4-12-2023

Genetic dissection of the pluripotent proteome through multi-omics data integration.

Selcan Aydin

Duy T Pham

Tian Zhang

Gregory R Keele

Daniel A Skelly

See next page for additional authors

Authors

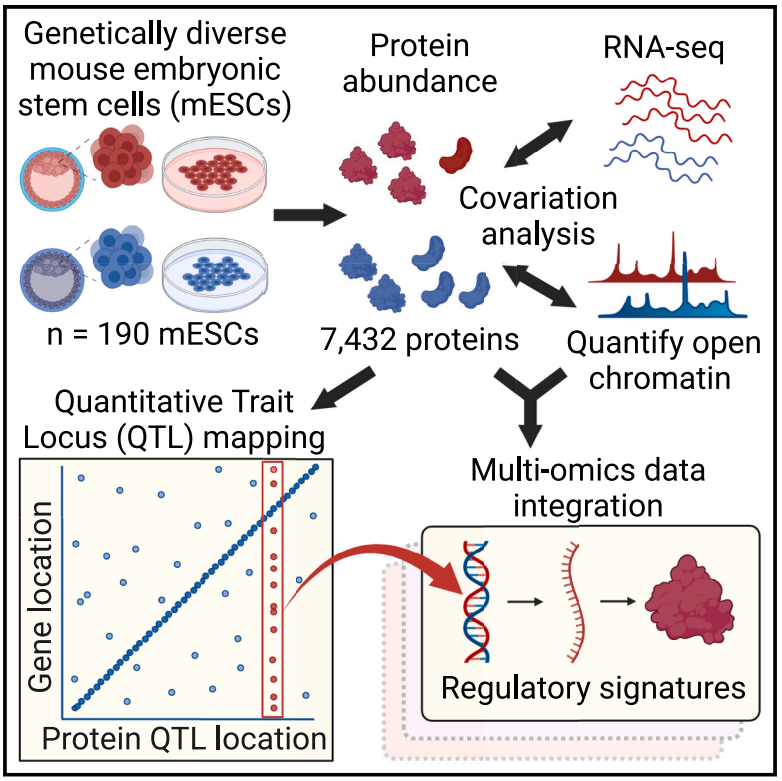
Selcan Aydin, Duy T Pham, Tian Zhang, Gregory R Keele, Daniel A Skelly, Joao A Paulo, Matthew Pankratz, Ted Choi, Steven P Gygi, Laura G Reinholdt, Christopher L. Baker, Gary Churchill, and Steven C. Munger

Cell Genomics

Article

Genetic dissection of the pluripotent proteome through multi-omics data integration

Graphical abstract



Highlights

- Proteomic profiling of 190 embryonic stem cells from genetically diverse mice
- Comparison with RNA-seq and open chromatin revealed variation unique to protein abundance
- Multi-omics data integration inferred regulatory signatures that co-vary among mESCs
- Genetic mapping identified genomic "hotspots" that drive multi-omic signatures

Authors

Selcan Aydin, Duy T. Pham, Tian Zhang, ..., Christopher L. Baker, Gary A. Churchill, Steven C. Munger

Correspondence

laura.reinholdt@jax.org (L.G.R.), christopher.baker@jax.org (C.L.B.), gary.churchill@jax.org (G.A.C.), steven.munger@jax.org (S.C.M.)

In brief

Genetic background drives phenotypic variability in pluripotent stem cells. Using a diverse panel of mouse ESCs, Aydin et al. performed quantitative proteomics and genetic mapping to identify QTL hotspots that drive variation in multi-omic regulatory signatures and numerous loci that influence the pluripotent proteome independent of the transcriptome.

Aydin et al., 2023, Cell Genomics *3*, 100283
April 12, 2023
https://doi.org/10.1016/j.xgen.2023.100283

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Genetic dissection of the pluripotent proteome through multi-omics data integration

Selcan Aydin,[1] Duy T. Pham,[1] Tian Zhang,[2] Gregory R. Keele,[1] Daniel A. Skelly,[1] Joao A. Paulo,[2] Matthew Pankratz,[3] Ted Choi,[3] Steven P. Gygi,[2] Laura G. Reinholdt,[1,4,*] Christopher L. Baker,[1,4,*] Gary A. Churchill,[1,4,*] and Steven C. Munger[1,4,5,*]

[1]The Jackson Laboratory, Bar Harbor, ME 04609, USA
[2]Harvard Medical School, Boston, MA 02115, USA
[3]Predictive Biology, Inc., Carlsbad, CA 92010, USA
[4]Graduate School of Biomedical Sciences, Tufts University, Boston, MA 02111, USA
[5]Lead contact
*Correspondence: laura.reinholdt@jax.org (L.G.R.), christopher.baker@jax.org (C.L.B.), gary.churchill@jax.org (G.A.C.), steven.munger@jax.org (S.C.M.)
https://doi.org/10.1016/j.xgen.2023.100283

SUMMARY

Genetic background drives phenotypic variability in pluripotent stem cells (PSCs). Most studies to date have used transcript abundance as the primary molecular readout of cell state in PSCs. We performed a comprehensive proteogenomics analysis of 190 genetically diverse mouse embryonic stem cell (mESC) lines. The quantitative proteome is highly variable across lines, and we identified pluripotency-associated pathways that were differentially activated in the proteomics data that were not evident in transcriptome data from the same lines. Integration of protein abundance to transcript levels and chromatin accessibility revealed broad co-variation across molecular layers as well as shared and unique drivers of quantitative variation in pluripotency-associated pathways. Quantitative trait locus (QTL) mapping localized the drivers of these multi-omic signatures to genomic hotspots. This study reveals post-transcriptional mechanisms and genetic interactions that underlie quantitative variability in the pluripotent proteome and provides a regulatory map for mESCs that can provide a basis for future mechanistic studies.

INTRODUCTION

Pluripotent stem cells (PSCs) hold great potential for regenerative medicine and modeling human disease,[2] but variation in the derivation, stability, and differentiation of individual cell lines impedes progress toward these goals.[3,4] Genetic background contributes significantly to phenotypic variation in human and mouse PSCs.[3,5] Systems genetics experiments can identify the loci that harbor genetic variants and can associate phenotypic variability with regulatory networks that are affected by these variants.[1,6–10]

Most PSC studies addressing phenotypic variability have focused on transcriptional regulation using measures of chromatin state and transcript abundance, due in part to the relatively low cost of RNA and DNA sequencing. However, cellular phenotypes are largely determined by proteins, and the effects from genetic variation on chromatin states and transcripts may be buffered, amplified, or even reversed by post-transcriptional processes acting on protein abundance.[9,11] Previous studies in cell and animal models have found a surprising level of disagreement between protein and transcript abundance[12–15]; this high discordance was also observed in differentiating mouse embryonic stem cells (mESCs[16]). Genetic analyses suggest that stoichiometric buffering of protein complex members may attenuate their individual transcriptional variation in adult mouse tissues,[11,17] and translational output was recently shown to relay back to chromatin state and transcription to drive mESC self-renewal.[18] These findings suggest that post-transcriptional regulation of protein abundance may play a significant role in pluripotency maintenance and differentiation in PSCs.

We previously derived a large panel of mESCs from Diversity Outbred (DO) mice. The outbred DO mice are derived from 8 inbred strains with high genetic diversity and a population structure optimal for genetic mapping.[1,19] We maintained mESCs in sensitized culture conditions to amplify genetic differences in the pluripotent ground state, and analyzed transcriptome and chromatin state data to map genetic loci underlying this variability.[1] We mapped thousands of quantitative trait loci (QTLs) that affected chromatin accessibility (caQTLs) and transcript abundance (eQTLs), and, of particular importance, we identified 1 locus on chromosome (Chr) 15 that was linked to the variable expression of 254 genes, many of which have known roles in pluripotency. Mediation analysis identified LIF receptor (*Lifr*) as the most likely candidate gene underlying this eQTL "hotspot," and further implicated a single causal SNP in an mESC-specific region of open chromatin ~10 kb upstream of *Lifr*. Luciferase assays and CRISPR-mediated allele swaps validated the importance of genotype at this enhancer SNP (hereafter referred to as

"*Lifr* genotype") on the expression of *Lifr*, downstream target genes, protein markers of pluripotency (NANOG), and capacity for self-renewal.[1] Finally, in a companion article, we showed that genetic background can bias differentiation propensity of mESCs through its effects on Wnt signaling activity.[20] Together, these studies demonstrated the power of this resource for discovery of genetic drivers and molecular mechanisms that underlie variation in the maintenance of the pluripotent ground state and differentiation propensity of mESCs.

In this study we expand to investigate how genetic effects on pluripotency are mediated by the proteome. We quantified proteins by multiplexed mass spectrometry across the same DO mESC lines. As with chromatin accessibility and transcript abundance, we find the quantitative proteome to be highly variable across lines. Genetic mapping identified protein QTLs (pQTLs) for 20% of all measured proteins, and one-third affect protein abundance independently from transcript levels, presumably through post-transcriptional mechanisms. These signatures of genetic effects on proteins were not detected in our earlier analysis of transcript abundance. The remaining pQTLs colocalize with previously identified eQTLs and/or caQTLs, consistent with transcriptional regulation. We applied multi-omics factor analysis (MOFA) to identify latent factors that account for the variability in gene regulatory signatures across these 3 layers of molecular data.[21,22] Genetic mapping of the latent factors identified the *Lifr* hotspot as well as novel loci. We show that multi-omics integration and dimensionality reduction with MOFA increases power to detect genetic drivers of broad regulatory signatures compared with QTL mapping of individual molecular traits. We further show how genetic variation affects transcriptional and post-transcriptional gene regulation to drive variation in ground-state pluripotency. The resulting regulatory map for mESCs can provide a rational basis for future mechanistic studies.

RESULTS

The pluripotent proteome of genetically diverse mESCs

We quantified relative protein abundance by mass spectrometry in 190 unique DO mESC lines (Figure 1A; Table S1). In total, we detected 7,432 proteins in at least half and 4,794 proteins in all the cell lines. Proteins detected in all mESCs are overrepresented for those involved in cellular metabolism, post-transcriptional processes, and protein complexes. By contrast, transmembrane proteins and transcription factors are overrepresented among the genes showing expression in the RNA sequencing (RNA-seq) data but not detected in the proteomics data (n = 5,492 out of 12,732 protein-coding genes) (Table S2). Transmembrane proteins contain both hydrophilic and hydrophobic subunits, making them less soluble[23] and therefore more difficult to isolate in untargeted proteomics analysis. The probability of detecting a given protein is linked to its transcript abundance (Figure S1A); expressed genes at the lower threshold for transcript abundance (average count = 1) have a protein detection rate of <60% (Figure 1B). This includes transcription factors, which, as a group, exhibit lower mean transcript abundance, presumably resulting in lower levels of detectable protein (Figures S1B and S1C). Proteins encoded by genes with high transcript expression—a group that includes many ribosomal and mitochondrial proteins—are detected at a much higher rate (>90%) (Figure 1B).

The mESC proteome is highly variable across cell lines (Figures S1D and S1E), and principal component analysis (PCA)[23] points to chromosomal sex as the largest driver of variance across samples (12.2%; Figure 1C). Most of this variation (principal component [PC] 1) stems from sexually dimorphic expression of X-linked proteins, likely due to 2 active X chromosomes in XX mESCs.[24,25] Over half of all proteins exhibit variable expression linked to sex (n = 4,106 out of 7,432, $p < 0.05$), including pluripotency factors SOX2, ESRRB, KLF2, KLF4, SALL4, UTF1, NR5A2, and LIN28A.[26] Next, we performed gene set variation analysis (GSVA,[27] see STAR Methods) to identify pathways that vary in their activity (expression) across lines. XX and XY mESCs vary in many cellular processes and protein complexes (Table S3); XY lines show higher activity of DNA methylation, histone modification, and chromatin remodeling pathways, consistent with previous studies[28,29] (Figure 1D). Ribosome biogenesis genes are overrepresented in weightings of PC1, with similarly higher expression in XY lines ($p < 5 \times 10^{-5}$) (Figure 1D). In addition, we find that *Lifr* genotype is associated with proteome activity for several biological processes (Table S3). For example, cell lines with at least 1 copy of the reference *Lifr* allele showed higher abundance of proteins with regulatory roles in ADP-ribosylation, the transfer of ADP-ribose moieties (derived from NAD) to protein amino acids (Figure 1D). This histone modification is catalyzed by poly-ADP-ribose polymerases, 2 of which (PARP1/ARTD1, PARP7/TIPARP) were shown in mESCs to occupy and maintain an active epigenetic state at key naive pluripotency genes including *Nanog*, *Oct4*/*Pou5f1*, *Sox2*, and *Rex1*/*Zfp42*.[30] Of note, GSVA enrichment of ADP-ribosylation proteins and several other pluripotency and differentiation pathways are observed only in the proteomics data and not evident in the transcriptome GSVA results (pathways highlighted in Table S3).

Extracellular proteins are overrepresented among the most variable proteins, while proteins that bind in complexes are among the least variable (false discovery rate [FDR] < 0.05; see STAR Methods); these differences in variance cannot be explained by differences in protein abundance levels (Figure S1F). Interestingly, targets of the transcription factor REX1, a marker of naive pluripotency,[26] are also overrepresented among the least variable proteins, despite REX1 protein itself being highly variable across mESCs (Figure S1G). REX1 is known to act as a repressor,[31] and the lowest REX1-expressing mESC lines may still exceed a threshold required to efficiently repress its target genes. Alternatively, the effects of variable REX1 protein abundance on downstream targets may be buffered. Indeed, REX1 may be dispensable for pluripotency maintenance,[32] supporting the existence of such downstream compensatory mechanisms.

Members of complexes vary less in their protein abundance overall than non-complex forming proteins (Figure S2A), and subunits of individual complexes co-vary in their abundance more than proteins not known to physically interact (Figure S2B; see STAR Methods), in line with previous studies[33] suggesting

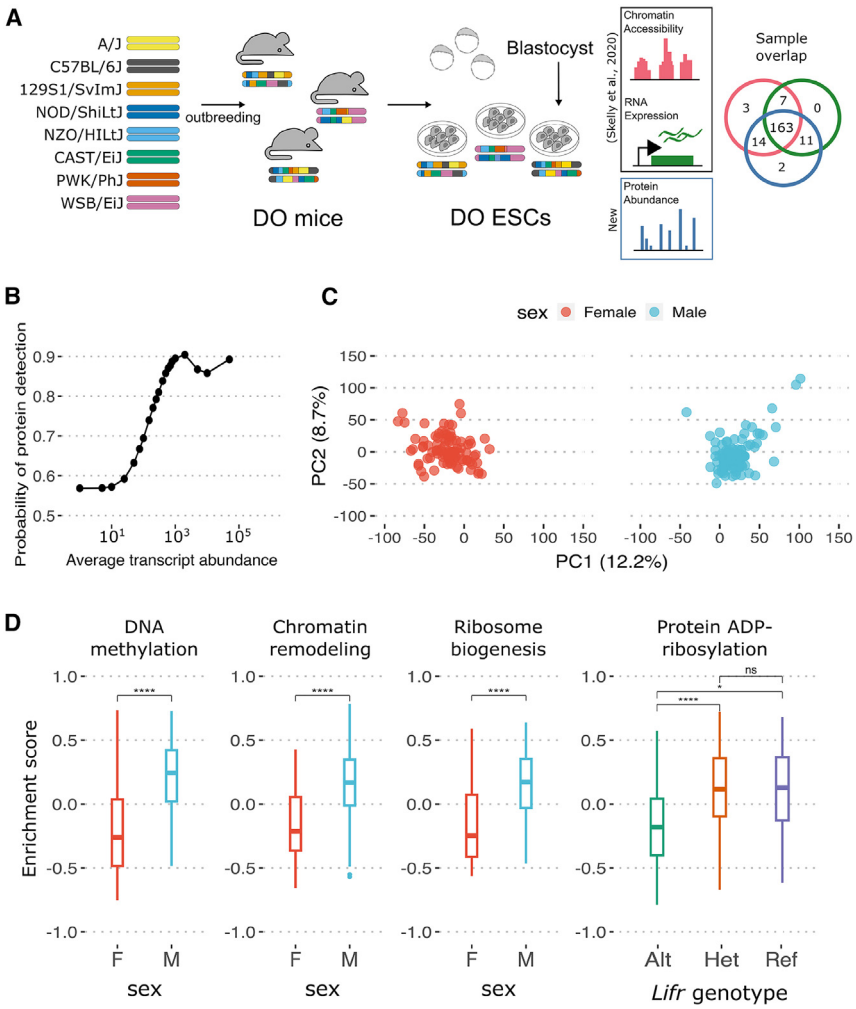


Figure 1. Overview of the quantitative proteome in Diversity Outbred mESC lines

(A) The proteomes of 190 mESCs were quantified and compared with published ATAC-seq and RNA-seq data.[1]

(B) The probability of detecting a protein by MS (plotted on y axis) is linked to the protein-encoding gene's average transcript abundance (x axis).

(C) Principal component analysis points to sex as the major source of proteome variation among mESCs. PC1 and PC2 are plotted and colored by sex.

(D) GO:BP categories including DNA methylation, chromatin remodeling, and ribosome biogenesis show significantly higher activity by GSVA in XY compared with XX lines. Protein ADP-ribosylation shows higher activity in mESCs having at least one copy of the reference *Lifr* allele (two-way ANOVA followed by Tukey's HSD, $^*p < 0.05$, $^{****}p < 5 \times 10^{-5}$). See also Figure S1 and Tables S1, S2, and S3.

Protein abundance co-varies with chromatin accessibility and transcript abundance

The pluripotent state is established and maintained by a gene regulatory cascade that orchestrates changes across multiple molecular layers from chromatin accessibility to transcript and protein abundance.[37] To better understand these multi-layered regulatory interactions and to identify proteins with potentially important roles, we looked at the co-variation of proteins with measures of chromatin accessibility (assay for transposase-accessible chromatin with sequencing [ATAC-seq]) and transcript abundance (RNA-seq) across the DO mESCs.

We first compared protein abundance (n = 7,148) with chromatin accessibility (n = 99,159 peaks) and found that many proteins were most highly correlated with chromatin in the region proximal to their protein-encoding gene (Figure S3A), consistent with our earlier observation of high concordance between transcript abundance and local open chromatin.[1] We identified 37 proteins whose abundance co-varied with chromatin accessibility at 100 or more ATAC-seq peaks genome-wide (abs(r) > 0.5; evident as horizontal bands in Figure S3A). The list includes well-characterized pluripotency regulators as well as proteins with no previously reported role in pluripotency maintenance (Table S4). For example, the abundance of ID1, a transcription factor critical in the maintenance of embryonic stem cell (ESC) self-renewal and regulation of lineage commitment,[38] co-varies significantly (both positively and negatively) with chromatin accessibility at 112 ATAC-seq peaks (Figure 3A). Other proteins with potential roles in pluripotency maintenance include AHDC1, a putative DNA-binding protein previously shown to physically interact with the transcription factor TCF7L1 (TCF3) involved in pluripotency regulation[39,40]; and UHRF2, a ubiquitin ligase identified as a target of epigenetic control during

that physical interactions among proteins may act to dampen their individual variation in abundance. We quantified co-variation between complex members and ranked complexes by their co-regulation, or "cohesiveness," of subunits. The most cohesive 10% of complexes were associated with the cell cycle, protein modification, and translation machinery, consistent with our analysis of individual proteins and published proteome studies of human and mouse cell lines and tissues[33–35] (Figure 2). Several complexes involved in protein trafficking and transcriptional regulation are similarly highly cohesive. By comparison, the least cohesive 10% of complexes are enriched for those associated with chromatin remodeling. Sex differences in complex cohesiveness are also observed; for example, protein constituents of the cytoplasmic small and large ribosomal subunits and mitochondrial small ribosomal subunit are more cohesive in XY than XX, while HOPS complex members are significantly more cohesive in XX than XY mESCs ($p < 5 \times 10^{-4}$) (Figure S2C). While the molecular basis of these sex differences in protein complex cohesiveness remains to be established, they are also observed in adult mouse liver[17] and heart proteomes[36] and are therefore unlikely to play a unique regulatory role in pluripotency.

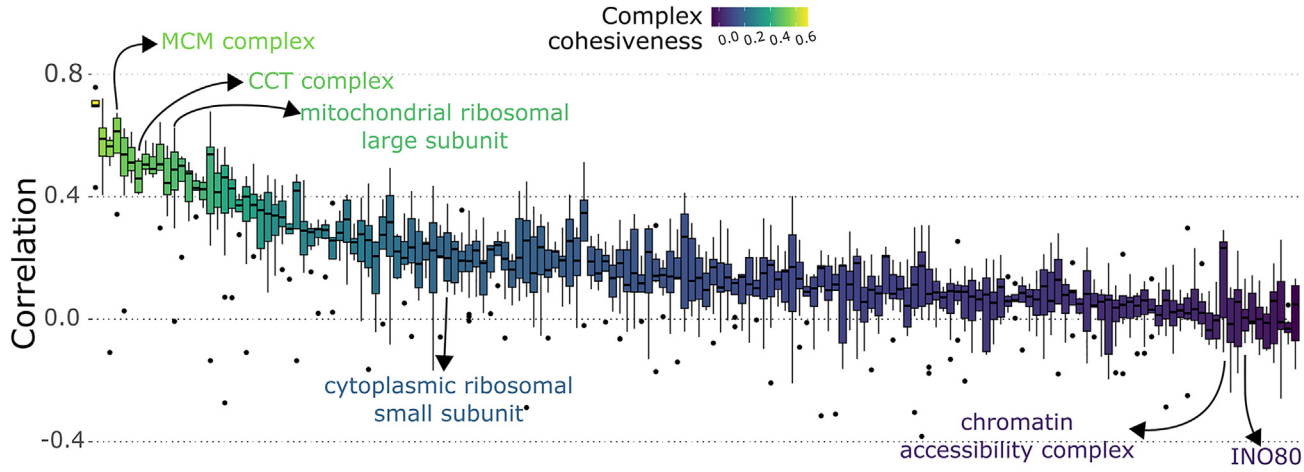


Figure 2. Subunit cohesiveness varies considerably among 164 protein complexes

For each complex, pairwise correlations between all subunits were calculated and summarized as a boxplot. Boxplots are ordered and colored based on their median pairwise correlation, with more cohesive complexes on the left. Specific examples are highlighted. See also Figure S2.

self-renewal.[41] For almost half of these proteins, we find that their covarying ATAC-seq peaks are overrepresented in binding sites active in ESCs for TRP53 (n = 28) and naive pluripotency factors NANOG, ESRRB, and PRDM14 (n = 19, 18, 16, respectively at FDR < 0.05).[26] Many of these covarying chromatin peaks are proximal to genes involved in cellular response to leukemia inhibitory factor (LIF), providing further evidence for their importance in establishing and/or maintaining pluripotency (FDR < 0.05). Notably, only 6 of the 37 proteins also co-vary with chromatin accessibility at the level of their transcript abundance; for the other 29, correlations to chromatin are observed only at the level of protein abundance, consistent with post-transcriptional regulation of these chromatin modifying proteins (Table S4). These data suggest 2 possibilities: these proteins may play active roles in chromatin remodeling and directly influence ground-state pluripotency, or, alternatively, they may serve as downstream quantitative protein biomarkers of upstream activities of known pluripotency regulators. The presence of known pluripotency regulators among this list supports the former scenario, while the relative lack of annotated TFs/chromatin remodelers suggests the latter. Future experiments will be required to validate direct or causal roles for these novel proteins.

We next examined the concordance between protein and transcript abundance in DO mESCs. For genes where we detect both (n = 7,241), protein and transcript abundance are broadly positively correlated in their magnitude and variance (r = 0.5, $p < 2.2 \times 10^{-16}$, see STAR Methods; Figures S3B and S3C). Similar studies in human iPSCs found that many proteins that varied in abundance did not show variation in their cognate RNAs.[9] We see a similar trend for a small number of proteins (n = 180) where protein abundance is highly variable across cell lines without similar variation at the transcript level. Conversely, genes exhibiting high variation in transcript abundance but lacking variation at the protein level (n = 111) are overrepresented for ribosomal proteins. Surprisingly, the overall agreement between protein and transcript levels within a cell line appears to vary considerably across the mESCs (r range 0.1–0.6) (Figure 3B). We ruled out sample mix-ups as a potential reason for the low concordance in some cell lines (Figure 3B), and even the lowest observed sample correlation is still well above the null distribution of correlation values from permuted sample assignments (Figure S3D). Looking at individual genes, we see a wide range of variation in the correlation between protein and transcript levels across mESC lines, where many are highly positively correlated while others are negatively correlated (Figure 3C). The larger group of genes showing positive transcript-protein correlation (n = 5,530, $r > 0.16$, $p < 0.05$) is enriched for proteins involved in X-linked inheritance, lipid metabolism, and membrane proteins (Figure 3C). The smaller group of genes with significantly negatively correlated transcript and protein levels (n = 82, $r < -0.16$, $p < 0.05$) are enriched for those with roles in cellular respiration and mitochondrial translation. Genes involved in mRNA splicing and cytoplasmic translation are enriched among those exhibiting low correlation in their transcript and protein levels (abs(r) < 0.05, n = 498). Genes that are not known to form protein complexes show stronger positive correlation in transcript and protein abundance (Figure S3E), further supporting the idea that complexes place physical constraints on protein abundance that can serve to buffer against transcriptional variation.[11,17]

Genetic characterization of the pluripotent proteome

Variation in protein abundance across DO mESC lines appears to be driven by genetic background, with more than 90% of measured proteins estimated to have non-zero heritability (median h^2 = 0.25). To identify genomic loci underlying this quantitative variation, we performed protein quantitative trait locus (pQTL) mapping (see STAR Methods; refer to Table S5 for a list of all significant pQTLs). We detected pQTLs for over 20% of expressed proteins (n = 1,555 out of 7,432), with a total of 1,677 pQTLs (LOD > 7.5, permutation $p < 0.05$, FDR = 0.058) (Figure 4A). Of these, nearly two-thirds (n = 1,056) are local pQTLs and map to within ±10 Mb of the midpoint of the corresponding gene. We found many fewer distant pQTLs (n = 621) that map

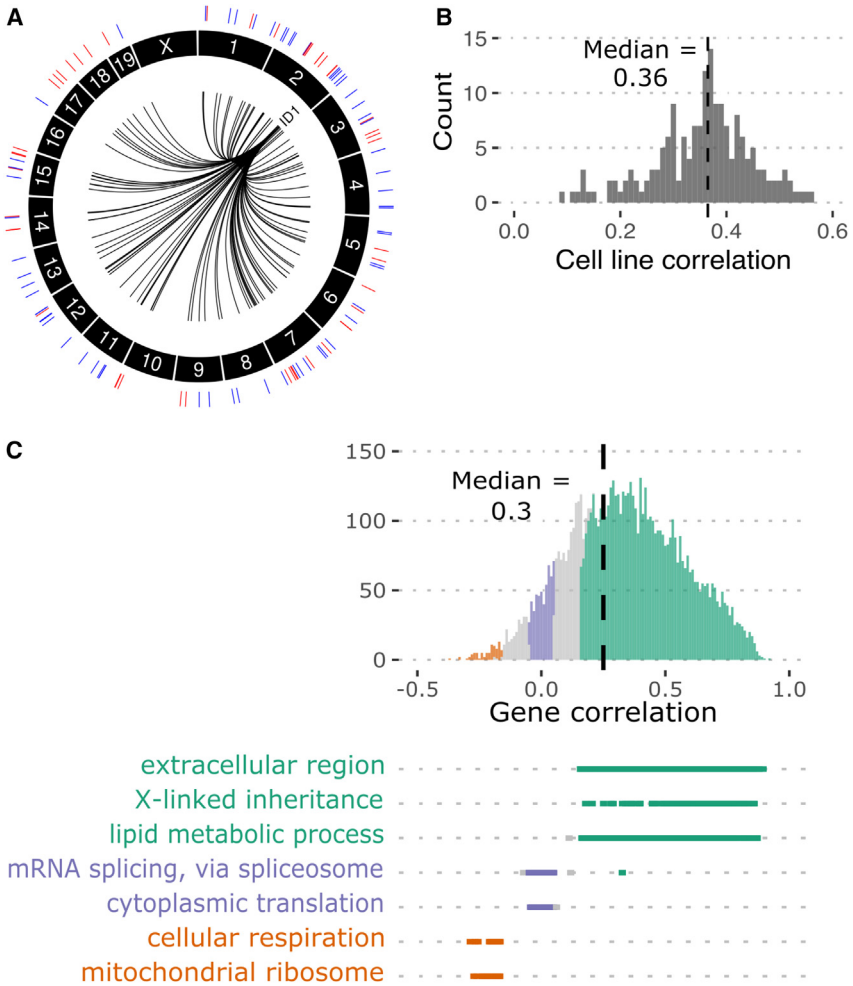


Figure 3. The quantitative proteome co-varies with chromatin accessibility and the transcriptome

(A) ID1 protein abundance is highly correlated with many regions of open chromatin genome wide. Circos plot showing ATAC-seq peaks where chromatin accessibility is positively (red) and negatively (blue) correlated with ID1 abundance; n = 112, abs(correlation) > 0.5.

(B) Overall agreement between the transcriptome and proteome within an mESC line is widely variable across lines, as shown by a histogram of sample-level Pearson correlations (n = 174).

(C) Agreement in transcript and protein abundance for a given gene also varies widely across lines. Histogram depicting the distribution of pairwise correlation coefficients between transcript and protein abundance of genes, with overrepresented GO terms annotated below in matching colors (green, positively correlated; orange, negatively correlated; purple, genes with little or no correlation). See also Figure S3 and Table S4.

outside of the local genomic window. As with previous pQTL studies of similar size in DO mice,[11,36] local pQTLs tend to be more significant and more reproducible than distant pQTLs (local median LOD = 10.8; distant median LOD = 7.9), and, for over 80% of genes that have a local pQTL, we also detected an eQTL for the cognate transcript. For most of these local eQTL-pQTL pairs, the founder strain allele effects at the peak SNP are highly correlated (75% of local pairs are significant at FDR < 0.05; median r = 0.9), consistent with a single causal variant affecting both transcript and protein abundance (Figure 4B). Correlation between chromatin accessibility (caQTLs) and co-mapping local pQTLs is more variable, with some proximal caQTLs showing strong correlation of allele effects and others showing little or even negative correlation to local pQTLs. For example, a chromatin region within the promoter of *Bspry*, a gene linked to pluripotency in mESCs and early embryonic development,[42] has a local caQTL with highly concordant founder allele effects on *Bspry* transcript and protein abundance (Figure 4C, top). Anti-correlated local caQTLs include a variable region in the promoter of the gene *Tfcp2l1*, which encodes a transcription factor that has critical roles in maintenance of naive pluripotency.[43,44] The founder allele effects at this caQTL are nearly opposite to those for the *Tfcp2l1* local eQTL and pQTL (r = −0.8 for both caQTL-eQTL and caQTL-pQTL pairs) (Figure 4C, bottom). Both strongly positively and negatively correlated local effects may implicate a single causal variant but with different molecular mechanisms—e.g., a promoter variant bound by a repressor could explain the anti-correlated caQTLs and pQTLs for *Tfcp2l1*—whereas uncorrelated founder effects suggest multiple causal variants with independent effects on chromatin and transcript/protein abundance.

Local pQTLs likely stem from *cis*-regulatory or nonsynonymous coding variants, whereas distant pQTLs reflect *trans* effects that are likely mediated through another protein. Distant pQTLs are not uniformly distributed across the genome and co-localize to hotspots, as we previously observed for caQTLs and eQTLs in DO mESCs.[1] We identified 3 pQTL hotspots on chromosomes (Chrs) 4, 9, and 15 (Figure 4D). Two of these were previously mapped as caQTLs and/or eQTLs (Chrs 4, 15),[1] while the Chr 9 hotspot uniquely affects protein levels (Table S6). The identity of the causal gene underlying the Chr 9 pQTL remains to be established, but targets of this pQTL-specific hotspot are enriched for proteins involved in translation initiation. This hotspot has not been detected in pQTL analyses of adult DO tissues and may point to a post-transcriptional regulatory mechanism that is unique to pluripotent mESCs. By contrast, we previously discovered a caQTL-eQTL hotspot on Chr 15 with shared transcriptional effects on hundreds of transcripts and chromatin peaks; the Chr 15 pQTL hotspot maps to the same region and exhibits similar properties. Indeed, we observe the same founder allele effects and we identified *Lifr* transcript as the top candidate mediator for most pQTLs that map to this locus (see STAR Methods; Figures S4A and S4B), consistent with previous

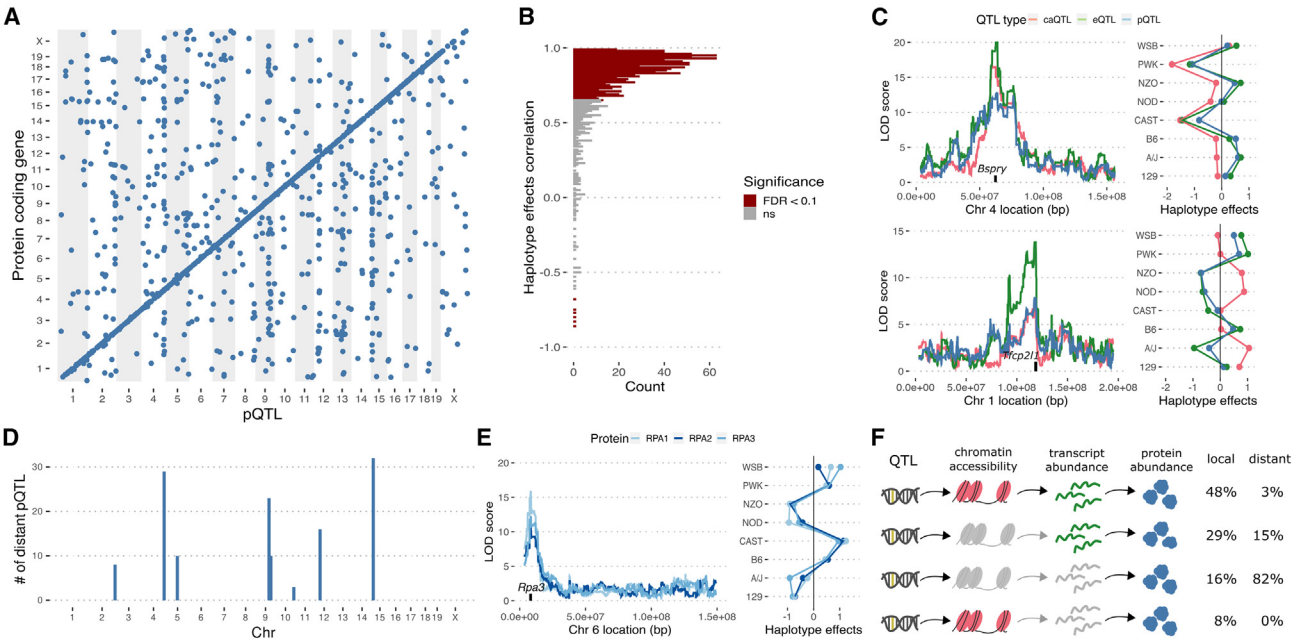


Figure 4. Genetic characterization of the pluripotent proteome

(A) Genetic mapping identifies 1,677 significant pQTLs including 1,056 local (diagonal line) and 621 distant loci. The location of the pQTL is plotted on the x axis against the midpoint of the protein-coding gene on the y axis.

(B) Most co-mapping eQTLs and pQTLs show high agreement in their haplotype effects. Histogram of pairwise correlation coefficients between inferred allele effects from eQTL and pQTL scans for all genes with co-mapping QTLs. Bars are colored by significance of the correlation.

(C) Examples of local pQTLs where the influence of genetic variation is seen at all three molecular layers. Left: LOD scores obtained from caQTL, eQTL, and pQTL scans for the target gene are plotted for the peak chromosome, with the target gene's location annotated on the x axis. Right: haplotype effects inferred at the caQTL, eQTL, and pQTL peaks are shown.

(D) Histogram showing that many distant pQTLs localize to specific genomic hotspots.

(E) The effect of one local pQTL is propagated across all protein subunits of the replication complex. Left: RPA1, RPA2, and RPA3 LOD scores are plotted for Chr 6 (x axis) and show a shared pQTL peak at the location of the *Rpa3* gene. Right: the inferred allele effects at the peak for all three proteins show high concordance.

(F) Graphical overview of the different classes of pQTLs based on their effects on one or more molecular layers. Layers lacking impact (no QTLs with LOD > 5 and matching allele effects) are depicted in gray.

See also Figure S4 and Tables S5 and S6.

findings for caQTLs and eQTLs.[1] We were unable to detect LIFR protein in our mass spectrometry data, likely because it is a transmembrane protein with low solubility.[45] Among the 32 significant pQTLs at this hotspot, 14 are found only for proteins, including TCF7L1, a regulator of exit from pluripotency.[46] These unique pQTLs could reflect post-transcriptional effects from LIFR; however, we find it more likely that transcript abundance for these genes is similarly affected by variation in *Lifr* expression but the eQTL failed to reach statistical significance. Likewise, of the 107 protein-coding genes with significant Chr 15 eQTLs we identified previously, only 9 are detected here as significant pQTLs. Again, many of these are likely false negatives due to the stringent genome-wide detection threshold. Finally, we treated our protein GSVA sample enrichment scores as quantitative traits for mapping and find that the protein ADP-ribosylation pathway maps with a near-significant QTL on proximal Chr 15 (LOD = 7.4, FDR = 0.06) that is best mediated by *Lifr* transcript abundance (Figure S4C), explaining its correlation to *Lifr* genotype in Figure 1D.

A growing body of evidence supports the idea that physical interactions among proteins can propagate or buffer the effects of transcriptional variation on protein abundance.[9,11,17] This "stoichiometric buffering" significantly affects proteins that bind in stable complexes and likely accounts for their increased co-variation and lower heritability in DO mESCs. Indeed, we map fewer pQTLs overall for protein complex members, consistent with previous reports.[17] We find abundant evidence for stoichiometric buffering of protein complexes, for example in ribosomal and chromatin remodeling complexes where subunits vary little in their protein abundance—and consequently do not map with any pQTLs—despite varying considerably in their transcript abundance and mapping with many significant eQTLs. In addition, we observe complexes that vary extensively across DO mESCs and whose subunits share a significant pQTL. In these cases, local genetic variation affecting a single subunit appears to propagate to other members of the complex. The replication complex provides such an example, where the subunits RPA1, RPA2, and RPA3 all map with a pQTL on Chr 6 and have concordant founder allele effects (Figure 4E). The *Rpa3* gene is located nearby, and *Rpa3* transcript levels are affected by a local eQTL that exhibits the same founder allele effects, suggesting that the causal variant acts in *cis* and influences transcript abundance

of *Rpa3* and protein abundance of all 3 subunits. Indeed, mediation analysis identifies RPA3 protein abundance as the best mediator of the RPA1 and RPA2 pQTLs. Rather than RPA3 being an active regulator of RPA1 and RPA2, though, the local variant likely decreases RPA3 expression and causes it to be the limiting subunit of the stable replication complex, with any unbound RPA1 and RPA2 proteins likely being degraded.

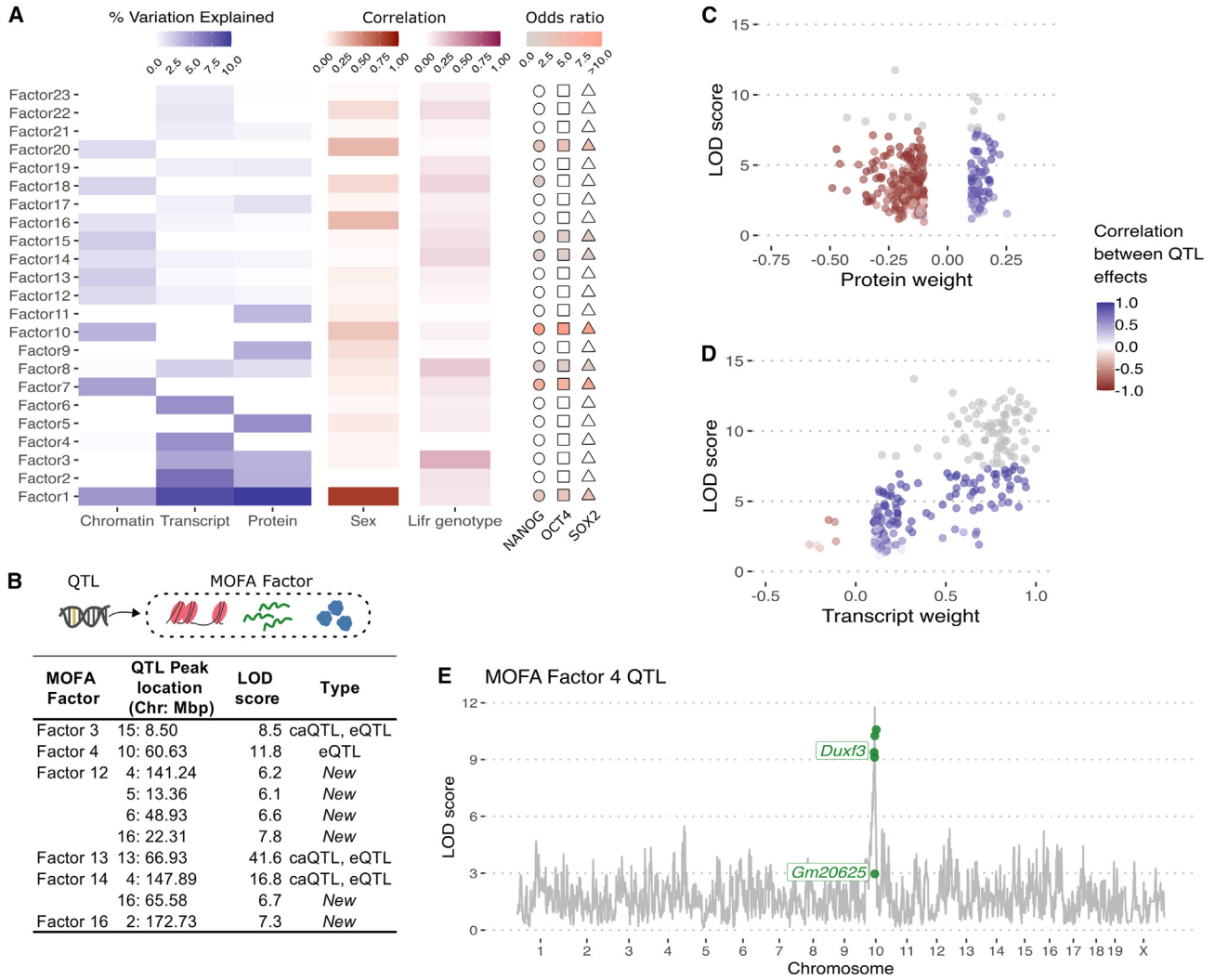
Protein QTLs can be broadly classified by their genomic location and whether they are most likely to affect transcriptional or post-transcriptional processes (Figure 4F). Most local pQTLs (84%, n = 851 out of 1,056) appear to stem from transcriptional variants acting in *cis* to affect local chromatin accessibility and/or transcript abundance of the protein-encoding gene; 48% (n = 483 out of 1,056) show similar genetic effects with local eQTLs and caQTLs, 8% (n = 80 out of 1,056) share a similar local caQTL, and 16% (n = 288 out of 1,056) share a similar local eQTL. In stark contrast, distant pQTLs primarily affect protein abundance without influencing transcript abundance or chromatin accessibility (82%, n = 476 out of 621), and mediation analysis suggests these unique *trans* effects on protein abundance can stem from physical interactions between binding partners and complex members. In summary, these data demonstrate that the high variability in the proteome observed across DO mESCs is highly heritable, genetic variants driving protein-level differences are numerous and widespread throughout the genome, and the genomic location of a pQTL relative to its target protein is predictive of its regulatory effects, with most local pQTLs influencing transcriptional processes, while most distant pQTLs confer post-transcriptional effects. These protein-specific effects of distant pQTLs highlight the importance of post-transcriptional regulation and physical interactions among proteins to the quantitative proteome in mESCs.

Integration of the proteome with the chromatin landscape and transcriptome reveals signatures spanning multiple layers of biological regulation

The extensive co-variation observed within and among the mESC proteome, transcriptome, and chromatin accessibility, along with numerous shared QTLs that appear to affect more than 1 of these regulatory layers, suggest the presence of 1 or more overarching regulatory signatures that co-vary among the genetically diverse DO mESC lines. To characterize these sources of variation more fully, we applied MOFA[21,22] to integrate and map our three genomic datasets onto a smaller set of latent factors—akin to principal components—that explain a significant proportion of the variation across mESC lines (see STAR Methods). For this analysis, we included a subset of the 15,000 most variable regions of open chromatin along with the complete sets of expressed transcripts (n = 14,405) and proteins (n = 7,432). We identified 23 latent factors that capture variation within and across the multi-omics data (Figure 5A; Table S7). Several of the latent factors correlate with biological variables that we previously identified as major drivers of variation, including chromosomal sex (factors 1, 10, 16, 18, 20; FDR < 0.05) and genotype at the *Lifr* locus (factors 3, 8, 14, 18, 22[1]). Factors differ in the degree of variation they explain both within and across datasets, and 7 factors capture variability spanning at least 2 or more layers of genomic data. For example, factor 4 captures 5.4% of the observed variation in transcript abundance but also explains 0.33% of variation in chromatin accessibility (Figure 5A). Factor 4 combines information across hundreds of transcripts with thousands of chromatin sites (Figure S5). Other factors capture variation across all 3 layers; e.g., factor 14 explains a small amount of variation for thousands of chromatin peaks (1.7%), transcripts (0.8%), and proteins (0.6%) (Figures 5A and S5). In all, the 23 MOFA factors explain 27%, 41%, and 36% of the variation in chromatin accessibility, transcript, and protein abundance, respectively.

We further dissected the regulatory signatures captured by each MOFA factor through functional annotation of their molecular drivers. This included enrichment of biological processes and pathways among protein and transcript drivers ranked by factor weights, and overrepresentation of transcription factor binding sites in the genomic sequences underlying chromatin peaks. Significantly, for 7 of the 23 factors, we find overrepresentation of binding sites associated with the core pluripotency factors NANOG, SOX2, and OCT4 in the sequences of their top ATAC-seq peak drivers (Figure 5A). For 3 factors, including factor 3, we find enrichment for genes involved in the regulation of pluripotency maintenance, such as response to LIF. Together, this functional evidence shows that MOFA factors are capturing variation across the molecular datasets that is relevant to pluripotency maintenance.

All but 1 of the 23 MOFA factors have a non-zero heritability (median $h^2 = 0.5$), indicating a strong genetic contribution to their observed variability across mESCs. To identify genetic loci driving these MOFA factors, we treated each factor as a quantitative trait and performed QTL mapping and mediation analysis (Figure 5B). Multiple published studies over the past decade applied dimensionality reduction techniques and QTL mapping to individual genomic (transcriptomic) datasets; e.g., mapping modifiers of module eigengenes derived from WGCNA.[47] We mapped 10 significant QTLs for 6 MOFA factors (Figure 5B). Five of these QTLs colocalize with molecular QTL hotspots described above, including factor 3, which mapped to the *Lifr* locus[1] (Figure 5A). MOFA[21,22] identified additional transcripts and proteins that individually did not have significant association with the Chr 15 QTL but were significant contributors to factor 3. Examination of their individual eQTLs and pQTLs showed evidence for sub-threshold genetic association and allele effects that are consistent with regulation by the *Lifr* locus (Figure 5C). MOFA factor 4, which captures a large amount of variation in transcript abundance, mapped to the eQTL hotspot on Chr 10. Genes mapping to this QTL include those that are upregulated in the rare two-cell-like cell (2CLC) state and are predicted to be regulated by *Duxf3*[1]. Based on their contribution to factor 4 and shared genetic effects at the locus, we identified additional target genes known to be upregulated in the 2CLC state (n = 13) including *Zscan4e* and *Tcstv1* that individually lack significant QTLs (Figure 5D).[48] Mediation analysis identifies *Gm20625* transcript abundance and not *Duxf3* as the best candidate regulator for this MOFA factor QTL on Chr 10 (Figure 5E). Further, we identified 2 single-nucleotide variants near *Gm20625* (rs49316493, rs265937729) that reside in annotated regulatory regions active in ESCs and that both exhibit a founder strain genotype pattern matching the observed genetic effects at



MOFA Factor	QTL Peak location (Chr: Mbp)	LOD score	Type
Factor 3	15: 8.50	8.5	caQTL, eQTL
Factor 4	10: 60.63	11.8	eQTL
Factor 12	4: 141.24	6.2	*New*
	5: 13.36	6.1	*New*
	6: 48.93	6.6	*New*
	16: 22.31	7.8	*New*
Factor 13	13: 66.93	41.6	caQTL, eQTL
Factor 14	4: 147.89	16.8	caQTL, eQTL
	16: 65.58	6.7	*New*
Factor 16	2: 172.73	7.3	*New*

Figure 5. MOFA reveals broad regulatory signatures that encompass multiple layers of data

(A) MOFA yielded 23 latent factors that capture variation in one or more layers of genomic data. For each factor, percentage of variation explained in chromatin accessibility, transcript abundance, and protein abundance is displayed as a heatmap, as is the correlation of each factor to sample covariates including sex and *Lifr* genotype. On the right, a heatmap indicates overrepresentation of pluripotency regulator binding sites (NANOG, OCT4 [*Pou5f1*], and SOX2) among the top chromatin drivers of each factor.

(B) Above: Depiction of QTL mapping with MOFA factors to identify genetic modifiers of shared molecular variation. Below: Table of QTL peaks for MOFA factors. Loci previously identified as QTL hotspots are denoted in the "Type" column.

(C) For all proteins, the LOD score calculated at the Chr 15 pQTL peak is plotted (y axis) relative to the protein's contribution (factor weight) to MOFA factor 3 (x axis). Proteins with absolute factor weights <0.1 were filtered. For each protein, color corresponds to the correlation between allele effects at the Chr 15 pQTL and the factor 3 QTL. Individual proteins that mapped with a significant pQTL are colored gray, and highlight that many proteins contribute substantially to factor 3 and show high agreement in allele effects at the Chr 15 peak (dark red and blue), despite individually not mapping with a significant pQTL.

(D) For each expressed transcript, LOD score at the Chr 10 eQTL peak is plotted (y axis) relative to that transcript's contribution to factor 4 (x axis). Transcripts with absolute factor weights <0.1 were filtered, and points are colored as described in (C). Many transcripts contribute to factor 4 and have correlated allele effects at the Chr 10 QTL, despite failing to map individually with a significant Chr 10 eQTL.

(E) Genome-wide LOD scores obtained from the factor 4 QTL scan are plotted with mediation results overlaid. *Duxf3* expression was previously identified as a strong candidate mediator for the eQTL hotspot in this region[1] but performs poorly as a mediator of the factor 4 QTL compared with *Gm20625*. Both genes are highlighted in green next to their corresponding LOD score drop.

See also Figure S5 and Table S7.

the QTL. These data implicate the predicted lncRNA *Gm20625* as potentially playing a regulatory role in the transition between the mESC and 2CLC states. Finally, we mapped novel QTLs for 2 of the MOFA factors (Figure 5B), including a QTL for factor 14 on Chr 16 that influences hundreds of features across all 3 molecular layers. Mediation analysis fails to identify strong transcript or protein candidates at these novel loci, perhaps suggesting that 1 or more may be due to causal variants that affect the

structure or function of the regulatory protein (e.g., missense variant) rather than its abundance in mESCs. Altogether, these examples highlight the power of multi-omics data integration and factor analysis to reveal higher-order regulatory signatures, identify additional genes as targets (factor 3) and mediators (factor 4) of previously mapped QTL hotspots, and discover novel loci that influence variation across all 3 molecular layers (factors 12 and 14).

DISCUSSION

We carried out a comprehensive genetic characterization of the pluripotent proteome in 190 genetically diverse DO mESC lines. Our data reveal that the proteome is highly variable across lines, and genetic background and sex are major drivers of this variation. We previously identified significant sex differences in gene expression stemming largely from X chromosome dosage,[1] and here we find that these differences are carried through to protein abundance.[28,29] GSVA identified multiple pluripotency and differentiation pathways that vary in activity across mESCs, including tRNA modification,[49] regulation of histone acetylation,[50] intermediate filament organization,[51] glutathione biosynthesis,[52–54] Golgi vesicle transport,[55] hippo signaling,[56,57] and JUN kinase activation[58] (Table S3). Of note, variation in these pathways is uniquely observed in the proteomics data.

Protein abundance is highly heritable, and we mapped pQTLs for more than 20% of all detected proteins. Most pQTLs map close to the protein-encoding gene (local pQTLs) and are also detected with concordant allele effects for gene transcript abundance and/or local chromatin accessibility. We found 680 protein-coding genes with significant local eQTLs but not corresponding local pQTLs. This discordance may stem in part from limitations in mapping power; however, it may also indicate buffering of protein levels against transcriptional variation. Indeed, post-transcriptional regulation is most evident among the 621 distant pQTLs, very few of which have corresponding distant eQTLs. We found evidence for stoichiometric buffering among the members of a number of complexes, including the replication complex, where genetic variation influencing 1 subunit (RPA3) is propagated to other members (RPA1, RPA2). More broadly, our observations of high variability in the quantitative proteome across lines and modest correlation between transcriptome and proteome within lines do not appear to be unique to mESCs. Similarly high discordance between the transcriptome and proteome was observed in a recent study of 217 human iPSC lines,[9] and the authors pointed to stoichiometry as a likely mechanism to propagate local genetic effects on a single subunit to other members of a complex. These unique distant pQTLs reveal post-transcriptional genetic interactions that are not detectable in transcriptome data, emphasizing recent findings of the importance of post-transcriptional regulation in pluripotency maintenance.[59]

Comparison of protein abundance with our earlier genetic study of transcript abundance and chromatin accessibility[1] revealed extensive co-variation across molecular layers. We utilized MOFA[21,22] to integrate the proteomics data with chromatin accessibility and transcript abundance to explore this co-variation more thoroughly. MOFA is a logical extension of common dimensionality reduction techniques such as WGCNA that identify co-expressed modules of genes in transcriptome data.[60] Characterization of the 23 MOFA factors revealed shared variation in gene regulatory signatures influencing pluripotency maintenance and correlated with chromosomal sex and *Lifr* genotype. Genetic mapping and mediation of the MOFA factors identified candidate regulatory genes underlying these multi-omics signatures. We mapped QTLs for MOFA factors that co-localize to both known molecular QTL hotspots and novel loci and in the process identified new genes as putative targets for QTL hotspots based on their significant contributions to MOFA factors and concordant allele effects between molecular and MOFA QTLs. With advances in technology and decreases in cost, multi-omics profiling has emerged as a popular tool for studying gene regulation. As demonstrated here and elsewhere, integration across multiple layers of genomic data can increase our power to detect and accurately quantitate regulatory signatures underlying cell state and developmental progression.[61]

Finally, this study revealed variation in protein levels in several known regulators of pluripotency and lineage differentiation, underscoring the labile nature of the pluripotent state across these genetically diverse mESC lines that may span cell states ranging from totipotent 2CLCs to those that are poised for differentiation to 1 or more cell lineages. The observed variability in levels of regulatory proteins and multi-omic signatures across these bulk mESC samples may reflect differences in cell state composition. Higher expression of 2CLC-associated genes, for example, likely indicate lines having a relatively higher proportion of cells in this rare totipotent state. Single-cell platforms will be required to measure the extent to which cellular heterogeneity contributes to the phenotypic variability observed across genetically diverse ESCs. While single-cell transcriptomics and chromatin profiling are now reasonably mature technologies, our study indicates that the picture may remain incomplete without the addition of single-cell proteomics data. Finally, previous work has suggested that differences in differentiation capacity and developmental progression can originate directly at the naive state.[20] How the genetic variation and variable gene regulatory states observed among DO mESCs influence their ability to differentiate into various cell lineages remains largely unexplored. Future studies will seek to characterize whether and how these molecular QTLs in mESCs act to bias cell fate decisions or transcriptional regulation in downstream cell lineages.

Limitations of the study

Our interpretation of the mESC proteome is tempered by known limitations inherent in the proteomics technology and genetic mapping methods. Current untargeted mass spectrometry platforms are biased against transmembrane proteins and lowly expressed genes (Figure 1B) including many transcription factors. As such, we failed to detect some important regulators of pluripotency, including LIFR and NANOG. Furthermore, sample size is a limitation in any QTL mapping study, as high numbers of samples are required to detect QTLs having subtle effects.[62] While our DO mESC panel is of similar size to previous DO mapping studies and well powered to detect local pQTLs, our ability to detect distant pQTLs is limited to those having the largest effects

on protein abundance; e.g., those proteins affected by the *Lifr* locus. Finally, mediation analysis is a powerful tool in the genetic toolkit and enabled us to predict numerous protein mediators of distant pQTL effects. However, its ability to detect mediators of distant QTLs is limited to those variants that affect the expression levels of the protein or transcript mediator. Distant QTLs arising from variants that disrupt the structure or function of a protein intermediate will be missed by mediation analysis. Thus, a complete understanding of the protein regulatory map underlying pluripotency will require targeted approaches or advances in the sensitivity of untargeted proteomics platforms, increased sample sizes in mapping studies to detect subtle genetic effects, and development of alternative approaches to fine-map QTLs and resolve their underlying causal variants.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Diversity Outbred mESC lines
- METHOD DETAILS
  - Sample preparation for proteomics analysis
  - Liquid chromatography and tandem mass spectrometry
  - Mass spectrometry data analysis
  - Protein abundance estimation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Diversity Outbred mESC RNA-seq
  - Diversity Outbred mESC ATAC-seq
  - Gene annotations and id matching across data sets
  - Correlation analysis
  - Gene set enrichment and overrepresentation analysis
  - Gene set variation analysis
  - Quantitative trait locus mapping
  - Defining QTL hotspots
  - Mediation analysis
  - Data integration and multi-omics factor analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100283.

ACKNOWLEDGMENTS

We thank Ann Wells, Greg Carter, and Martin Pera for helpful discussions and feedback on the project and manuscript. Funding sources included NIH grants R35GM133495 to S.C.M.; R01GM070683 to G.A.C.; R35GM133724 and T32HD007065 to C.L.B.; OD010921 and OD011102 to L.G.R.; NIEHS-National Toxicology Program (NTP) HHSN273201500196P to T.C.; F32GM134599 to G.R.K.; R01GM067945 to S.P.G.; and The Jackson Laboratory to S.C.M., C.L.B., L.G.R., and G.A.C. Graphical abstract created with BioRender.com.

AUTHOR CONTRIBUTIONS

Conceptualization, S.C.M., S.A., G.A.C., C.L.B., and L.G.R.; methodology, S.A., S.C.M., G.A.C., C.L.B., L.G.R., D.A.S., S.P.G., T.Z., J.A.P., G.R.K., and D.T.P.; investigation, S.A., D.T.P., T.Z., J.A.P., and M.P.; data curation & formal analysis, S.A., D.T.P., S.C.M., G.A.C.; visualization, S.A., S.C.M., and D.A.S.; writing – original draft, review & editing, S.A., S.C.M., G.A.C., C.L.B., L.G.R., S.P.G., T.Z., D.A.S., and T.C.; supervision, S.C.M., G.A.C., C.L.B., L.G.R., and S.P.G.; project administration, S.C.M., G.A.C., C.L.B., and L.G.R.; funding acquisition, S.C.M., G.A.C., C.L.B., L.G.R., T.C., and S.P.G.

DECLARATION OF INTERESTS

T.C. has an equity interest in Predictive Biology, Inc.

INCLUSION AND DIVERSITY

We worked to ensure sex balance in the selection of non-human subjects. We worked to ensure diversity in experimental samples through the selection of the cell lines. We worked to ensure diversity in experimental samples through the selection of the genomic datasets. One or more of the authors of this paper self-identifies as a gender minority in their field of research.

Received: April 26, 2022
Revised: September 12, 2022
Accepted: February 27, 2023
Published: March 23, 2023

REFERENCES

1. Skelly, D.A., Czechanski, A., Byers, C., Aydin, S., Spruce, C., Olivier, C., Choi, K., Gatti, D.M., Raghupathy, N., Keele, G.R., et al. (2020). Mapping the effects of genetic variation on chromatin state and gene expression reveals loci that control ground state pluripotency. Cell Stem Cell *27*, 459–469.e8. https://doi.org/10.1016/j.stem.2020.07.005.
2. Hamazaki, T., El Rouby, N., Fredette, N.C., Santostefano, K.E., and Terada, N. (2017). Concise Review: induced pluripotent stem cell research in the era of precision medicine. Stem Cell. *35*, 545–550. https://doi.org/10.1002/stem.2570.
3. Ortmann, D., and Vallier, L. (2017). Variability of human pluripotent stem cell lines. Curr. Opin. Genet. Dev. *46*, 179–185. https://doi.org/10.1016/j.gde.2017.07.004.
4. Volpato, V., and Webber, C. (2020). Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility. Dis. Model. Mech. *13*, dmm042317. https://doi.org/10.1242/dmm.042317.
5. Czechanski, A., Byers, C., Greenstein, I., Schrode, N., Donahue, L.R., Hadjantonakis, A.-K., and Reinholdt, L.G. (2014). Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. Nat. Protoc. *9*, 559–574. https://doi.org/10.1038/nprot.2014.030.
6. Byers, C., Spruce, C., Fortin, H.J., Hartig, E.I., Czechanski, A., Munger, S.C., Reinholdt, L.G., Skelly, D.A., and Baker, C.L. (2022). Genetic control of the pluripotency epigenome determines differentiation bias in mouse embryonic stem cells. EMBO J. *41*, e109445. https://doi.org/10.15252/embj.2021109445.
7. Carcamo-Orive, I., Hoffman, G.E., Cundiff, P., Beckmann, N.D., D'Souza, S.L., Knowles, J.W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G.M., et al. (2017). Analysis of transcriptional variability in a large human iPSC library reveals genetic and non-genetic determinants of heterogeneity. Cell Stem Cell *20*, 518–532.e9. https://doi.org/10.1016/j.stem.2016.11.005.
8. Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P., Culley, O.J., et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. Nature *546*, 370–375. https://doi.org/10.1038/nature22403.

9. Mirauta, B.A., Seaton, D.D., Bensaddek, D., Brenes, A., Bonder, M.J., Kilpinen, H., HipSci Consortium; Stegle, O., Lamond, A.I., Danecek, P., et al. (2020). Population-scale proteome variation in human induced pluripotent stem cells. Elife *9*, e57390. https://doi.org/10.7554/eLife.57390.

10. Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C., et al. (2017). iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. Stem Cell Rep. *8*, 1086–1100. https://doi.org/10.1016/j.stemcr.2017.03.012.

11. Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500–505. https://doi.org/10.1038/nature18270.

12. Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nat. Rev. Genet. *21*, 630–644. https://doi.org/10.1038/s41576-020-0258-4.

13. Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. *19*, 1720–1730. https://doi.org/10.1128/MCB.19.3.1720.

14. Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. *583*, 3966–3973. https://doi.org/10.1016/j.febslet.2009.10.036.

15. Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232. https://doi.org/10.1038/nrg3185.

16. van den Berg, P.R., Budnik, B., Slavov, N., and Semrau, S. (2017). Dynamic post-transcriptional regulation during embryonic stem cell differentiation. Syst. Biol. *10*, 1101–123497.

17. Keele, G.R., Zhang, T., Pham, D.T., Vincent, M., Bell, T.A., Hock, P., Shaw, G.D., Paulo, J.A., Munger, S.C., de Villena, F.P.M., et al. (2021). Regulation of protein abundance in genetically diverse mouse populations. Cell Genom. *1*, 100003. https://doi.org/10.1016/j.xgen.2021.100003.

18. Bulut-Karslioglu, A., Macrae, T.A., Oses-Prieto, J.A., Covarrubias, S., Percharde, M., Ku, G., Diaz, A., McManus, M.T., Burlingame, A.L., and Ramalho-Santos, M. (2018). The transcriptionally permissive chromatin state of embryonic stem cells is acutely tuned to translational output. Cell Stem Cell *22*, 369–383.e8. https://doi.org/10.1016/j.stem.2018.02.004.

19. Churchill, G.A., Gatti, D.M., Munger, S.C., and Svenson, K.L. (2012). The diversity outbred mouse population. Mamm. Genome *23*, 713–718. https://doi.org/10.1007/s00335-012-9414-2.

20. Ortmann, D., Brown, S., Czechanski, A., Aydin, S., Muraro, D., Huang, Y., Tomaz, R.A., Osnato, A., Canu, G., Wesley, B.T., et al. (2020). Naive pluripotent stem cells exhibit phenotypic variability that is driven by genetic variation. Cell Stem Cell *27*, 470–481.e6. https://doi.org/10.1016/j.stem.2020.07.019.

21. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124. https://doi.org/10.15252/msb.20178124.

22. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. *21*, 111. https://doi.org/10.1186/s13059-020-02015-1.

23. Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164–1167. https://doi.org/10.1093/bioinformatics/btm069.

24. Epstein, C.J., Smith, S., Travis, B., and Tucker, G. (1978). Both X chromosomes function before visible X-chromosome inactivation in female mouse embryos. Nature *274*, 500–503. https://doi.org/10.1038/274500a0.

25. Kratzer, P.G., and Gartler, S.M. (1978). HGPRT activity changes in preimplantation mouse embryos. Nature *274*, 503–504. https://doi.org/10.1038/274503a0.

26. Kalkan, T., Olova, N., Roode, M., Mulas, C., Lee, H.J., Nett, I., Marks, H., Walker, R., Stunnenberg, H.G., Lilley, K.S., et al. (2017). Tracking the embryonic stem cell transition from ground state pluripotency. Development *144*, 1221–1234. https://doi.org/10.1242/dev.142711.

27. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

28. Schulz, E.G., Meisig, J., Nakamura, T., Okamoto, I., Sieber, A., Picard, C., Borensztein, M., Saitou, M., Blüthgen, N., and Heard, E. (2014). The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating the ESC signaling network. Cell Stem Cell *14*, 203–216. https://doi.org/10.1016/j.stem.2013.11.022.

29. Werner, R.J., Schultz, B.M., Huhn, J.M., Jelinek, J., Madzo, J., and Engel, N. (2017). Sex chromosomes drive gene expression and regulatory dimorphisms in mouse embryonic stem cells. Biol. Sex Differ. *8*, 28. https://doi.org/10.1186/s13293-017-0150-x.

30. Roper, S.J., Chrysanthou, S., Senner, C.E., Sienerth, A., Gnan, S., Murray, A., Masutani, M., Latos, P., and Hemberger, M. (2014). ADP-ribosyltransferases Parp1 and Parp7 safeguard pluripotency of ES cells. Nucleic Acids Res. *42*, 8914–8927. https://doi.org/10.1093/nar/gku591.

31. Guallar, D., Pérez-Palacios, R., Climent, M., Martínez-Abadía, I., Larraga, A., Fernández-Juan, M., Vallejo, C., Muniesa, P., and Schoorlemmer, J. (2012). Expression of endogenous retroviruses is negatively regulated by the pluripotency marker Rex1/Zfp42. Nucleic Acids Res. *40*, 8993–9007. https://doi.org/10.1093/nar/gks686.

32. Masui, S., Ohtsuka, S., Yagi, R., Takahashi, K., Ko, M.S.H., and Niwa, H. (2008). Rex1/Zfp42 is dispensable for pluripotency in mouse ES cells. BMC Dev. Biol. *8*, 45. https://doi.org/10.1186/1471-213X-8-45.

33. Romanov, N., Kuhn, M., Aebersold, R., Ori, A., Beck, M., and Bork, P. (2019). Disentangling genetic and environmental effects on the proteotypes of individuals. Cell *177*, 1308–1318.e10. https://doi.org/10.1016/j.cell.2019.03.015.

34. Hansson, J., and Krijgsveld, J. (2013). Proteomic analysis of cell fate decision. Curr. Opin. Genet. Dev. *23*, 540–547. https://doi.org/10.1016/j.gde.2013.06.004.

35. Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. *17*, 47. https://doi.org/10.1186/s13059-016-0912-5.

36. Gerdes Gyuricza, I., Chick, J.M., Keele, G.R., Deighan, A.G., Munger, S.C., Korstanje, R., Gygi, S.P., and Churchill, G.A. (2022). Genome-wide transcript and protein analysis highlights the role of protein homeostasis in the aging mouse heart. Genome Res. *32*, 838–852. https://doi.org/10.1101/gr.275672.121.

37. Nichols, J., and Smith, A. (2009). Naive and primed pluripotent states. Cell Stem Cell *4*, 487–492. https://doi.org/10.1016/j.stem.2009.05.015.

38. Romero-Lanman, E.E., Pavlovic, S., Amlani, B., Chin, Y., and Benezra, R. (2012). Id1 maintains embryonic stem cell self-renewal by up-regulation of Nanog and repression of brachyury expression. Stem Cells Dev. *21*, 384–393. https://doi.org/10.1089/scd.2011.0428.

39. Moreira, S., Seo, C., Gordon, V., Xing, S., Wu, R., Polena, E., Fung, V., Ng, D., Wong, C.J., Larsen, B., et al. (2018). Endogenous BioID Elucidates TCF7L1 Interactome Modulation upon GSK-3 Inhibition in Mouse ESCs. https://doi.org/10.1101/431023.

40. Wray, J., Kalkan, T., Gomez-Lopez, S., Eckardt, D., Cook, A., Kemler, R., and Smith, A. (2011). Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. Nat. Cell Biol. *13*, 838–845. https://doi.org/10.1038/ncb2267.

41. Walker, E., Chang, W.Y., Hunkapiller, J., Cagney, G., Garcha, K., Torchia, J., Krogan, N.J., Reiter, J.F., and Stanford, W.L. (2010). Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. Cell Stem Cell *6*, 153–166. https://doi.org/10.1016/j.stem.2009.12.014.

42. Ikeda, M., Inoue, F., Ohkoshi, K., Yokoyama, S., Tatemizo, A., Tokunaga, T., and Furusawa, T. (2012). B-Box and SPRY domain containing protein (BSPRY) is associated with the maintenance of mouse embryonic stem cell pluripotency and early embryonic development. J. Reprod. Dev. *58*, 691–699. https://doi.org/10.1262/jrd.2011-009.

43. Qiu, D., Ye, S., Ruiz, B., Zhou, X., Liu, D., Zhang, Q., and Ying, Q.-L. (2015). Klf2 and Tfcp2l1, two Wnt/β-Catenin targets, act synergistically to induce and maintain naive pluripotency. Stem Cell Rep. *5*, 314–322. https://doi.org/10.1016/j.stemcr.2015.07.014.

44. Ye, S., Li, P., Tong, C., and Ying, Q.-L. (2013). Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. EMBO J. *32*, 2548–2560. https://doi.org/10.1038/emboj.2013.175.

45. Schey, K.L., Grey, A.C., and Nicklay, J.J. (2013). Mass spectrometry of membrane proteins: a focus on aquaporins. Biochemistry *52*, 3807–3817. https://doi.org/10.1021/bi301604j.

46. Kalkan, T., and Smith, A. (2014). Mapping the route from naive pluripotency to lineage specification. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130540. https://doi.org/10.1098/rstb.2013.0540.

47. Scott-Boyer, M.-P., Haibe-Kains, B., and Deschepper, C.F. (2013). Network statistics of genetically-driven gene co-expression modules in mouse crosses. Front. Genet. *4*, 291. https://doi.org/10.3389/fgene.2013.00291.

48. Hendrickson, P.G., Doráis, J.A., Grow, E.J., Whiddon, J.L., Lim, J.-W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L., et al. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat. Genet. *49*, 925–934. https://doi.org/10.1038/ng.3844.

49. Bornelöv, S., Selmi, T., Flad, S., Dietmann, S., and Frye, M. (2019). Codon usage optimization in pluripotent embryonic stem cells. Genome Biol. *20*, 119. https://doi.org/10.1186/s13059-019-1726-z.

50. Gonzales-Cope, M., Sidoli, S., Bhanu, N.V., Won, K.-J., and Garcia, B.A. (2016). Histone H4 acetylation and the epigenetic reader Brd4 are critical regulators of pluripotency in embryonic stem cells. BMC Genom. *17*, 95. https://doi.org/10.1186/s12864-016-2414-y.

51. Romero, J.J., De Rossi, M.C., Oses, C., Echegaray, C.V., Verneri, P., Francia, M., Guberman, A., and Levi, V. (2022). Nucleus-cytoskeleton communication impacts on OCT4-chromatin interactions in embryonic stem cells. BMC Biol. *20*, 6. https://doi.org/10.1186/s12915-021-01207-w.

52. Jagust, P., Alcalá, S., Sainz Jr, B., Heeschen, C., and Sancho, P. (2020). Glutathione metabolism is essential for self-renewal and chemoresistance of pancreatic cancer stem cells. World J. Stem Cells *12*, 1410–1428. https://doi.org/10.4252/wjsc.v12.i11.1410.

53. Xin, Y., Wang, Y., Zhong, L., Shi, B., Liang, H., and Han, J. (2019). Slc25a36 modulates pluripotency of mouse embryonic stem cells by regulating mitochondrial function and glutathione level. Biochem. J. *476*, 1585–1604. https://doi.org/10.1042/BCJ20190057.

54. Gu, W., Gaeta, X., Sahakyan, A., Chan, A.B., Hong, C.S., Kim, R., Braas, D., Plath, K., Lowry, W.E., and Christofk, H.R. (2016). Glycolytic metabolism plays a functional role in regulating human pluripotent stem cell state. Cell Stem Cell *19*, 476–490. https://doi.org/10.1016/j.stem.2016.08.008.

55. Cruz, L., Arevalo Romero, J.A., Brandão Prado, M., Santos, T.G., and Hohmuth Lopes, M. (2018). Evidence of extracellular vesicles biogenesis and release in mouse embryonic stem cells. Stem Cell Rev. Rep. *14*, 262–276. https://doi.org/10.1007/s12015-017-9776-7.

56. Frum, T., Murphy, T.M., and Ralston, A. (2018). HIPPO signaling resolves embryonic cell fate conflicts during establishment of pluripotency in vivo. Elife *7*, e42298. https://doi.org/10.7554/eLife.42298.

57. Sun, X., Ren, Z., Cun, Y., Zhao, C., Huang, X., Zhou, J., Hu, R., Su, X., Ji, L., Li, P., et al. (2020). Hippo-YAP signaling controls lineage differentiation of mouse embryonic stem cells through modulating the formation of super-enhancers. Nucleic Acids Res. *48*, 7182–7196. https://doi.org/10.1093/nar/gkaa482.

58. Li, Q.V., Dixon, G., Verma, N., Rosen, B.P., Gordillo, M., Luo, R., Xu, C., Wang, Q., Soh, C.-L., Yang, D., et al. (2019). Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. Nat. Genet. *51*, 999–1010. https://doi.org/10.1038/s41588-019-0408-9.

59. Chen, Q., and Hu, G. (2017). Post-transcriptional regulation of the pluripotent state. Curr. Opin. Genet. Dev. *46*, 15–23. https://doi.org/10.1016/j.gde.2017.06.010.

60. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. *9*, 559. https://doi.org/10.1186/1471-2105-9-559.

61. Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020). Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. *38*, 1007–1022. https://doi.org/10.1016/j.tibtech.2020.02.013.

62. Keele, G.R. (2022). Which Mouse Multiparental Population Is Right for Your Study? the Collaborative Cross Inbred Strains, Their F1 Hybrids, or the Diversity Outbred Population. https://doi.org/10.1101/2022.08.26.505416.

63. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. Preprint at arXiv. https://doi.org/10.48550/arXiv.1406.5823.

64. Wickham, H., Vaughan, D., and Girlich, M. (2023). tidyr: Tidy Messy Data. https://tidyr.tidyverse.org. https://github.com/tidyverse/tidyr.

65. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer). https://doi.org/10.1007/978-0-387-98141-3.

66. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

67. Choi, K., He, H., Gatti, D.M., Philip, V.M., Raghupathy, N., Gyuricza, I.G., Munger, S.C., Chesler, E.J., and Churchill, G.A. (2020). Genotype-free individual genome reconstruction of Multiparental Population Models by RNA sequencing data. Bioinformatics. https://doi.org/10.1101/2020.10.11.335323.

68. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

69. Gatti, D.M., Svenson, K.L., Shabalin, A., Wu, L.-Y., Valdar, W., Simecek, P., Goodwin, N., Cheng, R., Pomp, D., Palmer, A., et al. (2014). Quantitative trait locus mapping methods for diversity outbred mice. G3 *4*, 1623–1633. https://doi.org/10.1534/g3.114.013748.

70. Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145.

71. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.

72. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205. https://doi.org/10.1093/nar/gkz401.

73. Sheffield, N.C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics *32*, 587–589. https://doi.org/10.1093/bioinformatics/btv612.

74. Broman, K.W., Gatti, D.M., Simecek, P., Furlotte, N.A., Prins, P., Sen, Ś., Yandell, B.S., and Churchill, G.A. (2019). R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent

populations. Genetics *211*, 495–502. https://doi.org/10.1534/genetics.118.301595.

75. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
76. Navarrete-Perea, J., Yu, Q., Gygi, S.P., and Paulo, J.A. (2018). Streamlined tandem mass tag (SL-TMT) protocol: an efficient strategy for quantitative (Phospho)proteome profiling using tandem mass tag-synchronous precursor selection-MS3. J. Proteome Res. *17*, 2226–2236. https://doi.org/10.1021/acs.jproteome.8b00217.
77. Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189. https://doi.org/10.1016/j.cell.2010.12.001.
78. Raghupathy, N., Choi, K., Vincent, M.J., Beane, G.L., Sheppard, K.S., Munger, S.C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G.A. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. Bioinformatics *34*, 2177–2184. https://doi.org/10.1093/bioinformatics/bty078.
79. Churchill, G.A., and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. Genetics *138*, 963–971. https://doi.org/10.1093/genetics/138.3.963.
80. Storey, J.D., Taylor, J.E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. Roy. Stat. Soc. B *66*, 187–205.
81. Sen, S., and Churchill, G.A. (2001). A statistical framework for quantitative trait mapping. Genetics *159*, 371–387. https://doi.org/10.1093/genetics/159.1.371.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Recombinant Mouse LIF protein	Isolated from Chinese Hamster Ovary (CHO) cell line	N/A
CHIR99021 GSK-3 inhibitor	Tocris	Cat# 4423, CAS: 252917-06-9
PD0325901 MEK/ERK pathway inhibitor	STEMCELL Technologies	Cat# 72184, CAS: 391210-10-9
Pierce Protease Inhibitor Tablets	Thermo Fisher	A32963
Pierce Phosphatase Inhibitor Mini Tablets	Thermo Fisher	A32957
Trypsin Protease MS grade, Frozen	Thermo Fisher	90305R200
Lys-C, Mass Spectrometry Grade	Wako Chemicals	Barcode#4987481427648
TMT10plex Isobaric Label reagent Set plus TMT11-131C Label Reagent	Thermo Fisher	A34808
Critical commercial assays		
Pierce BCA Protein Assay Kit	Thermo Fisher	23227
Deposited data		
DO mESC proteomics	ProteomeXchange (http://www.proteomexchange.org)	PXD033001
DO mESC RNA-Seq and ATAC-Seq	Skelly et al.[1]; ArrayExpress (https://www.ebi.ac.uk/arrayexpress/)	E-MTAB-7728 (DO mESC RNA-Seq); E-MTAB-8759 (DO mESC ATAC-Seq)
Experimental models: Cell lines		
190 Diversity Outbred mESC lines	Predictive Biology	N/A
Experimental models: Organisms/strains		
Mouse: J:DO	The Jackson Laboratory	JAX: 009376
Software and algorithms		
lme4	Bates et al.[63]	https://cran.r-project.org/web/packages/lme4/index.html
R	The R Project	https://www.r-project.org
tidyr	Wickham et al.[64]	https://tidyr.tidyverse.org/
ggplot2	Wickham[65]	https://ggplot2.tidyverse.org/
pheatmap	N/A	https://cran.r-project.org/web/packages/pheatmap/index.html
bowtie v1.1.2	Langmead et al.[66]	https://bowtie-bio.sourceforge.net/index.shtml
gbrs v0.1.6	Choi et al.[67]	https://churchill-lab.github.io/gbrs/
ComBat	Johnson et al.[68]	https://doi.org/10.1093/biostatistics/kxj037
DOQTL	Gatti et al.[69]	https://doi.org/10.1534/g3.114.013748
ChIPseeker	Yu et al.[70]	https://guangchuangyu.github.io/software/ChIPseeker/
Hmisc	N/A	https://cran.r-project.org/web/packages/Hmisc/index.html
gProfiler2	Raudvere et al.[71]	https://biit.cs.ut.ee/gprofiler_archive3/e106_eg53_p16/gost
WebsGestaltR	Liao et al.[72]	http://www.webgestalt.org/
LOLA	Sheffield et al.[73]	https://code.databio.org/LOLA/
qvalue	N/A	https://github.com/StoreyLab/qvalue
GSVA	Hänzelmann et al.[27]	https://doi.org/10.18129/B9.bioc.GSVA
rstatix	N/A	https://rpkgs.datanovia.com/rstatix/
qtl2	Broman et al.[74]	https://github.com/rqtl/qtl2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bioconductor	Bioconductor	https://bioconductor.org
GenomicRanges	Lawrence et al.[75]	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
intermediate	N/A	https://github.com/churchill-lab/intermediate
pcaMethods	Stacklies et al.[23]	https://www.bioconductor.org/packages/release/bioc/html/pcaMethods.html
MOFA	Argelaguet et al.[21,22]	https://biofam.github.io/MOFA2/
Other		
Resource website for the publication containing all the code and data for reproducing figures and tables	http://do_mesc_proteomics.jax.org/	N/A
Processed data (e.g., proteomics, genotype probabilities) and code to reproduce figures.	https://doi.org/10.6084/m9.figshare.22012850	N/A
Gene Ontology Terms used in GSVA	Mouse Genome Informatics	http://www.informatics.jax.org/gotools/data/input/MGIgenes_by_GOid.txt

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Steven C. Munger (steven.munger@jax.org).

Materials availability

There are restrictions on the availability of Diversity Outbred mESCs used in this study due to overlap with intellectual property claims for the Predictive Biology *in vitro* genetics platform. Predictive Biology, Inc. offers access to these and additional lines on a commercial basis through their genetic screening and stem cell biology services.

Data and code availability

The DO mESC mass spectrometry proteomics data have been deposited in ProteomeXchange (http://www.proteomexchange.org/) via the PRIDE partner repository (ProteomeXchange: PXD033001). The DO mESC RNA-seq and ATAC-seq data were published in[1] and deposited to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) (RNA-seq ArrayExpress: E-MTAB-7728; ATAC-seq ArrayExpress: E-MTAB-8759).

All processed data and code to generate main and supplemental figures have been deposited to https://doi.org/10.6084/m9.figshare.22012850.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Diversity Outbred mESC lines

Mouse embryonic stem cell lines were derived from male and female Diversity Outbred mice (JR #009376, The Jackson Laboratory) and maintained at Predictive Biology, Inc. as previously described.[1] Briefly, at 24-26 days of age female DO mice were superovulated and mated to 7-15 week old males. Next, mESCs were derived from random blastocysts following previously described protocols.[5] The blastocysts were transferred to 96-well round-bottom ultra-low attachment plates containing 2i medium (2i + LIF: Dulbecco's Modified Eagle Medium (DMEM) supplemented with 15% fetal bovine serum, 100 U/mL Penicillin-Streptomycin, 2mM GlutaMAX, 0.1mM non-essential amino acids, 1mM sodium pyruvate, 0.1mM 2-mercaptoethanol, 500pM LIF, 1uM PD0325901, and 3uM CHIR99021) for 5-7 days. Blastocysts that showed inner cell mass outgrowth were dispersed and transferred onto 96-well flat-bottom tissue culture plates containing mitotically inactivated mouse embryonic fibroblasts (MEFs, C57BL/6J) in ES medium (ESM, 1i+LIF: Dulbecco's Modified Eagle Medium (DMEM) supplemented with 15% fetal bovine serum, 100 U/mL Penicillin-Streptomycin, 2mM GlutaMAX, 0.1mM non-essential amino acids, 1mM sodium pyruvate, 0.1mM 2-mercaptoethanol, approximately 2000U/ml LIF, and 3uM CHIR99021). The ES cells were weaned off feeder cells as they expanded via dilution by transferring into 24 well plates, followed by 6 well plates and finally into 10cM dishes containing ESM without feeder cells. LIF protein used in the experiments was produced by Predictive Biology using a Chinese Hamster Ovary cell line. For proteomics analysis, ~100,000 cryopreserved DO mESCs from each line were sent from Predictive Biology to the Gygi Lab at Harvard Medical School.

METHOD DETAILS

Sample preparation for proteomics analysis

Frozen cell pellets were resuspended in 8 M Urea, 200 mM EPPS, pH 8.5, with protease inhibitor, and lysed by passing through a 21-gauge needle with syringe. After centrifugation at 13,000 rpm at 4°C for 10min, supernatant was used for further analysis. BCA assay was performed to determine protein concentration of each sample. Samples were reduced in 5 mM TCEP for 15min, alkylated with 10 mM iodoacetamide for 15min, and quenched with 15 mM DTT for 15min. 200 μg protein was chloroform-methanol precipitated and re-suspended in 200 μL 200 mM EPPS (pH 8.5). Protein was digested by Lys-C at a 1:100 protease-to-peptide ratio overnight at room temperature with gentle shaking. Trypsin was used for further digestion for 6 hours at 37°C at 1:100. 100 μL of each sample were aliquoted. 30 μL acetonitrile (ACN) was added into each sample to 30% final volume. 200 μg TMT reagent (126, 127N, 127C, 128N, 128C, 129N, 129C, 130N, 130C, 131N) in 10 μL ACN was added to each sample. After 1 hour of labeling, 2 μL of each sample was combined, desalted, and analyzed using mass spectrometry. TMT labeling efficiency was calculated and over 99%. After quenching using 0.3% hydroxylamine, 10 samples in each TMT were combined and fractionated with basic pH reversed phase (BPRP) high performance liquid chromatography (HPLC), collected onto a 96 six well plate and combined for 24 fractions in total. Twelve fractions were desalted and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).[76]

Liquid chromatography and tandem mass spectrometry

For the BPRP fractions, mass spectrometric data were collected on an Orbitrap Fusion mass spectrometer coupled to a Proxeon NanoLC-1200 UHPLC. The 100 μm capillary column was packed with 35 cm of Accucore 50 resin (2.6 μm, 150Å; ThermoFisher Scientific). The mobile phase was 5% acetonitrile, 0.125% formic acid (A) and 95% acetonitrile, 0.125% formic acid (B). The data were collected using a DDA-SPS-MS3 method. Each fraction was eluted using a 150 min method over a gradient from 6% to 30% B. Peptides were ionized with a spray voltage of 2,600 kV. The instrument method included Orbitrap MS1 scans (resolution of 1.2 x105; mass range 350−1400 m/z; automatic gain control (AGC) target 5x105, max injection time of 100 ms and ion trap MS2 scans (CID collision energy of 35%; AGC target 2x104; rapid scan mode; max injection time of 120 ms). MS3 precursors were fragmented by HCD and analyzed using the Orbitrap (NCE 65%, AGC 1 x105, maximum injection time 150 ms, resolution was 5 x104 at 400 Th). Detailed parameters for MS2 and MS3 are embedded in the RAW files.

Mass spectrometry data analysis

Mass spectra were processed using a Sequest-based pipeline.[77] Spectra were converted to mzXML using a modified version of ReAdW.exe. Database search included all entries from an indexed Ensembl database version 98. This database was concatenated with one composed of all protein sequences in the reversed order. Searches were performed using a 50 ppm precursor ion tolerance for total protein level analysis. The product ion tolerance was set to 0.9 Da. TMT tags on lysine residues and peptide N termini (+229.163 Da) and carbamidomethylation of cysteine residues (+57.021 Da) were set as static modifications, while oxidation of methionine residues (+15.995 Da) was set as a variable modification. In addition, for phosphopeptide analysis, phosphorylation (+79.966 Da) on serine, threonine, and tyrosine are included as variable modifications. Peptide-spectrum matches (PSMs) were adjusted to a 1% false discovery rate (FDR). PSM filtering was performed using a linear discriminant analysis (LDA). For TMT-based reporter ion quantitation, we extracted the summed signal-to-noise (S:N) ratio for each TMT channel and found the closest matching centroid to the expected mass of the TMT reporter ion. For protein-level comparisons, PSMs were identified, quantified, and collapsed to a 1% peptide false discovery rate (FDR) and then collapsed further to a final protein-level FDR of 1%, which resulted in a final peptide level FDR of <0.1%. Moreover, protein assembly was guided by principles of parsimony to produce the smallest set of proteins necessary to account for all observed peptides. Proteins were quantified by summing reporter ion counts across all matching PSMs. PSMs with poor quality, MS3 spectra with less than 10 TMT reporter ion channels missing, MS3 spectra with TMT reporter summed signal-to-noise of less than 100 or having no MS3 spectra were excluded from quantification. Each reporter ion channel was summed across all quantified proteins and normalized assuming equal protein loading of all 10 samples. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with dataset identifier PXD033001.

Protein abundance estimation

Protein abundances were estimated as described previously.[17] Briefly, peptides that contain polymorphisms were filtered and TMT batch effects were removed from the filtered peptide data using a linear mixed model fit with the R/lme4 package.[63] Finally, protein abundances were estimated and normalized using the processed peptide data as described in detail in Keele et al.[17] Proteins missing values in more than 50% of the samples were removed from further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

All analyses and figures were generated with the R statistical programming language and are available at the following web resource [https://DO_mESC_proteomics.jax.org] and github [https://github.com/selcant/Aydin_et_al_website]. Unless otherwise stated R/tidyr package was used for data processing, R/ggplot2[65] for plotting and R/pheatmap for heatmap plots.

Diversity Outbred mESC RNA-seq

Raw RNA-seq data was retrieved (ArrayExpress: E-MTAB-7728) and analyzed as previously described,[1] but using both paired-end sequencing reads instead of single end. Briefly, we aligned paired-end 75 bp reads with bowtie v1.1.2[66] to a pooled ''8-way'' transcriptome containing strain-specific isoform sequences from all eight DO founder strains, then resolved multi-mapping reads and estimated transcript- and gene-level abundance for each sample using the EMASE method as implemented in gbrs v0.1.6.[67,78] Genes with a median TPM (transcripts per million) value smaller than 0.5 or zero value (i.e., not expressed) in more than half of the samples were filtered. Next, we normalized gene-level counts to the upper quartile value to account for differences in library size and then applied the ComBAT function from R/sva package to remove batch effects caused by library preparation.[68] For QTL mapping, we transformed normalized values to rank normal scores using rankZ normalization as implemented in the DOQTL R package.[69] Finally, sample mix-ups were resolved by comparing the genotypes inferred from the RNA-seq data using gbrs v0.1.6 (http://churchill-lab.github.io/gbrs/) to genotypes inferred from DNA microarrays (GigaMUGA platform, Neogen Geneseek).

Diversity Outbred mESC ATAC-seq

Normalized ATAC-seq peak values from Skelly et al.[1] were further processed using the ComBAT function in the R/sva package to remove any potential batch effects caused by library preparation.[68] Normalized, batch-corrected peak values were used in all correlation analyses. For QTL mapping, these values were further transformed to rank normal scores using the rankZ function from the DOQTL package.[69] For annotation of ATAC-seq peaks we utilized the ChIPseeker R package.[70]

Gene annotations and id matching across data sets

Transcript abundance data was annotated to Ensembl gene identifiers, proteomics data was annotated to Ensembl protein identifiers, and ATAC-seq data was annotated to Ensembl gene ids using ChIPSeeker R package. We used ENSEMBL v98 to add gene annotations such as MGI symbol, gene location, and gene biotype. MGI symbol was used as the identifier for all downstream analysis such as overrepresentation and gene set enrichment.

Correlation analysis

We used the rcorr function from the R/Hmisc package to calculate Pearson correlations. Individual p-values were adjusted for multiple testing using the p.adjust function in R/base and specifying the Benjamini-Hochberg ("BH") option to estimate the false discovery rate (FDR).

Sample-to-sample correlation for protein abundance

For proteome-to-proteome comparisons, we used the abundance of 7,432 proteins across 190 cell lines. To compare chromatin accessibility profiles to the proteome, we used 36,859 ATAC-seq peaks annotated to 6,865 proteins and their corresponding protein abundances in 163 cell lines for which ATAC-seq, transcriptomics and proteomics were profiled. Similarly, for comparing the transcriptome to the proteome across 174 cell lines that had both RNA-seq and proteomics data, we used the overlapping set of 7,241 genes with both transcript and protein abundance measures.

Correlation between chromatin accessibility and protein abundance

Pairwise Pearson correlations were calculated between the abundance of 7,148 autosomal proteins and the chromatin accessibility of 99,159 autosomal ATAC-seq peaks across 163 cell lines for which ATAC-seq, transcriptomics and proteomics were profiled. We excluded sex chromosome proteins and ATAC-seq peaks from the analysis.

Correlation between transcript and protein abundance for individual genes

Pairwise Pearson correlations were calculated for 7,241 genes with both transcript and protein abundance measures across 174 cell lines that had both RNA-seq and proteomics data.

Correlation between complex member and non-complex member proteins

The list of complex member proteins was retrieved from[33] which includes protein complexes manually curated using CORUM and COMPLEAT databases. Pairwise Pearson correlations between protein abundances of complex member and non-complex member genes were calculated for complexes with five or more subunits (n = 164) excluding proteins with significant pQTL to leave out large genetic effects that may not be shared among complex members.

Correlation between complex members

For each protein complex (n = 164), the median pairwise Pearson correlation between an individual protein subunit and the other members are calculated. Complex cohesiveness was then calculated as the median value of the correlations for the individual proteins. For each complex, sex effects were assessed by a one-way ANOVA followed by a t-test comparing the median complex co-abundances in XY and XX samples. P-values were adjusted using the p.adjust function in R/base and specifying the Benjamini-Hochberg (''BH'') option to estimate the false discovery rate (FDR) (n = 164 tests). Complexes that show a significant effect (p-value < 0.05) in both statistical tests were reported.

Gene set enrichment and overrepresentation analysis

We performed overrepresentation analysis using the 'gost' function in the gProfiler2 package by controlling the version using 'set_base_url(https://biit.cs.ut.ee/gprofiler_archive3/e106_eg53_p16)' in R[71] using an appropriate universal background on a case-by-case basis and 'fdr' option for p-value correction. For example, when looking at the functional enrichments in proteins

with high variation all genes identified in proteomics were used whereas only the shared set of genes between RNA-seq and proteomics was used when looking at genes with positive correlation between transcript and protein abundance. For gene set enrichment analysis, we used the WebsGestaltR R package.[72] To identify overrepresentation of genomic regions we utilized R package LOLA[73] which looks at the overlap between user data sets and public genomic data sets like transcription factor binding sites from ENCODE and the CODEX database. Following instructions of the R/LOLA package the p-values were transformed to q-values using the R/qvalue package to get FDR values.

Gene set variation analysis

We performed Gene Set Variation Analysis using the R/Bioconductor package GSVA[27] Gene Ontology terms with gene symbols were retrieved from MGI (http://www.informatics.jax.org/gotools/data/input/MGIgenes_by_GOid.txt) which included 8,436 GO Biological Process gene sets. List of protein complexes and subunits was retrieved from[33] which includes protein complexes manually curated using CORUM and COMPLEAT databases (n = 164). Enrichment scores were calculated using the abundance of 7,432 proteins across 190 cell lines for each gene set with at least 5 overlapping proteins (n = 900 GO terms and n = 158 complexes). Next, we evaluated the significance of enrichment scores across experimental covariates using a two-way ANOVA ($\sim$ sex + *Lifr* genotype + sex:*Lifr* genotype) where individual p values were corrected for multiple testing using the p.adjust function in R/base and specifying the Benjamini-Hochberg ("BH") option and followed by Tukey's HSD using R/rstatix package for pairwise comparisons (n = 1,058 tests). Categories that showed significance in both statistical tests were reported with the p value obtained from Tukey's HSD. Similarly, GSVA was performed using transcript abundances to calculate enrichment scores using the abundance of 14,405 genes across 184 cell lines for each gene set with at least 5 overlapping genes (n = 2,094 GO terms).

Quantitative trait locus mapping

Genetic mapping was performed using a linear-mixed model implemented as the 'scan1' function in R/qtl2 package.[74] We mapped using the normalized, transformed values with sex as a covariate and the Leave One Chromosome Out (loco) option for kinship correction.[69] To estimate genome-wide significance, we permuted genotypes 1000 times while maintaining the relationship between the phenotype and covariates. For each permutation we retained the maximum LOD score in order to generate a null distribution for the test statistic.[79] To calculate thresholds for pQTL, we repeated this permutation strategy for all proteins and estimated a significance cutoff at LOD > 7.5 (alpha = 0.05), and a suggestive cutoff at LOD > 6. False discovery rates (q-values) were determined for each permutation-derived p-value with R/qvalue software, using the bootstrap method to estimate π_0 and the default λ tuning parameters.[80] Support intervals for each QTL were defined by the 95% Bayesian credible interval.[81] We call a QTL 'local' if the QTL peak is within $\pm$10 Mbp to the midpoint of its corresponding gene and 'distal' if otherwise. Founder allele effects were estimated as best linear unbiased predictors (BLUPs) at the QTL using scan1blup function in R/qtl2 package. Previous work has estimated the genome-wide significance threshold at 7.6 and 7.5 for chromatin accessibility QTL (caQTL) and expression QTL (eQTL) respectively.[1] To identify overlaps with significant pQTL, we used a relaxed threshold of LOD > 5 for caQTL and eQTL. They were classified as shared if the QTL peaks were within +/-5 Mb of the significant pQTL peak and the absolute correlation between haplotype effects was higher than 0.5.

Defining QTL hotspots

We first identified distal QTL that reach genome-wide permutation-based threshold ($p < 0.05$; LOD 7.5). Next, we applied a sliding window method to identify hotspots as described in Skelly et al.[1] Briefly, we counted the number of distal QTL within 1cM windows (0.25 cM shift) across the genome and selected the top 0.5% of bins with the most distant pQTL (0.5% bin threshold $\geq$8 distant pQTLs). Final coordinates for each hotspot were determined using the Bioconductor package 'GenomicRanges' to merge adjacent bins into a single region.[75]

Mediation analysis

We used mediation analysis to identify regions of open chromatin, transcript, and protein abundance that were likely to be the causal mediator of a caQTL, eQTL, or pQTL. Mediation analysis was performed using the 'intermediate' package in R (https://github.com/simecek/intermediate) by regressing each target (T) on a mediator (M) at the QTL (Q) and adjusting for covariates. We applied the 'double-lod-diff' method to reduce the effects of missing values. For mediation of QTL with the matching data type we used the full sample set, e.g., pQTL mediation by proteins ($Q_{pQTL} \rightarrow Protein_M \rightarrow Protein_T$) were done using all the 190 samples. On the other hand, mediation across data types were done on common set of samples e.g., for mediation between protein and transcript ($Q_{pQTL} \rightarrow Transcript_M \rightarrow Protein_T \mid Q_{eQTL} \rightarrow Protein_M \rightarrow Transcript_T$) only the 174 samples with both protein and transcript measurements were used. To assess the significance of a LOD drop, we mediated the QTL against all of the mediator data, converted the recorded LOD scores to normal scores, and checked if the score fell below 6 standard deviations from the mean.[11] Mediators were further filtered to narrow down top candidates to include genes with midpoints that are found within 10Mb of the QTL peak.

Data integration and multi-omics factor analysis

For data integration we used Multi-Omics Factor Analysis (MOFA) implemented in Python (mofapy2) and in R (MOFA2).[21,22] MOFA integrates multi-omics data sets in an unsupervised fashion using a factor analysis model and infers interpretable latent factors. All

transcripts (n = 14,405), proteins (n = 7,432) and the most variable 15,000 ATAC-seq peaks based on total variance were used for integration from 163 cell lines with all three molecular measurements. All three datasets were log transformed using base R function log1p before modeling with MOFA. For model generation, we modified the following options from default: we set number of factors to 30, number of maximum iterations to 10,000, convergence mode to ''slow'' and scale views option to TRUE. The model with the best convergence based on the evidence lower bound statistic (ELBO) was saved for further analysis. Next, factors that showed a significant correlation to the total number of expressed features and that didn't explain more than 1% variation in at least one data set were removed resulting in 23 latent factors. We used the calculate_variance_explained and correlate_factors_with_covariates functions in the MOFA2 R package to calculate the proportion of variance explained by factor per data set and correlations between factors and experimental covariates, respectively. Functional characterization of MOFA Factors was done using the R/LOLA package for top ATAC-seq peak drivers and the R/WebsGestaltR package for transcripts and proteins. Top ATAC-seq drivers were obtained using the base R boxplot.stats function where the outliers correspond to data points that lie outside 1.5 times the interquartile range. MOFA factor weights were used to rank genes in enrichment analysis for transcripts and proteins. MOFA Factor values were rankZ transformed and QTL mapping, mediation, and permutation analysis with factors were done as described above using genotype probabilities from the 163 samples used in MOFA.