www.sciencemag.org/cgi/content/full/science.1260793/DC1

Supplementary Materials for

Impact of regulatory variation from RNA to protein

Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard,* Yoav Gilad*

*Corresponding author. E-mail: pritch@stanford.edu (J.K.P); gilad@uchicago.edu (Y.G.)

Published 18 December 2014 on *Science* Express
DOI: 10.1126/science.1260793

This PDF file includes

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1260793/DC1)

Databases S1 to S4 as zipped archives:

Data Table S1. cis-QTL mapping results.
Data Table S2. Significant psQTLs and esQTLs identified by the conditional model.
Data Table S3. psQTLs and esQTLs identified by the interaction model.
Data Table S4. Quantification for each gene in each individual for protein, RNA-seq, and ribosome profiling data.

Supplementary Information for

Impact of regulatory variation from RNA to protein

Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, Yoav Gilad

Supplementary Methods

Cell Culture

We collected protein abundance and ribosome occupancy data from Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) derived from humans (YRI from Coriell, NIGMS Human Genetic Cell Repository). The cells were donated by the Yoruba people of Ibadan, Nigeria. Cells were maintained at identical conditions of 37° with 5% CO_2 in RPMI media with 15% FBS, supplemented with 2 mM L-glutamate, 100 IU/ml penicillin, and 100 µg/ml streptomycin. An internal standard LCL (Coriell YRI, NIGMS Human Genetic Cell Repository, GM19238) was grown in RPMI minus L-Lysine and L-Arginine, 15% dialyzed FBS, and L-$^{13}C_6{}^{15}N_4$-arginine (Arg-10) and L-$^{13}C_6{}^{15}N_2$-lysine (Lys-8) (Cambridge Isotopes, Andover, MA, USA) supplemented with 2 mM L-glutamate, 100 IU/ml penicillin, and 100 µg/ml streptomycin under identical conditions as the unlabeled LCLs. The internal standard LCL (GM19238) was grown for 6 doublings to ensure complete SILAC label incorporation. Complete label incorporation was verified by analyzing the protein lysate from the labeled LCL alone by high-resolution LC-MS/MS.

Genotype Data

Genotype information was obtained as described in (*25*): Genotypes were imputed with BimBam (*26, 27*) for variants called in either HapMap (*28*) or 1000 Genomes (*29*). A reference panel of 210 YRI HapMap individuals was used for imputation. Using this approach we obtained genotypes at approximately 15.8 million variants -14.9M (94%) SNPs and 0.9M (6%) indels - for all individuals in our cohort. Genotypes can be obtained from:
http://eqtl.uchicago.edu/dsQTL_data/GENOTYPES/.

RNA-Sequencing

RNA-sequencing data (already processed and mapped) for 75 YRI HapMap individuals were available from Pickrell et al 2010a,b (*19, 30*), providing a median of 8.6M reads per individual uniquely mapped to Ensembl genes. Both mapped and unmapped reads can be obtained directly from:
http://eqtl.uchicago.edu/RNA_Seq_data/

Ribosome Profiling

Ribosome footprint profiling experiments were performed for 72 YRI HapMap individuals using ARTseq™ Ribosome Profiling kit for mammalian cells (RPHMR12126) following vendor's instructions. Cell lysates were prepared from flash frozen pellets of 30 to 50 million live cells by repeat pipetting in 1 ml cold lysis buffer. Sephacryl S400 spin column (GE; 27-5140-01) was used for monosome isolation. For rRNA depletion, Ribo-Zero Magnetic Kit (Epicentre; MRZH11124) was used. Ribosome footprint cDNA libraries were PCR amplified (12 to 15 thermo-cycles) and barcoded using ScriptMiner Index PCR Primers (Epicentre; SMIP2124). Indexed libraries were pooled to sequence on an Illumina HiSeq 2500. Sequence reads were processed and mapped to the human genome following the procedure used by Ingolia et al 2012 (*31*). Only uniquely mapped reads were used for downstream analysis. BEDTools (v2.8.0) was used to intersect (strand specific) mapped reads with Ensembl gene annotation (downloaded on 06/06/13) to generate a count table. Ribosome profiling libraries tend to contain large proportions of rRNA, tRNA and snoRNA reads (compared with the more standard RNA-

seq libraries). Yet, as we have shown, the gene-specific ribosome profiling read counts are highly consistent across technical replicates (Figure S2). In our data, approximately 80% of reads were from rRNA, tRNA or snoRNA, which is comparable to the previously reported levels of approximately 85% rRNA in other studies (*18*). We removed these reads from all subsequent analyses. After this filtering, we obtained a median of 12.1M reads per individual uniquely mapped to the genome (with at least 95% of reads in each individual passing Q30). We observed high correlations between sequencing runs for the same sample (Figure S2). Correlations between replicate runs (as well as the degree of expected periodicity across codons) were high regardless of the proportion of reads mapping to within coding regions (Figure S2, S3).

It should be noted that a certain degree of variation between RNA-seq and ribosomal profiling data is due to the fact that RNA-seq samples undergo poly-A selection, which somewhat bias the resulting estimates of expression levels. More recent RNA-seq protocols, which rely on effective ribosomal RNA exclusion rather than poly-A selection, can provide data that are less biased and probably more highly correlated to the ribosomal profiling data.

Quantitative, High-Resolution Mass Spectrometry

We measured protein abundance using mass spectrometry for 62 YRI HapMap individuals. We used the same protocol as applied previously by Khan et al (2013) (*14*). Specifically, the LCLs were washed in PBS three times and then lysed using the UPX Universal Protein Extraction Kit (Expedeon Inc., San Diego, CA, USA). Protein quantitation was performed using the Qubit fluorometry assay (Invitrogen) and the reducing agent-compatible (RAC) version of the BCA Protein Assay (Thermo Scientific Pierce). 12µg of each sample was combined with 12µg of the SILAC labeled lysate from human LCL GM19238. Note that the SILAC lysate was prepared once and used as an internal standard through the quantification of the cell lines. 24µg of each combined sample was then processed by SDS-PAGE using a 4-12% Bis Tris NuPage mini-gel (Invitrogen). Calibration was with Thermo PageRuler broad range markers. Each of 40 gel segments were processed by in-gel digestion using a ProGest robot (DigiLab) with the following protocol: wash with 25mM ammonium bicarbonate followed by acetonitrile, reduce with 10mM dithiothreitol at 60°C followed by alkylation with 50mM iodoacetamide at room temperature, digest with trypsin (Promega) at 37°C for 4h, and quench with formic acid. The supernatant was analyzed directly without further processing. Each gel digest was analyzed by nano-LC/MS/MS with a Waters NanoAcquity HPLC system interfaced to a ThermoFisher LTQ-Orbitrap Velos Pro. Peptides were loaded on a trapping column and eluted over a 75µm analytical column at 350nL/min using a 1-hour LC gradient. Both columns were packed with Jupiter Proteo resin (Phenomenex). The mass spectrometer was operated in data-dependent mode, with MS performed in the Orbitrap at 60,000 FWHM resolution and MS/MS performed in the LTQ. The fifteen most abundant ions were selected for MS/MS.

Computational Analysis of Mass Spectra

Low-level analysis was performed using the open-source proteomics software tool PVIEW (Release December 23, 2012; http://compbio.cs.princeton.edu/pview). As input to PVIEW, we generated *in silico* translations of coding genes from the UCSC Genome Browser database based on gene models from the human genome. Each protein sequence entry retained the corresponding Ensembl gene identifier and gene symbol. Database searches were performed using ±4 p.p.m. MS1 tolerance and an MS2 window tolerance of ±0.5 Da. Up to 2 missed tryptic cleavages were allowed during search. Peptide spectrum matches were obtained at a stringent false discovery rate (FDR) of 1%. As is standard practice, $\log_2$(sample/standard) peptide measurements were centered so their median was zero to adjust for any differences in sample loading. For the protein-level quantification, we used the median $\log_2$(sample/standard) ratio across all independent quantifications of a protein in the same individual (distinct peptides including duplicate peptide measurements across fractions and for differing charge states). We removed genes that were quantified at the protein level in fewer than half of the individuals. A total of 4,381 genes were quantified in at least half of the cell lines and thus used for further analysis.

Study design considerations

In our study, as in most other QTL studies of regulatory phenotypes, we chose to sacrifice quantitative accuracy to obtain data from tens of individuals. An alternative would have been to collect data from technical

and/or biological replicates from the same individual. Including replicates in our study design would have allowed us to account for much of the observed noise and ultimately obtain more accurate and precise measurements for all phenotypes. However, across genotype classes, most of the variation is found between individuals, not replicates. Thus, for equivalent size (total number of samples processed), studies that include more individuals generally have more power to detect regulatory QTLs.

Regardless of power considerations, another possible concern is that inaccurate measurements may lead to bias, especially when we compare patterns across phenotypes. The use of the SILAC ratio (with a common reference) is expected to minimize systematic bias in the proteomics data. Across all phenotypes, we used PCA to measure and account for general batch effects in the data, and in most of our analysis we used fold changes within a data class, which cancels out potential systematic bias that could obscure the true correlation between different data classes. Finally, we investigated an alternative approach to collect the transcript expression data (microarrays; Figure S17), and find the same patterns we observed with the RNA sequencing data, indicating that our approach is robust to noise properties of different technologies.

cis-QTL Mapping

To adjust for differences in sequencing depth, RNA-seq and ribosome profiling read counts were normalized by the number of mapped reads: #(reads mapping to a gene) / #(total mapped reads). Prior to mapping QTLs, we used the normalization and standardization approach developed previously by our group (*25*). Briefly, we centered and scaled each protein, ribosome occupancy, and mRNA measurement by subtracting out the mean value for the gene and dividing by the standard deviation. We then quantile-normalized the measurements of each individual to fit a standard normal distribution. We used principal components analysis (PCA) to regress out unidentified confounders from all three data sets. We regressed out 9 PCs from the protein abundance measurements, 18 PCs from the ribosome profiling measurements, and 15 PCs from the mRNA measurements prior to mapping cis-QTLs for each phenotype. The numbers of PCs corrected were chosen to maximize the number of detected QTLs in each data type. In order to apply PCA to the protein abundance measurements, we imputed missing values using K Nearest Neighbors imputation (using impute.knn from R package impute, with default settings). This was used only for the purposes of PC removal; values originally missing were excluded from eQTL mapping. To map cis-QTLs for each gene and each of the three phenotypes, we used the following procedure using a dense set of imputed SNPs obtained from HapMap3 data (*25*):

1. For every SNP within a ±20kbp window of the gene with minor allele frequency (MAF) of 0.1 or greater, record the Pearson correlation with the expression phenotype along with the corresponding p-value.
2. Take the minimum p-value p_{best} among all SNPs tested for the gene.
3. Repeat (1-2) for 10,000 permutations of the genotype sample labels, obtaining $p_{perm}(1)…p_{perm}(10000)$
4. Estimate an empirical gene-level p-value for the most significant QTL based on the permutations as the fraction of permutation p-values at least as significant as the original p-value: sum_i ($p_{perm}(i) \leq p_{best}$)/10,000.

Alternative measures of QTL effect sizes

An alternative measure of QTL effect sizes would be to report the % variance within vs. between genotypes. However, this statistic is affected by the amount of measurement error, which may differ among the three phenotypes. Thus, we decided to present the estimated effect size measured as fold-change across genotypes (this is unbiased with respect to measurement error), along with a p-value as a measure of significance. The p-values are, in effect, a test of the relative variances within and between classes.

Replication of QTLs between phenotypes

Replication rates across data sets were estimated by taking only the SNP-gene pairs identified as QTLs in one phenotype (e.g., transcript expression) and testing them in a second phenotype (e.g., protein expression). This analysis is less sensitive to power limitations than genome-wide testing. Here, since only one SNP is tested per gene, we simply recorded the nominal p-value for each QTL using a Pearson correlation test without

permutation analysis. Correction for multiple hypothesis testing was performed using qvalue (*32, 33*), and we report the fraction of tests passing an FDR threshold of 10%.

Replication of pQTLs from previous studies

Previously detected pQTLs were obtained from Wu et al. (*15*). We used the set of QTLs identified in the combined populations from their Supplementary Table 8. We considered the set of SNP-gene pairs for which the SNP is present in our data with MAF>10%, and for which we have gene-level data for all phenotypes. This analysis was based on the set of 60 individuals for which we measured all three phenotypes. Analysis was performed on the raw data (i.e., no PCs removed, no quantile normalization), and we recorded the nominal p-values from Pearson correlation tests.

QTL effect-size comparison

Comparing effect sizes across regulatory phenotypes is challenging because the initial identification of a QTL will bias the comparison (effect sizes tend to be overestimated in discovery samples due to the Winner's Curse). To address this challenge, we did not perform the initial QTL identification using our own data. Instead, we obtained the eQTLs reported online by the GEUVADIS study (*2*), using the set of genes reported with an eQTL in the CEU population at an FDR of 0.05 and available data in our study for all three phenotypes. Using the provided coordinates of the best SNP for each eQTL gene, we identified eQTLs with sufficient SNP frequency (MAF > 0.1) in our study, and recorded the reported direction of effect in GEUVADIS according to the non-reference allele for each. For each applicable eQTL, using the recorded SNP and gene, we then estimated the effect size in each of our three phenotypes, as follows. The RNA-seq and ribosome occupancy data for each individual were normalized for total read depth only, and $\log_2$-transformed to make them comparable to the protein data (which are measured on a $\log_2$ scale as fold-change relative to the internal standard). Our analysis proceeded without quantile normalization or PC removal, in order to avoid distortion of the effect sizes. Regression coefficients were recorded from a standard linear model with genotypes coded as 0,1,2 and the sign was adjusted according to the direction of the effect reported in GEUVADIS (thus we would expect replicating eQTLs to have positive effect sizes).

The effect sizes output from this model can be interpreted as measuring the average fold change (on $\log_2$ scale) per allele-copy of the higher expression allele. We chose this regression-based approach using raw data, rather than measures of correlation such as percent variance explained or R^2, to avoid undesired confounding influences such as different levels of technical noise for the three phenotypes, non-genetic biological factors, and MAF of the SNP. Any eQTLs that were false positives in GEUVADIS should have an expected effect size of zero in our data, and thus should slightly reduce the mean effect sizes that we estimate, but should have the same proportional effect on all three phenotypes. Significance of observed differences in the mean effects between phenotypes were assessed using a t-test.

Conditional model for finding phenotype-specific QTLs

We tested for phenotype-specific QTLs using a likelihood ratio test (LRT) to identify SNPs significantly associated with the phenotype of interest while including other phenotypes as covariates to account for effects fully mediated by another phenotype. Let P_{tij} represent the expression level of the phenotype *t* in individual *i* and gene *j*, g_{is} be the genotype for that individual at SNP *s*, and P_{uij} be the expression level in another phenotype *u* that we seek to control for. We compared two linear models

$$P_{tij} = \mu_{tj} + \beta_g g_{is} + \beta_p P_{uij} + \varepsilon_{i,j}$$

$$P_{tij} = \mu_{tj} + \beta_p P_{uij} + \varepsilon_{i,j}$$

using a likelihood ratio test to determine the contribution of SNP *s* not mediated by phenotype P_u. To identify psQTLs, we applied this model with $\underline{P_{tij}}$ representing protein levels and the covariate P_{uij} representing RNA. Similarly, to detect esQTLs, RNA is captured by P_{tij} and protein levels by P_{uij}. To assess the significance of genes with protein-specific and expression-specific QTLs, we employed the permutation approach described for standard QTL testing, with this test used in place of Pearson correlation. The conditional model was run using normalized and PC-corrected data, exactly as specified for *cis*-QTL mapping above, and was applied to all applicable SNPs in the cis-candidate region for each gene with the relevant data types.

Interaction Model for Comparison of Effect Sizes Across Phenotypes

To determine if there exists a significant difference between effect sizes across phenotypes (protein vs RNA), we tested for an interaction between a variable encoding a phenotype identifier and genotype. To test for this interaction, we compared the following two linear models:

$$P_{tij} = \mu_{tj} + \beta_{tg} g_{is} + \varepsilon_{tij}$$
$$P_{tij} = \mu_{tj} + \beta_{g} g_{is} + \varepsilon_{tij}$$

using a likelihood ratio test (LRT). Here P_{tij} is the phenotype value for phenotype *t* (*t*=protein or RNA), individual *i* and gene *j*, and g_{is} is the genotype for individual *i* at SNP *s*. The likelihoods were calculated over all individuals and both phenotypes for a given gene and SNP. In the full model, we allow for a phenotype-specific effect size β_{tg} for each phenotype *t*; in the reduced model we assume a single shared effect size β_g. A maximum likelihood fit of each of these models was obtained using generalized least squares to allow for unequal variances across phenotypes, and significance was calculated assuming one degree of freedom. This model was applied to the raw data (no normalization or PC correction) in order to more accurately reflect biologically meaningful effect sizes. To identify protein-specific QTLs (psQTLs) with this approach, we tested each of the best SNPs selected when mapping the 277 gene-level pQTLs quantified by all 3 methods for a significant interaction between phenotype identifier and genotype. Similarly, to identify RNA-specific QTLs (esQTLs), we tested the best SNPs chosen when mapping 902 gene-level eQTLs for a significant interaction between phenotype identifier and genotype.

QTL enrichment analysis

The enrichment results shown in Figures S18-S21 and Table 2 were generated based on the conditional model described above, used to identify psQTLs and esQTLs, in order to identify primarily functional effects not shared between phenotypes. Nominal p-values were recorded for every SNP-gene pair tested in the defined *cis* region of each gene, rather than a single best SNP per gene, with a relaxed MAF threshold of 0.05 for this analysis. The HLA region of chromosome 6 was excluded from this analysis due to extensive LD structure and strong associations that could dominate enrichment results. Note that some SNPs may be included in this analysis due to linkage disequilibrium with a distinct causal site; on average this effect will tend to slightly reduce the observed enrichments.

We evaluated the distribution of p-values for all SNP-gene pairs matching each of the annotations shown. We generated qq-plots using all recorded p-values. We evaluated significance of enrichment of each annotation (Table 2, Figure S21) based on Spearman's correlation between 1) the vector of recorded QTL p-values from the model and 2) a binary vector where each entry indicates whether the corresponding SNP matches the genomic annotation being tested. For many annotations, we restricted each of the tests to SNP-gene pairs matching an appropriate "background" – for instance, when testing for enrichment of non-synonymous SNPs, we considered only exonic SNPs when constructing the two vectors needed for the Spearman correlation test. This avoids confounding the enrichment of exonic SNPs (the background) when testing for significance of non-synonymous SNPs. Similarly, we generated p-values for all SNP-gene pairs from the interaction model testing for differences in effect between RNA and protein. In this case, for visualization (Figure S17), we assigned a sign to each p-value according to which phenotype displayed a larger absolute effect size (regression coefficient): SNPs with a greater effect on protein than RNA were assigned a positive sign, and greater effects on RNA were assigned a negative sign. In cases where the two regression coefficients indicated different directions of effect in addition to different magnitude, we treated them as a protein-specific effect (positive sign) as these SNPs are still indicated as candidates for post-transcriptional regulatory function.

Genomic annotations

Genomic annotations for use in the enrichment analysis were obtained from diverse public resources. Gene annotations were obtained from GENCODE annotation version 3c, and used to classify each SNP as belonging to intronic, exonic, 5' UTR, 3' UTR, or the intergenic region with respect to each tested gene. SNPs modifying, creating, or destroying an upstream ORF were obtained from Table S3 of Calvo et al. (*34*) SNPs predicted to modify mRNA secondary structure (riboSNitches) were obtained from Wan et al. (*35*), Supplementary Table 3. Nonsynonymous SNPs were obtained from the UCSC Genome Browser (*36*) variant table on 2/10/2014. Genomic positions of Pfam domains (*37*) were also obtained from the UCSC Genome

Browser table/track ucscGenePfam.txt downloaded on 6/21/14. Post-translational modifications (PTMs) were obtained from the 6/3/14 version of the PhosphositePlus Database (*38*). Regions of protein disorder were predicted using IUPred (*39*). Predictions of the deleteriousness of a non-synonymous SNP were obtained from PolyPhen-2 (*40*) column in the dbNSFP v2.5 database(*41*). For PTMs, disordered regions, and deleteriousness predictions, genomic positions of these annotations were obtained by mapping peptides back to protein sequences annotated with gene model coordinates. For each annotation considered, we restricted our analysis to SNPs polymorphic in our sample with MAF ≥0.05.

Data Release

All experimental data have been made publicly available. The protein data have been deposited into ProteomExchange (http://www.proteomexchange.org) under accession PXD001406 and the ribosome profiling data have been deposited into GEO (http://www.ncbi.nlm.nih.gov/geo/) under accession GSE61742.

Supplementary Figures

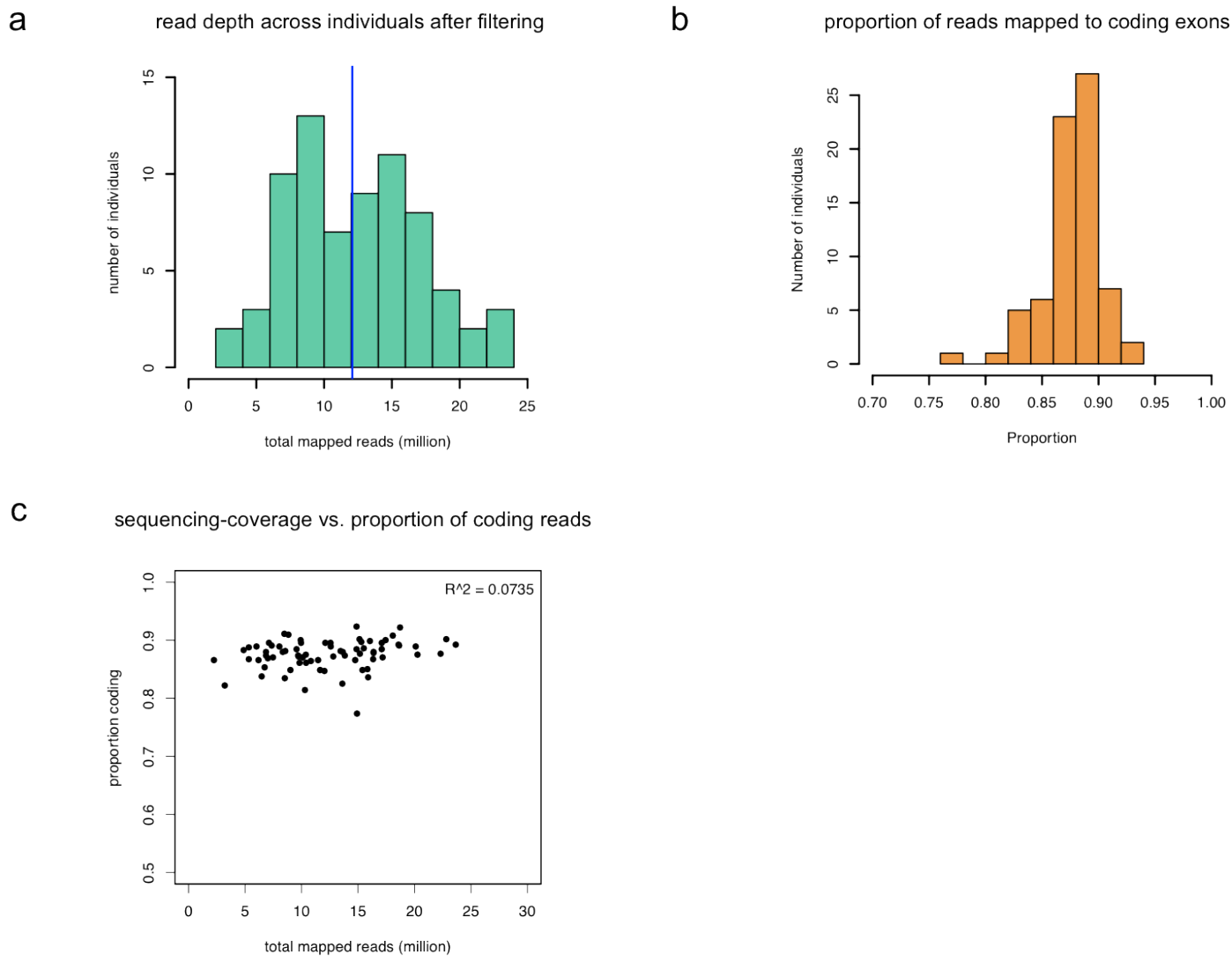


Figure S1. Global properties of ribo-seq data. (a) Read depth distribution across individuals, after filtering of rRNA, snoRNA and tRNA reads. The blue line indicates the median level at 12.1 million uniquely mapped reads. (b) Proportion of reads mapped to coding exons from uniquely mapped reads for each cell line. (c) Scatter plot correlating sequencing-coverage and the proportion of coding reads by individual cell lines.

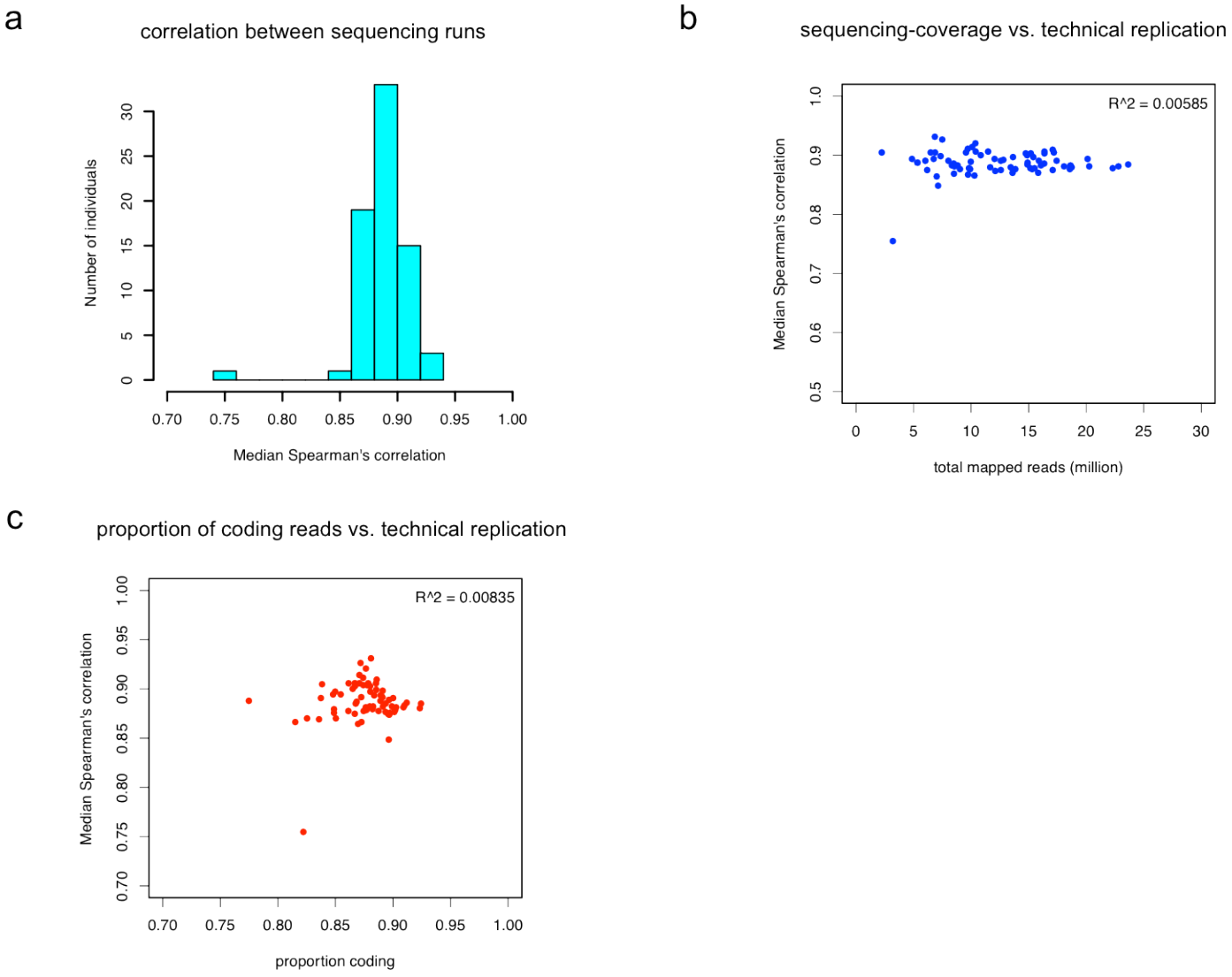


Figure S2. Quality Control Analyses of Ribosome Profiling Data: technical replications. (a) Median Spearman's correlation for translation level estimates (read count per gene) between sequencing runs for each cell line. (b) Scatter plot correlating sequencing-coverage and technical replication of ribosome occupancy estimates by individual cell lines. (c) Scatter plot of the proportion of coding reads and correlation of ribosome occupancy estimates across technical replicates of individual cell lines.

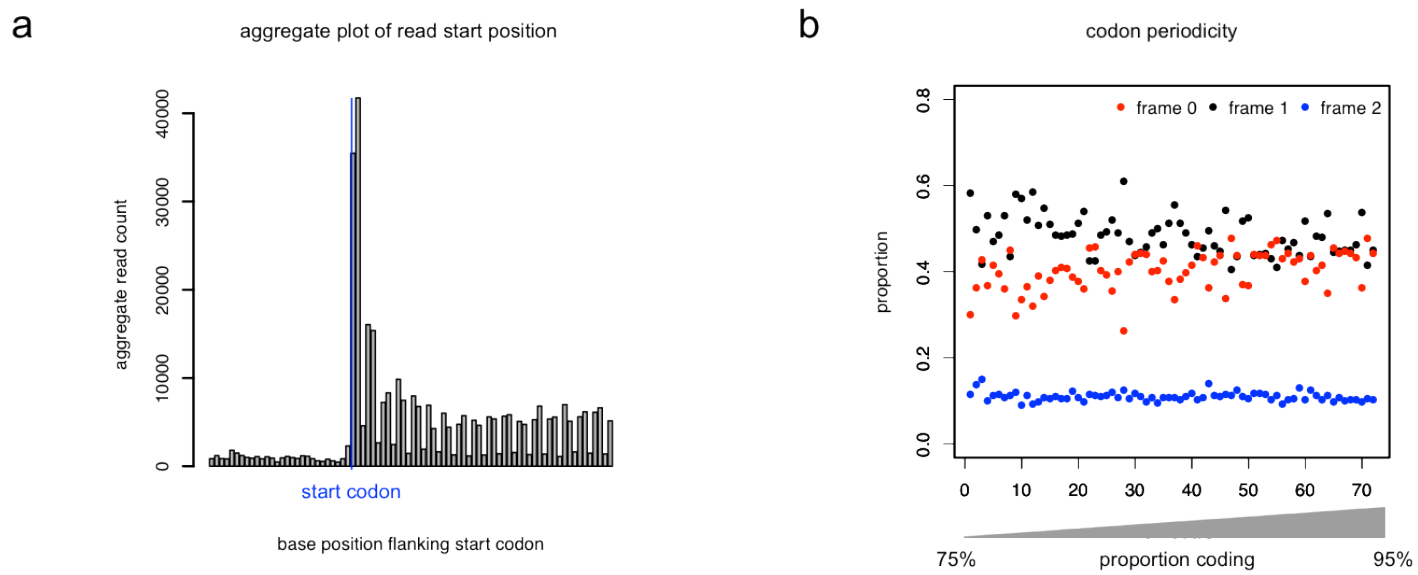


Figure S3. Quality Control Analyses of Ribosome Profiling Data: codon periodicity. (a) An aggregate plot for 5' position of all sequence reads from GM19116 that fall in 80 bp windows flanking all annotated start codon on the plus strand. Notice the subcodon periodicity beginning at annotated start codon. (b) Proportion of reads (5' end) that fall in each reading frame of the start codon window described in (a) for all individuals, sorted by percentage of reads mapped to coding exons (from low to high as indicated beneath the plot).

This plot (as well as figures S1 and S2), show that variation in the proportion of reads mapped to reading frame (coding exons) is not associated with levels of sequencing depth, codon periodicity, or correlation between technical replicates.

Protein quantification strategy.

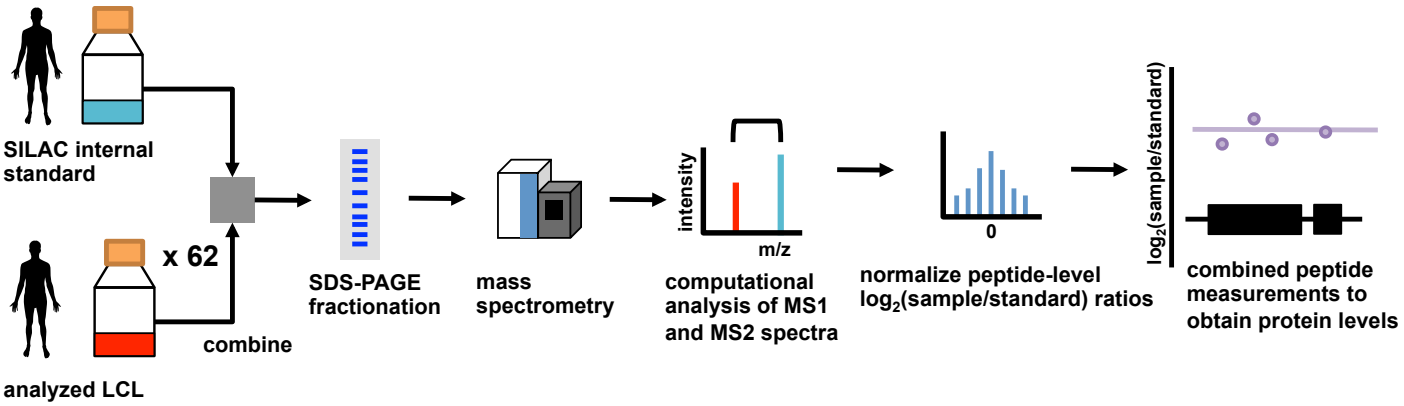


Figure S4. Protein quantification strategy. Proteins were quantified using a single stable-isotope labeled protein sample obtained from a batch culture of human lymphoblastoid cell line (LCL) GM19238, which served as a common internal standard. Protein samples from unlabeled analyzed cell lines were combined with aliquots of the SILAC internal standard protein sample from GM19238. Denatured proteins were separated by one-dimensional SDS-PAGE. A total of 40 gel fractions were manually excised and trypsin digested by robot. Each gel fraction was analyzed by high-resolution LC-MS/MS using 1-hour long LC gradients. Computational analysis of mass spectra provided $\log_2$(sample/standard) ratios for each detected tryptic peptide. Peptide ratios were subsequently normalized to adjust for small differences in sample input for the analyzed LCL and the internal standard protein sample. The median peptide ratio across a gene was used as the protein-level quantification. Protein levels, relative to the common internal standard, were compared across individuals for each detected protein.

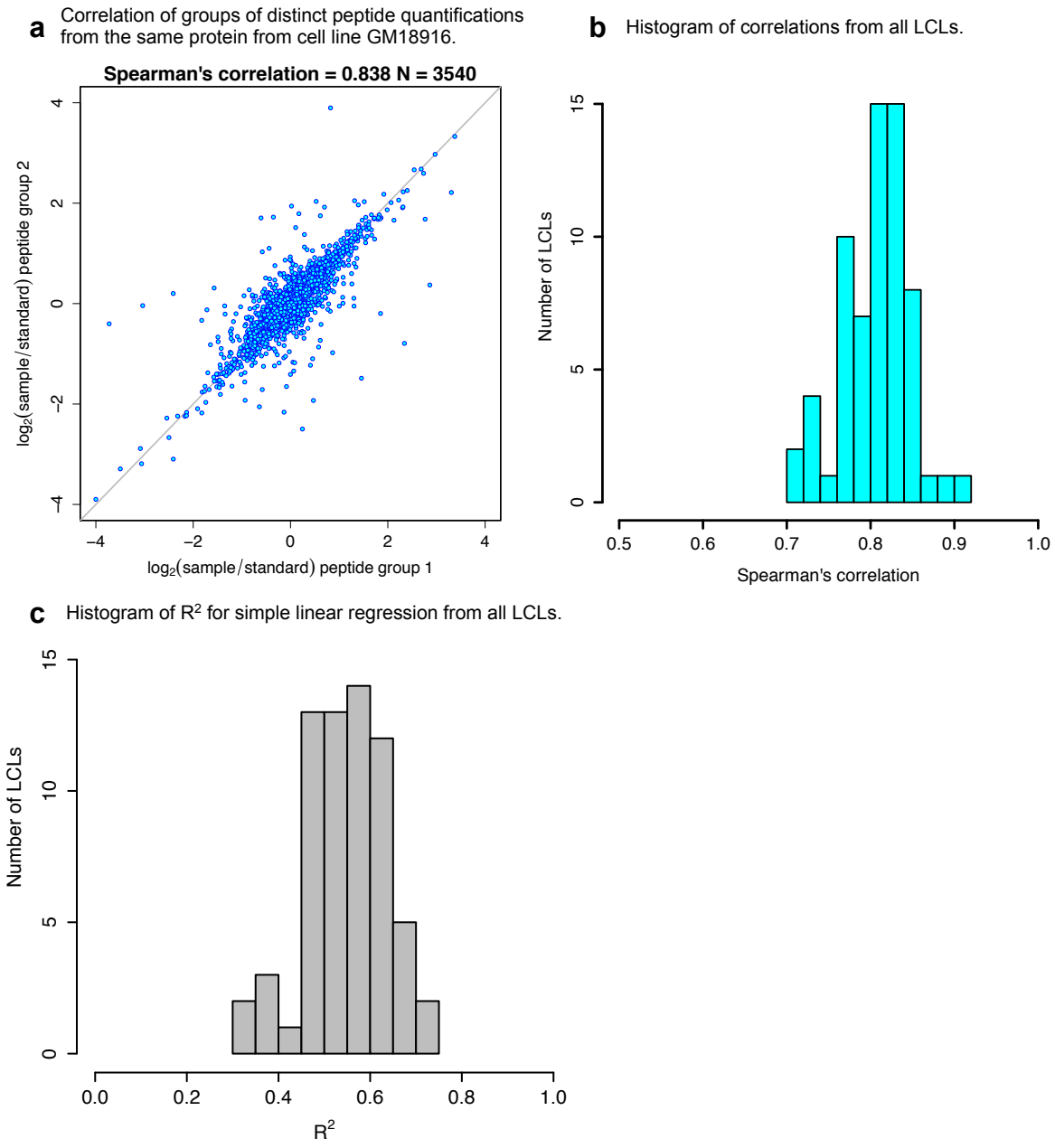


Figure S5. Assessment of protein identification accuracy and quantification precision. We assessed the accuracy of peptide identifications and the precision of peptide quantifications, relative to the common internal standard, by correlating $\log_2$(sample/standard) ratios from distinct peptides from the same protein. Distinct peptide quantifications for proteins quantified by two or more peptides were randomly grouped into one of two groups. The median quantification from group 1 was correlated with group 2. If the peptides are accurately identified and quantified, the median quantifications of the randomly generated groups will be correlated. Off diagonal points reflect a combination of technical and biological noise. (a) Illustrates the correlation of distinct peptide quantifications from the same protein for LCL GM18916 for 3540 proteins that were quantified with 2 peptides or more. Each point corresponds to a protein. (b) provides a histogram of Spearman's rank correlations obtained for the LCLs analyzed in this study. (c) provides a histogram of R^2 values for simple linear regression between the two group quantifications for all LCLs analyzed.

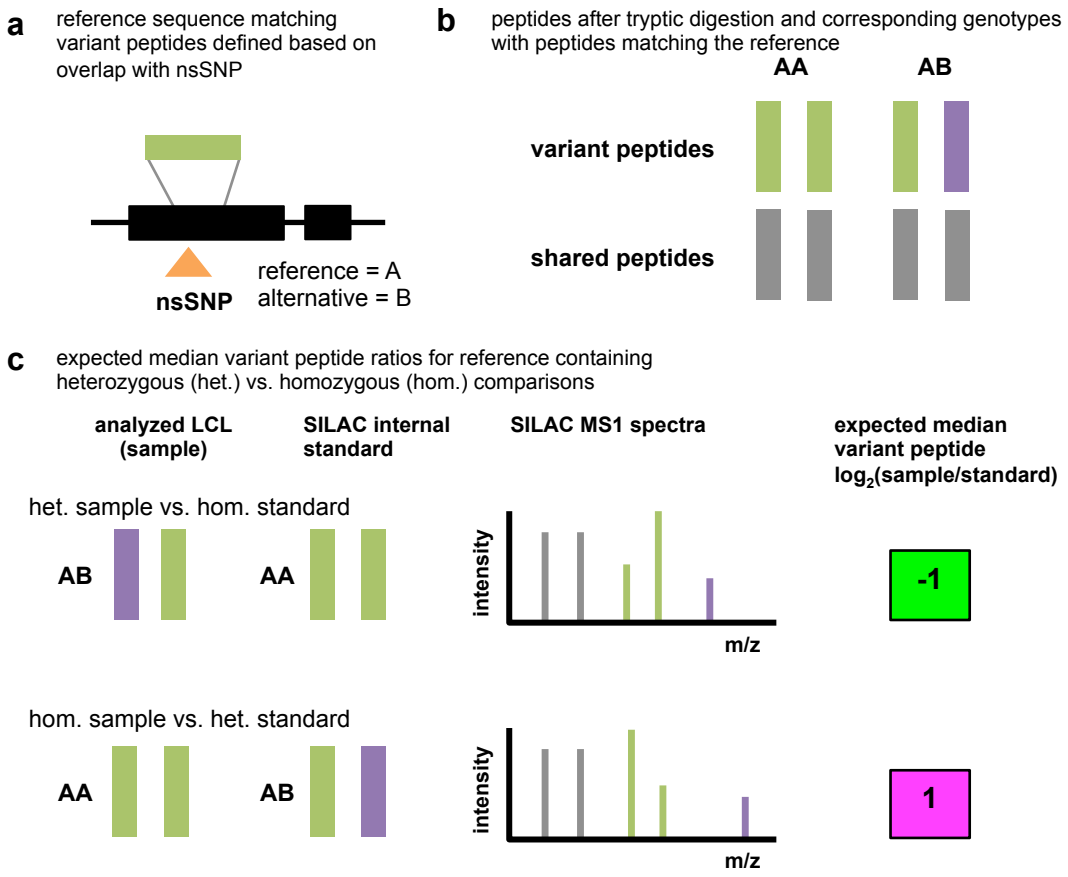


Figure S6. Evaluation of peptide ratios for variant peptides overlapping non-synonymous SNPs (nsSNPs) in comparisons between homozygotes and heterozygotes. (a) Variant peptides mapping to the reference sequence were identified on the basis of their overlap of an nsSNP within a coding region. The two possible variant peptide alleles are colored in green for matching the genome reference allele and purple for alternative, above. (b) If a gene contains a single nsSNP, tryptic peptides that are shared (gray) as well as those corresponding to the two alleles (green, reference, and purple, alternative) will be generated after tryptic digestion. We identify only peptides mapping to the reference sequence. The purple peptides mapping to the alternative allele will not be identified or quantified. Yet, in heterozygous versus homozygous quantifications, the presence of the alternative allele will affect the median quantification of the corresponding reference matching peptides. (c) illustrates comparisons at a specific nsSNP across peptides heterozygous in either the internal standard or the analyzed cell line. Assuming that on average the two protein alleles are expressed at the same level in a heterozygote, the expected median ratio should reflect the copy number of the variant peptide in the heterozygote, as illustrated to the right.

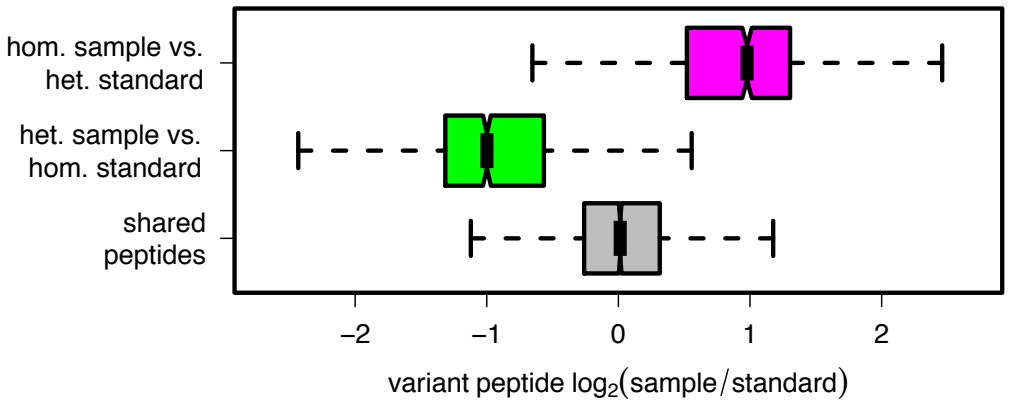


Figure S7. Analysis of variant peptides in heterozygous to homozygous comparisons at nsSNPs shows no evidence of ratio compression. In total, we identified 10,782 variant peptide spectrum matches at a highly stringent false discovery rate (FDR) of 0.1% that overlapped genotyped nsSNPs across the 62 individuals. Among these high confidence variant peptides, 1136 were homozygous in the labeled standard and heterozygous in the analyzed cell line (quantifications in magenta). 1551 were homozygous in the sample and heterozygous in the internal standard (quantifications in green). Quantifications of all remaining shared peptides that also overlap known nsSNPs sites, homozygous in both the sample and the standard, are shown in gray. Bold lines in the boxplots designate the median. Hinges correspond to the first and third quartiles. Whiskers extend 1.5 times the corresponding quartile. Consistent with expectation, the overall median protein levels, measured as $\log_2$(sample/standard) from primarily shared peptides, were centered at zero. The expected median $\log_2$(sample/standard) measured from variant peptides that overlapped an nsSNP that was homozygous in the sample and heterozygous were centered approximately at one as in the heterozygote both protein alleles are expressed (*42*). Similarly, variant peptides that overlapped an nsSNP heterozygous in the sample and homozygous in the standard were centered approximately at minus one.

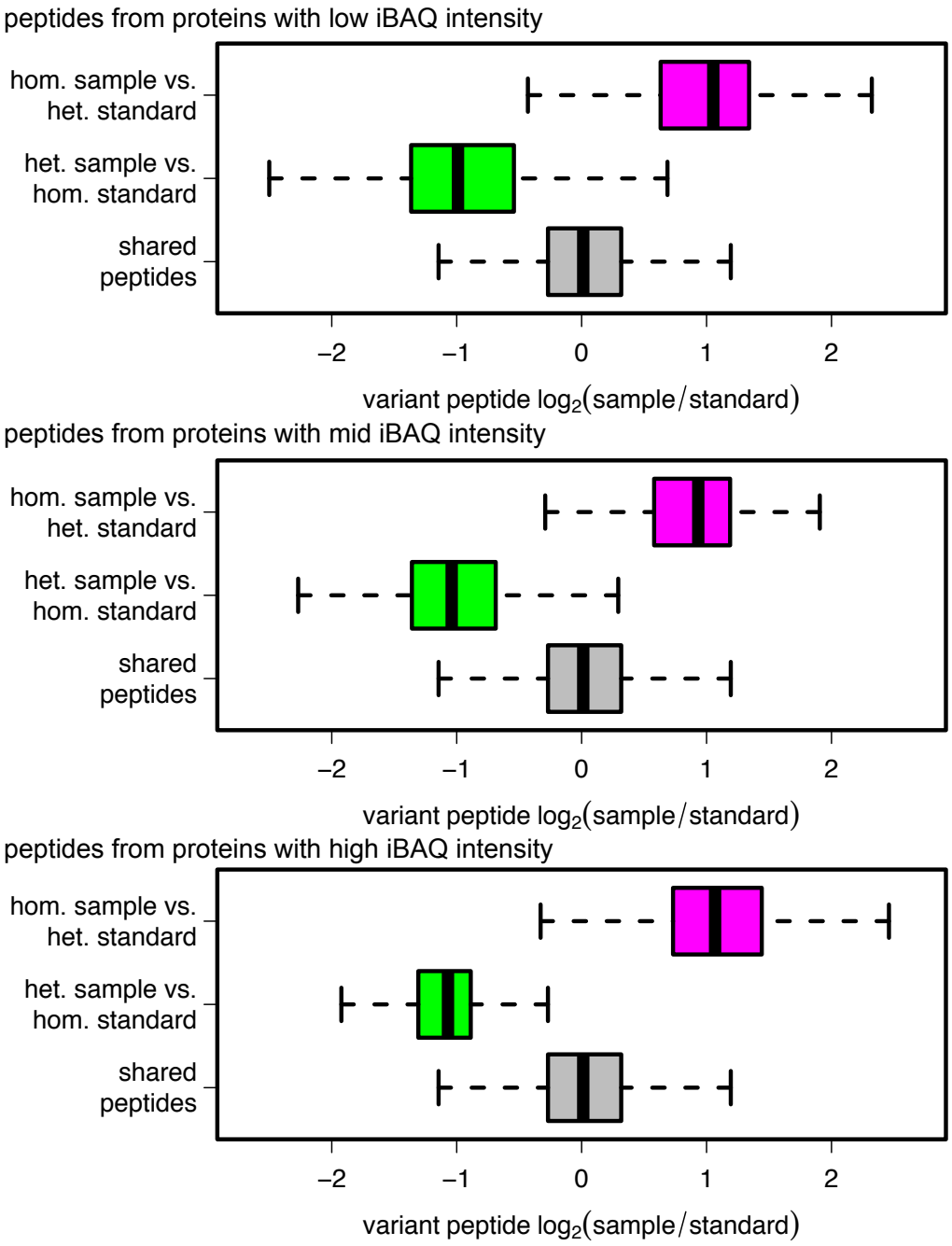


Figure S8. Variant peptide quantifications in heterozygous to homozygous comparisons at nsSNPs stratified by iBAQ protein intensity. To evaluate whether effect sizes might display ratio compression at some—but not all—iBAQ intensity levels, we repeated the analysis from Figure S5 after stratifying genes according to intensity. Here, genes are grouped into 3 equally-sized buckets according to the median iBAQ intensity across individuals.

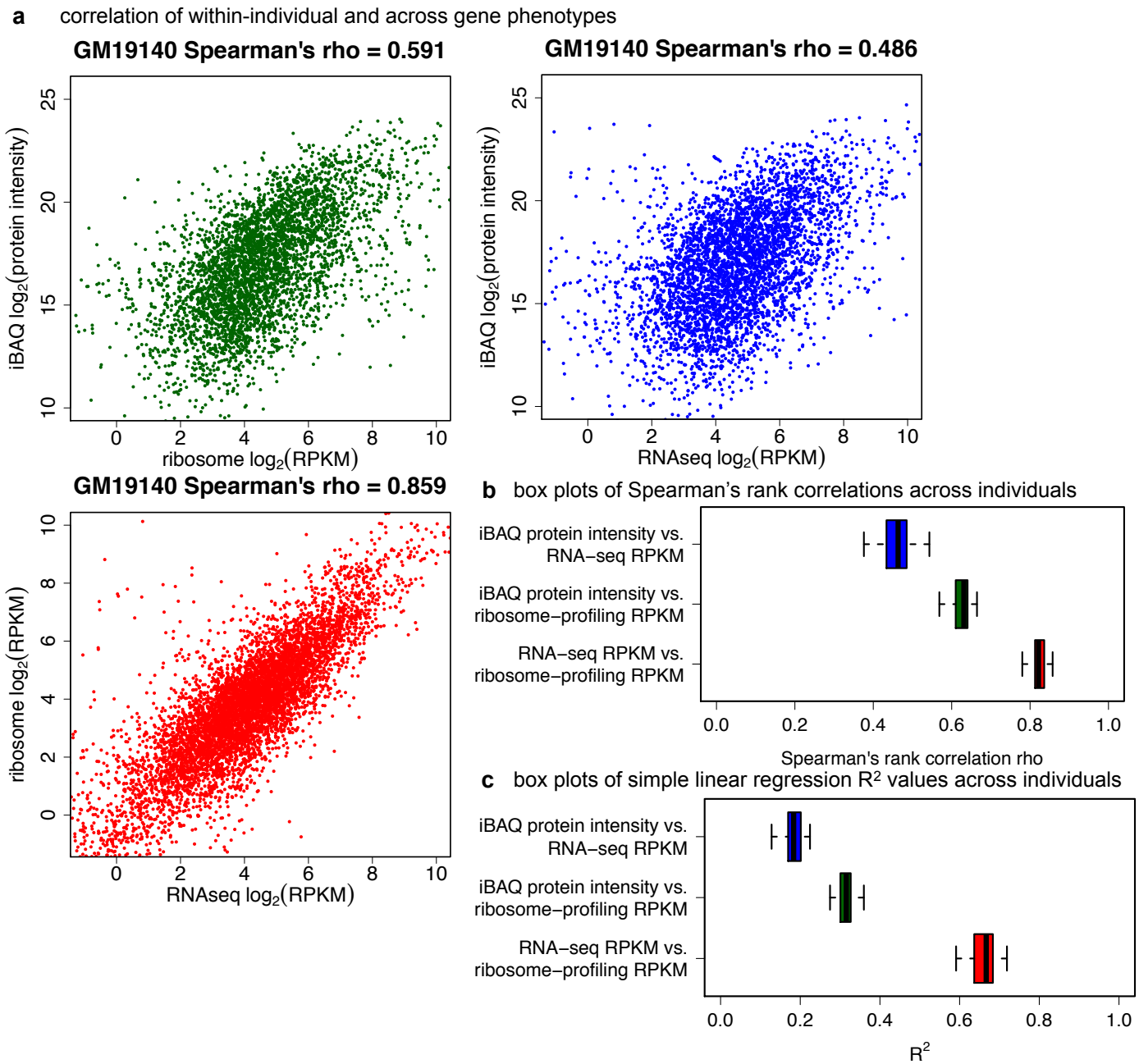


Figure S9. Across-gene and within-individual correlation of phenotypes. To compare protein data with sequencing-based phenotypes within each sample (across genes), an absolute rather than relative quantification was required. Therefore, here, iBAQ protein intensity, an intensity weighted spectral counting measure, was computed as previously described (*22*) using the unlabeled channel in the SILAC data sets. We focused on 4,092 genes from 61 individuals where we could compute iBAQ protein intensity values, ribosome profiling RPKM and RNA-seq RPKM. (a) Scatter plots for individual GM19140 correlating each phenotype. (b) Boxplots of Spearman's rank correlation values computed for each individual. (c) provides boxplots of simple linear regression R^2 values computed for each individual across phenotypes. Ends of boxes designate quartiles, and the thick line designates the median. Whiskers extend to 1.5 times each respective quartile.

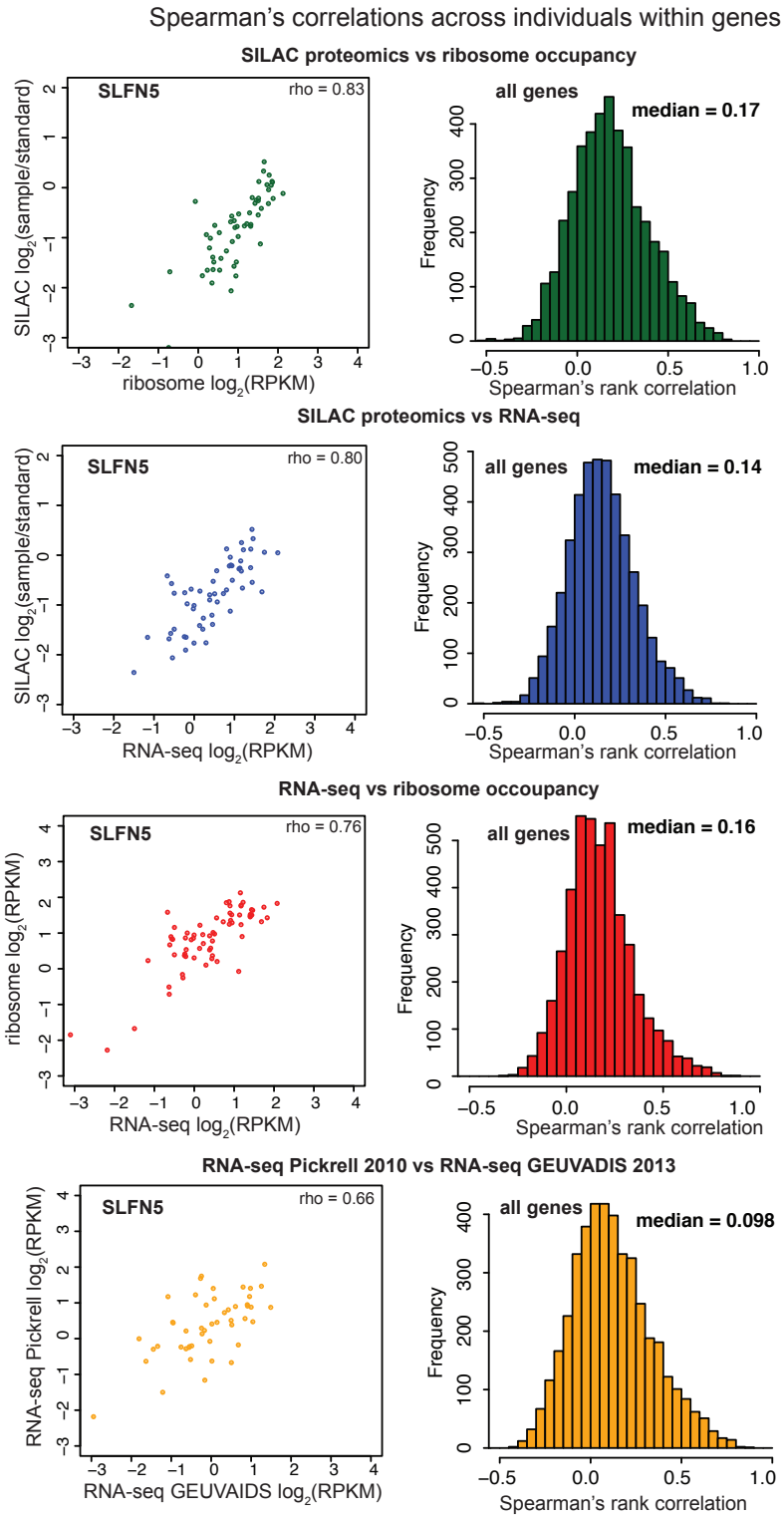
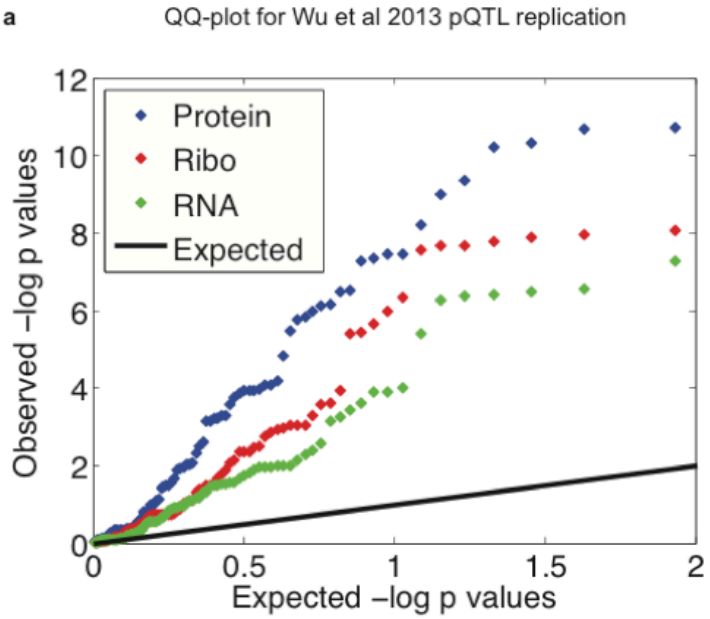


Figure S10. Correlations across individual (within genes). For each gene, we computed Spearman's correlation between phenotypes across individuals. Here, it was appropriate to use SILAC $\log_2$ ratios for protein data. The left column shows scatter plot of individual expression values for SLFN5/ENSG00000166750 comparing each pair of phenotypes as well as comparing RNA-seq data from Pickrell et al 2010 (*30*) versus RNA-seq data from the GEUVADIS study (*2*) for 52 lymphoblastoid cell lines shared between these two studies. The right column shows the Spearman's rank correlations for 4,322 genes quantified in all three phenotypes. Note that there is only little regulatory variation between individuals. For most genes there is little change in phenotype between individuals and thus the relationship between phenotypes is predominated by error and no significant correlation is expected. For a subset of genes, there are differences in the phenotypes across individuals resulting in significant correlations.

a



b **Replication of known human pQTLs**

Phenotype	Replication (FDR 0.05)	Replication (FDR 0.1)	Replication (FDR 0.2)
RNA	0.42	0.50	0.60
Ribosome occupancy	0.47	0.51	0.71
Protein level	**0.70**	**0.88**	**1.00**

Figure S11. Replication of pQTLs from Wu et al (2013) in protein, ribosome occupancy, and RNA-seq measurements. Previously detected pQTLs were obtained from Wu et al. (*15*), using QTLs identified in the combined populations, and filtered for MAF > 0.10 and expression in all three phenotypes in our population. We tested each of the resulting 86 SNP-gene pairs in data from our three phenotypes, using raw data (no PCs removed, no quantile normalization), recording the nominal p-values from Pearson correlation tests. (a) QQ-plot of all p-values obtained in replication testing, displaying enrichment in all three phenotypes, with the strongest replication in protein. (b) Replication rate of previously detected pQTLs in each of our three phenotypes at varying FDR cutoffs.

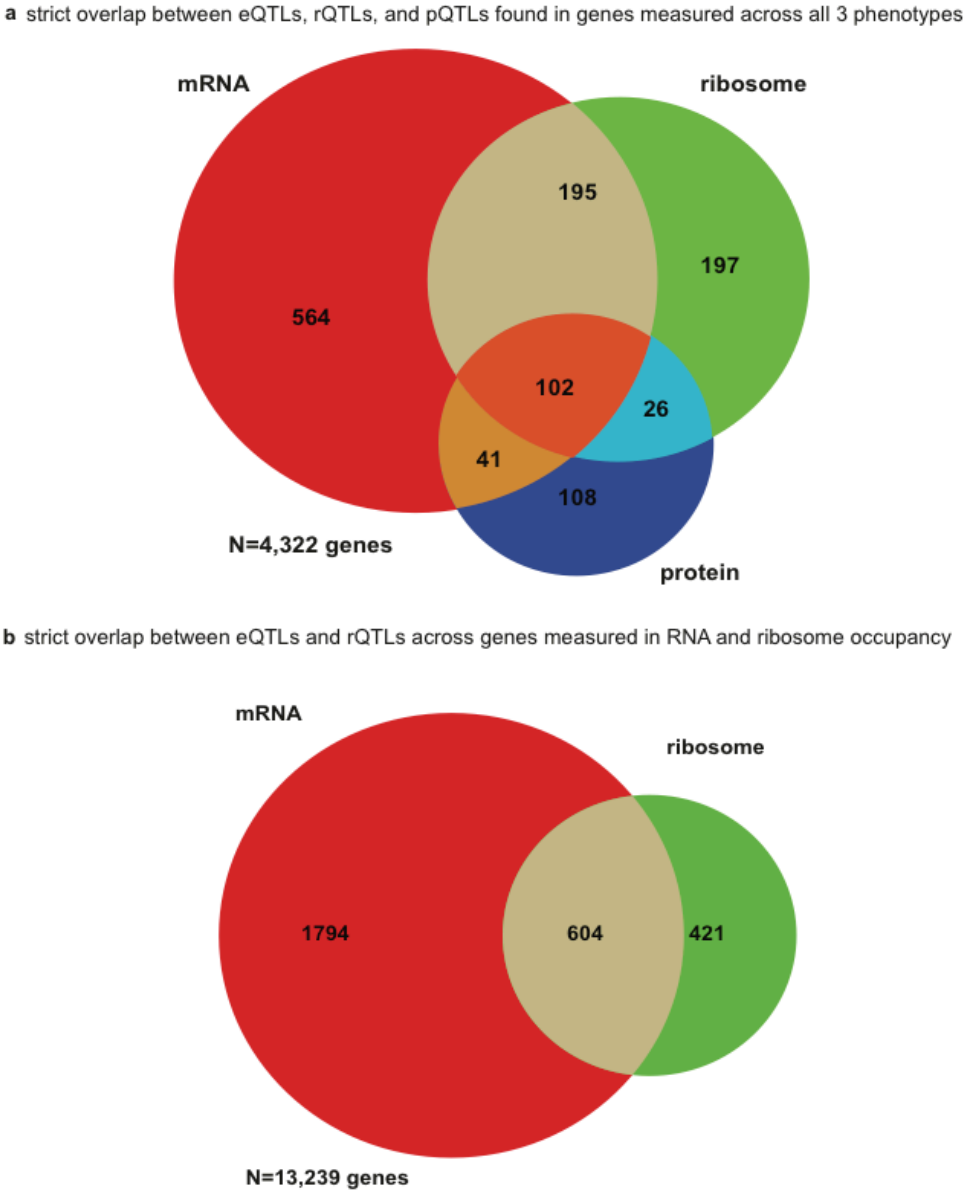


Figure S12. Overlap between genes with cis-eQTLs, cis-rQTLs, and cis-pQTLs. Here, we define overlap as detection of QTL for the same gene, without requiring the same SNP to be selected for different phenotypes. (a) QTL overlap among 4,322 genes tested in all three phenotypes. (b) QTL overlap among 13,239 genes tested in mRNA and ribosome occupancy.

Replication rates of cis-eQTLs, cis-rQTLs for N=13,329 genes quantified by ribosome profiling and RNA-seq

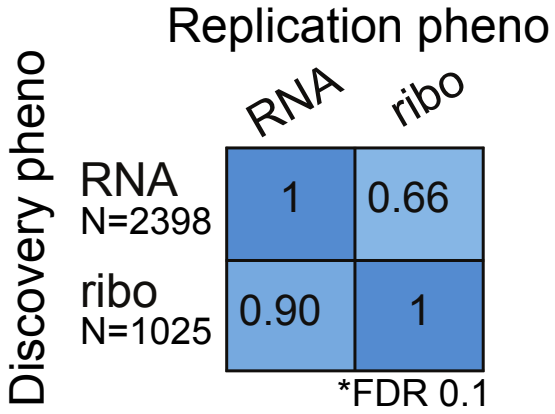


Figure S13. *cis*-QTL replication rate between RNA and ribosome occupancy. Here, we consider the 13,239 genes expressed in both, without considering protein (thereby allowing us to consider a larger set of genes). QTLs detected for the phenotype labeled on each row are tested in the phenotype listed for each column. FDR is estimated across resulting p-values only for the specific SNP-gene pairs included in each replication test.

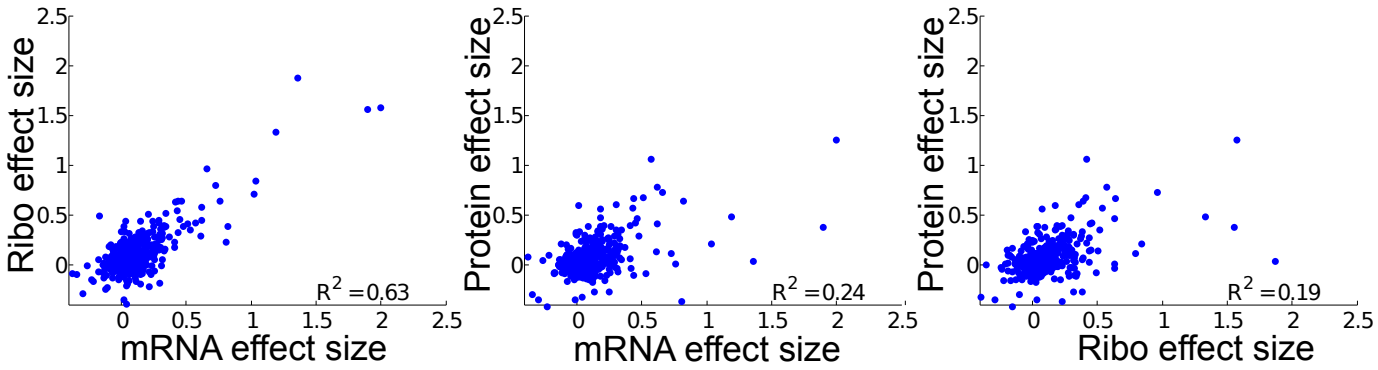


Figure S14. Effect sizes of GEUVADIS eQTLs compared in mRNA, ribosome occupancy, and protein. Here, we compare effect sizes for each individual eQTL previously ascertained in the GEUVADIS study, with comparisons between each pair of expression phenotypes (plots show fold change on $\log_2$ scale). We measured effect size as the regression coefficient obtained from linear regression using raw data for each of our three phenotypes (no quantile normalization or PC-correction, but mRNA and ribosome occupancy values were log-transformed). One outlier data point is not shown on the protein scatter plots, with an effect size of 3.5 in protein, and 1.0 and 0.7 in mRNA and ribo respectively. Strong correlation and no effect size compression is observed between mRNA and ribosome occupancy, whereas effect sizes appear compressed in protein data.

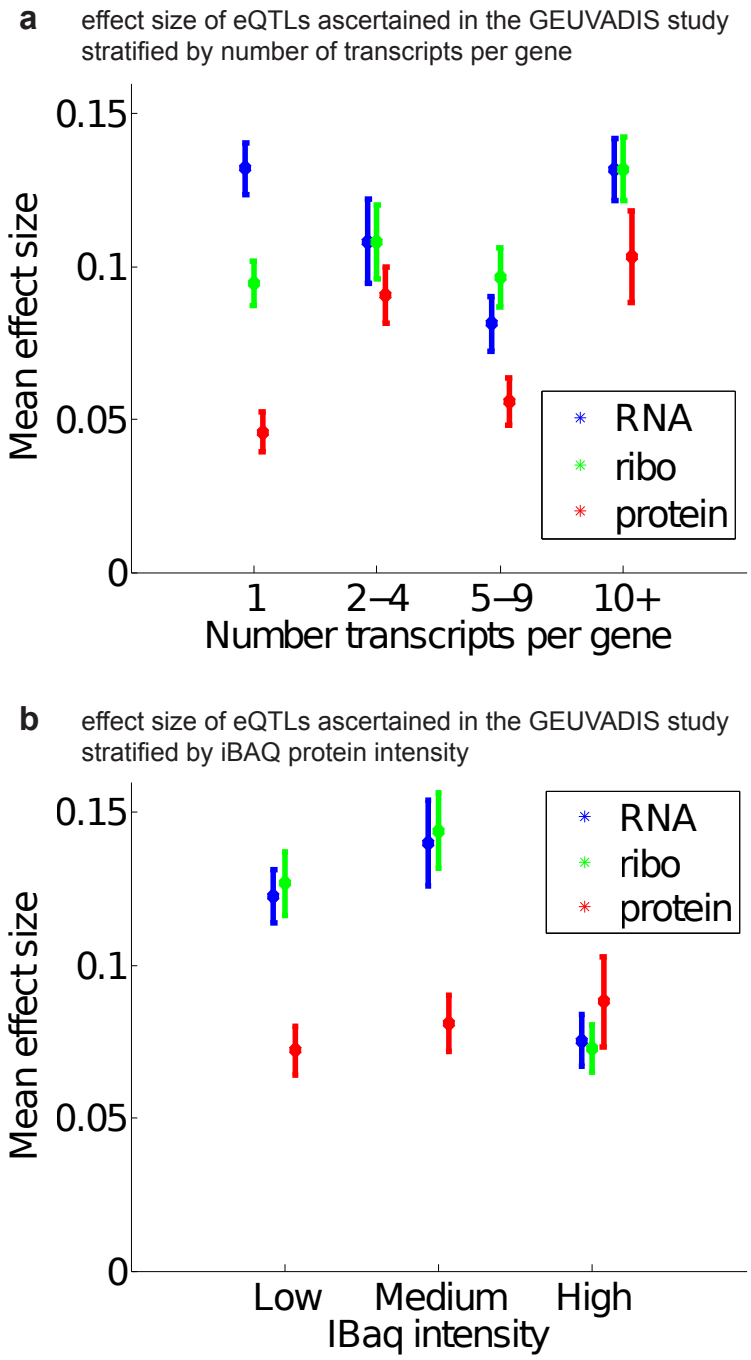


Figure S15. Mean effect sizes of eQTLs ascertained in the GEUVADIS study (a) stratified by number of expressed transcripts per gene (as determined by the number of expressed transcripts quantified in the GEUVADIS study itself) and (b) stratified by iBAQ protein intensity, where genes were grouped into three equally sized buckets based on median intensity across measured individuals. Effect sizes are estimated using standard linear regression on data without quantile normalization or PC correction. Error bars represent standard error of the mean.

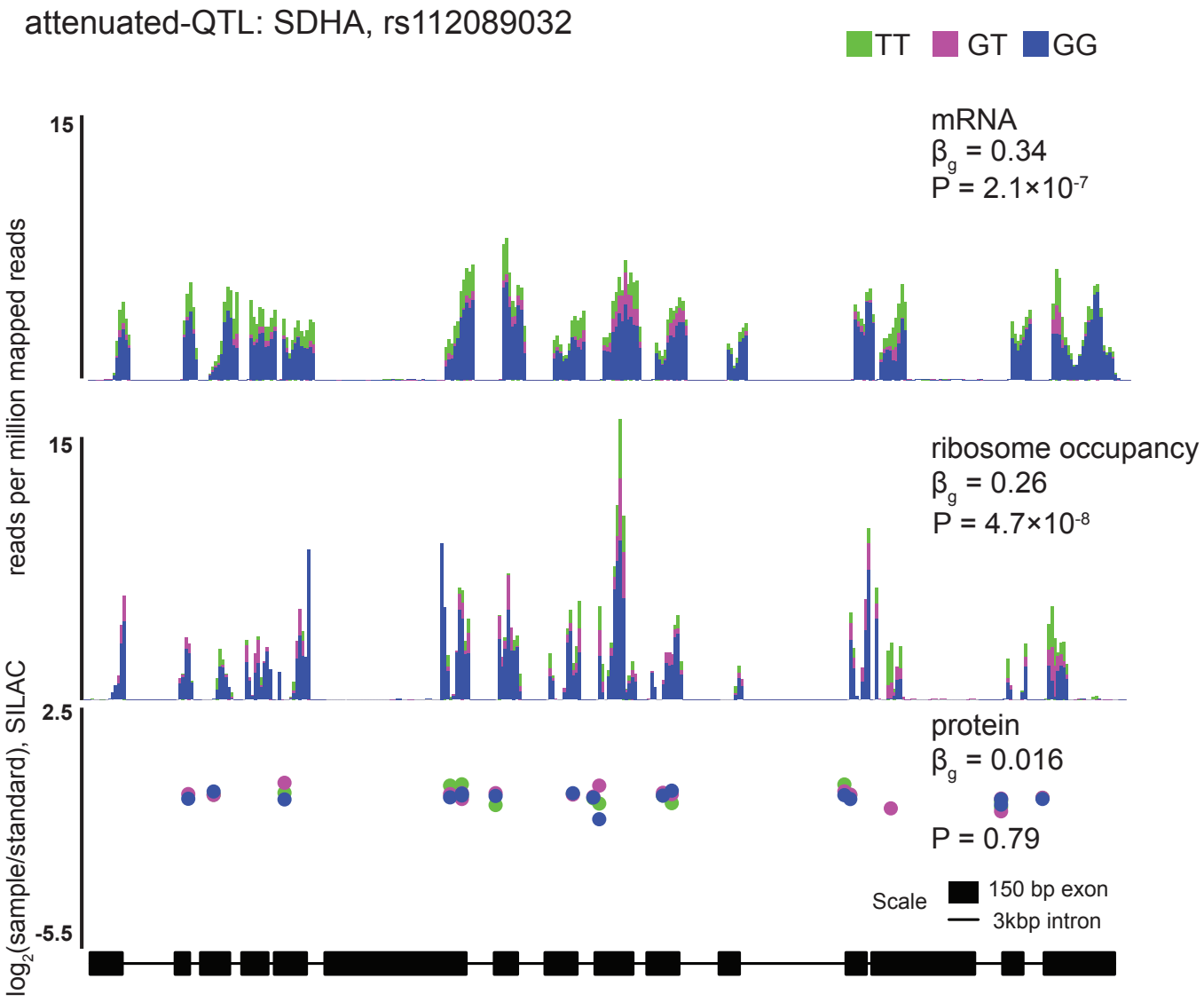
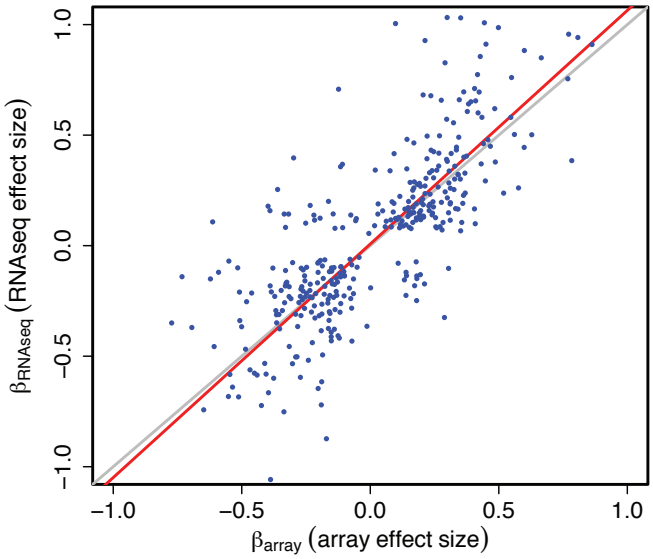


Figure S16. Example of a gene where the effect of an eQTL is attenuated at the protein level. Here, for mRNA and ribosome occupancy, RPKM at each 15-bp window along the gene is shown, using the mean among individuals with each genotype for the QTL SNP. SILAC log ratios at each detected peptide are shown for protein data, using the median among individuals with each genotype. Effect sizes (β) are shown for each phenotype derived from linear regression in data without quantile normalization or PC-correction.

a comparison of eQTL effect sizes estimated using microarray and RNA-seq data



b overlap between psQTLs

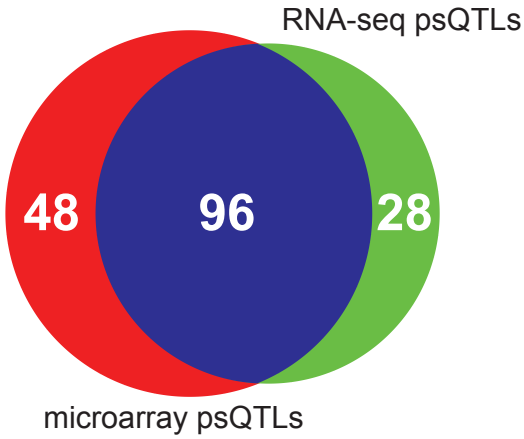


Figure S17. Comparison of eQTL effect sizes across technologies. (a) We used normalized microarray data to identify eQTLs at FDR of 10% [untreated time point from (*43*)]. To allow comparison of cases where a QTL effect could be detected by both technologies, we additionally required that these array-identified eQTLs also have a nominal p-value of less than 0.2 in RNA-seq data. For these eQTLs present in both data sets, we observed a Spearman's correlation of 0.82 and an R^2 0.47 across technologies and slope of 1.05. Gray shows a line with slope 1. Red line is the line of best fit to the effect size data. (b) illustrates the overlap between psQTLs identified across technologies using the conditional model. Out of 146 sequencing-based psQTLs identified by the conditional model, 124 were among the genes tested using arrays, and of these, 96 (77%) were among the array-based psQTLs at FDR 10%. Note that the array data were collected from independently thawed cultures of the same LCLs. Thus, this analysis indicates that our results are robust with respect to cell culturing, cell line passage numbers, RNA extractions, and the specific technology used to collect the transcript expression data.

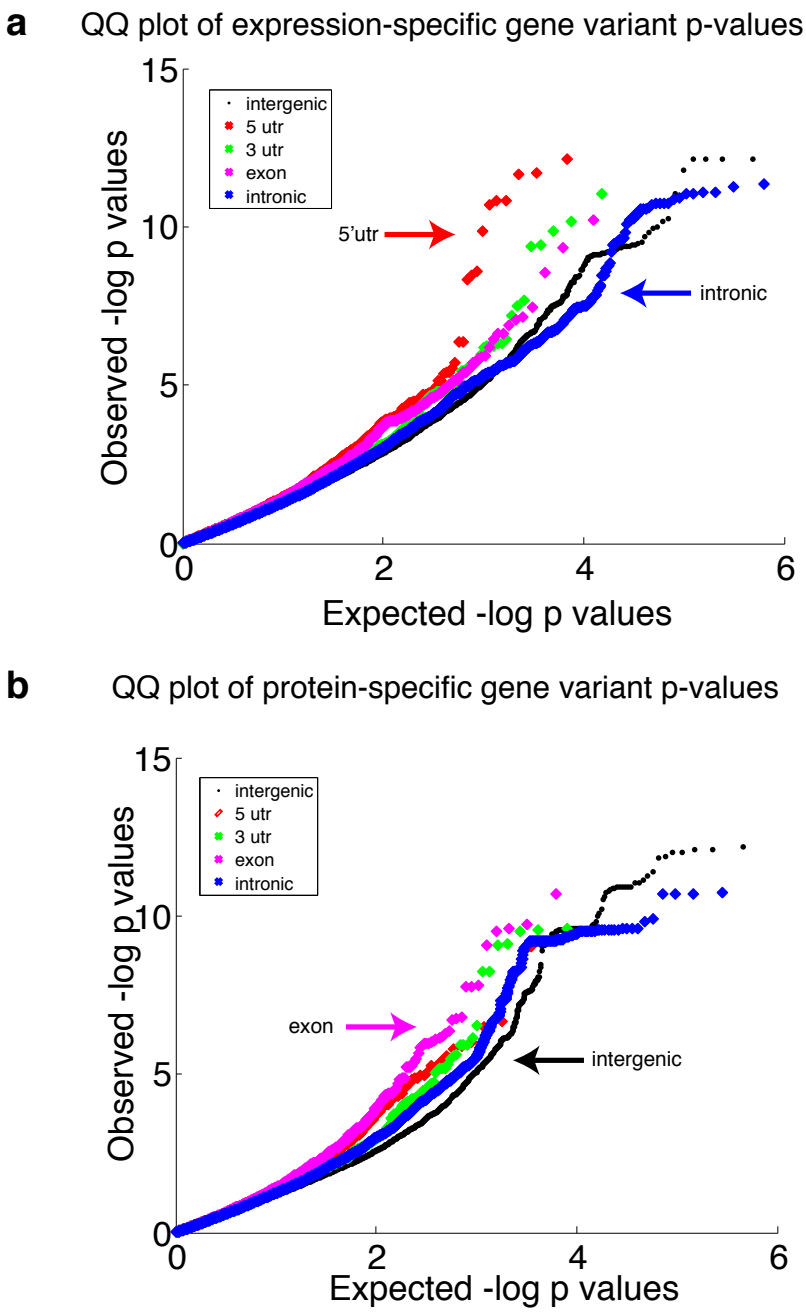


Figure S18. Enrichment of genomic annotations of SNPs associated with expression-specific or protein-specific QTLs. Using the conditional model of phenotype-specificity, nominal p-values were recorded for every SNP-gene pair tested (MAF >= 0.05, excluding HLA region of chromosome 6) for both protein specific and mRNA-specific QTLs. Subsequently, we evaluated the distribution of recorded p-values for all SNP-gene pairs matching each of the annotations shown.

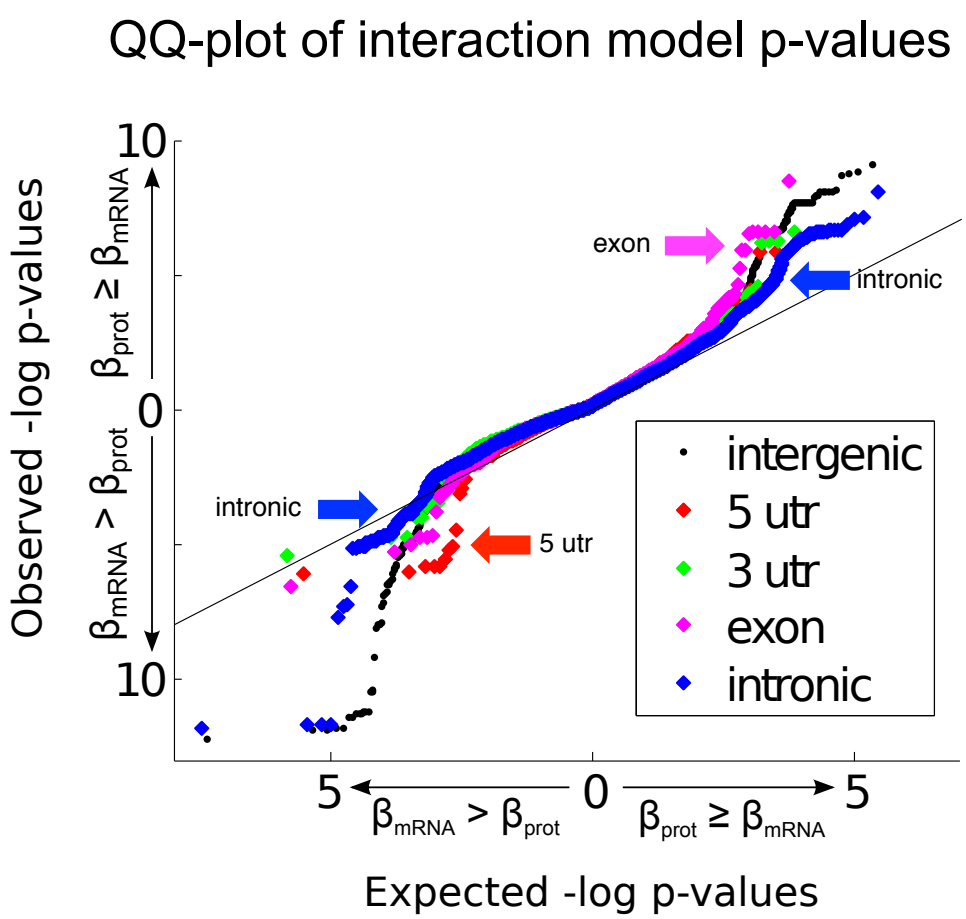


Figure S19. Enrichment analysis of genomic annotations of variants associated with differing effect sizes between mRNA and protein. Nominal p-values were recorded for every SNP-gene pair tested (MAF >= 0.05, excluding HLA region of chromosome 6) using the interaction model to detect significantly different effect sizes between mRNA and protein. Sign (positive or negative) was assigned according to the phenotype with larger magnitude of effect – cases with larger effect in protein are recorded as positive, and larger effect in mRNA as negative.

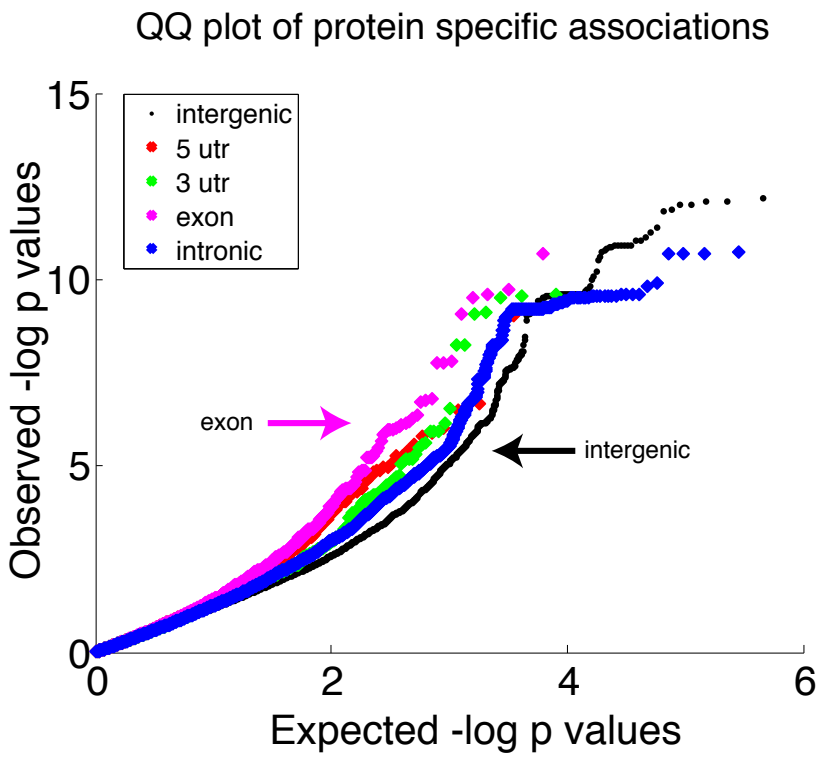


Figure S20. Enrichment analysis of genomic annotations associated with protein-specific changes excluding genes potentially biased by undetected peptides at non-synonymous sites. Genes for which more than 5 individuals were quantified using fewer than three peptides were excluded from this analysis, in order to reduce the chance that non-synonymous variants could bias protein quantification, which relies on a median across detected peptides.

Non-significant genomic annotations among expression and protein specific QTLs

Annotation	N	Background	psQTLs	esQTLs
PolyPhen deleterious	505	Non-synonymous	5.7e-1	8.9e-2
Disordered regions	586	Non-synonymous	1.5e-1	2.8e-1
Phosphorylation sites	211	Non-synonymous	8.5e-1	4.7e-1
Ubiquitination sites	61	Non-synonymous	6.5e-1	9.1e-1
Post-translational mod sites	267	Non-synonymous	8.7e-1	5.3e-1
Pfam domains	370	Non-synonymous	4.5e-1	3.1e-1
uORF	38	Transcript	2.1e-1	7.5e-1
N-terminus	87	Non-synonymous	8.8e-1	4.6e-1

Figure S21. Additional enrichment analyses performed using conditional model p-values. Enrichment for expression-specific and protein-specific QTLs according to genomic annotation, based on QTL results from the conditional model. Columns in order describe: the name of the tested annotation, the number of SNPs matching this annotation, the set of SNPs used as background control for the corresponding enrichment test, the enrichment p-value for protein-specific QTLs, and the enrichment p-value for expression-specific QTLs.

Sample	Total Number of Reads Collected	Reads mapping to rRNA, tRNA, and snoRNA	Uniquely mapped reads after filtering	Filtered reads mapping to coding regions	Sample	Total Number of Reads Collected	Reads mapping to rRNA, tRNA, and snoRNA	Uniquely mapped reads after filtering	Filtered reads mapping to coding regions
GM18486	114849676	95870594	15149205	13665725	**GM19108**	62493733	47201512	12108455	10859786
GM18498	108354773	89811996	9568928	8472492	**GM19114**	99694179	78387573	12595765	11219777
GM18499	97711371	58554141	15841149	13472359	**GM19116**	106004041	84782231	12047102	10217929
GM18501	117867014	90640652	14883115	13179159	**GM19119**	70884831	54877585	8347070	7345633
GM18502	87589080	53986775	13622030	11244210	**GM19127**	59469673	47628158	6029934	5369655
GM18504	117650230	85940870	12816442	11182798	**GM19128**	90142213	79809359	5335942	4634869
GM18505	110993723	72859933	11662850	9899988	**GM19130**	95905734	77907252	10107057	8803403
GM18507	70862603	48428696	8498245	7750134	**GM19131**	81415181	62776198	10389168	9107523
GM18508	79497338	45627550	15309194	13739557	**GM19137**	67313591	54179691	6751387	5770633
GM18510	94734968	85006651	7022039	6107812	**GM19138**	81424910	64812465	9833272	8469830
GM18511	87102615	72935936	9925784	8945070	**GM19140**	118901192	92905578	15409363	13095012
GM18516	70754219	52745035	4884145	4316681	**GM19141**	98866261	87029213	5351023	4757622
GM18517	120662231	90834551	17446164	15706403	**GM19143**	96997074	80173987	9717545	8494974
GM18519	63271895	51454381	6489051	5440274	**GM19144**	67112164	49685014	8566627	7561704
GM18520	34300235	26896376	3219692	2647057	**GM19147**	78446248	59853977	9976432	8944314
GM18522	88476111	57632090	10305027	8401217	**GM19152**	56397052	50389500	2253707	1954417
GM18523	99732920	78171783	18584382	16606148	**GM19153**	83053311	65613921	10842200	9379727
GM18852	127377957	95313202	17198446	14980214	**GM19159**	115668557	91211119	20276984	17770932
GM18853	116487029	96882898	16074475	14454285	**GM19160**	68430182	55253623	7492201	6533224
GM18855	91568161	74662954	8038673	7160668	**GM19171**	102890455	75210939	23663605	21144721
GM18856	88402910	77586558	8844436	8049358	**GM19172**	122558606	101256936	17110575	15326023
GM18858	128130772	94371404	20123552	17897079	**GM19190**	101063259	88863471	9039899	7673884
GM18861	102275229	77648081	14757463	12797893	**GM19192**	80871130	69094795	6868160	6050863
GM18862	125693244	97082633	16389881	14398950	**GM19193**	75791434	47293898	9747870	8506471
GM18870	111445793	88891169	18738947	17305453	**GM19200**	48673370	38874114	6196680	5371588
GM18907	98179931	82597710	13458930	11875594	**GM19201**	116680175	89859034	22310111	19566118
GM18909	98438571	74825478	13654499	12020903	**GM19204**	115848739	85527209	17105214	15153312
GM18912	100576982	79804970	11481320	9954180	**GM19207**	104820602	84050629	18090266	16446813
GM18913	101583481	73580311	16389513	14427004	**GM19209**	60944423	46907141	7129643	6390912
GM18916	130144045	102452968	14914501	11557128	**GM19210**	86032295	67247537	16353653	14191205
GM19092	114308515	92815559	18645907	16619712	**GM19222**	114396231	95294203	15526067	13779240
GM19093	102326488	83531712	10405480	8964584	**GM19223**	107528516	91856999	12585714	11279977
GM19098	127674251	100491923	15896845	13313218	**GM19225**	74918709	57385775	14890125	13765614
GM19099	108056102	82334557	15193312	13333467	**GM19238**	107922509	91092210	13844684	12107124
GM19101	80114177	66702659	7364753	6562651	**GM19239**	107600409	81935336	22812808	20596330
GM19102	85178108	74529690	8527635	7127863	**GM19257**	95685320	82720501	6870258	6006713

Table S1. Total number of reads collected using ribosome profiling, before and after filtering of reads mapping to rRNA, tRNA, and snoRNA. Data were collected using an Illumina HiSeq 2500.

Sample	MS1 Spectra	MS2 Spectra	Sample	MS1 Spectra	MS2 Spectra
GM18486	79641	406135	**GM19093**	77666	401160
GM18498	82156	360025	**GM19098**	94825	430867
GM18499	82665	397846	**GM19099**	78158	405295
GM18501	76419	416852	**GM19101**	82097	383419
GM18502	78810	418257	**GM19108**	79505	387955
GM18504	77981	413280	**GM19114**	86912	367480
GM18505	62497	473147	**GM19116**	88452	371406
GM18507	91999	422500	**GM19119**	79117	397599
GM18508	78615	410975	**GM19127**	80729	377168
GM18510	83102	384761	**GM19128**	91685	400370
GM18511	83716	372522	**GM19130**	87111	358985
GM18516	110208	405340	**GM19131**	84555	375284
GM18517	81348	385496	**GM19137**	75884	395996
GM18519	78710	392546	**GM19138**	79146	384863
GM18520	80456	379870	**GM19140**	82386	395242
GM18522	76100	403298	**GM19143**	85758	332275
GM18523	80074	379407	**GM19144**	82333	366447
GM18852	86952	360181	**GM19147**	76623	430195
GM18855	83532	387870	**GM19152**	80940	403950
GM18858	80586	393731	**GM19153**	80990	381426
GM18861	78838	415000	**GM19160**	83937	397055
GM18862	79315	413004	**GM19172**	96629	341215
GM18870	76820	423566	**GM19192**	83069	373635
GM18871	85244	378408	**GM19193**	72124	382871
GM18907	78646	384382	**GM19200**	97266	339044
GM18909	81281	378786	**GM19203**	79353	437706
GM18912	85516	385926	**GM19204**	70686	447980
GM18913	78426	392084	**GM19207**	76317	400830
GM18916	78402	384338	**GM19209**	83676	386591
GM19102	83476	362698	**GM19222**	86711	390913
GM19092	83221	364612	**GM19257**	79493	423819
GM19138 run2	82586	363808	**GM19114 run2**	80185	404282
GM18855 run2	79353	437706		-	-

Table S2. Raw protein mass spectrometry data. Total number of high-resolution MS1 spectra and MS2 fragmentation spectra collected using an LTQ-Orbitrap Velos instrument. Spectra were collected using a 1-hour LC gradient where the fifteen most abundant ions were selected for fragmentation.

Sample	Events	Peptides	Proteins	Sample	Events	Peptides	Proteins
GM18486	73785	46500	5163	**GM19093**	66681	44315	4992
GM18498	54942	36665	4596	**GM19098**	77520	45441	4882
GM18499	71667	45319	5013	**GM19099**	66752	43598	4929
GM18501	69763	46412	4964	**GM19101**	62859	40543	4740
GM18502	76046	49914	5286	**GM19108**	60831	40330	4692
GM18504	73184	48263	5027	**GM19114**	60402	40635	4779
GM18505	81596	47755	5021	**GM19116**	63588	39341	4651
GM18507	72836	44737	4863	**GM19119**	68839	43792	4894
GM18508	76934	48204	5100	**GM19127**	67693	42482	4825
GM18510	67716	42571	5031	**GM19128**	63811	40067	4793
GM18511	65073	40999	4923	**GM19130**	60296	38074	4549
GM18516	64492	38810	4728	**GM19131**	56263	36954	4601
GM18517	59146	39040	4805	**GM19137**	64600	41479	4953
GM18519	60890	40917	4795	**GM19138**	65020	39867	4602
GM18520	63443	42322	4948	**GM19140**	66568	43814	4946
GM18522	57648	38531	4671	**GM19143**	59370	36690	4510
GM18523	63404	40339	4693	**GM19144**	58209	36466	4441
GM18852	59645	38107	4694	**GM19147**	78452	46417	4972
GM18855	52655	36050	4603	**GM19152**	64853	41436	4763
GM18858	60317	39792	4679	**GM19153**	67665	40682	4851
GM18861	74539	47119	5097	**GM19160**	72656	43289	5099
GM18862	68326	44469	4893	**GM19172**	54113	37388	4684
GM18870	73122	46000	5062	**GM19192**	53097	33128	4674
GM18871	52245	31175	3985	**GM19193**	62661	41012	4650
GM18907	58743	39189	4771	**GM19200**	56703	38073	4676
GM18909	63256	41418	4828	**GM19203**	74613	46538	4989
GM18912	68487	43192	4785	**GM19204**	76908	46565	5052
GM18913	64458	41856	4753	**GM19207**	67086	41311	4895
GM18916	62296	41596	4785	**GM19209**	60033	36390	4808
GM19012	67048	40593	4822	**GM19222**	71263	42620	5011
GM19092	59946	39892	4744	**GM19257**	71362	41377	4808
GM19138 run2	65154	41812	4981				
GM18855 run2	76585	46721	5057	**GM19114 run2**	72305	45130	4906

Table S3. Quantified peptide and protein levels. Peptide spectrum matches were assigned at a false discovery rate of 1% and the number of quantification events, distinct quantified peptides, and distinct quantified proteins per sample were tabulated above. See methods for low-level peptide and protein quantification parameters.

Supplementary Data Table Descriptions

Supplementary Data Table 1. cis-QTL mapping results.

ENSG: Ensembl gene identifier.
chr: chromosome identifier.
perm.p.value: empirical gene-level p-value for the most significant SNP based on 10,000 permutations of genotype sample labels
snp.pvalue: Pearson correlation p-value between the expression phenotype and genotype at the most significant SNP for the gene
snp.R.value: Pearson correlation between the expression phenotype and the genotype at the most significant SNP for the gene
hg18.pos: position of the most significant SNP

eQTLs tab: Permutation p-values obtained for normalized and PC-corrected RNA-seq data.
rQTLs tab**:** Same as above, but for ribosome profiling data.
pQTLs tab**:** Same as above, but for quantitative proteomics data.

Supplementary Data Table 2. Significant psQTLs and esQTLs identified by the conditional model

ENSG: Ensembl gene identifier.
chr: chromosome identifier.
perm.p.value: empirical gene-level p-value for the most significant SNP based on 10,000 permutations of genotype sample labels
hg18.pos: position of the most significant SNP

esQTLs_protein tab: expression-specific QTLs identified by conditioning on protein level
psQTLs_rna tab: protein-specific QTLs identified by conditioning on RNA level
psQTLs_ribo_rna tab: protein-specific QTLs identified by conditioning both on RNA and ribosome profiling data

Supplementary Data Table 3. psQTLs and esQTLs identified by the interaction model

pQTLs and eQTLs where a significant difference between effect sizes across phenotypes was detected.

ENSG: Ensembl gene identifier.
snp.hg18.position: hg18 position of the SNP.
p.inter.rna.pro: interaction model p-value
beta.pro: Estimated regression effect size from quantitative proteomics data.
beta.rna: Same as above, but for RNA-seq data.
beta.ribo: Same as above, but for ribosome profiling data.
beta.pro.low.95, beta.pro.high.95: 95% confidence interval for regression beta, proteomics data
beta.rna.low.95, beta.rna.high.95: Same as above, but for RNA-seq.
beta.ribo.low.95, beta.ribo.high.95: Same as above, but for ribosome profiling data.

Supplementary Data Table 4. Quantitation for each gene in each individual for protein, RNA-seq and ribosome profiling data.

ENSG: Ensembl gene identifier.
GMxxxxx: designates HapMap LCL identifier
txStart: hg18 position of transcription start site
txEnd: hg18 position of transcription end site
chr: chromosome identifier
gene.len: length of the transcribed region of the gene for canonical Ensembl transcript in base pairs

RNA-seq and ribosome profiling quantifications are expressed as $\log_2$(counts per million mapped reads). They can be converted to $\log_2$(reads per kilobase transcript per million mapped reads) by subtracting $\log_2$(gene.len/10^3). For both RNA-seq and ribosome profiling data genes where the median row counts equal to zero were removed. Protein level quantifications are expressed as the $\log_2$(sample/standard) SILAC ratio for the gene.

REFERENCES

1. A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, D. Koller, Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014). Medline doi:10.1101/gr.155192.113

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

3. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector; Multiple Tissue Human Expression Resource (MuTHER) Consortium, Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012). Medline doi:10.1038/ng.2394

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

5. A. Ghazalpour, B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P. Z. Wen, H. Brewer, K. Weitz, D. G. Camp 2nd, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, A. J. Lusis, Comparative analysis of proteome and transcriptome variation in mouse. *PLOS Genet.* **7**, e1001393 (2011). Medline doi:10.1371/journal.pgen.1001393

6. E. J. Foss, D. Radulovic, S. A. Shaffer, D. M. Ruderfer, A. Bedalov, D. R. Goodlett, L. Kruglyak, Genetic basis of proteome variation in yeast. *Nat. Genet.* **39**, 1369–1375 (2007). Medline doi:10.1038/ng.2007.22

7. P. Picotti, M. Clément-Ziza, H. Lam, D. S. Campbell, A. Schmidt, E. W. Deutsch, H. Röst, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J. J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R. L. Moritz, A. Beyer, R. Aebersold, A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494**, 266–270 (2013). Medline doi:10.1038/nature11835

8. F. W. Albert, S. Treusch, A. H. Shockley, J. S. Bloom, L. Kruglyak, Genetics of single-cell protein abundance variation in large yeast populations. *Nature* **506**, 494–497 (2014). Medline doi:10.1038/nature12904

9. J. M. Laurent, C. Vogel, T. Kwon, S. A. Craig, D. R. Boutz, H. K. Huse, K. Nozue, H. Walia, M. Whiteley, P. C. Ronald, E. M. Marcotte, Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10**, 4209–4212 (2010). Medline doi:10.1002/pmic.201000327

10. S. P. Schrimpf, M. Weiss, L. Reiter, C. H. Ahrens, M. Jovanovic, J. Malmström, E. Brunner, S. Mohanty, M. J. Lercher, P. E. Hunziker, R. Aebersold, C. von Mering, M. O. Hengartner, Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLOS Biol.* **7**, e48 (2009). Medline doi:10.1371/journal.pbio.1000048

11. M. Stadler, A. Fire, Conserved translatome remodeling in nematode species executing a shared developmental transition. *PLOS Genet.* **9**, e1003739 (2013). Medline doi:10.1371/journal.pgen.1003739

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

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

14. Z. Khan, M. J. Ford, D. A. Cusanovich, A. Mitrano, J. K. Pritchard, Y. Gilad, Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**, 1100–1104 (2013). Medline doi:10.1126/science.1242379

15. L. Wu, S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, H. Tang, M. Snyder, Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013). Medline doi:10.1038/nature12223

16. R. J. Hause, A. L. Stark, N. N. Antao, L. K. Gorsic, S. H. Chung, C. D. Brown, S. S. Wong, D. F. Gill, J. L. Myers, L. A. To, K. P. White, M. E. Dolan, R. B. Jones, Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. *Am. J. Hum. Genet.* **95**, 194–208 (2014). Medline doi:10.1016/j.ajhg.2014.07.005

17. N. Garge, H. Pan, M. D. Rowland, B. J. Cargile, X. Zhang, P. C. Cooley, G. P. Page, M. K. Bunger, Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol. Cell. Proteomics* **9**, 1383–1399 (2010). Medline doi:10.1074/mcp.M900378-MCP200

18. N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009). Medline doi:10.1126/science.1168978

19. J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J. B. Veyrieras, M. Stephens, Y. Gilad, J. K. Pritchard, Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010). Medline doi:10.1038/nature08872

20. S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, M. Mann, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002). Medline doi:10.1074/mcp.M200025-MCP200

21. M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014). Medline doi:10.1038/nature13319

22. B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011). Medline doi:10.1038/nature10098

23. X. J. Yang, E. Seto, Lysine acetylation: Codified crosstalk with other posttranslational modifications. *Mol. Cell* **31**, 449–461 (2008). Medline doi:10.1016/j.molcel.2008.07.002

24. M. N. Lee, C. Ye, A. C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, L. D. Ward, K. Raddassi, C. McCabe, M. H. Lee, I. Y. Frohlich, D. A. Hafler, M. Kellis, S. Raychaudhuri, F. Zhang, B. E. Stranger, C. O. Benoist, P. L. De Jager, A. Regev, N. Hacohen, Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014). Medline doi:10.1126/science.1246980

25. J. F. Degner, A. A. Pai, R. Pique-Regi, J. B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, J. K. Pritchard, DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012). Medline doi:10.1038/nature10808

26. Y. Guan, M. Stephens, Practical issues in imputation-based association mapping. *PLOS Genet.* **4**, e1000279 (2008). Medline doi:10.1371/journal.pgen.1000279

27. P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006). Medline doi:10.1086/502802

28. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu,

H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. Vernon Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart; International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). Medline doi:10.1038/nature06258

29. G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean; 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). Medline doi:10.1038/nature09534

30. J. K. Pickrell, A. A. Pai, Y. Gilad, J. K. Pritchard, Noisy splicing drives mRNA isoform diversity in human cells. *PLOS Genet.* **6**, e1001236 (2010). Medline doi:10.1371/journal.pgen.1001236

31. N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, J. S. Weissman, The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012). Medline doi:10.1038/nprot.2012.086

32. J. D. Storey, A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002). doi:10.1111/1467-9868.00346

33. J. D. Storey, J. E. Taylor, D. Siegmund, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Series B Stat. Methodol.* **66**, 187–205 (2004). doi:10.1111/j.1467-9868.2004.00439.x

34. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 7507–7512 (2009). Medline doi:10.1073/pnas.0810916106

35. Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, H. Y. Chang, Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014). Medline doi:10.1038/nature12946

36. D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42** (D1), D764–D770 (2014). Medline doi:10.1093/nar/gkt1168

37. R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, Pfam: The protein families database. *Nucleic Acids Res.* **42** (D1), D222–D230 (2014). Medline doi:10.1093/nar/gkt1223

38. P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, M. Sullivan, PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40** (D1), D261–D270 (2012). Medline doi:10.1093/nar/gkr1122

39. Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005). Medline doi:10.1093/bioinformatics/bti541

40. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010). Medline doi:10.1038/nmeth0410-248

41. X. Liu, X. Jian, E. Boerwinkle, dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013). Medline doi:10.1002/humu.22376

42. Z. Khan, J. S. Bloom, S. Amini, M. Singh, D. H. Perlman, A. A. Caudy, L. Kruglyak, Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Mol. Syst. Biol.* **8**, 602 (2012). 10.1038/msb.2012.34 Medline doi:10.1038/msb.2012.34

43. A. A. Pai, C. E. Cain, O. Mizrahi-Man, S. De Leon, N. Lewellen, J. B. Veyrieras, J. F. Degner, D. J. Gaffney, J. K. Pickrell, M. Stephens, J. K. Pritchard, Y. Gilad, The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLOS Genet.* **8**, e1003000 (2012). Medline doi:10.1371/journal.pgen.1003000