Genetic Architecture of Protein Expression and its Regulation in the Mouse Brain

Alyssa Erickson
University of North Dakota https://orcid.org/0000-0003-4528-314X
Suiping Zhou
St Jude Children's Research Hospital
Jie Luo
Zhejiang Academy of Agricultural Sciences
Ling Li
University of North Dakota https://orcid.org/0000-0001-6295-2128
He Huang
University of North Dakota
Hai-Ming Xu
Zhejiang University College of Agriculture and Biotechnology https://orcid.org/0000-0002-0675-4087
Junmin Peng
St Jude Children's Research Hospital
Lu Lu
UTHSC: The University of Tennessee Health Science Center
Xusheng Wang (✉ xusheng.wang@und.edu)
University of North Dakota https://orcid.org/0000-0002-1759-9588

Research article

Keywords: Proteome, Mouse, Brain, Protein expression, Expression pattern, Protein regulation

Posted Date: June 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-608734/v1

Version of Record: A version of this preprint was published at BMC Genomics on December 1st, 2021. See the published version at https://doi.org/10.1186/s12864-021-08168-y.

Abstract

Background

Natural variation in protein expression is common in all organisms and contributes to phenotypic differences among individuals. While variation in gene expression at the transcript level has been extensively investigated, the genetic mechanisms underlying variation in protein expression have lagged considerably behind. Here we investigate genetic architecture of protein expression by profiling a deep mouse brain proteome of two inbred strains, C57BL/6J (B6) and DBA/2J (D2), and their reciprocal F1 hybrids using two-dimensional liquid chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) technology.

Results

By comparing protein expression levels in the four mouse strains, we observed 329 statistically significant differentially expressed proteins between the two parental strains and identified four common inheritance patterns, including 1,133 dominant, 980 additive, 63 over- and 62 under-dominant expression. We further applied the proteogenomic approach to detect variant peptides and define protein allele-specific expression (pASE), identifying 33 variant peptides with *cis*-effects and 17 variant peptides showing *trans*-effects. Comparison of regulation at transcript and protein levels show a significant divergence.

Conclusions

The results provide a comprehensive analysis of genetic architecture of protein expression and the contribution of *cis*- and *trans*-acting regulatory differences to protein expression.

Background

One of the fundamental goals of biological research is to understand the genetic basis of phenotypic variation. The phenotypic variation is substantially contributed by the regulation of gene expression at both transcriptional and protein levels (1). Previous studies revealed that the regulation of gene expression is ubiquitous in humans and other organisms and is controlled by the interplay between genetic and environmental factors (2, 3). The regulation of gene expression at the cell-type and single-cell level has also recently been investigated owing to advances in single-cell RNA-seq technology (4, 5). While proteins are more relevant to phenotypic variation than transcripts, the regulation of protein expression has lagged behind considerably.

Recently, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) technology has become a powerful platform for profiling deep proteomes, enabling us to investigate the regulation of protein expression. Genome-wide analysis of mRNA and protein expression in mice revealed a discrepancy between their regulations using quantitative trait loci (QTLs) mapping (3, 6). In addition, allele-specific expression (ASE) is also used to further dissect the regulation into *cis*- and *trans*-components. Although ASE at the transcript level has been extensive explored and pervasive allelic imbalance across different tissues were identified (7, 8), only one study to date examined protein allele-specific expression (pASE) in yeast using stable isotope labeling by amino acids in cell culture (SILAC) technology (9).

Recombinant inbred (RI) strains are a useful resource for identifying genetic variation in phenotypic traits (10). The BXD RI panel, derived from C57BL/6J (B6) and DBA/2J (D2), exhibits high and uniform levels of genetic and phenotypic variations (11). Genetic regulation of gene expression and ASE at the transcript level have been studied in multiple tissues from the BXD RI panel. For example, over 50% of transcripts showing ASE in the liver were detected in the two reciprocal F1 hybrids (B6D2F1 and D2B6F1) (12, 13). However, the transcript level is often not an accurate indicator of the protein abundance (14, 15). More importantly, genetic inheritance of protein expression and protein ASE (pASE) are not well defined.

To characterize the genetic regulation of protein expression, we first perform deep proteome profiling of brain tissue from B6 and D2 mouse strains, as well as their two reciprocal F1 hybrids using 11-plex tandem mass tag (TMT)-based LC/LC-MS/MS (Fig. 1A). We then characterize differentially expressed proteins between the two parental strains (Fig. 1B) and systematically determine the inheritance patterns of protein expression using protein co-expression network analysis (Fig. 1C). We finally define *cis*- and *trans*-regulation of protein expression using the proteogenomics approach (Fig. 1D) and examine the difference in the regulation at transcript and protein levels (Fig. 1E).

Results

Comprehensive and quantitative proteome profiling of the mouse brain

To characterize genetic architecture of protein expression, we generate deep brain proteome of four mouse strains, including B6 and D2, and their two reciprocal F1 hybrids (i.e., B6D2F1 and B6D2F1) (Fig. 2A). By using 11-plex TMT-based LC/LC-MS/MS with extensive fractionation, we identified a total of 273,063 peptide-spectrum matches (PSMs) and 87,892 peptides, corresponding to 9,979 proteins (9,688 genes) at protein FDR < 1% (Fig. 2A; **Additional file 1: Table S1**). Principal component analysis shows that two replicates of four mouse strain grouped well (Fig. 2B). Pearson correlation analysis also shows a high correlation (r^2 = ~ 0.99) between the two replicates (**Figure S1**). The agreement of biological replicates indicates a high quality of the proteomics data.

We next ask to what extent mRNAs can be detected by our proteomic data and differences between mRNAs across four mouse strains reflect at the level of proteins. To this end, we compared our proteomic data with transcriptome from the same mouse strains generated by RNA sequencing (7) with respect to protein coverage and absolute abundance (i.e., concentration). The proteins identified in this study cover most (80.5%) of highly abundant genes (i.e., $\log_2$(tpm) > 5), indicating deep coverage of the expressed proteome (Fig. 2C). Consistent with previous reports (14, 15), comparison of absolute abundance showed a modest correlation between mRNA and proteins (correlation coefficient $r^2 = 0.459$; p value < 2.2×10^{-16}) (Fig. 2D), suggesting that ~ 45% of the variance of expression levels between different proteins could be explained by mRNA levels. The discrepancy could be ascribed to protein translation rates and post-translational modifications as well as biases of RNA-sequencing and mass spectrometry technologies.

Genetic variation can lead to difference in expression level of the same protein across different mouse strains. To determine which proteins are influenced by the genetic variation, we calculated the coefficient of variation (CV) for each protein. We found that a subset of proteins (1,347/9,979) showed high variation in protein expression (Fig. 2E), defined as two standard deviation above the average of the CV. Gene Ontology (GO) enrichment analysis showed that these variable proteins were enriched in chromatin modification and protein secretion.

Genetic difference in protein expression

We next sought to identify differentially expressed proteins (DEPs) between B6 and D2 strains, as they are highly polymorphic in genotypes and phenotypes. We identified 329 DEPs at the FDR of 0.05 and log2 fold change ($\log_2$FC) cut-off of 1.5 (Fig. 3A; **Additional file 2: Figure S2**), including 113 and 216 proteins with higher expression in B6 and in D2, respectively (**Additional file 1: Table S2**). A large proportion (71.4%) of proteins show a modest level of expression alteration ($\log_2$FC between 1.5 and 2). Among these 322 out of 329 DE proteins on autosomes, we identified 25 proteins with single parent expression (SPE), defined as an extreme form of differential expression in which either B6 or D2 shows a high expression abundance ($\log_2$ expression level z-score > 25th percentile) while the other is silent ($\log_2$ expression level z-score < 1st percentile) (Fig. 3B). Enrichment analysis revealed that DEPs with higher relative expression in B6 were significantly enriched for GO terms related to prostaglandin response (Fig. 3C, **Additional file 1: Table S3**). In contrast, DEPs with higher relative expression in D2 were significantly enriched for terms related to mitochondrial function, electron transfer activity, and cytochrome-c oxidase activity (Fig. 3C; **Additional file 1: Table S3**).

We also define a correlation of the relative expression (B6 vs. D2) between mRNA and protein levels. We observed a fairly low correlation of relative expression ratio in mRNA compared to protein (Fig. 3D), indicating potential buffering at the protein level caused by genetic variation. Despite their low correlation, we confirmed consistent changes at both transcript and protein levels, such as ALAD and HDHD3 (Fig. 3E,F), for which, in our previous study, we found that ALAD and HDHD3 span with a CNV and are associated with high variation in mRNA expression in multiple brain regions between B6 and D2 strains (11).

Systematic characterization of inheritance pattern of protein expression

The differences in protein expression across strains can be further partitioned into heritable and non-heritable variation. The proportion of heritable variation (genetic) contributing to the total observed variation is known as the broad sense heritability (H^2). We calculated heritability as the ratio of strain variance to the total variance for each protein. We found that 8,470 (84.9% of the total) proteins showed heritability > 50% (Fig. 4A). The median heritability of all proteins is 0.82, which is higher than that of transcripts in BXD strains (16). We evaluated the inheritance patterns of protein expression using the distribution of D/A (i.e., dominance / additivity), revealing that the dominant expression pattern is more common than the additive pattern (Fig. 4B).

We then determined inheritance patterns by comparing protein expression in the F1 hybrids and in two parent strains. The patterns include additive effect (A), dominant effect (D), over-dominance, and under-dominance. We next applied the weighted gene co-expression network analysis (WGCNA) to define the inheritance patterns of protein expression. We only selected proteins with highly variable expression across samples by removing proteins with a coefficient of variation (CV) less than that of the mean of the lower CV distribution (Fig. 4C). Considering the reference protein database is largely derived from the B6 genome, we only focused on proteins with higher expression in B6, resulting in a total of 2,238 proteins to be used as input for co-expression network analysis (Fig. 4D). With soft-threshold power ($\beta = 18$) in WGCNA, we identified a total of four major patterns: additive, dominant, over- and under-dominant expressions. The largest pattern is dominant expression ($n = 1{,}133$), followed by additive expression ($n = 980$) (Fig. 4E; **Additional file 1: Table S4**). In addition, we also observed 63 over- and 62 under-dominant expressions. To assess the biological functions of each inheritance pattern, we performed gene ontology enrichment analysis using the R package clusterProfiler (version 3.18.1). We found that proteins in the additive pattern are enriched for signal transduction pathways (**Additional file 1: Table S5, Additional file 2: Figure S3A**), whereas dominant expression are enriched for pathways such as mRNA processes and autophagy (**Additional file 1: Table S5, Additional file 2: Figure S3B**).

Identification of protein allele specific expression in *cis*- and *trans*-regulations

The highly variable genome sequences between B6 and D2 strains provide an opportunity to investigate allele-specific expression. While several studies have investigated ASE in mice using transcriptomics data, there is no research for protein allele specific expression (pASE). Recently, the proteogenomics approach that integrates genomic and proteomics data has been proven to be a valuable method in detecting variant peptides (17–19). We performed the proteogenomics analysis to detect variant peptides using JUMPg, a proteogenomics pipeline we recently developed.

Using 11,115 missense variants detected in our previous D2 sequencing project, we identified a total of 286 variant peptides, including 169 and 205 D-allele and B-allele peptides, respectively, at the peptide FDR of 1% (Fig. 5A; **Additional file 1: Table S6**). By comparing B-allele and D-allele peptides, we found a total of 88 pairs of variant peptides (Fig. 5B, **Additional file 1: Table S7**). Two examples of B-allele peptide and D-allele peptide are shown in Fig. 5C. The B-allele peptide can only be detected in B6 strain and both F1 hybrids, whereas the D allele peptide can be detected in D2 strain and both F1 hybrids. Even though the signal of the two allelic peptides cannot be directly compared due to different chemical properties that alter ionization efficiency, the ratio of the two alleles in parents and F1 hybrids can be calculated, allowing us to determine pASE. By comparing the allelic ratio in parents and F1 hybrids, the regulation of protein expression are classified into five different categories: *cis*-, *trans*-, compensatory, conserved, and unexpected biased (Fig. 5D). The ratio of the two parental alleles is contributed by a combination of *cis*- and *trans*-regulatory effects. If the allelic ratio in the F1 is similar to the parental proportions, the differential expression in the parents is likely due to variation in *cis*-acting elements because the common *trans*- effect is present in the F1 hybrids. In contrast, if the change in the allelic ratio is only observed in parental strains but not in F1s ($\log_2$ ratio < 1), it is likely to be caused by *trans*-acting factors. If there is no change in parental strains, but with significant changes in F1 hybrids, the *cis*- and *trans*- effects in the parental are compensatory. We define conserved regulation as there are no changes in both parental strains and F1 hybrids.

Among those 88 pair variant peptides with both B- and D-types (Fig. 5E), we found 33 *cis*-regulation, followed by 25 conserved regulation. In addition, there are 17 trans-regulation and 5 compensatory regulations. In addition, two proteins showed unexpected biased, which could be due to peptide false identification or quantitative measurement error in the shot-gun proteomics.

Comparison of regulations at transcript and protein expression levels

We next sought to compare regulation at transcript and protein expression levels. To identify ASE at the transcript level, we analyzed the transcriptome of the hippocampus of matched samples (i.e., B6, D2, and B6D2F1). By mapping RNA-seq data to both B6 reference and D2 customized genomes, we identified 2,630 protein-coding transcripts with ASE expression (Fig. 6A), including 215 *cis*-, 500 *trans*-, 213 compensatory, 1,666 conserved, and 36 unexpected bias regulation. By comparing regulation between transcripts and proteins, we found that there is a significant overlap between transcripts and proteins showing ASE (Fisher's exact test $p = 2.2 \times 10^{-9}$). The conserved regulation showed the highest overlapped in both levels. However, only two genes/proteins were found to show ASE in *trans*-regulations (Fig. 6B).

Discussion

In this study, we profiled the mouse brain proteome of B6 and D2 strains, as well as their two reciprocal F1 hybrids, allowing us to investigate the regulation of protein expression. With this deep proteomic data, we identified 329 DEPs between B6 and D2 strains and 25 proteins with SPE. We further detected additive, dominant, and over- and under-dominant inheritance patterns of protein expression. We finally defined the allele specific expression and *cis*- and *trans*-regulation of variant peptides detected by the proteogenomics approach. The deep proteomic data provides a unique opportunity to investigate the genetic regulation of protein expression in the mouse brain.

Our experimental design of including two inbred mice and their F1 hybrids enabled us to analyze the heterosis of protein expression. As expected, we found that the non-additive expressions (e.g., dominance, over- and under-dominance) are more prevalent than the additive expression. However, surprisingly, we observed only 25 proteins exhibiting single parent expression (SPE) at the protein level. In contrast, previous studies found 347 genes showing SPE at the transcript level in the mouse brain (20). One possible reason for this discrepancy is that divergent translation regulation buffers mis-expression of mRNA abundance (21). In addition, TMT ion suppression in the shot-gun proteomics could alleviate the difference in protein expression.

Compared with gene ASE analysis of transcriptomic data, protein ASE has not been well studied and only one study to date examined it in yeast using SILAC technology (9). Instead of using the SILAC approach, we used a TMT-based approach to quantify the expression of variant peptides. There are several advantages of the TMT method over the SILAC when applied to the pASE study: (1) it is capable of including a larger number of samples that can be compared, which can increase protein coverage and reduce batch effect. For example, we used 10 samples in one batch of a TMT experiment, whereas it requires at least five batches for a SILAC experiment; (2) since isobaric labeling methods (i.e., TMT, iTRAQ) are a chemical labeling approach, they can be applied to human samples and are less expensive to use in small mammals.

While the proteogenomics approach has been widely used to detect variant peptides (17–19), we extended its application to define ASE for variant peptides detected in quantitative proteomic data. Comparing with the fact that ASE can be defined by read counts at the transcript level, one of the major challenges in defining protein *cis*- and *trans*-regulation is that the expression level of variant peptides cannot be directly compared because they have different amino acid composition, which alters their mass and ionization efficiency, which significantly complicates accurate analysis of pASE. In this study, we propose to define the regulation using relative ratios of expression level between two types of variant peptides in F1 hybrids and their expression in their two parental strains.

One of the limitations of this study is the number of ASE events detected at the protein/peptide levels due to the constraints of the current shot-gun proteomics technology. While we generated, to the best of our knowledge, the deepest proteome coverage (~ 10,000 unique proteins) in mouse, we only detected 374 variant peptides and 88 shared B-type and D-type variant peptides out of 11,115 missense genomic variants detected from the whole genome sequencing. The number is also substantially fewer than the number of variants (2,630) detected from RNA-seq based transcriptomic data. With further advances in mass spectrometry technology, it will be possible to improve the ability to detect more variant peptides and ultimately define genome-wide pASE events.

Conclusion

In summary, our study provides a framework for investigating patterns of protein expression and allele-specific expression. Allele-specific expression could be caused by epigenetic regulation, such as methylation. Further investigations are needed to understand the mechanisms underlying allele-specific expression by methylome profiling. Further investigation may provide important insights into the pathways that protein allele-specific expression contributes to phenotypic and disease variation. With the development of cell-type proteomics technology (22), genetic regulation of protein expression and pASE will be eventually defined at the cell-type or even single cell level.

Materials And Methods

Animals

The following mouse strains were used in this study: B6, D2, and two reciprocal F1 hybrids (i.e. B6D2F1 and D2B6F1). The B6D2F1 hybrid is created by mating a B6 female to a D2 male mouse, whereas the D2B6F1 is created by mating a D2 female to a B6 male mouse. Both male and female mice of each strain were used as biological replicates in this study ($n = 2$). All animals were bred at the University of Tennessee Health Science Center (UTHSC). Animals were housed and maintained on a 12 : 12 light/dark cycle, with ad libitum access to food and water. Mice at 12-week-old were sacrificed, and whole brain tissue samples (Left and right olfactory bulbs were removed) were dissected rapidly, frozen in liquid nitrogen, and stored at − 80°C for the subsequent proteome profiling. The euthanasia was carried out by cervical dislocation. Criteria for euthanasia were based on an assessment by our veterinary staff following AAALAC guidelines.

Protein Extraction and Quantification

The frozen samples were weighed and homogenized in the lysis buffer (50 mM HEPES, pH 8.5, 8 M urea, and 0.5% sodium deoxycholate, 100 µl buffer per 10 mg tissue) with 1x PhosSTOP phosphatase inhibitor cocktail (Sigma-Aldrich). Protein concentration was measured by the BCA assay (Thermo Fisher) and then confirmed by Coomassie-stained short SDS gels.

Protein Digestion and Tandem-Mass-Tag (TMT) Labeling

Quantified protein samples (∼1 mg in the lysis buffer with 8 M urea) for each TMT channel were proteolyzed with Lys-C (Wako, 1:100 w/w) at 21°C for 2 h, diluted by 4-fold to reduce urea to 2 M for the addition of trypsin (Promega, 1:50 w/w) to continue the digestion at 21°C overnight. The digestion was terminated by the addition of 1% trifluoroacetic acid. After centrifugation, the supernatant was desalted with the Sep-Pak C18 cartridge (Waters), and then dried by Speedvac. Each sample was resuspended in 50 mM HEPES (pH 8.5) for TMT labeling, and then mixed equally, followed by desalting for the subsequent fractionation. For whole proteome analysis alone, 0.1 mg protein per sample was used.

Extensive Two-Dimensional Liquid Chromatography-Tandem Mass Spectrometry (LC/LC-MS/MS)

The TMT labeled samples were fractionated by offline basic pH reverse phase LC, yielding 40 fractions. Each fraction was analyzed by the acidic pH reverse phase LC-MS/MS (Wang, 2015, Journal of Proteome

Research). In the acidic pH LC-MS/MS analysis, each fraction was run sequentially on a column (75 µm x 20 cm for the whole proteome, 50 µm x ∼30 cm for phosphoproteome, 1.9 µm C18 resin from Dr. Maisch GmbH, 65°C to reduce backpressure) interfaced with a Q Exactive HF Orbitrap or Fusion MS (Thermo Fisher). Peptides were eluted by a 2–3 hr gradient (buffer A: 0.2% formic acid, 5% DMSO; buffer B: buffer A plus 65% acetonitrile). MS settings included the MS1 scan (410–1600 m/z, 60,000 or 120,000 resolution, 1×10^6 AGC and 50 ms maximal ion time) and 20 data-dependent MS2 scans (fixed first mass of 120 m/z, 60,000 resolution, 1×10^5 AGC, 100–150 ms maximal ion time, HCD, 35–38% normalized collision energy, ∼1.0 m/z isolation window).

Identification of Proteins by Database Search with JUMP Software

We used JUMP search engine (23) to search MS/MS raw data against a composite target/decoy database to evaluate FDR (24). All original target protein sequences were reversed to generate a decoy database that was concatenated to the target database. FDR in the target database was estimated by the number of decoy matches (nd) and the number of target matches (nt), according to the equation (FDR = nd/nt), assuming mismatches in the target database were the same as in the decoy database. The target database was downloaded from the UniProt mouse database (59,423 entries), and decoy database was generated by reversing targeted protein sequences. Major parameters included precursor and product ion mass tolerance (± 15 ppm), full trypticity, static mass shift for the TMT tags (+ 229.16293) and carbamidomethyl modification of 57.02146 on cysteine, dynamic mass shift for Met oxidation (+ 15.99491), maximal missed cleavage ($n = 2$), and maximal modification sites ($n = 3$). Putative PSMs were filtered by mass accuracy and then grouped by precursor ion charge state and filtered by JUMP-based matching scores (Jscore and ΔJn) to reduce FDR below 1% for proteins during the whole proteome analysis. If one peptide could be generated from multiple homologous proteins, based on the rule of parsimony, the peptide was assigned to the canonical protein form in the manually curated Swiss-Prot database. If no canonical form was defined, the peptide was assigned to the protein with the highest PSM number.

TMT-based Peptide/Protein Quantification by JUMP Software Suite

Protein expression was quantified using the following steps with JUMP software suite: (i) TMT reporter ion intensities were extracted for each PSM; (ii) the raw intensities were corrected based on isotopic distribution of each labeling reagent; (iii) PSMs with very low intensities (e.g. minimum intensity of 1,000 and median intensity of 5,000) were excluded from the subsequent analysis; (iv) Sample loading bias was normalized with the trimmed median intensity of all PSMs; (v) the mean-centered intensities across samples was calculated, (vi) protein relative intensities by averaging related PSMs was calculated; (vii) protein absolute intensities by multiplying the relative intensities by the grand-mean of three most highly abundant PSMs was computed.

Principal Component Analysis

Principal component analysis (PCA) was used to visualize the differences among samples. All gene and metabolite abundance were used as features of PCA. The pairwise Euclidean distance between features was calculated. PCA was performed using the R package prcomp (version 3.4.0).

Differential Expression Analysis

Differentially expressed proteins between the two strains were identified using the limma R package (version 3.46.0). The Benjamini-Hochberg method for false discovery rate correction was used, and proteins with an adjusted *p*-value < 0.05 and $\log_2$ fold change > 1.5 were defined as differentially expressed between the B6 and D2 strains.

Pathway Enrichment

To assess the functional relevance of the differentially expressed proteins, the R package clusterProfiler (version 3.18.1) was used for gene ontology enrichment analysis. Gene ontology terms with a Benjamini-Hochberg adjusted *p*-value < 0.05 were defined as significantly enriched.

Analysis of Patterns of Genetic Inheritance

The additive expression is defined as a pattern where the expression in the F1s does not differ from the average of the two parental strains. The dominant expression shows the F1s deviate substantially from the mid-parent value. Over-dominance and under-dominance are the patterns in which protein expression in F1s is either significantly higher or lower, respectively, than that of the parental strains.

We defined patterns of genetic inheritance using the R package WGCNA (version 1.69). The product was a weighted adjacency matrix that provided continuous connection strength ([0, 1]) based on the β parameter for each condition to meet the scale-free topology criterion. The concept of the scale-free network has emerged as a powerful paradigm in the study of network biology. Most biological networks, such as metabolic, protein, and gene interaction networks, have been reported to exhibit scale-free behavior based on the analysis of the distribution of the number of connections of the network nodes. A scale-free network is one whose majority nodes has only a few connections to other nodes, whereas some nodes (hubs) are connected to many other nodes in the network. The number of connections each node has is called its degree. If we represent the degree distribution of a scale-free network in a logarithmic scale, we can see how it fits with a line (they fit a power-law), having a small number of nodes with high degree (the hubs), and a large number of nodes with a low degree. Subsequently, the co-expression matrix and the topological overlap matrix (TOM) were constructed. For TOM, we assessed the interconnectedness of two genes by the degree of their shared neighbors across the global network. We detected the gene modules by average linkage hierarchical clustering for each group. The intra-modular connectivity of each gene was also computed using the intra-modular connectivity function in R. The module eigengene (ME) is the first principal component of a given module, and it was used to evaluate the module membership, which assesses the importance of genes in the network. We used the β power parameter of 18 for proteins with high B6 expression.

Characterization of Additive and Dominance Inheritance

The additive effect, A, is estimated as half of the observed difference between the parental strains. The dominance effect, D, was estimated as the difference between the F1 and the mid-parent values. We defined the scaled difference in expression levels between F1s and mid-parent strains as follows,

$$D/A = \frac{\left[\frac{(B6D2F1 + D2B6F1)}{2} - \frac{(B6 + D2)}{2}\right]}{\max(B6, D2) - \frac{(B6 + D2)}{2}}$$

Heritability estimation

To calculate heritability, we first estimated the variance components for strain (Vg), sex (Vs), and residue (Ve). These variances were estimated using two-way ANOVA using the following model:

$$y_{ij} = \mu + G_i + S_j + \epsilon_{ij}$$

where y*ij* is the protein expression level, G*i* is the effect associated with strain *i*, S*j* is the effect associated with sex *j*, and ε is the residual error. Heritability was then computed as the proportion of variance attributed to strain using the estimated strain, sex, and residual variance components for each protein as:

$$h^2 = \frac{V_g}{V_g + V_s + V_e}$$

Protein ASE Detection

D2 SNPs (dbSNP version: 142) were downloaded from UCSC genome browser database and were re-annotated using the genome annotation tool ANNOVAR (25) based on the GRCm38 (mm10) genome assembly. A customized protein database was constructed by appending mouse UniProt database with SNPs with the amino acid sequences of nonsynonymous variants. MS data was searched by JUMPg (19), a proteogenomic tool we recently developed. The false discovery rate (FDR) for variant peptide identification was set to 1% at peptide level.

B6 variant peptides were identified from the original peptides quantified using JUMP if they contained a nonsynonymous variant. Variant peptides with their respective alleles detected in both B6 and D2 were retained for detection of protein allele-specific expression (pASE). An empirical bayes-moderated *t*-test between alleles with a Benjamini-Hochberg adjusted *p*-value < 0.05 was used to detect peptides displaying pASE. To compare *cis*- and *trans*- regulation of pASE, allelic expression ratios were calculated as $\log_2$(B6 peptide abundance) – $\log_2$(D2 peptide abundance) in both the parental strains and F1 strains.

Transcriptomic Analysis and ASE Analysis at the Transcript Level

Paired-end RNA-seq data was downloaded from the European Nucleotide Archive for parental strains B6 and D2 whole brain tissue (accession number ERP000614) and for B6xD2 hybrid whole brain tissue (accession number ERP000591) [7]. Reads were trimmed to remove low quality sequences using Trimmomatic (version 0.39), resulting in ~ 134m read pairs for parental strains and ~ 148m read pairs for the hybrid strain (2 x 30–76 bp).

A reference sequence for D2 was created using vcftools (version 0.1.17) by merging D2 SNPs with the current GRCm38 (mm10) reference assembly. Trimmed RNA-seq reads from all samples were aligned to both the consensus D2 and the GRCm38 reference sequences using STAR (version 2.7.1) with the parameter "–outFilterMultiMapNmax 1" to only retain uniquely mapped reads. Reads that aligned to regions containing SNPs were sorted based on mapping quality to either the B6 (GRCm38) or D2 allele using a python script [26]. Genes displaying allele specific expression (ASE) were identified in the hybrid samples using a binomial test with a Benjamini-Hochberg adjusted p-value < 0.05.

Declarations

Ethics Approval and Consent to Participate

All experimental procedures were in accordance with the Guidelines for the Care and Use of Laboratory Animals published by the National Institutes of Health and were approved by the Animal Care and Use Committee at the University of Tennessee Health Science Center (UTHSC; Memphis, TN, USA; IACUC Protocol#: 18-104).

Consent for Publication

Not applicable.

Availability of Data and Supporting Materials

The dataset supporting the conclusions of this article is available in the ProteomeXchange database, with identifier PXD025830.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was partially supported by the UND Center for Biomedical Research Excellence (CoBRE) for Epigenomics of Development and Disease (5P20GM104360-07; PI: Vaughan, Roxanne A.) Pilot Grant for X.W.

Authors' Contributions

X.W. contributed to the conception and design of the project. A.E., J.L., L.L., H.H. and H.X. performed data analysis. S.Z. and J.P. performed proteomics experiments. L.L. provided mouse samples. X.W. and A.E. wrote the manuscript.

References

1. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature reviews Genetics. 2015;16(4):197–212.
2. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660–5.
3. Williams EG, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann CA, et al. Systems proteomics of liver mitochondria function. Science. 2016;352(6291):aad0189.
4. Nitzan Rosenfeld JWY, Uri Alon, Peter S, Swain MB. Elowitz. Gene Regulation at the Single-Cell Level. Science. 2005;307:1962–5.
5. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, LifeLines Cohort S, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50(4):493–7.
6. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016;534(7608):500–5.
7. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477(7364):289–94.
8. Piotrowski A, Xie J, Liu YF, Poplawski AB, Gomes AR, Madanecki P, et al. Germline loss-of-function mutations in LZTR1 predispose to an inherited disorder of multiple schwannomas. Nat Genet. 2014;46(2):182–7.
9. Khan Z, Bloom JS, Amini S, Singh M, Perlman DH, Caudy AA, et al. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Molecular systems biology. 2012;8:602.
10. Ashbrook DG, Arends D, Prins P, Mulligan MK, Roy S, Williams EG, et al. A platform for experimental precision medicine: The extended BXD mouse family. Cell Systems. 2021;12(3):235 – 47.e9.
11. Wang X, Pandey AK, Mulligan MK, Williams EG, Mozhui K, Li Z, et al. Joint mouse–human phenome-wide association to test gene function and disease risk. Nat Commun. 2016;7(1):10464.
12. Ciobanu DC, Lu L, Mozhui K, Wang X, Jagalur M, Morris JA, et al. Detection, Validation, and Downstream Analysis of Allelic Variation in Gene Expression. Genetics. 2010;184(1):119–28.
13. Pandey AK, Williams RW. Genomic analysis of allele-specific expression in the mouse liver. BioRXiv. 2015.
14. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, et al. Editing DNA Methylation in the Mammalian Genome Cell. 2016;167(1):233–47. e17.

15. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–7.
16. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet. 2005;37(3):233–42.
17. Zhang W, Mazzarello R, Wuttig M, Ma E. Designing crystallization in phase-change materials for universal memory and neuro-inspired computing. Nature Reviews Materials. 2019;4(3):150–68.
18. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014;11(11):1114–25.
19. Li Y, Wang X, Cho J-H, Shaw TI, Wu Z, Bai B, et al. JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. J Proteome Res. 2016;15(7):2309–20.
20. Christopher Gregg JZ, Brandon Weissbourd S, Luo GP, Schroth. David Haig, Catherine Dulac. High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain. Science. 2010;329(5992):643–8.
21. McManus CJ, May GE, Spealman P, Shteyman A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome research. 2014;24(3):422–30.
22. Sharma K, Schmitt S, Bergner CG, Tyanova S, Kannaiyan N, Manrique-Hoyos N, et al. Cell type- and brain region-resolved mouse brain proteome. Nature neuroscience. 2015;18(12):1819–31.
23. Wang X, Li Y, Wu Z, Wang H, Tan H, Peng J. JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. Molecular cellular proteomics: MCP. 2014;13(12):3663–73.
24. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC – MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. J Proteome Res. 2003;2(1):43–50.
25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164-e.

Figures

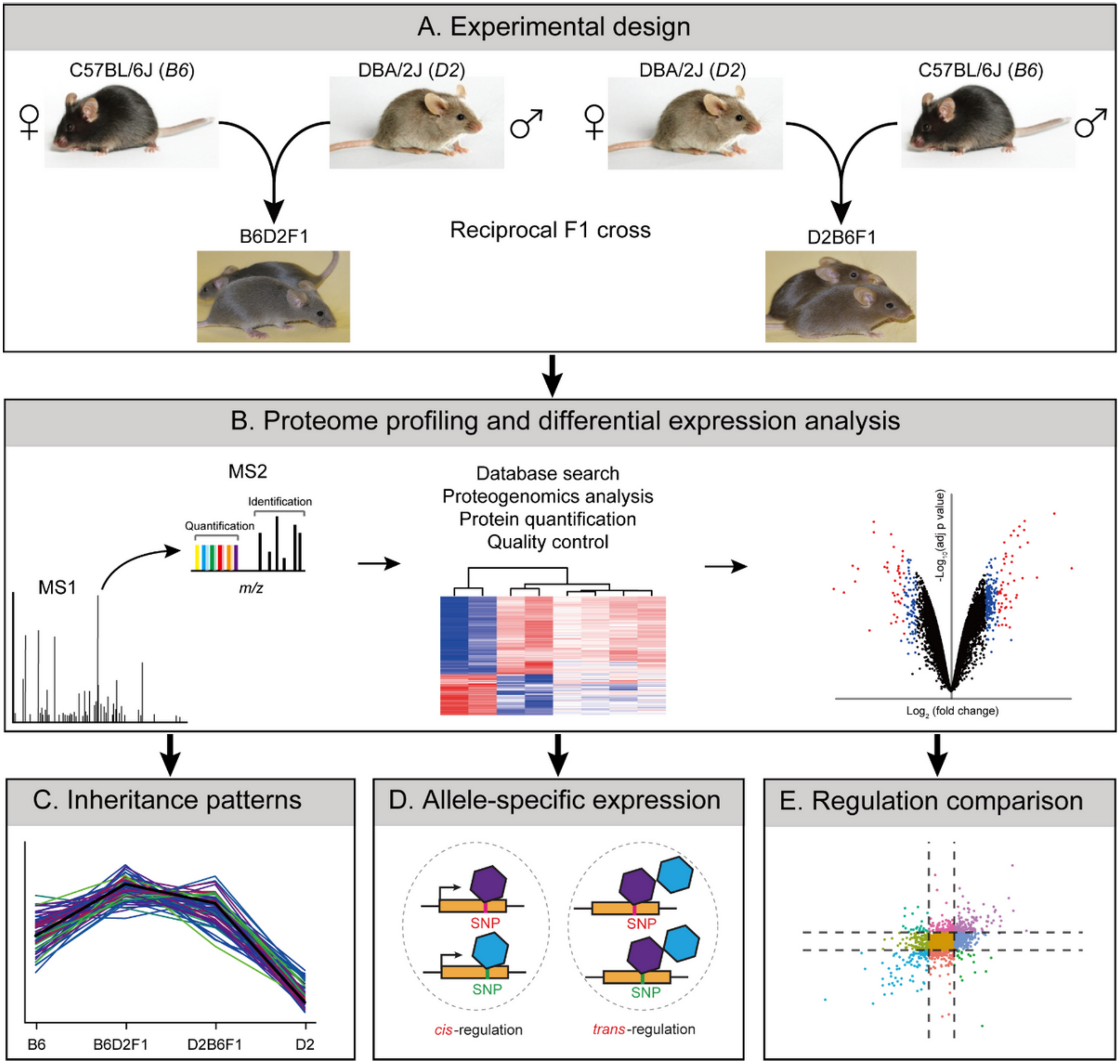

Figure 1

Schematic diagram of the experimental design and data analysis of this study. (A) Experimental scheme. Four strains, including B6, D2, and the two reciprocal F1s (B6D2F1 and D2B6F1), were used. (B) Mouse brain proteome was profiled by TMT-based proteomics, followed by data quality control and differential expression analysis. (C) Inheritance patterns were detected by WGCNA. (D) allele-specific expression was defined by the proteogenomics approach. (E) Regulation at the transcript and protein levels were compared.

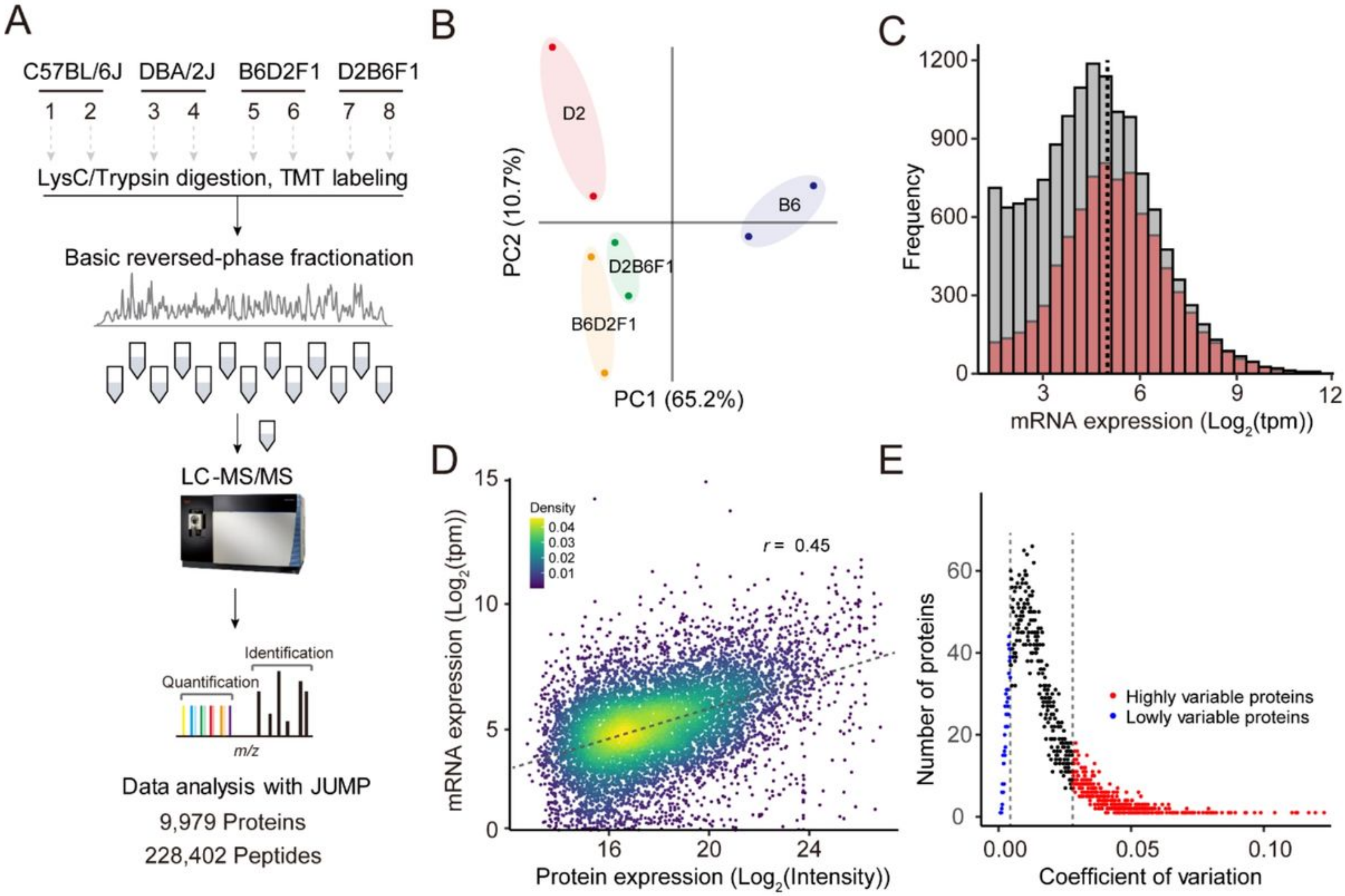

Figure 2

Proteome-wide profiling of mouse brain tissue. (A) 10-plex TMT-based global proteome analysis workflow. A total of 10 samples were analyzed by LC/LC-MS/MS. All proteomic data were analyzed using JUMP software. More than 228,000 distinct peptides, corresponding to 9,979 proteins, were identified and quantified. (B) Principal-component analysis of all quantified proteins. (C) Histogram showing the coverage of proteomic data compared to RNAseq data from B6 and D2 mice. The open bar represents the distribution of protein coding genes detected by RNAseq, whereas the red bar indicates the distribution of protein coding genes from proteomic data. (D) Scatter plot showing a comparison of absolute expression between proteins and protein-coding transcripts. (E) Distribution of coefficient of variation (CV) for all proteins.

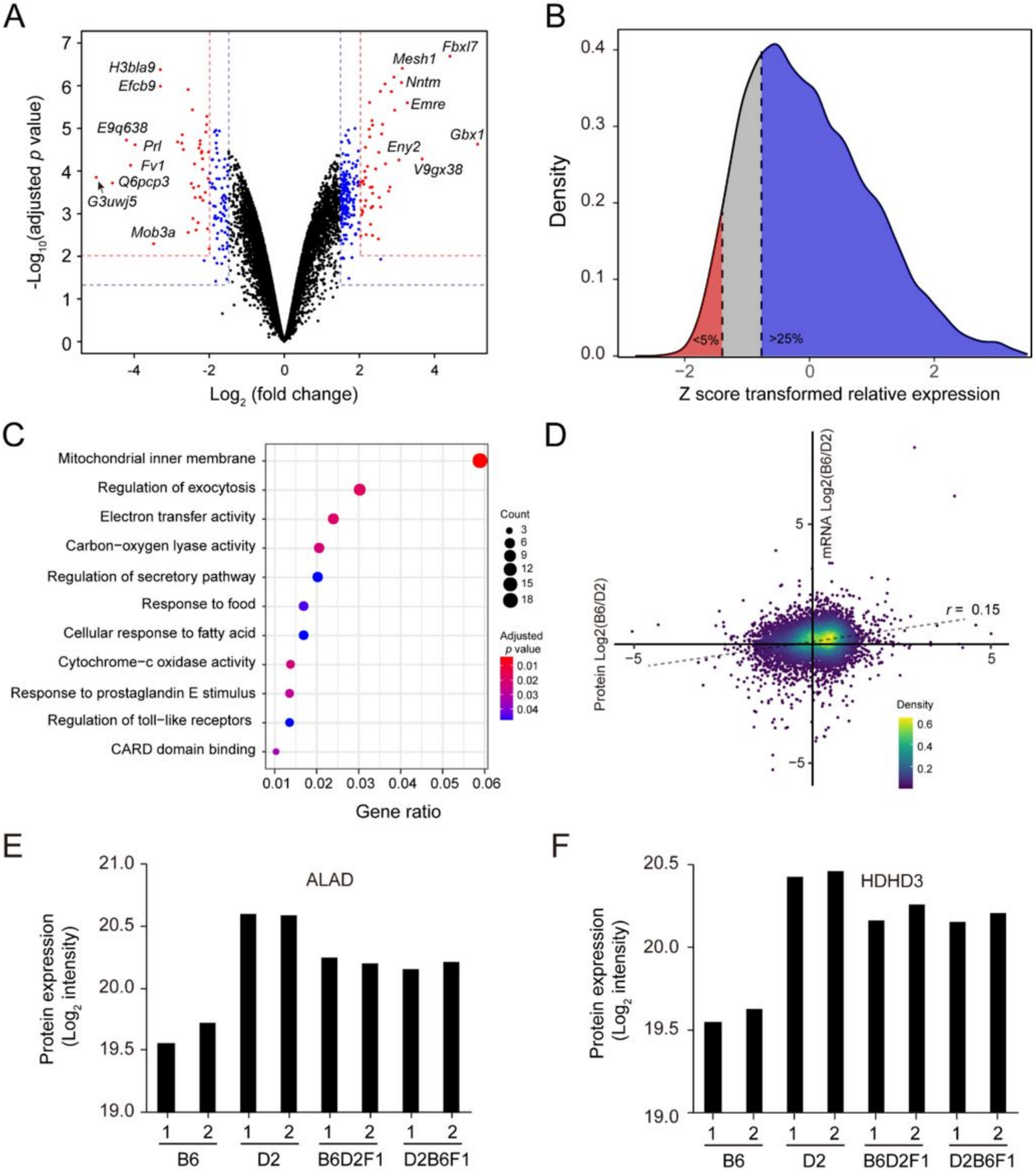

Figure 3

Analysis of differentially expressed (DE) proteins between B6 and D2. (A) The volcano plot of the differentially expressed proteins. Log2 fold change was plotted against the −log10 adjusted p-value with two criteria: (1) 4-fold change and 1% FDR; (2) 2-fold change and 5% FDR. (B) Distribution of z-score transformed relative expression between B6 and D2. An extreme form of differential expression was used to define single parent expression. (C) Enrichment analysis of DE proteins. (D) Scatter plot showing a comparison of relative expression between proteins and protein-coding transcripts. (E-F) Expression levels of ALAD and HDHD3 between B6, D2 and the two F1s.

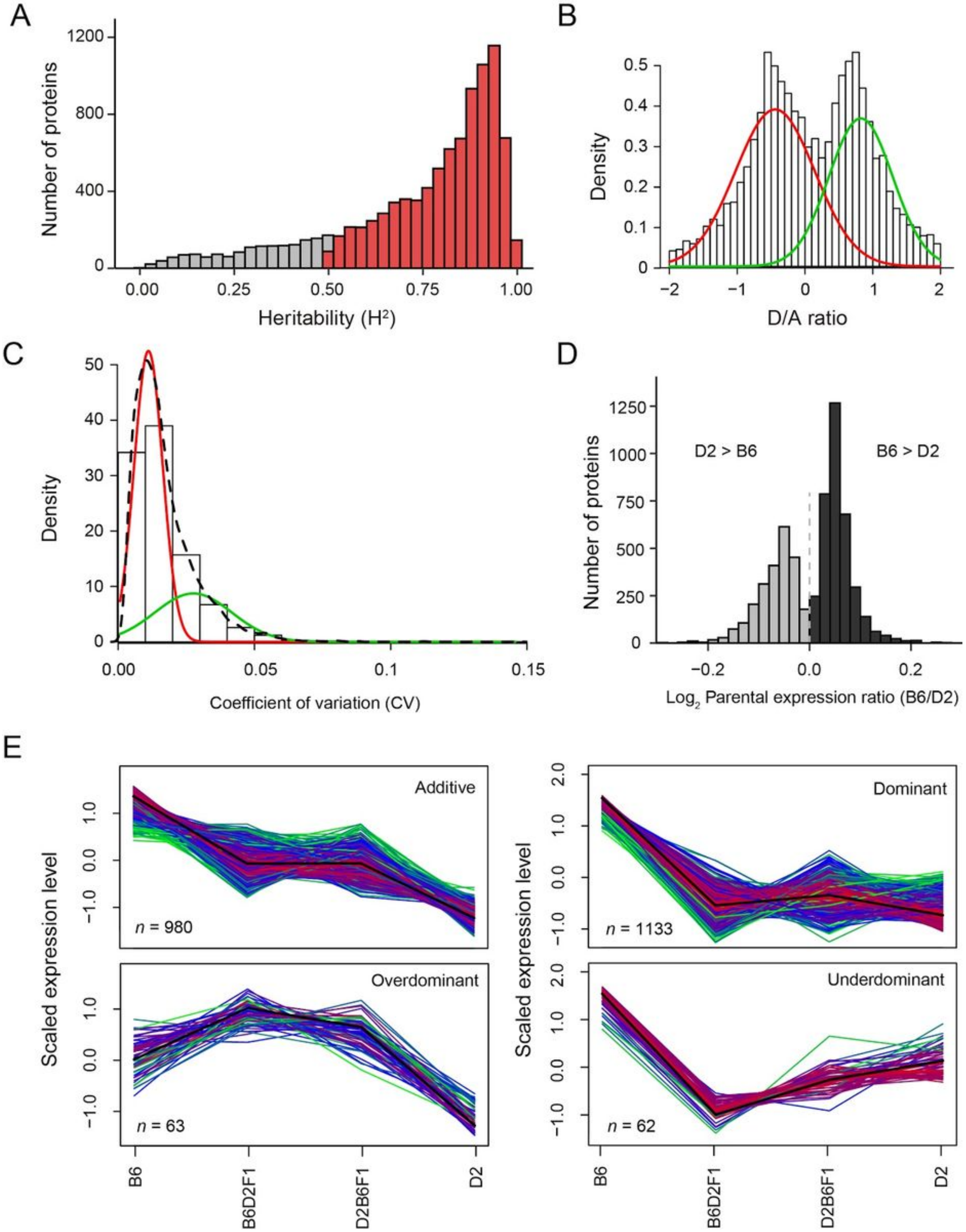

Figure 4

Inheritance patterns of protein expression. (A) Distribution of broad sense heritability (H2) of protein expression. (B) Distribution of dominance to additivity (D/A) ratio. Dominance is the difference between the observed F1 transcript abundance (in this case averaged over the two reciprocal F1 genotypes) and the midpoint of the two parents. Additivity is the absolute value of the difference between the parental means relative protein abundance. (C) Distribution of coefficient of variation (CV) and fits with two normal distributions with the EM algorithm. (D) Distribution of protein expression ratios (log2-transformed) between two strains. Negative values indicate higher expression in D2 (D2 > B6), whereas positive values indicate higher expression in B6 (B6 > D2). (E) Inheritance patterns of protein expression. Four inheritance patterns were identified: additive, dominant, over-dominant, and over-dominant. The patterns were defined by comparing the protein expression between the two reciprocal F1 hybrids and those of their parental strains.

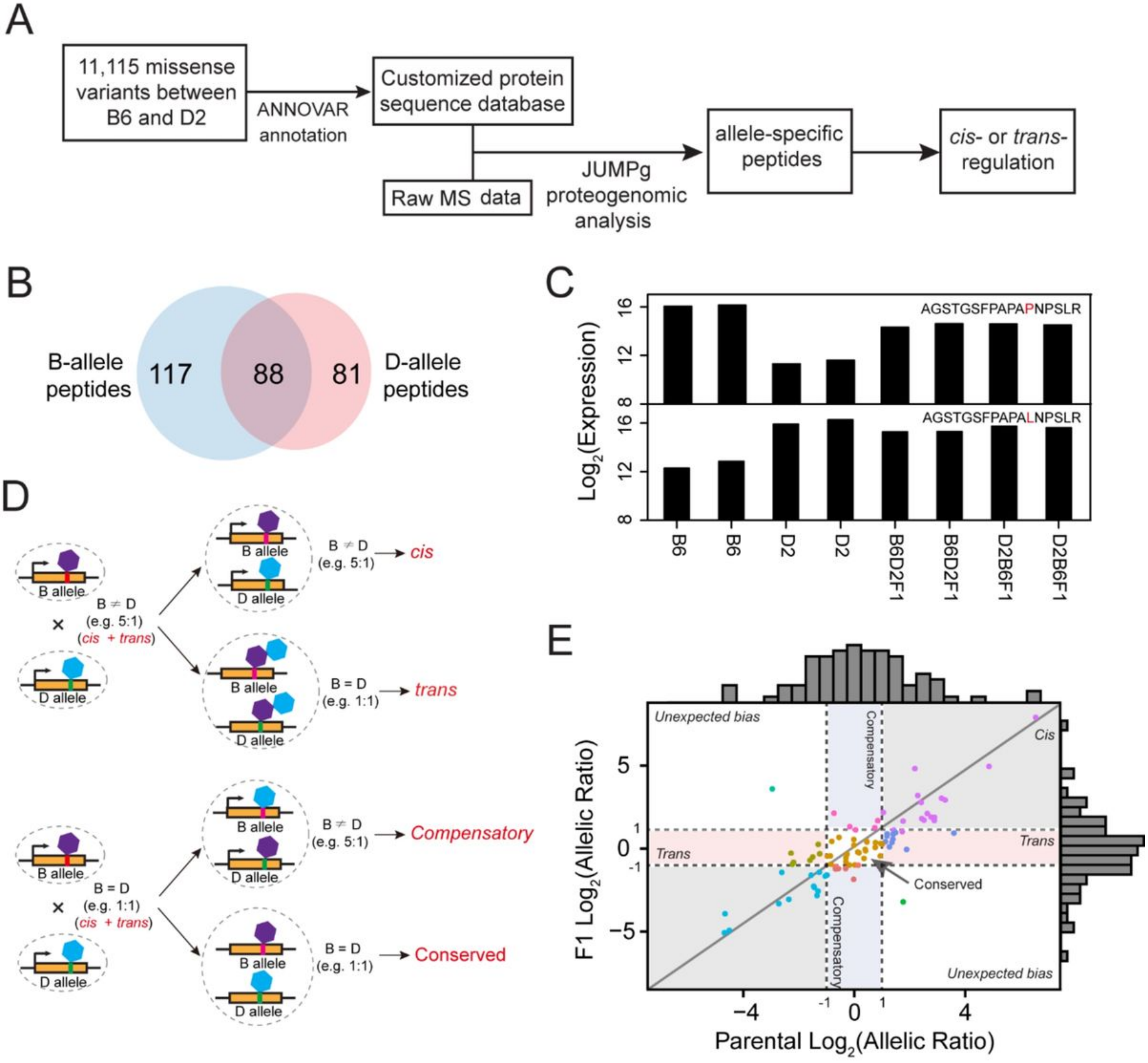

Figure 5

Protein allele-specific expression. (A) Workflow of allele-specific expression detection using the proteogenomic approach. (B) Venn diagram showing the overlap number of B-type and D-type variant peptides. (C) Two examples showing expression pattern of B-type and D-type variant peptides. (D) Conceptual diagram of cis-, trans-, compensatory, and conserved regulation. (E) Scatterplot of protein allelic ratios in parental and F1 strains showing different regulations: cis-, trans-, compensatory, and conserved.

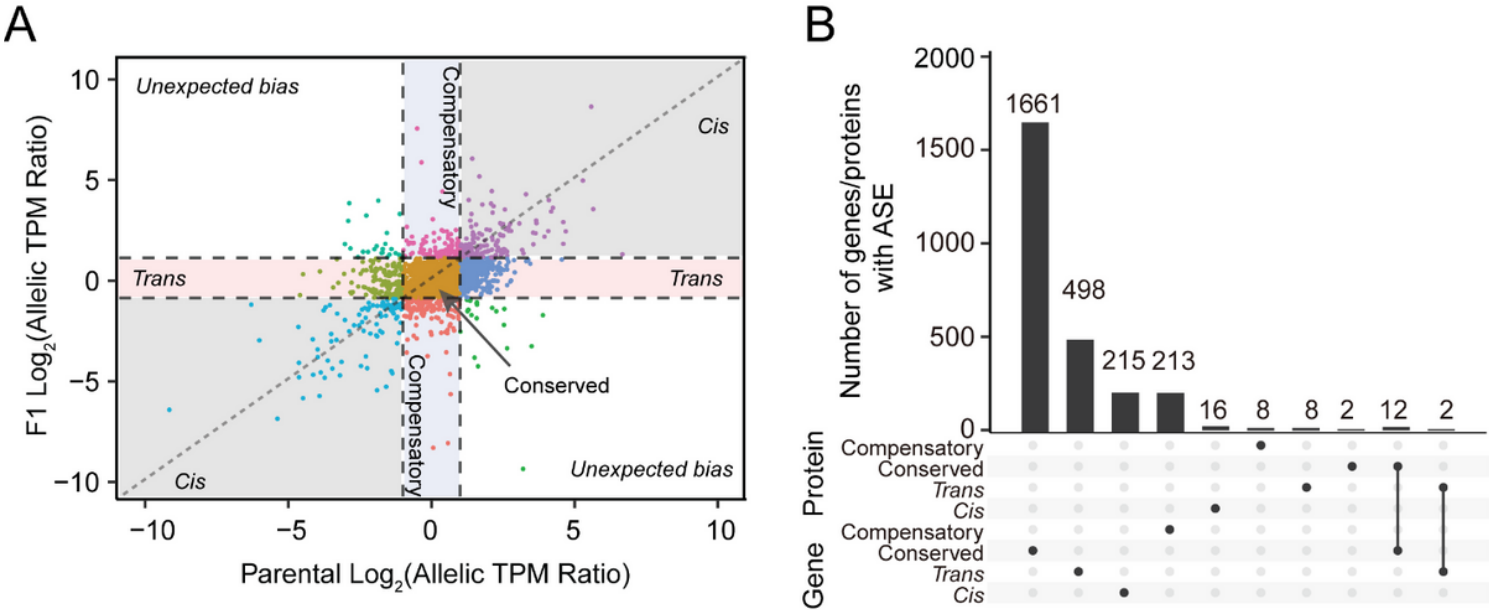

Figure 6

Comparison of allele-specific expression at transcript and protein levels. (A) Scatterplot showing regulations at the transcript level. (B) UpSet plot showing the number of different regulations between transcript and protein levels.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ARRIVEChecklistv3.pdf
- additionalfile1.xlsx
- additionalfile2.pdf