction variation in data matrix explained							
0.05		0.04		0.04		0.02	
Contribution of features to eigenvector (sorted)							
Potential_PEST	0.64	log2mRNA	-0.67	Potential_PEST	-0.60	log2UTR3Length	0.61
log2UTR3Length	0.38	AA_IsoelectricPoint	-0.35	log2mRNA	-0.55	log2CSLength	-0.52
log2TotalLength	0.35	Nc	0.32	log2UTR3Length	0.44	Local_3UTR_3end_window40_Significance_best	0.21
log2CSLength	0.34	log2TotalLength	-0.23	AA_IsoelectricPoint	0.29	log2TotalLength	-0.21
log2mRNA	-0.26	log2CSLength	-0.22	log2TotalLength	0.17	Local_3UTR_5end_window40_Significance_best	0.20
log2UTR5_TotalLength	-0.22	Potential_PEST	0.20	log2UTR5Length	0.09	log2mRNA	0.19
Nc	-0.18	log2UTR5Length	-0.18	log2UTR5_TotalLength	-0.08	Potential_PEST	0.17
log2UTR5Length	0.13	Local_3UTR_5end_window40_FoldingEnergy_best	-0.17	Mean_3UTR_5end_window40_Energy_avg	0.07	Mean_3UTR_5end_window40_Energy_stdev	-0.13
AA_IsoelectricPoint	-0.12	Local_3UTR_3end_window40_FoldingEnergy_best	-0.16	Mean_3UTR_3end_window40_Energy_stdev	-0.05	tt_5UTR	0.12
Local_3UTR_5end_window40_FoldingEnergy_best	0.10	log2UTR3Length	-0.15	Local_3UTR_5end_window40_FoldingEnergy_best	0.04	Mean_3UTR_5end_window40_Energy_avg	-0.12
Local_3UTR_5end_window40_Stability_best	0.07	Mean_3UTR_5end_window40_Energy_avg	-0.11	Nc	-0.04	aa_5UTR	0.11
Local_3UTR_3end_window40_Stability_best	0.05	aa_5UTR	-0.07	log2CSLength	0.03	log2UTR5_TotalLength	0.11
Local_3UTR_3end_window40_FoldingEnergy_best	0.05	Local_3UTR_5end_window40_Significance_best	-0.06	Phosphorylatedyes1	-0.03	log2UTR5Length	-0.10
tg_5UTR	0.03	ag_5UTR	-0.06	Local_3UTR_3end_window40_Significance_best	0.02	Local_3UTR_3end_window40_FoldingEnergy_best	-0.08
Mean_3UTR_5end_window40_Energy_stdev_Table21	-0.03	Mean_3UTR_5end_window40_Significance_avg	-0.06	gg_3UTR	-0.02	Mean_3UTR_5end_window40_Significance_stdev	-0.07
Mean_3UTR_3end_window40_Energy_stdev	-0.03	tt_5UTR	-0.05	tt_5UTR	0.02	Mean_3UTR_3end_window40_Energy_stdev	-0.07
ag_3UTR	0.03	AA_hydrophobicmoment_EISD860102	-0.05	Local_3UTR_5end_window40_Significance_best	-0.02	Local_3UTR_5end_window40_Stability_best	0.07

tc_5UTR	0.03	tc_5UTR	-0.05	tc_5UTR	-0.02	Mean_3UTR_5end_window40_Stability_stdev	-0.07
CS_ag	0.02	log2UTR5_TotalLength	0.05	gc_5UTR	0.02	Mean_3UTR_3end_window40_Significance_avg	0.06
Local_3UTR_5end_window40_Significance_best	0.02	aa_3UTR	-0.05	Local_3UTR_3end_window40_Stability_best	-0.02	ca_5UTR	-0.06

5 References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19:** 716–723.

Bodenmiller B, Campbell D, Gerrits B, Lam H, Jovanovic M, Picotti P, Schlapbach R, Aebersold R (2008) PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nat Biotechnol* **26:** 1339-1340.

Charif D, Thioulouse J, Lobry JR, Perriere G (2005) Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* **21:** 545-547.

CodonW http://bioweb.pasteur.fr/seqanal/interfaces/condonw.html.

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer Math* **31:** 317-403.

Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23:** 327-337.

Friedman JH (1991) Multivariate Adaptive Regression Splines. *Annals of Statistics* **19:** 1-67.

Friedman JH, Roosen CB (1995) An introduction to multivariate adaptive regression splines. *Stat Methods Med Res* **4:** 197-217.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154-158.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91-105.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL (2008) The Vienna RNA websuite. *Nucleic Acids Res* **36:** W70-74.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11**.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717-728.

Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* **8:** R183.

Karlin S, Cardon LR (1994) Computational DNA sequence analysis. *Annu Rev Microbiol* **48:** 619-654.

Kozak M (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15:** 8125-8148.

Kuntumalla S, Braisted JC, Huang ST, Parmar PP, Clark DJ, Alami H, Zhang Q, Donohue-Rolfe A, Tzipori S, Fleischmann RD, Peterson SN, Pieper R (2009) Comparison of two label-free global quantitation methods, APEX and 2D gel electrophoresis, applied to the Shigella dysenteriae proteome. *Proteome Sci* **7:** 22.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Le SY, Chen JH, Maizel JV (1990a) Efficient searches for unusual folding regions in RNA sequences. In *Structure and Methods: Human Genome Initiative and DNA Recombination*, Sarma RH, Sarma MH (eds), Vol. 1, pp 127-136: Adenine Press.

Le SY, Malim MH, Cullen BR, Maizel JV (1990b) A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res* **18:** 1613-1623.

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27:** 652-658.

Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25:** 117-124.

Mazan-Mamczarz K, Kawai T, Martindale JL, Gorospe M (2005) En masse analysis of nascent translation using microarrays. *Biotechniques* **39:** 61-62, 64, 66-67.

Milborrow S (2009) earth: Multivariate Adaptive Regression Spline Models.

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.

Nie L, Wu G, Zhang W (2006) Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. *Genetics* **174:** 2229-2243.

Tabaska JE, Zhang MQ (1999) Detection of polyadenylation signals in human DNA sequences. *Gene* **231:** 77-86.

Team RDC (2004) R: A Language and Environment for Statistical Computing. . *R Foundation for Statistical Computing, Vienna, Austria* **ISBN 3-900051- 07-0:** http://www.R-project.org.

Team RDC (2009) R: A language and environment for statistical computing.

Vogel C, Marcotte EM (2008) Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nat Protoc* **3:** 1444-1451.

von der Haar T (2008) A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* **2:** 87.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20:** 2138-2139.

Wu G, Nie L, Zhang W (2006) Relation between mRNA expression and sequence information in Desulfovibrio vulgaris: combinatorial contributions of upstream regulatory motifs and coding sequence features to variations in mRNA abundance. *Biochem Biophys Res Commun* **344:** 114-121.

Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE, Jr. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13:** 1863-1872.

Yen HC, Xu Q, Chou DM, Zhao Z, Elledge SJ (2008) Global protein stability profiling in mammalian cells. *Science* **322:** 918-923.

Yoon K, Ko D, Doderer M, Livi CB, Penalva LO (2008) Over-represented sequences located on 3' UTRs are potentially involved in regulatory functions. *RNA Biol* **5**.