Northumbria Research Link

Citation: Li, Hao, Wang, Xu, Rukina, Daria, Huang, Qingyao, Lin, Tao, Sorrentino, Vincenzo, Zhang, Hongbo, Bou Sleiman, Maroun, Arends, Danny, McDaid, Aaron, Luan, Peiling, Ziari, Naveed, Velázquez-Villegas, Laura A., Gariani, Karim, Kutalik, Zoltan, Schoonjans, Kristina, Radcliffe, Richard A., Prins, Pjotr, Morgenthaler, Stephan, Williams, Robert W. and Auwerx, Johan (2018) An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. Cell Systems, 6 (1). 90-102.e4. ISSN 2405-4712

Published by: Elsevier

URL: https://doi.org/10.1016/j.cels.2017.10.016 <https://doi.org/10.1016/j.cels.2017.10.016>

This version was downloaded from Northumbria Research Link: https://nrl.northumbria.ac.uk/id/eprint/50160/

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: http://nrl.northumbria.ac.uk/policies.html

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

Northumbria University NEWCASTLE

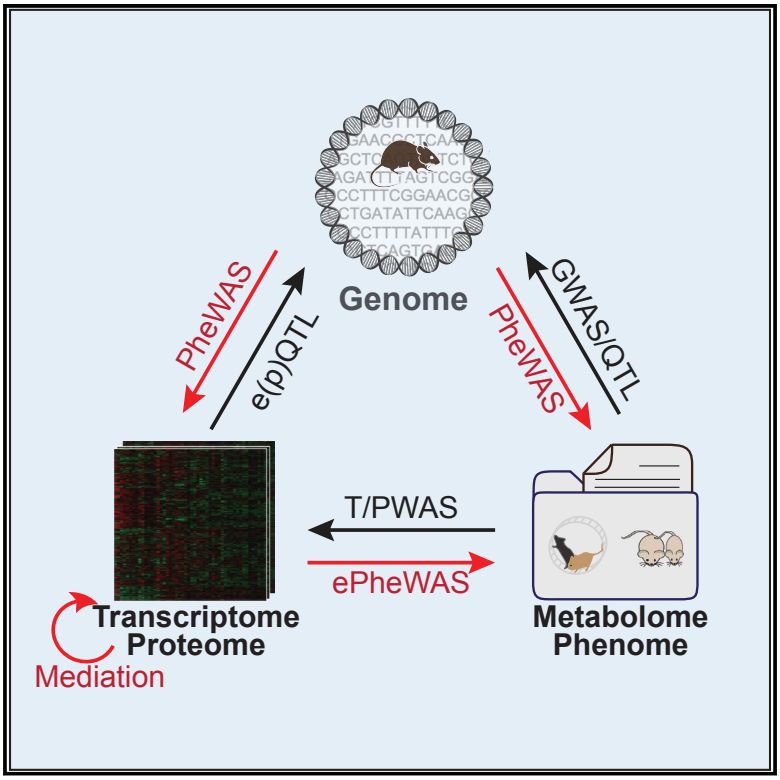
UniversityLibrary

Article

Cell Systems

An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function

Graphical Abstract



Authors

Hao Li, Xu Wang, Daria Rukina, ...,
Stephan Morgenthaler,
Robert W. Williams, Johan Auwerx

Correspondence

admin.auwerx@epfl.ch

In Brief

Li et al. here develop and implement a series of systems tools and establish a web resource using multi-omics datasets of the BXD mouse cohort to identify novel associations between genes and phenotypes.

Highlights

- The BXD mouse cohort is one of the largest resources for multi-omics analysis
- Integrative and complimentary approaches allow the dissection of complex traits
- The open-access analysis platform expedites *in silico* gene function prediction
- New genes influencing metabolism can be identified and validated

Li et al., 2018, Cell Systems *6*, 90–102
January 24, 2018 © 2017 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.10.016

CellPress

OPEN ACCESS
CellPress

Cell Systems
Article

An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function

Hao Li,[1,10] Xu Wang,[1,10] Daria Rukina,[2] Qingyao Huang,[3] Tao Lin,[1] Vincenzo Sorrentino,[1] Hongbo Zhang,[1] Maroun Bou Sleiman,[1] Danny Arends,[4] Aaron McDaid,[5,6] Peiling Luan,[1] Naveed Ziari,[1] Laura A. Velázquez-Villegas,[3] Karim Gariani,[1] Zoltan Kutalik,[5,6] Kristina Schoonjans,[3] Richard A. Radcliffe,[7] Pjotr Prins,[8,9] Stephan Morgenthaler,[2] Robert W. Williams,[9] and Johan Auwerx[1,11,*]

[1]Laboratory for Integrative and Systems Physiology, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland
[2]Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland
[3]Laboratory of Metabolic Signaling, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland
[4]Albrecht Daniel Thaer-Institut für Agrar- und Gartenbauwissenschaften, Humboldt-Universität zu Berlin, D-10115 Berlin, Germany
[5]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
[6]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne 1010, Switzerland
[7]Department of Pharmaceutical Sciences, University of Colorado, Aurora, CO 80045, USA
[8]University Medical Center Utrecht, 3584CT Utrecht, the Netherlands
[9]Department of Genetics, Genomics and Informatics, University of Tennessee, Memphis, TN 38163, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: admin.auwerx@epfl.ch
https://doi.org/10.1016/j.cels.2017.10.016

SUMMARY

Identifying genetic and environmental factors that impact complex traits and common diseases is a high biomedical priority. Here, we developed, validated, and implemented a series of multi-layered systems approaches, including (expression-based) phenome-wide association, transcriptome-/proteome-wide association, and (reverse-) mediation analysis, in an open-access web server (systems-genetics.org) to expedite the systems dissection of gene function. We applied these approaches to multi-omics datasets from the BXD mouse genetic reference population, and identified and validated associations between genes and clinical and molecular phenotypes, including previously unreported links between *Rpl26* and body weight, and *Cpt1a* and lipid metabolism. Furthermore, through mediation and reverse-mediation analysis we established regulatory relations between genes, such as the co-regulation of BCKDHA and BCKDHB protein levels, and identified targets of transcription factors E2F6, ZFP277, and ZKSCAN1. Our multifaceted toolkit enabled the identification of gene-gene and gene-phenotype links that are robust and that translate well across populations and species, and can be universally applied to any populations with multi-omics datasets.

INTRODUCTION

Unraveling the genetic basis of complex traits is crucial to understand the pathogenesis of disease and to develop effective therapies. Genetic studies using human populations have successfully discovered many gene-to-phenotype (G2P) associations, but this approach falls short in controlling for environmental influences and is constrained by limited access to relevant deep tissue samples for mechanistic validation studies (Altshuler et al., 2008; Williams and Auwerx, 2015). Genetically diverse cohorts of model organisms, ranging from yeast, *Caenorhabditis elegans*, *Drosophila melanogaster*, to mouse and rat, can model the complex genetics of human populations, while providing tight control over environmental factors to study gene-by-environmental interactions (GXE), and allowing access to deep tissues at different ages and treatments (Aitman et al., 2011; Cook et al., 2017; Ehrenreich et al., 2010; Flint and Mackay, 2009; Williams and Auwerx, 2015).

In principle, systems genetics approaches for complex trait analysis employ either forward or reverse genetic strategies. Forward genetic tools, such as genome-wide association studies (GWAS) and quantitative trait loci (QTL) linkage studies have been successfully applied to dissect complex traits (Flint and Eskin, 2012; McCarthy et al., 2008). To reveal potential pleiotropic phenotypes associated with gene variants and QTLs, phenome-wide association studies (PheWAS) have emerged as a viable reverse genetic strategy in humans (Bush et al., 2016; Denny et al., 2016). We recently applied PheWAS in the BXDs, enabling the discovery of novel G2P associations, which were then validated in independent human cohorts or by experimental approaches (Wang et al., 2016). These early approaches, however, do not exploit the full spectrum of possible relationships between genotypes, intermediate phenotypes, and clinical phenotypes (see Figures 1A and 1B). Furthermore, exploring this space is difficult in humans because of limited availability of populations with deep genome, transcriptome, proteome, and phenome data. This is, however, less of an issue in populations of model organisms, such as the BXD mouse, the DGRP fly, or the 1001 Genomes *A. thaliana* genetic reference

 © 2017 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

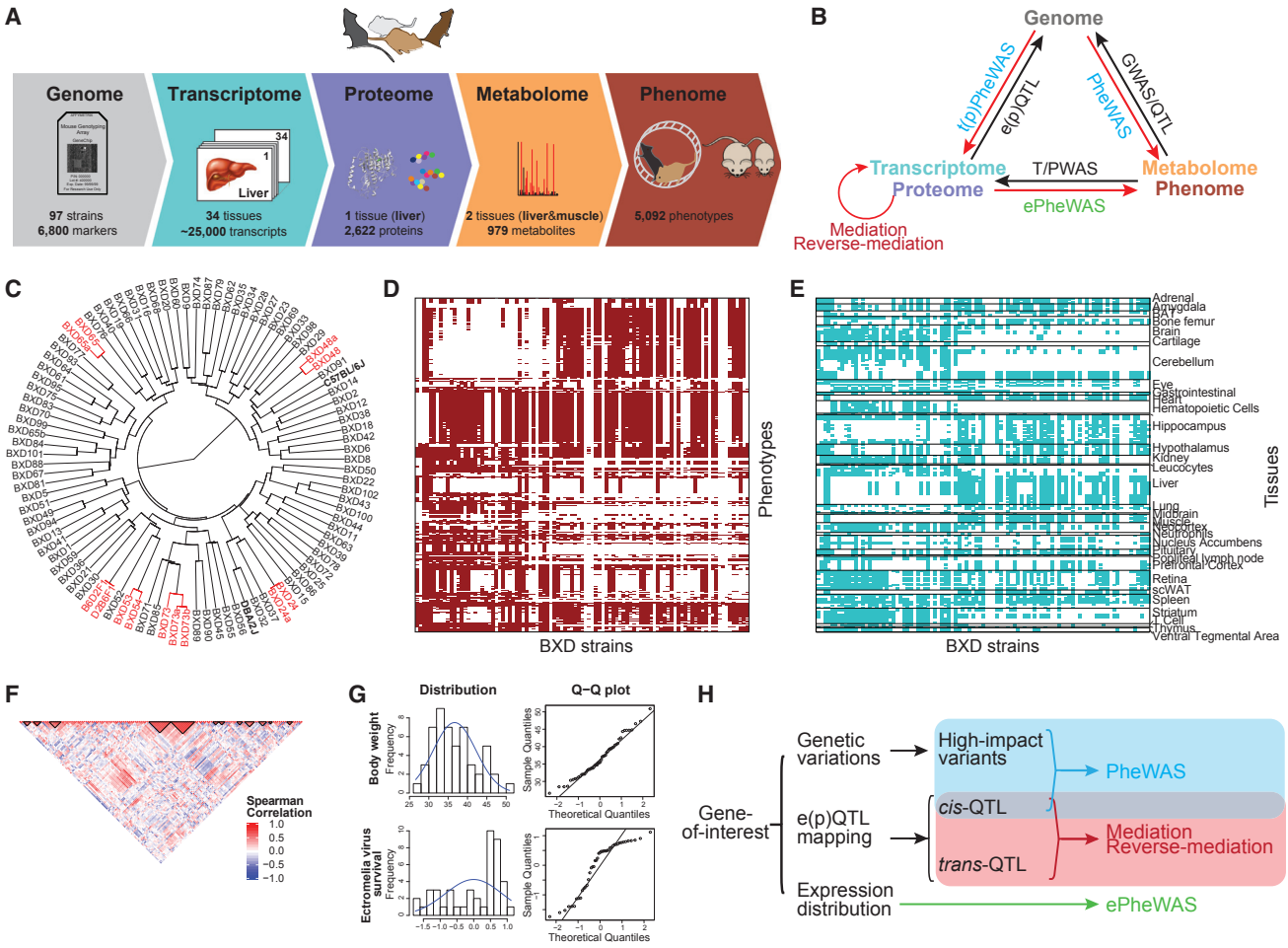


Figure 1. Overview of Multi-omic Data from the BXD Population, and the Scheme of Applied Systems Approaches

(A) Multi-omic data of the BXD population. Genome: genotype data were collected for 6,800 markers. Transcriptome: levels of ~25,000 transcripts have been measured from 34 tissues (see also Figure S1B, Table S2). Proteome: expression of ~2,600 proteins were quantified in livers by mass spectrometry (Williams et al., 2016). Metabolome: ~980 metabolites have been measured in both liver and muscle (Williams et al., 2016). Phenome: ~5,000 clinical phenotypes have been collected by more than 200 research groups (see also Figure S1A, Table S1).

(B) Systems approaches that can be applied using the multi-omic data in BXDs. Approaches developed in this study are highlighted with red arrows and the same colors as corresponding text in (H).

(C) Circular dendrogram showing the genetic relatedness among BXD strains. Sister strains with over 80% identical by descent are highlighted in red, and parental strains (C57BL/6J and DBA/2J) are in bold.

(D) An overview of BXD phenome. Phenotypes were aligned (vertical) based on the groups where the phenotypes were measured. Red blocks indicate that phenotypic data of the particular strain are available, while white blocks show that data are missing or not measured.

(E) Distribution of transcriptome datasets across 34 tissues. Blue blocks indicate that transcript data are available, while white blocks show missing or unmeasured data.

(F) Relatedness of phenotypes. Phenotypes from (Andreux et al., 2012) were clustered based on the correlation between phenotypes. PL blocks were indicated using black triangles.

(G) Normality of two phenotype examples, body weight (upper panel) showing normal distribution and ectromelia virus survival (lower panel) showing non-normal distribution, were represented using histogram and Q-Q plot.

(H) Flowchart for the systems approaches using the multi-omic BXD data. The gene-of-interest is first inspected on three aspects in the BXD GRP, i.e., the existence of genetic variations, e(p)QTLs, and its expression across strains. PheWAS can be applied on genes that possess high-impact variants or *cis*-QTLs to identify the associated traits. Genes that have *cis*- or *trans*-QTLs can be analyzed to reveal the regulatory mechanism of gene expression through (reverse-) mediation analysis. ePheWAS investigates the association between gene expression and phenotypic traits.

See also Figure S1 and Tables S1 and S2.

populations (GRPs), where such data are readily available. We hence exploited the full complexity of G2P relationships in the BXDs, one of the most widely used mouse GRPs, and developed an easy-to-use resource (systems-genetics.org) for the research community.

First, we systematized and improved the PheWAS method both to detect G2P links and validate putative associations from independent studies. We also developed a set of methods to analyze the different layers of omics data that contribute to complex traits. In particular, intermediate phenotypes, including

transcripts, proteins, and metabolites (Gagneur et al., 2013; Williams et al., 2016; Wu et al., 2014) were exploited to consolidate G2P and GXE connections. Despite their potential, transcriptome-/proteome-wide association studies (T/PWAS), which test the associations between a phenotype and all transcripts or proteins of a given tissue, have not been fully explored (Gusev et al., 2016; Okada et al., 2016), largely because of the limited availability of cohorts with such data (see above). With transcriptome/proteome data from over 30 tissues available, the BXD cohort serves as a perfect resource for such analysis. Similarly, reversal of such T/PWAS approaches, i.e., expression-based PheWAS (ePheWAS), may help in revealing pleiotropic functions of intermediate phenotypes across multiple tissues. In addition, some intermediate phenotypes are controlled by distant genetic variants, so-called *trans*-QTLs. Here we have also implemented mediation analysis to identify mediators (genes within the locus of the *trans*-QTLs) that potentially modulate downstream gene expression (Chick et al., 2016), and proposed reverse-mediation analysis to reveal potential transcriptional targets.

This multi-layered toolkit is easily accessible through systems-genetics.org, and will expedite the systems dissection of gene function. This will not only provide full leverage of the large historical and rapidly expanding datasets available in the BXD mouse GRP, but will also be universally applicable to any other population.

RESULTS

Structure and Pre-processing of Multi-layer Data from BXD Population

Over the past decades, hundreds of studies on the BXD population have created a wealth of multi-layered omics data, ranging from genomic, transcriptomic, proteomic, metabolomic, to phenomic data (Figure 1A). All the data have been archived and are publicly available in GeneNetwork (www.genenetwork.org/). We focus here on data from 93 BXD strains (including BXD1–BXD102), the parental C57BL/6J and DBA/2J strains, and reciprocal F1 hybrids (i.e., B6D2F1 and D2B6F1), that collectively encompass the vast majority of all BXD data.

Genome

Five million sequence variants segregate in the BXD family (Wang et al., 2016). A phylogenetic tree of the 97 BXD strains, inferred from whole-genome SNP analysis, was used to evaluate family substructure. Several strains have strong genetic similarities, such as BXD48 and BXD48a, which are 93.2% identical by descent (Figure 1C). There is also more subtle genetic similarity among those BXDs (BXD43–BXD102) that were produced by inbreeding advanced intercross progeny (Peirce et al., 2004). We compensated for this kinship in our statistical analyses.

Phenome

Since the first publication on the BXDs (Taylor et al., 1973), well over 200 research groups have generated behavioral, neurological, pharmacological, immunological, and, more recently, metabolic phenotypes, for this family. The size and variety of the BXD phenome has increased exponentially since 2010, to ~5,000 quantitative clinical phenotypes as of December 2016 (Figure S1A; Table S1). We identified three confounding factors that require correction to improve phenome-wide analyses. (1) Since different groups worked with different subsets of the BXDs, variable overlap of strains across traits (missing phenotypic data for subset of strains) is a general problem (Figure 1D). Therefore, data from different groups were analyzed separately. (2) The BXD phenome contains batches of strongly correlated phenotypes, a phenomenon we termed as "phenome linkage" (PL). As an example, multiple measurements of body weight and blood glucose levels over time formed two big PL blocks (Figure 1F) (Andreux et al., 2012). Therefore, the effective number of independent phenotypes (N_{eff}) was used to estimate the significance of phenome-wide association. (3) Although most phenotypes follow an approximately normal distribution (Figure 1G, top), others do not and contain outliers (Figure 1G, bottom). To establish a robust analysis pipeline, we transformed the phenotypes into a standard normal distribution.

Transcriptome, Proteome, and Metabolome

Approximately 200 transcriptome datasets from 34 BXD tissues existed (Figures 1E and S1B; Table S2). One or two datasets were selected to represent the transcript profiles in each tissue (yellow labeled in Table S2). Furthermore, other molecular data in our analyses include ~2,600 liver proteins quantified by SWATH-MS, and ~980 metabolites measured in liver (Williams et al., 2016) and muscle, released with the current study at systems-genetics.org, as well as at GeneNetwork.

In combination, we employed deep phenome data consisting of ~5,000 phenotypic traits, and more than 200 transcriptome, proteome, and metabolome datasets for the BXD GRP (by far the largest coherent multi-omics data assembled for any animal population) as the foundation to identify the genetic architecture underlying complex traits and diseases. Here we integrated these multi-omic data collected over the last decades, and assembled a series of state-of-the-art systems tools (Figure 1B) into a streamlined workflow (Figure 1H) to identify gene function. In the prioritization of PheWAS candidate genes, we included not only genes with high-impact variants (Wang et al., 2016), but also genes that had *cis*-QTLs for transcripts and proteins, since functional effects of genetic variants on phenotypic traits are mediated through both coding and non-coding sequences (Alexander et al., 2010). Genes with *trans*- or *cis*-QTLs could be analyzed using mediation or reverse-mediation analysis, to determine the regulatory mechanisms of gene expression. With expression patterns of target genes in various BXD tissues, it is practical to carry out ePheWAS to reveal associated phenotypic traits. This analytical toolkit and its power to identify potential gene functions are described in detail below.

PheWAS Reveals G2P Associations and Facilitates the Detection of Pleiotropic Effects

Linkage analysis and GWAS have successfully identified gene variants and QTLs associated with complex traits. The same data can also be analyzed in a reverse fashion, i.e., testing the phenotypes that are associated with the gene of interest, using PheWAS (Bush et al., 2016; Denny et al., 2016) enabling the detection of pleiotropic effects of genetic variants (Figure S2A). We recently applied PheWAS to the BXDs using Pearson's correlation (Wang et al., 2016), but this analysis did not account for the non-normality or outliers in the data, or the population substructure among strains. To improve PheWAS, we: (1) transformed all phenotypes to a normal distribution; (2) used linear mixed models to correct for kinship; and (3) adjusted for the PL

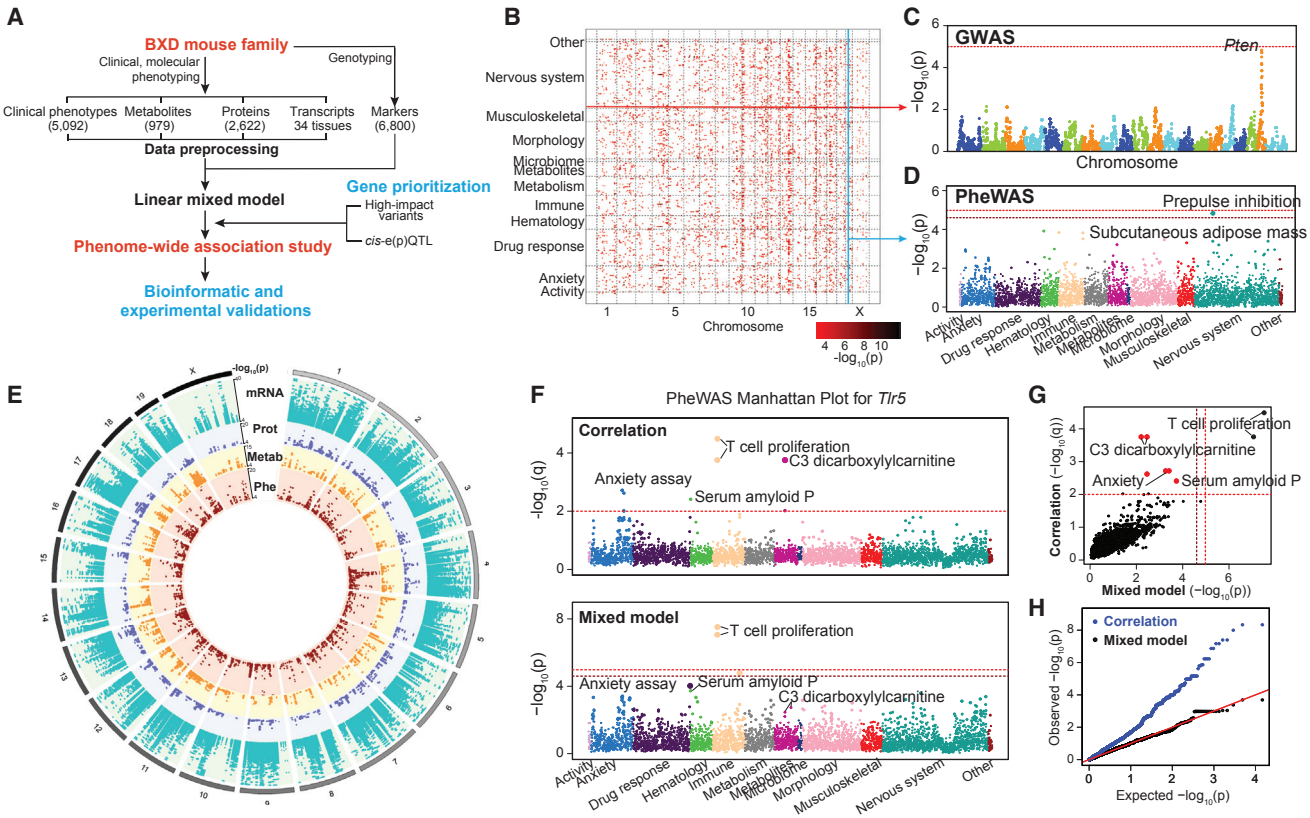


Figure 2. Phenome-wide Association Analysis

(A) Flowchart explaining the steps for PheWAS in the BXD GRP (see text).

(B) Whole-genome PheWAS with genes arranged horizontally based on their genetic locations and phenotypes arranged vertically based on phenotypic categories as defined in the STAR Methods. Significance of the G2P associations is reflected by the color of the dots.

(C) GWAS of prepulse inhibition detected *Pten* as a top candidate gene. Genome-wide significance threshold ($0.05/6{,}800 = 7.4 \times 10^{-6}$) was corrected by the number of tested SNPs.

(D) PheWAS on *Pten* unveiled its association with a list of phenotypes, including prepulse inhibition and heart rate. Phenotypes were arranged and colored according to respective phenotypic categories. Phenome-wide significance was determined based on Bonferroni correction using the total number ($0.05/4{,}784 = 1.0 \times 10^{-5}$, red dashed line) as well as the effective number ($0.05/2{,}754 = 1.8 \times 10^{-5}$, dark red dashed line) of phenotypes.

(E) Circos plot showing all the significant associations of prioritized genes in the multi-layered PheWAS. Genomic positions of genes on chromosomes are labeled on the outer edge, with multi-layered PheWAS data of transcripts (turquoise), proteins (periwinkle), metabolites (orange), and clinical phenotypes (brown) assigned from the outmost to the innermost tracks.

(F) Comparison of the PheWAS results on *Tlr5* from Pearson's correlation (top) and mixed model (bottom), the method employed in this paper. A q value of 0.01 after false discovery rate correction was used as the phenome-wide significance threshold for results from Pearson's correlations. Phenome-wide significance for results from mixed model was determined by Bonferroni correction for the total and effective numbers of phenotypes.

(G and H) Simple correlation results in false-positives. Some significant associations obtained by Pearson's correlation, e.g., C3 dicarboxylylcarnitine levels, are not significant with the mixed model (G), because of the failure in controlling for the population structure, as indicated by the inflated observed p values from correlation analysis in the Q-Q plot (H).

See also Figures S2 and S3, and Table S3.

in the phenome to improve the statistical power of detection. We calculated the effective number of independent phenotypes (N_{eff}) to adjust for the redundancy and to control family-wise error rate in the following analysis (Li and Ji, 2005) (see STAR Methods). This correction estimated that there were ~2,700 effective phenotypes from the ~5,000 initial phenotypes. In total, 4,682 genes with high-impact variants and 9,558 genes with *cis*-QTLs were prioritized (a total of 11,548 genes for PheWAS analysis) (Figure S2B). Associations between the genetic variants of each gene and clinical and molecular phenotypes (transcripts, proteins, and metabolites) were performed using EMMA (Kang et al., 2008). A simplified flowchart representing our updated PheWAS approach is depicted in Figure 2A.

We performed both forward (e.g., GWAS) and reverse genetic approaches (e.g., PheWAS) on genome and phenome data in the BXDs (Figure 2B). For example, GWAS on the prepulse inhibition (PPI) of acoustic startle response mapped a significant signal on Chr 19. Phosphatase and tensin homolog (*Pten*), a gene known to be associated with a wide spectrum of neurodevelopmental diseases stood out as one of the top candidates (Figure 2C). PheWAS for *Pten* revealed several associated traits, including PPI and subcutaneous white

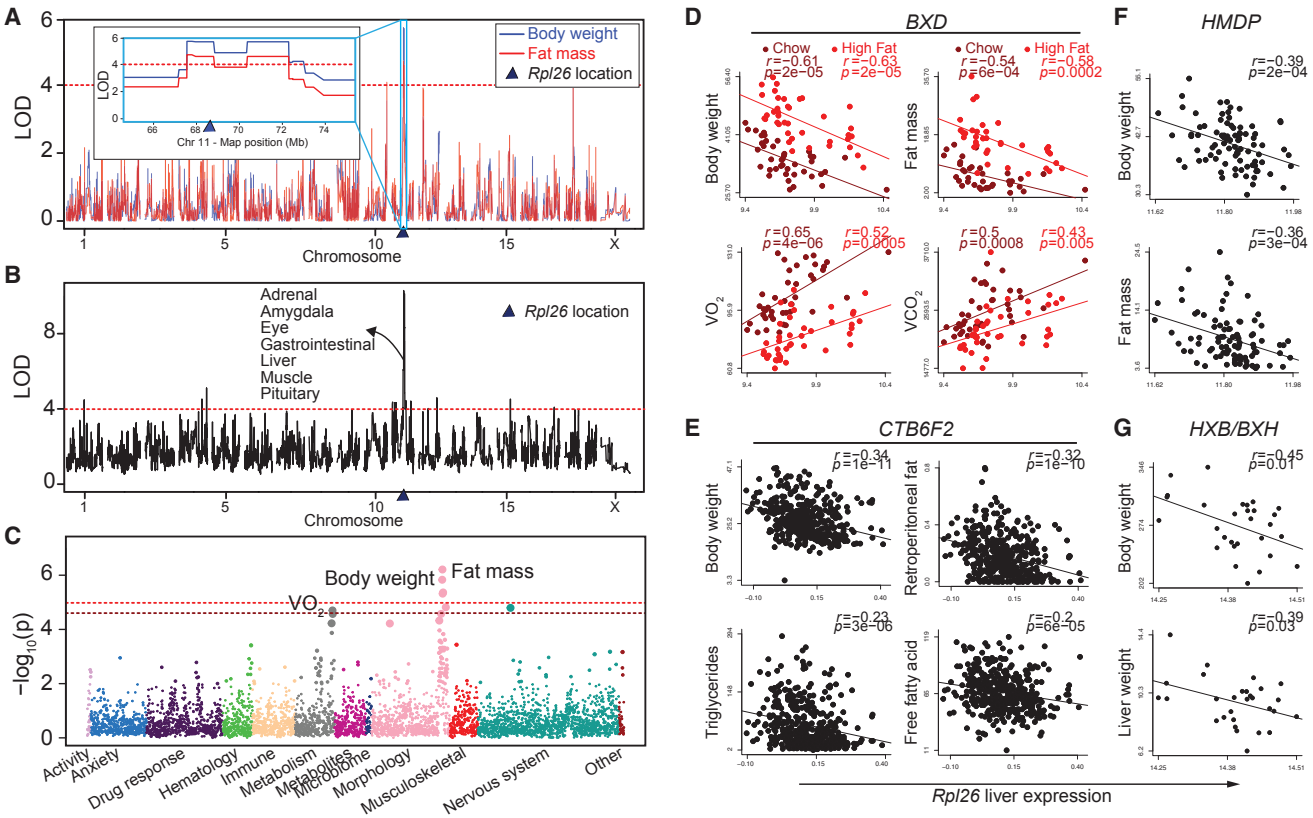


Figure 3. Genotype-Phenotype Associations Revealed by PheWAS

(A) QTL mapping of body weight and fat mass showed a common QTL on Chr 11, where *Rpl26* locates (indicated by a blue triangle).

(B) *Rpl26* possesses *cis*-eQTLs in the tissues listed.

(C) PheWAS reveals the genetic association between *Rpl26* and metabolic traits. Phenome-wide significance was determined as in Figure 2D.

(D) *Rpl26* liver transcripts correlate with a series of metabolic traits, such as body weight, fat mass, VO_2, and VCO_2.

(E–G) Data from CTB6F2 (E) and HMDP (F) mouse cohorts, and the HXB/BXH rat cohort (G) indicate significant negative correlations between liver *Rpl26* levels and body weight, and other metabolic traits.

adipose tissue (subWAT) mass (Figure 2D), suggesting pleiotropic effects of *Pten*. The links between *Pten* and neurobiological and metabolic phenotypes have been confirmed by independent studies (Kwon et al., 2006; Ortega-Molina et al., 2012). Overall, PheWAS showed that 4,230 out of 11,548 genes were associated with at least one phenotypic trait and all genes had significant associated molecular traits after phenome-wide correction (Figures 2E; Table S3).

We compared the performance of the original and updated PheWAS methods (Wang et al., 2016), taking *Tlr5* (Toll-like receptor 5) as an example (Figure 2F). Both methods associated *Tlr5* with T cell proliferation, befitting its known function in immune response (Caron et al., 2005). However, our initial method yielded C3 dicarboxylylcarnitine, anxiety assay, and serum amyloid P component as false-positives (Figure 2G), due to the failure to control for population structure, as indicated by the inflation of p values in the Q-Q plot (Figure 2H).

A few more examples illustrate the use of PheWAS in revealing G2P associations. Obesity/overweight is a global health problem and a leading risk factor for diabetes, cardiovascular diseases, and cancer. We found a QTL for body weight and fat mass on Chr 11 (Figure 3A). From this region, *Rpl26* (ribosomal protein L26) stood out as a strongest candidate with *cis*-eQTLs in many tissues, including liver (Figure 3B). Through PheWAS, we identified a link between *Rpl26* and body weight, fat mass, as well as oxygen consumption (VO_2) (Figure 3C). The genetic association was confirmed by the negative correlations of *Rpl26* liver transcripts with these metabolic traits in BXDs on both chow (CD) (GeneNetwork Accession: GN432) and high fat diet (HFD) (GN431) (Figure 3D), and further validated in several independent datasets, including an F2 cross between CAST/EiJ and C57BL/6J (CTB6F2, GN172) (Schadt et al., 2008) (Figure 3E), the Hybrid Mouse Diversity Panel (HMDP) (Bennett et al., 2015) (Figure 3F), as well as in the HXB/BXH rat cohort (Hubner et al., 2005) (Figure 3G). The correlations of *Rpl26* with metabolic traits translate well across populations and species, and suggest a role of *Rpl26* in regulating body weight.

Through PheWAS, we also confirmed the link between *Oprm1* (opioid receptor, mu 1) and morphine response (Uhl et al., 1999) (Figures S3A and S3B). A nonsynonymous variant (rs8256412) in *Oprm1* associated with morphine response traits as well as the *Oprm1* expression in neural tissues, including hippocampus (GN110) and ventral tegmental area (VTA, GN228). Further evidence was provided by the negative correlations of *Oprm1*

with locomotion activity after morphine injection in the BXDs (Figure S3C).

Expression-Based PheWAS: A Tool to Discover Gene Functions

Despite the success of GWAS and PheWAS to uncover novel genetic variants associated with complex traits and diseases, these variants only explain a limited proportion of the heritability of the phenotypic traits (Manolio et al., 2009). Intermediate phenotypes, including transcript and protein levels, integrate the effects from genetic factors, including those poorly captured or hidden in common association studies (Gagneur et al., 2013), as well as those from environmental factors. A few recent studies have explored the use of transcriptome-/proteome-wide association using either imputed transcript expression (Gusev et al., 2016; Mancuso et al., 2017) or proteomic data (Okada et al., 2016). Given that transcriptome data are available for over 30 tissues, the BXDs are a perfect resource for such analysis. A linear mixed model was applied to find associations between gene expression and clinical phenotypes while accounting for population structure across strains (Kang et al., 2008) (Figure 4A). Forward genetics strategies could link phenotypes to tissue-specific transcript levels in T/PWAS (Figure 4B, red line). Conversely, reverse approaches starting from expression of the gene of interest toward the phenome, i.e., ePheWAS, could reveal the gene's potential pleiotropic functions, especially when considering its expression across multiple tissues (Figure 4B, blue line). The numbers of G2P associations that survive the phenome-wide significance threshold differed across tissues and across phenotypic categories (Figure 4C). For example, phenotypes from the "Morphology" category were enriched in brown adipose tissue and liver, while phenotypes from the "Drug response" and "Nervous system" categories were more correlated with genes from hippocampus and hypothalamus. These data coincide with the results from human studies (Emilsson et al., 2008), suggesting that many phenotypic traits are under tissue-specific regulations.

TWAS in liver (GN432) identified *Slc25a10* as a potential regulator for VO_2 max (Figures 4B and 4D). *Slc25a10* exports malonate, malate, and succinate across the mitochondrial inner membrane for fatty acid synthesis in the cytosol (Mizuarai et al., 2005). Through ePheWAS, we found that *Slc25a10* not only associated with VO_2, but also with body weight and fat mass (Figures 4B and 4E). Furthermore, *Slc25a10* liver expression correlated positively with body weight, fat mass, and sub-WAT mass, and negatively with VO_2 in both CD and HFD fed BXDs (Figure 4F). We found comparable correlations with similar metabolic traits in the CTB6F2 (Figure 4G) and the HMDP (Figure 4H), corroborating the role of *Slc25a10* as a dicarboxylate carrier.

Fasting is an efficient way to induce weight loss; however, its effects vary across populations, suggesting potentially genetic influences (Wing and Hill, 2001). There are notable differences in weight loss after an overnight fast across the BXDs (ranging from 0.8 to 3.9 g on CD and 0.3 to 3.5 g on HFD), although there was no significant difference between CD and HFD cohorts (Figure 5A). However, no genetic variant was found to be associated with fasting weight loss using QTL mapping (Figure 5B). Through TWAS using liver transcripts, we detected *Cpt1a* as the top candidate associated with fasting weight loss in both CD (Figure 5C) and HFD cohorts (data not shown). ePheWAS showed further associations of *Cpt1a* with plasma acylcarnitine levels (Figure 5D). Liver *Cpt1a* levels correlated positively with acylcarnitines (Figure 5E, upper), which corresponds to the recognized function of *Cpt1a* in transferring the acyl group of long-chain fatty acyl-CoA to carnitine for further β-oxidation in mitochondria (Pande, 1975). Strains with higher *Cpt1a* expression tend to lose more weight upon fasting, and have lower plasma triglycerides (Figure 5E, bottom). Furthermore, we validated the correlations between *Cpt1a* and metabolic phenotypes in another independent mouse population, i.e., the HMDP (Figure 5F), and in *C. elegans*, where feeding an RNAi targeting *cpt-1*, the *Cpt1a* worm homolog, lowered lipid content (Figure 5G); this highlights the cross-species conservation of *Cpt1a*'s role in lipid metabolism.

Using ePheWAS, we also identified associations between *Cd36* liver transcripts and fat mass and acid β-glucosidase activity (Figure S4A). BXD and HMDP mouse strains, as well as HXB/BXH rat strains, with higher *Cd36* expression had increased fat mass and body weight, as well as decreased VO_2 and liver acid β-glucosidase activity (Figures S4B and S4C), confirming the involvement of *Cd36* in metabolism (Silverstein and Febbraio, 2009) and suggesting a potential role in Gaucher's disease, which results from the deficiency of acid β-glucosidase (Grabowski, 2008). An association between *Abca8a* liver transcripts and triglyceride levels was also revealed (Figure S4D). Increased liver *Abca8a* levels correlated with the increase of plasma triglycerides, free fatty acid, cholesterol, glucose levels, and fat mass, as well as lower plasma acylcarnitine levels in the BXD, HMDP, and HXB/BXH GRPs (Figures S4E and S4F). This substantiates a role for this poorly characterized ABCA protein in lipid transport, similar to many other ABCA transporters (Dean et al., 2001).

Evaluation of PheWAS and ePheWAS in Detecting Associations

We observed that datasets with a larger cohort size tend to have more power in detecting G2P associations (Figures 6A and 6D; Table S3). To test the influence of cohort size on the number of significant associations and to estimate the robustness of associations detected by PheWAS or ePheWAS, we used a subsampling approach on the actual BXD data. The eye transcriptome dataset, which has the largest cohort size of 72 strains, was used as an illustration to detect the phenome-wide association signals against the BXD genome. We randomly sampled subset cohorts with different sizes and then performed association analysis on each set. Then we calculated the number of recovered hits: the significant associations that are common between each random subsample and the full set. The total number of detected PheWAS hits (blue curve, Figure 6B) linearly increased with the number of strains sampled; so did the number of recovered hits (red curve, Figure 6B). In all subsamples, more than ~75% of the hits are recovered hits, which implies that the associations are robust (Figure 6B). We also assessed the robustness of the associations by comparing the significant hits obtained from subsamples of the same size. As expected, simulated cohorts with larger size had relatively high probability to detect the same G2P association signals (Figure 6C). Interestingly, subsamples of as few as 20 strains

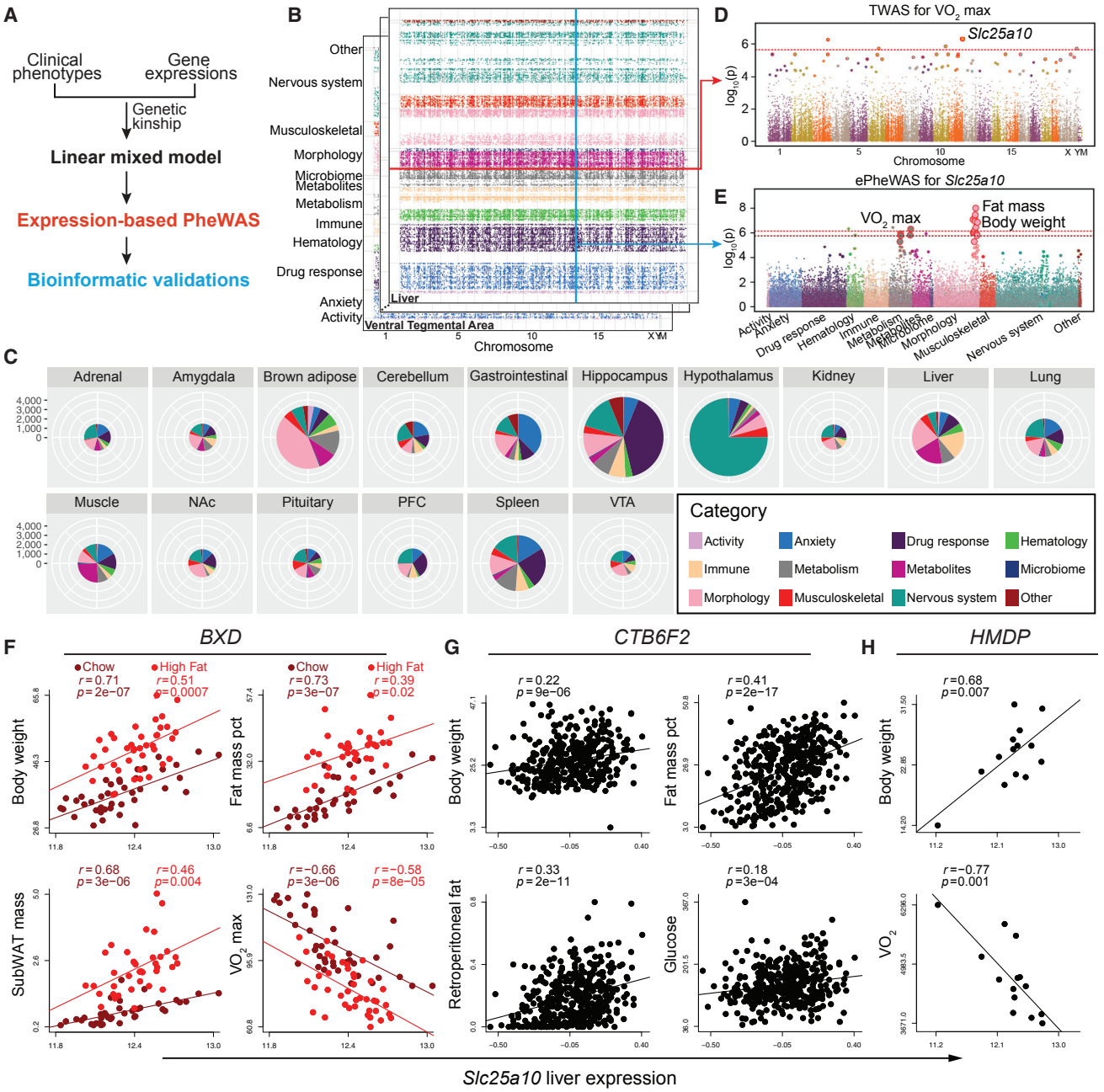


Figure 4. ePheWAS Displays Tissue-Specific Regulators

(A) Flowchart explaining the steps for ePheWAS (see text).

(B) Whole-transcriptome ePheWAS scheme showing the complementary findings of TWAS. All significant associations are displayed, with genes arranged horizontally based on their genetic locations and phenotypes arranged vertically based on phenotypic categories. Phenotypes from each category are labeled with the corresponding color as in (C and E). Major gaps in the plot are due to the limited numbers of phenotyped strains with expression data available.

(C) Statistical summary of significant ePheWAS associations across 16 major tissues. The number of identified significant associations in each tissue is represented by pie plots, with phenotypes from each category indicated by their respective colors. Muscle, gastrocnemius muscle; NAc, nucleus accumbens; PFC, prefrontal cortex.

(D) TWAS identifies *Slc25a10* as the best candidate to explain changes in VO_2 max across the BXDs. Transcripts were arranged by their genetic location. Transcriptome-wide significance ($0.05/25{,}000 = 2 \times 10^{-6}$) was adjusted by the number of transcripts tested in the analysis.

(E) ePheWAS of *Slc25a10* reveals its pleiotropic functions on fat and body mass, as well as VO_2. Phenome-wide significance was adjusted by Bonferroni correction for the number of used tissues, 16 major tissues as listed in (B), together with the total number ($0.05/4{,}784/16 = 6.5 \times 10^{-7}$), as well as the effective number of phenotypes ($0.05/2{,}754/16 = 1.1 \times 10^{-6}$), indicated by red and dark red dashed line, respectively.

(F–H) Liver *Slc25a10* transcripts correlate with relevant metabolic phenotypes, such as body weight, fat mass, VO_2, in the BXD (F), CTB6F2 (G), and HMDP (H) cohorts. See also Figure S4.

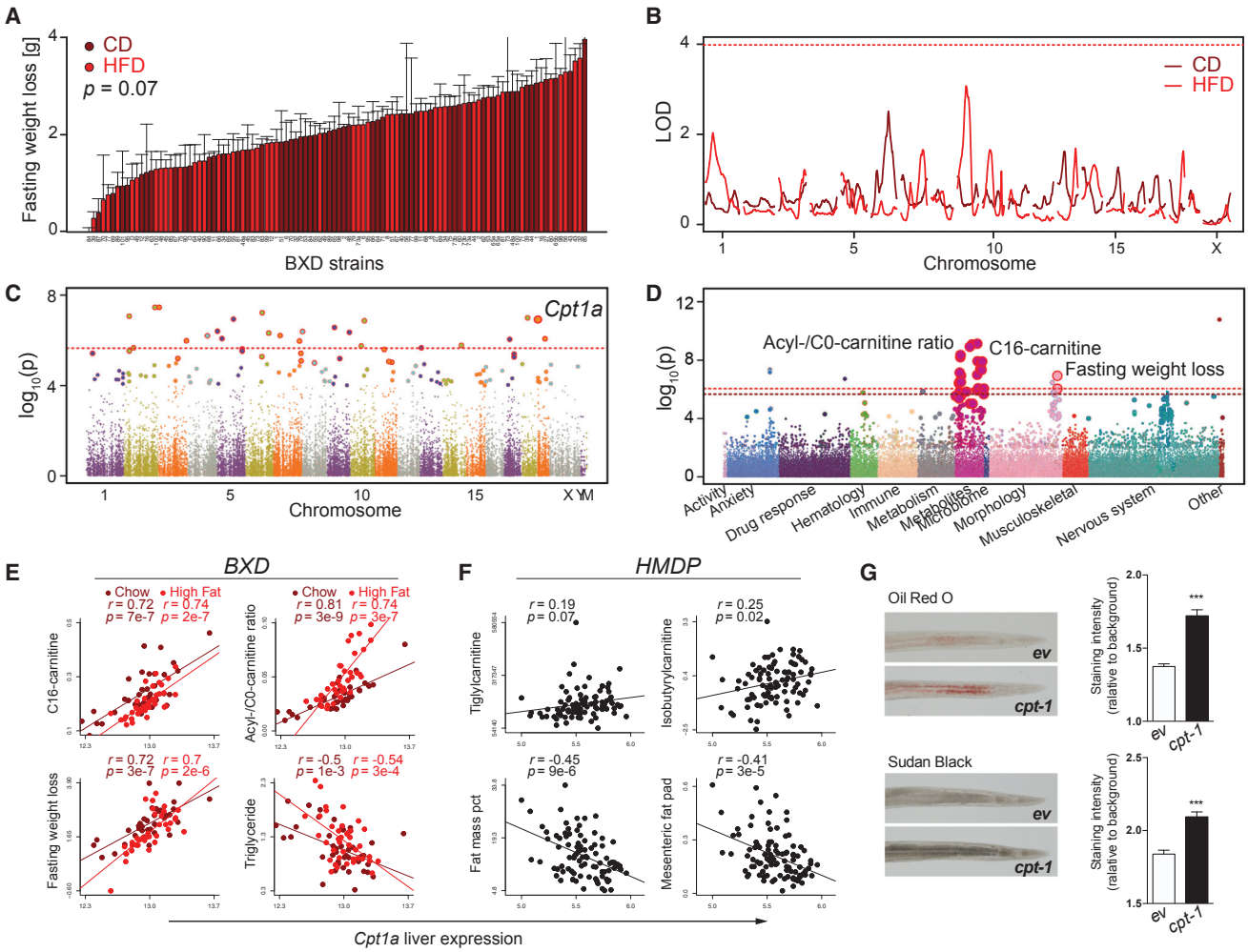


Figure 5. ePheWAS Reveals *Cpt1a* as a Regulator of Fasting Weight Loss

(A) Body weight loss upon fasting across the BXDs fed with either chow (CD, dark red) or high fat diet (HFD, red). Error bars represent mean ± SEM.

(B) Genetic mapping failed to detect significant QTLs for fasting weight loss in both CD and HFD cohorts.

(C) TWAS for fasting weight loss in liver in CD fed BXD mice. Transcriptome-wide significance was adjusted as in Figure 4D.

(D) ePheWAS for *Cpt1a* identified its association with carnitine levels and fasting weight loss. Same as Figure 4E, phenome-wide significance was adjusted by Bonferroni correction for the numbers of used tissues and phenotypes.

(E and F) Correlations between liver *Cpt1a* expression and metabolic phenotypes, including carnitine levels, fasting weight loss, triglycerides, and fat mass, in the BXD (E) and HMDP (F) populations.

(G) Knock down of *cpt-1*, the *C. elegans* ortholog of *Cpt1a*, leads to the accumulation of lipid droplets, revealed by staining with oil red O or Sudan black. Data are represented as mean ± SEM. *Ev*, empty vector. ***p < 0.001.

share ~40%–50% of their associations. We also performed subsampling analysis on ePheWAS looking for significant associations between gene expression in the eye and the clinical phenome, and observed a similar influence of sample size on performance (Figures 6E and 6F). However, there was a more rapid reduction of true-positives with decreasing sample size compared with PheWAS (Figure 6B), mainly due to the incompleteness of the phenotype data (i.e., different laboratories sampling different lines, shown in Figure 1D).

Overall, the almost linear dependence between the sample size and number of significant hits suggests that while the current BXD cohort sizes enable the detection of robust associations, larger cohorts can identify even more G2P associations.

Mediation Analysis Identifies Regulatory Mechanism of Gene Expression

The regulation of transcript and protein abundance is crucial for cellular, and organismal homeostasis. Mediation analysis was developed to identify the mediating effects of a mediator between an independent variable and a dependent variable (MacKinnon et al., 2007). This concept has also been applied to reveal the mediating role of gene expression in the association between SNPs and clinical phenotypic variations (Yao et al., 2017) or *trans*-regulated genes (Chick et al., 2016; Pierce et al., 2014).

We first identified QTLs for all genes in the transcriptome and proteome datasets across all tissues. The number of *cis*- and *trans*-QTLs varied across tissues, gender and treatments

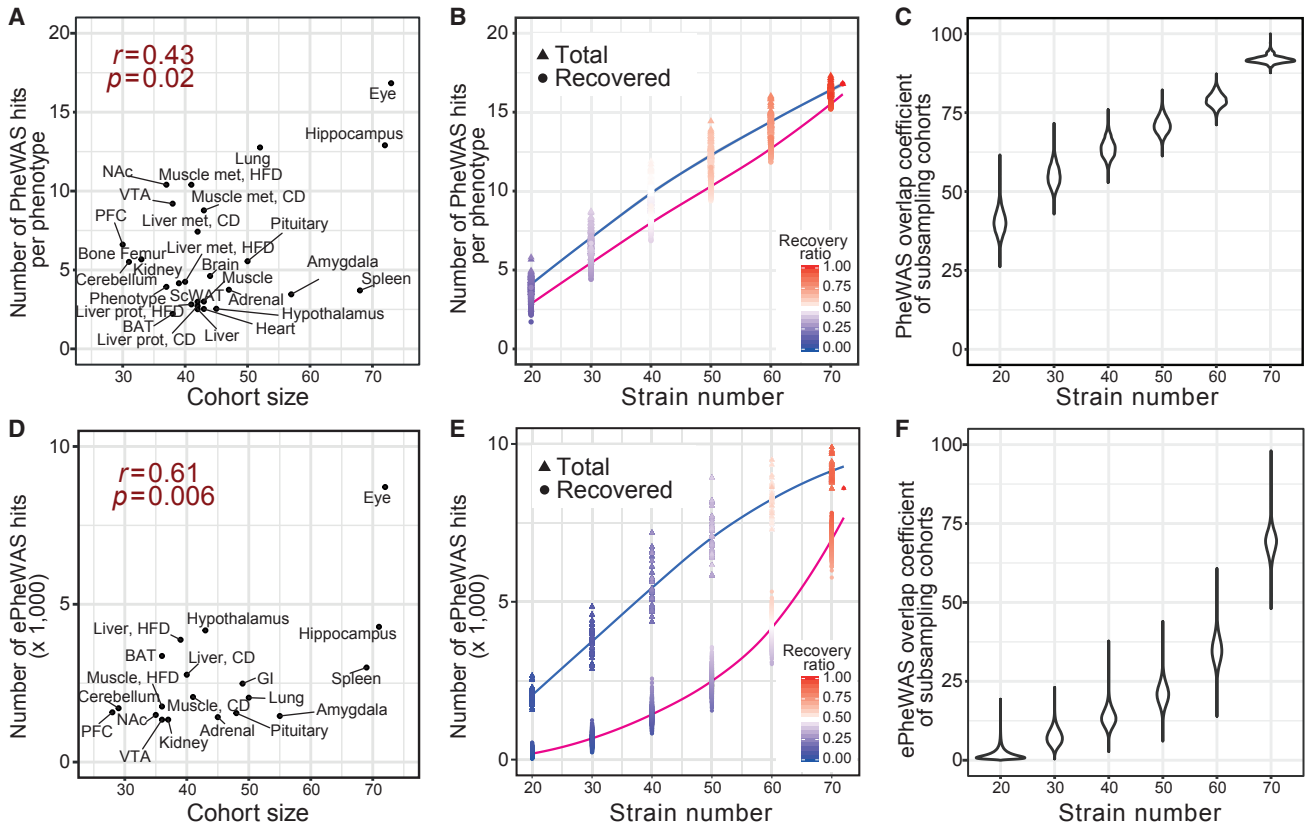


Figure 6. Performance of PheWAS and ePheWAS in Detecting G2P Associations

(A and D) Correlations between cohort sizes of each omics dataset versus the numbers of significant PheWAS hits normalized per phenotype (A) or the numbers of significant ePheWAS hits (D). BAT, brown adipose tissue. GI, gastrointestinal; Liver met, liver metabolite; Liver prot, liver protein; Muscle met, muscle metabolite.

(B, C, E, and F) Random subsampling analysis from the 72 strains of the eye transcriptome dataset to investigate the performance of PheWAS (B and C) and ePheWAS (E and F).

(B and E) The influence of strain number on the number of total detected (triangles) as well as recovered (circles) PheWAS (B) or ePheWAS (E) hits was revealed through random subsampling. The recovered ratio in detecting "real" significant associations was also indicated.

(C and F) Overlap coefficient of PheWAS (C) or ePheWAS (F) associations between subsampling subset cohorts of the same size.

(Figure 7A), suggesting that the modulation of gene expression is tissue and environment specific (Dimas et al., 2009; Grundberg et al., 2012). For example, the eye transcriptome shows a *trans*-eQTL hotspot on Chr 1. *Pou2f1*, a gene involved in lens placode development (Donner et al., 2007), has a strong *cis*-QTL in this locus (indicated by arrow) and could explain this tissue-specific *trans*-eQTL hotspot (including *Atf4*, *Faim2*, *Fkbp1b*, *Gab1*, etc.) in the eye. We applied mediation analysis on the transcriptome and proteome datasets to elucidate the genetic modulation of gene expression. The concept of mediation allows the application of two reciprocal approaches. Mediation starts from the dependent variable (a gene with a *trans*-QTL) and aims to find its mediator (a gene with a *cis*-QTL in the locus of the *trans*-QTL) (Figure 7B). While, on the contrary, reverse mediation investigates the mediated variables of a potential mediator (gene with a *cis*-QTL) (Figure 7I).

The power of mediation analysis was illustrated by using the BCKDHA protein as an example. Together with BCKDHB, BCKDHA composes the branched-chain alpha-keto acid dehydrogenase (BCKD) E1 complex that breaks down branched-chain amino acids. BCKDHA protein levels in liver (GN704) mapped a *trans*-pQTL on Chr 9, the same locus of the *Bckdhb* and BCKDHB *cis*-pQTL (Figure 7C). Mediation analysis revealed that BCKDHB is a potential mediator of BCKDHA protein levels (Figure 7D). The mediation results were further confirmed in the liver protein dataset (GN705) from BXDs fed with HFD (Figure S5), as well as in the diversity outbred (DO) mouse cohort (Chick et al., 2016) (Figures 7E and 7F). BCKDHB correlated with BCKDHA protein levels in both BXD (Figure 6G) and DO cohorts (Figure 6H). This demonstrates that the mediation effect of BCKDHB on BCKDHA is independent of environmental influences (e.g., diet), conserved across populations, and most likely the consequence of the impaired assembly of the BCKDH complex when the protein abundance of BCKDHB reduces.

Mediation analysis was also performed on *Rpsa* and *Rps2*, components of the 40S ribosomal subunits, to determine the upstream regulation factors (Figure S6). *Rpsa* and *Rps2* have *trans*-eQTLs on Chr 12 in many tissues, including brain and hippocampus (Figures S6B and S6E), suggesting a shared regulation. Mediation revealed *Zfp277* as the potential regulator of *Rpsa* and *Rps2* (Figures S6C and S6F). *Zfp277* strongly co-expressed with *Rpsa* and *Rps2* (Figures S6D and S6G). Furthermore, the

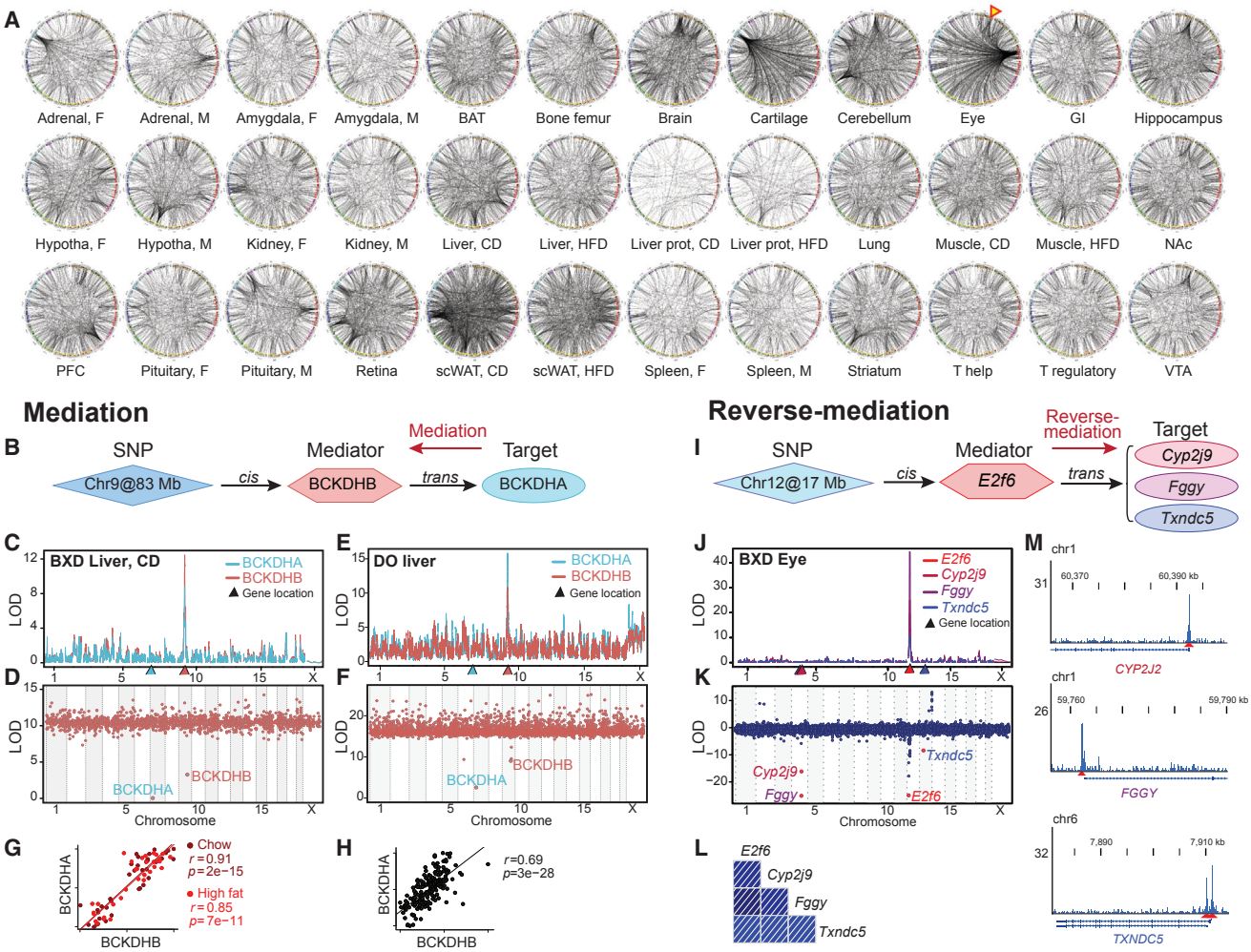


Figure 7. Mediation and Reverse-Mediation Analysis Discovers Gene Interactions

(A) Circos plots showing the QTLs in transcriptome and proteome datasets from BXDs with different sex (F, female; M, male) or diets (CD, chow; HFD, high fat diet) across tissues. Each transcript dataset is represented by a single circos plot. *trans*-QTLs are illustrated by curves connecting the genetic loci of these genes and their respective *trans*-QTLs, with arrows pointing to the *trans*-QTLs. The position of *Pou2f1* and the *trans*-eQTL hotspot mapping to its location in eye transcriptome data is indicated by an arrow. Hypotha, hypothalamus.

(B) Conceptual scheme of mediation analysis. The causal SNP, mediator, and dependent variable (target) are represented in rhombus, hexagon, and oval, respectively. The mediator of the dependent variable (gene with *trans*-QTL) can be identified by mediation analysis. The red arrow shows the direction of mediation analysis, i.e., from the target to find the potential mediator. As an example, the *trans*-pQTL of BCKDHA acts through affecting the BCKDHB protein level in *cis*.

(C and E) pQTL mapping of BCKDHA and BCKDHB in livers from CD (C) fed BXD mice, and DO mice (E). BCKDHA exhibits a *trans*-pQTL that maps on Chr 9, where the BCKDHB *cis*-pQTL locates.

(D and F) Mediation plot of BCKDHA in liver proteomic datasets showing that BCKDHB is a mediator of BCKDHA.

(G and H) Significant correlation between BCKDHA and BCKDHB protein levels in livers of either CD or HFD fed BXDs, as well as in the DO mice (Chick et al., 2016).

(I) Conceptual scheme of reverse-mediation analysis. The dependent variable (target) of a given mediator (gene with *cis*-QTL) can be detected using reverse-mediation analysis. The red arrow shows the direction of mediation analysis, i.e., from the mediator to find the potential targets. As an example, the genetic variant underlying the *cis*-eQTL of *E2f6* influences the expression of *Cyp2j9*, *Fggy*, and *Txndc5* in *trans*.

(J) eQTL mapping of *E2f6* transcript levels and some potential *E2f6* transcriptional targets, including *Cyp2j9*, *Fggy*, and *Txndc5* in transcriptome from BXD eye (GN207). All these target genes map *trans*-QTLs in the same locus of the *E2f6* *cis*-eQTL.

(K) Reverse-mediation plot of *E2f6* showing its mediation effects on *Cyp2j9*, *Fggy*, and *Txndc5*, which are pulled down from the background in the plot.

(L) Correlation between the expression of *E2f6* and its target genes in the eye.

(M) Binding of E2F6 on the promoter of the human orthologs of the mouse *E2f6* target genes in human ENCODE (indicated in blue). Chromosome numbers relate to human chromosomes. The predicted binding site is indicated in red.

See also Figures S5–S7.

mediating role of *Zfp277* on *Rpsa* and *Rps2* was confirmed using data from prefrontal cortex (Figures S6H–S6J, GN130) and brain (Figures S6K–S6M, GN784) of the LXS GRP (Williams et al., 2004). As all three genes have been linked to cancer, we then tested whether they co-express in cancer. Based on RNA sequencing data from 35 cancer types in The Cancer Genome Atlas (TCGA), *ZNF277* positively correlated with expression of *RPSA* and *RPS2* and the majority of the ribosomal protein family (Cerami et al., 2012; Gao et al., 2013) (Figure S6N), suggesting a potential role of *ZNF277-RPSA/RPS2* pathway in cancer.

Because transcription factors (TFs) regulate the expression of distal genes, we described reverse-mediation analysis, a strategy to validate the transcriptional regulation of target genes by a given TF *in silico* (Figure 7I). *E2f6* is a known TF with a *cis*-eQTL in the eye (GN207). A large number of genes, including *Cyp2j9*, *Fggy,* and *Txndc5*, also exhibited *trans*-eQTLs in the locus of *E2f6* on Chr 12 (Figure 7J). Reverse mediation revealed the mediating role of *E2f6* on the expression of these genes (Figure 7K). In addition, *E2f6* transcripts positively correlated with these genes in the eye (Figure 7L). This finding led to the hypothesis that *E2f6* binds to the regulatory regions of these genes, which was confirmed by human chromatin immunoprecipitation sequencing (ChIP-seq) data from the Encyclopedia of DNA Elements (ENCODE) (ENCODE Consortium, 2012) (Figure 7M), illustrating the cross-species translational value of studies in the BXDs.

Reverse mediation also exposed the transcriptional regulation of *Zkscan1* on its potential targets, e.g., *Adam10*, *Atl2, Phf3,* etc., in the hippocampus (GN110) (Figures S7A–S7C). These target genes showed *trans*-eQTLs mapping to the genetic locus of *Zkscan1* (Figure S7B), and were tightly co-expressed with *Zkscan1* (Figure S7D). ENCODE ChIP-seq data confirmed the binding of ZKSCAN1 on the promoter region of the human orthologs of the identified candidates (ENCODE Consortium, 2012) (Figure S7E).

DISCUSSION

In this study, we developed, applied, and validated a series of systems approaches (including PheWAS, T/PWAS, ePheWAS, mediation, and reverse-mediation analysis) using multi-omic datasets from the BXD mouse population. We provide examples of each approach to predict gene function across layers of multi-omics data, by focusing on complex metabolic traits. All the data and analysis tools are archived in the open-access systems genetics resource webpage (systems-genetics.org).

Compared with the original PheWAS methodology in mouse (Wang et al., 2016), we developed a more robust strategy by applying a mixed-model approach on normalized phenotypic traits and by considering high-impact genetic variants from both coding and non-coding regions. The effective number of independent phenotypes based on PL was applied to more accurately estimate the significance threshold based on permutation testing (Sham and Purcell, 2014). However, since different groups used different subsets of BXD strains for phenotyping, there are missing gaps in the data, leading to an inability to fully account for PL. Therefore, we expect the effective number of phenotypes (N_{eff}) to be lower than our estimate, implying that our phenome-wide significance threshold ($0.05/N_{eff}$) may be too conservative.

mRNA and protein are the integrators of intrinsic genetic differences and external environmental factors. In many cases, such intermediate phenotypes may be even better predictors of complex traits than genetic variation *per se*, and therefore allow the identification of G2P associations that are not evident through classical approaches. T/PWAS were used to discover the association between traits and transcripts or proteins levels using a forward genetics approach. In addition, we introduced ePheWAS, the reverse approach to T/PWAS, to identify phenotypes associated with expression of the gene of interest. Researchers interested in a certain gene can quickly investigate the relationship between its expression levels in certain tissues and a wide range of phenotypes.

e(p)QTL analyses allow the integration of genetic information with expression levels. While they can be useful in detecting *cis*- and *trans*-genetic associations, they cannot infer causality between genes. We tackled this issue by implementing mediation analysis (Chick et al., 2016), an efficient way to determine the mediators of genes with *trans*-QTLs. Reverse-mediation analysis, as an inverse approach, investigates potential mediated genes by designated mediators (genes with *cis*-QTLs). One can exploit the mediating effects through (reverse-) mediation to infer the most probable route from genetic variants to gene expression levels.

Despite the success in revealing gene functions through applying our suite of systems tools on data collected from the BXD cohort, this population possesses some inherent disadvantages. By a random sampling analysis, we revealed that cohorts with larger size tend to have better performance in detecting (e)PheWAS associations. The relatively small cohort size and limited genetic variance and recombination across the BXD strains are in fact limiting factors. However, this disadvantage is offset by the tight control of the experimental conditions during phenotyping and sample collection. Moreover, our analytical approaches will be powerful on other genetic reference panels, including those in yeast, worm, fly, mouse, rat, and plants, where environmental confounding factors could be well controlled.

Human cohorts or cohorts from other species, which are larger and have a higher genetic diversity, e.g., GTEx (GTEx Consortium, 2015), or TCGA (Cerami et al., 2012), or the 1001 Genomes Project for *A. thaliana* (1001 Genomes Consortium, 2016), may even be better suited for similar analyses. In the case of humans, however, it is almost impossible to simultaneously phenotype individuals and sample multi-tissue and multi-omic data, while controlling the environmental sources of variation. Assessing the use of these tools may require cohorts that have extensive multi-omics datasets available or have relevant samples biobanked, e.g., the Framingham Heart Study (Mahmood et al., 2014). Imputation of gene expression in deep tissues from either reference transcriptome datasets (Gamazon et al., 2015) or GWAS summary statistics (Gusev et al., 2016) could be used to facilitate the applications of our tools, especially ePheWAS, in such human cohorts.

Altogether, this integrated systems genetics toolkit, which is freely accessible on systems-genetics.org, can expedite *in silico* hypothesis generation and testing, facilitating the identification and validation of new gene functions and gene networks in populations, which generally are robust and translate well across

populations and species, unlike many connections seen in classic loss-of-function studies.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - *C. elegans* Lines
- METHOD DETAILS
  - BXD Multi-Omics Datasets
  - Data from Other Mouse and Rat Populations
  - The Cancer Genome Atlas Data
  - The Encyclopedia of DNA Elements Data
  - *C. elegans* RNAi and Lipid Staining
  - Statistical Analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and three tables and can be found with this article online at https://doi.org/10.1016/j.cels.2017.10.016.

AUTHOR CONTRIBUTIONS

Conceptualization, H.L. and J.A.; Methodology, H.L., D.R., M.B.S., Z.K., S.M., R.W.W., and J.A.; Software, H.L., and T.L.; Validation, H.L., X.W., V.S., H.Z., and P.L.; Formal Analysis, H.L., D.R., S.M., and J.A.; Resources, H.L., Q.H., V.S., T.L., H.Z., D.A., A.M., N.Z., L.A.V.-V., K.G., K.S., P.P., R.A.R., S.M., R.W.W., and J.A.; Data Curation, H.L., R.W.W., and J.A.; Writing – Original Draft, H.L., and J.A.; Writing – Review & Editing; H.L., X.W., R.W.W., and J.A.; Visualization, H.L., and J.A.; Supervision, J.A.; Project Administration and Funding Acquisition, Z.K., K.S., R.W.W., and J.A.

ACKNOWLEDGMENTS

We are grateful to the BXD community for generating the valuable resource for systems biology research and thank A.J. Lusis, G.A. Churchill, S.P. Gygi, M. Miles, M. Pravenec, and T.J. Aitman for making data from the CTB6F2 and HMDP, the DO, the LXS, and the HXB/BXH GRPs available. We thank the entire J.A. lab for comments and discussions. H.L. is the recipient of a doctoral scholarship from the China Scholarship Council (CSC). This work was supported by grants from the École Polytechnique Fédérale de Lausanne, the Swiss National Science Foundation (31003A-140780), the Velux Stiftung, the Kristian Gerhard Jebsen Foundation; the AgingX program of the Swiss Initiative for Systems Biology (51RTP0-151019), and the NIH (R01AG043930, R01AA016957).

Received: May 5, 2017
Revised: August 31, 2017
Accepted: October 25, 2017
Published: November 29, 2017

REFERENCES

1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell *166*, 481–491.

Aitman, T.J., Boone, C., Churchill, G.A., Hengartner, M.O., Mackay, T.F., and Stemple, D.L. (2011). The future of model organisms in human disease research. Nat. Rev. Genet. *12*, 575–582.

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010). Annotating non-coding regions of the genome. Nat. Rev. Genet. *11*, 559–571.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881–888.

Andreux, P.A., Williams, E.G., Koutnikova, H., Houtkooper, R.H., Champy, M.F., Henry, H., Schoonjans, K., Williams, R.W., and Auwerx, J. (2012). Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. Cell *150*, 1287–1299.

Bennett, B.J., Davis, R.C., Civelek, M., Orozco, L., Wu, J., Qi, H., Pan, C., Packard, R.R., Eskin, E., Yan, M., et al. (2015). Genetic architecture of atherosclerosis in mice: a systems genetics analysis of common inbred strains. PLoS Genet. *11*, e1005711.

Breiman, L. (2001). Random forests. Machine Learn. *45*, 5–32.

Broman, K.W., Wu, H., Sen, S., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics *19*, 889–890.

Bush, W.S., Oetjens, M.T., and Crawford, D.C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat. Rev. Genet. *17*, 129–145.

Caron, G., Duluc, D., Fremaux, I., Jeannin, P., David, C., Gascan, H., and Delneste, Y. (2005). Direct stimulation of human T cells via TLR5 and TLR7/8: flagellin and R-848 up-regulate proliferation and IFN-gamma production by memory CD4+ T cells. J. Immunol. *175*, 1551–1557.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

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.

Cook, D.E., Zdraljevic, S., Roberts, J.P., and Andersen, E.C. (2017). CeNDR, the *Caenorhabditis elegans* natural diversity resource. Nucleic Acids Res. *45*, D650–D657.

Dean, M., Rzhetsky, A., and Allikmets, R. (2001). The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. *11*, 1156–1166.

Denny, J.C., Bastarache, L., and Roden, D.M. (2016). Phenome-wide association studies as a tool to advance precision medicine. Annu. Rev. Genomics Hum. Genet. *17*, 353–373.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246–1250.

Donner, A.L., Episkopou, V., and Maas, R.L. (2007). Sox2 and Pou2f1 interact to control lens and olfactory placode development. Dev. Biol. *303*, 784–799.

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A., and Kruglyak, L. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature *464*, 1039–1042.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

ENCODE Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Flint, J., and Eskin, E. (2012). Genome-wide association studies in mice. Nat. Rev. Genet. *13*, 807–817.

Flint, J., and Mackay, T.F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res. *19*, 723–733.

Gagneur, J., Stegle, O., Zhu, C., Jakob, P., Tekkedil, M.M., Aiyar, R.S., Schuon, A.K., Pe'er, D., and Steinmetz, L.M. (2013). Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. PLoS Genet. *9*, e1003803.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L.,

Cox, N.J., and Im, H.K. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Grabowski, G.A. (2008). Phenotype, diagnosis, and treatment of Gaucher's disease. Lancet *372*, 1263–1271.

Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., et al. (2012). Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.

GTEx Consortium (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. *37*, 243–253.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kwon, C.H., Luikart, B.W., Powell, C.M., Zhou, J., Matheny, S.A., Zhang, W., Li, Y., Baker, S.J., and Parada, L.F. (2006). Pten regulates neuronal arborization and social interaction in mice. Neuron *50*, 377–388.

Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb.) *95*, 221–227.

MacKinnon, D.P., Fairchild, A.J., and Fritz, M.S. (2007). Mediation analysis. Annu. Rev. Psychol. *58*, 593–614.

Mahmood, S.S., Levy, D., Vasan, R.S., and Wang, T.J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet *383*, 999–1008.

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. *100*, 473–487.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

Mizuarai, S., Miki, S., Araki, H., Takahashi, K., and Kotani, H. (2005). Identification of dicarboxylate carrier Slc25a10 as malate transporter in de novo fatty acid synthesis. J. Biol. Chem. *280*, 32434–32441.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics *19*, 2088–2096.

Okada, H., Ebhardt, H.A., Vonesch, S.C., Aebersold, R., and Hafen, E. (2016). Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in *Drosophila melanogaster*. Nat. Commun. *7*, 12649.

Ortega-Molina, A., Efeyan, A., Lopez-Guadamillas, E., Munoz-Martin, M., Gomez-Lopez, G., Canamero, M., Mulero, F., Pastor, J., Martinez, S., Romanos, E., et al. (2012). Pten positively regulates brown adipose function, energy expenditure, and longevity. Cell Metab. *15*, 382–394.

Pande, S.V. (1975). A mitochondrial carnitine acylcarnitine translocase system. Proc. Natl. Acad. Sci. USA *72*, 883–887.

Peirce, J.L., Lu, L., Gu, J., Silver, L.M., and Williams, R.W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet. *5*, 7.

Pierce, B.L., Tong, L., Chen, L.S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.J., Franke, L., et al. (2014). Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet. *10*, e1004818.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. *6*, e107.

Segura, V., Vilhjalmsson, B.J., Platt, A., Korte, A., Seren, U., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. *44*, 825–830.

Sham, P.C., and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. *15*, 335–346.

Silverstein, R.L., and Febbraio, M. (2009). CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. Sci. Signal. *2*, re3.

Stekhoven, D.J., and Buhlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics *28*, 112–118.

Taylor, B.A., Heiniger, H.J., and Meier, H. (1973). Genetic analysis of resistance to cadmium-induced testicular damage in mice. Proc. Soc. Exp. Biol. Med. *143*, 629–633.

Uhl, G.R., Sora, I., and Wang, Z. (1999). The mu opiate receptor as a candidate gene for pain: polymorphisms, variations in expression, nociception, and opiate responses. Proc. Natl. Acad. Sci. USA *96*, 7752–7755.

Wang, X., Pandey, A.K., Mulligan, M.K., Williams, E.G., Mozhui, K., Li, Z., Jovaisaite, V., Quarles, L.D., Xiao, Z., Huang, J., et al. (2016). Joint mouse-human phenome-wide association to test gene function and disease risk. Nat. Commun. *7*, 10464.

Williams, E.G., and Auwerx, J. (2015). The convergence of systems and reductionist approaches in complex trait analysis. Cell *162*, 23–32.

Williams, R.W., Bennett, B., Lu, L., Gu, J., DeFries, J.C., Carosone-Link, P.J., Rikke, B.A., Belknap, J.K., and Johnson, T.E. (2004). Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. Mamm. Genome *15*, 637–647.

Williams, E.G., Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C.A., Houten, S.M., Amariuta, T., Wolski, W., Zamboni, N., et al. (2016). Systems proteomics of liver mitochondria function. Science *352*, aad0189.

Wing, R.R., and Hill, J.O. (2001). Successful weight loss maintenance. Annu. Rev. Nutr. *21*, 323–341.

Wu, Y., Williams, E.G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S.M., Argmann, C.A., Faridi, P., Wolski, W., Kutalik, Z., et al. (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell *158*, 1415–1430.

Yao, C., Joehanes, R., Johnson, A.D., Huan, T., Liu, C., Freedman, J.E., Munson, P.J., Hill, D.E., Vidal, M., and Levy, D. (2017). Dynamic role of trans regulation of gene expression in relation to complex traits. Am. J. Hum. Genet. *100*, 985–986.

Yen, K., Le, T.T., Bansal, A., Narasimhan, S.D., Cheng, J.X., and Tissenbaum, H.A. (2010). A comparative study of fat storage quantitation in nematode *Caenorhabditis elegans* using label and label-free methods. PLoS One *5*, e12810.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
cpt-1 RNAi clone	Ahringer RNAi library	Y46G5A.17
Chemicals, Peptides, and Recombinant Proteins		
Sudan black	Sigma	Cat #199664
Oil red O	Sigma	Cat #0625
Deposited Data		
BXD mouse transcriptome data	http://www.genenetwork.org/	Summarized in Table S2
DO mouse proteome data	(Chick et al., 2016)	http://www.nature.com/nature/journal/v534/n7608/extref/nature18270-s1.zip
HMDP mouse liver transcriptome & phenotype data	(Bennett et al., 2015)	http://phenome.jax.org/
CTB6F2 mouse liver transcriptome & phenotype data	http://www.genenetwork.org/	GN Accession: GN172
HXB/BXH rat liver transcriptome & phenotype data	http://www.genenetwork.org/	GN Accession: GN222
LXS mouse prefrontal cortex transcriptome data	http://www.genenetwork.org/	GN Accession: GN110
LXS mouse brain transcriptome data	http://www.genenetwork.org/	Data provided by Richard Radcliffe
ENCODE	(ENCODE Consortium, 2012)	https://www.encodeproject.org/
TCGA	(Cerami et al., 2012; Gao et al., 2013)	http://www.cbioportal.org/
Experimental Models: Organisms/Strains		
Caenorhabditis elegans	Caenorhabditis Genetics Center (Minneapolis, MN)	Bristol strain (N2)
Software and Algorithms		
R	The R Foundation	https://www.r-project.org/
corrgram	The R Foundation	https://cran.r-project.org/web/packages/corrgram/index.html
VennDiagram	The R Foundation	https://cran.r-project.org/web/packages/VennDiagram/index.html
qtl	(Broman et al., 2003)	http://www.rqtl.org/
EMMA	(Kang et al., 2008)	http://mouse.cs.ucla.edu/emma/
intermediate	(Chick et al., 2016)	https://github.com/simecek/intermediate
Circos	(Krzywinski et al., 2009)	http://circos.ca/
ImageJ	National Institutes of Health	https://imagej.nih.gov/ij/index.html
IGV	Broad Institute	http://software.broadinstitute.org/software/igv/
Other		
Olympus AX70	Olympus	http://www.olympusmicro.com/
Resource website for the described tools	This paper	http://www.systems-genetics.org/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Johan Auwerx (admin.auwerx@epfl.ch).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

C. elegans Lines

Wild-type Bristol N2 *C. elegans* were cultured at 20°C on nematode growth media (NGM) plates and sustained on the OP50 *E. coli* strain. Strains were provided by the Caenorhabditis Genetics Center (University of Minnesota).

METHOD DETAILS

BXD Multi-Omics Datasets

Data from 5,092 clinical phenotypes in BXD mouse population were retrieved from GeneNetwork database (http://www.genenetwork.org) on November 1, 2016 (Table S1). Furthermore, molecular data include transcriptomes by microarrays from 34 tissues, ~2,600 proteins quantified by Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH-MS) in liver, ~980 metabolites measured in liver and muscle, and metagenome data that were collected in both feces and caecum of all animals. In summary, we have assembled a deep phenome data set consisting of over 5,000 metabolic, physiological, pharmacological, and behavioral traits, and more than 200 transcriptomic, proteomic, and metabolomic datasets—by far the largest coherent phenome for any animal experimental cohort.

Data from Other Mouse and Rat Populations

Phenotype and transcription data from CTB6F2 (Schadt et al., 2008) and LXS mouse populations (Williams et al., 2004), as well as the HXB/BXH rat cohort (Hubner et al., 2005) were retrieved from GeneNetwork. Data from HMDP was downloaded from http://phenome.jax.org/ and supplemental materials of (Bennett et al., 2015). Proteome data from DO population was downloaded from the supplemental materials of (Chick et al., 2016).

The Cancer Genome Atlas Data

Expression data of *ZNF277* and ribosomal protein genes in cancer samples from 46 datasets, including 35 different cancer types, with RNA-seq available was downloaded from TCGA (http://www.cbioportal.org/) (Cerami et al., 2012; Gao et al., 2013). Datasets with less than 30 samples were removed from the analysis.

The Encyclopedia of DNA Elements Data

Human ChIP-seq data from the Encyclopedia of DNA Elements (ENCODE) was downloaded from www.encodeproject.org. The ENCODE track ID for *ZKSCAN1* is: HeLa ZKSCN1 IgR. The ENCODE track ID for *E2F6* is: hESC E2F6 V11 1. The coverage histograms were generated by using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

C. elegans RNAi and Lipid Staining

RNAi in C. elegans

RNA interference (RNAi) in worms was performed on 90 mm Petri dishes containing NGM agar. Plates were induced overnight with 1mM IPTG at room temperature and seeded with HT115 bacteria expressing either empty vector or the RNAi clones for *cpt-1*. RNAi experiments were performed using L1 larvae synchronized after bleaching of adult worms.

Worm Fixation and Lipid Content Staining

N2 worms were grown on regular NGM plates at 20°C until reaching adulthood, then bleached and the eggs collected and let hatch in M9 medium. L1 larvae were then transferred to RNAi plates for *cpt-1* or to empty vector control plates. At Day 1 of adulthood, worms were collected, washed twice with 1 x PBS and then suspended in 120 μl of PBS to which an equal volume of 2X MRWB buffer (160 mM KCl, 40 mM NaCl, 14 mM Na_2EGTA, 30 mM PIPES pH 7.4, 1 mM Spermidine, 0.4 mM Spermine, 2% paraformaldehyde, 0.2% beta- mercaptoethanol) was added. The worms were taken through 3 freeze-thaw cycles between dry ice/ethanol and warm running tap water, followed by spinning 1 minute at 14,000g washing once in PBS to remove paraformaldehyde. Sudan Black and Oil Red O stainings of stored fat were performed after fixation. For Sudan Black staining, worms were sequentially dehydrated by washes in 25%, 50% and 70% ethanol. Saturated Sudan Black solution was prepared fresh in 70% ethanol. The fixed worms were incubated overnight in 250 μl of Sudan Black, on a shaker at room temperature. Worms were washed twice in 70% ethanol after staining. For Oil Red O staining, worms were re-suspended and dehydrated in 60% isopropanol. 250 μl of 60% Oil Red O stain was added to each sample, and samples were incubated overnight at room temperature. Worms were washed twice in 60% isopropanol solution after Oil Red O staining. The region immediately behind the pharynx of each animal was used for imaging of the lipid droplets (Yen et al., 2010).

Statistical Analysis

BXD Multi-Omics Data Preprocessing

Clinical phenotypes that were measured in less than 15 BXD strains were removed, resulting in 4,784 phenotypes for further analysis in this paper. The clinical phenome has been subdivided into 13 categories based on general biological ontologies through manual inspection.

To obtain the effective number of phenotypes, the whole phenome data was divided based on the respective groups where the animals were raised, to avoid the problem caused by missing of overlapping phenotyped strains across different labs. Imputation was performed to estimate the missing data within groups. Considering that observed phenotypes not necessarily have parametric distributions, we have chosen a promising non-parametric imputation scheme (Stekhoven and Buhlmann, 2012) based on random forests (Breiman, 2001). All phenotypes that had <20% of missing values were imputed and ones with normalized root mean squared error (NRMSE) < 15% have been considered for further reduction (Oba et al., 2003). Based on the correlation matrix of the phenotypes in each group, we took the first m eigenvalues that explain 99.5% of the total variance as the effective number of phenotypes

(N_{eff}) in this group (Li and Ji, 2005). The total number of independent phenotypes across the phenome, a sum of N_{eff} over all studies, was used to estimate the phenome-wide significance threshold ($0.05/N_{eff}$). The same technique was applied to other omics data, including metabolomic, proteomic, and transcriptomic datasets across all tissues. For microarrays, to reduce the burden of multiple testing, we included only probes targeting known transcripts. For genes with multiple probes, probe sets with the highest expression were used in subsequent analysis. This eliminates most intronic probes and those that generally have poor signal-to-noise ratios.

To avoid model misspecification, clinical and molecular phenotypes were transformed into normal shape for the following association analysis.

QTL Mapping

QTL mapping was performed by R/qtl package using Haley-Knott regression (Broman et al., 2003). Local or *cis*-QTLs were determined within a range of 2 Mb up- and down-stream of the gene position, and QTLs that located 5 Mb away from the gene were considered as distant or *trans*-QTLs. A LOD score of 4 was used as the threshold of significance in *trans*-QTLs, and 3 was used in *cis*-QTLs, because for the *cis*-QTLs we have no need to correct for the entire genome multiple testing.

Phenome-Wide Association Analysis

Genes that contain high-impact variants, including missense, nonsense, splice site, frameshift mutations, copy number variations (CNVs), as well as genes that have significant *cis*-e(p)QTLs in the BXD transcriptome and proteome datasets were included in the PheWAS analysis. Genetic variants of each gene are represented by the SNPs within the genes as well as their *cis*-QTLs. About 5,000 clinical phenotypes and over 3,000 metabolites from liver and muscle were used to study the association between genes and phenotypes. Similarly, expression datasets representing 34 different tissues of the BXD strains were used to explore the genetic basis of variation at protein or transcript levels. A mixed model was applied to account for the population structure of the BXD strains (Kang et al., 2008). It is important to take into account that traits are influenced by many genetic loci and, therefore, doing single locus association study can be misleading. In this paper we used a multi-locus mixed-model approach (mlmm) (Segura et al., 2012) to estimate the associations between each gene (represented by the genetic variants of the gene) and clinical and molecular phenotypes (transcripts, proteins and metabolites). This step-wise mixed-model regression with forward inclusion and backward elimination of causative confounding polymorphisms along with the population structure enables to add as a covariates multiple loci, that in turn leads to higher power and lower FDR. Kinship matrix of the BXD strains was estimated using EMMA (Kang et al., 2008). Phenotype y is modeled by mixed effect model as

$$y = X\beta + u + \varepsilon,$$

where X represents a matrix of fixed effects (genotypes), β is a vector of the effect sizes, u is a vector of random effects due to the population structure (its covariance matrix is estimated as $\sigma_u^2 K$) and ε is an error term which is normally distributed around zero with the variance σ_e^2. At each step the variances of each component are recomputed and the most significant loci are added as cofactors until the contribution of the variance of the genetic component, $\frac{\sigma_g^2}{Var(y)}$, is not zero. After re-computation backward stepwise regression eliminate excessive cofactors. The correction for multiple testing was performed with stringent Bonferroni method using both the total number and the effective number of tests. PheWAS results from the clinical phenome are represented as 13 categories based on general biological ontologies, and those from transcriptome and proteome are divided according to the genetic location of the gene across different chromosomes.

Transcriptome/Proteome-Wide Association Analysis

To reduce multiple testing burdens, only probes targeting known transcripts were included in the analysis. For the genes with multiple probes, the highest expressed probe was selected to represent the expression of the gene. Association between transcripts/proteins and traits were evaluated using correlations and corrected for population structure through mixed effect model as described above.

Expression-Based Phenome-Wide Association Analysis

One or two datasets of each tissue from animals cultured in normal or challenged conditions were selected to represent the gene expression profiles in this tissue in our analysis. Associations between transcripts/protein and phenotypic traits were estimated using mixed model regression analysis (Kang et al., 2008). Transcript-trait pairs that had less than 15 overlapping strains were removed from the analysis. Phenome-wide significance was performed using stringent Bonferroni correction using both the total number and the effective number of phenotypes and the number of tissues used in the analysis.

Evaluation of PheWAS and ePheWAS in Detecting associations

The BXD eye transcriptome dataset with 72 strains was used as the molecular phenome data to estimate the performance of PheWAS in detecting associations against the genotypes. The significant PheWAS hits were considered as the "real" positive hits. We then randomly sampled subset cohorts of 20, 30, 40, 50, 60, 70 strains, and performed PheWAS using the actual phenotype data of these cohorts. The random sampling was performed 100 times, and the significant PheWAS hits, as well as the "real" positive hits recovered from these subset cohorts were recorded. Recovery ratio is defined as the ratio of the number of the "real" positive hits recovered and the number of all significant hits from the subset cohort. Overlap coefficient is defined by the number of common significant hits from two subset cohorts divided by the smaller size of significant hits from the two sets.

For ePheWAS, the eye transcriptome dataset was used to represent the gene levels to identify associations against the clinical phenome. Random sampling of ePheWAS was performed 100 times using a similar approach as PheWAS (see above).

Mediation and Reverse-Mediation Analysis

Mediation Analysis. For transcripts that have *trans*-eQTLs, mediation analysis was performed to verify which of the transcripts localizing in the same region are more likely to be the mediators of the target *trans*-eQTLs (Chick et al., 2016). The basic principle is that

each individual transcript level was included as the additive covariate in the QTL mapping of the target gene expression, and regression analysis was performed only at the peak SNP of the QTL. LOD scores of the peak SNP after taking all the transcripts as covariates were used as the significance of the mediation effects of these transcripts on the target *trans*-eQTLs. We performed the same analysis on the proteome datasets as well to determine the causal mediators for *trans*-pQTLs.

Reverse-Mediation Analysis. Using the similar principle, we reversed the mediation approach to determine whether the *cis*-QTL could mediate the *trans*-QTLs that map to the same locus. Specifically, the transcripts/proteins that have *cis*-QTLs were included as additive covariates in the QTL mapping for all transcripts/proteins, with the decrease of QTL LOD scores used as the significance of reverse-mediation effects.

The mediation and reverse-mediation analysis were performed from the R package ''intermediate'' (Chick et al., 2016).

Quantification of Worm Lipid Content Staining

Sudan Black and Oil Red O stained worm Images were taken using Olympus AX70 and quantified with Fiji (ImageJ). We measured the average pixel intensity for an 85-pixel radius immediately behind the pharynx of each animal. In addition, we measured the pixel intensity of the area without worm as background, which was later divided from the values obtained from the staining. A minimum of 26 animals was measured for each strain. Significance was determined by Student's t-test.

DATA AND SOFTWARE AVAILABILITY

All the strategies and data included in this paper are available from systems-genetics.org. Source codes for the analyses described in the paper are available on github.com/auwerxlab/PheWAS.