Published in final edited form as:

Nature. 2016 June 23; 534(7608): 500–505. doi:10.1038/nature18270.

Defining the consequences of genetic variation on a proteome–wide scale

Joel M. Chick[1,*], **Steven C. Munger**[2,*], **Petr Simecek**[2], **Edward L. Huttlin**[1], **Kwangbom Choi**[2], **Daniel M. Gatti**[2], **Narayanan Raghupathy**[2], **Karen L. Svenson**[2], **Gary A. Churchill**[2,§], and **Steven P. Gygi**[1,§]

[1]Harvard Medical School, Boston, Massachusetts 02115, USA.

[2]The Jackson Laboratory, Bar Harbor, Maine 04609, USA.

Abstract

Genetic variation modulates protein expression through both transcriptional and post-transcriptional mechanisms. To characterize the consequences of natural genetic diversity on the proteome, here we combine a multiplexed, mass spectrometry-based method for protein quantification with an emerging outbred mouse model containing extensive genetic variation from eight inbred founder strains. By measuring genome-wide transcript and protein expression in livers from 192 Diversity outbred mice, we identify 2,866 protein quantitative trait loci (pQTL) with twice as many local as distant genetic variants. These data support distinct transcriptional and post-transcriptional models underlying the observed pQTL effects. Using a sensitive approach to mediation analysis, we often identified a second protein or transcript as the causal mediator of distant pQTL. Our analysis reveals an extensive network of direct protein–protein interactions. Finally, we show that local genotype can provide accurate predictions of protein abundance in an independent cohort of collaborative cross mice.

Regulation of protein abundance is vital to cellular functions and environmental response. According to the central dogma[1], the coding sequence of DNA is transcribed into mRNA

Reprints and permissions information is available at www.nature.com/reprints

Correspondence and requests for materials should be addressed to S.P.G. (steven_gygi@hms.harvard.edu) or G.A.C. (gary.churchill@jax.org).

[*]These authors contributed equally to this work.

[§]These authors jointly supervised this work.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Supplementary Information is available in the online version of the paper.

Author Contributions S.P.G. and G.A.C. conceived and planned the study. J.M.C. performed the proteomics sample preparation and mass spectrometry analysis. S.C.M., K.C., N.R., P.S. and D.M.G. performed the RNA-seq, genetic mapping, model selection and mediation analyses. S.C.M. and G.A.C. conceived the mediation analysis. E.L.H. performed the network analysis for distant pQTL. K.L.S. and G.A.C. provided DO and CC mice. J.M.C., S.C.M., G.A.C. and S.P.G. interpreted the scientific findings, discussed and wrote the paper with input from all authors.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD002801 (http://www.proteomexchange.org/). Raw RNA-seq fastq files and processed gene-level data are archived at Gene Expression Omnibus (GEO) under accession number GSE72759. We implemented our mediation method as the R package, intermediate, which can be freely downloaded from http://github.com/churchill-lab/intermediate. The Genotyping by RNA-seq (GBRS) software is available for download from https://github.com/churchill-lab/gbrs.

The authors declare no competing financial interests.

(transcript), which in turn is translated into protein. Although rates of transcription, translation and degradation of both transcript and protein vary, under this simplest model of regulation, the cellular pool of a protein is determined by the abundance of its corresponding transcript. Genetic or environmental perturbations that alter transcription would directly affect protein abundance. In reality, many layers of regulation intervene in this process, and numerous studies have been carried out to determine whether and to what extent transcript abundance is a predictor of protein abundance[2–6]. Several studies have reported that there is generally a low correlation between the two. An emerging consensus is that much of the protein constituent of the cell is buffered against transcriptional variation[4,7], but a global perspective of protein buffering has not been put forward.

Genetic variants can influence transcript and protein levels in a quantitative manner. Mapping quantitative trait loci (QTL) that affect transcript (eQTL) or protein (pQTL) abundance in model organisms or human cell lines can identify causal variants and provide a tool to dissect the mode of regulation of gene expression[8]. Analyses of eQTL have yielded a global but incomplete understanding of the regulatory mechanisms associated with gene expression[9–13]. Until now, pQTL analysis has been applied to a modest set of proteins through shotgun proteomics or targeted protein analysis[5,7,14–19]. Much of the pioneering work behind pQTL analysis has been conducted in yeast crosses using mass spectrometry[14–16] or green fluorescent protein (GFP) fusions[20]. Recent advances in quantitative proteomics[21,22] present the possibility of near-comprehensive, genome-wide pQTL analysis.

To investigate how genetic variation affects transcript and protein abundance globally requires a broad set of perturbations. The diversity outbred (DO) mouse model is a heterogeneous stock derived from the same eight founder strains as the collaborative cross (CC) mice[23–25] (Fig. 1a). The founder strains are fully sequenced[26] and capture a considerable cross-section of the genetic variation present in laboratory and wild mouse populations. The balanced allele frequencies and simple population structure of the DO mice provides high power and precision for mapping QTL with relatively small sample sizes relative to human mapping studies. We designed a QTL mapping approach that takes advantage of these unique properties of the DO and our knowledge of the founder genomes[27]. For each individual DO mouse, we imputed the founder strain ancestry at 64,000 evenly spaced loci across the diploid genome.

Gene and protein expression profiling

We first applied multiplexed proteomics to evaluate the extent of protein abundance variation among the eight DO/CC founder strains. Founder strain liver proteins were analysed in duplicate from both sexes (Extended Data Fig. 1a, Supplementary Table 1). Protein abundance was highly variable across the eight founder strains; hierarchical clustering and principal component analysis suggested that strain was the major factor driving variation followed by sex. This analysis confirmed our expectation that the wild-derived founder strains CAST/EiJ (CAST) and PWK/PhJ (PWK) were most distinct, underlying much of the genetic variability in protein expression (Extended Data Fig. 1b–d).

We next profiled protein and transcript levels in liver tissue from 192 DO mice, including both females and males, with half of the animals fed standard rodent chow and the other half fed a high-fat diet (Methods, Fig. 1b and Supplementary Tables 2 and 3). In total, we measured 6,756 proteins and 16,921 transcripts with detection in at least half of the samples. Both transcript and protein abundance were highly variable, and principal component analysis identified sex and diet as major drivers of this variation (Extended Data Fig. 2a). As expected, many proteins displayed sex- or diet-specific protein expression. Known female- and male-specific proteins were selectively expressed in a sex-dependent manner (Extended Data Fig. 2b, c). Likewise, many proteins showed diet-specific expression such as PPAR signalling, fat and cholesterol metabolism enzymes (Extended Data Fig. 2d, e), and many of these had concordant transcriptional responses (Extended Data Fig. 2f–j). These results demonstrate that sex and diet induced expected changes in transcript and protein expression.

Genetic regulation of protein abundance

In the subsequent analyses, we focused on 6,707 proteins for which both the protein and its corresponding transcript were detected in at least half of the DO liver samples. Genetic factors explained a substantial portion of variation in the abundance of protein and transcripts in the DO population (Extended Data Fig. 3a–f). To identify these, we performed QTL mapping analysis on transcript (eQTL, Supplementary Table 4) and protein (pQTL, Supplementary Table 5) abundance.

We identified 2,866 pQTL for 2,552 distinct proteins at a genomewide significance level of $P < 0.1$ (Fig. 2a). This is the largest set of pQTL identified so far, with tenfold greater numbers than other mass spectrometry (MS)-based approaches. Significant local pQTL were more common than distant pQTL (1,736 local and 1,130 distant pQTL) (Extended Data Fig. 3g). In addition, we identified 4,188 significant eQTL among 3,706 genes, with threefold more local than distant associations at the transcript level (3,211 local and 977 distant eQTL; Fig. 2a, Extended Data Fig. 3h, i). Finally, to examine the replication rate, we analysed a replication set of 192 separate DO mice treated under identical conditions for eQTL (see Methods and Extended Data Fig. 4).

To determine whether the same genetic loci acted on transcript and protein abundance, we first compared the QTL maps. We observed a significant overlap of proteins with pQTL and eQTL ($n = 1,400$; hypergeometric $P < 1 \times 10^{-16}$; Fig. 2a). As expected, genes with concordant QTL had generally higher correlations between protein and transcript abundance compared to those having only pQTL, only eQTL or neither (Fig. 2b). Among local QTL only, we observed a high degree of overlap with 80% of local pQTL having a corresponding local eQTL. The small number of local pQTL that lack corresponding eQTL ($n = 344$) could result from genetic variation that regulated protein abundance via post-transcriptional mechanisms such as coding variation that affected protein stability without altering transcript levels. In contrast, distant genetic variants that affected both transcript and protein levels seem to be nearly mutually exclusive (Fig. 2a). This observation leads to the intriguing hypothesis that most distant pQTL affected the abundance of a target protein via post-transcriptional mechanism(s).

For each of the 6,707 expressed proteins, we chose the most significant local and distant QTL, regardless of whether the log odds ratio (LOD) scores at each locus exceeded the pQTL detection threshold. We regressed out the transcript abundance and examined the effect on the peak LOD scores (Fig. 2c). Proteins with pQTL that are mediated through their corresponding transcript should show a reduced LOD score when transcript abundance is included in the regression model. Most local pQTL had significantly lower LOD scores after conditioning on their corresponding transcript (1,136 out of 1,736 dropped by ≥20%), while most distant pQTL were unaffected after conditioning on their transcript (164 out of 1,007 dropped by ≥20%). This suggests that local pQTL were largely mediated through transcriptional mechanisms, whereas distant pQTL were more likely to regulate protein abundance without affecting transcript abundance.

We carried out a model selection analysis using Bayesian information criterion (BIC) to identify the most probable path relating a locus genotype to a protein and its corresponding transcript. We evaluated all 6,707 proteins using the best local and distant markers identified in the pQTL mapping, and recorded the path that best explained the observed expression data (Fig. 2d, Extended Data Fig. 5 and Supplementary Tables 6 and 7). We illustrate these models in Fig. 2d in a more simplified form and present a more complete version of these models in Extended Data Fig. 5. Three of the models had no path connecting the locus to protein abundance. For most proteins, these were the best-fitting models for the local QTL ($n = 4{,}505$) and for the distant QTL ($n = 5{,}944$). The remaining models linked the abundance of a protein to either a local QTL or a distant QTL. Among local QTL, we found that most had effects that were mediated at least partially through the transcript ($n = 1{,}579$), while a minority affected protein abundance independently of the transcript ($n = 623$). Among distant QTL, a much smaller proportion acted through the transcript ($n = 17$), and most affected protein abundance independently of the transcript ($n = 746$). We conclude that most local pQTL affected both protein and transcript abundance, consistent with a transcriptional mode of regulation. However, distant pQTL affected protein abundance independently of the transcript, consistent with a posttranscriptional mode of regulation.

Local pQTL effect on protein abundance

We highlight two examples that illustrate the most common models of local regulation. DHTKD1 exemplifies a pQTL in which a local genetic variant affected transcript abundance that was transmitted to the protein (Fig. 3a, b). This simple transcript-to-protein model of regulation was evidenced by the high correlation between transcript and protein abundance (Fig. 3b, inset) and loss of the pQTL when transcript abundance was added as a covariate in the regression model (Supplementary Table 8). Founder strain allelic contributions derived from the pQTL mapping model suggested that four founder strain alleles (129S1, CAST, PWK and WSB) shared the genetic variant and exhibited higher protein expression levels. To validate these findings, a comparison of these expression coefficients to founder strain data showed the same expression profiles (Fig. 3c). Using genome sequences of the founder strains[26], we identified a candidate causal genetic variant—a 1-kb deletion in intron 1 of the gene. The same variant was previously reported as a pQTL in the DBA mouse strain[17]. DHTKD1 was just one of almost 1,600 cases in which QTL-to-transcript-to-protein regulation was identified as the best local model. Additional examples include *Ces2h* and *Pipox* (Extended Data Fig. 6).

A total of 623 proteins had local pQTL that affected protein abundance directly, including OMA1 (Fig. 3d–f). These proteins were uncoupled from their transcript, as evidenced by the lack of correlation between protein and transcript abundance (Fig. 3e, inset). For *Oma1*, founder allele contributions in the DO population pointed to an allele from the CAST strain causing reduced protein levels. This was validated by protein expression in the founder strains (Fig. 3f). Genome analysis identified four missense mutations in *Oma1* (H73N, R97Q, I127K and V283L), suggesting that protein structure may be affected and not the transcript. Other examples of variants that influenced protein expression that were not mediated through transcripts include *Entpd5* and *Lars2* (Extended Data Fig. 6).

Causal intermediates of distant pQTL

Unlike local pQTL, in which the causative variant is directly linked to the target protein-coding gene, distant pQTL exert their effects on target proteins in *trans* through a causal intermediate. To determine whether a distant pQTL acts proximally through the transcript of the affected protein or directly on the protein bypassing the transcript, we used mediation analysis (see Methods). We examined 1,130 distant pQTL and identified at least one candidate protein or transcript mediator for 743 (Supplementary Table 8). In total, we found 618 unique protein/transcript mediators, of which 534 regulated a single protein, 61 regulated two proteins, and 23 regulated three or more proteins. Furthermore, 84% of the top candidate protein mediators were themselves driven by a local pQTL. This illustrates that a single local QTL, acting proximally on a transcript or protein intermediate, can effectively control the abundance of a distant protein or multiple distant proteins, uncoupling them from their transcriptional control mechanisms.

We highlight examples in which mediation analysis identified the regulatory protein or transcript underlying the distant pQTL. TMEM68 protein exemplified a post-transcriptional model of regulation (Fig. 4a). TMEM68 has a distant pQTL peak on chromosome 13, and the *Tmem68* transcript has a local QTL on chromosome 4 (Fig. 4b, Supplementary Table 4). The protein and transcript levels were uncorrelated (Fig. 4d, left). We identified both NNT protein and *Nnt* transcript on chromosome 13 as candidate mediators of the distant pQTL for TMEM68 (Fig. 4c). The *Nnt* protein and transcript shared a local QTL indicating a transcriptional mechanism. Both *Nnt* protein and transcript were highly correlated with TMEM68 abundance (Fig. 4d). Founder allele expression patterns inferred at the distant pQTL suggest that a variant in B6 mice causes a downregulation in NNT protein levels, which was validated by proteomic analysis of the founder strains (Fig. 4e). This effect on *Nnt* expression has been previously attributed to a small exonic deletion found only in the B6 strain[28–30]. Using this same approach, we reconfirmed numerous known protein–protein associations including SNX7–SNX4, PGAM1–PGAM2 and LRRFIP1–FLII (refs 31–33), and inferred many new associations (Extended Data Fig. 7).

The chaperonin containing TCP1 (CC T) complex illustrates how mediation analysis can reveal larger co-regulated complexes and pathways (Fig. 4f). All eight subunits of the CC T complex shared a distant pQTL (but not distant eQTL) on chromosome 5 with the same pattern of allele effects. We identified the transcript and protein abundance of *Cct6a* as mediators of this post-transcriptional distant effect (Fig. 4g, h). This relationship is evident by the high correlation in protein–protein and protein–transcript abundance between *Cct6a* and other complex members (Fig. 4i, Extended Data Fig. 8). Founder strain allele effects inferred at the distant pQTL showed that DO animals containing the NOD strain allele on chromosome 5 expressed lower overall levels of the entire complex. This same pattern was observed in the founder strains (Fig. 4j). Genome sequence analysis identified a variant (rs228180583) in a conserved KLF4-binding domain in the *Cct6a* promoter region that was present only in the NOD strain. From these data, we propose that the variant lowers *Cct6a* transcript and protein abundance, which results in a stoichiometric imbalance and degradation of excess unbound complex members. These examples highlight the power of mediation analysis to identify protein–protein associations and co-regulated groups of proteins.

Genetic perturbations reveal protein networks

By leveraging the large number of distant pQTL and mediation analysis of each, we created a network of pQTL-regulated proteins (Extended Data Fig. 9a). Each distant pQTL is causally linked to its target protein with mediators and other co-regulated proteins to form a network. When merged across all 1,130 distant pQTL, the network comprises 5,794 causal or co-regulatory relationships among 3,938 proteins or QTL. Markov cluster algorithm (MCL) clustering defined 671 clusters of variable sizes (Extended Data Fig. 9b). Approximately 44% of clusters included members with shared biological functions as assessed by Gene Ontology (GO) enrichment (Extended Data Fig. 9c). As an example, almost all cholesterol synthesis enzymes were determined to be co-regulated and associated with just two distant pQTL that affected the protein expression for *Lss* and *Cyp51* (Supplementary Table 8). Clusters found within the larger regulatory network tended to associate proteins with shared biological properties. Some clusters grouped proteins according to subcellular localization, as seen for complex I of the electron transport chain (Extended Data Fig. 9d), SUCLG1/SUCLG2 and associated mitochondrial proteins (Extended Data Fig. 9e), and IMMT/SAMM50 with other mitochondrial proteins (Extended Data Fig. 9f). Each corresponds to a well-studied complex, suggesting that the regulatory network emerging from mediation analysis provides an accurate snapshot of mouse liver gene regulation.

To probe further the correspondence between protein co-regulation and physical association, each pQTL and its co-regulated proteins were mapped onto an ongoing and recently published human interactome network[34]. Physical associations accounted for a significant subset of protein regulatory networks, especially among distant QTL (Extended Data Fig. 9g–l). Through these findings, we propose that a considerable fraction of distant pQTL were the direct result of post-transcriptional regulation of proteins that had similar biological functions, cell locations, and/or complex membership.

Genotype is a predictor of protein abundance

For many genes with pQTL, founder strain allele patterns inferred from the DO pQTL mapping model closely matched protein abundance measured in the founder strains themselves. To determine the extent to which genotype can be a predictor of protein abundance, we examined all significant pQTL and compared the founder strain coefficients observed at the pQTL location to the protein levels measured in the founder strains (Fig. 5a). We found that predictive power increased with the significance of the pQTL (Fig. 5a, Extended Data Fig. 10). Because of their tight linkage to the controlled gene, local pQTL tended to have higher predictive power than distant loci (local pQTL median r = 0.72; distant pQTL median $r = 0.11$). However, highly significant distant pQTL (>10 LOD) have comparable predictive power to local pQTL of similar significance.

We further validated our strains predictions using the quantitation of ~6,500 proteins from four CC strains (Supplementary Table 9). For each pQTL, we identified the genotype in the CC strains and predicted the protein abundance using the DO proteomics data. Our data suggest that strain genotype is also predictive of protein abundance in the CC strains (Fig. 5b). The predictive power was higher for local pQTL than distant ones. As an example, LYPLAL1 was identified with a local pQTL in the DO population and was predicted to have lower protein abundance in the CC 001 and CC 003 strains (Fig. 5c). For distant pQTL with high LOD scores, the predictive power was also high. For distant pQTL, these predictions were made by comparing the measured protein and the genotype at the QTL location. For example, GLYCTK protein abundance was predicted using the genotype at the *Nags* gene location where the variant was detected (Fig. 5d).

This study quantified both protein and transcript abundance in a genetically diverse population of mice, mapping their genetic architecture. We identified the largest catalogue of pQTL so far, which can be attributed to two variables in our experimental design. First, we have improved the accuracy and sensitivity of quantification for both protein and transcript abundance. Second, our experimental population captured genetic diversity far in excess of the human population and standard laboratory mouse strains. Earlier studies reported a disconnect between transcript and protein abundance[2,3,6], which has also been a conclusion drawn from several recent eQTL–pQTL analyses[4,7,17,35]. Data here show that local QTL tend to abide by the central dogma as demonstrated by concordant effects on transcripts and proteins, whereas distant pQTL are conferred by post-transcriptional mechanisms. Our mediation analysis provided the ability to identify causal protein intermediates underlying distant pQTL and led to the identification of hundreds of protein–protein associations. Our experimental design provides an advantage over protein interaction maps because genetic mapping is not dependent on physical interactions. This conclusion is further exemplified by the co-regulation of protein complexes or biochemical pathways in this study. Stoichiometric buffering provides one explanation for co-regulation of protein complexes and may account for earlier observations that protein abundances (but not transcript abundances) of orthologues are well-conserved across large evolutionary distances[36,37].

These findings suggest a new predictive genomics framework in which quantitative proteomics and transcriptomics are combined in the analysis of a discovery population like the DO to identify genetic interactions. Next, pathways relevant to the tissue/physiological phenotype of interest are intersected with the list of significant pQTL. Pathways enriched for proteins with significant pQTL should be amenable to manipulation in the founder and CC strains. That is, the founder allele effects inferred at the pQTL can be combined in such a way via crosses of CC strains to tune pathway output. Moreover, as we better understand the types of mutation that can affect protein abundance, we can introduce specific mutations with gene editing into sensitized or robust genetic backgrounds. We foresee this strategy being used to design reproducible rodent models that span a range of human-relevant phenotypes, for example, in drug metabolism or toxicology studies.

Methods

The sample size (192 animals) was calculated based on previous experimental RNA-seq data and was determined to be sufficient to detect genetics effects that explain 10% or more genetic variation with 90% power and 10^{-6} type I error rate. Randomization was used to assign mice to treatments and samples to batches, bar codes, and TMT tags in both the RNA-seq and proteomics experiments. Data collection was carried out by automation, and as such there was no need for blinding the sample identifiers.

Animals and genotyping: DO mice

Diversity Outbred mice (DO, stock no. 009376) were obtained from The Jackson Laboratory (JAX) at 3 weeks of age, housed at JAX, and fed either standard rodent chow (6% fat by weight, LabDiet 5K52; LabDiet, Scott Distributing) or a high-fat diet (44.6% kcal fat and 34% kcal sucrose by weight, TD.08811, Harlan Laboratories) from wean age throughout the study. In total, 192 DO mice were analysed in the current study, including 50 females and 48 males raised on standard chow, and 48 females and 46 males raised on the high fat diet. At 26 weeks of age, animals were euthanized, dissected, and liver samples were sent for RNA-seq analysis at JAX (samples stored in RNAlater solution; Life Technologies) and proteomics analysis at Harvard Medical School (HMS; samples sent as snap frozen tissue).

Animals and genotyping: founder strain and CC mice

Two male and two female mice from each of the eight DO/CC founder inbred strains and four (3 males and 3 females) CC recombinant inbred strains (CC strains CC 001, CC 003, CC 004, and CC 017) were obtained from and housed at JAX, raised on the standard chow diet. Founder strain mice were euthanized at 26 weeks of age, and the CC mice were euthanized at 8–16 weeks of age. Liver samples were dissected from each mouse, snap frozen and sent to HMS for proteomics analysis. All procedures on mice were approved by the Animal Care and Use Committee at JAX.

Multiplexed quantitative proteomic analysis of mouse livers: sample preparation and TMT labelling

A total of 192 DO mouse livers (~50 mg), 32 founder strains livers (8 founders strains, 2 male and 2 female replicates for each strain) and 24 CC strain livers (4 strains, 3 male and 3 female replicates for each strain) were homogenized in 1 ml lysis buffer (1% SDS, 50 mM Tris, pH 8.8 and Roche complete protease inhibitors). Samples were reduced with 5 mM dithiothreitol for 30 min at 37 °C followed by alkylation with 15 mM for 30 min at room temperature in the dark. The alkylation reaction was quenched by adding 5 mM dithiothreitol for 15 min at room temperature in the dark. A 500 μl aliquot was then methanol/chloroform precipitated. The samples were allowed to air dry before being resuspended in 1 ml of 8 M urea and 50 mM Tris, pH 8.8. The urea concentration was diluted down to ~1.5 M urea with 50 mM Tris. Proteins were quantified using a BCA assay. Protein was then digested using a combination of Lys-C/trypsin at an enzyme-to-protein ratio of 1:100. First, protein was digested overnight with Lys-C followed by 6-h digestion with trypsin all at 37 °C. Samples were then acidified using formic acid to approximately pH 3. Samples were then desalted using a SepPak column. Eluents were then dried using a vacuum centrifuge. Peptide pellets were resuspended in 110 μl of 200 mM HEPES buffer, pH 8, and peptides were quantified by a BCA assay. Approximately 70 μg of peptides (100 μl of sample + 30 μl of 100% acetonitrile) were then labelled with 15 μl of 20 μg μl^{-1} of the corresponding TMT 10-plex reagent (DO or founder strains) or TMT 8-plex reagent (CC strains) for 2 h at room temperature. The reaction was quenched using 8 μl of 5% hydroxylamine for 15 min. Peptides were then acidified using 150 μl of 1% formic acid, each set of 10 samples were mixed and desalted using a SepPak column. In total, 25 TMT 10-plex reactions and 3 8-plex reactions were performed (21 DO mice, 4 founder strains and 3 CC strains). The full labelling schemes for the DO mice, the founder strains and CC strains are provided as supplementary tables (Supplementary Tables 1, 3 and 7).

Basic reverse-phase fractionation

Each of the 28 TMT experiments was separated by basic, reversed-phase chromatography. Samples were loaded onto an Agilent 300 Extend C18 column (5 μm particles, 4.6 mm ID and 220 mm in length). Using an Agilent 1100 quaternary pump equipped with a degasser and a photodiode array detector (set at 220- and 280-nm wavelength), peptides were separated using a 50 min linear gradient from 18% to 40% acetonitrile in 10 mM ammonium bicarbonate, pH 8, at a flow rate of 0.8 ml $\min^{-1}$. Peptides were separated into a total of 96 fractions that were consolidated into 24. Samples were subsequently acidified with 1% formic acid and vacuum centrifuged to near dryness. Each fraction was desalted via StageTip, dried via vacuum centrifugation, and reconstituted in 1% formic acid for liquid chromatography tandem mass spectrometry (LC–MS/MS) processing.

Liquid chromatography electrospray ionization tandem mass spectrometry (LC–ESI-MS/MS)

Peptides from every odd fraction (12 fractions total) from basic reverse-phase fractionation were analysed using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) equipped with a Proxeon ultra high pressure liquid chromatography unit. Peptide mixtures were separated on a 100 μm ID microcapillary column packed first with ~0.5 cm of 5 μm Magic C18 resin followed by 40 cm of 1.8 μm GP-C18 resin. Peptides were separated using a 3-h gradient of 6–30% acetonitrile gradient in 0.125% formic acid with a flow rate of ~400 nl $\min^{-1}$. In each data collection cycle, one full MS scan (400–1,400 *m/z*) was acquired in the Orbitrap (1.2×10^5 resolution setting and an automatic gain control (AGC) setting of 2×10^5). The subsequent MS2–MS3 analysis was conducted with a top 10 setting or a top speed approach using a 2-s duration. The most abundant ions were selected for fragmentation by collision induced dissociation (CID). CID was performed with a collision energy of 35%, an AGC setting of 4 × 10^3, an isolation window of 0.5 Da, a maximum ion accumulation time of 150 ms and the rapid ion trap setting. Previously analysed precursor ions were dynamically excluded for 40 s.

During the MS3 analyses for TMT quantification, precursors were isolated using a 2.5-Da *m/z* window and fragmented by 35% CID in the ion trap. Multiple fragment ions (SPS ions) were co-selected and further fragmented by HCD. Precursor ion selection was based on the previous MS2 scan and the MS2–MS3 was conducting using sequential precursor selection (SPS) methodology. HCD used for the MS3 was performed using 55% collision energy and reporter ions were detected using the Orbitrap with a resolution setting of 60,000, an AGC setting of 50,000 and a maximum ion accumulation time of 150 ms.

Database searching and reporter ion quantification

Software tools were used to convert mass spectrometric data from raw file to the mzxml format[34]. Erroneous charge state and monoisotopic *m/z* values were corrected as per previous publication[34]. MS/MS spectra assignments were made with the Sequest algorithm[41] using an indexed Ensembl database (mouse: Mus_musculus NCBIM37.61). Databases were prepared with forward and reversed sequences concatenated according to the target-decoy strategy[42]. All searches were performed using a static modification for cysteine alkylation (57.0215 Da) and TMT on the peptide N termini and lysines. Methionine oxidation (15.9949 Da) was considered a dynamic modification. Mass spectra were searched with trypsin specificity using a precursor ion tolerance of 10 p.p.m. and a fragment ion tolerance of 0.8 Da. Sequest matches were filtered by linear discriminant analysis as described previously, first to a data set level error of 1% at the peptide level based on matches to reversed sequences[42]. Peptide probabilities were then multiplied to create protein rankings and the data set was again filtered to a final data set level error of 1% false discovery rate (FDR) at the protein level. The final peptide-level FDR fell well below 1% (~0.2% peptide level). Peptides were then assigned to protein matches using a reductionist model, where all peptides were explained using the least number of proteins.

Peptide quantitation using TMT reporter ions was accomplished as previously published[21,22]. In brief, a 0.003 Da *m/z* window centred on the theoretical *m/z* value of each reporter ion was monitored for each of the 8–10 reporter ions, and the intensity of the signal closest to the theoretical *m/z* value was recorded. TMT signals were also corrected for isotope impurities based on the manufacturer’s instructions. Peptides were only considered quantifiable if the total signal-to-noise for all channels was >200 and an isolation specificity of >0.75. Within each TMT experiment, peptide quantitation was normalized by summing the values across each channel and then each channel was corrected so that each channel had the same summed value. Protein quantitation was performed by summing the signal-to-noise for all peptides for a given protein. Protein quantitative measurements were then scaled to 100 (equal expression across all channels would be a value of 10). Normalization across each of the 10plex experiments was then performed using quantile normalization.

Statistical analyses

Principal component analysis was performed using Cluster 3.0 (ref. 43). Hierarchical clustering, *K*-means clustering and ANOVA were performed using Multi experiment Viewer. Analysis on the founder strains proteomics data sets was performed using an ANOVA and adjusted for multiple testing using the Benjamini–Hochberg FDR procedure.

Implications of multiplexed quantitative proteomics platform

Improvements in several aspects of the analysis pipeline enabled the increase in scale. Our quantitative proteomics technology proved instrumental as it supported multiplexing with ten different mouse livers in the same analysis. Accurate expression measurements were obtained by applying a notched isolation waveform on an Orbitrap Fusion instrument. The time required to collect expression profiles from each 10-plex was 36 h or ~4 h per mouse liver of mass spectrometry analysis time. The proteome-wide analysis of 192 livers thus required 35 days. As a result of these methodology improvements, we detected tenfold more pQTL than previous MS-based reports.

Genotyping of DO and CC samples: DO samples

Genomic DNA was extracted from each DO mouse ($n = 192$ total samples) and genotyped at 57,973 single nucleotide polymorphisms (SNPs) on the Mega-MUGA platform (Geneseek)[44]. A total of 177 out of 192 samples passed SNP quality control metrics. For these samples, founder haplotypes were inferred from SNP probe intensities using a hidden Markov model implemented in the DOQTL R package[27,45], and then used to interpolate a grid of 64,000 evenly-spaced genetic intervals. In addition, founder haplotypes were independently inferred from the RNA-seq data by genotyping by RNA-seq (GBRS) protocol (see next section) and interpolated to the same 64,000 interval grid.

For each sample, we verified that the haplotype reconstructions agreed between the DNA Mega-MUGA and GBRS reconstructions by calculating the Pearson correlation between each pair of samples. When a Mega-MUGA sample had a correlation below 0.4 with the same sample ID in the RNA-seq data, we assumed that this sample was mismatched. We searched the RNA-seq data for the correct match to the Mega-MUGA sample by looking for another sample that was more highly correlated. If we found an RNA-seq sample with a correlation >0.4 that was not assigned to another sample, we matched it with the Mega-MUGA sample. When a sample was removed from the Mega-MUGA data for technical reasons, we used the GBRS haplotype reconstructions (samples F326, F328, F362, F363, F368, M377, M388, M392, M393, M394, M404, M408, M411, M419 and M425).

Genotyping of DO and CC samples: CC samples

Founder haplotypes for the CC strains were downloaded from the CC strain database (csbio.unc.edu/CCstatus/gstemp/AllImageHapAndGenotypeFiles.zip) maintained at the University of North Carolina.

Transcriptome profiling and GBRS

Total liver RNA was isolated from each of the 192 DO mice and sequenced by single-end RNA-seq as previously described[46]. We aligned raw reads against pooled transcriptomes of the eight founder strains. To construct the pooled transcriptome, we incorporated founder strain-specific SNPs and insertions/deletions (Sanger REL-1410) into the reference strain genome sequence (GRCm38/mm10) to produce strain-specific genomes. We derived transcript sequences for all annotated genes (Ensembl version 75 gene annotation) from each strain genome, and then combined the eight founder allele sequences for each transcript into one pooled transcriptome for read alignment. After alignment, we quantified expected read counts expressed from each transcript allele using an expectation maximization algorithm (EMASE, https://github.com/churchill-lab/emase). We repeated the same process for liver RNA-seq data from the eight founder strains to assess how specifically each founder read aligns back to their origin strain when exposed to all other founder alleles simultaneously in the alignment pool. We then evaluated the genotype probability of each transcript using a hidden Markov Model (HMM), where we bring those read counts together and calculate (1) how likely allele-specific read counts are generated from a specific genotype, and (2) how much those likelihoods comply within the context of neighbouring transcripts. Finally, we re-quantified total and allele-specific expression with EMASE by repeating the similar process but using individualized diploid transcriptomes reconstructed along our genotype calls.

QTL mapping of transcript and protein abundance

Quantitative proteomics combined with transcript quantitation by RNA-seq makes it possible to define the relative contributions of transcriptional versus post-transcriptional mechanisms and local versus distant effects on protein abundance. For example, a local QTL is a genetic variant near the target gene that influences its expression; it might be expected to act in *cis* and affect both transcript and protein levels. By contrast, distant QTL exert their effect on a target gene's expression in *trans*, most likely via a causal intermediate such as another protein or RNA species. Identifying causal intermediates of distant QTL effects may reveal novel protein–protein associations and their biological consequences. Our comprehensive pQTL analysis yielded a global network of interactions that shed new light on the regulation of protein abundance.

QTL mapping

For mapping of pQTL and eQTL, we included only proteins that were present (non-0) in ≥96 samples and corresponded to gene identifiers in the RNA-seq data that were also expressed in ≥96 samples. A total of 6,707 proteins met these criteria. For pQTL mapping with the proteomics data, protein abundance values were first quantile-normalized and transformed to rank normal scores, and then pQTL were mapped with the R package DOQTL[27], using a linear mixed model with sex, diet and TMT tag as additive covariates and a random polygenic term to account for genetic relatedness among the DO animals[47]. For eQTL mapping from the RNA-seq data, gene-level counts were first normalized to the upper quartile value and transformed to rank normal scores, and then eQTL were mapped with DOQTL including sex, diet and batch as additive covariates and a random polygenic term to account for relatedness. We used the 64 k genotype matrix derived from Mega-MUGA DNA genotypes as input for pQTL and eQTL mapping, with the exception of samples with missing or low quality DNA genotype results where we used GBRS-derived genotypes.

Statistical analyses

Significance thresholds were established by performing 10,000 permutations and fitting an extreme value distribution to the maximum LOD scores[48]. Permutation derived *P* values were then converted to *q*-values with the QVALUE R package, using the bootstrap method to estimate π_0 and the default λ tuning parameters[49]. The significance threshold for declaring a QTL was set at a genome-wide significance level of $P < 0.1$ (FDR = 10%).

eQTL replication analysis

To detect a pQTL and eQTL requires a strong statistical signal to exceed stringent genome-wide significance thresholds. We considered the possibility that lack of concordance between distant pQTL and eQTL could be explained by low power, especially for the distant pQTL. The proteomics data in this study were obtained on a subset (discovery set) of DO mice from an earlier study[46]. We created a replication set for the eQTL by random sampling of 192 additional DO samples. As expected, the likelihood of replicating an eQTL depended on the significance of the QTL in the discovery set (Supplementary Fig. 4a). Local eQTL tend to be more significant and replicated well across experiments (76% replication, n = 2,448), while distant eQTL replicated poorly (5% replication, n = 52; Supplementary Fig. 4a). The distribution of LOD scores is similar for distant pQTL and distant eQTL (Supplementary Fig. 4b), suggesting that we had similar low power to detect distant pQTL as distant eQTL. While the overlap between distant pQTL and eQTL is lower than what we had expected (<1%, n = 9; Supplementary Fig. 4c), it is still difficult to rule out low rate of detection as a possible explanation. We provide additional evidence that distant pQTL act through post-transcriptional mechanisms.

Model selection by BIC

For each of the 6,707 proteins in the discovery set with detectable transcript and protein abundance, we identified (1) the locus within ±10 Mb of the gene midpoint with the highest LOD score (local), and (2) the locus on a separate chromosome with the highest LOD score (distant), regardless of their statistical significance. Next, for each local and distant locus, we considered all possible relationships among locus genotype, transcript abundance, and protein abundance. We computed the BIC score for each of eight possible models. For each protein, we recorded the optimal local and distant locus model (that is, model that yields the lowest BIC score). In addition, we calculated the Bayesian posterior probability (assuming a uniform prior over relationships), and from these posterior probabilities estimated the expected number of proteins for each model.

Mediation analysis to identify distant regulators and co-regulated proteins

For proteins with distant pQTL, mediation analysis was used to identify proteins and transcripts in that region that were likely to be the causal mediator of the QTL. Mediation analysis in this context is adapted from the general approach outlined previously[50] to differentiate moderator from mediator variables in social psychology research[51]. We implemented our method as the function ‘intermediate’ for the open statistical language R. In brief, for a given distant pQTL, we first identified all expressed proteins and transcripts within 10 Mb of the peak SNP—these genes are candidate mediators of the distant pQTL. We then included the protein abundance of each candidate individually as an additive covariate in the pQTL mapping model and re-ran the regression at the peak distant SNP. We performed the same analysis with transcript abundance as the additive covariate. Our expectation was that many distant pQTL would be mediated by the protein and/or transcript abundance of a gene in that locus. For distant pQTL where this is true, including the abundance of the mediator protein/transcript in the pQTL mapping model should significantly decrease or abolish the distant pQTL effect—as evidenced by a decrease in LOD score. We calculate LOD scores using the ‘double-lod-diff ’ method in r/intermediate to minimize the effects of missing data in the proteomics and RNA-seq data sets.

Statistical analysis

To assess the significance of the LOD drop for a given candidate mediator on a given distant pQTL, a null distribution of LOD scores was estimated by re-running the regression at the peak SNP and including all expressed proteins and transcripts outside of the candidate regions as additive covariates. In total, this yields mediation LOD scores for 8,050 proteins and 21,454 transcripts for each distant pQTL. Mediation LOD scores are then scaled to z-scores, and any candidate with a conservative *z*-score ≤ −6 is recorded as a potential causal mediator. Further, any protein/transcript outside of the pQTL window with a *z*-score ≤ −6 is recorded as a potential co-regulated partner of the target protein. We examined 1,130 distant pQTL and identified at least one candidate protein or transcript mediator for 743. In total, we found 618 unique protein/transcript mediators, of which 534 regulated a single protein, 61 regulated two proteins, and 23 regulated three or more proteins. Furthermore, 84% of the top candidate protein mediators were themselves driven by a local pQTL.

Analysis of distant pQTL for transcriptional modes of regulation

We observed that a small subset of local pQTL and nearly all distant pQTL lacked corresponding eQTL. For these proteins, transcript and protein abundance appeared to be largely uncoupled (buffered). For the minority of local pQTL lacking corresponding local eQTL, we expected that mutations altering protein stabilization but not affecting transcript abundance conferred this effect. The paucity of distant pQTL with corresponding eQTL is especially puzzling given our initial expectation that *trans* effects on protein abundance would likely stem from transcription factors or chromatin modifying proteins. We detected few transcription factors and fewer transcription factor pQTL in our protein data set ($n = 132$ expressed out of 2,243 annotated transcription factors; $n = 21$ out of 132 transcription factors with pQTL; $n = 9$ local transcription factor pQTL, $n = 12$ distant transcription factor pQTL), suggesting (as others have noted[52]) that their regulation is more evolutionarily constrained and less tolerant of genetic variation, or alternatively, that the effects of any individual polymorphism in a transcription factor may be buffered by other transcriptional components. Results from recent large population genetics data sets[53] support the former explanation, and consequently distant effects from transcription factors may resist detection by genetic mapping methods and account for the lack of distant pQTL that affect both transcript and protein abundance

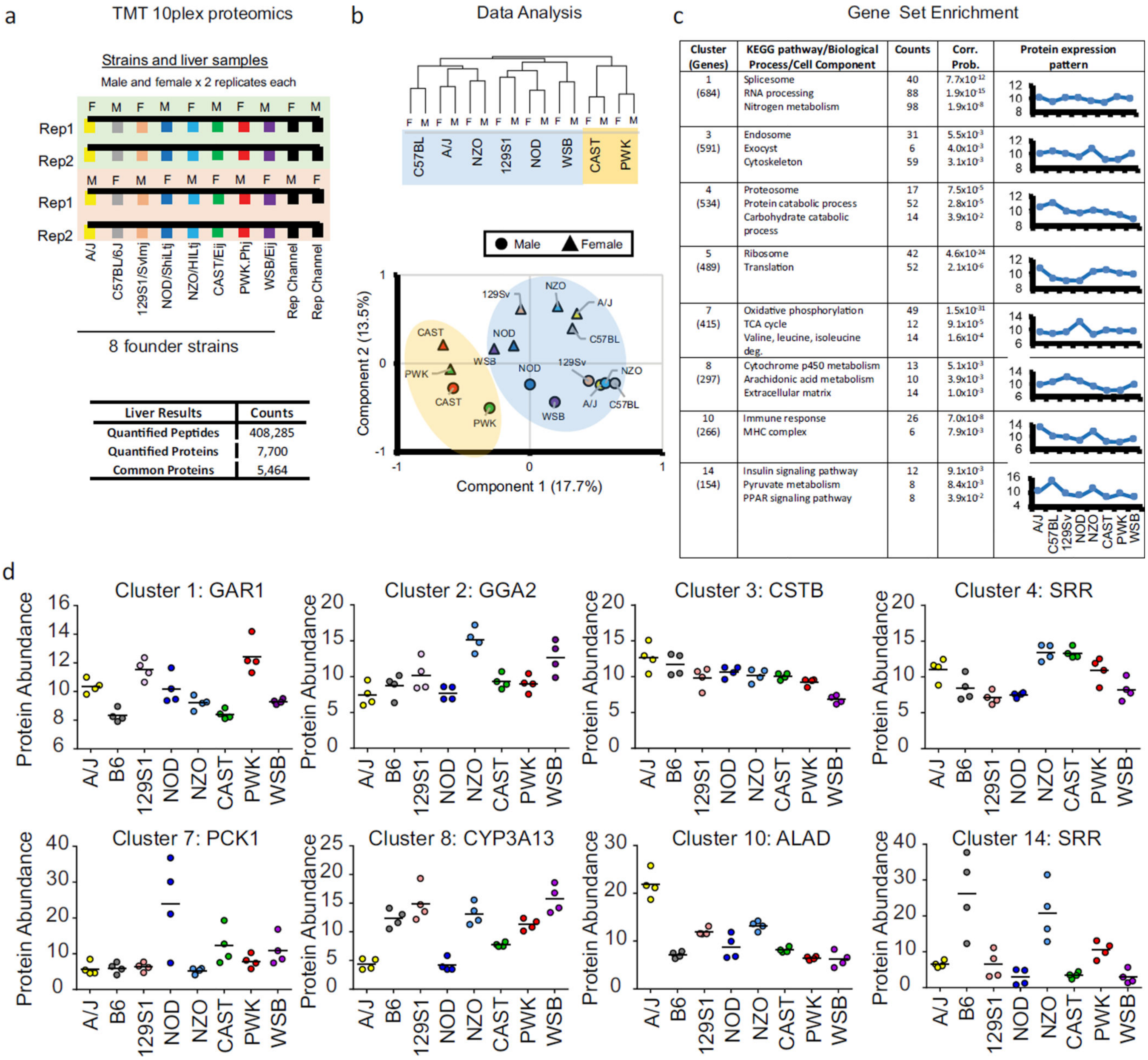
Assembly and clustering of the distant pQTL regulatory network

We assembled the distant pQTL regulatory network by drawing directed edges to connect each *trans*-pQTL with its primary target protein. Each target protein was then connected to co-regulated proteins via directed edges. For purposes of graph assembly, each distant pQTL was represented by the protein most likely to be responsible for the effects of the QTL as indicated by mediation analysis. To identify clusters of co-regulated proteins, the directed network was converted to undirected form and subjected to MCL clustering[54] using an inflation parameter of 1.5. Each cluster was then evaluated for enrichment of PFAM domains[55], subcellular localizations[56], or GO categories[57] using a hypergeometric test with subsequent multiple testing correction[58]. $P < 0.05$ after multiple testing correction was considered indicative of enrichment.

Mapping distant pQTL and co-regulated proteins onto the BioPlex protein interaction network

To quantify the extent to which direct physical interactions could explain distant pQTL regulation, each distant pQTL and its regulated proteins were associated with their human homologues using official gene symbols and mapped to the BioPlex network of human protein interactions[31]. Any protein that could not be mapped to the BioPlex network, either because a human homologue was not known or because the protein did not occur in the network, was excluded. Physical interactions connecting the pQTL and its co-regulated proteins were counted and compared against the maximum number of pairwise connections to calculate the density of physical interactions. A binomial model was used to identify sets with unusually high numbers of interactions assuming the probability of an interaction occurring between two randomly selected proteins in the BioPlex network was 9.42×10^{-4} (the BioPlex graph density). P values were adjusted for multiple hypothesis testing using the method of Benjamini–Hochberg[58] and those smaller than 0.05 after correction were taken to be significant.

Extended Data

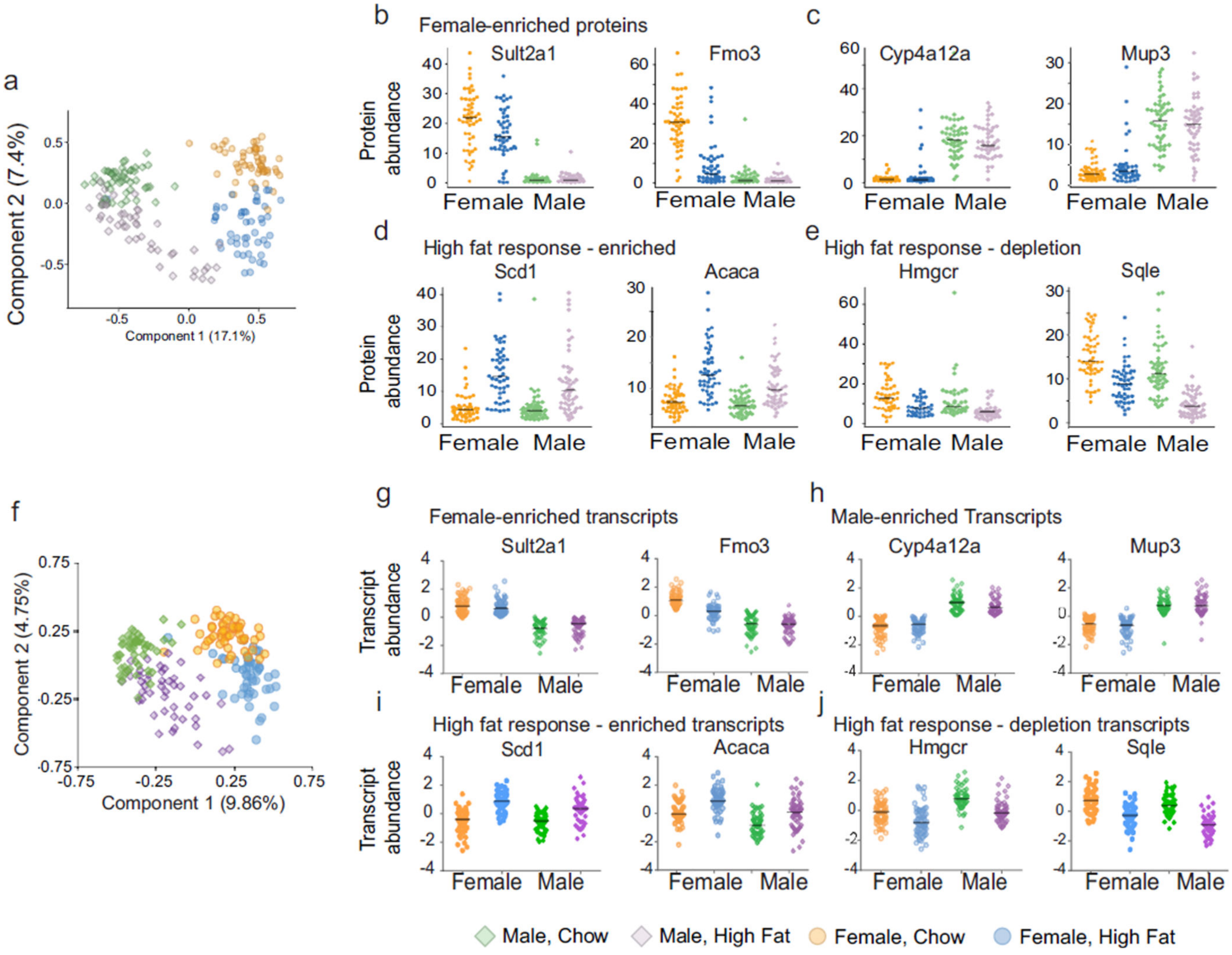


Liver Results	Counts
Quantified Peptides	408,285
Quantified Proteins	7,700
Common Proteins	5,464

Cluster (Genes)	KEGG pathway/Biological Process/Cell Component	Counts	Corr. Prob.	Protein expression pattern
1 (684)	Splicesome RNA processing Nitrogen metabolism	40 88 98	$7.7x10^{-12}$ $1.9x10^{-15}$ $1.9x10^{-8}$	12 10 8
3 (591)	Endosome Exocyst Cytoskeleton	31 6 59	$5.5x10^{-3}$ $4.0x10^{-3}$ $3.1x10^{-3}$	12 10 8
4 (534)	Proteosome Protein catabolic process Carbohydrate catabolic process	17 52 14	$7.5x10^{-5}$ $2.8x10^{-5}$ $3.9x10^{-2}$	12 10 8
5 (489)	Ribosome Translation	42 52	$4.6x10^{-24}$ $2.1x10^{-6}$	12 10 8
7 (415)	Oxidative phosphorylation TCA cycle Valine, leucine, isoleucine deg.	49 12 14	$1.5x10^{-31}$ $9.1x10^{-5}$ $1.6x10^{-4}$	14 10 6
8 (297)	Cytochrome p450 metabolism Arachidonic acid metabolism Extracellular matrix	13 10 14	$5.1x10^{-3}$ $3.9x10^{-3}$ $1.0x10^{-3}$	14 10 6
10 (266)	Immune response MHC complex	26 6	$7.0x10^{-8}$ $7.9x10^{-3}$	14 10 6
14 (154)	Insulin signaling pathway Pyruvate metabolism PPAR signaling pathway	12 8 8	$9.1x10^{-3}$ $8.4x10^{-3}$ $3.9x10^{-2}$	16 10 4

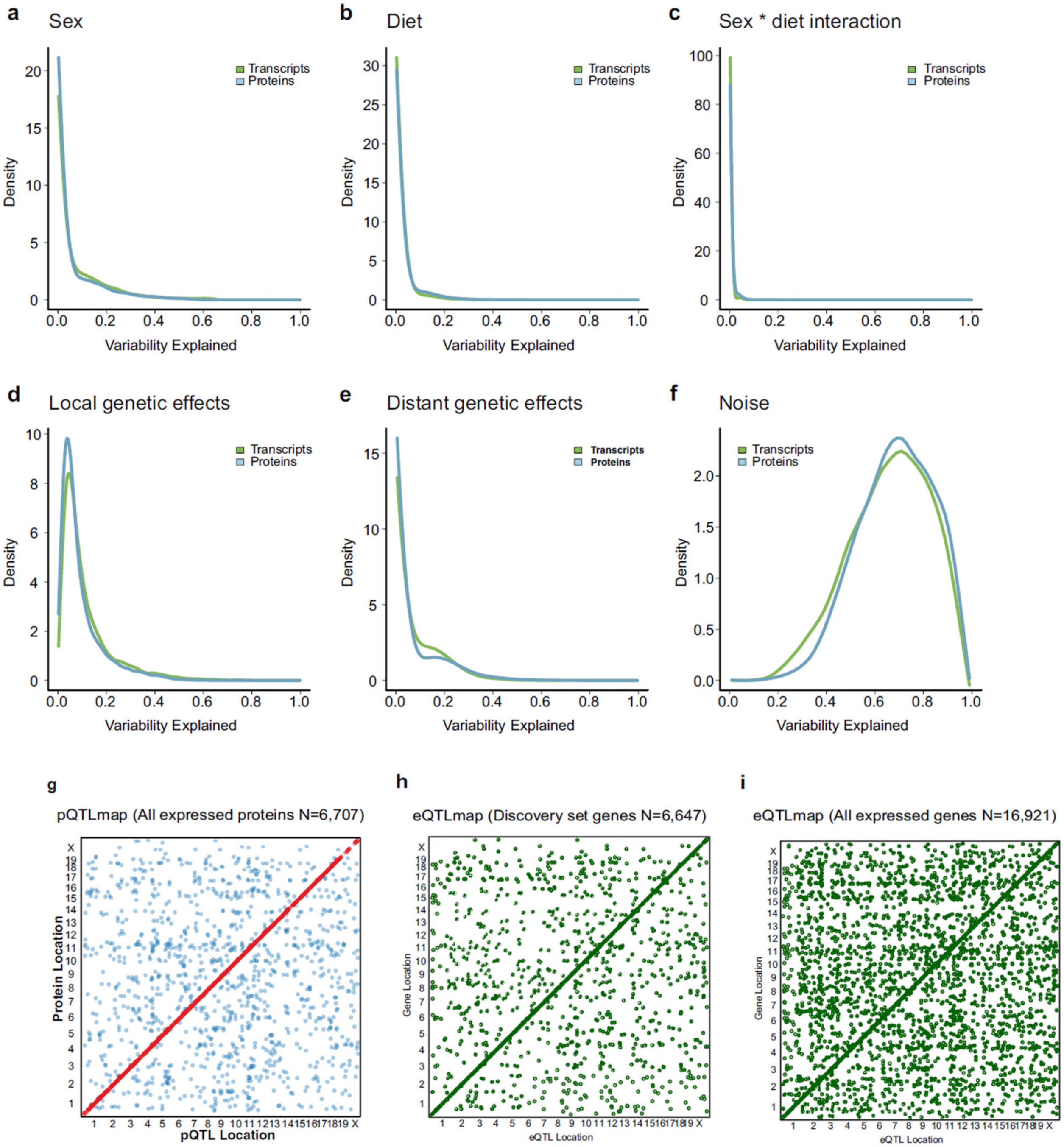
Extended Data Figure 1. Proteomic profiling of the eight founder strains used to create the DO mouse population

a, A multiplexed TMT proteomics method was used to characterize protein expression for the eight founder strains with two biological replicates for each strain using both sexes. In total, just over 400,000 peptides were quantified corresponding to 7,699 proteins. **b**, Hierarchical clustering and principal component analysis determined that the major source of variation in protein expression is due to genetic variation among the eight strains and the sex within strains. **c**, *K*-means clustering and gene set enrichment determined that each of the clusters was specifically enriched for metabolic pathways, biological process or cellular components. **d**, Proteins representing each of the displayed clusters from **c**. These proteins have specific patterns of expression as exemplified by PCK1, which was highly expressed in the NOD strain. Other examples include SCD1, which was highly expressed in C57BL/6J and NZO strains ($n = 4$ mice for each founder, 2 male and 2 female, black bars represent median values). Protein abundance is shown as the percentage contribution of that mouse's protein levels to its respective 10-plex.



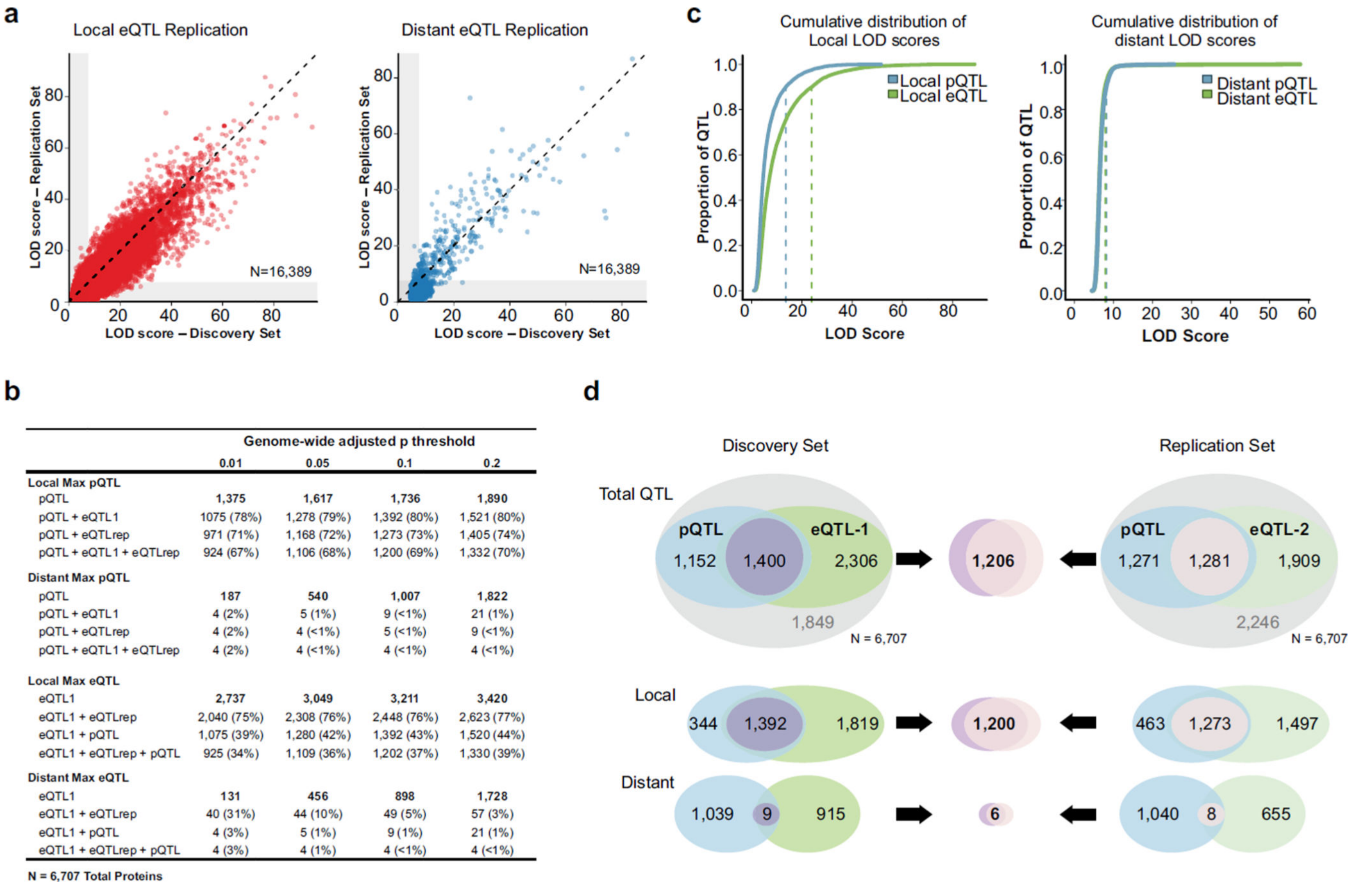
Extended Data Figure 2. The influence of sex and diet on protein and transcript abundance

a, Principal component analysis aligns well with sex and diet as major experimental contributors of variation in protein abundance. **b**, Female-specific protein abundance profiles for SULT2A1 and FMO3. **c**, Male-specific protein abundance profiles for CYP4A12A and MUP3. **d, e**, Diet also resulted in the regulation of many proteins, which are represented by proteins such as SCD1 and ACACA that increased in abundance and proteins such as HMGCR and SQLE that decreased in abundance. **f**, Principal component analysis aligns well with sex and diet as major experimental contributors of variation in transcript abundance. **g–j**. Transcript scatter plots for the proteins in **b–e**. Transcript abundance data were transformed to rank normal scores for plotting.



Extended Data Figure 3. Genetic effects drive much of the observed expression variance in the RNA-seq and proteomics data

Liver transcript and protein abundance are highly variable in the DO population. Among the discovery set (*n* = 6,707 proteins, 6,647 genes), much of this variance can be attributed to one or more experimental variables and/or genetic effects. **a–c**, The experimental covariates sex and diet influence many transcripts and proteins in an additive manner, however, the interaction of sex and diet does not seem to affect many genes. The effects from sex and diet are not biased towards one molecular species—that is, similar numbers of transcripts and proteins are similarly affected by these experimental variables. Genetic variation underlies many of the most variable transcripts and proteins. **d, e**, Local genetic variation in particular is a strong driver of expression variation for many genes, while distant genetic effects are observed but more subtle. Among the discovery set, we observe more and larger genetic effects (both local and distant) on transcript abundance than protein abundance. **f**, For most transcripts and proteins detected in this study, expression variation is minimal, cannot be attributed to a known experimental or genetic variable, and is plotted as noise. **g**, pQTL map for all 6,707 proteins tested from genetic linkage analysis. **h, i**, QTL mapping identified the genetic loci that underlie variability in transcript abundance (eQTL). For the discovery set of transcripts with detected proteins and the larger set of all expressed genes, the location of the eQTL is plotted on the *x* axis and the location of the controlled gene is plotted on the *y* axis. Most genetic effects are local and map to the same location as the gene, as evidenced by the prominent diagonal line in both maps.

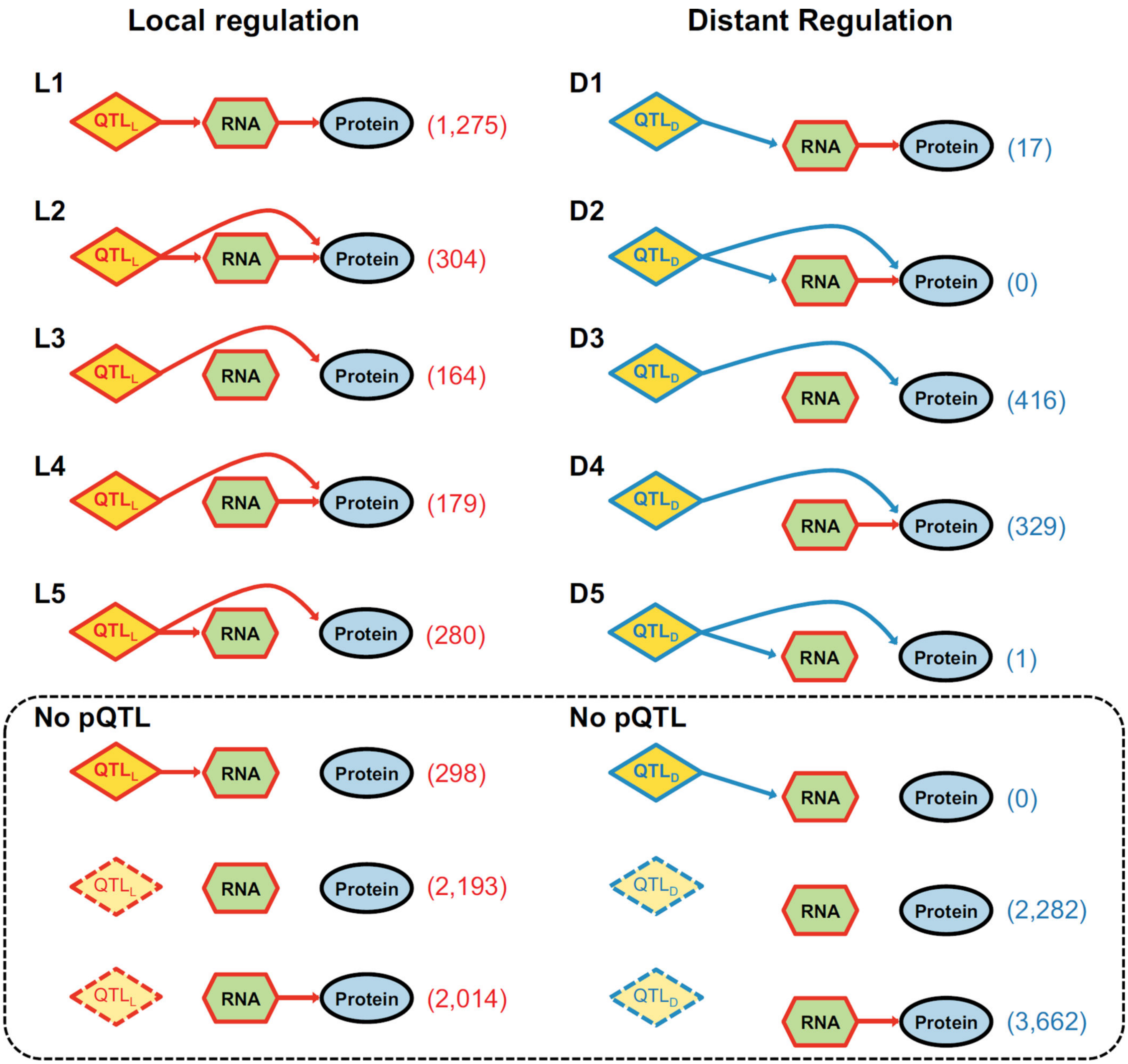


	Genome-wide adjusted p threshold			
	0.01	0.05	0.1	0.2
Local Max pQTL				
pQTL	**1,375**	**1,617**	**1,736**	**1,890**
pQTL + eQTL1	1075 (78%)	1,278 (79%)	1,392 (80%)	1,521 (80%)
pQTL + eQTLrep	971 (71%)	1,168 (72%)	1,273 (73%)	1,405 (74%)
pQTL + eQTL1 + eQTLrep	924 (67%)	1,106 (68%)	1,200 (69%)	1,332 (70%)
Distant Max pQTL				
pQTL	**187**	**540**	**1,007**	**1,822**
pQTL + eQTL1	4 (2%)	5 (1%)	9 (<1%)	21 (1%)
pQTL + eQTLrep	4 (2%)	4 (<1%)	5 (<1%)	9 (<1%)
pQTL + eQTL1 + eQTLrep	4 (2%)	4 (<1%)	4 (<1%)	4 (<1%)
Local Max eQTL				
eQTL1	**2,737**	**3,049**	**3,211**	**3,420**
eQTL1 + eQTLrep	2,040 (75%)	2,308 (76%)	2,448 (76%)	2,623 (77%)
eQTL1 + pQTL	1,075 (39%)	1,280 (42%)	1,392 (43%)	1,520 (44%)
eQTL1 + eQTLrep + pQTL	925 (34%)	1,109 (36%)	1,202 (37%)	1,330 (39%)
Distant Max eQTL				
eQTL1	**131**	**456**	**898**	**1,728**
eQTL1 + eQTLrep	40 (31%)	44 (10%)	49 (5%)	57 (3%)
eQTL1 + pQTL	4 (3%)	5 (1%)	9 (1%)	21 (1%)
eQTL1 + eQTLrep + pQTL	4 (3%)	4 (1%)	4 (<1%)	4 (<1%)

N = 6,707 Total Proteins

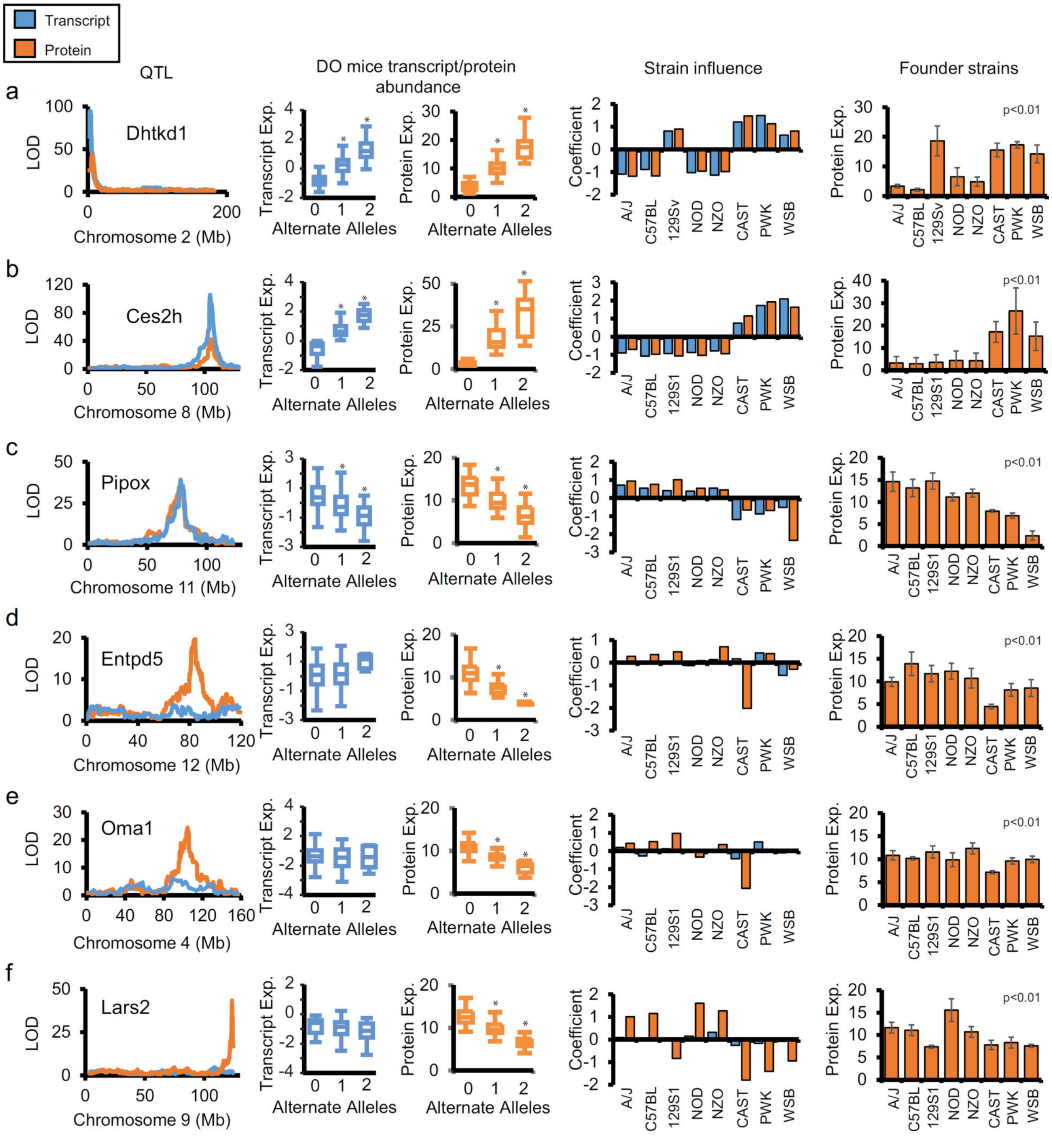
Extended Data Figure 4. Replication rates for eQTL are highly correlated with effect size, and local eQTL replicate at higher rates than distant eQTL

a, To assess replication of eQTL, an independent set of 192 DO liver RNA-seq samples was analysed ('replication set') and compared to the discovery set. A total of 16,839 genes were expressed in half or more samples in both data sets. For each gene, the most significant proximal locus (within ± 10 Mb of gene) and distant locus (located on a different chromosome from the gene) were identified from the discovery set—LOD scores at these loci are plotted on the *x* axis (local in red; distant in blue). Next, the most significant loci within a 10-Mb window flanking the local and distant loci from the discovery set were identified in the replication set and plotted on the *y* axis. LOD scores are highly correlated at these peak loci (local Pearson *r* = 0.91; distant *r* = 0.84). **b**, For the core set of 6,707 proteins (6,647 gene ids), pQTL and eQTL overlap were compared at multiple genome-wide *P* value thresholds from 0.01 to 0.2. Again, one maximum proximal locus and one maximum distant locus were identified for each gene/protein, and recorded if it met the *P* value cut off. Local pQTL exhibit high overlap with both the discovery eQTL set and replication eQTL set, regardless of *P* value threshold (67–80%). Distant pQTL exhibit slightly higher overlap with eQTL at the most stringent *P* value cut off, however, overlap is consistently low for distant pQTL (<1–2%). Local eQTL overlap well with the replication eQTL set regardless of *P* value threshold (75–77%). Distant eQTL replicate poorly overall (3–31%), but overlap rate is highest (31%) at the most stringent *P* value threshold, suggesting that larger sample sizes will be required to fully and accurately characterize distant effects on gene expression. **c**, The maximum proximal locus and distant locus were identified for each of the 6,707 proteins and transcripts, and the cumulative distribution of their LOD scores is plotted (blue = proteins, green = transcripts). LOD score is plotted on the *x* axis, and the proportion of total QTL is plotted on the *y* axis. Local eQTL as a group exhibit higher LOD scores (consistent with higher effect sizes) than local pQTL (ninetieth percentile LOD = 23.9 for local eQTL, 13.6 for pQTL), while distant eQTL and pQTL are of similar scale (ninetieth percentile LOD = 7.9 for distant eQTL, 8.2 for distant pQTL). **d**, Comparison of pQTL from the discovery set to eQTL from the discovery set (left set of Venn diagrams) and eQTL from the replication set (right). As expected given that they derive from the same samples, local pQTL and eQTL overlap is observed to be higher in the discovery set (1,392 out of 1,736 = 80%), however, local pQTL still overlap well with eQTL from the replication set (1,273 out of 1,736 = 73%). Distant pQTL overlap poorly with both eQTL sets (9 out of 1,048 in discovery set); 8 out of 1,048 in replication set), however, 6 of 9 distant pQTL that do overlap with eQTL in the discovery set are also identified as overlapping in the replication set.


Local regulation
Distant Regulation
L1
QTL$_L$ RNA Protein (1,275)
L2
QTL$_L$ RNA Protein (304)
L3
QTL$_L$ RNA Protein (164)
L4
QTL$_L$ RNA Protein (179)
L5
QTL$_L$ RNA Protein (280)
D1
QTL$_D$ RNA Protein (17)
D2
QTL$_D$ RNA Protein (0)
D3
QTL$_D$ RNA Protein (416)
D4
QTL$_D$ RNA Protein (329)
D5
QTL$_D$ RNA Protein (1)
No pQTL
QTL$_L$ RNA Protein (298)
QTL$_L$ RNA Protein (2,193)
QTL$_L$ RNA Protein (2,014)
No pQTL
QTL$_D$ RNA Protein (0)
QTL$_D$ RNA Protein (2,282)
QTL$_D$ RNA Protein (3,662)

Extended Data Figure 5. BIC model selection reveals transcriptional mechanisms driving most local pQTL and post-transcriptional mechanisms underlying most distant pQTL

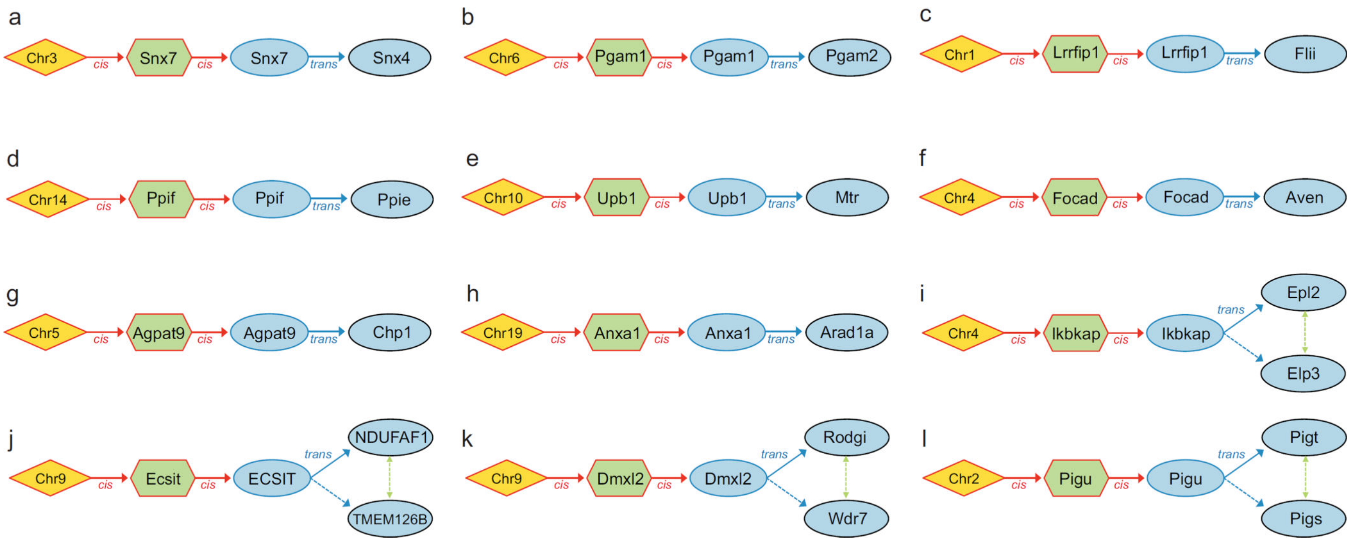
We identified the local and distant QTL with the maximum LOD score (regardless of significance) for each of the 6,707 proteins, and used BIC to assess eight models linking QTL genotype to transcript and protein abundance. Most proteins are not affected by the local or distant QTL, and fall in one of the three groups below outlined by the dotted line. Among the five models where a QTL effect on protein abundance is detected, two are transcriptional in nature (L1, L2; D1, D2); the QTL effect on protein abundance is conferred at least partially through the transcript. The remaining three genetic models are post-transcriptional (L3–5; D3–5); the QTL effect on protein abundance is not mediated through the transcript. The transcriptional L1 and L2 models are identified as the best models for most local pQTL, while the post-transcriptional D3 and D4 models are optimal for most distant pQTL.



Extended Data Figure 6. Examples of local pQTL that are due to an underlying eQTL and those that are due to post-transcriptional mechanisms

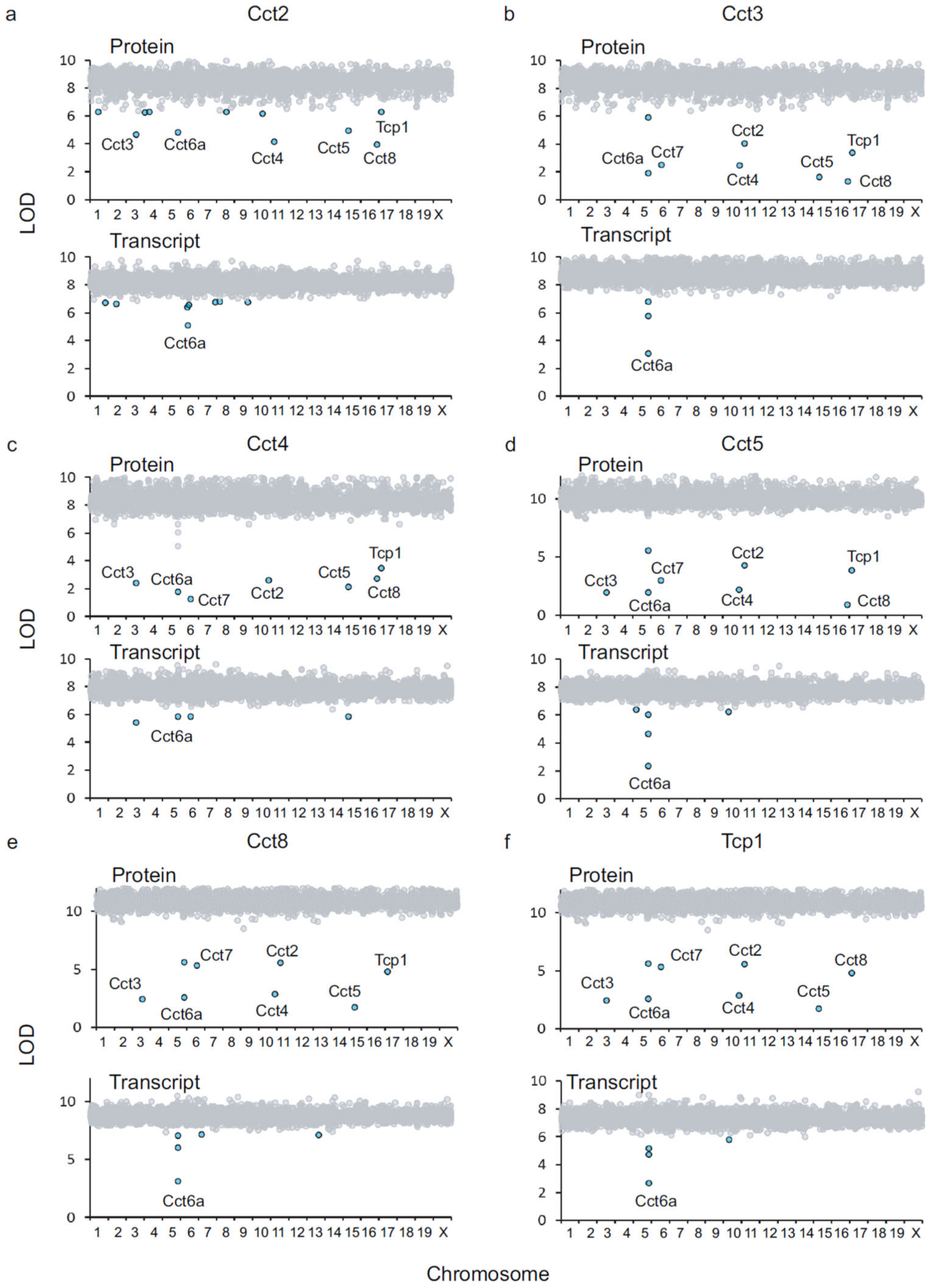
a, The protein DHTKD1 contained a local acting eQTL and pQTL, which was associated with increased transcript and protein abundance derived from 129S1/SvImJ, CAST/EiJ, PWK/PhJ and WSB/EiJ strains. Mice were divided into three groups depending on whether or not their genomes contained 0, 1 or 2 of the alleles found to be associated with the pQTL.

These increases in protein abundance were further validated using the proteomic analysis of the founder strains. **b, c**, Similarly, *Ces2h* and *Pipox* had both a local acting eQTL and pQTL that could be associated with specific strains (CAST/EiJ, PWK/PhJ and WSB/EiJ). These protein abundance measurements were further validated using the founder strains data set. **d, e**, Alternatively, 10% of the genes had local pQTL but lacked local eQTLs, which is evident in proteins such as ENTPD5 and OMA1. The founder allele expression patterns inferred at the pQTL were validated by protein abundance measurements in the founder strains, which could be explained CAST/EiJ specific missense mutations in both genes. **f**, Likewise, *Lars2* also contained a pQTL that had no observable eQTL that showed a decrease in protein abundance in the 129S1/SvImJ, CAST/EiJ, PWK/PhJ and WSB/EiJ strains. Genome sequencing determined that these strains share four missense mutations (*$P < 0.01$ using a Student's *t*-test; for founder strains, $n = 4$ mice for each founder, 2 male and 2 female, error bars represent s.d.).



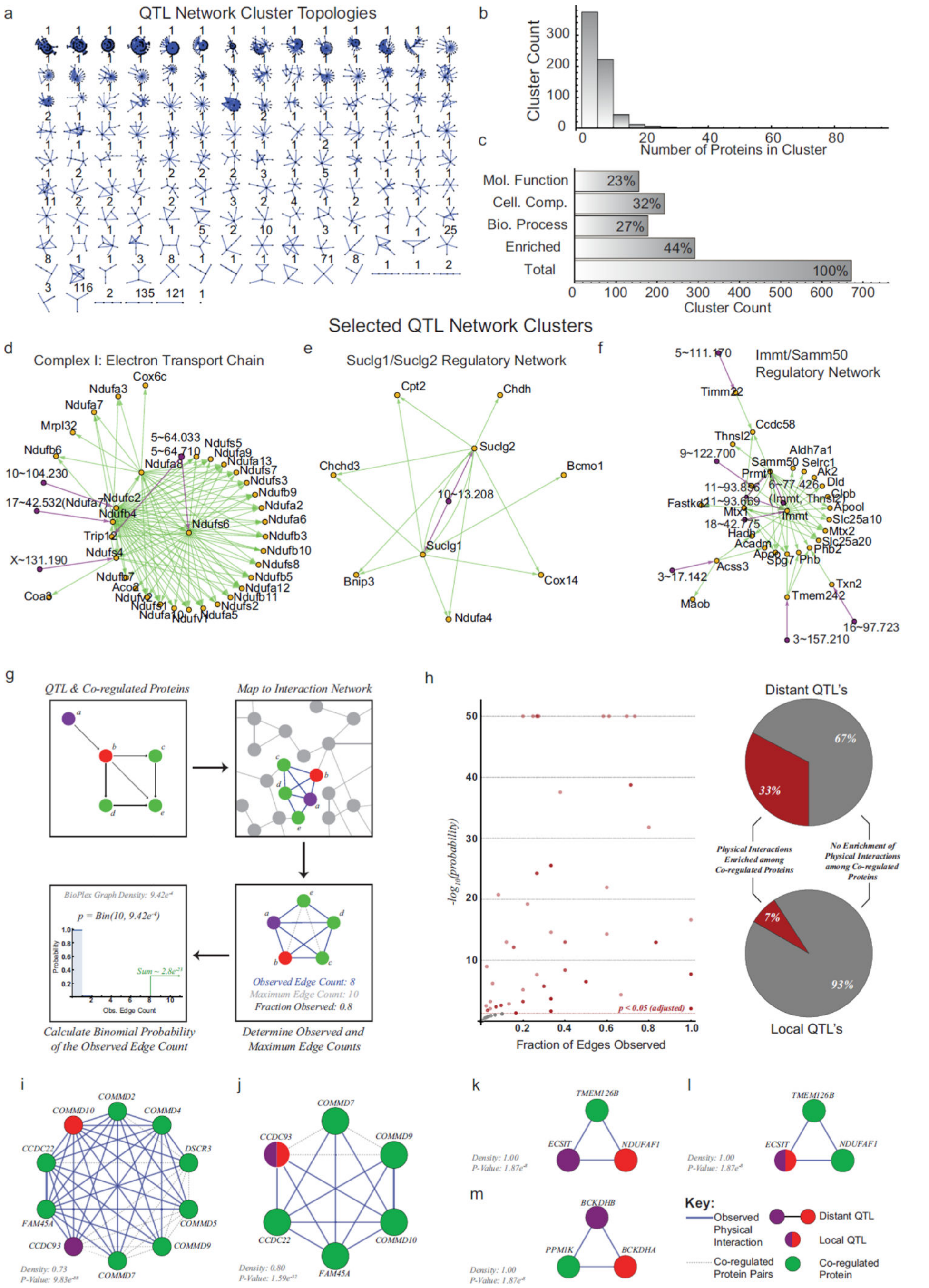
Extended Data Figure 7. The causal relationship between genetic variation and protein expression was determined for over 700 proteins as inferred by mediation analysis

a–d, Many of the causal relationships between proteins have been previously documented such as the associations between SNX7–SNX4, PGAM1–PGAM2, LRRFIP1–FLII and PPIF–PPIE. **e–h**, In addition, many of the protein associations had not be previously documented such as UPB1–MTR, FOC AD–AVEN, AGPAT9–CHP1 and ANXA1–ARAD1A. **i–l**, Protein associations were also identified for multimeric complexes such as ECSIT–NDUFAF1–TMEM126B, DMXL2–ROGDI–WDR7, PIGU–PIGT–PIGS and IKBKAP–ELP2–ELP3.



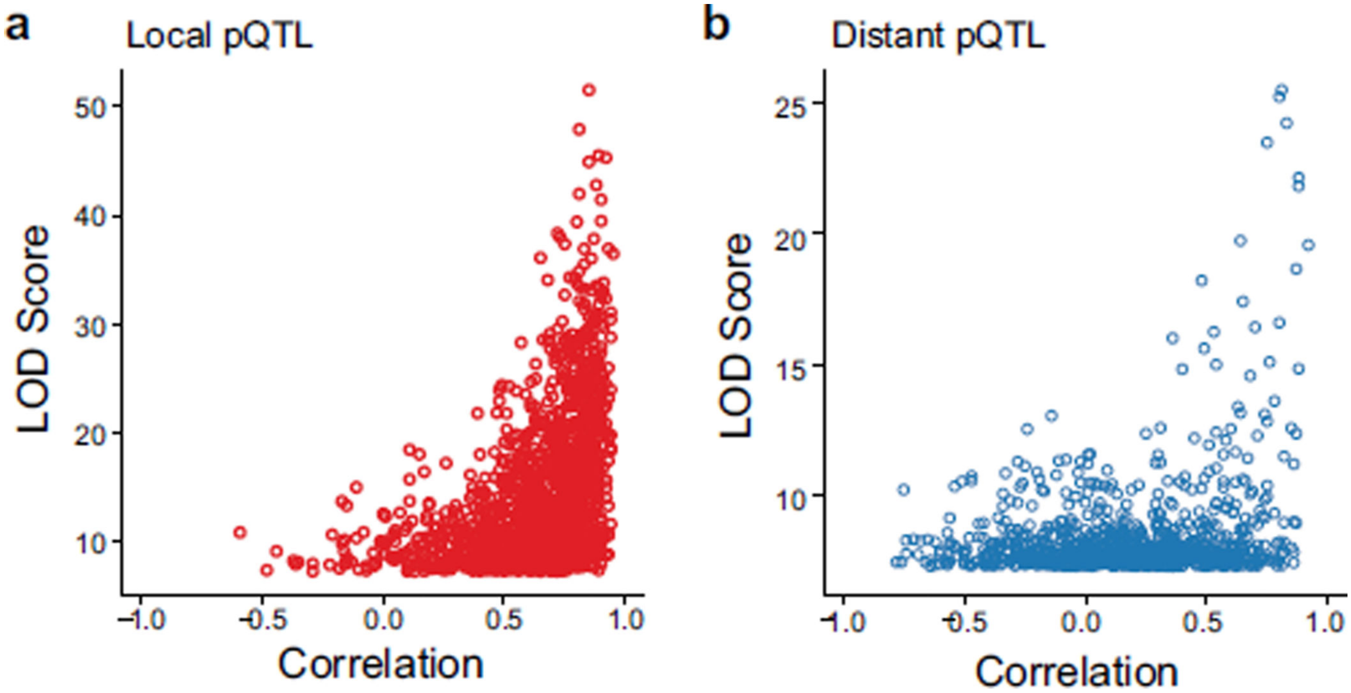
Extended Data Figure 8. Mediation analysis for CCT complex members details the effects of a QTL in *Cct6a* on protein abundance through post-transcriptional protein buffering

a–f, Mediation analysis for each of the Cct complex identifies *Cct6a* as the causal intermediate. A local QTL for *Cct6a* affects transcript and protein abundance, and CC T6A abundance sets the abundance of other CC T proteins regardless of variation in their transcripts. For each of the complex members tested, all other complex members are confirmed to be co-regulated providing additional supporting evidence for stoichiometric buffering.



Extended Data Figure 9. Distant pQTL and co-regulated proteins frequently correspond to complexes of physically interacting proteins

a, Distant pQTL and co-regulated proteins assemble to form a regulatory network, which is defined by protein clusters with distinct topologies. A total of 3,938 proteins/QTL are linked by 5,794 associations. Distant pQTL are depicted as purple arrows pointing from the inferred causal protein to its regulated pair. Co-regulated proteins are connected with green arrows emanating from the primary target protein. **b**, MCL clustering decomposes the distant pQTL network into 671 clusters. Cluster size varies considerably, although most clusters contain fewer than 20 proteins. **c**, Clusters extracted from the distant pQTL network frequently associate proteins with shared biological functions. More than half of clusters are enriched for at least one GO category, as depicted in the bar chart above. **d–f**, Three selected clusters of distant pQTL and co-regulated proteins. **g**, To understand the relationship between the distant pQTL associations and protein interactions, each distant pQTL and its co-regulated proteins were mapped to their human homologues in the BioPlex network of human protein interactions. To assess the tendency for these co-regulated proteins to cluster together, the median graph distance separating all pairs of co-regulated proteins was determined. The distribution of median distances observed for equal numbers of randomly selected proteins was also determined and used to assign a *Z*-score to each distant pQTL and its co-regulated proteins. **h**, Histogram depicting the *Z*-score distribution for distant pQTL and co-regulated proteins. *Z*-scores below −2.5 (highlighted in red) indicated that co-regulated proteins were unusually close within the BioPlex network. **i–l**, Selected distant pQTL and co-regulated proteins, mapped onto the BioPlex network of protein interactions. All shortest paths connecting distant pQTL and their regulated proteins have been extracted from the BioPlex network and displayed. Proteins inferred to be responsible for each QTL are purple, while primary regulated proteins are red and secondary co-regulated proteins are green. Grey circles represent neighbouring proteins in the BioPlex network that were not found to be co-regulated. Grey edges indicate BioPlex interactions, while Blue edges denote co-regulation uncovered from trans-QTL analysis.



Extended Data Figure 10. Comparison of protein abundance in the DO and founder strains reveals a positive correlation between pQTL significance and predictive power

a, b, For all detected liver pQTL in the DO population, founder strain allelic contributions were derived from the mapping model and compared to protein abundance measured directly from the eight founder strains. Pearson correlations are plotted against the LOD score of the pQTL for both local and distant pQTL. Predictive power tracks well with pQTL significance. Local pQTL tend to be more significant and yield higher predictive power than distant pQTL, however highly significant distant pQTL (>10 LOD) have comparable predictive power to local pQTL of similar significance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank L. Somes and M. Strobel for breeding the mice; S. Ciciotte, S. Daigle, J. Pereira, C. Snow, R. Lynch and H. Munger for extracting RNA and performing RNA-seq experiments; and A. Manichaikul for discussion on mediation analysis. Collaborative Cross strains used in this study were imported to JAX from the Systems Genetics Core Facility at the University of North Carolina (USA)[38]. Previous to their relocation to UNC, CC lines CC001, CC003 and CC017 were generated and bred at Oak Ridge National Laboratory (USA)[39]; CC line CC004 was generated and bred at Tel Aviv University (Israel)[40]. Research reported here was supported by Harvard Medical School, The Jackson Laboratory, and National Institutes of Health (NIH) grants under awards P50GM076468 (to G.A.C.), F32HD074299 (to S.C.M.), GM67945 (to S.P.G.) and U41HG006673 (to S.P.G. and E.L.H).

References

1. Crick F. Central dogma of molecular biology. Nature. 1970; 227:561–563. [PubMed: 4913914]
2. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 1999; 19:1720–1730. [PubMed: 10022859]
3. Schwanhäusser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]
4. Ghazalpour A, et al. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 2011; 7:e1001393. [PubMed: 21695224]
5. Skelly DA, et al. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. 2013; 23:1496–1504. [PubMed: 23720455]
6. Wühr M, et al. Deep proteomics of the *Xenopus laevis* egg using an mRNA-derived reference database. Curr. Biol. 2014; 24:1467–1475. [PubMed: 24954049]
7. Fu J, et al. System-wide molecular evidence for phenotypic buffering in *Arabidopsis*. Nat. Genet. 2009; 41:166–167. [PubMed: 19169256]
8. Rockman MV, Kruglyak L. Genetics of global gene expression. Nat. Rev. Genet. 2006; 7:862–872. [PubMed: 17047685]
9. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296:752–755. [PubMed: 11923494]
10. Morley M, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430:743–747. [PubMed: 15269782]
11. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422:297–302. [PubMed: 12646919]
12. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet. 2001; 17:388–391. [PubMed: 11418218]
13. Chesler EJ, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet. 2005; 37:233–242. [PubMed: 15711545]
14. Foss EJ, et al. Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. PLoS Biol. 2011; 9:e1001144. [PubMed: 21909241]
15. Foss EJ, et al. Genetic basis of proteome variation in yeast. Nat. Genet. 2007; 39:1369–1375. [PubMed: 17952072]

16. Khan Z, Bloom JS, Garcia BA, Singh M, Kruglyak L. Protein quantification across hundreds of experimental conditions. Proc. Natl Acad. Sci. USA. 2009; 106:15544–15548. [PubMed: 19717460]

17. Wu Y, et al. Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell. 2014; 158:1415–1430. [PubMed: 25215496]

18. Wu L, et al. Variation and genetic control of protein abundance in humans. Nature. 2013; 499:79–82. [PubMed: 23676674]

19. Damerval C, Maurice A, Josse JM, de Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics. 1994; 137:289–301. [PubMed: 7914503]

20. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L. Genetics of single-cell protein abundance variation in large yeast populations. Nature. 2014; 506:494–497. [PubMed: 24402228]

21. Ting L, Rad R, Gygi SP, Haas W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. Nat. Methods. 2011; 8:937–940. [PubMed: 21963607]

22. McAlister GC, et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal. Chem. 2014; 86:7150–7158. [PubMed: 24927332]

23. Churchill GA, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 2004; 36:1133–1137. [PubMed: 15514660]

24. Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. Mamm. Genome. 2012; 23:713–718. [PubMed: 22892839]

25. Threadgill DW, Churchill GA. Ten years of the collaborative cross. Genetics. 2012; 190:291–294. [PubMed: 22345604]

26. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477:289–294. [PubMed: 21921910]

27. Gatti DM, et al. Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda). 2014; 4:1623–1633. [PubMed: 25237114]

28. Toye AA, et al. A genetic and physiological study of impaired glucose homeostasis control in C57BL/6J mice. Diabetologia. 2005; 48:675–686. [PubMed: 15729571]

29. Ronchi JA, et al. A spontaneous mutation in the nicotinamide nucleotide transhydrogenase gene of C57BL/6J mice results in mitochondrial redox abnormalities. Free Radic. Biol. Med. 2013; 63:446–456. [PubMed: 23747984]

30. Freeman HC, Hugill A, Dear NT, Ashcroft FM, Cox RD. Deletion of nicotinamide nucleotide transhydrogenase: a new quantitive trait locus accounting for glucose intolerance in C57BL/6J mice. Diabetes. 2006; 55:2153–2156. [PubMed: 16804088]

31. Huttlin EL, et al. The BioPlex Network: a systematic exploration of the human interactome. Cell. 2015; 162:425–440. [PubMed: 26186194]

32. van Weering JRT, et al. Molecular basis for SNX-BAR-mediated assembly of distinct endosomal sorting tubules. EMBO J. 2012; 31:4466–4480. [PubMed: 23085988]

33. Liu Y-T, Yin HL. Identification of the binding partners for flightless I, A novel protein bridging the leucine-rich repeat and the gelsolin superfamilies. J. Biol. Chem. 1998; 273:7920–7927. [PubMed: 9525888]

34. Huttlin EL, et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell. 2010; 143s:1174–1189.

35. Battle A, et al. Genomic varation. Impact of regulatory variation from RNA to protein. Science. 2013; 347:664–667.

36. Laurent JM, et al. Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics. 2010; 10:4209–4212. [PubMed: 21089048]

37. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 2012; 13:227–232. [PubMed: 22411467]

38. Welsh CE, et al. Status and access to the Collaborative Cross population. Mamm. Genome. 2012; 23:706–712. [PubMed: 22847377]

39. Chesler EJ, et al. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. Mamm. Genome. 2008; 19:382–389. [PubMed: 18716833]

40. Iraqi FA, Churchill G, Mott R. The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. Mamm. Genome. 2008; 19:379–381. [PubMed: 18521666]

41. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 1994; 5:976–989. [PubMed: 24226387]

42. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods. 2007; 4:207–214. [PubMed: 17327847]

43. de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004; 20:1453–1454. [PubMed: 14871861]

44. Welsh CE, McMillan L. Accelerating the inbreeding of multi-parental recombinant inbred lines generated by sibling matings. G3 (Bethesda). 2012; 2:191–198. [PubMed: 22384397]

45. Broman KW, et al. Haplotype probabilities in advanced intercross populations. G3 (Bethesda). 2012; 2:199–202. [PubMed: 22384398]

46. Munger SC, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014; 198:59–73. [PubMed: 25236449]

47. Cheng R, Abney M, Palmer AA, Skol AD. QTLRel: an R package for genome-wide association studies in which relatedness is a concern. BMC Genet. 2011; 12:66. [PubMed: 21794153]

48. Dudbridge F, Koeleman BPC. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am. J. Hum. Genet. 2004; 75:424–435. [PubMed: 15266393]

49. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J. R. Stat. Soc. Series B. 2004; 66:187–205.

50. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. 1986; 51:1173–1182. [PubMed: 3806354]

51. Fritz MS, Mackinnon DP. Required sample size to detect the mediated effect. Psychol. Sci. 2007; 18:233–239. [PubMed: 17444920]

52. Yvert G, et al. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat. Genet. 2003; 35:57–64.

53. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013; 9:e1003709. [PubMed: 23990802]

54. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30:1575–1584. [PubMed: 11917018]

55. Finn RD, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–D230. [PubMed: 24288371]

56. Magrane M. UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011; 2011:bar009. [PubMed: 21447597]

57. Ashburner M, et al. Gene ontology: tool for the unification of biology. Nat. Genet. 2000; 25:25–29. [PubMed: 10802651]

58. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing on JSTOR. J. R. Stat. Soc. B. 1995; 57:289–300.

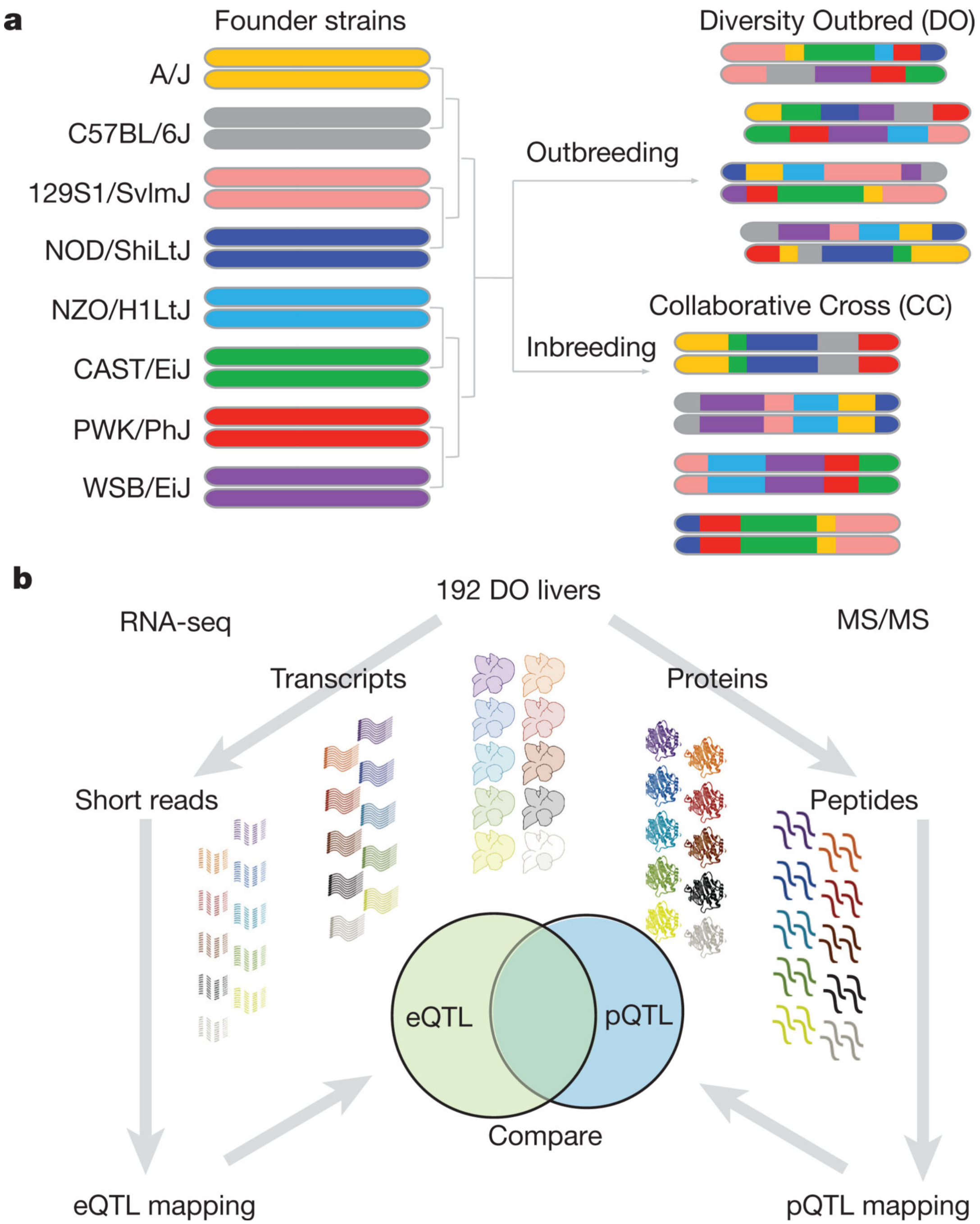


Figure 1. Tandem mass tag (TMT)-based liver proteomics in 192 DO mice

a, Overview of the breeding scheme to create the DO and CC mouse strains. **b**, Experimental overview of the genotyping, transcriptomics and proteomic analysis on 192 DO mouse livers from both sexes on a high-fat or chow diet.

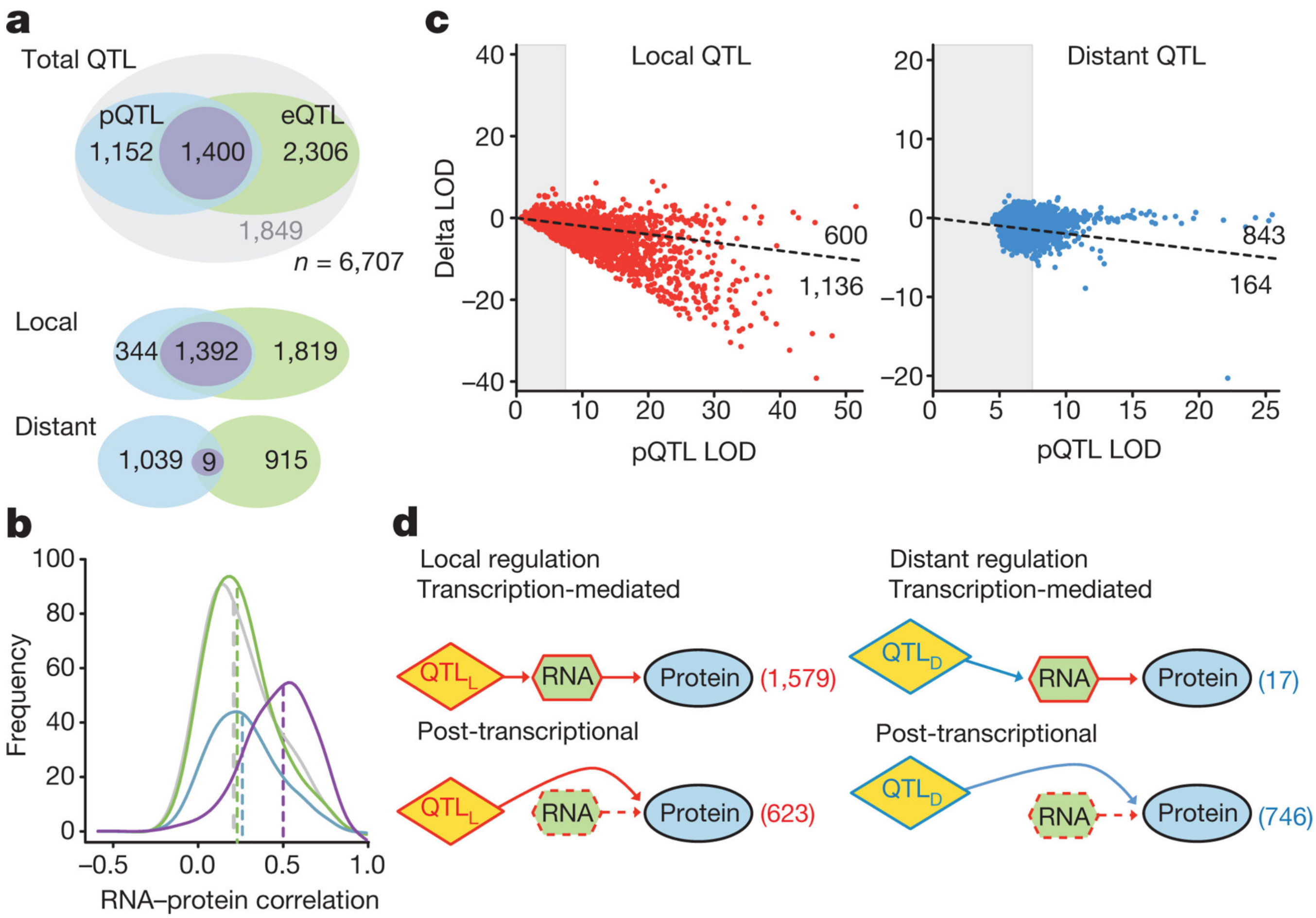


Figure 2. Global view of the liver proteome reveals distinct genetic models of protein regulation

a, Venn diagram showing the distribution of transcripts and proteins broken down into local or distant QTL. **b**, Histograms of Pearson correlations for each gene’s protein and transcript measurements after segregating into four groups (eQTL–pQTL (purple), pQTL–no eQTL (blue), eQTL–no pQTL (green) and no QTL (grey)). **c**, Local and distant pQTL LOD scores after transcript measurements were used as a covariate in the regression model showing that local pQTL were mediated through their cognate transcripts unlike distant pQTL. **d**, Model selection by Bayesian information criterion (BIC). Local pQTL (QTL$_L$) were mostly transcriptionally controlled, whereas distant pQTL (QTL$_D$) were regulated generally by post-transcriptional mechanisms.

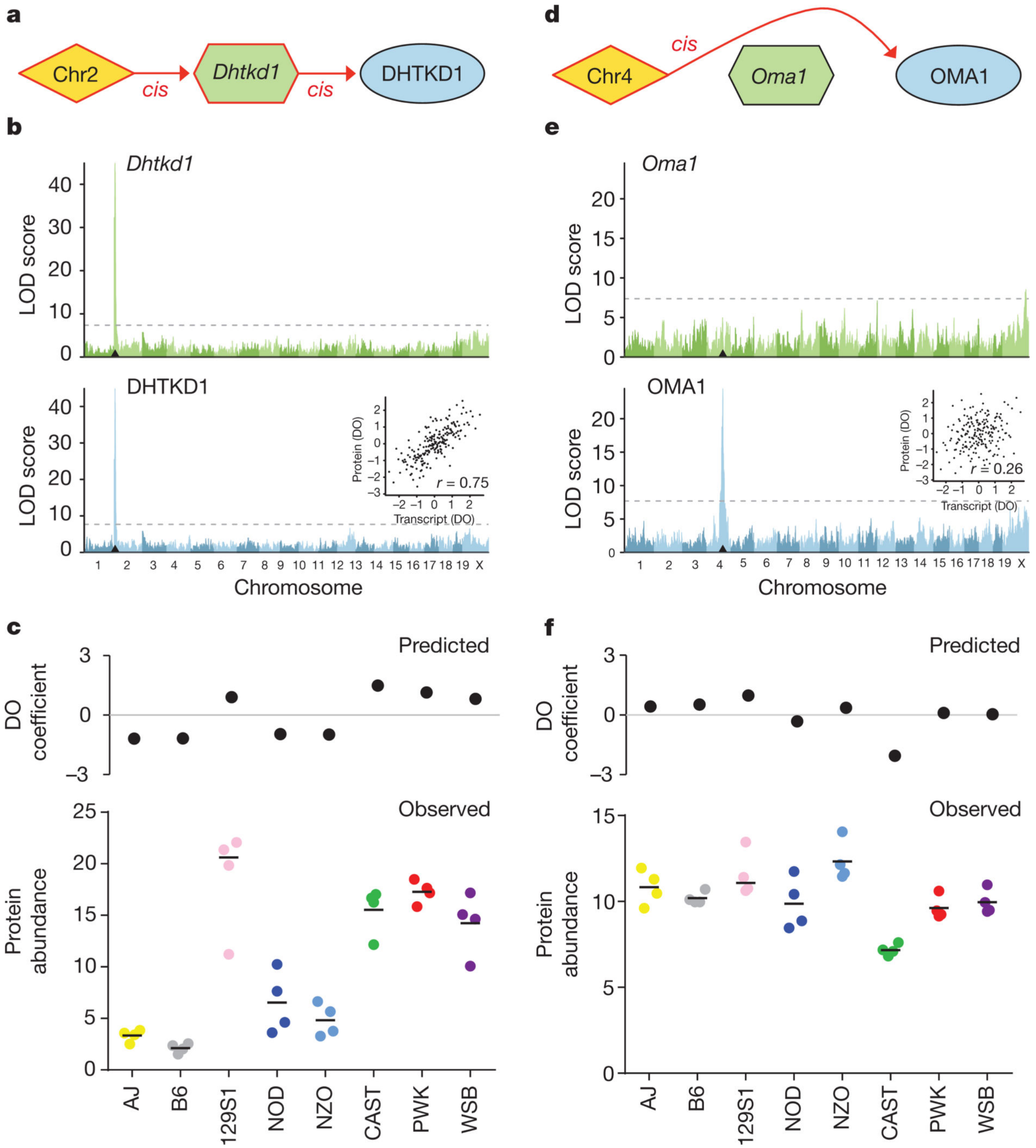


Figure 3. Examples of local pQTL that illustrate different models of regulation

a, DHTKD1 abundance is regulated by a local pQTL that probably acts proximally on transcript abundance. **b**, *Dhtkd1* has a strong local eQTL (green) and local pQTL (blue), which corresponds to high correlation between transcript and protein abundance (inset; abundance data transformed to rank normal scores for comparison). **c**, The predicted founder strain abundance of DHTKD1 in the DO population mirrors the measured abundance of DHTKD1 in the founder strains (*n* = 4 mice for each founder, 2 male and 2 female, black bars represent median values). **d**, OMA1 follows a mode of regulation in which the pQTL acts directly on protein abundance without affecting transcript levels. **e**, OMA1 protein abundance is controlled by a strong local pQTL without a corresponding local eQTL, leading to low correlation (inset) observed between protein and transcript abundance. **f**, The predicted founder strain expression in the DO population is highly correlated to measured OMA1 abundance in the founder strains (*n* = 4 mice for each founder, 2 male and 2 female, black bars represent median values).

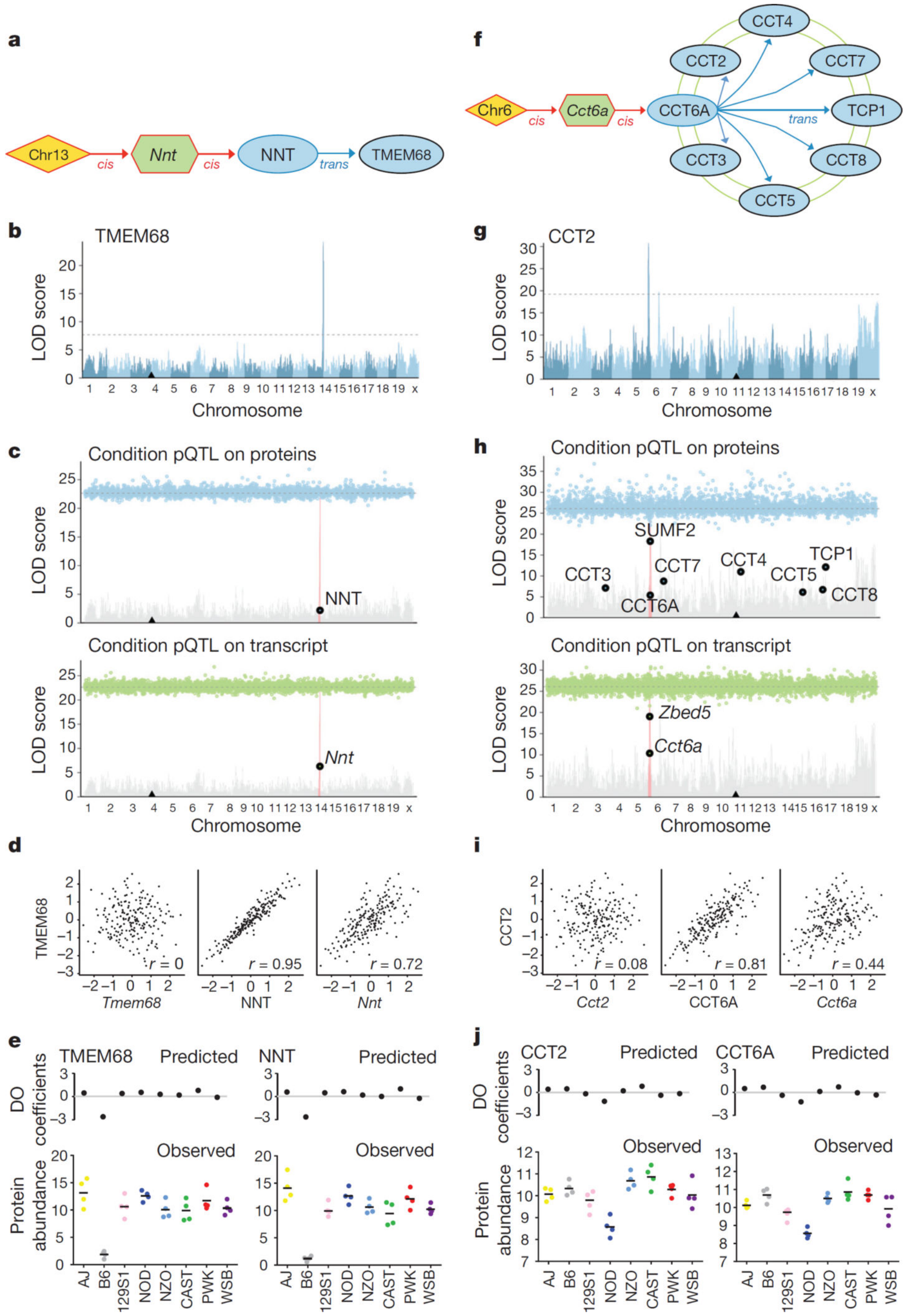


Figure 4. Mediation of distant pQTL reveals network interactions in the liver proteome

a, The genetic variant underlying the distant TMEM68 pQTL acts proximally in *cis* on *Nnt* transcript and protein abundance. **b**, TMEM68 protein abundance is buffered against local genetic variation affecting transcript levels by a distant regulator on chromosome 13. **c**, Mediation analysis identified NNT protein and *Nnt* transcript as the likely mediator. **d**, TMEM68 protein is poorly correlated to its corresponding transcript, but highly correlated with both NNT protein and *Nnt* transcript abundance. **e**, TMEM68 strain abundance predicted at the chromosome 13 distant pQTL in the DO population is highly correlated to

TMEM68 and NNT abundance measured in the founder strains, and matches the predicted NNT strain abundance in the DO population (*n* = 4 mice for each founder, 2 male and 2 female, black bars represent median values). In all cases the C57BL/6J allele is observed to be the low expressor. **f**, The chromosome 5 variant responsible for the distant effect on CC T2 abundance acts proximally in *cis* on *Cct6a* transcript and protein abundance. **g**, All members of the chaperonin containing Tcp1 (CC T) complex including CC T2 exhibit a distant pQTL that maps to distal chromosome 5. **h**, Mediation analysis identified *Cct6a*/CC T6A as the probable mediator of this effect. Protein mediation shows that the protein abundance of all CC T complex members is highly correlated as all members are pulled down in the background of the mediation plot. **i**, CC T2 protein abundance is highly correlated to CC T6A protein and *Cct6a* transcript abundance. All other CC T complex members show this same pattern. **j**, CC T2 abundance predicted at the chromosome 5 distant pQTL is highly correlated with CC T2 and CC T6A abundance measured in the founder strains, and tracks with CC T6A abundance predicted at the pQTL in the DO population (*n* = 4 mice for each founder, 2 male and 2 female, black bars represent median values). DO animals that derive the chromosome 5 region from NOD/ShiLtJ have lower abundance of all CC T proteins.

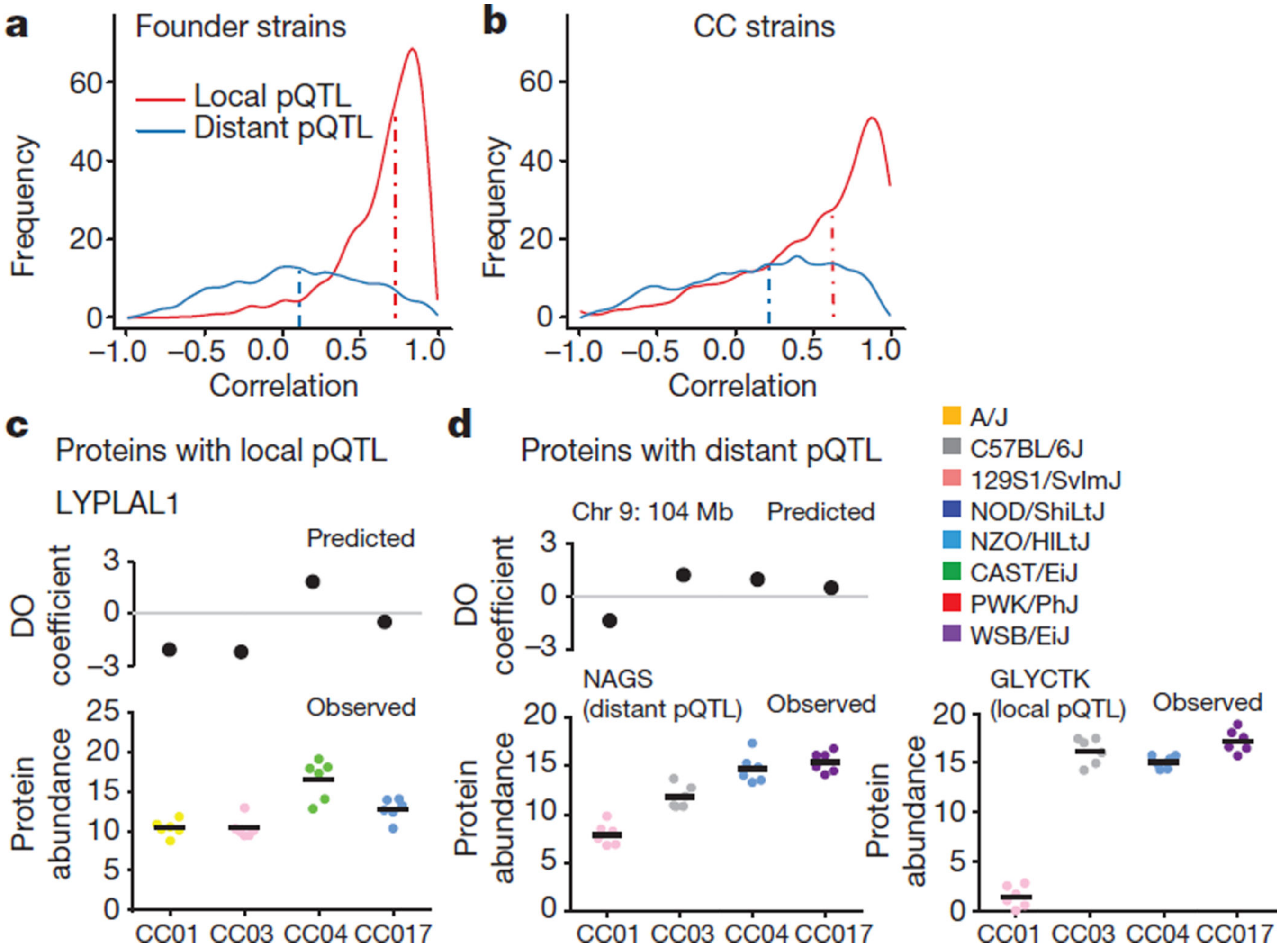


Figure 5. Genotype can be an accurate predictor of protein abundance

a, Founder strain protein abundance values inferred at significant pQTL in the DO population closely match measured abundance values from the founder strains themselves. The distributions of Pearson correlations are plotted for local pQTL and distant pQTL. Local pQTL are generally more predictive of abundance values in the founder strains (local median $r = 0.72$, distant median $r = 0.11$). **b**, Founder strain allele predictions from the DO were also assessed against protein abundance data collected from four CC strains ($n = 6$ mice per strain). We observe that local pQTL are more predictive of protein abundance in the CC strains (local median $r = 0.63$; distant median $r = 0.22$). **c**, Predictive power depends largely on the significance of the pQTL. Local pQTL generally had higher LOD scores, and as such we had higher power to predict these proteins ($n = 4$ mice for each founder, 2 male and 2 female, black bars represent median values). An example is shown for LYPLAL1. **d**, Protein abundance could also be predicted for genes with significant distant pQTL in the DO population; however, as a group these predictions were modest compared to local pQTL. As an example, NAGS abundance in the CC strains could be predicted based on the local genotype at its mediator protein, GLYCTK ($n = 6$ mice for each CC strain, 3 male and 3 female, black bars represent median values).