Systems genetics approaches to probe gene function

Thèse N° 9813

Présentée le 12 décembre 2019

à la Faculté des sciences de la vie
Chaire Nestlé en métabolisme énergétique
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

Hao LI

Acceptée sur proposition du jury

Prof. M. Dal Peraro, président du jury
Prof. J. Auwerx, directeur de thèse
Prof. D. Pagliarini, rapporteur
Prof. D. Wegmann, rapporteur
Prof. B. Deplancke, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

To my parents

Acknowledgements

First of all, I would like to thank my supervisor Prof. Johan Auwerx for giving me the opportunity to work in his lab and providing great atmosphere and resources to learn everything new for my research projects. I joined the lab with a background of molecular biology and limited statistical and programming skills. Johan has always been trusting in me and guiding me through all the difficulties encountered during my study. What I learned most from Johan over these years is his passion on science. The quote on the lab website serves as the best description of Johan: "*Passion wakes us up in the morning after keeping us up all night dreaming of how to be better. Passion drives us to push our mind bodies, and souls to the limit*." I will absolutely benefit a great deal from the experience working with Johan in my future career. I thank Prof. Kristina Schoonjans for the guidance in the collaboration projects and for her support and advice throughout my PhD.

I am grateful to my mentor Prof. Stephan Morgenthaler for his great help during 5 years of my PhD. We meet every Monday afternoon at 1:30 pm to discuss any of my problems in either statistics or other aspects of my study and life. And thanks to Daria Rukina from Stephan's group for the discussions on statistical problems.

I acknowledge Prof. Matteo Dal Peraro, Prof. David J. Pagliarini, Prof. Daniel Wegmann, and Prof. Bart Deplancke, for accepting to be the jury of my thesis and for the insightful questions and comments on my thesis.

I would like to express my thanks to the current and previous members of the Auwerx and Schoonjans labs. Thanks to Terytty Yang Li, Hongbo Zhang, Chang-Myung Oh, Vincenzo Sorrentino, Arwen Gao, and Peiling Luan for the help in the experimental validations. Thanks to Evan Williams, Maroun Bou Sleiman, Alexis Bachmann, and Tao Lin for helpful discussions in bioinformatics and computational problems. Thanks to Alessia Perino, Qingyao Huang, Laura Velazquez Villegas, Karim Gariani, Vera Lemos, Davide D'Amico, Hadrien Demagny, Pan Xu and Adrienne Mottis for the help in animal experiments. Thanks to Fabrice David for the help in the creation of the web source. Thanks to Norman Moullan, Thibaud Clerc, Sabrina Bichet, Andréane Fouassier, as well as the staff members of the EPFL animal facility for the technical assistance. And thanks to Valérie Stengel and Rita Heiniger for the administrative assistance.

I thank the collaborators outside of EPFL, especially Prof. Zoltán Kutalik, Prof. Marc Robinson-Rechavi and Andrea Komljenovic from University of Lausanne, and Prof. Robert W. Williams from University of Tennessee, for their insightful discussions on my research projects.

Thanks to all my friends for the memorable moments we shared and all the fun and laughter.

At the end, I want to thank my parents for their unconditional love and support. 爸爸妈妈，谢谢你们！

A special thank you to my wife, Adi Zheng, who consistently supports and encourages me to follow my heart and pursue what I want. I would not have been able to go through the whole process without her. Adi, having you in my life is the best thing that ever happened to me.

Hao Li / 李昊

Lausanne, November 20, 2019

Abstract

Genes are the functional units of heredity. However, the functions of many genes remain unknown, which impedes the understanding of the underlying mechanism of complex traits and diseases. Systems genetics approaches try to understand the complexity underlining phenotypic traits using high-throughput experimental and computational approaches. In this thesis, I describe a list of novel systems genetics approaches to identify gene functions based on publicly available datasets.

In Chapter 2, I focused on the multi-omics datasets collected from the BXD mouse cohort, which is one of the most studied mouse genetic reference populations. Over the past 40 years, the BXD community has generated tremendous amounts of data covering different omics layers, making the BXD GRP a perfect data source to conduct systems genetics analyses to discover biological insights. By integrating the data from different omics layers, I developed phenome-wide association study (PheWAS) and expression-based PheWAS (ePheWAS) to reveal the associations between genes and phenotypic traits, as well as mediation and reverse-mediation analysis to identify the regulatory connections between genes.

In Chapter 3, I analyzed the transcriptome datasets from six different model organisms, ranging from yeast to human. It is commonly believed that genes with similar functions tend to have similar expression patterns. Therefore by using the co-expression patterns of genes, one can annotate a gene from its correlating genes with known functions. I proposed here a new systems genetics method, termed gene-module association determination (G-MAD), which assigns novel functions of genes and proposes new components of pathway modules. Several new associations, including DDT as a novel mitochondrial protein, were experimentally validated. In addition, G-MAD was further extended to determine the interconnection between pathway modules, for example those between mitochondria and proteasome, as well as ribosome and lipid biosynthesis.

Altogether, this thesis described several novel systems genetics approaches to identify the associations between genes, pathway modules, phenotypic traits, and diseases. The approaches and data described in this thesis have been deposited in a publicly accessible web source at www.systems-genetics.org, and will hopefully facilitate the identification of new gene functions.

Keywords

Systems genetics ; Genetic reference population ; BXDs ; Phenome-wide association study (PheWAS) ; Expression-based phenome-wide association study (ePheWAS) ; Mediation analysis ; Gene set analysis ; Gene annotation ; Module connection ; GeneBridge ; Gene-module association determination (G-MAD) ; Module-module association determination (M-MAD)

Résumé

Les gènes sont les unités fonctionnelles de l'hérédité. Cependant, les fonctions de nombreux gènes restent inconnues, ce qui empêche de comprendre le mécanisme sous-jacent des traits complexes et des maladies. Les approches de la génétique des systèmes tentent de comprendre la complexité soulignant les traits phénotypiques à l'aide d'approches expérimentales et informatiques à haut débit. Dans cette thèse, je décris une liste de nouvelles approches de la génétique des systèmes pour identifier les fonctions des gènes sur la base d'ensembles de données disponibles au public.

Au chapitre 2, je me suis concentré sur les jeux de données multi-omiques collectés dans la cohorte de souris BXD, qui est l'une des populations de référence génétique de souris les plus étudiées. Au cours des 40 dernières années, la communauté BXD a généré d'énormes quantités de données couvrant différentes couches omiques, faisant du BXD GRP une source de données idéale pour effectuer des analyses génétiques des systèmes afin de découvrir des informations biologiques. En intégrant les données de différentes couches omiques, j'ai développé une étude d'association à l'échelle du phénome (PheWAS) et PheWAS basée sur l'expression (ePheWAS) afin de révéler les associations entre gènes et traits phénotypiques, ainsi qu'une analyse de médiation et de médiation inverse pour identifier connexions entre les gènes.

Au chapitre 3, j'ai analysé les ensembles de données du transcriptome de six organismes modèles différents, allant de la levure à l'homme. Il est communément admis que les gènes ayant des fonctions similaires ont tendance à avoir des schémas d'expression similaires. Par conséquent, en utilisant les modèles de co-expression de gènes, on peut annoter un gène à partir de ses gènes en corrélation avec des fonctions connues. J'ai proposé ici une nouvelle méthode génétique des systèmes, appelée détermination d'association gène-module (G-MAD), qui attribue de nouvelles fonctions aux gènes et propose de nouveaux composants de modules de voies. Plusieurs nouvelles associations, y compris le DDT en tant que nouvelle protéine mitochondriale, ont été validées expérimentalement. De plus, G-MAD a été étendu pour déterminer l'interconnexion entre les modules de la voie, par exemple ceux entre les mitochondries et le protéasome, ainsi que la biosynthèse des ribosomes et des lipides.

Au total, cette thèse a décrit plusieurs approches novatrices de la génétique des systèmes pour identifier les associations entre gènes, modules de voies, caractéristiques phénotypiques et maladies. Les approches et les données décrites dans cette thèse ont été déposées dans une source Web accessible au public sur www.systems-genetics.org. Elles faciliteront, espérons-le, l'identification de nouvelles fonctions géniques.

Mots-clés

Génétique des systèmes; Population de référence génétique; BXDs ; Etude d'association à l'échelle du phénome (PheWAS); Étude d'association à l'échelle du phénome à base d'expression (ePheWAS) ; Analyse de médiation; Analyse de jeux de gènes; Annotation génique; Connexion du module; GeneBridge ; Détermination de l'association gène-module (G-MAD); Détermination de l'association module-module (M-MAD).

Contents

Acknowledgements ..... **v**
Abstract ..... **vii**
Keywords ..... **vii**
Résumé ..... **viii**
Mots-clés ..... **viii**
List of Figures ..... **xii**
Chapter 1 Introduction **13**
1.1 Mouse systems genetics as a prelude to precision medicine 14
1.1.1 Introduction 14
1.1.2 The essentiality of mouse studies in human precision medicine 14
1.1.3 Use the correct models in mouse studies 15
1.1.4 Panels and resources in mouse systems genetics 17
1.1.5 Systems genetics approaches to analyze multi-omics data 19
1.1.6 Concluding remarks and future perspectives 21
1.2 Most of the genes remain poorly annotated 22
1.3 Aims of the thesis 23
Chapter 2 An integrated systems genetics and omics toolkit to probe gene function 25
2.1 Graphic abstract 25
2.2 Abstract 25
2.3 Introduction 26
2.4 Methods 26
2.4.1 Experimental model and subject details 26
2.4.2 Method details 27
2.4.3 Statistical analysis 28
2.4.4 Data and software availability 30
2.5 Results 30
2.5.1 Structure and pre-processing of multi-layer data from BXD population 30
2.5.2 PheWAS reveals G2P associations and facilitates the detection of pleiotropic effects 32

2.5.3 Expression-based PheWAS (ePheWAS)—a tool to discover gene functions 34
2.5.4 Evaluation of PheWAS and ePheWAS in detecting associations 37
2.5.5 Mediation analysis identifies regulatory mechanism of gene expression 38
2.6 Discussion 41
2.7 Supplemental figures 42
2.8 Acknowledgements 46
Chapter 3 Identifying gene function and module connections by the integration of multi-species expression compendia 47
3.1 Abstract 48
3.2 Introduction 48
3.3 Methods 49
3.3.1 Gene annotations / Modules 49
3.3.2 Module similarity calculation 49
3.3.3 Gene expression across tissues 49
3.3.4 Transcriptome datasets 49
3.3.5 Data preprocessing of transcriptome datasets 49
3.3.6 Gene-Module Association Determination (G-MAD) 50
3.3.7 Module-Module Association Determination (M-MAD) 51
3.3.8 Module network analysis 51
3.3.9 Gene correlation network analysis 52
3.3.10 Cross validation 52
3.3.11 Gene set enrichment analysis 52
3.3.12 Transcript-phenotype correlation analysis in mouse cohorts 52
3.3.13 Cell culture and siRNA transfection 52
3.3.14 Mitochondrial function assay 53
3.3.15 Mitochondrial localization 53
3.3.16 *C. elegans* experiments 53
3.3.17 Data access 54
3.4 Results 54
3.4.1 Gene-Module Association Determination (G-MAD) 54
3.4.2 G-MAD identifies tissue-specific associations 56
3.4.3 G-MAD determines novel genes linked to mitochondria 58
3.4.4 Module-Module Association Determination (M-MAD) 60
3.5 Discussion 62
3.6 Supplemental figures 65
3.7 Acknowledgements 77

Chapter 4 Conclusion and perspectives 79
4.1 Research summary 79
4.2 Perspectives 80
References 83
List of abbreviations 91
Curriculum Vitae 92

List of Figures

Figure 1.1 Genetic difference across inbred mouse strains 16

Figure 1.2 The influence of genetic background on the phenotypic response to genetic and environmental perturbations 17

Figure 1.3 Breeding scheme of mouse genetic reference populations 18

Figure 1.4 Systems genetics approaches 20

Figure 1.5 GO annotations from different evidences 22

Figure 1.6 Known annotations for human genes 23

Figure 2.1 Overview of multi-omic data from the BXD population, and the scheme of applied systems approaches 31

Figure 2.2 Phenome-wide association analysis 33

Figure 2.3 Genotype-phenotype associations revealed by PheWAS 34

Figure 2.4 ePheWAS displays tissue-specific regulators 35

Figure 2.5 ePheWAS reveals *Cpt1a* as a regulator of fasting weight loss 36

Figure 2.6 Performance of PheWAS and ePheWAS in detecting G2P associations 38

Figure 2.7 Mediation and reverse-mediation analysis discovers gene interactions 40

Figure 3.1 Gene-Module Association Determination (G-MAD) 55

Figure 3.2 Predicting tissue-specificity of modules 57

Figure 3.3 G-MAD identifies tissue-specific associated modules for EHHADH by using datasets from different tissues 58

Figure 3.4 G-MAD predicts novel genes linked to mitochondria 59

Figure 3.5 Module-Module Association Determination (M-MAD) reveals module connections 61

Figure 3.6 M-MAD reveals a negative association between the ribosome and lipid biosynthetic modules 63

Chapter 1 Introduction

This part of work is adapted from a review article written by Hao Li and Johan Auwerx.

Contribution to this work: I prepared the figures and wrote the manuscript with the guidance of Prof. Johan Auwerx.

1.1 Mouse systems genetics as a prelude to precision medicine

Mouse models have been instrumental in understanding human disease biology and proposing possible new treatments. The precise control of the environment and genetic composition of mice allows more rigorous observations, but limits the generalizability and translatability of the results into human applications. In the era of precision medicine, strategies using mouse models have to be revisited to effectively emulate human populations. Systems genetics is one promising paradigm that may promote the transition to novel precision medicine strategies. Here, we review the-state-of-the-art and discuss how applying systems genetics in mouse populations helps to understand complex traits and diseases, with a particular emphasis on the existing resources and strategies.

1.1.1 Introduction

Most complex traits and diseases, such as height, longevity, and diabetes, are heritable and influenced by various genetic factors [1], while being modulated by environmental stimuli. Due to the fact that every individual has a unique genetic makeup, response to drugs [2], nutrition [3], and life-style [4] changes, vary considerably from person to person. This uniqueness of every human being underpins the purpose of precision medicine, which posits that disease prediction, diagnosis and treatment for each individual is based on personal genomic variations and external environments [5]. Precision medicine is an innovative approach that takes the variability in genetics, environment, and lifestyle of each individual into account in disease prevention and treatment, and provides better prediction of effective treatments, while concurrently minimizing the possibility of drug side effects [6]. Therefore, precision medicine requires a good understanding of the genetic bases of variation in phenotypes and their interaction with the environment in health and disease.

Due to practical and ethical issues, model organisms have been used as simplified models for human to study the genetic, molecular, and physiological basis of complex traits and to find therapeutic targets for human diseases. Mice have been the most studied animal models because of many reasons, including their similarity to human and the possibility to control the environmental factors. In recent years, more and more systems genetics studies have been performed on mouse populations and proved that mice from different genetic backgrounds exhibit distinct phenotypic responses, corroborating the principles that form the basis of precision medicine. A list of genetic determinants of complex traits have been identified and verified in human cohorts [7, 8]. We review here the recent developments in mouse systems genetics studies on complex traits and diseases, and summarize the existing resources and strategies and discuss how they may help with the implementation of personalized and precision medicine approaches.

1.1.2 The essentiality of mouse studies in human precision medicine

For many decades, research studies using model organisms have been conducted to guide our understanding of biological processes, with the mouse being one of the most extensively used models. In 2011, 61% of all the animals used for experimental and other scientific purposes in the European Union were mice (http://eara.eu/en/animal-research/animal-research-statistics-europe/#eu-statistical-report). Recently, however, there has been increasing doubts about the translational potential of findings in mouse models [9, 10]. In another review, we argued against this opinion and demonstrated through evidence the contributions of mouse studies in human drug discovery and in the general understanding of human biology [11].

One major criticism against mouse models is that results from mouse experiments do not always reflect human diseases. For example, there is no single model that recapitulates the pathophysiological and molecular aspects of non-alcoholic steatohepatitis (NASH) [12]. This could be partially explained by the fact that most mouse models are built on inbred strains with a fixed genome, while individuals with different genetic backgrounds may behave and progress differently in disease conditions [13, 14]. In addition, laboratory mice housed in well-controlled and hygienic cages do not experience the dynamic real-life environment as wild mice or humans [15]. Therefore the expectations in many cases are too high given the imperfect mouse models. However, the process of finding novel and refining existing mouse models is an ongoing iterative process [16-18]. For instance, new mouse models have been recently proposed for the most common form of heart failure

in humans [19]. In addition, the emergence of new technologies such as CRISPR/Cas9 could unlock novel or refined mouse models [20]. Finally, we argue that many of the shortcomings of existing mouse models can be blamed on the extreme standardization of mouse experiments, where most research is performed on mice of one or a few genetic backgrounds, usually inbred strains. Whereas human individuals in reality are genetically diverse and heterozygous in most genetic loci.

Here we listed the advantages of using mouse systems genetics to study human diseases:

1. Mice are similar to humans in many aspects, including genetics, anatomy and physiology. The pathophysiology of disease in mice is also similar to that in humans.
2. The genomes of many commonly used mouse strains have been sequenced, and there are developed tools available to manipulate the mouse genome and record their phenotypes.
3. There are well-established mouse models for many diseases, as well as genetic reference panels for systems genetics studies.
4. Mice are cost effective due to their relatively short lifespan (two to three years) and generation time, and are easy to handle and breed.
5. The external environment of mouse models can be well controlled and monitored, which also facilitate the study of gene-environment interactions.
6. Studies using inbred mice allow resampling isogenic individuals to replicate the same experiment or perform multiple experiments to better estimate the influence of genetics and environment on phenotypes.
7. Researchers have access to all the tissue samples in mice, especially those highly relevant in diseases, which is impossible in most human studies because of ethical issues.
8. Mouse models can be used to capture the disease progression stages in longitudinal studies.
9. Mouse genetic populations are able to model the genetic diversity of human populations, and require fewer individuals for genetic association analyses.
10. Unlike human genetic studies where data should always be kept highly confidential, data from mouse studies can be made public available to facilitate its re-analysis to the fullest extent.

1.1.3 Use the correct models in mouse studies

The choice of genetic background in biomedical research is a crucial but often disregarded step. However, increasing evidence shows that individuals with different genetic backgrounds may behave and progress differently in disease conditions and can even react in opposite directions to external stimuli and treatments [13, 14]. The response to morphine or cocaine [21], body weight gain upon high fat diet [22], and lifespan changes after caloric restriction [23] are just a few examples. C57BL/6J is the most extensively used mouse strain in biomedical research [24]; however, many of the findings from C57BL/6J cannot be even be generalized to its substrains like C57BL/6N, which has only 51 coding variants differing from C57BL/6J [25, 26]. From a genomic standpoint, C57BL/6J carries the minor alleles for 19% of the high-impact variants among the 30 sequenced inbred mouse strains [27, 28] (Figure 1.1A), demonstrating that studies focusing on genes with these variants using C57BL/6J might not be well translated to most of the other strains. Furthermore, the percentage of high-impact variants with the minor allele is similar in other mouse strains (Figure 1.1A), implying that choosing one strain over the others may lead to serious biases due to these naturally occurring variants. This would be the equivalent to studying diseases using a single human individual and then extrapolating results to the entire population.

In addition, the high-impact genetic mutations in some strains lead to the disruption of genes crucial in certain biological processes (Figure 1.1B), and therefore more attention should be paid when planning for animal experiments. For example, C57BL/6J is known to carry a large deletion in the *Nnt* gene, which associates with impaired insulin secretion and glucose tolerance [26]. Some strains, including the widely used C57BL/6J and BALB/cJ strains, have a 6-bp deletion of *Cox7a2l* causing a two amino acids truncation and its inactivation, disrupting the formation of mitochondrial supercomplexes [22, 29]. DBA/2J possesses coding and non-coding variants in *Oprm1*, a known opioid receptor, and therefore exhibits weaker morphine preference and response compared to C57BL/6J [7, 30]. Known mutations in genes that are relevant to the phenotype-of-interest must therefore be avoided in order to preclude unwanted biases. However, the genes crucial for these traits and diseases are not always known. Therefore, mouse populations, instead of mice of a single genetic background, would serve as a natural starting point to finding adequate models as well as to study the genetic basis of physiological traits and diseases. In this experimental setting, mouse strains with deleterious variants can serve as the counterpart model of human individuals with rare disease mutations to test their response to external challenges, for instance relevant drugs.

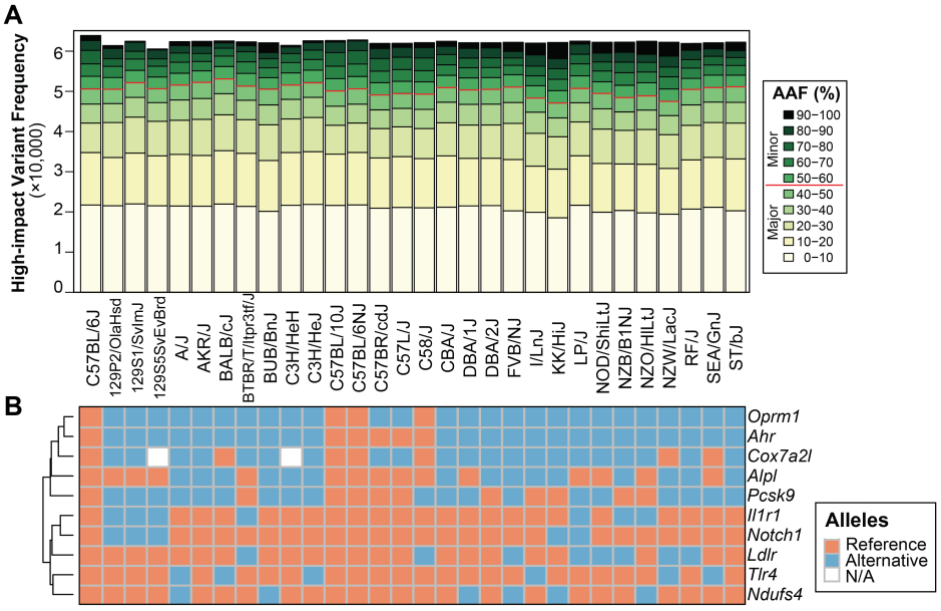


Figure 1.1 Genetic difference across inbred mouse strains

A, The alternative allele frequency (AAF) of the high-impact variants across 30 inbred mouse strains whose genome was sequenced. Data were downloaded from the Mouse Genomes Project (www.sanger.ac.uk/sanger/Mouse_SnpViewer/). Wild-derived strains were removed from the analysis. The variation consequences were predicted with the Variant Effect Predictor (VEP). High-impact genetic variants were counted based on their AAF in these mouse strains and separated into 10 bins. The proportions of the genetic variants possessing the minor alleles in respective strains were indicated by the red line.

B, The genetic diversity of high-impact variants for a set of genes that are crucial in physiology and diseases across 30 inbred strains. The alleles of C57BL/6J were used as the reference allele and other alleles were indicated as alternative alleles.

Recently, several studies have been conducted to study the effects of disease-causing genetic mutations or environmental stimuli in different mouse strains, and observed strong influence of genetic background on phenotypic responses (Figure 1.2). For instance, the phenotypic effects of *Cacna1c* and *Tcf7l2* mutations were evaluated in different genetic backgrounds by breeding heterozygous males to females from 30 inbred strains (Figure 1.2A) [13]. The phenotypic responses to these two mutations varied across different genetic backgrounds and in several cases there were even opposite effects [13], demonstrating that the genetic effects observed in animal models with a single genetic background are not generalizable to the whole population. A similar strategy was used to study the translatability of Alzheimer's disease (AD) mouse models by crossing a heterozygous AD transgenic line with 28 genetically diverse BXD recombinant inbred strains (Figure 1.2A) [14]. Although most of the mice with transgenic alleles exhibited impaired cognitive function, the impact of transgene varied widely depending on the genetic background of the strains [14]. The translatability of AD mouse models was also tested by backcrossing AD animals to three wild-derived mouse strain; significant

phenotypic variations in the neuropathological performance of the animals from different strains as well as genders was observed [31]. Similarly, over 100 inbred strains of mice from the HMDP were crossed with a strain with dyslipidemia-inducing mutations and the obtained F1 progeny were further exposed to a high-fat high-cholesterol diet to promote atherosclerosis development [32]. Animals with different genetic backgrounds exhibited distinct susceptibility to atherosclerosis induced by hyperlipidemia, which is consistent with the results in human epidemiologic studies. Candidate genes underlying the atherosclerosis-related traits were then identified through association mapping and correlation analyses [32]. Likewise, a penetrant prostate cancer mouse model was crossed to the Diversity Outbred cohort, and the obtained F1 males were used to study the effects of genetic variation on the susceptibility to prostate cancer [33]. Further integrative analyses identified several genes as aggressive prostate cancer modifiers, which were then validated in human [33]. The responses to four human-comparable mouse diets (American diet, Mediterranean, Japanese and Maasai/ketogenic) were evaluated in four inbred mouse strains in an effort to find the best alternative to the American diet (Figure 1.2B) [34]. Of note, the best diet was shown to be strain-dependent and it was proposed that health outcomes could be improved through a precision dietetics approach. Altogether, these studies highlight the importance of genetic diversity of animal models in biomedical research. Considering that most of the initial discoveries were made using mouse models of single genetic background, it is therefore explicable that some findings from mouse studies were not well translated into humans [35].

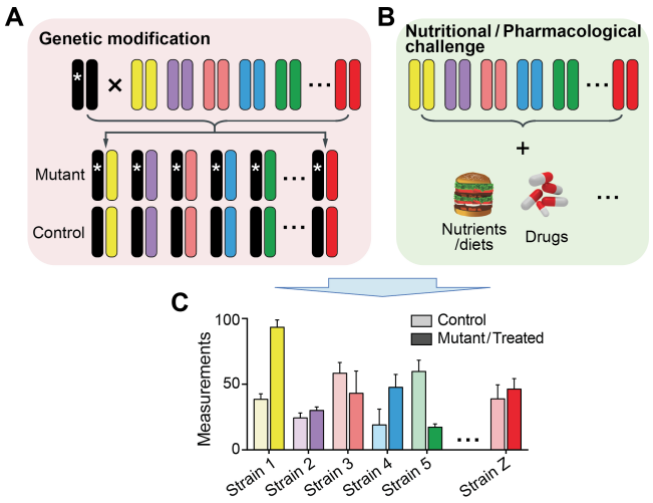


Figure 1.2 The influence of genetic background on the phenotypic response to genetic and environmental perturbations

A, An inbred strain heterozygous for a disease-causing mutation is crossed to a panel of inbred strains to generate genetically diverse, but isogenic, F1 offspring. The progeny inheriting the mutation can be compared against their littermates to identify the influence of genetic background on disease pathogenesis.

B, The panel of inbred strains is challenged by either nutritional or pharmacological approaches to assess their respective response.

C, The response to genetic and environmental perturbations is highly affected by the genetic background of the strains.

1.1.4 Panels and resources in mouse systems genetics

For decades, research groups have generated various mouse genetic reference panels (GRPs), to study the genetic bases of phenotypic traits and diseases [36]. These mouse populations are often derived from several different parental strains that have distinct phenotypic performances. For example, the BXD cohort was derived from the C57BL/6J and DBA/2J strains with different response to drugs and diet-induced obesity and thus this population is commonly used for neuropharmacological and metabolic research [37]. While the LXS cohort was generated from the Inbred Long-Sleep (ILS) and Inbred Short-Sleep (ISS) strains, and is often used in neural and behavioral studies [38]. Unlike common cohorts originated from two inbred strains or individuals, the collaborative cross (CC) and diversity outbred (DO) cohorts are more recently established advanced diversity panels that derived from eight parental strains through a community effort [39]. By including three

wild-derived strains, the CC/DO founder strains capture nearly 90% (vs ~13% in BXDs) of the common genetic variations in *Mus musculus* strains [39, 40], approximating human genetic diversity.

In general, mouse genetic panels can be divided into different categories depending on the breeding strategies, including inbred, F1 hybrids, F2 hybrids, outbred, heterogeneous, recombinant inbred, recombinant inbred cross, recombinant inbred backcross, congenic, consomic, and conplastic strains (Figure 1.3, Table 1.1). Different mouse cohorts have different genetic origins, availability, and usability [41], therefore attention should be paid when considering the cohort for specific experimental settings and research questions. Hybrid diversity panels, for example the HMDP, rely on the available strains and combine the inbred strains to increase mapping resolution and recombinant inbred strains for the mapping power [42, 43].

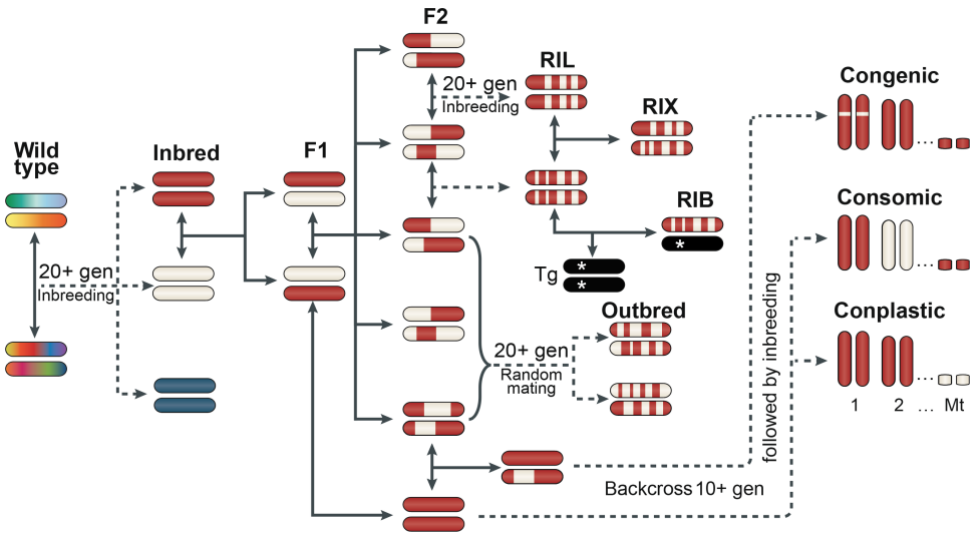


Figure 1.3 Breeding scheme of mouse genetic reference populations

A brief breeding scheme summarizing the common mouse population types (text in bold). Different colors represent the genotypes of the chromosomes. The scheme mainly focuses on cohorts derived from two parental strains; multiparental populations employ a similar but more complex breeding strategies [39]. Inbred strains are derived from at least 20 generations of brother-sister mating of wild type mice. Individuals of an inbred strain are considered as isogenic. F1 hybrids are generated by crossing mice of two different inbred strains, and F2 hybrids are produced by crossing F1 mice. Recombinant inbred lines (RILs) are derived from long-term inbreeding (usually over 20 generations) of F2 progenies. Recombinant inbred intercrosses (RIXs) are established by crossing mice from different RILs, while recombinant inbred backcrosses (RIBs) are produced by creating F1 hybrids from a transgenic strain (Tg) and RILs. Outbred mice can be generated through random mating of F2 progenies. The congenic strain is an inbred strain with a chromosomal segment substituted by the corresponding segment of another strain. Special types of congenic strains include consomic and conplastic strains, where a whole chromosome or the mitochondria are substituted by that of another strain. Consomic strains are also called chromosome substitution strains. Figure adapted from [36].

Despite the advantages of mouse systems genetics, the cost and resources for such studies needed remain one of the major obstacles, especially for research labs with limited budget. However, researchers can benefit from existing data of previous systems genetics studies to generate and verify research hypotheses in their projects. With the development of high-throughput molecular technologies, the collection of omics data has become a routine in biomedical research, especially in systems genetics studies using mouse genetic populations. Contributed by research groups around the world, large scale of omics data have been collected, ranging from epigenomics [44, 45], transcriptomics (data from the BXD and HMDP cohorts were partially summarized in [7, 43]), proteomics [22, 46, 47], lipidomics [48-50], metabolomics [22, 51], microbiome [52-54], as well as phenomics [7, 8]. As mouse systems genetics studies often are unbiased towards the gene targets, therefore data from such studies can be re-used to analyze any gene-of-interest for different research groups. There are various resources that provide access to the mouse systems genetics datasets (Table 1.1), as well as the systems approaches, including GeneNetwork (www.genenetwork.org), the Mouse Phenome Database (MPD, https://phenome.jax.org/), the Systems Genetics Resource (https://systems.genetics.ucla.edu/), the Attie Lab Diabetes Database (http://diabetes.wisc.edu/), the Diversity Outbred Database (www.jax.org/research-and-faculty/genetic-diversity-initiative/tools-data/diversity-outbred-database), the

Swiss-BXD web interface (https://bxd.vital-it.ch), and Systems-Genetics.org (www.systems-genetics.org/). These systems genetics resources enable the possibility to reuse historically collected data to identify novel biological insights, such as done previously [7, 8, 55, 56].

Table 1.1. Commonly used mouse genetic cohorts and data sources

Cohort type	Example cohort name	Parental strains	Data source	Key references
Inbred	-	-	https://phenome.jax.org/panels	[57]
Outbred	CFW	-	https://wp.cs.ucl.ac.uk/outbredmice/ http://dx.doi.org/10.5061/dryad.2rs41	[58, 59]
	DO	A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ	https://phenome.jax.org/panels/DO%20population www.jax.org/research-and-faculty/genetic-diversity-initiative/tools-data/diversity-outbred-database	[47, 60, 61]
	AIL	LG/J x SM/J	https://palmerlab.org/protocols-data/	[62]
Heterogeneous	HS	A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL/6J, DBA/2J, LP/J	https://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/	[63]
	ITP	BALB/cByJ, C57BL/6J, C3H/HeJ, DBA/2J	https://phenome.jax.org/projects/ITP1	[64]
Recombinant inbred	BXD	C57BL/6J, DBA/2J	http://www.genenetwork.org	[22, 46, 55]
	LXS	ILS, ISS	http://www.genenetwork.org	[38]
	CC	A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ	https://phenome.jax.org/panels/CC	[39, 40]
Hybrid diversity panel	HMDP	C57BL/6J, DBA/2J, A/J	https://systems.genetics.ucla.edu/data/hmdp	[43, 50, 56]
F1 hybrids	-	Two inbred strains	-	
F2 hybrids	B6BTBRF2	C57BL/6J, BTBR T+tf/J	http://diabetes.wisc.edu/	[65, 66]
	CASTB6F2	C57BL/6J, CAST/EiJ	https://systems.genetics.ucla.edu/data/B6_CAST	[67]
	BHF2	C57BL/6J, C3H/HeJ	https://systems.genetics.ucla.edu/data/C3H_B6	[67]
RIX	CC-RIX	A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ	-	[68]
RIB	AD-BXD	5XFAD, BXDs	-	[14]
	Ath-HMDP	CETP, ApoE3-Leiden, HMDP	https://systems.genetics.ucla.edu/data/hmdp_apoe_leiden	[32]
Congenic	-	Two inbred strains	-	
Consomic	-	Two inbred strains	-	[69]
Conplastic	-	Two inbred strains	-	

CFW, Carworth Farms Swiss Webster; DO, Diversity outbred; AIL, Advanced intercross line; HS, Heterogeneous stock; ITP, Interventions testing program; BXD, recombinant inbred cohort by crossing C57BL/6J (B) with DBA/2J (D); CC, Collaborative cross; HMDP, the hybrid mouse diversity panel; B6BTBRF2, F2 hybrids by crossing C57BL/6J (B6) with BTBR T+tf/J (BTBR); RIX, recombinant inbred cross; RIB, recombinant inbred backcross; 5XFAD, B6SJL-Tg(APPSwFlLon, PSEN1*M146L*L286V) 6799Vas/Mmjax (Stock Number: 006554); CETP, B6.CBA-Tg(CETP)5203Tall/J (Stock Number: 003904); ApoE3-Leiden, mice carrying the human ApoE3 Leiden variant.

1.1.5 Systems genetics approaches to analyze multi-omics data

The accumulating multi-omics datasets from mouse systems genetics studies provide valuable resources, which can form the foundation of systems genetics approaches to discover novel biological findings. Here we summarized the commonly used, as well as newly described systems approaches analyzing these multi-omics datasets.

Genetics approaches connecting genes with phenotypes can be generally separated into forward and reverse genetics approaches. Forward genetics approaches identify the causal genes or genetic variants that contribute to the phenotypic variance, while reverse genetics approaches start with a gene-of-interest or a genetic variant and try to identify its impact on downstream traits (Figure 1.4A).

Common forward genetics analyses include quantitative trait locus (QTL) mapping (linkage studies applied on related individuals) and genome-wide association study (GWAS, association studies using a large number of related or unrelated individuals), which are widely used to map the genetic loci that correlate with a particular trait, ranging from phenotypes, metabolites, proteins, transcripts, to epigenetic markers [44, 45]. In recent years, studies have been performed to study the genes involved in the various diseases and complex traits, including insulin secretion [60], diabetes [37], hepatic steatosis or fibrosis [43, 44], blood pressure [70], bone density [57].

Epigenetics, such as DNA methylation and histone modifications, affects complex traits through regulating gene expression and activity. The DNA methylation levels were for instance measured in the livers of 90 HMDP mouse strains, allowing epigenome-wide association studies (EWAS) to determine the association between variation in DNA methylation and complex phenotypic traits [45].

Phenome-wide association study (PheWAS), as a complementary approach to GWAS, examines the associations between one genetic variant and a large number of phenotypes (termed pleiotropy) [71]. PheWAS was first used to analyze electronic health records in humans [72], and was later applied in mouse populations, more in particular in the BXD cohort [7, 8]. Genetic reference panels that are composed of inbred strains or recombinant inbred strains, which can be easily reproduced and extensively phenotyped, allow the accumulation of huge phenomic datasets and therefore are perfect sources for such reverse-genetics analyses.

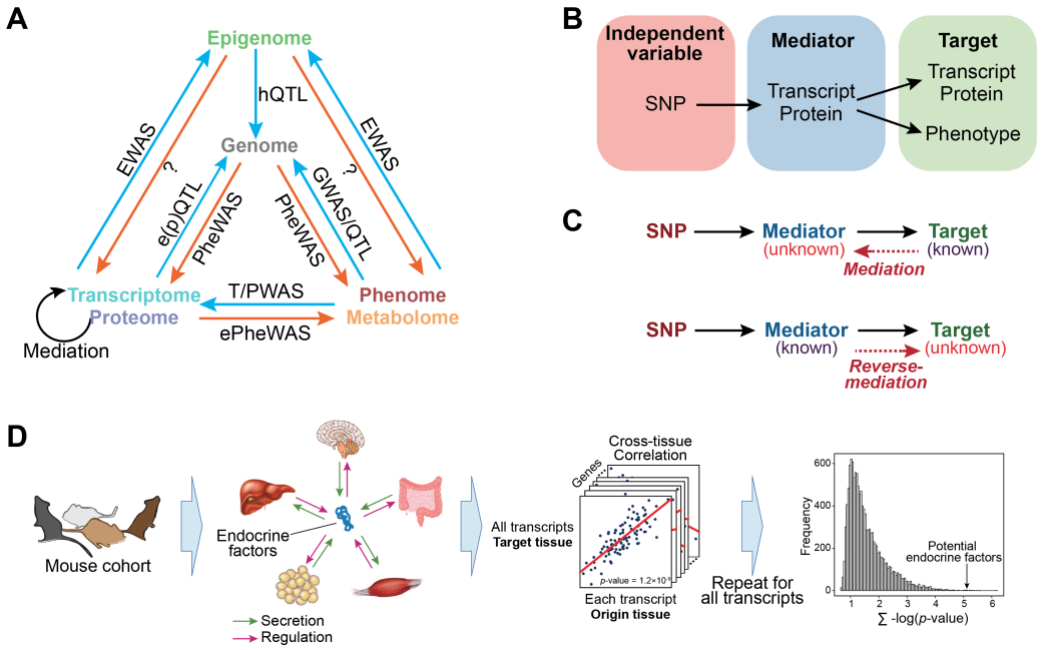


Figure 1.4 Systems genetics approaches

A, A scheme of systems genetics approaches that can be applied on multi-omics data. Multi-omics datasets are divided into four categories, i.e. the genome, the epigenome, the transcriptome/proteome, and the phenome/metabolome. Arrows connecting different omics layers are colored blue if the respective approaches are forward genetics methods and orange if they are reverse genetics methods.

B, Flow of the biological information. SNPs are independent variables that affect transcripts/protein or phenotypes (target) through influencing intermediate molecules (mediator), such as transcripts or proteins.

C, Mediation analysis starts with a known target and identifies the unknown mediator (Upper), while reverse-mediation analysis starts from a known mediator to discover its downstream targets (Lower).

D, Cross-tissue correlation to uncover endocrine factors. Expression datasets obtained from different tissues of the same mouse cohort can be used to identify the endocrine factors that regulate gene expression in other tissues. The p-values of the correlation between the expression levels of each gene in the origin tissue and those of all genes in the target tissue are calculated and then aggregated after applying logarithm transformation. Genes with higher $\sum$ -log(p-value) are potential endocrine factors.

Intermediate molecules, such as mRNA and protein, integrate the effects from genetic factors, including those poorly captured or hidden in common association studies [73], as well as effects from environmental factors. Several studies explored the use of these intermediate phenotypes and introduced the concept of transcriptome- or proteome-wide association study (T/PWAS), which suggest candidate genes by associating the phenotypic traits to the expression (transcript or protein) levels of the gene [74, 75]. Conversely, a reverse approach (expression-based PheWAS, ePheWAS) that identifies the associations between one gene and multiple phenotypic traits based on its gene expression has been proposed [7]. Different from common correlation analyses, T/PWAS and ePheWAS exploit mixed effect models to account the population structure when exploring the connections between genes and phenotypes [76]. By applying on the data from the BXDs, T/PWAS and ePheWAS uncovered a number of gene-phenotype associations, many of which were not recognized using genetic associations [7].

QTL mapping of traits in mouse cohorts often ends up with a genetic locus, composed of a list of candidate genes. Several studies proposed the use of mediation analysis to identify the causal gene (mediator) between the genetic variant (independent variable) and the trait-of-interest (dependent variable) (Figure 1.4B) [7, 47, 61, 77]. Mediation analysis can be used either on gene expression levels to identify the regulatory mechanisms [7, 47, 61], or on phenotypic traits to discover the potential causal drivers contributing to the phenotypic variances [77] (Figure 1.4C upper). Contrary to mediation analysis, reverse-mediation analysis starts with the mediator (the gene with *cis*-QTL) and identifies its downstream targets [7] (Figure 1.4C lower).

Additional computational methods will surely emerge that exploit such huge datasets. For instance, using transcriptome datasets obtained from different tissues of the HMDP cohort, a new strategy to identify important endocrine factors in the communication between tissues was developed (Figure 1.4D) [56]. Using expression datasets from large cohort studies, novel systems approaches, including the GeneBridge toolset (www.systems-genetics.org), have also been developed to identify the novel function of genes or new members of pathway modules [78]. Other studies applied gene network modeling algorithms to identify the potential regulators in complex diseases, for example cardiomyopathy [79], hepatic steatosis [80], as well as coronary artery disease [81].

Finally, there are many other integrative approaches available for the analysis of multi-omics data, but have not yet been applied in mouse systems genetics studies. Examples include the transcriptome-wide association study (TWAS) that integrates GWAS with expression datasets from other independent cohorts to prioritize candidate gene for phenotypic traits. In addition, Mendelian randomization, which estimates the causal associations between a risk factor and diseases [82] or those between gene expression and complex traits [83] can be similarly applied in mouse genetic cohorts.

1.1.6 Concluding remarks and future perspectives

Mouse models have long been used to study the basis of human diseases, to screen for potential drug targets, and to test the safety and efficiency of drugs in pre-clinical trials. We review here the recent advances applying systems genetics in mouse populations to understand the basis of complex traits and diseases. The resources of available archived datasets as well as the commonly used systems approaches are described. However, the infrastructures for the data generation, storage and integrative analyses in mouse systems genetics are not yet standardized and will require further work.

There are arguments that mouse models only poorly mimic human diseases and predict disease outcomes in human; we do not adhere to this opinion. It is clear that phenotypic traits as well as the response to disease-

causing variants or environmental stimuli are strongly affected by the genetic background of the individuals. This exposes the disadvantages of the use of animals from single genetic background in traditional animal studies and argues for the use of genetically diverse cohort in assessing the effects of external factors, such as done in the ITP studies, where the effects of various treatments on aging were tested in a large panel of genetically heterogeneous mice [64]. We hence propose here the concept of “mouse precision medicine” and argue that it can serve as better prototype for future mouse studies and as such provide valuable insights for human precision medicine.

1.2 Most of the genes remain poorly annotated

Annotation of gene function is a main focus for biological research. However, despite great efforts to annotate the cellular and physiological role of genes, many of their functions remain poorly understood [84]. Here we summarized the current status of existing gene annotations from several of the major resources, which were manually curated and maintained by research consortia.

One of the most widely used resources of gene annotation is the Gene Ontology (GO), which characterizes gene function into three ontologies, i.e. biological process (BP), molecular function (MF), and cellular component (CC) [85]. The version retrieved on Oct 4, 2017 covered 19,440 protein-coding genes with over 400,000 annotations. GO annotations include evidence codes to indicate how the annotations are supported. We summarized the sources of evidence for all the available GO annotation in human and found out that the uncurated IEA (Inferred from electronic annotation) contributed considerable amount (18.84%) to all the annotations (Figure 1.5). It should be noted that annotations from IEA are not manually reviewed and curated; therefore we only focused on the annotations in the analyses described in Chapter 3.

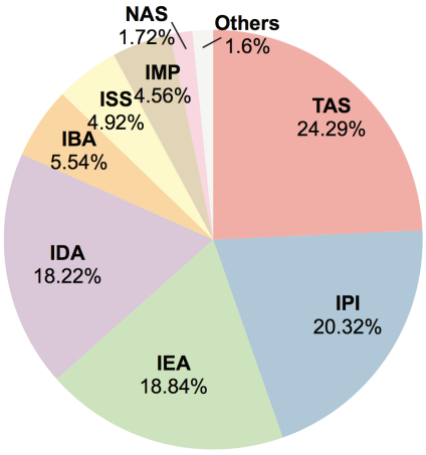


Figure 1.5 GO annotations from different evidences

TAS, Traceable Author Statement; IPI, Inferred from Physical Interaction; IEA, Inferred from Electronic Annotation; IDA, Inferred from Direct Assay; IBA, Inferred from Biological aspect of Ancestor; ISS, Inferred from Sequence or structural Similarity; IMP, Inferred from Mutant Phenotype; NAS, Non-traceable Author Statement.

We then summarized the number of annotations for each gene, and found that over 54% (10,543 genes) of all the protein-coding genes in humans have no more than 10 annotations (Figure 1.6A), whereas the most annotated gene *TP53* has more than 800 annotations (Figure 1.6B). In fact, the top 20% most annotated genes have more than 64% of all annotations in GO (Figure 1.6C). It makes sense that these genes received more attention, because many of these genes have crucial roles in different biological processes and are often mutated in various diseases, for example *TP53* in cancer. However, research attentions should also be paid to genes that are poorly studied to investigate the underlying mechanism of diseases.

The situation that most genes remain under-studied is also true for annotations retrieved from other sources, such as GeneRIF (Gene Reference Into Function, www.ncbi.nlm.nih.gov/gene/about-generif), which is a manually curated functional annotation source for genes based on literature [86]. Over 64% (12,449 genes) of genes have no more than 10 GeneRIF annotations, while TP53 has around 8,000 annotations; the 20% best annotated genes have in total 77.74% of all annotations (Figure 1.6D-F). The coverage of genes in the literature has also similar patterns [87]. We summarized the number of papers for each gene based on publications archived in PubMed, and saw nearly 40% (7,655 genes) of all protein coding genes were covered

in no more than 10 papers (Figure 1.6G). With no surprise, many of the well-known genes, including *TP53*, *TNF*, and *EGFR*, are among the most studied genes (Figure 1.6H). In summary, more than 80% of the publications, or research efforts have been focusing on only 20% of the genes (Figure 1.6I).

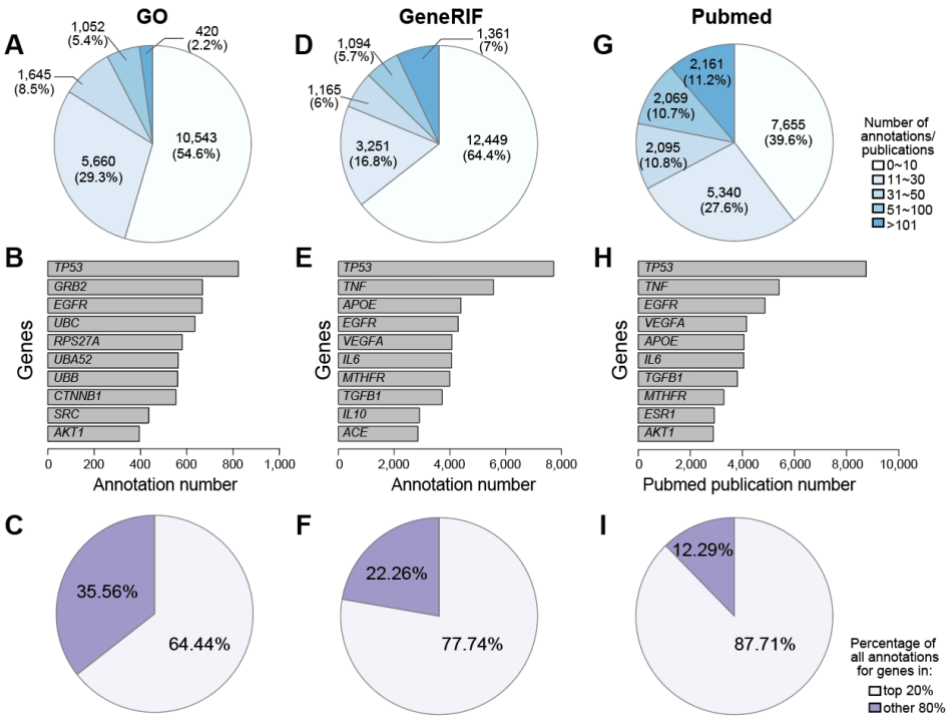


Figure 1.6 Known annotations for human genes

The number of annotations per gene for human genes in GO (**A**), GeneRIF (**D**), and number of publications in Pubmed (**G**). The top 10 genes with the most annotations/publications in GO (**B**), GeneRIF (**E**), and PubMed (**H**) are rank ordered. The percentage of all annotations/publications covering the top 20% most annotated genes in human (**C**, **F**, **I**).

From this perspective, it is clear that most human genes are still poorly annotated. The pattern is the same in other model species. Specifically, over 48% (10,166 genes) in mouse, 60% (11,833 genes) in rat, 61% (8,514 genes) in fly, 29% (5,885 genes) in worm, and 26% (1,566 genes) in yeast, have fewer than 10 entries in GO. The phenomenon that many genes are ignored in biological research has been pointed out before [88-90]. Several possible reasons for this bias, such as prior knowledge, publication bias, and priorities of funding support have been raised [88, 89, 91]. Therefore, an unbiased approach for gene function analysis would most likely provide many novel insights for future research.

1.3 Aims of the thesis

Given the fact that most of the genes are not characterized, I proposed, developed, and applied several systems genetics approaches to identify the potential gene functions using publicly available datasets with an unbiased manner. The aims of my studies were :

Aim 1 : To reveal the associations between genes and complex traits using multi-omics datasets from the BXD mouse cohort through novel systems genetics strategies.

Aim 2 : To identify the potential functions of genes and propose new components of pathway modules using large-scale transcriptome datasets.

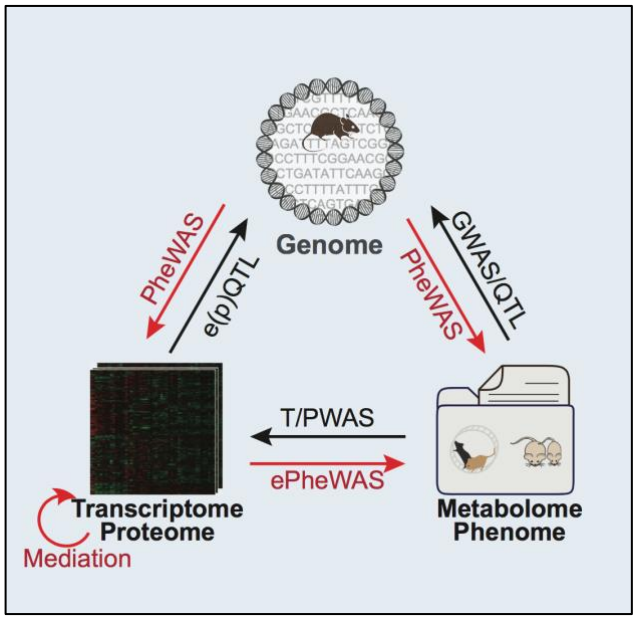
Chapter 2 An integrated systems genetics and omics toolkit to probe gene function

The postprint version of this part of work has been published in ***Cell Systems***.

Li H*, Wang X*, Rukina D, Huang Q, Lin T, Sorrentino V, Zhang H, Bou Sleiman M, Arends D, McDaid A, Luan P, Ziari N, Velázquez-Villegas LA, Gariani K, Kutalik Z, Schoonjans K, Radcliffe RA, Prins P, Morgenthaler S, Williams RW, Auwerx J. ***Cell Syst***. 2018 Jan 24;6(1):90-102.e4. doi: 10.1016/j.cels.2017.10.016.

Contribution to this work : I conceptually designed the study, curated data, developed the methodology, performed the analysis, created the web source, and wrote the manuscript.

2.1 Graphic abstract



2.2 Abstract

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 multi-facetted 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.

2.3 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 [36, 92]. 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 [36, 93-96].

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 [97, 98]. 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 [71, 99]. 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 [100]. These early approaches, however, do not exploit the full spectrum of possible relationships between genotypes, intermediate phenotypes, and clinical phenotypes (see Figure 2.1A-B). 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, DGRP fly, or the 1001 Genomes Project *A. thaliana* genetic reference 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 [22, 46, 73] 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 [74, 75], 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 [47], 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.

2.4 Methods

2.4.1 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).

2.4.2 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. 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 [67] and LXS mouse populations [38], as well as the HXB/BXH rat cohort [101] were retrieved from GeneNetwork. Data from HMDP was downloaded from http://phenome.jax.org/ and supplemental materials of [32]. Proteome data from DO population was downloaded from the supplemental materials of [47].

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/) [102, 103]. 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) [104].

***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* (Y46G5A.17). 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 [105].

2.4.3 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 [106] based on random forests [107]. 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 [108]. 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 [109]. 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 [110]. 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 (PheWAS)

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 [76]. 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) [111] 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 efficient mixed-model association (EMMA) [76]. 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 (T/PWAS)

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 (ePheWAS)

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 [76]. 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 [47]. 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” [47].

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.

2.4.4 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/lihaone/PheWAS.

2.5 Results

2.5.1 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 2.1A). 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, D2B6F1), that collectively encompass the vast majority of all BXD data.

Genome—Five million sequence variants segregate in the BXD family [100]. 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 2.1C). There is also more subtle genetic similarity among those BXDs (BXD43 to BXD102) that were produced by inbreeding advanced intercross progeny [112]. We compensated for this kinship in our statistical analyses.

Phenome—Since the first publication on the BXDs [113], 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 S2.1A). 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 2.1D). 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 2.1F) [114]. 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 2.1G, top), others do not and contain outliers (Figure 2.1G, bottom). To establish a robust analysis pipeline, we transformed the phenotypes into a standard normal distribution.

Transcriptome, proteome, *and metabolome*— ~200 transcriptome datasets from 34 BXD tissues existed (Figure 2.1E, Figure S2.1B). One or two datasets were selected to represent the transcript profiles in each tissue. Furthermore, other molecular data in our analyses include ~2,600 liver proteins quantified by SWATH-MS, and ~980 metabolites measured in liver [22] 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 2.1B) into a streamlined workflow (Figure 2.1H) to identify gene function. In the prioritization of PheWAS candidate genes, we included not only genes with high impact variants [100], 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 [115]. 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.

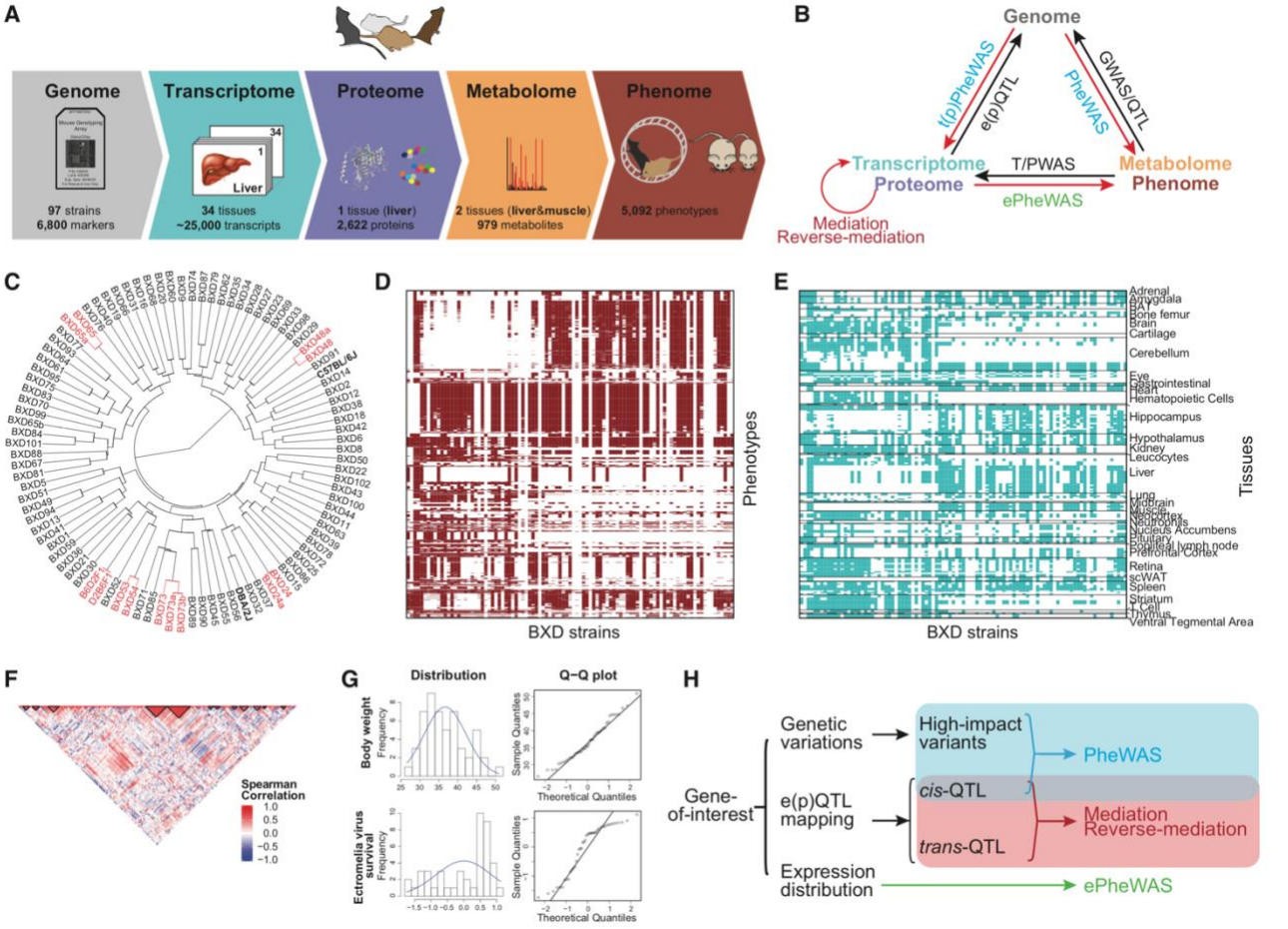


Figure 2.1 Overview of multi-omic data from the BXD population, and the scheme of applied systems approaches

A, Multi-omic data of the BXD population. Genome: genotype data was collected for 6,800 markers. Transcriptome: levels of ~25,000 transcripts have been measured from 34 tissues (See also Figure **S2.1B**). Proteome: expression of ~2,600 proteins were quantified in livers by mass spectrometry [22]. Metabolome: ~980 metabolites have been measured in both liver and muscle [22]. Phenome: ~5,000 clinical phenotypes have been collected by more than 200 research groups (See also Figure **S2.1A**).

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 [114] 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.

2.5.2 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 [71, 99] enabling the detection of pleiotropic effects of genetic variants (Figure S2.2A). We recently applied PheWAS to the BXDs using Pearson's correlation [100], 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 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 [109]. 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 S2.2B). Associations between the genetic variants of each gene and clinical and molecular phenotypes (transcripts, proteins and metabolites) were performed using EMMA [76]. A simplified flowchart representing our updated PheWAS approach is depicted in Figure 2.2A.

We performed both forward (e.g. GWAS) and reverse genetic approaches (e.g. PheWAS) on genome and phenome data in the BXDs (Figure 2.2B). For example, GWAS on the prepulse inhibition (PPI) of acoustic startle response mapped a significant signal on Chr19. 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 2.2C). PheWAS for *Pten* revealed several associated traits, including PPI and subcutaneous white adipose tissue (subWAT) mass (Figure 2.2D), suggesting pleiotropic effects of *Pten*. The links between *Pten* and neurobiological and metabolic phenotypes have been confirmed by independent studies [116, 117]. 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 (Figure 2.2E).

We compared the performance of the original and updated PheWAS methods [100], taking *Tlr5* (Toll-like receptor 5) as an example (Figure 2.2F). Both methods associated *Tlr5* with T cell proliferation, befitting its known function in immune response [118]. However, our initial method yielded C3 dicarboxylylcarnitine, anxiety assay, and serum amyloid P component as false positives (Figure 2.2G), due to the failure to control for population structure, as indicated by the inflation of p values in the QQ plot (Figure 2.2H).

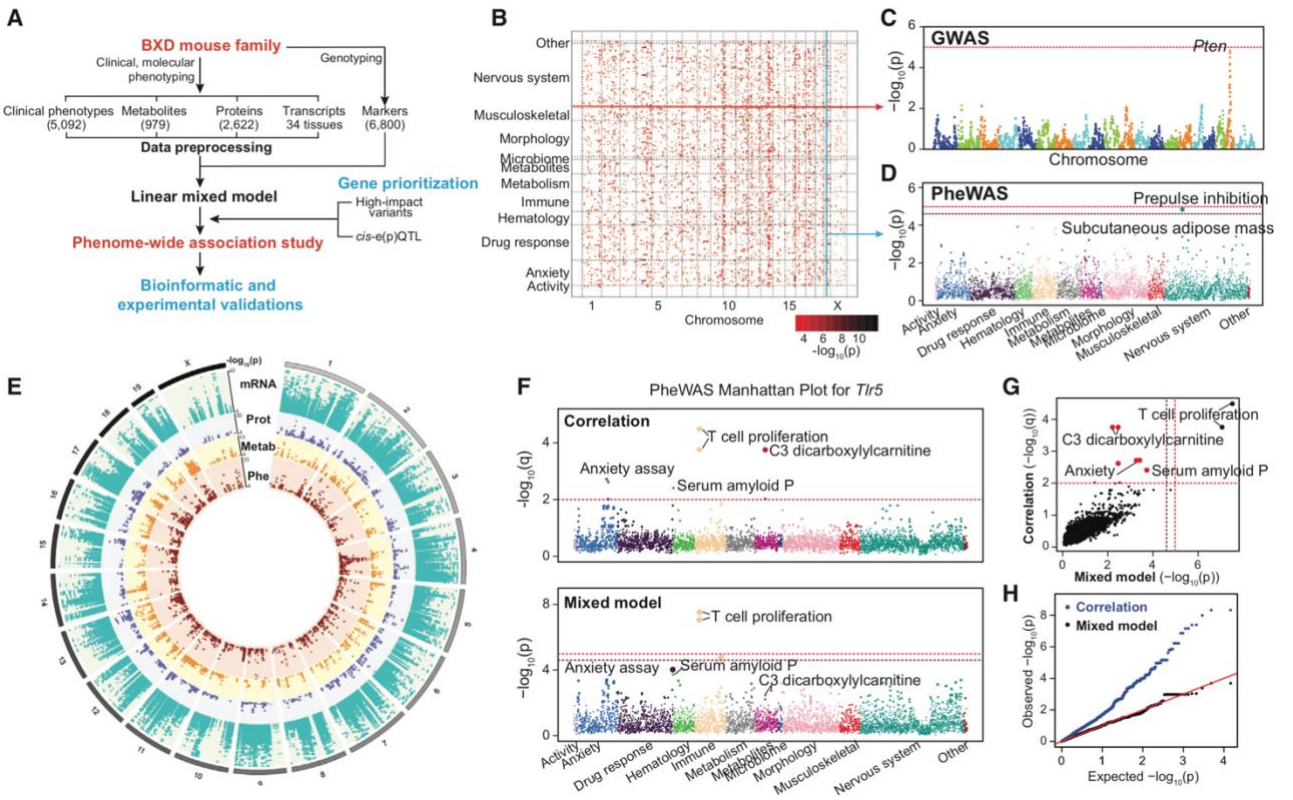


Figure 2.2 Phenome-wide association analysis

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. 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 × 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 × 10^{-5}, red dashed line) as well as the effective number (0.05/2,754 = 1.8 × 10^{-5}, dark red dashed line) of phenotypes.

E, Circos plot showing all the significant associations of prioritized genes in the multilayered PheWAS. Genomic positions of genes on chromosomes are labeled on the outer edge, with multilayered 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 FDR 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-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 QQ plot (**H**).

A few more examples illustrate the utility 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 Chr11 (Figure 2.3A). From this region, *Rpl26* (ribosomal protein L26) stood out as a strongest candidate with *cis*-eQTLs in many tissues, including liver (Figure 2.3B). Through PheWAS, we identified a link between *Rpl26* and body weight, fat mass, as well as oxygen consumption (VO_2) (Figure 2.3C). 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 2.3D), and further validated in several independent datasets, including an F2 cross between CAST/EiJ and C57BL/6J (CTB6F2, GN172) [67] (Figure 2.3E), the Hybrid Mouse Diversity Panel (HMDP) [32] (Figure 2.3F), as well as in the HXB/BXH rat cohort [101] (Figure 2.3G). The correlations of *Rpl26* with metabolic traits translate well across populations and species, and suggest a role of *Rpl26* in regulating

body weight.

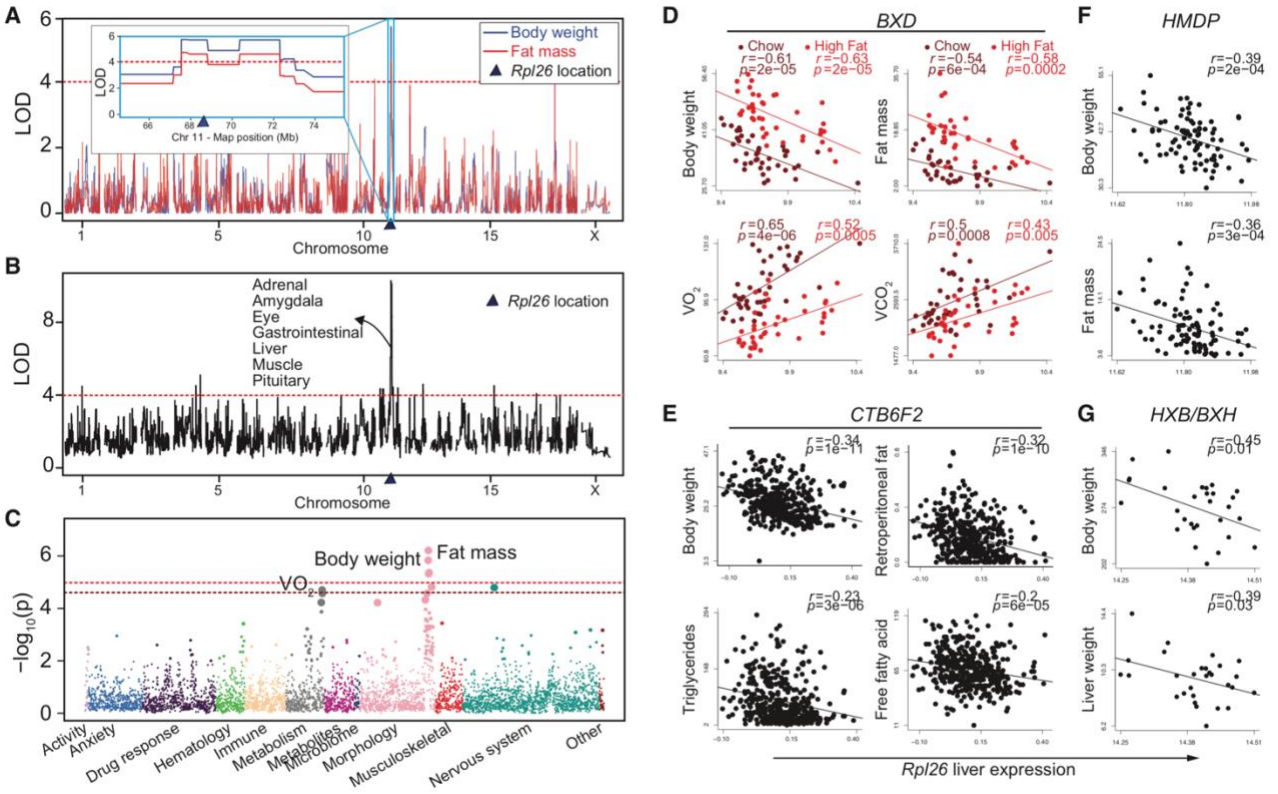


Figure 2.3 Genotype-phenotype associations revealed by PheWAS

A, QTL mapping of body weight and fat mass showed a common QTL on Chr11, 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 **2.2D**.

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.

Through PheWAS, we also confirmed the link between *Oprm1* (opioid receptor, mu 1) and morphine response [119] (Figure S2.3A-B). 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 S2.3C).

2.5.3 Expression-based PheWAS (ePheWAS)—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 [120]. Intermediate phenotypes, including transcript and protein levels, integrate the effects from genetic factors, including those poorly captured or hidden in common association studies [73], as well as from environmental factors. A few recent studies have explored the use of transcriptome-/proteome-wide association using either imputed transcript expression [74, 121] or proteomic data [75]. Given that transcriptome data are available for over 30 tissues, the BXDs are a perfect resource for such analysis. Linear mixed model was applied to find associations between gene expression and clinical phenotypes while accounting for population structure across strains [76] (Figure 2.4A). Forward genetics strategies could link phenotypes to tissue-specific transcript levels in T/PWAS (Figure 2.4B- red line). Conversely, reverse approaches starting from expression of the gene-of-interest towards the phenome, i.e. ePheWAS, could reveal the gene's potential pleiotropic functions, especially when considering its expression across multiple tissues (Figure 2.4B- blue line). The numbers of

G2P associations that survive the phenome-wide significance threshold differed across tissues and across phenotypic categories (Figure 2.4C). 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 [122], suggesting that many phenotypic traits are under tissue-specific regulations.

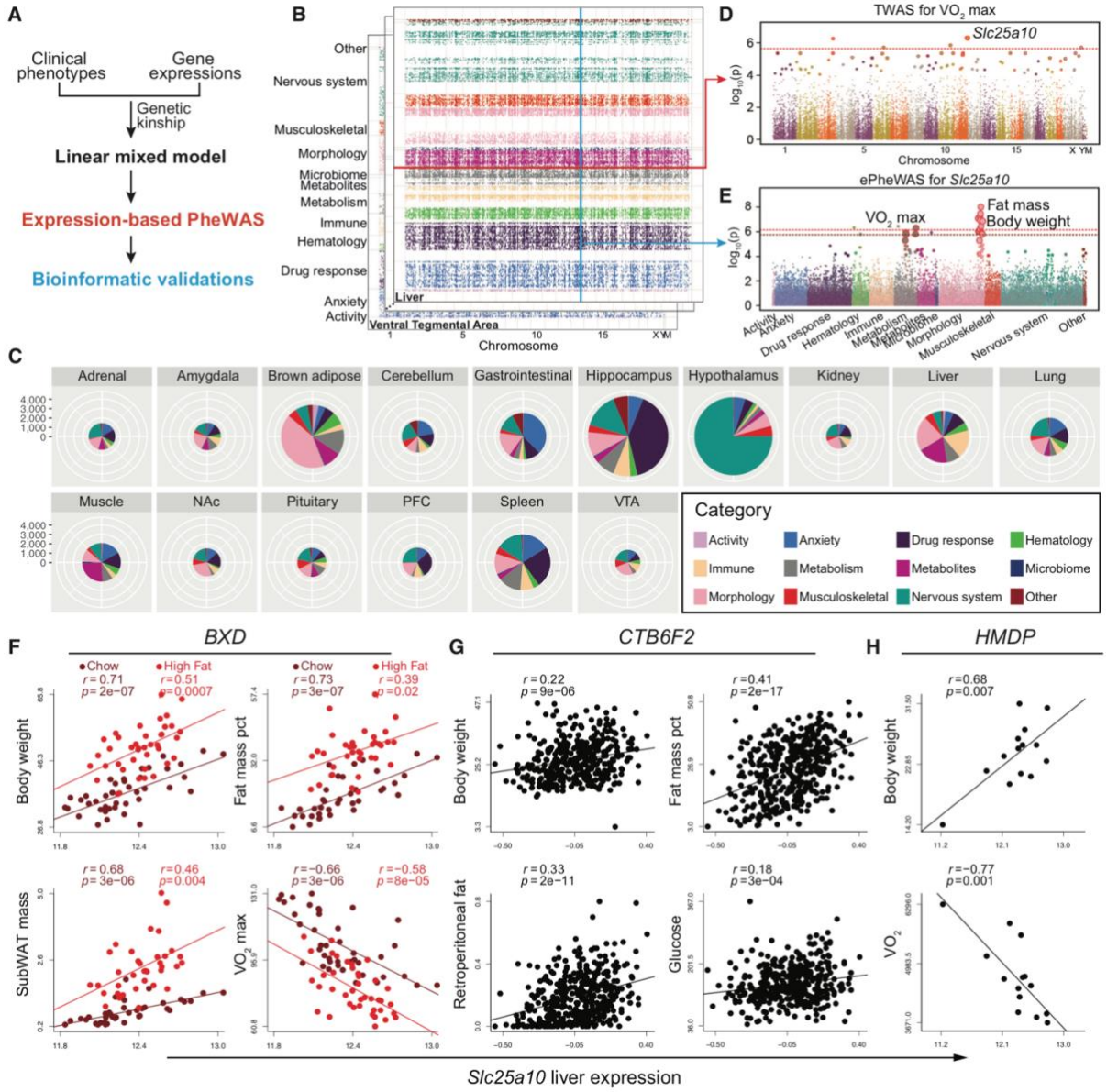


Figure 2.4 ePheWAS displays tissue-specific regulators

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 plot, 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 candidate to explain changes in VO_2 max across the BXDs. Transcripts were arranged by the 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 Figure **2.4C**), together with the total number

(0.05/4,784/16 = 6.5 × 10^{-7}), as well as the effective number of phenotypes (0.05/2,754/16 = 1.1 × 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.

TWAS in liver (GN432) identified *Slc25a10* as a potential regulator for VO_2 max (Figure 2.4B and D). *Slc25a10* exports malonate, malate, and succinate across the mitochondrial inner membrane for fatty acid synthesis in the cytosol [123]. Through ePheWAS, we found that *Slc25a10* not only associated with VO_2, but also with body weight and fat mass (Figure 2.4B and E). Furthermore, *Slc25a10* liver expression correlated positively with body weight, fat mass, and subWAT mass, and negatively with VO_2 in both CD and HFD fed BXDs (Figure 2.4F). We found comparable correlations with similar metabolic traits in the CTB6F2 (Figure 2.4G) and the HMDP (Figure 2.4H), 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 [124]. There are notable differences in weight loss after an overnight fast across the BXDs (ranging from 0.8-3.9 on CD and 0.3-3.5 grams on HFD), although there was no significant difference between CD and HFD cohorts (Figure 2.5A). However, no genetic variant was found to be associated with fasting weight loss using QTL mapping (Figure 2.5B). Through TWAS using liver transcripts, we detected *Cpt1a* as the top candidate associated with fasting weight loss in both CD (Figure 2.5C) and HFD cohorts (data not shown). ePheWAS showed further associations of *Cpt1a* with plasma acylcarnitine levels (Figure 2.5D). Liver *Cpt1a* levels correlated positively with acylcarnitines (Figure 2.5E, upper), which corresponds to the recognized function of *Cpt1a* in transferring the acyl group of long-chain fatty acyl-CoA to carnitine for further *beta*-oxidation in mitochondria [125]. Strains with higher *Cpt1a* expression tend to lose more weight upon fasting, and have lower plasma triglycerides (Figure 2.5E, bottom). Furthermore, we validated the correlations between *Cpt1a* and metabolic phenotypes in another independent mouse population, i.e. the HMDP (Figure 2.5F), and in *C. elegans*, where feeding an RNAi targeting *cpt-1*, the *Cpt1a* worm homolog, lowered lipid content (Figure 2.5G); this highlights the cross-species conservation of its role in lipid metabolism.

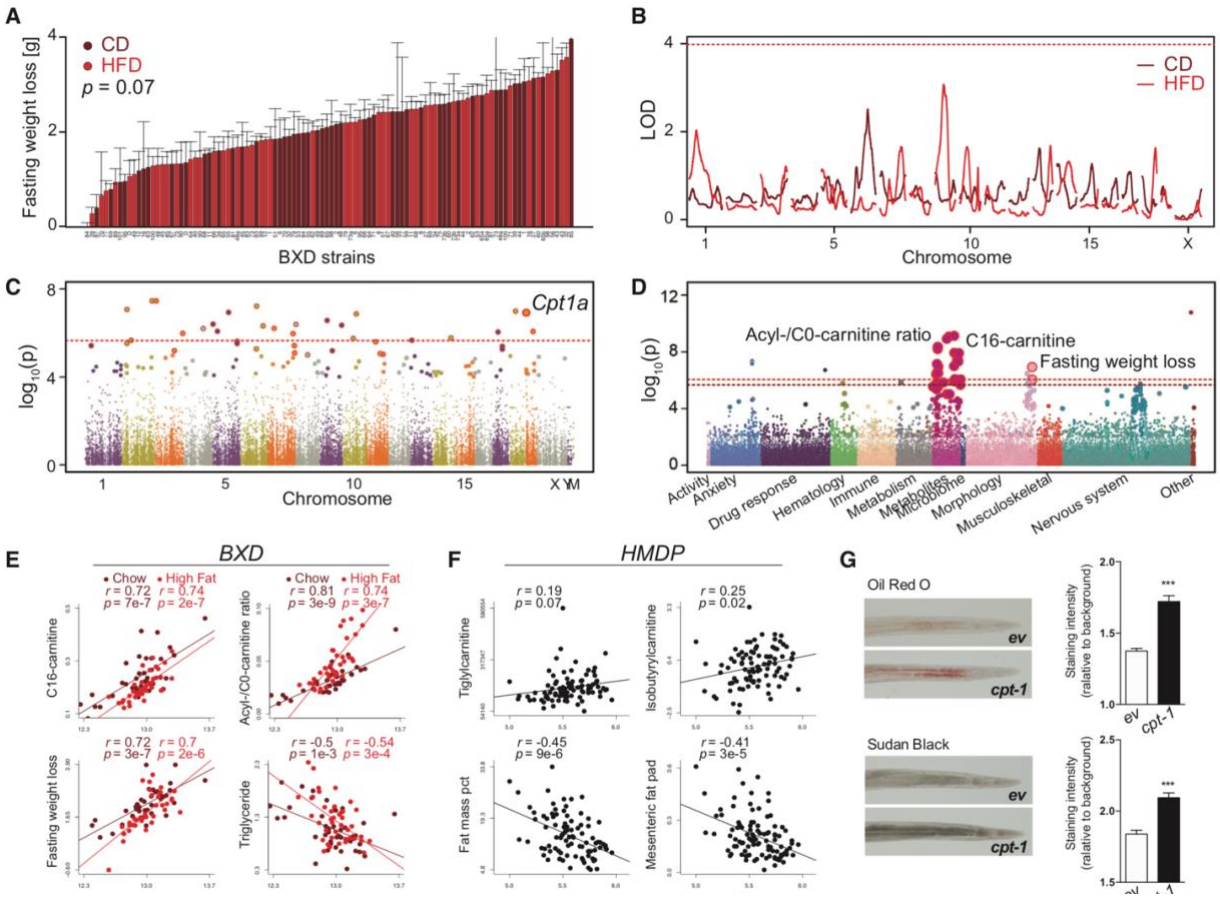


Figure 2.5 ePheWAS reveals *Cpt1a* as a regulator of fasting weight loss

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 **2.4D**.

D, ePheWAS for *Cpt1a* identified its association with carnitine levels and fasting weight loss. Same as Figure **2.4E**, phenome-wide significance was adjusted by Bonferroni correction for the numbers of used tissues and phenotypes.

E-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, Knockdown 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.

Using ePheWAS, we also identified associations between *Cd36* liver transcripts and fat mass and acid beta-glucosidase activity (Figure S2.4A). 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 beta−glucosidase activity (Figure S2.4B-C), confirming the involvement of *Cd36* in metabolism [126] and suggesting a potential role in Gaucher's disease, which results from the deficiency of acid beta−glucosidase [127]. An association between *Abca8a* liver transcripts and triglyceride levels was also revealed (Figure S2.4D). 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 (Figure S2.4E-F). This substantiates a role for this poorly characterized ABCA protein in lipid transport, similar to many other ABCA transporters [128].

2.5.4 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 (Figure 2.6A, D). 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 2.6B) linearly increased with the number of strains sampled; so did the number of recovered hits (red curve, Figure 2.6B). In all subsamples, more than ~75% of the hits are recovered hits, which implies that the associations are robust (Figure 2.6B). 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 2.6C). Interestingly, subsamples of as few as 20 strains share ~40-50% of their associations. We also performed subsampling analysis on ePheWAS looking for significant associations between gene expression in eye and the clinical phenome, and observed a similar influence of sample size on performance (Figure 2.6E-F). However, there was a more rapid reduction of true positives with decreasing sample size compared with PheWAS (Figure 2.6B), mainly due to the incompleteness of the phenotype data (i.e. different laboratories sampling different lines, shown in Figure 2.1D).

Over all, 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.

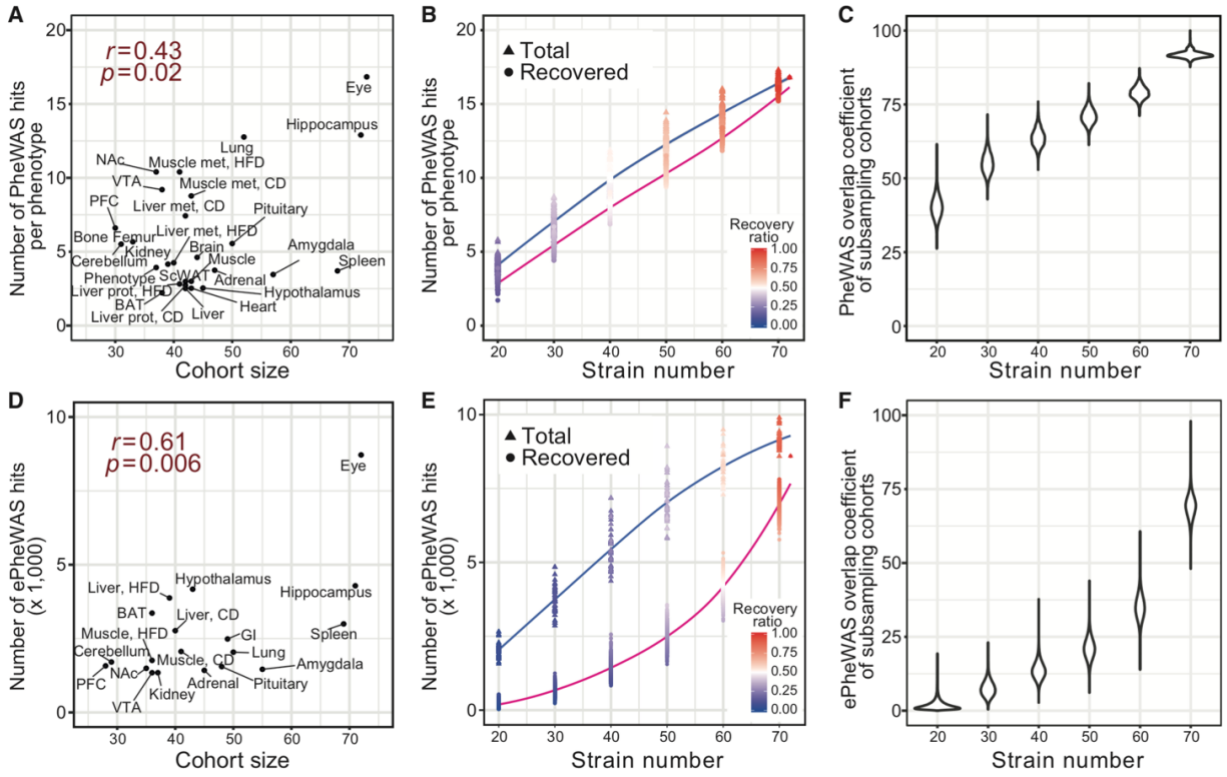


Figure 2.6 Performance of PheWAS and ePheWAS in detecting G2P associations

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

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

2.5.5 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 [129]. This concept has also been applied to reveal the mediating role of gene expression in the association between SNPs and clinical phenotypic variations [130] or *trans*-regulated genes [47, 131].

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 2.7A), suggesting that the modulation of gene expression is tissue- and environment-specific [132, 133]. For example, the eye transcriptome shows a *trans*-eQTL hotspot on Chr 1. *Pou2f1*, a gene involved in lens placode development [134], 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 2.7B). While on the contrary, reverse-mediation investigates the mediated variables of a potential mediator (gene with a *cis*-QTL) (Figure 2.7I).

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 (BCKDH) 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 *Bckdhb* and BCKDHB *cis*-pQTL (Figure 2.7C). Mediation analysis revealed that BCKDHB is a potential mediator of BCKDHA protein levels (Figure 2.7D). The mediation results were further confirmed in the liver protein dataset (GN705) from BXDs fed with HFD (Figure S2.5), as well as in the diversity outbred (DO) mouse cohort [47] (Figure 2.7E-F). BCKDHB correlated with BCKDHA protein levels in both BXD (Figure 2.7G) and DO cohorts (Figure 2.7H). This demonstrates that the mediation effect of BCKDHB on BCKDHA is independent of environmental influences (e.g. diet) and conserved across populations.

Mediation analysis was also performed on *Rpsa* and *Rps2*, components of the 40S ribosomal subunits, to determine the upstream regulation factors (Figure S2.6). *Rpsa* and *Rps2* have *trans*-eQTLs on Chr 12 in many tissues, including brain and hippocampus (Figure S2.6B, E), suggesting a shared regulation. Mediation revealed *Zfp277* as the potential regulator of *Rpsa* and *Rps2* (Figure S2.6C, F). *Zfp277* strongly co-expressed with *Rpsa* and *Rps2* (Figure S2.6D, G). Furthermore, the mediating role of *Zfp277* on *Rpsa* and *Rps2* was confirmed using data from prefrontal cortex (Figure S2.6H-J, GN130) and brain (Figure S2.6K-M, GN784) of the LXS GRP [38]. As all three genes have been linked to cancer, we then tested whether they co-express in cancer. Based on RNA-seq 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 [102, 103] (Figure S2.6N), 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 2.7I). *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 2.7J). Reverse-mediation revealed the mediating role of *E2f6* on the expression of these genes (Figure 2.7K). In addition, *E2f6* transcripts positively correlated with these genes in the eye (Figure 2.7L). This finding led to the hypothesis that *E2f6* binds to the regulatory regions of these genes, which was confirmed by human ChIP-seq data from ENCODE [135] (Figure 2.7M), 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) (Figure S2.7A-C). These target genes showed *trans*-eQTLs mapping to the genetic locus of *Zkscan1* (Figure S2.7B), and were tightly co-expressed with *Zkscan1* (Figure S2.7D). ENCODE ChIP-seq data confirmed the binding of ZKSCAN1 on the promoter region of the human orthologs of the identified candidates [135] (Figure S2.7E).

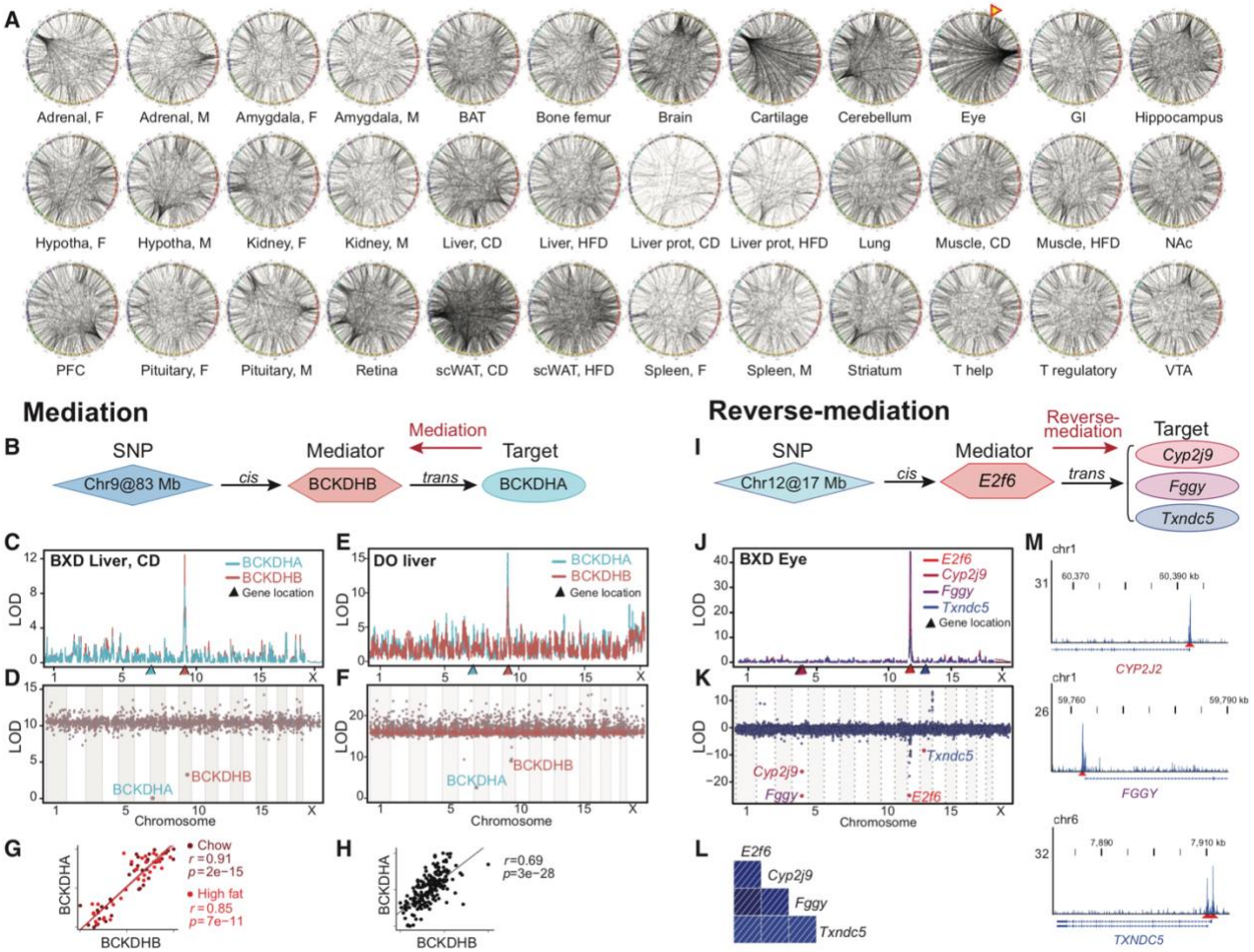


Figure 2.7 Mediation and reverse-mediation analysis discovers gene interactions

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. BAT, brown adipose tissue. GI, Gastrointestinal. Hypotha, hypothalamus. Liver prot, liver protein. NAc: nucleus accumbens. PFC: prefrontal cortex. ScWAT, subcutaneous white adipose tissue. VTA, ventral tegmental area.

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, 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, F, Mediation plot of BCKDHA in liver proteomic datasets from BXD (**D**) and DO (**F**) mice showing that BCKDHB is a mediator of BCKDHA.

G, H, Significant correlation between BCKDHA and BCKDHB protein levels in livers of either CD or HFD fed BXDs (**G**), as well as in the DO mice (**H**) [47].

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*, *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.

2.6 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 to the original PheWAS methodology in mouse [100], 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 [136]. 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 [47], 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 [137], or TCGA [102], or the 1001 Genomes Project for *A. thaliana* [138]—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 utility of these tools may require cohorts that have extensive multi-omics datasets available or have relevant samples biobanked, e.g. the Framingham Heart Study [139]. Imputation of gene expression in deep tissues from either reference transcriptome data sets [140] or GWAS summary statistics [74] 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.

2.7 Supplemental figures

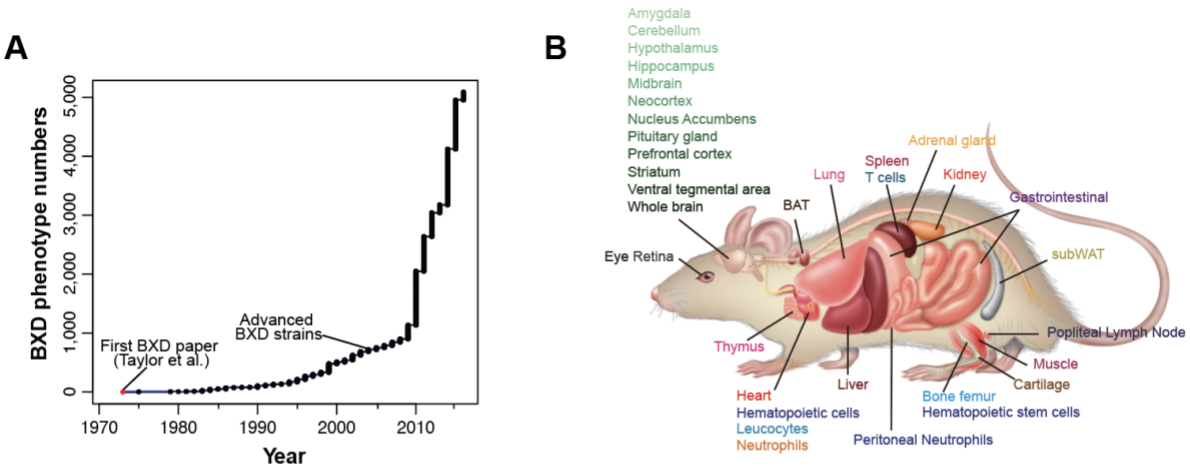


Figure S2.1. BXD phenome and transcriptome, Related to Figure 2.1.
A, Number of phenotypes collected on the BXD population increases over years.
B, Graphical illustration of the tissues profiled at the transcriptomic levels from the BXDs.

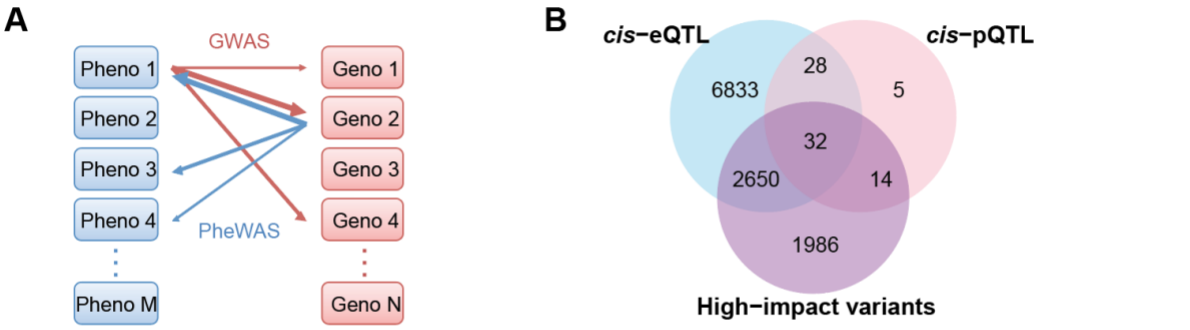


Figure S2.2. Scheme and statistical summary for PheWAS, Related to Figure 2.1, Figure 2.2.
A, Conceptual scheme comparing GWAS and PheWAS. Arrows represent connections between phenotypes and genotypes, and line weights represent the significance of associations.
B, Number of genes containing high-impact genetic variants and *cis*-QTLs that fit the requirement of PheWAS in BXD datasets.

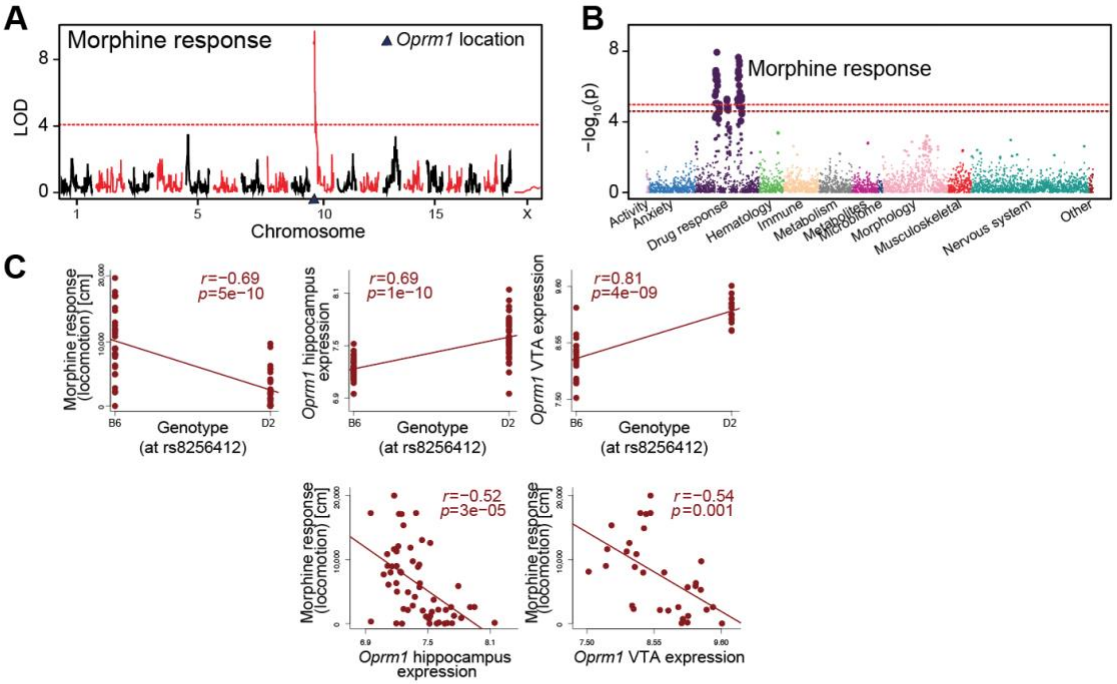


Figure S2.3. PheWAS reveals the association of *Oprm1* and morphine response, Related to Figure 2.2.
A, QTL mapping suggests the involvement of *Oprm1* in morphine response.
B, PheWAS for *Oprm1* identifies its association with a series of morphine response phenotypes.
C, The D allele of *Oprm1* genotype associates with decreased locomotion activity after morphine injection and up-regulation of *Oprm1* transcript levels in many neurological tissues, including the hippocampus and ventral tegmental area (VTA). Transcript levels of *Oprm1* in hippocampus and VTA correlate with morphine response phenotypes in the BXDs.

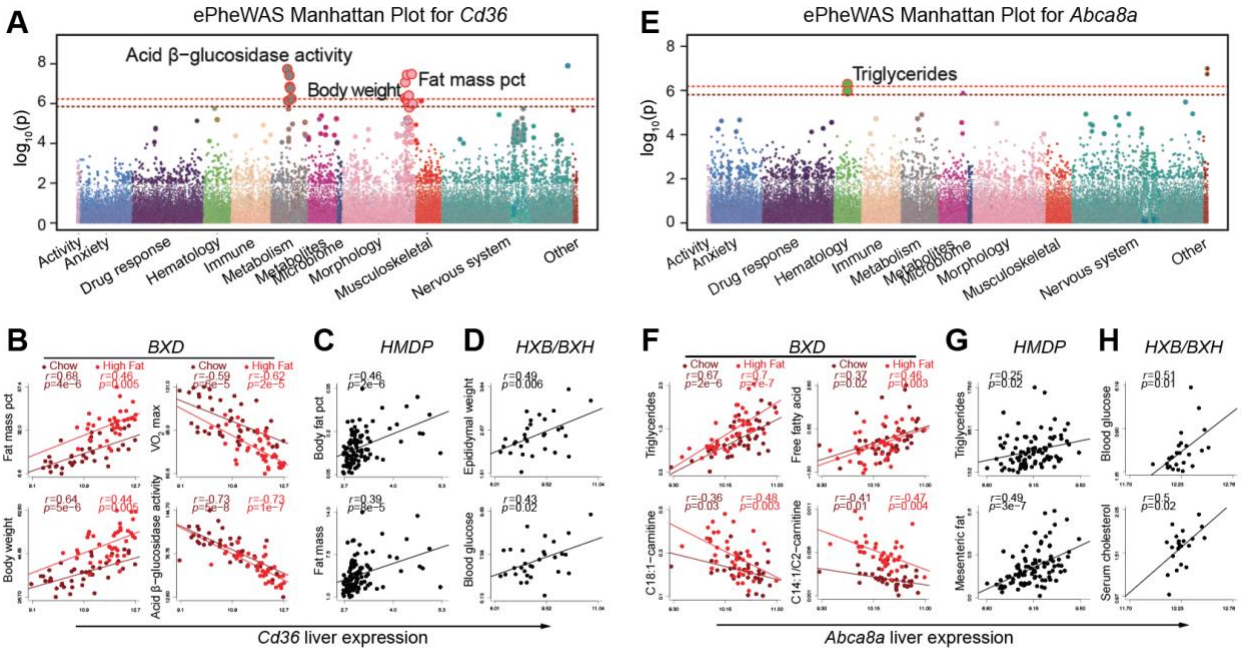


Figure S2.4. ePheWAS demonstrates metabolic functions of *Cd36* and *Abca8a*, Related to Figure 2.4.
A, ePheWAS for *Cd36* identifies its metabolic role in the modulation of fat mass and acid beta-glucosidase activity.
B-D, Correlation analysis of liver *Cd36* transcripts and metabolic phenotypes, e.g. fat mass, body weight, and oxygen consumption, in the BXD (B), HMDP (C) and HXB/BXH (D) populations.
E, ePheWAS reveals the association of *Abca8a* and triglycerides.
F-H, Liver *Abca8a* correlates with a list of metabolic traits, including plasma triglycerides, free fatty acids, and carnitine levels in the BXD (F), HMDP (G), HXB/BXH (H) populations.

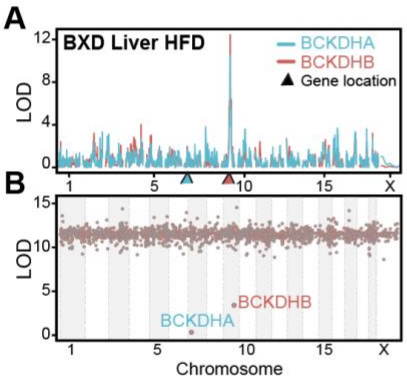


Figure S2.5. Mediation analysis of BCKDHA in BXD liver HFD proteomic datasets, Related to Figure 2.7.
A, pQTL mapping of BCKDHA and BCKDHB in livers from HFD fed BXD mice. BCKDHA exhibits a *trans*-pQTL that maps to chromosome 9, where the BCKDHB *cis*-pQTL locates.
B, Mediation plot of BCKDHA in liver HFD proteomic datasets showing that BCKDHB is a mediator of BCKDHA.

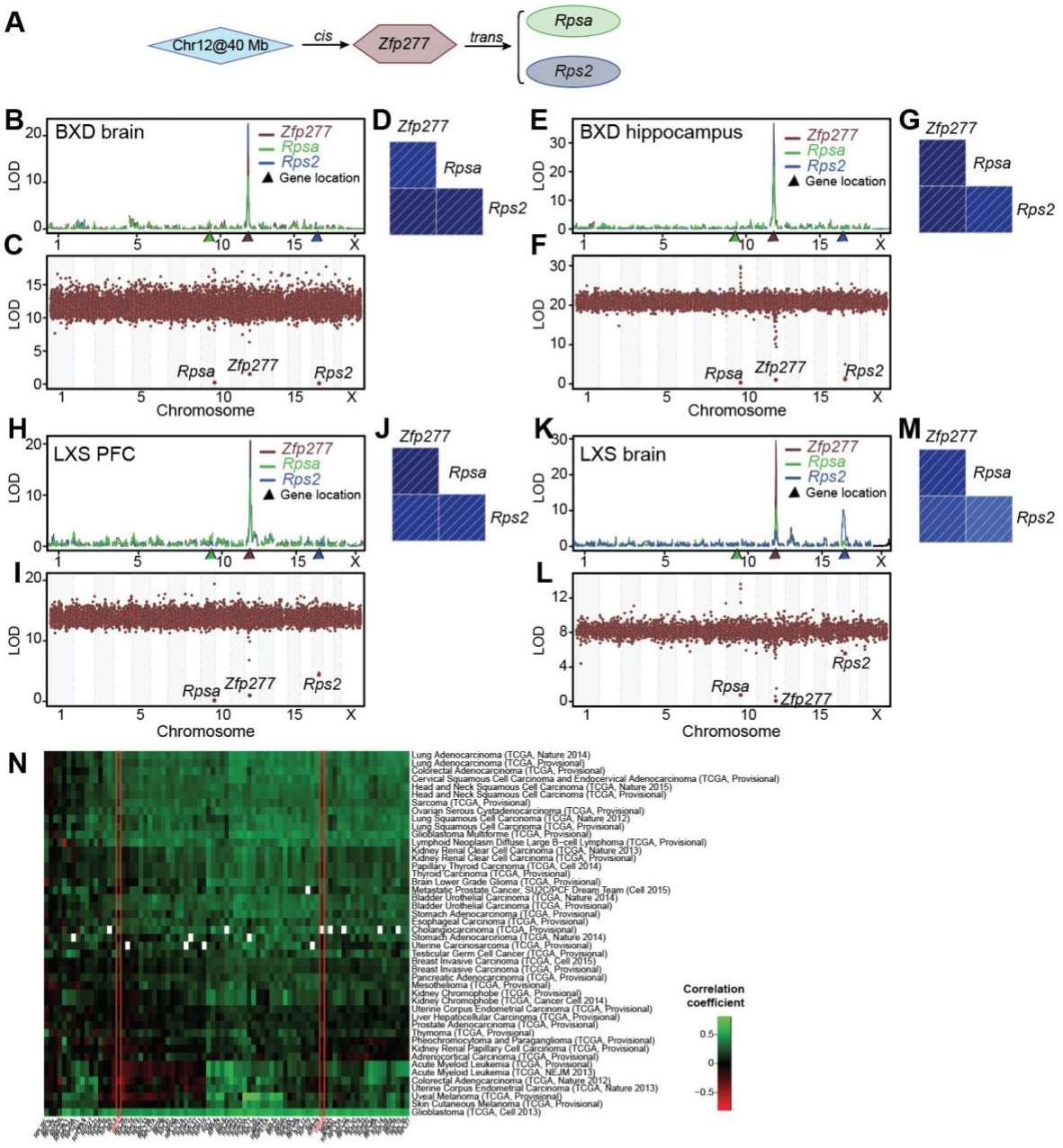


Figure S2.6. Mediation analysis reveals the influence of *Zfp277* on *Rpsa* and *Rps2*, Related to Figure 2.7.
A, Scheme showing the mediation model. The genetic variant underlying the *cis*-eQTL of *Zfp277* influences the expression levels of several genes, including *Rpsa* and *Rps2*, in *trans*.
B-J, Mediation analysis revealed *Zfp277* as the mediator of *Rpsa* and *Rps2* in transcriptome data from brains (B-D, GN113) and hippocampi (E-G, GN110) of the BXD cohort, as well as in transcriptome data from prefrontal cortex (PFC, H-J, GN130) and brains (K-M, GN784) of LXS cohort.
B, E, H, K, eQTL mapping of *Zfp277*, *Rpsa* and *Rps2* in the respective transcriptome datasets. *Rpsa* and *Rps2* possess significant trans-eQTLs that map to the same locus of *Zfp277* cis-eQTL.
C, F, I, L, Mediation plots of *Rpsa* validate the potential mediating effects of *Zfp277* on *Rpsa* and *Rps2*.
D, G, J, M, *Zfp277* expression levels significantly correlate with *Rpsa* and *Rps2*.
N, *ZNF277* co-expresses with the ribosomal protein family genes, including *RPSA* and *RPS2* (highlighted in red boxes), in biopsies from 35 different cancer types. Pearson's correlation coefficients between the expression of *ZNF277* and the ribosomal protein family genes in 46 RNA-seq data sets from 35 different cancer types are indicated in the heatmap.

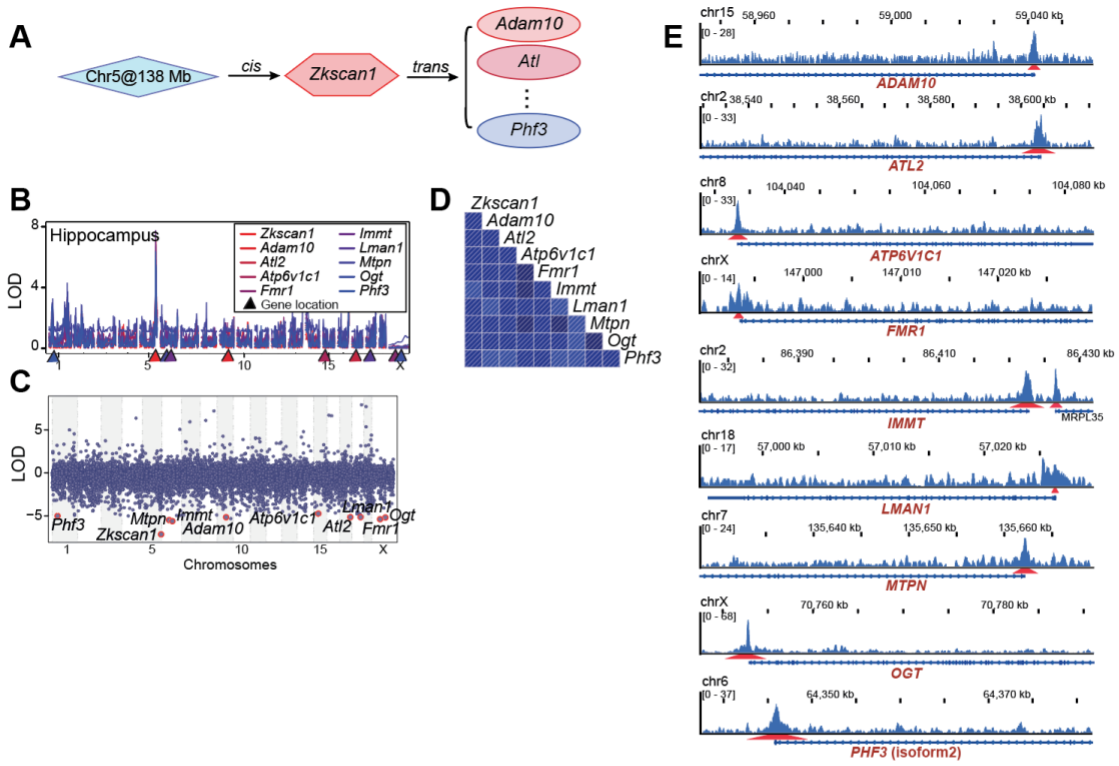


Figure S2.7. Mediation effects of *Zkscan1* on its target genes by reverse-mediation analysis, Related to Figure 2.7
A, Scheme showing the mediation model of *Zkscan1*. The genetic variant underlying the *cis*-eQTL of *Zkscan1* influences the expression levels of a list of genes, including *Adam10* and *Atl*, in *trans*.
B, eQTL mapping of transcripts of *Zkscan1* and several potential *Zkscan1* transcriptional targets, including *Adam10* and *Atl*, in BXD hippocampi (GN110). All these target genes map significant *trans*-QTLs to chromosome 5, where the *Zkscan1 cis*-eQTL locates.
C, Reverse-mediation plot of *Zkscan1* showing its mediation effects on its target genes, which are highlighted in red.
D, Correlation between the expression levels of *Zkscan1* and its target genes in the hippocampus transcriptome.
E, Binding of ZKSCAN1 on the promoter loci of the human orthologs of the target genes in ENCODE. The binding activity of ZKSCAN1 on the DNA of its target genes is shown and the predicted binding sequence of ZKSCAN1 is indicated in red. Genes are arranged based on their genetic position in human chromosomes.

2.8 Acknowledgements

We are grateful to the BXD community for generating the valuable resource for systems biology research and thank A.J. Lusis, G.A. Churchill and S.P. Gygi, M. Miles, and 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).

Chapter 3 Identifying gene function and module connections by the integration of multi-species expression compendia

The postprint version of this part of work has been published in ***Genome Research***.

Li H, Rukina D, David F, Li TY, Oh CM, Gao AW, Katsyuba E, Bou Sleiman M, Komljenovic A, Huang Q, Williams RW, Robinson-Rechavi M, Schoonjans K, Morgenthaler S, Auwerx J. ***Genome Res***. 2019. doi: 10.1101/gr.251983.119.

Contribution to this work : I conceptually designed the study, collected and curated data, developed the methodology, performed the analysis, created the web source, and wrote the manuscript.

3.1 Abstract

The functions of many eukaryotic genes are still poorly understood. Here we developed and validated a new method, termed GeneBridge, which is based on two linked approaches to impute gene function and bridge genes with biological processes. First, Gene-Module Association Determination (G-MAD) allows the annotation of gene function. Second, Module-Module Association Determination (M-MAD) allows predicting connectivity among modules. We applied the GeneBridge tools to large-scale multi-species expression compendia—1,700 datasets with over 300,000 samples from human, mouse, rat, fly, worm, and yeast—collected in this study. G-MAD identifies novel functions of genes, for example *DDT* in mitochondrial respiration and *WDFY4* in T cell activation, and also suggests novel components for modules, for example for cholesterol biosynthesis. By applying G-MAD on datasets from respective tissues, tissue-specific functions of genes were identified, for instance the roles of *EHHADH* in liver and kidney, as well as *SLC6A1* in brain and liver. Using M-MAD, we identified a list of module-module associations, such as those between mitochondria and proteasome, mitochondria and histone demethylation, as well as ribosomes and lipid biosynthesis. The GeneBridge tools together with the expression compendia are available at systems-genetics.org, to facilitate the identification of connections linking genes, modules, phenotypes, and diseases.

3.2 Introduction

The identification of gene function and the integrated understanding of their roles in physiology are core aims of many biological and biomedical research projects — an effort that is still far from being complete [87-90]. Traditionally, gene function has been elucidated through experimental approaches, including the evaluation of the phenotypic consequences of gain- or loss-of-function (G/LOF) mutations [141, 142], or by genetic linkage or association studies [36]. A large number of bioinformatics tools have been developed to predict gene function based on sequence homology [143-145], protein structure [144-146], phylogenetic profiles [147-149], protein-protein interactions [150-152], genetic interactions [153-155], and co-expression [156-162].

With the development of transcriptome profiling technologies, thousands of high-throughput studies have generated a wealth of genome-wide data that has become a valuable resource for systems genetics analyses. A few web resources, including GEO [163], ArrayExpress [164], GeneNetwork [165], and Bgee [166] amongst others, have created repositories of such expression data for curation, reuse, and integration. Several tools, such as GeneMANIA [156], GIANT [157], SEEK [167], GeneFriends [158], WeGET [159], COXPRESdb [160], WGCNA [162], and CLIC [161], are able to assign putative new functions to genes by means of correlations or co-expression networks. At their core, these methods rely on the concept of guilt-by-association – that transcripts or proteins exhibiting similar expression patterns tend to be functionally related [168]. By using over-representation analyses on sub-networks or modules, one can then deduce aspects of gene functions.

However, these approaches generally depend on discrete subsets of genes whose expression correlations exceed either a hard or soft threshold, which would strongly influence the final results. In addition, such analyses typically focus on positive or absolute values of correlations among datasets. The key polarity of interactions is often lost among gene products and linked modules [156-158, 161, 167]. Gene set analyses, such as gene set enrichment analysis (GSEA) [169], have been developed to identify processes or modules that are affected by certain genetic or environmental perturbations [170]. While GSEA uses all measured genes in the analysis, its application has mainly been limited to studying G/LOF models or environmental perturbations, where comparisons are inherently among discrete categories. This limits its applicability in most populations, in which variations among individuals are often subtle and continuous [36].

Here we developed the GeneBridge toolkit that uses two interconnected approaches to improve upon the identification of gene function and to bridge genes to phenotypes using large-scale cross-species transcriptome compendia collected for this study. *First*, we describe a computational approach, named Gene-Module Association Determination (G-MAD), to impute gene function. G-MAD considers expression as a continuous variable and identifies the associations between genes and modules. *Second*, we developed the Module-Module Association Determination (M-MAD) method to identify connections between modules based on the transcriptome compendia. The data and GeneBridge tools described here are available at systems-

genetics.org, an open resource, which will facilitate the identification of novel connections between genes, modules, phenotypes, and diseases.

3.3 Methods

3.3.1 Gene annotations / Modules

Gene ontology (GO) annotations [85] were downloaded from www.geneontology.org/ on Oct 4, 2017, with versions indicated by submission date below. Gene Reference Into Function (GeneRIF) [86] was downloaded from ftp://ftp.ncbi.nih.gov/gene/GeneRIF/ on Oct 11, 2017. Literature data for the genes was downloaded from PubMed at ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz on Mar 15, 2018.

Module data for all the species were retrieved from GO [85], Kyoto Encyclopedia of Genes and Genomes (KEGG) [171], and Reactome [172]. Annotations from GO with evidence codes of IEA (inferred from electronic annotation), ND (No biological data available), NR (Not recorded), NAS (Non-traceable author statement) were removed from the analysis. The parent-child hierarchical structure of GO was ignored. All modules, including the redundant modules (modules with similar gene components), as well as parent-child modules, were considered as independent in the analysis.

Modules with less than 15 genes or larger than 1,000 genes were excluded, resulting in 6,979, 7,489, 7,462, 3,811, 2,495, and 2,381 modules for human, mouse, rat, fly, worm, and yeast, respectively, for the analysis.

3.3.2 Module similarity calculation

Similarity between two modules were defined as the Jaccard index $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, i.e. the number of genes in A and B divided by the number of genes in A or B. It measures the intersection between the modules as a fraction of the total size.

3.3.3 Gene expression across tissues

Expression patterns of *EHHADH* and *SLC6A1* in mRNA and protein levels across human tissues were obtained from the Human Protein Atlas [173], and are available from v18.proteinatlas.org/ENSG00000113790-EHHADH/tissue and v18.proteinatlas.org/ENSG00000157103-SLC6A1/tissue, respectively.

3.3.4 Transcriptome datasets

Human GTEx transcriptome datasets were downloaded from https://www.gtexportal.org [174]. Most of the microarray and RNAseq datasets were downloaded from GEO [163] and ArrayExpress [164], with processed human and mouse RNAseq datasets obtained from ARCHS4 [175]. The rest of the datasets were downloaded from other sources, including the database of Genotypes and Phenotypes (dbGaP) [176], Mouse phenome database[177], and other data repository websites. Data from single cell RNA-seq were excluded from the study because they contain the many zero counts. Detailed information can be found at systems-genetics.org/datasets.

3.3.5 Data preprocessing of transcriptome datasets

For microarray datasets, the expression for a given gene with more than one probe set was represented by the average values of all its probe sets. Un-annotated probe sets were removed in the data pre-processing step. Only protein coding genes were considered in the analysis, as non-coding genes are often not well measured in microarray platforms. For RNAseq datasets, CPM (Count Per Million) were calculated to normalize the gene expression across samples and $\log_2$(CPM) were used for further analysis. Only protein coding genes were considered in the analysis to match the data in microarray datasets.

Transcriptome data were standardized by quantile-transformation to fit a normal distribution to avoid model misspecification when performing gene-level statistics. The expression values of all genes were normalized to

the range of 0 to 1. Samples and genes with more than 30% missing values were removed from the analysis, and the remaining missing data were imputed using nearest neighbor averaging by the *impute.knn* function in the "impute" R package.

For all the datasets, covariates were manually annotated and curated based on the metadata available from the respective data sources. Datasets containing data from different tissues were separated into single tissues. To account for confounding sources of expression variations, the effects of known covariates, including age, gender, genotype, platform, disease, treatment, batch, etc., as well as hidden determinants of gene expression were estimated and removed by using PEER (probabilistic estimation of expression residuals) [178], and the expression residuals were used for further analysis.

3.3.6 Gene-Module Association Determination (G-MAD)

G-MAD makes use of the PEER resulted expression residuals [178] of transcriptome datasets from large cohorts (datasets with over 80 samples). The expression levels of the gene-of-interest (target gene T) are used as a continuous trait to test whether a module M is enriched when T is highly expressed or, alternatively, whether it is depleted. The analysis uses the competitive gene set testing method CAMERA, which adjusts for inter-gene correlations [179]. This adjustment is important, because left unadjusted too many significant results would emerge. To perform CAMERA, we first regress all genes G on T according to the following relationship

$$G = \mu + \beta_{T\to G} T + e.$$

The fitting of this model equation to the observations is done separately for each data set by using the least squares method. The result is one fitted value $\beta_{T\to G}$ per gene. These coefficients define a set of statistics numerically characterizing the connection between the target gene T and any gene G. CAMERA provides a test of the null hypothesis that the average values of the β coefficients for the genes G in the module M are equal to the values for the genes not in the module. In order to correct for the inter-gene correlations a variance inflation factor is computed based on the average correlation coefficient $\bar{\rho}_M$ computed from the expression residuals obtained and only using the genes in the module M. When the average association scores between genes in the set and genes outside the set, $\frac{\sum_{G\in M}\beta_{T\to G}}{|M|}$ and $\frac{\sum_{G\notin M}\beta_{T\to G}}{|Genes\setminus M|}$, are compared on the final step, $\bar{\rho}_M$is included in the variance inflation factor. The resulting statistic revealing the association between the target gene T and M we refer to as the enrichment score $ES_M(T)$.

The same procedure was conducted for all the genes in the analyzed datasets to obtain the enrichment *p*-value matrix between genes and modules in all the datasets. Two types of analyses can be applied on the gene-module *p*-value matrix. One can extract the *p*-values for one gene against all modules across the datasets to obtain the association between this gene and all modules; or extract the *p*-values for one module against all genes to check the association between this module and all genes. To restrict the final scores into the range of (-1, 1), we converted the *p*-values to 1/0/-1 based on the significance threshold using Bonferroni corrections for each dataset (i.e. the thresholds are either $\frac{0.05}{\#\ genes}$ when assessing genes for a given module or $\frac{0.05}{\#\ modules}$ when assessing modules for a given gene). Gene-module associations with *p*-values that survived multiple testing corrections were set to 1 or −1, based on the enrichment direction, and 0 otherwise: $S(p_{G|M}) = \begin{cases} \pm 1, & p_{G|M} < \frac{0.05}{\#\ modules} \\ 0, & otherwise \end{cases}$, where $p_{G|M}$ are one-sided *p*-values, corresponding to either positive or negative associations. The resulting $S(p_{G|M})$ values were then meta-analyzed across the datasets, and the gene-module association scores (GMAS) were computed as the weighted averages of the scores with the weights functions of the sample sizes combined with the inter-gene correlation coefficients within modules. In this way, datasets with more samples and with higher co-expression of genes in modules are given more weight. Denote $D_j,\ j = 1, \dots, J$ available datasets with corresponding sample sizes $n_j,\ j = 1, \dots, J$, and average inter-gene correlations $\bar{\rho}_j\ , j = 1, \dots, J$. Let the *p*-value obtained for the j^{th} dataset is $p_{G|M}(j)$. The final association score is then computed as

$$\text{GMAS} = \frac{\sum_{j=1}^{J} w_j S(p_{G|M}(j))}{\sum_{j=1}^{J} w_j},$$

where weight for the j^{th} dataset is $w_j = \sqrt{n_j}\bar{\rho}_j$. Under the null hypothesis, if we consider the positive and negative associations separately, the random variables $S\left(p_{G|M}(j)\right)$ follow a Bernoulli distribution with probability of success $= \frac{0.05}{\#\, modules}$. Therefore, statistic GMAS is the weighted sum of Bernoulli variables, whose theoretical distribution is hard to establish. The weight is proportional to the square root of the sample size in the j^{th} dataset. Another important component of w_j is the average correlation coefficients among genes in the module in the j^{th} dataset, $\bar{\rho}_j$, which reflects the co-expression or "level of activation" of the module for this dataset.

For the final decision, we computed the true positive rate (percentage of known genes above the threshold against all known genes) and false positive rate (percentage of unknown genes above the threshold against all unknown genes) by varying the threshold of significance. We noticed that decreasing the threshold would increase the true positive rate (TPR) but also the false positive rate (FPR). Therefore, we selected a very stringent threshold for GMAS of 0.268, where only 10% of the known (TPR) and 0.24% of the unknown (FPR) gene-module connections are recovered.

3.3.7 Module-Module Association Determination (M-MAD)

M-MAD takes the association *p*-value matrix between a target module and all genes computed by CAMERA in all datasets (Figure 3.1A bottom-left), and uses the -log$_{10}$(*p*) values as a continuous measure to test whether other biological modules are enriched by having genes that are highly associated with the target module. As CAMERA generates p-values that are uniformly distributed, -log$_{10}$(*p*) transformed values have an exponential distribution skewed towards 0. The following analysis again uses the competitive gene set testing method CAMERA to compute a *p*-value for testing the equality of the average transformed values for the genes in the other biological modules compared to all other genes. It will result in a small *p*-value when many of the genes in the other biological modules are relatively highly connected to the target module. The same analysis is performed for all modules to achieve a final association *p*-value matrix between modules. The Bonferroni correction was used to correct for the multiple testing errors with $\frac{0.05}{\#_{modules}}$ as the significance threshold. To constrain the final score into the range between -1 and 1, module-module connections with enrichment *p*-values that survived multiple testing corrections were allocated 1 or -1, based on the enrichment directions, and 0 otherwise. The results were then meta-analyzed across the datasets, and the module-module association scores (MMAS) were computed as the weighted averages of the connection scores by the sample sizes and inter-gene correlation coefficients within modules across datasets.

3.3.8 Module network analysis

Module networks were constructed using Gephi 0.9.2 [180] based on either the module similarities or module connections from M-MAD. The Fruchterman-Reingold algorithm [181] was used to create the network layout with a gravity value of 10. Iterations were stopped when the network reached stability. The node colors were obtained using the community detection algorithm [182] embedded as the modularity tool in Gephi. Clusters with more than 20 nodes were colored to illustrate the module communities. The most frequent 10 biological terms (excluding biological meaningless words, such as "of", "in", or "and") were used to represent the modules of these communities. The statistical characteristics of the module networks were computed using Gephi. For the network visualization of G-MAD results for one gene, modules were plotted according to their x and y coordinates of the module similarity network, and the gene-module association scores (GMAS) against all modules were used to color the modules using indicated color codes.

3.3.9 Gene correlation network analysis

Gene correlation networks were constructed based on the Pearson correlation among genes of indicated modules in respective datasets using the "*layout_with_fr*" function in the *igraph* R package. Edges with correlation *p*-values lower than the indicated cutoffs in the figure panels were plotted.

3.3.10 Cross validation

In order to test the predictive performance of G-MAD and compare it with the other methods using co-expression (including WeGET, COXPRESdb and average *r*), we performed a cross validation analysis by removing groups of genes from modules, re-computing the associations between the removed genes and the reduced module and testing if we can rediscover the removed genes [159]. We applied leave-one-out cross validation for modules with no more than 50 genes, and 10-fold cross validation for larger modules. The area under the receiver operating characteristic (ROC) curve (AUC) is used to estimate the performance of prediction, with an AUC of 1 indicating perfect prediction and 0.5 indicating random guess. Details of these methods are described below.

WeGET. The WeGET pre-computed results for around 7,000 modules from GO, KEGG, and Reactome were downloaded from https://coexpression.cmbi.umcn.nl/downloads [159].

COXPRESdb. The correlation table for all genes from human datasets was downloaded from https://coxpresdb.jp/download/ (coexpression version: Hsa-r.c4-0, release date: 2019.02.25) [160]. The algorithm described in https://coxpresdb.jp/top_search/#CoExSearch was implemented in R to test the predictive performance of COXPRESdb.

*Average r***.** A simpler method (average *r*) based on average of correlation coefficient was applied on the same expression compendia collected in this study to compare with G-MAD. Specifically, the coexpression between two genes was calculated by taking the average of their correlation coefficients in all datasets. Such calculation was repeated for all gene pairs to obtain the coexpression table across all genes. For a given module, the association with a gene was computed by averaging its correlation coefficients (average *r*) with all genes in this module. The final average *r* was used as the final score to estimate the gene-module association.

3.3.11 Gene set enrichment analysis

Transcriptome data of uterus-specific *Arid1a* knockout mice were downloaded from GEO under the accession number GSE72200 [183]. For enrichment analysis, genes were ranked based on their fold changes between *Arid1a* knockout and control samples, and gene set enrichment analysis (GSEA) was performed to identify the enriched gene sets using the R/fgsea package [169, 184].

3.3.12 Transcript-phenotype correlation analysis in mouse cohorts

Phenotype data, as well as transcriptome data of liver and white adipose tissue, from the BXD [46] and CTB6F2 [67] mouse cohorts were downloaded from GeneNetwork (www.genenetwork.org). Spearman's correlation coefficient *rho* was used to calculate the correlation between the transcript levels of ribosomal protein genes and metabolic phenotypes.

3.3.13 Cell culture and siRNA transfection

Human embryonic kidney (HEK) 293 cells were cultured in DMEM supplemented with 10% fetal bovine serum, 100 IU/ml penicillin and 100 µg/ml streptomycin. HEK 293 cells were grown to approximately 70% confluence in 12-well plate. The cells were treated with either scrambled siRNA, or human *DDT* / *BOLA3* siRNA (Dharmacon) mixed with lipofectamine 2000 to yield a final concentration of 100nM according to the supplier's protocol. After siRNA treatment for 48 hours, cells were collected for quantitative real-time PCR assay. Primers used in this assay are listed below (Table 3.1). Statistical significance was determined by two-tailed Student's t-test.

Table 3.1. Sequences of primers used for RT-PCR

Gene symbol	Direction	Sequence
GAPDH	Forward	TTGGTATCGTGGAAGGACTC
	Reverse	ACAGTCTTCTGGGTGGCAGT
DDT	Forward	CGCCCACTTCTTTGAGTTTC
	Reverse	GGAAGAAGCAGCCAGTTCAC
BOLA3	Forward	GGAGCTCAGAGTGACCCAAA
	Reverse	CAAGGTCTTAAGCAGCAGCA
NDUFA2	Forward	CAAGCTCTGGGCCCGCTACG
	Reverse	CCCAGGCTCTGGGGCTGTTG
NDUFB7	Forward	CCTGCAGATGCCAACCTT
	Reverse	GCGTCCATCATCTCCTGCT
SDHB	Forward	TCTATCGATGGGACCCAGAC
	Reverse	AAGCATCCAATACCATGGGG
SDHD	Forward	GATGGACTATTCCCTGGCTG
	Reverse	AAGGCATCCCCATGAACATA
UQCRC1	Forward	TACCGGGAGCTGGTCAAG
	Reverse	GGTACCCAGTCCAGGATCAG
ATP6	Forward	TAGCCCACTTCTTACCACAAGGCA
	Reverse	TGAGTAGGTGGCCTGCAGTAATGT

3.3.14 Mitochondrial function assay

Mitochondrial oxygen consumption rate (OCR) was measured on a Seahorse XFe96 analyzer (Agilent) according to the manufacturer's protocol. HEK 293 cells were seeded on to 96-well XF analyzer assay plate. Cells were treated with scrambled siRNA or human *DDT* / *BOLA3* siRNA. After 48 hours siRNA treatment, Seahorse XFe96 analyzer was used to measure OCR of the cells. After basal OCR levels were measured, HEK 293 cells were cumulatively treated with 1μM Oligomycin (ATP synthase inhibitor), then 3μM carbonyl cyanide 4-(trifluoromethoxy) phenylhydrazone (FCCP, mitochondrial uncoupler). Then, a mixture of 1μM Antimycin A (mitochondrial respiratory chain Complex III inhibitor) and 1μM Rotenone (Complex I inhibitor) was added. OCR levels were normalized to total protein content per well determined by Lowry protein assay. Statistical significance was determined by two-tailed Student's t-test.

3.3.15 Mitochondrial localization

Mouse embryonic fibroblasts (MEFs) were used to determine the subcellular location of DDT. DDT antibody was purchased from Invitrogen (PA5-62071). MitoTracker® Red CMXRos was purchased from ThermoFisher (M7512). The cells were stained in culture medium containing 100 nM MitoTracker for 30 minutes and then fixed. Cells were then stained with DDT antibody (1:200 dilution) and DAPI, and imaged using the ZEISS LSM 700 microscope. Mitochondrial localization was confirmed by overlaying the signals of DDT and MitoTracker. Cells stained with only MitoTracker or DDT antibody were also included, and no interference signals between the red and green channels were detected.

3.3.16 *C. elegans* experiments

Lipid droplets were stained in *C. elegans* as described previously [7]. Inhibition of ribosome in early stage of worms affects their development and growth, so RNAi was performed after the worms reached

adulthood. Specifically, L1 larvae of N2 worms were grown on regular nematode growth media (NGM) plates at 20°C for 2 days until reaching adulthood. Then worms were then transferred to RNAi plates with 1mM IPTG containing HT115 bacteria expressing RNAi clones for ribosomal genes or empty vector. After 2 days of RNAi treatment, worms were collected, washed twice with 1x PBS and then suspended in 120 µl of PBS. Then 120 µl 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 mixture and warm running tap water, followed by 1 minute spinning at 14,000g. Worms were then washed once using PBS to remove paraformaldehyde. Oil Red O staining of lipid droplets was performed after fixation. 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 worm was used for imaging of the lipid droplets [7]. The lipid droplets were quantified using Fiji (ImageJ) as previously described [7]. Statistical significance was determined by two-tailed Student's t-test.

3.3.17 Data access

Data Availability. All data included in the study is available from https://systems-genetics.org/.

Code Availability. Source code used in this study is available from https://github.com/lihaone/GeneBridge.

3.4 Results

3.4.1 Gene-Module Association Determination (G-MAD)

Owing to the fact that a large number of genes are still not well annotated or even uncharacterized (Figure 1.6, Figure S3.1), we propose here a new computational strategy, "Gene-Module Association Determination" (G-MAD), which uses expression data from large-scale cohorts to propose potential functions of genes. We use the term "modules" to refer the knowledge-based gene sets, ontology terms, and biological pathways from different resources for simplicity in the rest of the paper. The differences between gene sets or directed or undirected pathways are important in many contexts, but for our purpose they can be treated in the same manner as modules and will not be distinguished. The basic concept is similar to classic pathway/gene set analysis, i.e. genes that possess similar functions tend to have similar expression patterns [169]. However, instead of using binary group settings (e.g., control *vs.* treatment, or wild-type *vs.* knockout) as commonly used in gene set analysis, we consider the continuous expression levels of the gene-of-interest across a population and determine its possible functions based on its co-expression patterns against all genes.

In this study, we collected transcriptome datasets with over 80 samples obtained from 6 species (human, mouse, rat, fly, worm and yeast) from GEO, ArrayExpress, dbGaP, GeneNetwork, and other data repository sources. For example, 1,337 datasets containing over 265,000 human samples with whole genome transcript levels were analyzed in this study. The expression datasets were preprocessed using PEER [178] to remove the known and hidden covariates that would influence the analysis (Figure S3.2). We applied a competitive gene set testing method — Correlation Adjusted MEan RAnk gene set test (CAMERA), which adjusts for inter-gene correlations [179] — to compute the enrichment between gene-of-interest and biological modules. Gene-module connections with enrichment *p*-values that survived multiple testing corrections of the gene or module numbers were allocated connection scores of 1 or -1, based on the enrichment direction, and 0 otherwise. The results were then meta-analyzed across datasets, and gene-module association scores (GMAS) were computed as the averages of the connection scores weighted by the sample sizes and inter-gene correlation coefficients within modules ($\bar{\rho}$) (Figure 3.1A).

One should be aware of the fact that modules can overlap partially or completely. For example, GO categories have a hierarchical structure [85], and modules from different sources can be very similar in composition. Therefore, we computed the similarities across all modules, and generated a global module similarity network. As expected, redundant modules formed clusters in the network, and we were able to extract 62 distinct module clusters in the human module similarity network (Figure 3.1B). This network can be used as a way to visualize the results of gene-module associations.

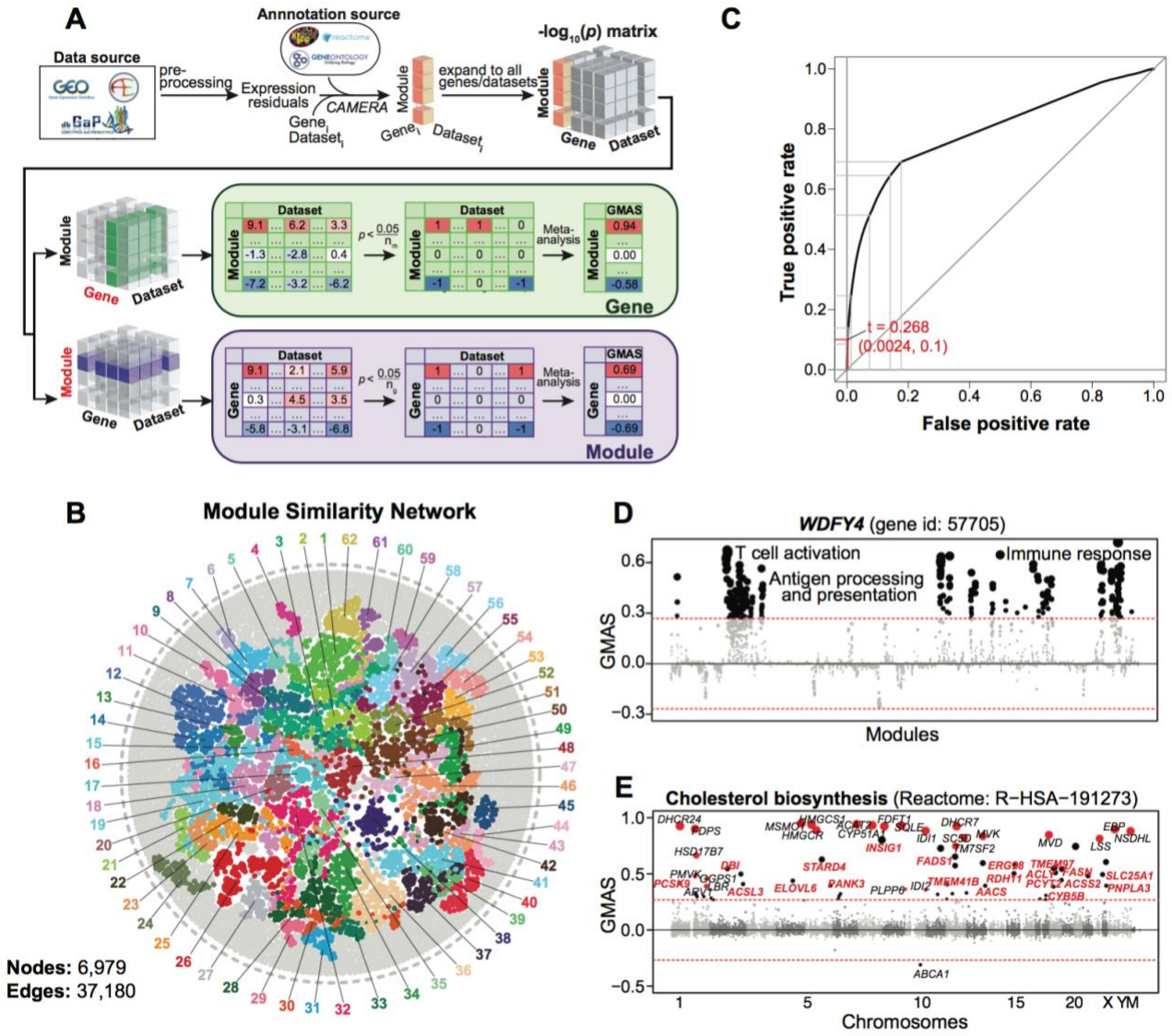


Figure 3.1 Gene-Module Association Determination (G-MAD)

A, G-MAD methodology. See text and Materials and Methods for detailed description.

B, Module similarity network showing the composition similarities across all module pairs. Modules were detected using community detection algorithm embedded in Gephi and indicated in different colors.

C, Influence of the GMAS threshold (t) on the true positive rate (TPR) and false positive rate (FPR) of G-MAD. Using a threshold of 0.268, G-MAD identified 10% of true positives and 0.24% of false positives (reflected by the red lines intersecting x- and y-axis).

D, G-MAD revealed the potential role of *WDFY4* in T cell activation and immune response. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by the module similarities. Known modules connected to *WDFY4* from annotations are shown in red dots (there is no known connected module for *WDFY4*), and other modules with GMAS over the threshold are shown in black dots. Dot sizes reflect the GMAS of *WDFY4* against the respective modules. Detailed information of all the modules are available at www.systems-genetics.org/modules_by_gene/WDFY4?organism=human.

E, G-MAD identified the involvement of known as well as 20 novel genes in cholesterol biosynthesis. The threshold of significant gene-module association is indicated by the red dashed line. Genes are organized by the genetic positions across chromosomes. Genes annotated to be involved in cholesterol biosynthesis are shown in red dots, and novel genes with GMAS over the threshold are shown in black dots. Novel genes conserved in human, mouse and rat are highlighted in red bold text.

We assessed the performance of G-MAD in prioritizing known genes for modules through cross validation. We then compared the area under the receiver operating characteristic (ROC) curve (AUC) with the ones obtained from WeGET [159], and COXPRESdb [160]. G-MAD exhibits better predictive performance than WeGET and COXPRESdb, especially for larger modules (e.g. those with more than 50 genes), as well as a much simpler method based on the average of correlation coefficient between gene pairs using the same expression compendium of our method (average *r*) (Figure S3.3A-B). To estimate if the performance gained from larger dataset numbers in our study, we repeated G-MAD using a subset (800) of the datasets (G-MADsub). We observed that G-MADsub had similar performance as G-MAD and better than COXPRESdb and WeGET, where around 1,000 datasets were used. We investigated the influence of the inter-gene correlations within modules ($\bar{\rho}$) on the predictive performance of G-MAD, and noticed that modules with higher inter-gene correlations and smaller modules tend to have better performances (Figure S3.3C-D).

Furthermore, in order to determine the threshold of significance of gene-module associations, we computed the GMAS of all the gene-module pairs, including both known and unknown pairs. We then created the ROC curve by varying the threshold of significance and calculating the true positive rate (percentage of known genes above the threshold against all known genes) and false positive rate (percentage of unknown genes above the threshold against all unknown genes) (Figure 3.1C). Detecting more true positives by lowering the threshold comes with a cost of higher false positive rate. Therefore, to be stringent in proposing novel gene-module associations and restraint the likelihood of raising false positives, we considered a true positive rate of 0.1 (only 10% of all the known gene-module pairs as significant), and used a GMAS threshold of 0.268 (Figure 3.1C). With this threshold, we saw only 0.24% of unknown gene-module pairs are significant, which is 40 times less than the known pairs.

The gene-module connections predicted by G-MAD provide a resource, which researchers can use as a reference when annotating gene functions. We describe below some examples on how the G-MAD results can be used to facilitate the discovery of novel gene functions or the identification of new members of modules. *WDFY4* was recently annotated as a crucial gene in activating immunological T cells in antiviral and antitumor immunity through a functional CRISPR screen [185]. Through G-MAD, we found that *WDFY4* indeed associated with antigen processing, T cell activation, and immune response in human, mouse, and rat (Figure 3.1D, Figure S3.4A-B), verifying its functions conserved across species. Cholesterol is critical in cell differentiation and growth. We identified 20 genes (*AACS*, *ACLY*, *ACSL3*, *ACSS2*, *CYB5B*, *DBI*, *ELOVL6*, *ERG28*, *FADS1*, *FASN*, *INSIG1*, *PANK3*, *PCSK9*, *PCYT2*, *PNPLA3*, *RDH11*, *SLC25A1*, *STARD4*, *TMEM41B*, *TMEM97)* associated to cholesterol biosynthesis conserved in human, mouse, and rat (Figure 3.1E, Figure S3.4C-E). Several of these genes, including *FASN* [186] and *TMEM97* [187], have already been described to have relevant functions in cholesterol metabolism.

3.4.2 G-MAD identifies tissue-specific associations

Using the expression compendia, we noticed that genes annotated to some modules have higher co-expression in datasets from certain tissues than others (Figure 3.2A), suggesting the tissue-specific activation of these modules. For instance, genes involved in “pancreatic secretion” have much higher co-expressions in datasets obtained from pancreas (Figure 3.2B). To predict the tissue specificity of modules, we compared the inter-gene correlations within each module ($\bar{\rho}$) in every tissue against those in the other tissues using the non-parametric Kolmogorov–Smirnov (K-S) test. The resulting p-values are used as a measure to indicate tissue specificity of the modules (Figure 3.2C). As an example, the “pancreatic secretion” module has higher specificity in the pancreas than in other tissues, for example the blood (Figure 3.2D). Similarly, genes belonging to “collecting duct acid secretion” module are highly co-expressed in kidney (Figure S3.5A-E), while genes in the “lamellar body” module are highly co-expressed in lung (Figure S3.5F-J).

Therefore, G-MAD can also highlight tissue-specific gene-module associations using datasets from specific tissues. *EHHADH* is a peroxisomal protein highly expressed in liver and kidney (Figure 3.3A) [173]. Although best known for its key role in the peroxisomal oxidation pathway, recent report demonstrated that *EHHADH* mutations cause renal Fanconi's syndrome [188]. G-MAD of *EHHADH* in liver and kidney identifies its conserved role in peroxisome and fatty acid oxidation, and also recovers its specific functions in liver (e.g. bile

acid biosynthesis) and kidney (e.g. brush border membrane) (Figure 3.3B-E). *SLC6A1* is one of the major gamma-aminobutyric acid (GABA) transporters in the neurotransmitter release cycle in brain [189]. However, *SLC6A1* is also highly expressed in the liver (Figure S3.6A), and its function in liver remains poorly understood. G-MAD of *SLC6A1* in all datasets and only datasets from brain confirms its function as neurotransmitter transporters in GABA release cycle (Figure S3.6B-C), while G-MAD using datasets from liver identifies its possible role in carboxylic acid transport and metabolism (Figure S3.6D-E).

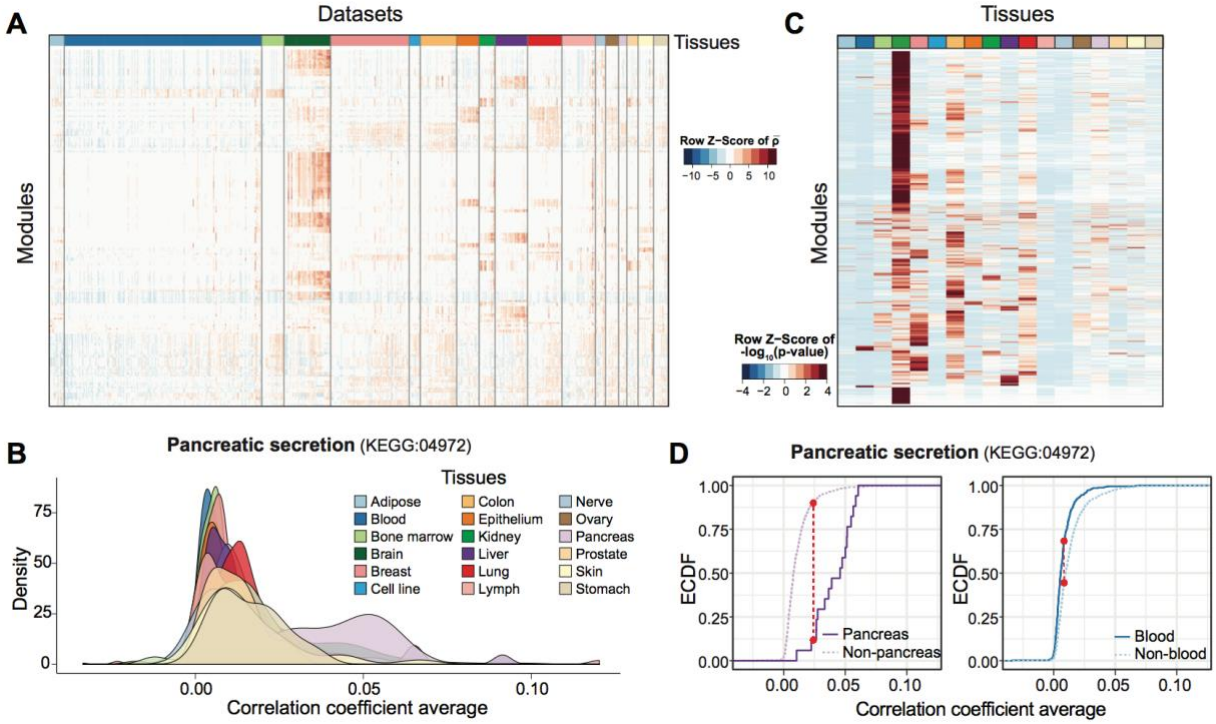


Figure 3.2 Predicting tissue-specificity of modules

A, Heatmap showing the correlation coefficient averages of genes ($\bar{\rho}$) in modules from expression data of a subset of human datasets. Datasets from different tissues are arranged and colored (top bar). Modules are clustered in rows using hierarchical clustering. $\bar{\rho}$ values for each module are centered and scaled.

B, Co-expressions among genes of pancreatic secretion module across tissues in human. The average correlation coefficients across the genes in the pancreatic secretion module in human datasets are used to illustrate the co-expressions of this module across tissues. Genes in the pancreatic secretion module have higher co-expression in datasets from the pancreas compared to those from other tissues.

C, Heatmap showing the tissue-specificity of modules inferred from the correlation coefficient of respective tissues against the other tissues. Modules are clustered in rows using hierarchical clustering. The -$\log_{10}$(p-values) obtained from the K–S test are centered and scaled for each module.

D, The tissue specificity of pancreatic secretion in pancreas (left) and blood (right) is illustrated by the empirical cumulative distribution function (ECDF). The red dotted lines indicate the K-S statistic, which is based on the maximum distance between the two curves. Curves shifting towards the right indicate that datasets from the respective tissue have a higher correlation coefficient, therefore greater specificity for this tissue. In this case, the steeply rising part of the ECDF, also shown as the peak of the density of the correlations in Fig. **3.2B** is shifted towards higher correlations.

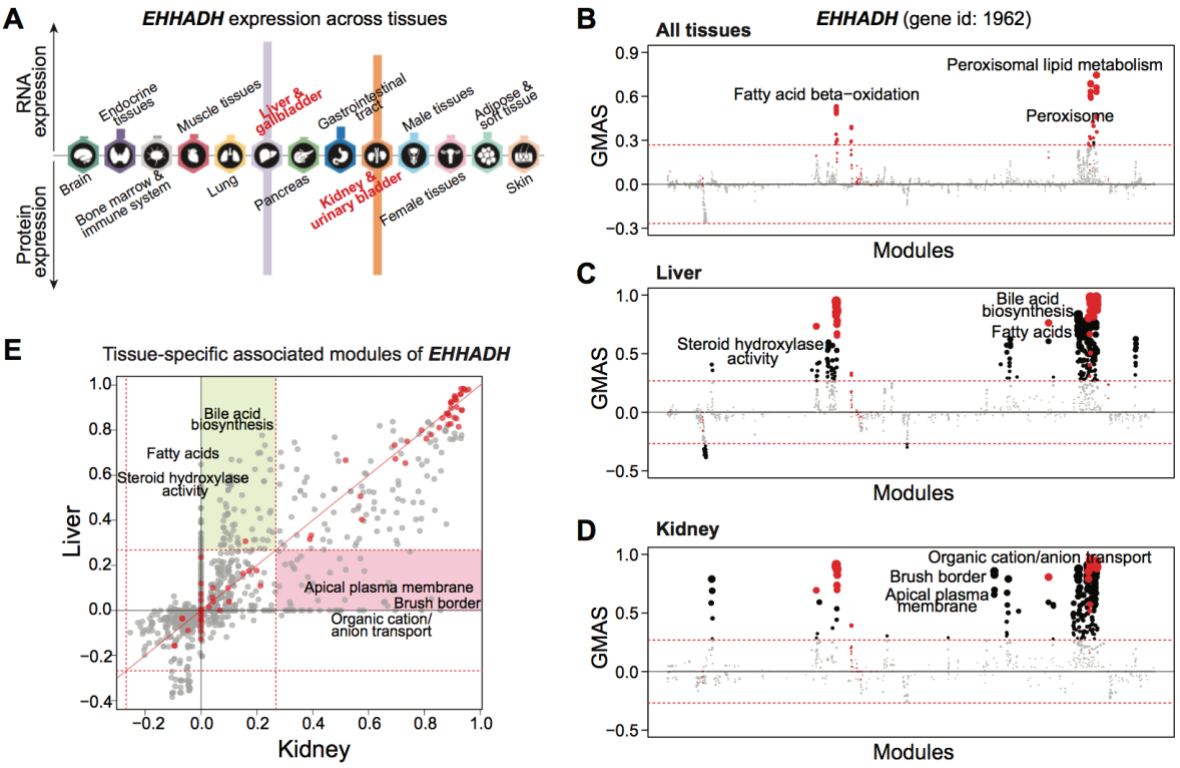


Figure 3.3 G-MAD identifies tissue-specific associated modules for EHHADH by using datasets from different tissues

A, Expression patterns of *EHHADH* across tissues. The figure was adapted from the Human Protein Atlas (www.proteinatlas.org/).

B-**D**, G-MAD of *EHHADH* in human using datasets from all tissues (**B**), from liver (**C**), or from kidney (**D**). The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by their similarities. Known modules connected to *EHHADH* from gene annotations are shown in red dots, and other modules with GMAS over the threshold are shown by black dots.

E, Comparison of G-MAD results of *EHHADH* in liver and kidney. Known modules connected to *EHHADH* are shown in red dots. The threshold of significant gene-module association is indicated by the red dashed line. Modules significantly associated with *EHHADH* only in one specific tissue are highlighted.

3.4.3 G-MAD determines novel genes linked to mitochondria

Mitochondria are the main powerhouses of cells and harvest energy in the form of ATP through mitochondrial respiration. There are around 1,100 genes known to encode mitochondria-localized proteins (mito-proteins), depending on the source used (e.g. 1,158 mito-proteins in Mitocarta [190], 1,074 in Human Protein Atlas [173]); however, many of these genes remain uncharacterized, and the list of mito-proteins is still incomplete [191].

By using the genes annotated to be involved in respiratory electron transport chain (ETC, Reactome: R-HSA-611105), we searched for genes potentially related to respiratory electron transport, by applying G-MAD to expression datasets in human, mouse, and rat. As expected, genes annotated in the ETC module are strongly enriched; moreover, other known ETC genes that were not included in the module were also positively enriched, providing proof that G-MAD can recover known gene functions (Figure 3.4A, Figure S3.7A-B). Based on G-MAD results from human, mouse and rat, there were 707 genes showing conserved associations with the ETC (Figure 3.4B). Many of these genes, for example *DMAC1* (previously known as *C9orf123*) [155, 192, 193], *NDUFAF8* (*C17orf89*) *[194]*, and *FMC1* (*C7orf55*) [161, 195] were not included in the respiratory electron transport module, but have been recently validated to be involved in mitochondrial respiration (Figure 3.4B). *DDT* is among the top genes associated with the ETC (Figure 3.4A-B), and there is no previous study linking it to mitochondria. G-MAD reveals that *DDT* is strongly associated with mitochondrial respiration across different species, including the invertebrate *C. elegans* (Figure 3.4C-D, Figure S3.7C-G), suggesting a conserved role of *DDT* in mitochondria. We validated the mitochondrial localization of *DDT* through immunocytochemistry *in vitro* (Figure 3.4E). The function of *DDT* was confirmed through RNAi-mediated knockdown in HEK293 cells, which led to reduced transcript levels of genes encoding for the ETC subunits

(Figure S3.7H) and decreased oxygen consumption rate (OCR) (Figure 3.4F, Figure S3.7I), verifying that *DDT* impacts mitochondrial respiration. Similarly, we also validated the involvement of *BOLA3* in the ETC using G-MAD and further experimental validations [196] (Figure S3.8).

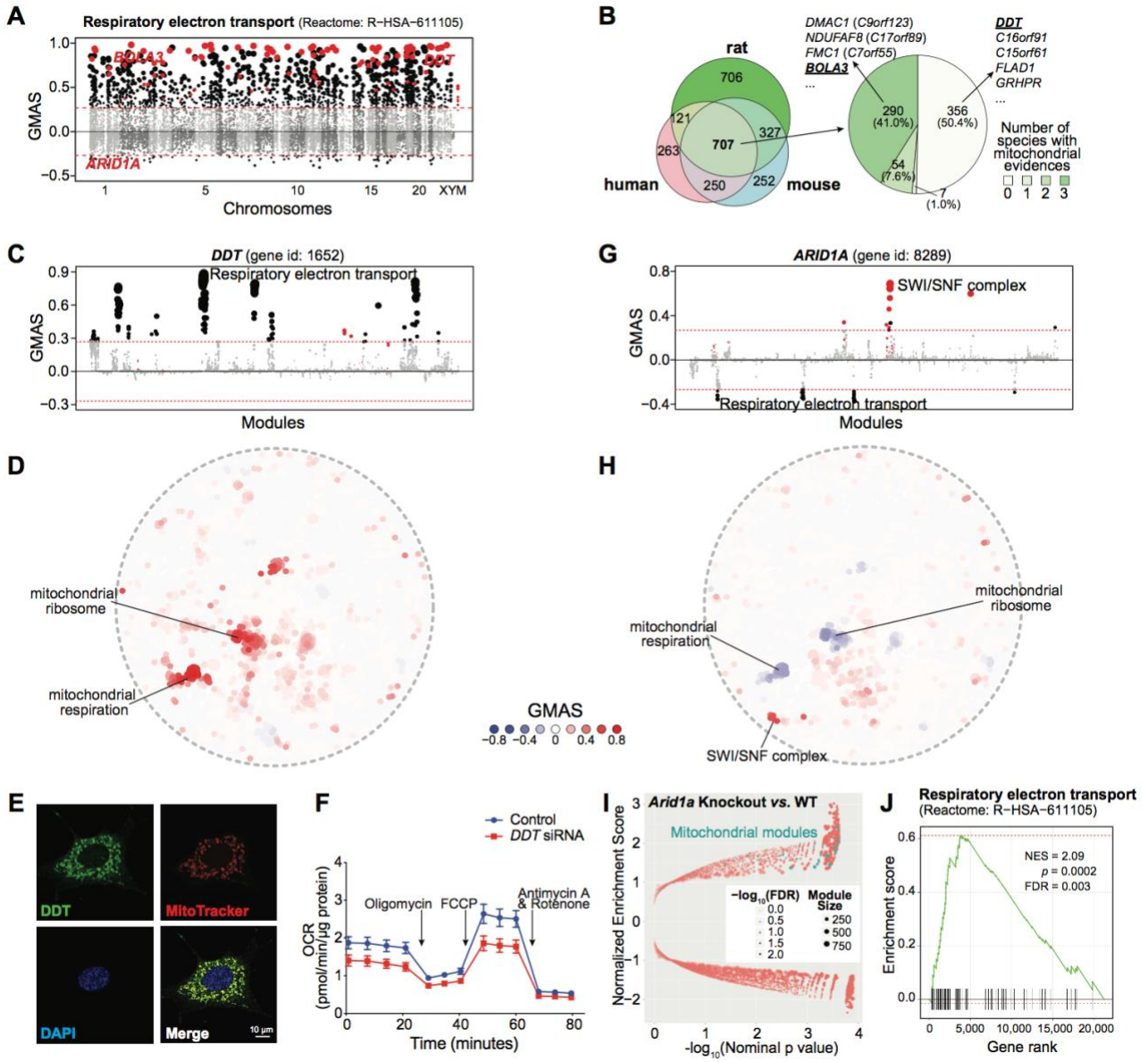


Figure 3.4 G-MAD predicts novel genes linked to mitochondria

A, G-MAD Manhattan plot of the respiratory electron transport (Reactome: R-HSA-611105) module in human. Genes are arranged based on their genetic positions, and genes annotated to be involved in the module are colored red. Genes with absolute GMAS over 0.268 are considered significantly associated. *DDT*, *BOLA3*, and *ARID1A* are labeled.

B, Venn diagram of novel genes associated with respiratory electron transport module in human, mouse and rat. 707 genes were predicted to be mito-proteins by G-MAD in all three species. 351 genes, including *DMAC1* (previously known as *C9orf123*), *NDUFAF8* (*C17orf89*), *FMC1* (*C7orf55*), and *BOLA3,* were recently annotated to be involved in mitochondrial respiration in at least one species, whereas 356 genes, including *DDT*, *C16orf91*, *C15orf61*, *FLAD1*, and *GRHPR*, have not been previously linked with mitochondria based on the current annotations.

C, *DDT* associates with mitochondrial respiratory chain modules in human. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by module similarities. Known modules connected to *DDT* from annotations are highlighted in red, and other modules with GMAS over the threshold are colored in black. Dot sizes reflect the GMAS of *DDT* against the respective modules.

D, Module similarity network showing the modules associated with *DDT*. Modules are plotted based on their layout in Figure **3.1B** and colored based on their GMAS against *DDT*.

E, Mitochondrial localization of *DDT* in mouse embryonic fibroblasts (MEFs). DDT expression is overlapped with the Mitotracker red label.

F, *DDT* knockdown leads to the reduction of oxygen consumption rate (OCR) as a reflection of mitochondrial respiration in human HEK293 cells. Addition of specific mitochondrial inhibitors, including the oligomycin (ATPase inhibitor), FCCP (uncoupling agent), and rotenone/antimycin A (electron transport chain inhibitors) are indicated by arrows.

G, *ARID1A* negatively associates with mitochondrial respiratory chain in human. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by the module similarities. Known modules connected to *ARID1A* from extant annotations are highlighted in red, and other modules with GMAS over the threshold are colored in black. Dot sizes are proportional to GMAS of the respective modules.

H, Module similarity network showing the modules associated with *ARID1A*. Modules are colored based on their GMAS against *ARID1A*.

I, Mice with the uterine-specific *Arid1a* knockout showed positive enrichment in mitochondrial respiration modules. Nominal *p*-values from the GSEA results are used to plot against normalized enrichment score (NES), with dot sizes indicating the number of genes in the modules and transparencies indicating the false discovery rate (FDR).

J, Enrichment plot showing the enrichment of genes included in respiratory electron transport in uterus-specific *Arid1a* knockout mice compared to wild-type controls. Genes are ranked based on the fold change between *Arid1a* knockout and wild-type mice, and the ranking positions of genes in respiratory electron transport are labeled as vertical black bars. NES, normalized enrichment score. FDR, false discovery rate.

Contrary to most of the existing sources that predict only positive gene-module associations, G-MAD is also able to exploit negative associations. For example, *ARID1A* exhibits significant negative associations with the respiratory electron transport in human and mouse (Figure 3.4A, G-H, Figure S3.9). *ARID1A* is a known member of the SWI/SNF family, and the inactivating mutations of SWI/SNF complex genes (mainly *SMARCA4* and *ARID1A*) have recently been linked to increased expression of ETC genes and mitochondrial respiration [197]. To further validate its regulatory role, we checked an extant public dataset from mice with uterus-specific *Arid1a* knock-out [183], and confirmed that dysfunction of *Arid1a* led to the increased expression of mitochondrial genes (Figure 3.4I), especially those involved in respiratory electron transport (Figure 3.4J).

3.4.4 Module-Module Association Determination (M-MAD)

Biological processes and modules, such as metabolism, cellular signaling, biogenesis, and degradation are interconnected and coordinated [198]. However, there are few reports exploring the connections between modules in a systematic fashion [199]. Here we extend G-MAD to develop Module-Module Association Determination (M-MAD) to investigate the connections between modules based on the expression compendia. Results for individual modules against all genes, obtained from G-MAD, were used to compute their associations against all modules. The enrichment scores of all genes for the target module were used as the gene-level statistics to calculate the enrichment against all modules using CAMERA [179]. The resulting enrichment *p*-values across modules were transformed to 1, 0, or –1 based on the Bonferroni threshold, and then meta-analyzed across all datasets to obtain the module-module association scores (MMAS) (Figure 3.5A).

Module-module associations with an absolute MMAS of over 0.268, corresponding to 4% of the total number of module pairs, were considered significant and were used to construct a module association network (Figure 3.5B). Modules were represented as nodes with the same colors as the module clusters from Figure 3.1B. While the module *similarity* network in Figure 3.1B is based solely on existing gene annotations, the module *association* network relies on analyzing the full expression datasets. It can thus reveal new biological connections among modules, which were not included in literature-based annotations. We compared the two networks (Figure S3.10) obtained from module similarity (Figure 3.1B) and module association (Figure 3.5B). Interestingly, there are numerous module pairs with no similarity (overlap of annotated genes), but with high association based on expression (M-MAD) (Figure 3.5C). Moreover, many module pairs have predicted negative associations (Figure 3.5C). Therefore, these results provide a resource for hypothesis generation and validation of the module connections.

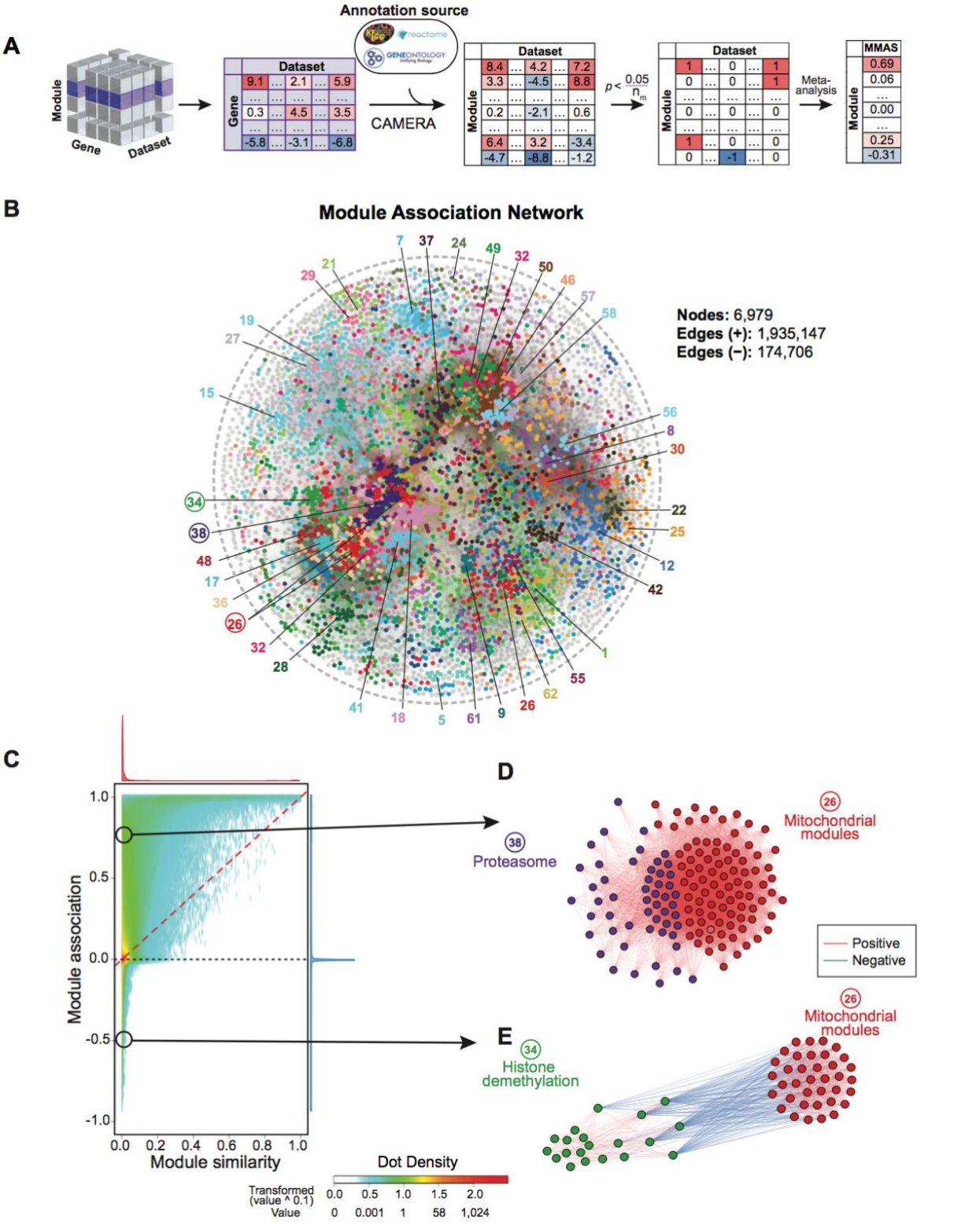


Figure 3.5 Module-Module Association Determination (M-MAD) reveals module connections

A, Scheme of the M-MAD methodology in detecting module connections. Intermediate results of G-MAD for all modules are further processed and used as the basis of M-MAD. The $-\log_{10}(p)$ values of G-MAD for the target module against all genes in each dataset are used as the gene statistic for the module, and connections between the target module and all modules are calculated using CAMERA. The results are then meta-analyzed by taking the sample sizes and inter-gene correlations of all datasets to compute the module-module association score (MMAS) between modules.

B, Module association network showing the connections across all modules. Colors of nodes represent the modules defined in the global module similarity network in Figure **3.1B**. Module clusters with respective colors are identified and labeled. Modules used as examples in the following figure panels are highlighted in circle.

C, Comparison of pairwise module connections derived from module similarities in Figure **3.1B** and associations (from M-MAD) in Figure **3.5B**. A red dashed line is plotted when the pairwise module similarity equals association. The distributions of module similarity and association scores are illustrated in the top and at the right of the plot and are colored in red and blue, respectively. Two examples of novel module connections are encircled.

D-**E**, Subnetworks showing the association between mitochondrial and proteasomal modules (**D**), and mitochondrial and histone demethylation modules (**E**). Edges colors indicate the significance of module connections, with red as positive and blue as negative.

By applying M-MAD, we observed a strong positive link between mitochondrial modules and the proteasome (Figure 3.5D, Figure S3.11A-C). Most of the genes encoding for proteasomal subunits exhibit remarkable association with the ETC in human and mouse (Figure S3.11G), indicating a conserved co-regulatory mechanism. Dysfunction of mitochondria and the ubiquitin-proteasome system (UPS) are hallmarks of aging and aging-related neurodegenerative diseases, such as Alzheimer's, Parkinson's, and Huntington's diseases [200-202]. Abnormalities that perturb the crosstalk between these two modules have been demonstrated to contribute to the pathogenesis of these diseases and several mechanisms have been proposed [202, 203]. It has also been shown that ETC disruption leads to proteasome impairment [202], while conversely the inhibition of the UPS causes mitochondrial dysfunction [201].

Similar to G-MAD, M-MAD can also predict negative connections between modules. For example, we found strong negative connections between histone demethylation processes and mitochondrial modules (Figure 3.5E, Figure S3.11D-F). The link between epigenetics and mitochondria is a research focus for many groups, including ours [204-206]. It has been reported that mitochondrial dysfunction affects histone methylation, and conversely histone lysine demethylases can impact mitochondrial functions [206]. Most of the histone lysine demethylases showed negative associations with the ETC in human and mouse (Figure S3.11G), suggesting a conserved negative connection between histone demethylation and mitochondrial function.

As another example of M-MAD, we investigated modules connected with lipid biosynthetic modules. Interestingly, ribosome modules exhibited strong negative association with lipid biosynthetic modules (Figure 3.6A-B, Figure S3.12A-B). This is in line with our previous finding that a ribosomal protein, *Rpl26*, negatively correlates with body weight and fat mass [7]. In support of this connection, liver and adipose transcripts of most ribosomal protein genes negatively correlated with metabolic phenotypes, such as body weight, fat mass, and cholesterol levels in the BXD mouse cohort [46] (Figure 3.6C, Figure S3.12C), as well as in a CAST/EiJ and C57BL/6J F2 intercross [67] (Figure 3.6D, Figure S3.12D). Finally, RNAi targeting 9 of the identified ribosomal protein genes out of total 13 tested led to the accumulation of lipid droplets in *C. elegans* (Figure 3.6E, Figure S3.12E-G), further validating the robustness of the lipid synthesis-ribosome connection across species.

3.5 Discussion

Significant efforts in biological research have been devoted to defining the molecular and physiological functions of genes. However, many genes are still not well annotated and even remain uncharacterized [87-89]. Here we developed an approach, termed G-MAD, to facilitate the identification of novel gene functions and to establish robust connections between genes and modules. Using transcriptome datasets from cohorts ranging from human to mouse, rat, fly, worm, and yeast, we identified millions of gene-module connections, many of which are novel. Unlike most available sources relying on co-expression to predict gene functions, G-MAD can identify not only positive gene-module connections, but also negative associations between genes and modules or processes. We illustrated the predictive power of G-MAD in revealing potential gene-module connections using the mitochondrial electron transport chain (ETC) module as an example. 707 genes were consistently associated with the ETC in human, mouse and rat, of which *DDT* and *BOLA3* were validated through experiments. A negative connection between *ARID1A*, a member of the SWI/SNF family, and the ETC was also identified using G-MAD, which was consistent with a recent report that inactivation of SWI/SNF complex increased mitochondrial respiration [197]. Meanwhile, tissue-specific functions of genes, for example *EHHADH* and *SLC6A1*, can also be identified using datasets derived from respective tissues.

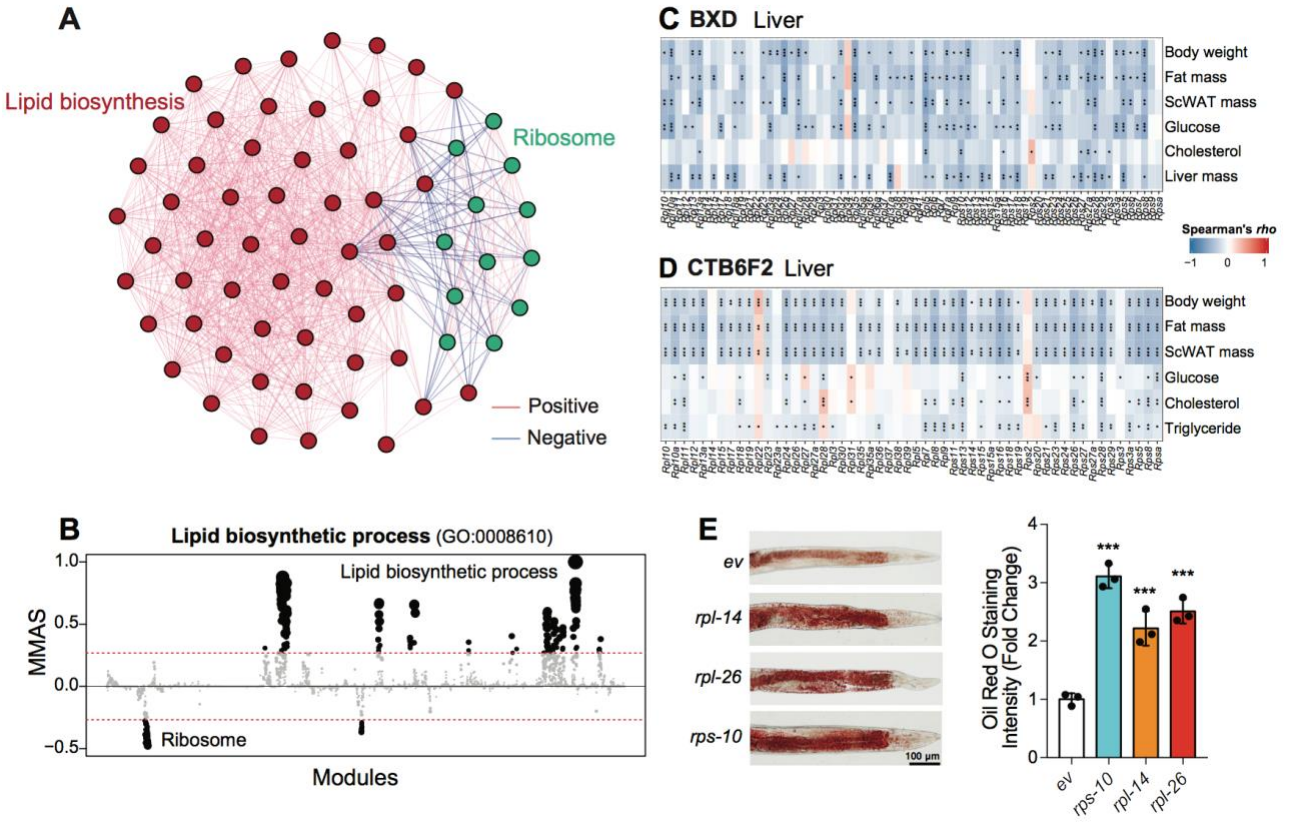


Figure 3.6 M-MAD reveals a negative association between the ribosome and lipid biosynthetic modules

A, Subnetwork for the ribosome and lipid biosynthetic modules. The colors of the edges indicate the significance of module connections, with red as positive and blue as negative.

B, Lipid biosynthetic process negatively connected with ribosomal modules in human. The threshold of significant module-module connection is indicated by the red dashed line. Modules are organized by the module similarities. Dot sizes are proportional to MMAS of the respective modules.

C-**D**, Transcripts of genes encoding for ribosomal proteins in the liver negatively correlate with metabolic traits, such as body weight, fat mass, plasma glucose and cholesterol levels, in the BXD (**C**) and CTB6F2 (**D**) mouse cohorts. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

E, Feeding adult *C. elegans* with RNAi clones of ribosomal proteins, including *rps-10*, *rpl-14*, and *rpl-26*, results in the accumulation of lipids, as reflected by Oil Red O staining. Experimental scheme and additional examples are shown in Figure **S3.12**. ***, $p < 0.001$. *ev*, empty vector. n=3.

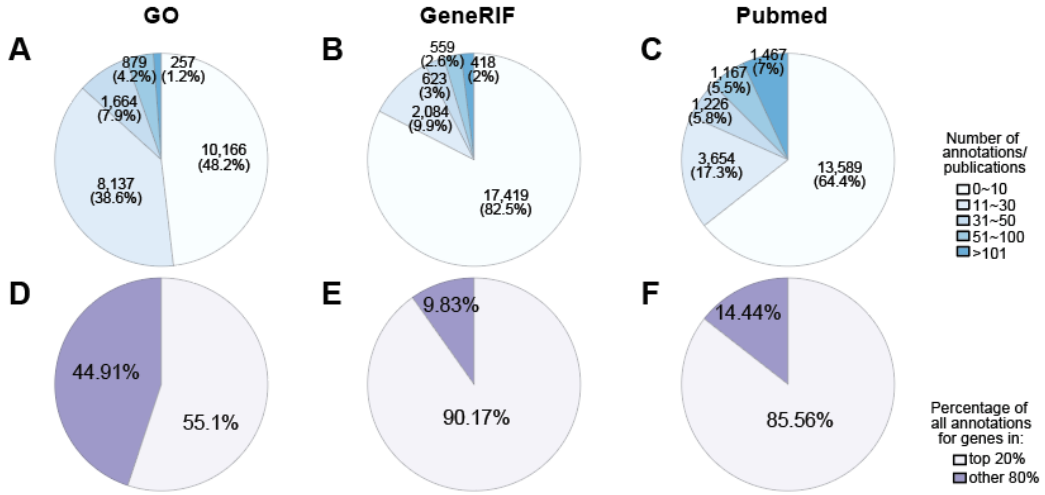
In addition, we extended G-MAD to M-MAD, to uncover connections between modules. Association scores of one module against all genes from G-MAD were used to compute its associations with all modules. Similar to G-MAD, M-MAD can identify both positive and negative module associations. For example, in humans we identified around 2,000,000 associations between all modules, over 170,000 of which were negative. We constructed a module association network based on these connected modules, and compared it to the module similarity network. Interestingly, many of the associated module pairs have low or no similarities in gene compositions. By applying M-MAD on the ETC module, we discovered a conserved connection between mitochondria and the proteasome in various organisms [202]. In addition, we identified negative associations between histone lysine demethylation and mitochondrial modules, underscoring the inverse connection between epigenetic regulation and mitochondrial function [204-206]. Moreover, we discovered and validated a novel negative regulatory role of ribosomal proteins on lipid biosynthesis [7].

In summary, we described here a set of approaches to identify gene function and module connectivity, that we collectively termed GeneBridge, to reflect their capacity to bridge genes to biological functions and phenotypes. The GeneBridge toolset is accessible through our open web resource (systems-genetics.org) to the research community for hypothesis generation or validation. It should be noted that we selected a stringent threshold of 0.268 to limit the probability of detecting false positives. Researchers, however, have the possibility to fully explore the results by altering the thresholds on the open web resource. Although only protein-coding genes were included in our analysis, the same approach can be applied to non-coding genes

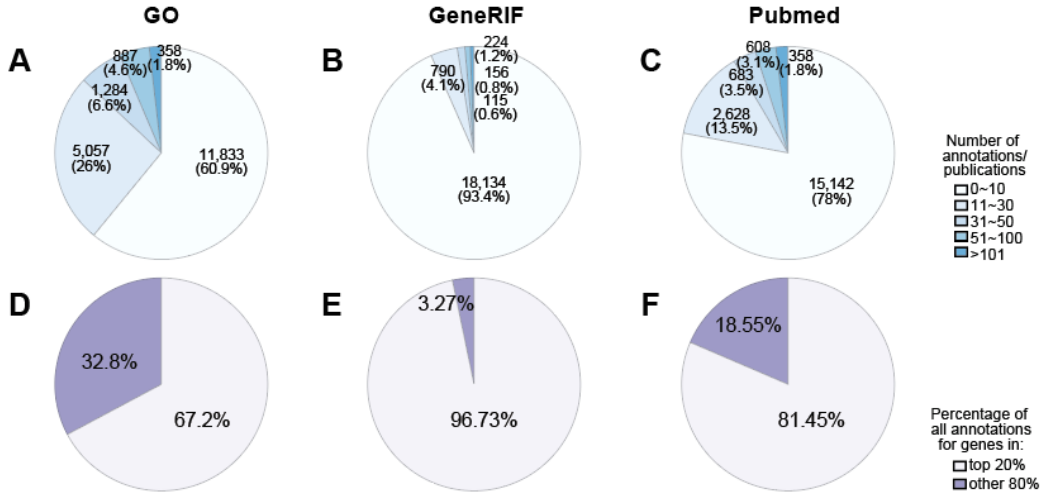
to reveal their potential functions. Similarly, GeneBridge can also be utilized to identify novel gene-disease associations based on known disease-associated genes from databases, such as the Human Disease Ontology (DO) [207] or DisGeNET [208]. The GeneBridge toolkit could also be applied to large-scale proteomics datasets after correcting for the background of all measured proteins. Integration of GeneBridge with other well-established databases, such as BioGRID [209] and STRING [210], will facilitate the investigation of the connections between genes, modules, and diseases.

3.6 Supplemental figures

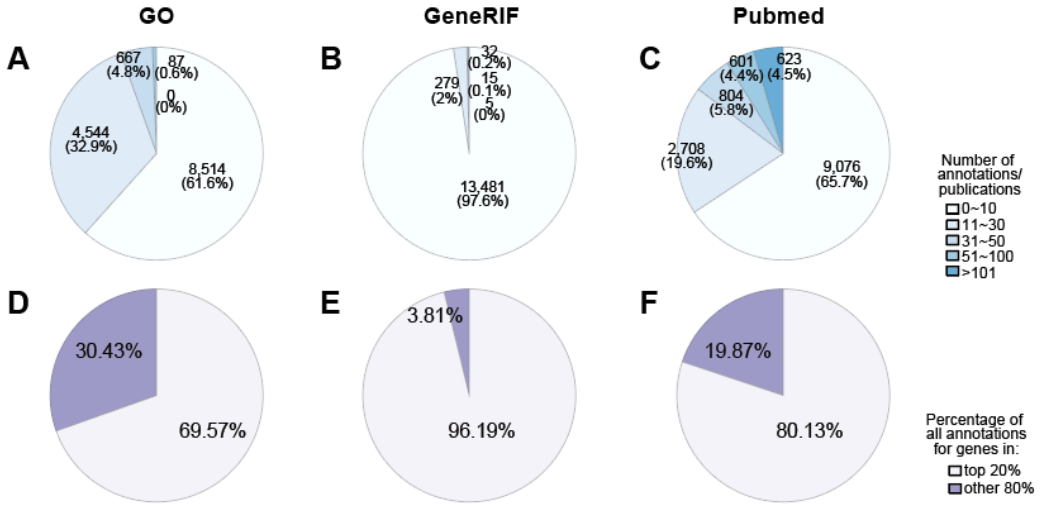
S1-1. *Mus musculus*



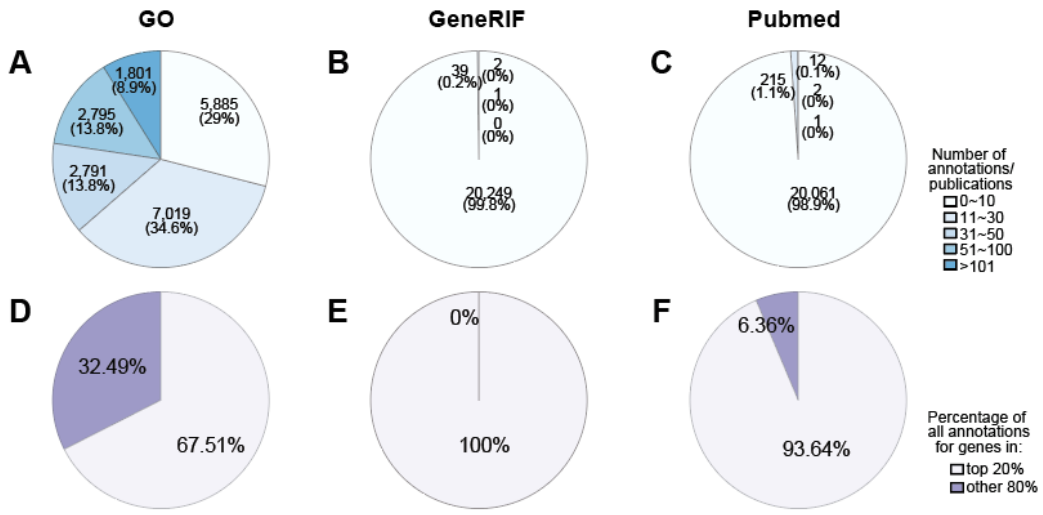
S1-2. *Rattus norvegicus*



S1-3. *Drosophila melanogaster*



S1-4. *Caenorhabditis elegans*



S1-5. *Saccharomyces cerevisiae*

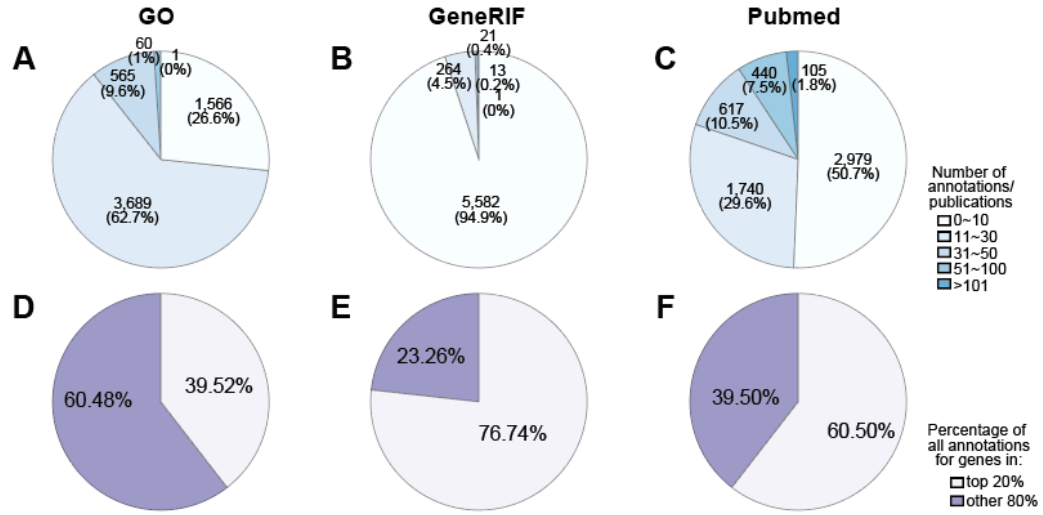


Figure S3.1. Statistical summary of available annotations for genes in *M. musculus* (S1-1), *R. norvegicus* (S1-2), *D. melanogaster* (S1-3), *C. elegans* (S1-4), and *S. cerevisiae* (S1-5).

The number of annotations per gene in GO (**A**), GeneRIF (**B**), and number of publications in PubMed (**C**). The percentage of all annotations/publications covering the top 20% most annotated genes in respective species (**D**, **E**, **F**).

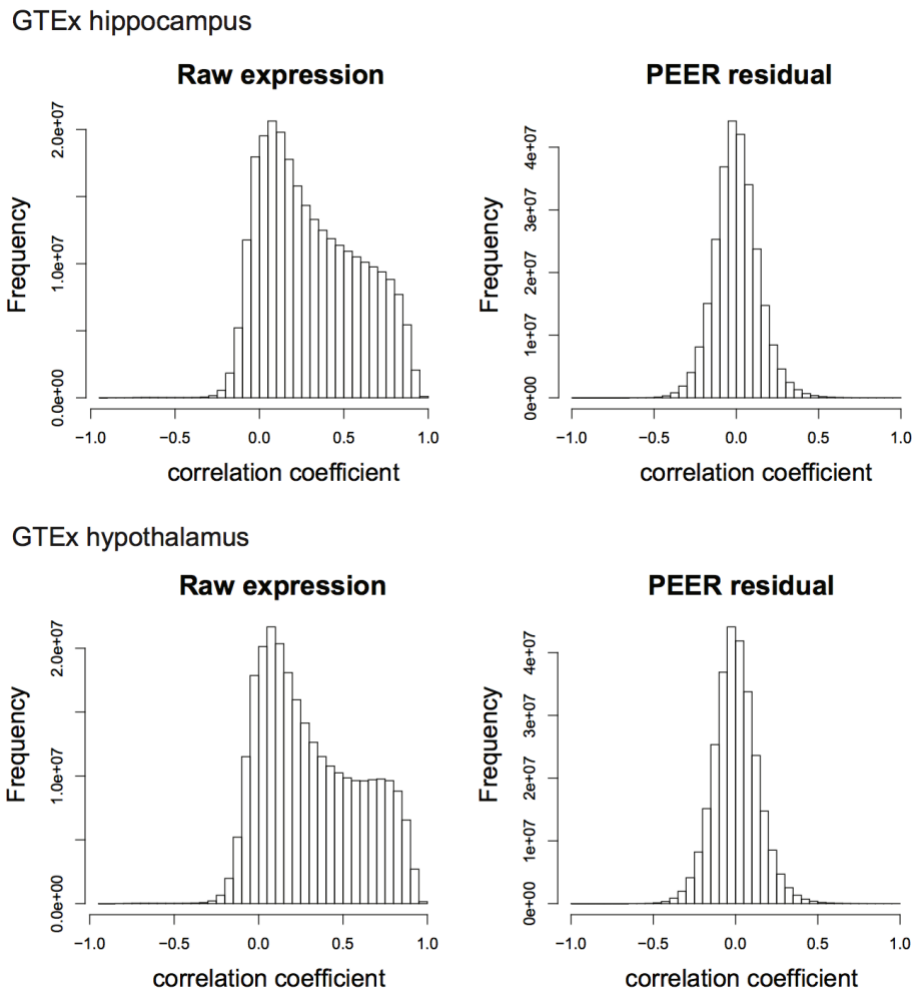


Figure S3.2. Covariates in expression datasets affects the co-expressions between genes.

The hippocampus (**upper**) and hypothalamus (**lower**) datasets from GTEx were used as examples to illustrate the influence of covariates in the co-expressions between genes. The distributions of correlation coefficient of all gene pairs from either the raw expression data (**left**) or expression residual from PEER (**right**) were compared.

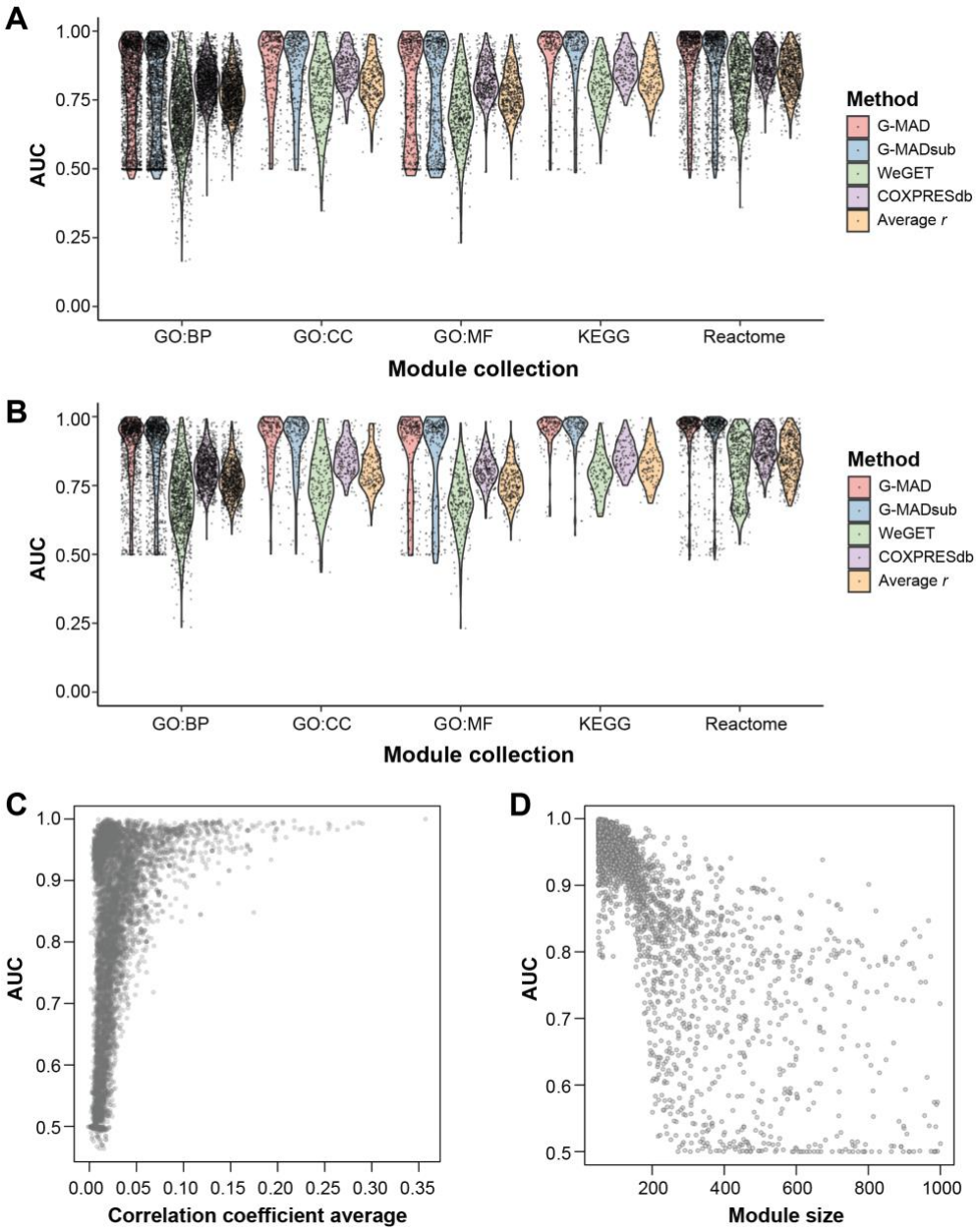


Figure S3.3. Comparison of the predictive performance of G-MAD with available methods.

A, **B**, The performances of G-MAD, as well as other existing methods in all modules (**A**) or modules with more than 50 genes (**B**) are summarized based on the collection source of the modules. The predictive performance of G-MAD is compared to WeGET and COXPRESdb, as well as a simpler method based on average of correlation coefficient (average *r*) using the same expression compendium of G-MAD, using cross-validation. In addition, we repeated G-MAD with a subset (G-MADsub, using 800 datasets) of the datasets to test if the number of datasets brought the higher performance than WeGET, which has around 1,000 datasets. Cross validation evaluates the robustness of the methods by removing the genes from the query module and test the performance in redetecting them. Performance of the method is computed as the area under the receiver operating characteristic curve (AUC) for each module. A high AUC indicates that most of the genes in the module are rediscovered when they are removed from the module in the analysis.

C, **D**, The intergene correlations (correlation coefficient average) (**C**) and module size (**D**) have strong influence on the predictive performance of modules.

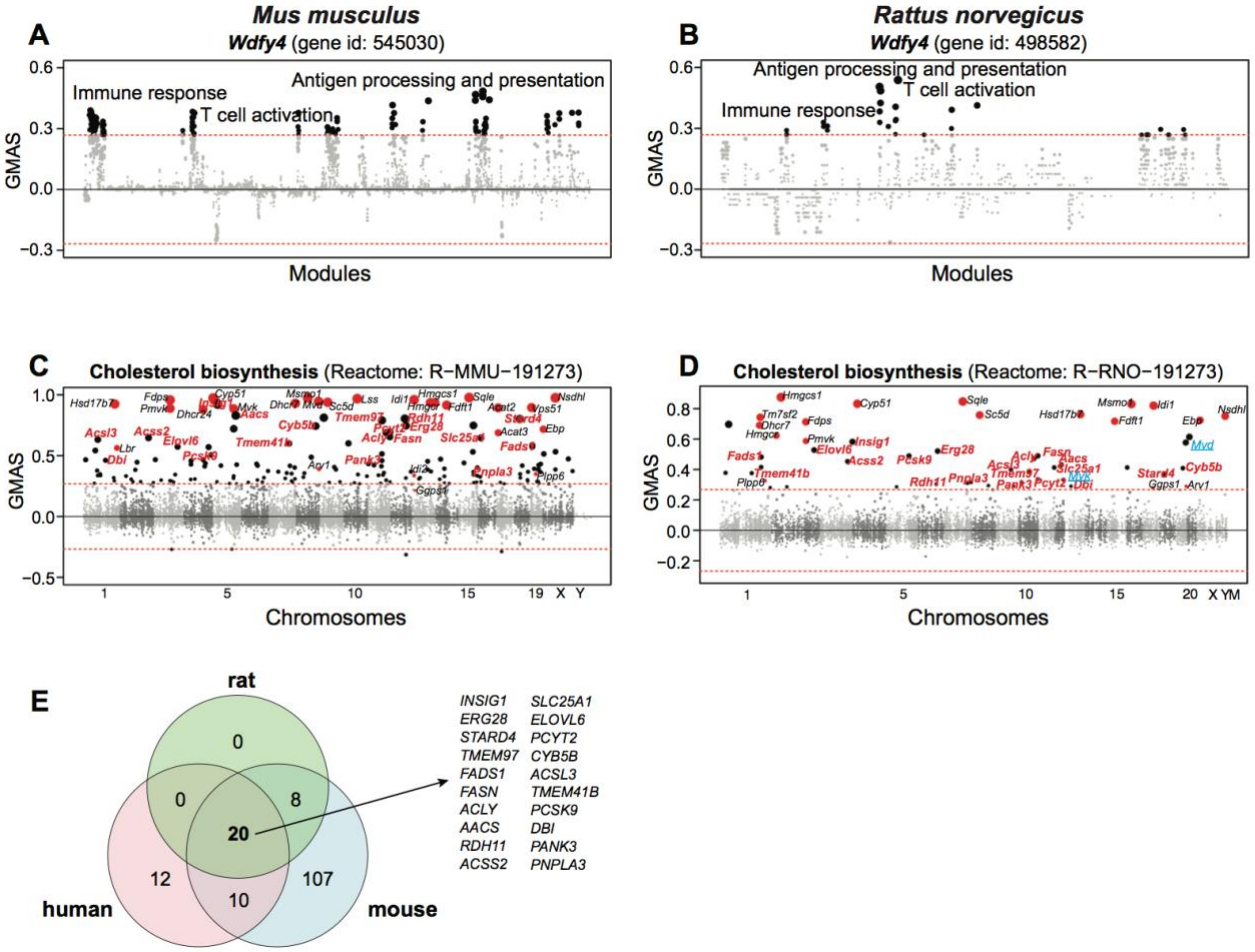


Figure S3.4. G-MAD in mouse and rat confirms the gene-module connections between *WDFY4* and T cell activation, and links 20 new genes with cholesterol biosynthesis.

A, B, G-MAD of *Wdfy4* in mouse (**A**) and rat (**B**) confirms its involvement in T cell activation and immune response. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by the module similarities. Known modules connected to *Wdfy4* from annotations are shown in red dots (no connected modules for *Wdfy4*), and other modules with GMAS over the threshold are shown with black dots.

C, D, G-MAD confirms the involvement of novel genes in cholesterol biosynthesis in mouse (**C**) and rat (**D**). The threshold of significant gene-module association is indicated by the red dashed line. Genes are arranged based on their genetic positions. Genes annotated to be involved in cholesterol biosynthesis are shown in red dots, and genes with GMAS over 0.268 are shown in black dots. Novel genes conserved in human, mouse and rat are highlighted in red bold text. *Mvd* and *Mvk* (highlighted in blue text in **D**) are included in the annotation of cholesterol biosynthesis module in human and mouse, but not in rat.

E, Venn diagram comparing G-MAD results of cholesterol biosynthesis in human, mouse, and rat. 20 novel genes were identified with conserved associations with cholesterol biosynthesis in all 3 species.

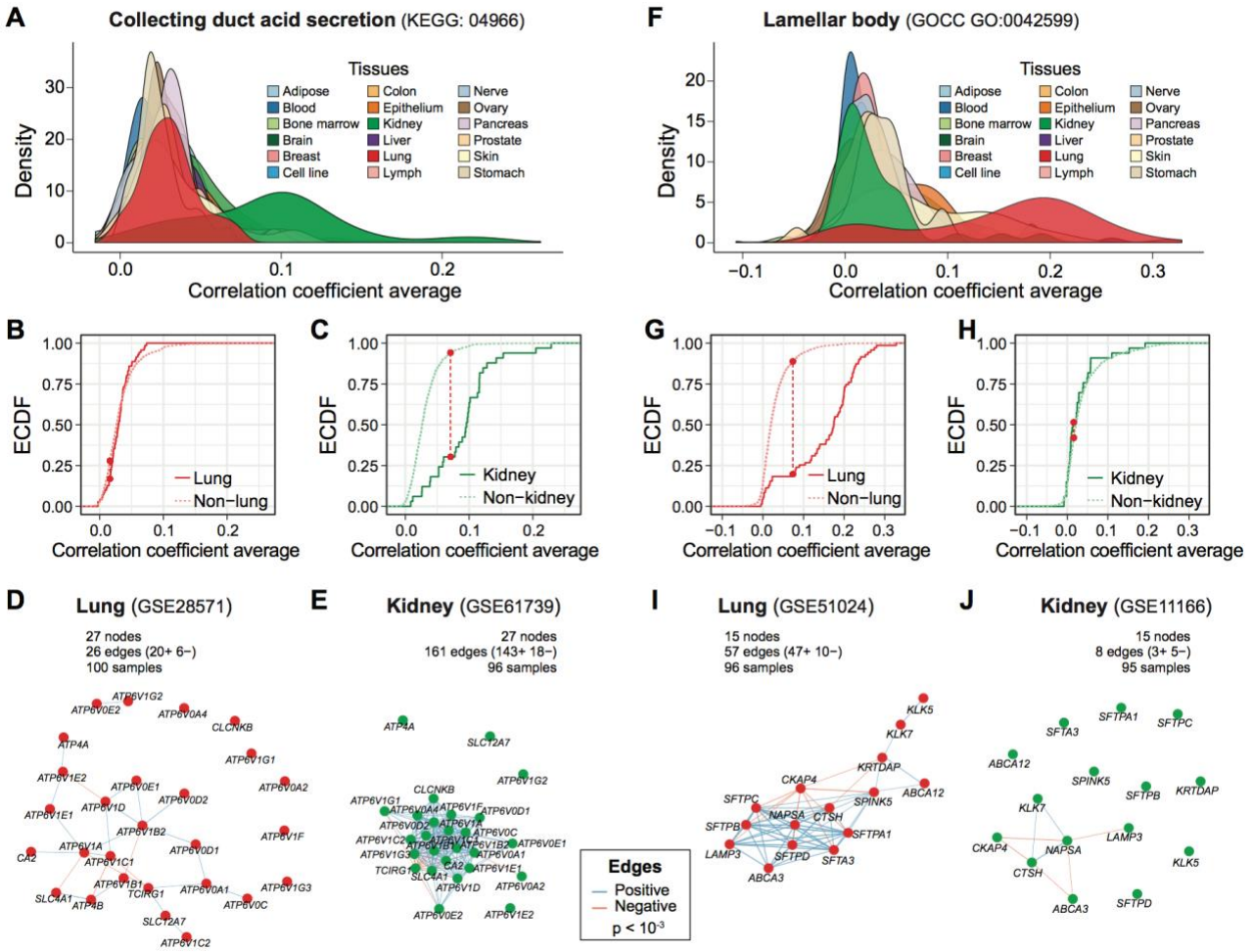


Figure S3.5. Tissue specific co-expression of modules.

A, **F**, Distribution of the co-expression of genes in the “collecting duct acid secretion” (**A**) or “lamellar body” (**F**) module across tissues in human. The average correlation coefficient of the gene pairs of the module in 1,300 human expression datasets from 18 major tissues was used to illustrate the co-expressions of this module across tissues. Genes in “collecting duct acid secretion” (**A**) and “lamellar body” (**F**) module have higher co-expression in datasets from kidney and lung, respectively, indicating the potential to assign tissue-specificity.

B-**C**, **G**-**H**, The tissue specificity of “collecting duct acid secretion” (**B**-**C**) or “lamellar body” (**G**-**H**) module in lung (**B**, **G**) and kidney (**C**, **H**) is illustrated by the empirical cumulative distribution function (ECDF). The red dotted lines indicate the Kolmogorov–Smirnov statistic, which is based on the maximum distance between the two curves. Curves shifting towards the right indicate that datasets from the respective tissue have a higher correlation coefficient, therefore greater specificity for this tissue. In this case, the steeply rising part of the ECDF, also shown as the peak of the density of the correlations in Fig. **S3.5A, F** is shifted towards higher correlations.

D-**E**, **I**-**J**, Pearson correlation network of genes in the “collecting duct acid secretion” (**D**, **E**) or “lamellar body” (**I**, **J**) module in representative datasets of lung (**D**, **I**) and kidney (**E**, **J**). The number of genes (nodes) and gene pairs (edges) that survive the indicated threshold of correlation significance are shown. Genes in the “collecting duct acid secretion” module have higher co-expression in datasets from the kidney (**E**) than the lung (**D**), while genes in the “lamellar body” module have higher co-expression in datasets from the lung (**I**) than the kidney (**J**).

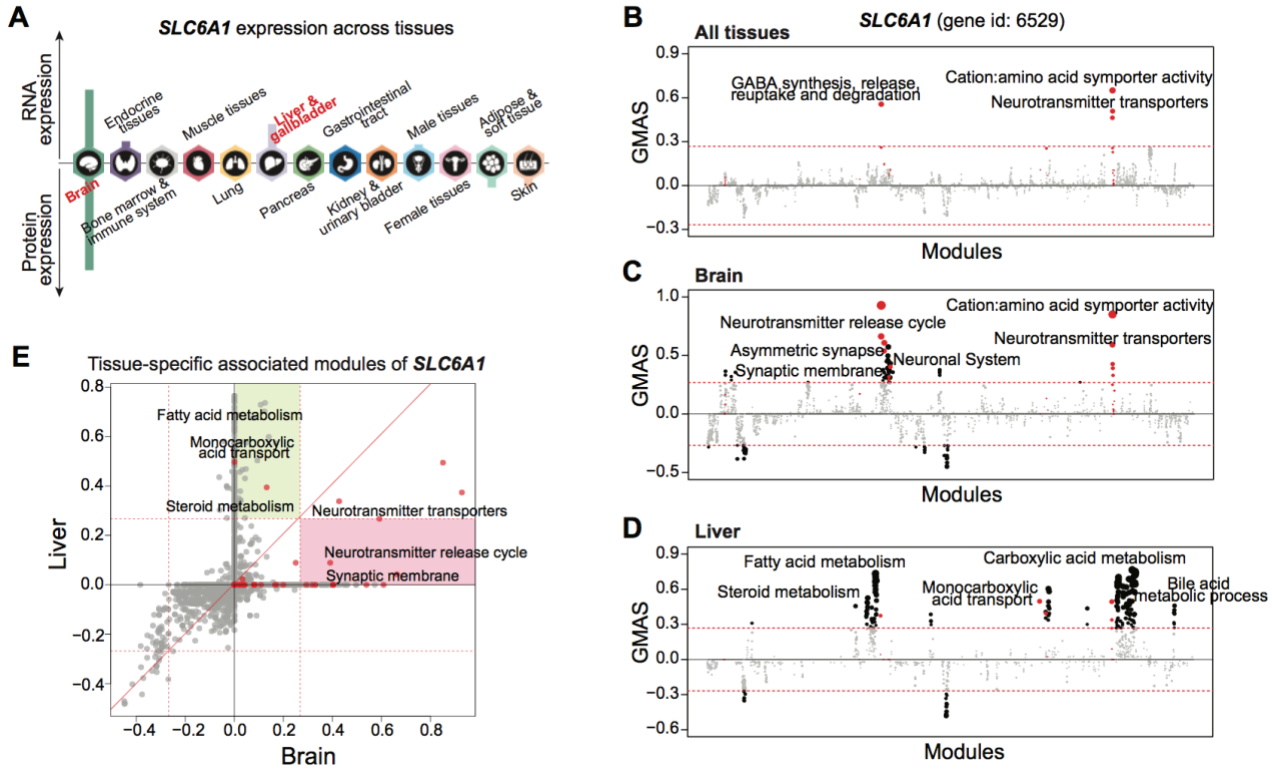


Figure S3.6. G-MAD identifies tissue-specific associated modules for *SLC6A1* by using datasets from different tissues.

A, Expression patterns of *SLC6A1* across tissues. The figure was adapted from the Human Protein Atlas.

B-**D**, G-MAD of *SLC6A1* in human using datasets from all tissues (**B**), from brain (**C**), or from liver (**D**). The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by their similarities. Known modules connected to *SLC6A1* from gene annotations are shown in red dots, and other modules with GMAS over the threshold are shown by black dots.

E, Comparison of G-MAD results of *SLC6A1* in brain and liver. Known modules connected to *SLC6A1* are shown in red dots. The threshold of significant gene-module association is indicated by the red dashed line. Modules significantly associated with *SLC6A1* only in one specific tissue are highlighted.

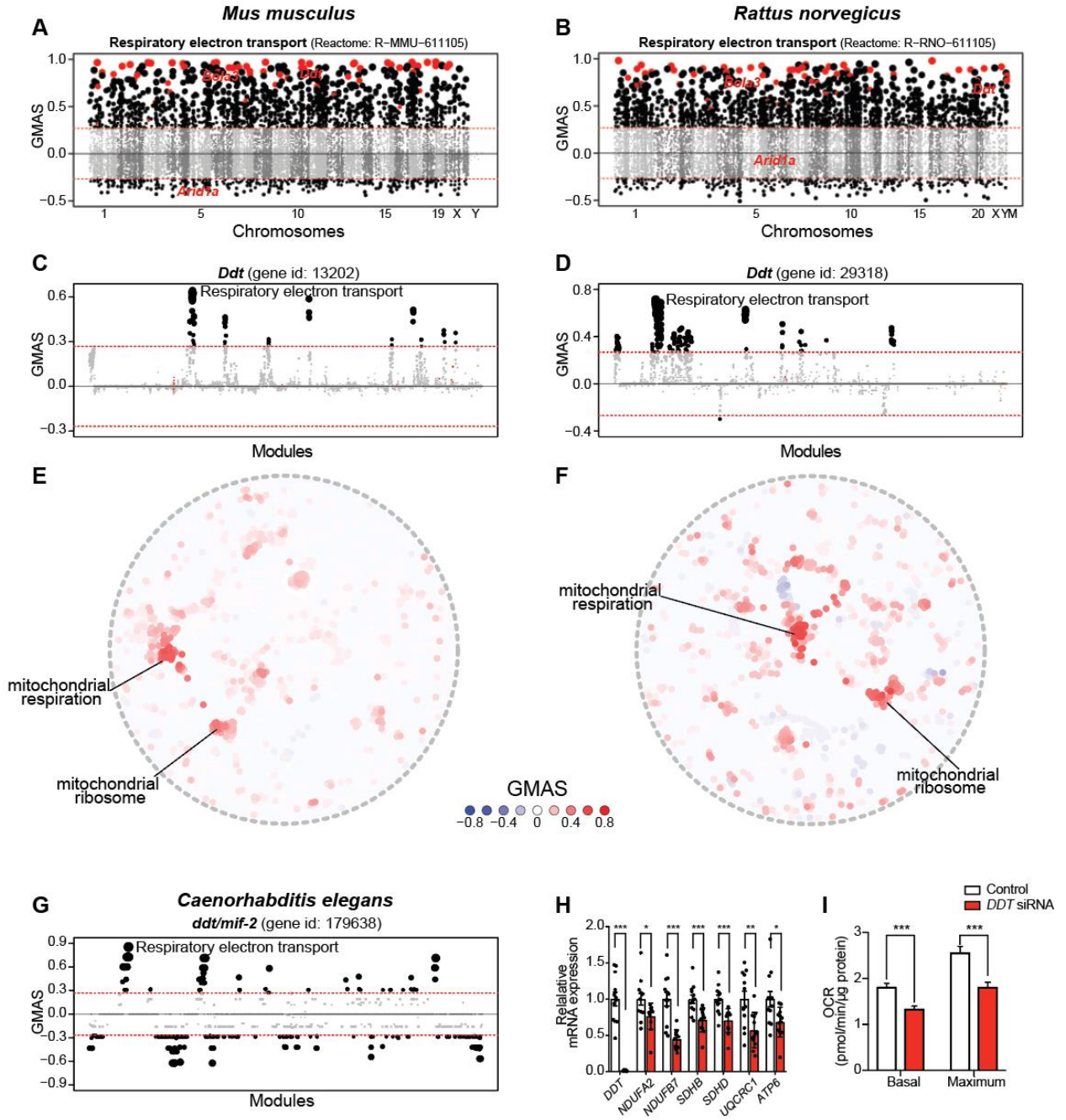


Figure S3.7. G-MAD verifies the potential involvement of *DDT* in mitochondrial respiration in mouse and rat.

A, B, G-MAD of respiratory electron transport in mouse (**A**) and rat (**B**). The threshold of significant gene-module association is indicated by the red dashed line. Genes are arranged based on their genetic positions. Genes annotated to be involved in respiratory electron transport are shown in red dots, and other genes with GMAS over 0.268 are highlighted by black dots.

C, D, G-MAD of *Ddt* in mouse (**C**) and rat (**D**) confirms its involvement in mitochondrial respiratory chain. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by their similarities. Known modules connected to *Ddt* from gene annotations are shown in red dots, and other modules with GMAS over the threshold are shown by black dots.

E, F, Network plots showing the significantly connected modules of *Ddt* in mouse (**E**) and rat (**F**). Modules are colored based on their GMAS against *Ddt*.

G, G-MAD of *ddt/mif-2* in *C. elegans* confirms its involvement in mitochondrial respiratory chain function also in invertebrates. The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by their similarities. Known modules connected to *ddt/mif-2* from gene annotations are shown in red dots, and other modules with GMAS over the threshold are shown by black dots.

H, Silencing *DDT* expression in HEK293 cells decreases expression levels of indicated genes involved in mitochondrial respiratory chain complexes. Error bars represent standard errors. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. n=12.

I, *DDT* RNAi reduced oxygen consumption rate (OCR) in HEK293 cells. Results were computed from Figure **3.4F**. ***, $p < 0.001$.

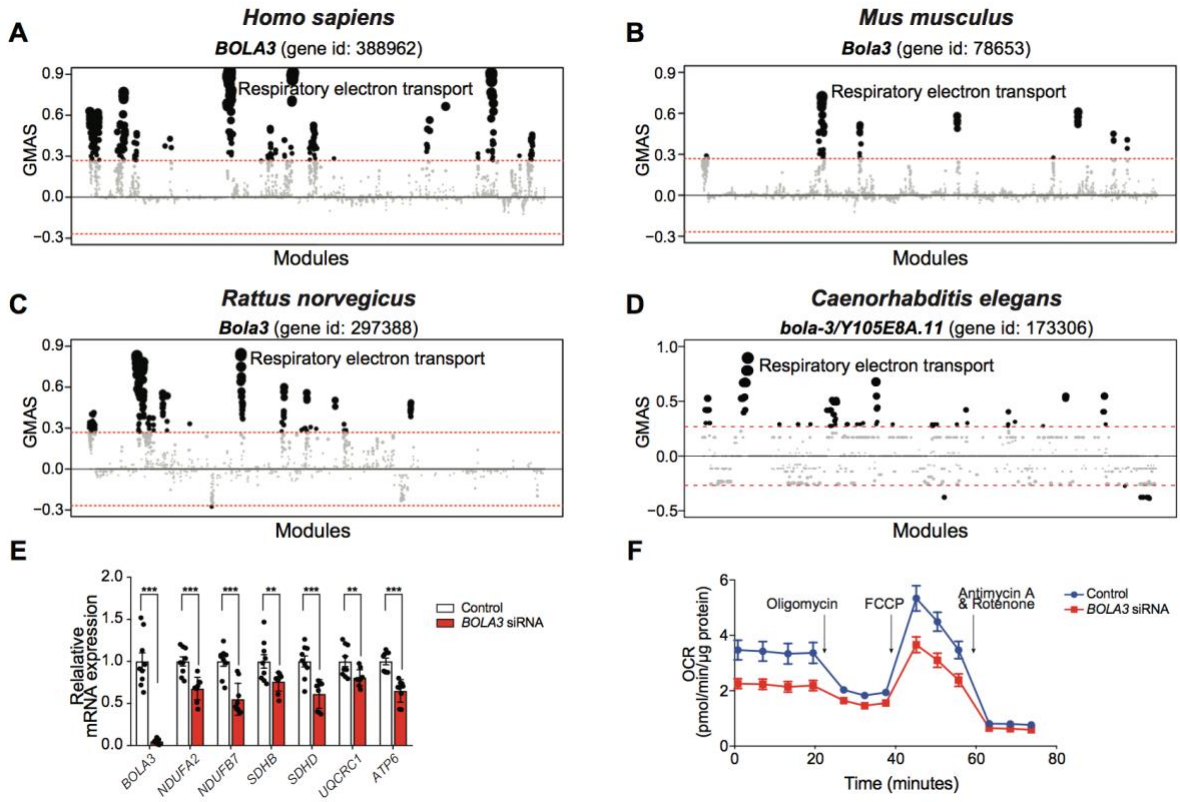


Figure S3.8. G-MAD confirms the involvement of *BOLA3* in mitochondrial respiration.

A-D, *BOLA3/Bola3/bola-3* associates with mitochondrial respiratory chain modules in human (**A**), mouse (**B**), rat (**C**), and *C. elegans* (**D**). The threshold of significant gene-module association is indicated by the red dashed line. Modules are organized by module similarities. Known modules connected to *BOLA3/Bola3/bola-3* from annotations are highlighted in red (no connected modules for *BOLA3/Bola3/bola-3*), and other modules with GMAS over the threshold are colored in black. Dot sizes reflect the GMAS of *BOLA3* against the respective modules.

E, Silencing *BOLA3* expression in HEK293 cells decreases expression levels of indicated genes involved in mitochondrial respiratory chain complexes. Error bars represent standard errors. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. n=9.

F, *BOLA3* knockdown leads to the reduction of oxygen consumption rate (OCR) as a reflection of mitochondrial respiration in human HEK293 cells. Addition of specific mitochondrial inhibitors, including the oligomycin (ATPase inhibitor), FCCP (uncoupling agent), and rotenone/antimycin A (electron transport chain inhibitors) are indicated.

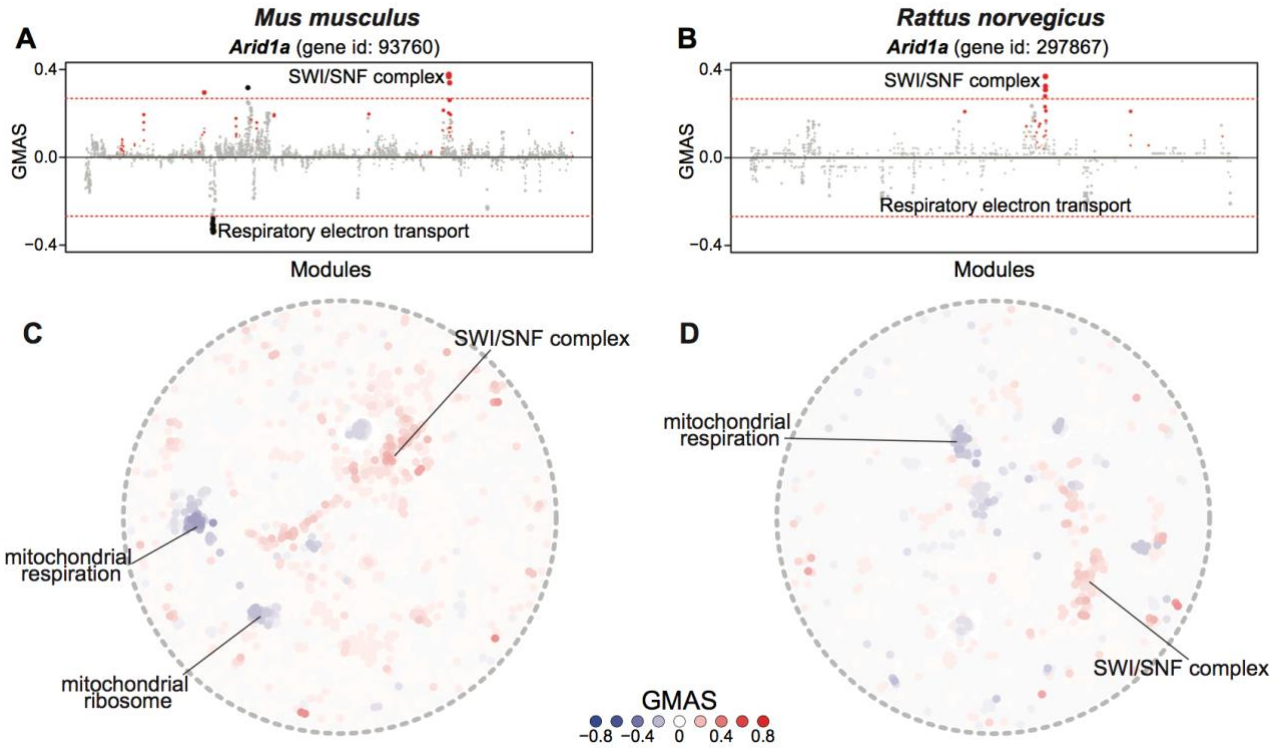


Figure S3.9. G-MAD verifies the negative association of *Arid1a* with mitochondrial respiration in mouse and rat.

A, B, G-MAD of *Arid1a* in mouse (**A**) and rat (**B**) confirms its negative association with mitochondrial respiratory chain. The threshold of significant gene-module association is indicated by the red dashed line; note that the associations are only suggestive in the mouse. Modules are organized by their similarities. Known modules connected to *Arid1a* from gene annotations are shown in red dots, and other modules with GMAS over the threshold are shown by black dots.

C, D, Network plot showing the significantly connected modules of *Arid1a* in mouse (**C**) and rat (**D**). Modules are colored based on their GMAS against *Arid1a*.

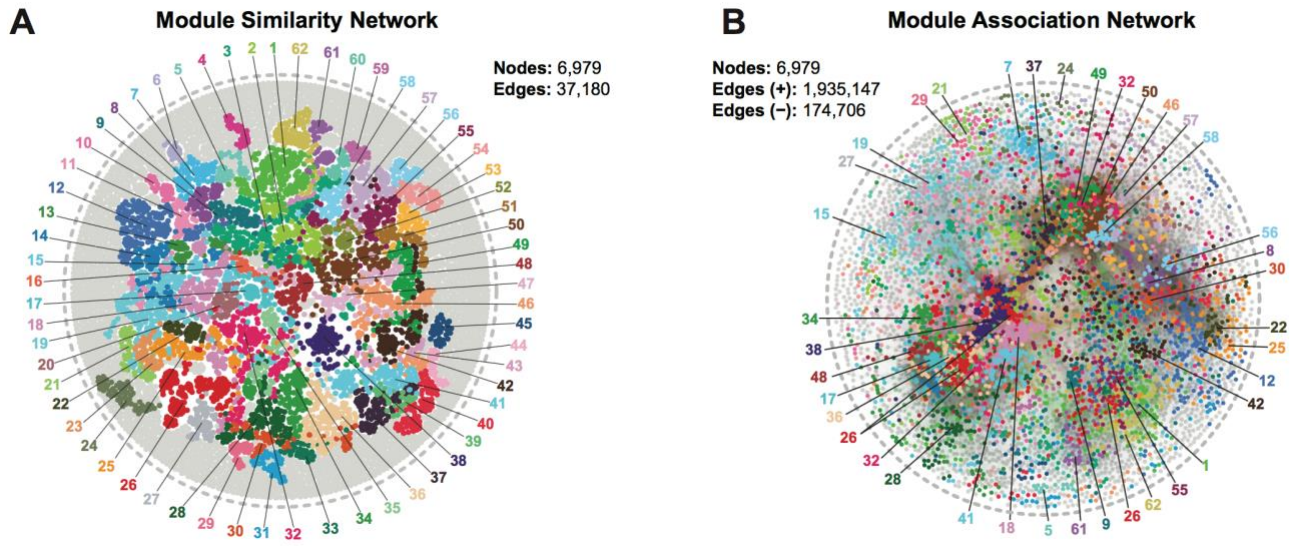


Figure S3.10. Comparison of module similarity network and module association network.

Module network obtained from module similarity (**A**) and association (**B**; M-MAD) were put together to facilitate the comparison.

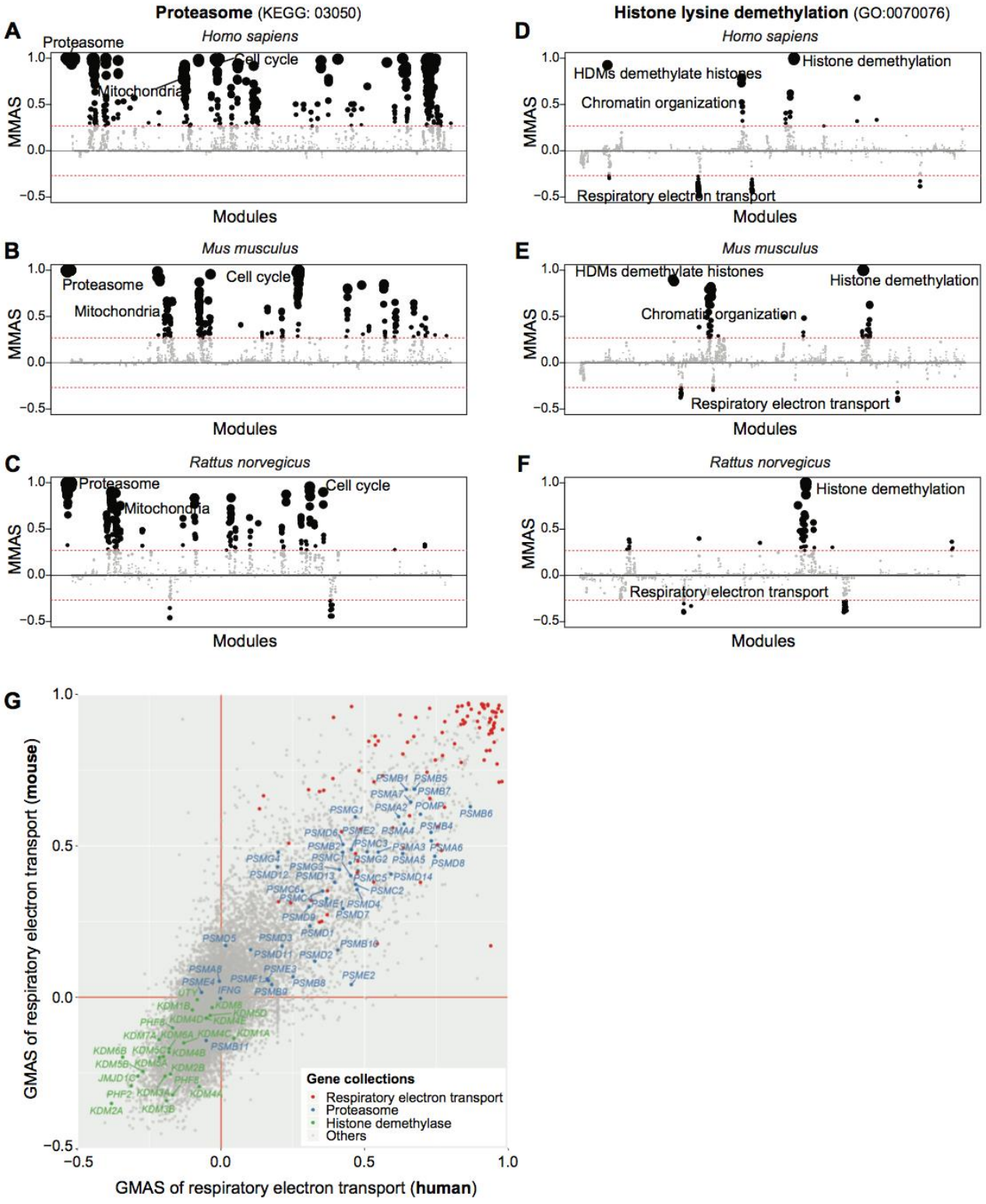


Figure S3.11. M-MAD and G-MAD identify the connections between respiratory electron transport chain and proteasome, as well as histone lysine demethylation.

A-F, M-MAD results for proteasome (**A-C**) and histone lysine demethylation (**D-F**) in human (**A, D**), mouse (**B, E**), and rat (**C, F**). The threshold of significant module-module connection is indicated by the red dashed line. Modules are organized by the module similarities. Dot sizes are proportional to MMAS of the respective modules.

G, G-MAD results for electron transport chain from human and mouse are plotted in the x and y axes, respectively. Genes annotated to be involved in respective modules are indicated in different colors.

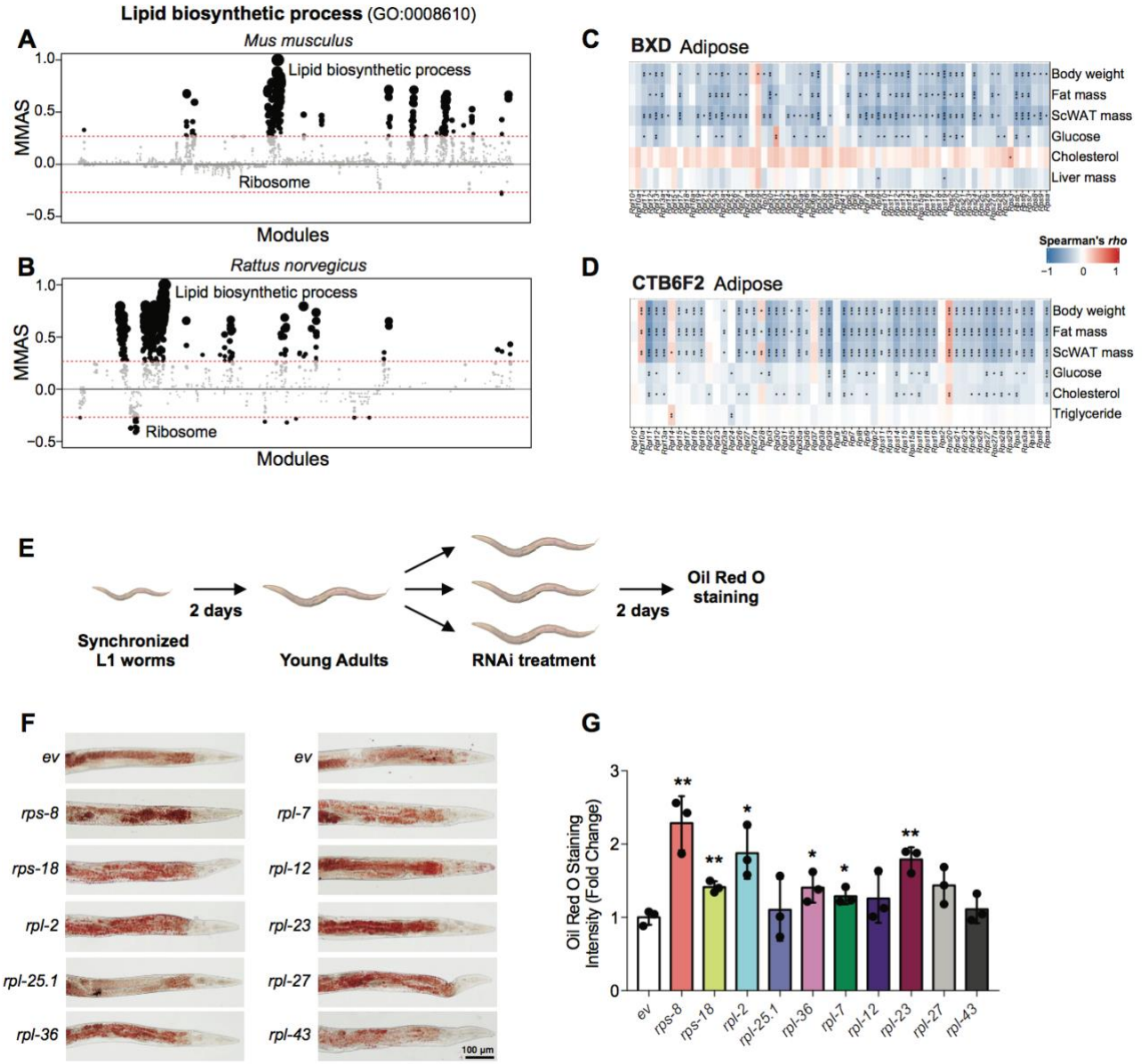


Figure S3.12. Validations of the negative association between ribosome and lipid biosynthetic modules.

A-B, M-MAD results for lipid biosynthetic process in mouse (**A**) and rat (**B**). The threshold of significant module-module connection is indicated by the red dashed line; note that the associations are only suggestive in the mouse. Modules are organized by the module similarities. Dot sizes are proportional to MMAS of the respective modules.

C-D, Transcripts of genes encoding for ribosomal proteins in the white adipose tissue negatively correlate with metabolic traits in the BXD (**C**) and CTB6F2 (**D**) mouse cohorts. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. ScWAT, subcutaneous white adipose tissue.

E, Scheme of the experimental design. L1 worm larvae were grown on regular NGM plates at 20°C for 2 days and then transferred to RNAi plates with 1mM IPTG containing HT115 bacteria expressing RNAi clones for ribosomal genes or empty vector (*ev*). After 2 days, worms were collected and lipid droplets were stained using Oil Red O.

F, Representative images of worm lipid droplet staining after RNAi of ribosomal genes.

G, Quantification of the lipid droplet staining intensity in **F**. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. n=3.

3.7 Acknowledgements

We are grateful to the research groups who made these data publicly available for systems biology research. We thank N. Agarwal for help in data preprocessing. 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. This work was

supported by grants from the EPFL, the ERC (AdG-787702), the SNSF (310030B-160318), the AgingX program of the Swiss Initiative for Systems Biology (RTD 2013/153), and the NIH (R01AG043930).

Chapter 4 Conclusion and perspectives

4.1 Research summary

In this thesis, I developed and explored several systems genetics approaches to identify the potential function of genes by integrating public available datasets either from a mouse genetic reference population (**Chapter 2**) or from six most commonly used model organisms (**Chapter 3**). Data obtained from these approaches has been made publicly accessible through systems-genetics.org/, a systems genetics resource for the associations between genetic variants, gene, module (pathway), and phenotype.

In **Chapter 2**, by using the multi-omics datasets from the BXD mouse genetic reference population, I proposed and developed a series of integrative systems approaches. PheWAS (Phenome-Wide Association Study), as a reverse approach of the commonly used GWAS (Genome-Wide Association Study), can be used to identify the phenotypic traits associated to the genetic variants of a gene-of-interest. PheWAS identifies numerous gene-phenotype associations, including those between *Pten* and prepulse inhibition, *Tlr5* and T cell proliferation, *Oprm1* and morphine response. As an example, I illustrated the novel associations between *Rpl26* and body weight as well as fat mass. The connection of ribosomal protein genes and lipid biosynthesis was confirmed using GeneBridge and also validated in *C. elegans* in Chapter 3, suggesting a conserved connection between the ribosome and lipid biosynthesis. T/PWAS (Transcriptome/Proteome-Wide Association Study) reveals the association between one gene and phenotypes based on its expression levels either at the transcript (mRNA) or protein level. The BXDs with expression datasets from more than 30 tissues available from the same mouse strains serve as a perfect source for such analysis. In addition, the phenome of over 5,000 phenotypic traits enabled us to propose ePheWAS (expression-based PheWAS), the reverse approach of T/PWAS, to associate the gene-of-interest with phenotypes based on its expression levels. Unlike PheWAS, ePheWAS associates a gene with phenotypes based on its expression levels in respective tissues and does not require the target gene to have high impact genetic variants. Therefore, ePheWAS can reveal gene-phenotype associations that cannot be identified through PheWAS, for example *Slc25a10* was associated with VO_2, body weight and fat mass, *Cpt1a* with acylcarnitines and fasting weight loss, *Cd36* with fat mass and acid beta-glucosidase activity, *Abca8a* with plasma triglycerides. By summarizing the gene-phenotype associations identified through ePheWAS using expression data from different tissues, I showed that phenotypic traits were under tissue-specific regulations. In this thesis, mediation analysis was also applied and further extended to reverse-mediation to elucidate the regulatory mechanism of gene expression using the expression datasets in the BXD cohort. Mediation analysis tries to identify the mediating effects of potential mediators between an independent variable and a dependent variable, and was applied in bioinformatic analysis to identify the mediators between SNPs and *trans*-regulated genes. The regulatory effects of genes, including BCKDHB on BCKDHA, *Zfp277* on *Rpsa* and *Rps2*, *E2f6* on *Cyp2j9*, *Fggy,* and *Txndc5*, as well as *Zkscan1* on *Adam10*, *Atl2*, and *Phf3* were revealed. In summary, PheWAS and ePheWAS identified hundreds of thousands of potential gene-phenotype associations, many of which are novel. Mediation and reverse-mediation suggested the regulatory mechanism between genes. Altogether, these approaches provide a systems genetics resource for researchers to generate and validate hypotheses related to their research interests.

In **Chapter 3**, I proposed a novel systems approach, termed gene-module association determination (G-MAD), to impute gene function from large-scale high-throughput expression datasets from 6 different model organisms. In contrast to existing methods that predict gene functions based on co-expression analysis, G-MAD does not require an arbitrary threshold for gene filtering and makes use of all the genes for the analysis. Therefore G-MAD is capable to identify potential negative associations between genes and modules, which is impossible for most of the existing methods. In addition, G-MAD has much better predictive performance than existing methods, such as WeGET [159] and COXPRESdb [160], and the improved performance does not rely on the larger dataset numbers in this study. Using G-MAD, researchers are able to discover novel gene functions and identify new members of given modules. As an example, I confirmed the recently identified role of *WDFY4* in activating T cells and immune response, and proposed 20 genes to be associated to cholesterol

biosynthesis. I also identified about 700 genes associated to mitochondria, the main power source of cells, and experimentally validated one of the top genes *DDT in vitro*. Furthermore, I confirmed the negative association between *ARID1A*, a known member of the SWI/SNF family, and mitochondrial respiration. From the expression compendia collected in this thesis, I observed tissue-specific co-expression or activation of modules, for example genes in the pancreatic secretion module are co-expressed to a higher extent in pancreatic datasets. Hence G-MAD can also extract tissue-specific gene-module associations by using datasets from specific tissues. As an example, I identified the novel function of *EHHADH* in brush border membrane in kidney besides its well-known role in peroxisomal oxidation pathway in liver. The novel liver specific fatty acid transport function of *SLC6A1* was also discovered in addition to its role as a major neurotransmitter transporter in brain. I then further extended G-MAD to develop Module-Module Association Determination (M-MAD) to investigate the connections between modules based on the expression compendia. Interestingly, there are numerous module pairs with no overlap of annotated genes, but with high association based on M-MAD, for example mitochondrial modules and the proteasome, indicating a conserved co-regulatory mechanism. Similar to G-MAD, M-MAD can also reveal negative associations between modules, for instance histone demethylation and mitochondria, underscoring the inverse connection between epigenetic regulation and mitochondrial function. We also discovered and experimentally verified the negative connection between lipid biosynthetic process and ribosomal modules, which is in line with our previous finding in Chapter 2 that a ribosomal protein, *Rpl26*, negatively correlates with body weight and fat mass.

4.2 Perspectives

Systems approaches developed in this thesis can not only be used on the data described previously, but are fairly easy to be employed and adapted for other data or purposes. There are several potential updates and improvements that can be built on what has been accomplished in this thesis.

In **Chapter 2**, I used the averaged data of phenotypic measures for each strain downloaded from genenetwork.org to employ the systems genetics approaches, since individual measurements are not available for most of the phenotypic traits collected in the BXD animals. The use of individual measurements would significantly increase the statistical power of detecting associations [76]. It would be great if the individual data from these historical traits could be included in the analysis pipeline, such that these data can be better exploited and appreciated. In addition, there are large amount of data collected by the mouse genetics community from other mouse populations, for example the HMDP [43]. These data could not only serve as independent resources for validation analyses, but also provide the possibility to apply meta-analytical approaches to discover novel biological findings, as now commonly done in human genetics studies [211].

In recent years, transcriptome-wide association studies (TWAS) was proposed to integrate human GWAS and gene expression datasets to identify gene-phenotype associations with the aim to prioritize causal genes [212-229]. This approach does not require the gene expression datasets to be collected from the same population of GWAS, and therefore is perfect for human genetics studies, where accessing samples from deep tissues of participants is nearly impossible. Such approach can also be applied in mouse studies to use transcriptome data from one cohort to associate with phenotypes from other cohorts. In this scenario, the BXDs with genotype and expression data available from over 30 tissues can serve as a perfect gene expression reference panel.

Furthermore, integration of other layers of omics data, e.g. epigenomics [44, 230] or microbiomics [54], into the existing analytic methods would be very motivating to study the influence of other factors on phenotypic traits. Approaches including GWAS and PheWAS normally utilize specific thresholds based on Bonferroni correction according to the number of genetic variants or phenotypes to determine the significance of the associations. One future direction is to propose alternative statistical methods that could avoid such reliance to compute a common statistics as a measure of the association between two omics layers. In addition, the approaches applied in this thesis mostly explore the pair-wise connections between two layers. More sophisticated computational approaches, for example artificial intelligence or machine learning [231-233], can be applied to integrate the multi-layered omics datasets to establish predictive models or networks in a more

holistic framework. Such integration will optimistically allow a better understanding of the basis of physiological traits or diseases.

In **Chapter 3**, the accuracy of the GeneBridge will be gradually improved with the further understanding of genes and annotations. In this thesis, I only included transcriptome datasets with over 80 samples. It would be interesting to see how the predictive performance changes when including datasets with less samples (e.g. 50). This is valuable especially for organisms where such large-scale studies are less available, for example *C. elegans*. Bonferroni thresholding was applied based on the number of modules in GeneBridge to scale the final association score into a range of [-1, 1]. Owing to the fact that some modules have very high similarity in their gene compositions, Bonferroni correction is too much stringent and might cause the problem of false negatives. One potential improvement of this approach here is to compute the number of independent or effective modules based on their similarities. In addition, some gene-module pairs might have enrichment p values very close to the Bonferroni threshold in most of the datasets, but end up with very low final association scores. In order to avoid such situation, it is possible to apply a series of thresholds to transfer the continuous enrichment p values into a smaller number of intervals between -1 and 1, for example -1, -0.5, 0, 0.5, 1. In this case, we may be able to capture the nonsignificant but consistent connections across datasets.

I used several recently annotated genes in this thesis to prove that our method is capable of identifying novel gene-module associations. However, this approach may appear to be biased to the audience, since only very few genes are demonstrated and they might be selected subjectively. Actually, I tested many newly published gene functions, for example the negative control of *REST* on neuronal activity [234], the regulation of *DWORF* on skeletal muscle function [235, 236], the negative influence of *PUM2* on mitochondrial function [237, 238], the impact of *LINC00116* / *MTLN* on mitochondrial respiration [239-241], and the involvement of *PUSL1* in mitochondrial ribosome [242]. It would be interesting to systematically evaluate the performance of G-MAD on either all the recently published links between genes and modules, or through experimental validations retrospectively after a few years, as done in the CAFA (critical assessment of functional annotation) [145].

Meanwhile, one of the issues and limitations of modules (gene set or pathway) related methods is that some of the modules are poorly annotated. It is possible to pinpoint such modules using the source and approaches described in this thesis, for example through the gene correlations within modules using the collected expression compendia. The modules with low correlations across genes are most likely badly annotated modules. Modules that have poor predictive performance in G-MAD also tend to be poorly annotated, since ideally all the known gene would be ranked in the top of the analysis if a module is well annotated.

What should be noted is that one gene is able to encode multiple protein isoforms and thus have different biological functions. This process is regulated by alternative splicing in the production of mRNA [243, 244]. However, common microarrays are not capable to detect the different RNA splicing isoforms [245] and large part of the datasets are obtained from microarrays. Therefore, in this thesis I ignored the RNA splicing isoforms and only focused on gene level quantifications in RNA-seq datasets to match with the microarray datasets. It would be interesting to extract the abundance of the alternative splicing isoforms from RNA-seq datasets and predict their respective functions. Besides, due to the incapability of microarray in determining expression of non-protein coding genes, I only focused on protein coding genes in the analysis. However, the approaches described in this thesis can be easily utilized to determine the possible functions of non-protein coding genes with such data available.

In recent years, more and more single cell RNA sequencing (scRNA-seq) studies covering gene expression from hundreds or thousands of cells have been conducted. But due to several remaining issues of scRNA-seq, for instance over-abundance of zero-values and data normalization [246-248], I did not include data from scRNA-seq in the current analysis. It would, however, be very interesting to include these datasets when these issues are solved, considering the large sample sizes from single cell analyses. Another way to include scRNA-

seq data in the analysis is to split the data based on the cell types, so that the lowly expressed genes for the respective cell types could be ignored in the analysis.

Several recent studies demonstrated that proteome profiling outperforms transcriptome profiling in predicting gene functions using co-expression analysis [249-252]. These results make biological sense, since protein, the next phase of mRNA in the central dogma of molecular biology, provides a more accurate estimation of gene activity. However, the current quantitative proteomics techniques are not able to measure all proteins (limited to about 6,000 proteins), especially those lowly expressed and the membrane proteins. Such proteome datasets can be utilized to predict gene functions using a “guilt-by-association” strategy that relies on gene subsets that fulfill certain criteria, but are not suitable in the approach described in this thesis, where genome-wide measurements are required. Therefore, a future direction is to propose new methods that can correct for the background of measured proteins. Besides, there are many existing tools that predict gene functions based on other gene features, including sequence homology [143-145], protein structure [144-146], phylogenetic profiles [147-149], protein-protein interactions [150-152], and genetic interactions [153-155]. In addition, data from linkage or association studies (e.g. GWAS), or from high-throughput genetic screening experiments (e.g. CRISPR screening), or from animal gain-or-loss- of function studies, or from the gene-drug interactions, can also be exploited to predict potential gene functions. Integration of GeneBridge with data from these sources will further enhance the performance for gene function prediction, as is done in STRING [253], GeneMANIA [254] and Mitocarta [190, 255].

To summarize, the work presented in this thesis proposed and applied a series of novel approaches for systems genetics analysis, and provided an open and easy-to-use web resource (systems-genetics.org) to the research community for *in silico* hypothesis generation and validation. Systems approaches established in this thesis can be also easily applied on other cohorts with multi-omics data available, for example the HMDP mouse cohort [43], the GTEx project [256] and the UK Biobank study [257]. One question is how could the systems genetics approaches and findings described in this thesis contribute to the development of precision medicine. Recently, several concerns have been raised that precision medicine has so far not been very successfully, especially for common complex diseases [258]. To my understanding, what remains one of the prime obstacles for the implementation of precision medicine is that the function for the majority of the genes still remain neglected and poorly studied [78, 84, 89], while most of the research efforts have been devoted to only a small set of genes. The systems genetics tools kit and data sources can directly help by proposing gene annotations in an unbiased manner. The performance and applicability of the findings from these approaches can be tested in human diseases and drug targets. Given that PheWAS and ePheWAS can reveal the potential pleiotropic functions of genes, (e)PheWAS can be applied on human drug target-disease connections to explore the likelihood of drug side effects or drug repurposing. The gene-module associations identified by GeneBridge can also be similarly employed to identify possible drug mechanisms and suggest drug adverse effects and facilitate drug repositioning.

References

1. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.
2. Roden, D.M., et al., *Pharmacogenomics: the genetics of variable drug responses.* Circulation, 2011. **123**(15): p. 1661-70.
3. Zeevi, D., et al., *Personalized Nutrition by Prediction of Glycemic Responses.* Cell, 2015. **163**(5): p. 1079-1094.
4. Buford, T.W., M.D. Roberts, and T.S. Church, *Toward exercise as personalized medicine.* Sports Med, 2013. **43**(3): p. 157-65.
5. Hamburg, M.A. and F.S. Collins, *The path to personalized medicine.* N Engl J Med, 2010. **363**(4): p. 301-4.
6. Collins, F.S. and H. Varmus, *A new initiative on precision medicine.* N Engl J Med, 2015. **372**(9): p. 793-5.
7. Li, H., et al., *An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function.* Cell Syst, 2018. **6**(1): p. 90-102 e4.
8. Wang, X., et al., *Joint mouse-human phenome-wide association to test gene function and disease risk.* Nat Commun, 2016. **7**: p. 10464.
9. Seok, J., et al., *Genomic responses in mouse models poorly mimic human inflammatory diseases.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. 3507-12.
10. Takao, K. and T. Miyakawa, *Genomic responses in mouse models greatly mimic human inflammatory diseases.* Proc Natl Acad Sci U S A, 2015. **112**(4): p. 1167-72.
11. Nadeau, J.H. and J. Auwerx, *The virtuous cycle of human genetics and mouse models in drug discovery.* Nat Rev Drug Discov, 2019. **18**(4): p. 255-272.
12. Teufel, A., et al., *Comparison of Gene Expression Patterns Between Mouse Models of Nonalcoholic Fatty Liver Disease and Liver Tissues From Patients.* Gastroenterology, 2016. **151**(3): p. 513-525 e0.
13. Sittig, L.J., et al., *Genetic Background Limits Generalizability of Genotype-Phenotype Relationships.* Neuron, 2016. **91**(6): p. 1253-1259.
14. Neuner, S.M., et al., *Harnessing Genetic Complexity to Enhance Translatability of Alzheimer's Disease Mouse Models: A Path toward Precision Medicine.* Neuron, 2019. **101**(3): p. 399-411 e5.
15. Beura, L.K., et al., *Normalizing the environment recapitulates adult human immune traits in laboratory mice.* Nature, 2016. **532**(7600): p. 512-6.
16. Kleinert, M., et al., *Animal models of obesity and diabetes mellitus.* Nat Rev Endocrinol, 2018. **14**(3): p. 140-162.
17. Kebede, M.A. and A.D. Attie, *Insights into obesity and diabetes at the intersection of mouse and human genetics.* Trends Endocrinol Metab, 2014. **25**(10): p. 493-501.
18. von Scheidt, M., et al., *Applications and Limitations of Mouse Models for Understanding Human Atherosclerosis.* Cell Metab, 2017. **25**(2): p. 248-261.
19. Schiattarella, G.G., et al., *Nitrosative stress drives heart failure with preserved ejection fraction.* Nature, 2019. **568**(7752): p. 351-356.
20. Zuberi, A. and C. Lutz, *Mouse Models for Drug Discovery. Can New Tools and Technology Improve Translational Power?* ILAR J, 2016. **57**(2): p. 178-185.
21. Philip, V.M., et al., *High-throughput behavioral phenotyping in the expanded panel of BXD recombinant inbred strains.* Genes Brain Behav, 2010. **9**(2): p. 129-59.
22. Williams, E.G., et al., *Systems proteomics of liver mitochondria function.* Science, 2016. **352**(6291): p. aad0189.
23. Liao, C.Y., et al., *Genetic variation in the murine lifespan response to dietary restriction: from life extension to life shortening.* Aging Cell, 2010. **9**(1): p. 92-5.
24. Johnson, M., *Laboratory Mice and Rats.* Mater. Methods, 2012. **2**: p. 113.
25. Fontaine, D.A. and D.B. Davis, *Attention to Background Strain Is Essential for Metabolic Research: C57BL/6 and the International Knockout Mouse Consortium.* Diabetes, 2016. **65**(1): p. 25-33.
26. Simon, M.M., et al., *A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains.* Genome Biol, 2013. **14**(7): p. R82.
27. Lilue, J., et al., *Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci.* Nat Genet, 2018. **50**(11): p. 1574-1583.
28. Keane, T.M., et al., *Mouse genomic variation and its effect on phenotypes and gene regulation.* Nature, 2011. **477**(7364): p. 289-94.
29. Lapuente-Brun, E., et al., *Supercomplex assembly determines electron flux in the mitochondrial electron transport chain.* Science, 2013. **340**(6140): p. 1567-70.
30. Berrettini, W.H., et al., *Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains.* Nat Genet, 1994. **7**(1): p. 54-8.
31. Onos, K.D., et al., *Enhancing face validity of mouse models of Alzheimer's disease with natural genetic variation.* PLoS Genet, 2019. **15**(5): p. e1008155.
32. Bennett, B.J., et al., *Genetic Architecture of Atherosclerosis in Mice: A Systems Genetics Analysis of Common Inbred Strains.* PLoS Genet, 2015. **11**(12): p. e1005711.
33. Winter, J.M., et al., *Mapping Complex Traits in a Diversity Outbred F1 Mouse Population Identifies Germline Modifiers of Metastasis in Human Prostate Cancer.* Cell Syst, 2017. **4**(1): p. 31-45 e6.

34. Barrington, W.T., et al., *Improving Metabolic Health Through Precision Dietetics in Mice.* Genetics, 2018. **208**(1): p. 399-417.
35. Pound, P. and M.B. Bracken, *Is animal research sufficiently evidence based to be a cornerstone of biomedical research?* BMJ, 2014. **348**: p. g3387.
36. Williams, E.G. and J. Auwerx, *The Convergence of Systems and Reductionist Approaches in Complex Trait Analysis.* Cell, 2015. **162**(1): p. 23-32.
37. Ashbrook, D.G., et al., *The expanded BXD family of mice: A cohort for experimental systems genetics and precision medicine.* bioRxiv, 2019: p. 672097.
38. Williams, R.W., et al., *Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis.* Mamm Genome, 2004. **15**(8): p. 637-47.
39. Saul, M.C., et al., *High-Diversity Mouse Populations for Complex Traits.* Trends Genet, 2019. **35**(7): p. 501-514.
40. Collaborative Cross, C., *The genome architecture of the Collaborative Cross mouse genetic reference population.* Genetics, 2012. **190**(2): p. 389-401.
41. Buchner, D.A. and J.H. Nadeau, *Contrasting genetic architectures in different mouse reference populations used for studying complex traits.* Genome Res, 2015. **25**(6): p. 775-91.
42. Bennett, B.J., et al., *A high-resolution association mapping panel for the dissection of complex traits in mice.* Genome Res, 2010. **20**(2): p. 281-90.
43. Lusis, A.J., et al., *The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits.* J Lipid Res, 2016. **57**(6): p. 925-42.
44. Baker, C.L., et al., *Tissue-Specific Trans Regulation of the Mouse Epigenome.* Genetics, 2019. **211**(3): p. 831-845.
45. Orozco, L.D., et al., *Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice.* Cell Metab, 2015. **21**(6): p. 905-17.
46. Wu, Y., et al., *Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population.* Cell, 2014. **158**(6): p. 1415-1430.
47. Chick, J.M., et al., *Defining the consequences of genetic variation on a proteome-wide scale.* Nature, 2016. **534**(7608): p. 500-5.
48. Jha, P., et al., *Systems Analyses Reveal Physiological Roles and Genetic Regulators of Liver Lipid Species.* Cell Syst, 2018. **6**(6): p. 722-733 e6.
49. Jha, P., et al., *Genetic Regulation of Plasma Lipid Species and Their Association with Metabolic Phenotypes.* Cell Syst, 2018. **6**(6): p. 709-721 e6.
50. Parker, B.L., et al., *An integrative systems genetic analysis of mammalian lipid metabolism.* Nature, 2019. **567**(7747): p. 187-193.
51. Ghazalpour, A., et al., *Genetic regulation of mouse liver metabolite levels.* Mol Syst Biol, 2014. **10**: p. 730.
52. Parks, B.W., et al., *Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice.* Cell Metab, 2013. **17**(1): p. 141-52.
53. Org, E., et al., *Genetic and environmental control of host-gut microbiota interactions.* Genome Res, 2015. **25**(10): p. 1558-69.
54. McKnite, A.M., et al., *Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits.* PLoS One, 2012. **7**(6): p. e39191.
55. Houtkooper, R.H., et al., *Mitonuclear protein imbalance as a conserved longevity mechanism.* Nature, 2013. **497**(7450): p. 451-7.
56. Seldin, M.M., et al., *A Strategy for Discovery of Endocrine Interactions with Application to Whole-Body Metabolism.* Cell Metab, 2018. **27**(5): p. 1138-1155 e6.
57. Mesner, L.D., et al., *Mouse genome-wide association and systems genetics identifies Lhfp as a regulator of bone mass.* PLoS Genet, 2019. **15**(5): p. e1008123.
58. Parker, C.C., et al., *Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice.* Nat Genet, 2016. **48**(8): p. 919-26.
59. Nicod, J., et al., *Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing.* Nat Genet, 2016. **48**(8): p. 912-8.
60. Keller, M.P., et al., *Gene loci associated with insulin secretion in islets from non-diabetic mice.* J Clin Invest, 2019. **130**.
61. Keller, M.P., et al., *Genetic Drivers of Pancreatic Islet Function.* Genetics, 2018. **209**(1): p. 335-356.
62. Gonzales, N.M., et al., *Genome wide association analysis in a mouse advanced intercross line.* Nat Commun, 2018. **9**(1): p. 5162.
63. Valdar, W., et al., *Genome-wide genetic association of complex traits in heterogeneous stock mice.* Nat Genet, 2006. **38**(8): p. 879-87.
64. Nadon, N.L., et al., *NIA Interventions Testing Program: Investigating Putative Aging Intervention Agents in a Genetically Heterogeneous Mouse Model.* EBioMedicine, 2017. **21**: p. 3-4.
65. Keller, M.P., et al., *The Transcription Factor Nfatc2 Regulates beta-Cell Proliferation and Genes Associated with Type 2 Diabetes in Mouse and Human Islets.* PLoS Genet, 2016. **12**(12): p. e1006466.
66. Tu, Z., et al., *Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets.* PLoS Genet, 2012. **8**(12): p. e1003107.
67. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.
68. Rasmussen, A.L., et al., *Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance.* Science, 2014. **346**(6212): p. 987-91.

69. Nadeau, J.H., et al., *Chromosome substitution strains: gene discovery, functional analysis, and systems studies.* Mamm Genome, 2012. **23**(9-10): p. 693-705.
70. Koutnikova, H., et al., *Identification of the UBP1 locus as a critical blood pressure determinant using a combination of mouse and human genetics.* PLoS Genet, 2009. **5**(8): p. e1000591.
71. Denny, J.C., L. Bastarache, and D.M. Roden, *Phenome-Wide Association Studies as a Tool to Advance Precision Medicine.* Annu Rev Genomics Hum Genet, 2016. **17**: p. 353-73.
72. Denny, J.C., et al., *PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.* Bioinformatics, 2010. **26**(9): p. 1205-10.
73. Gagneur, J., et al., *Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype.* PLoS Genet, 2013. **9**(9): p. e1003803.
74. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nat Genet, 2016. **48**(3): p. 245-52.
75. Okada, H., et al., *Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster.* Nat Commun, 2016. **7**: p. 12649.
76. Kang, H.M., et al., *Efficient control of population structure in model organism association mapping.* Genetics, 2008. **178**(3): p. 1709-23.
77. Gatti, D.M., et al., *The Effects of Sex and Diet on Physiology and Liver Gene Expression in Diversity Outbred Mice.* bioRxiv, 2017: p. 098657.
78. Li, H., et al., *Identifying gene function and module connections by the integration of multi-species expression compendia.* bioRxiv, 2019: p. 649079.
79. Rau, C.D., et al., *Systems Genetics Approach Identifies Gene Pathways and Adamts2 as Drivers of Isoproterenol-Induced Cardiac Hypertrophy and Cardiomyopathy in Mice.* Cell Syst, 2017. **4**(1): p. 121-128 e4.
80. Chella Krishnan, K., et al., *Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease.* Cell Syst, 2018. **6**(1): p. 103-115 e7.
81. Talukdar, H.A., et al., *Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease.* Cell Syst, 2016. **2**(3): p. 196-208.
82. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.
83. Porcu, E., et al., *Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits.* Nat Commun, 2019. **10**(1): p. 3300.
84. Haynes, W.A., A. Tomczak, and P. Khatri, *Gene annotation bias impedes biomedical research.* Sci Rep, 2018. **8**(1): p. 1362.
85. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
86. Mitchell, J.A., et al., *Gene indexing: characterization and analysis of NLM's GeneRIFs.* AMIA Annu Symp Proc, 2003: p. 460-4.
87. Dolgin, E., *The most popular genes in the human genome.* Nature, 2017. **551**(7681): p. 427-431.
88. Edwards, A.M., et al., *Too many roads not taken.* Nature, 2011. **470**(7333): p. 163-5.
89. Stoeger, T., et al., *Large-scale investigation of the reasons why potentially important genes are ignored.* PLoS Biol, 2018. **16**(9): p. e2006643.
90. Pandey, A.K., et al., *Functionally enigmatic genes: a case study of the brain ignorome.* PLoS One, 2014. **9**(2): p. e88889.
91. Greene, C.S. and O.G. Troyanskaya, *Accurate evaluation and analysis of functional genomics data and methods.* Ann N Y Acad Sci, 2012. **1260**: p. 95-100.
92. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
93. Aitman, T.J., et al., *The future of model organisms in human disease research.* Nat Rev Genet, 2011. **12**(8): p. 575-82.
94. Flint, J. and T.F. Mackay, *Genetic architecture of quantitative traits in mice, flies, and humans.* Genome Res, 2009. **19**(5): p. 723-33.
95. Cook, D.E., et al., *CeNDR, the Caenorhabditis elegans natural diversity resource.* Nucleic Acids Res, 2017. **45**(D1): p. D650-D657.
96. Ehrenreich, I.M., et al., *Dissection of genetically complex traits with extremely large pools of yeast segregants.* Nature, 2010. **464**(7291): p. 1039-42.
97. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
98. Flint, J. and E. Eskin, *Genome-wide association studies in mice.* Nat Rev Genet, 2012. **13**(11): p. 807-17.
99. Bush, W.S., M.T. Oetjens, and D.C. Crawford, *Unravelling the human genome-phenome relationship using phenome-wide association studies.* Nat Rev Genet, 2016. **17**(3): p. 129-45.
100. Wang, X., et al., *Joint mouse-human phenome-wide association to test gene function and disease risk.* Nat Commun, 2016. **7**.
101. Hubner, N., et al., *Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease.* Nat Genet, 2005. **37**(3): p. 243-53.
102. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.
103. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.

104. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
105. Yen, K., et al., *A comparative study of fat storage quantitation in nematode Caenorhabditis elegans using label and label-free methods.* PLoS One, 2010. **5**(9).
106. Stekhoven, D.J. and P. Buhlmann, *MissForest--non-parametric missing value imputation for mixed-type data.* Bioinformatics, 2012. **28**(1): p. 112-8.
107. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
108. Oba, S., et al., *A Bayesian missing value estimation method for gene expression profile data.* Bioinformatics, 2003. **19**(16): p. 2088-96.
109. Li, J. and L. Ji, *Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix.* Heredity (Edinb), 2005. **95**(3): p. 221-7.
110. Broman, K.W., et al., *R/qtl: QTL mapping in experimental crosses.* Bioinformatics, 2003. **19**(7): p. 889-90.
111. Segura, V., et al., *An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations.* Nat Genet, 2012. **44**(7): p. 825-30.
112. Peirce, J.L., et al., *A new set of BXD recombinant inbred lines from advanced intercross populations in mice.* BMC Genet, 2004. **5**: p. 7.
113. Taylor, B.A., H.J. Heiniger, and H. Meier, *Genetic analysis of resistance to cadmium-induced testicular damage in mice.* Proc Soc Exp Biol Med, 1973. **143**(3): p. 629-33.
114. Andreux, P.A., et al., *Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits.* Cell, 2012. **150**(6): p. 1287-99.
115. Alexander, R.P., et al., *Annotating non-coding regions of the genome.* Nat Rev Genet, 2010. **11**(8): p. 559-71.
116. Kwon, C.H., et al., *Pten regulates neuronal arborization and social interaction in mice.* Neuron, 2006. **50**(3): p. 377-88.
117. Ortega-Molina, A., et al., *Pten positively regulates brown adipose function, energy expenditure, and longevity.* Cell Metab, 2012. **15**(3): p. 382-94.
118. Caron, G., et al., *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, 2005. **175**(3): p. 1551-7.
119. Uhl, G.R., I. Sora, and Z. Wang, *The mu opiate receptor as a candidate gene for pain: polymorphisms, variations in expression, nociception, and opiate responses.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 7752-5.
120. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
121. Mancuso, N., et al., *Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits.* Am J Hum Genet, 2017. **100**(3): p. 473-487.
122. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.
123. Mizuarai, S., et al., *Identification of dicarboxylate carrier Slc25a10 as malate transporter in de novo fatty acid synthesis.* J Biol Chem, 2005. **280**(37): p. 32434-41.
124. Wing, R.R. and J.O. Hill, *Successful weight loss maintenance.* Annu Rev Nutr, 2001. **21**: p. 323-41.
125. Pande, S.V., *A mitochondrial carnitine acylcarnitine translocase system.* Proc Natl Acad Sci U S A, 1975. **72**(3): p. 883-7.
126. Silverstein, R.L. and M. Febbraio, *CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior.* Sci Signal, 2009. **2**(72): p. re3.
127. Grabowski, G.A., *Phenotype, diagnosis, and treatment of Gaucher's disease.* Lancet, 2008. **372**(9645): p. 1263-71.
128. Dean, M., A. Rzhetsky, and R. Allikmets, *The human ATP-binding cassette (ABC) transporter superfamily.* Genome Res, 2001. **11**(7): p. 1156-66.
129. MacKinnon, D.P., A.J. Fairchild, and M.S. Fritz, *Mediation analysis.* Annu Rev Psychol, 2007. **58**: p. 593-614.
130. Yao, C., et al., *Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits.* Am J Hum Genet, 2017. **100**(6): p. 985-986.
131. Pierce, B.L., et al., *Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians.* PLoS Genet, 2014. **10**(12): p. e1004818.
132. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.
133. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.
134. Donner, A.L., V. Episkopou, and R.L. Maas, *Sox2 and Pou2f1 interact to control lens and olfactory placode development.* Dev Biol, 2007. **303**(2): p. 784-99.
135. ENCODE Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
136. Sham, P.C. and S.M. Purcell, *Statistical power and significance testing in large-scale genetic studies.* Nat Rev Genet, 2014. **15**(5): p. 335-46.
137. GTEx Consortium, *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
138. 1001 Genomes Consortium, *1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana.* Cell, 2016. **166**(2): p. 481-491.
139. Mahmood, S.S., et al., *The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective.* Lancet, 2014. **383**(9921): p. 999-1008.
140. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nat Genet, 2015. **47**(9): p. 1091-8.

141. Dickinson, M.E., et al., *High-throughput discovery of novel developmental phenotypes.* Nature, 2016. **537**(7621): p. 508-514.
142. Austin, C.P., et al., *The knockout mouse project.* Nat Genet, 2004. **36**(9): p. 921-4.
143. Marcotte, E.M., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-3.
144. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.
145. Jiang, Y., et al., *An expanded evaluation of protein function prediction methods shows an improvement in accuracy.* Genome Biol, 2016. **17**(1): p. 184.
146. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nat Protoc, 2010. **5**(4): p. 725-38.
147. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.
148. Li, Y., et al., *Expansion of biological pathways based on evolutionary inference.* Cell, 2014. **158**(1): p. 213-25.
149. Tabach, Y., et al., *Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence.* Nature, 2013. **493**(7434): p. 694-8.
150. Huttlin, E.L., et al., *Architecture of the human interactome defines protein communities and disease networks.* Nature, 2017. **545**(7655): p. 505-509.
151. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-1226.
152. Hein, M.Y., et al., *A human interactome in three quantitative dimensions organized by stoichiometries and abundances.* Cell, 2015. **163**(3): p. 712-23.
153. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.
154. Tong, A.H., et al., *Global mapping of the yeast genetic interaction network.* Science, 2004. **303**(5659): p. 808-13.
155. Horlbeck, M.A., et al., *Mapping the Genetic Landscape of Human Cells.* Cell, 2018. **174**(4): p. 953-967 e22.
156. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W214-20.
157. Greene, C.S., et al., *Understanding multicellular function and disease with human tissue-specific networks.* Nat Genet, 2015. **47**(6): p. 569-76.
158. van Dam, S., T. Craig, and J.P. de Magalhaes, *GeneFriends: a human RNA-seq-based gene and transcript co-expression database.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1124-32.
159. Szklarczyk, R., et al., *WeGET: predicting new genes for molecular systems by weighted co-expression.* Nucleic Acids Res, 2016. **44**(D1): p. D567-73.
160. Obayashi, T., et al., *COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference.* Nucleic Acids Res, 2019. **47**(D1): p. D55-D62.
161. Li, Y., et al., *CLIC, a tool for expanding biological pathways based on co-expression across thousands of datasets.* PLoS Comput Biol, 2017. **13**(7): p. e1005653.
162. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
163. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
164. Kolesnikov, N., et al., *ArrayExpress update--simplifying data submissions.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1113-6.
165. Chesler, E.J., et al., *WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior.* Nat Neurosci, 2004. **7**(5): p. 485-6.
166. Bastian, F., et al. *Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species*. in *Data Integration in the Life Sciences*. 2008. Berlin, Heidelberg: Springer Berlin Heidelberg.
167. Zhu, Q., et al., *Targeted exploration and analysis of large cross-platform human transcriptomic compendia.* Nat Methods, 2015. **12**(3): p. 211-4, 3 p following 214.
168. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
169. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
170. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2): p. e1002375.
171. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
172. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.
173. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
174. GTEx Consortium, *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
175. Lachmann, A., et al., *Massive mining of publicly available RNA-seq data from human and mouse.* Nat Commun, 2018. **9**(1): p. 1366.
176. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes.* Nat Genet, 2007. **39**(10): p. 1181-6.
177. Bogue, M.A., et al., *Mouse Phenome Database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice.* Nucleic Acids Res, 2018. **46**(D1): p. D843-D850.

178. Stegle, O., et al., *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nat Protoc, 2012. **7**(3): p. 500-7.
179. Wu, D. and G.K. Smyth, *Camera: a competitive gene set test accounting for inter-gene correlation.* Nucleic Acids Res, 2012. **40**(17): p. e133.
180. Mathieu, B., H. Sebastien, and J. Mathieu, *Gephi: An Open Source Software for Exploring and Manipulating Networks.* International AAAI Conference on Weblogs and Social Media, 2009.
181. Fruchterman, T.M.J. and E.M. Reingold, *Graph drawing by force-directed placement.* Software: Practice and Experience, 1991. **21**(11): p. 1129-1164.
182. Vincent, D.B., et al., *Fast unfolding of communities in large networks.* J. Stat. Mech, 2008. **2008**(10): p. P10008.
183. Kim, T.H., et al., *ARID1A Is Essential for Endometrial Function during Early Pregnancy.* PLoS Genet, 2015. **11**(9): p. e1005537.
184. Sergushichev, A., *An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation.* bioRxiv, 2016.
185. Theisen, D.J., et al., *WDFY4 is required for cross-presentation in response to viral and tumor antigens.* Science, 2018. **362**(6415): p. 694-699.
186. Carroll, R.G., et al., *An unexpected link between fatty acid synthase and cholesterol synthesis in proinflammatory macrophage activation.* J Biol Chem, 2018. **293**(15): p. 5509-5521.
187. Bartz, F., et al., *Identification of cholesterol-regulating genes by targeted RNAi screening.* Cell Metab, 2009. **10**(1): p. 63-75.
188. Klootwijk, E.D., et al., *Mistargeting of peroxisomal EHHADH and inherited renal Fanconi's syndrome.* N Engl J Med, 2014. **370**(2): p. 129-38.
189. Carvill, G.L., et al., *Mutations in the GABA Transporter SLC6A1 Cause Epilepsy with Myoclonic-Atonic Seizures.* Am J Hum Genet, 2015. **96**(5): p. 808-15.
190. Calvo, S.E., K.R. Clauser, and V.K. Mootha, *MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins.* Nucleic Acids Res, 2016. **44**(D1): p. D1251-7.
191. Williams, E.G., et al., *Quantifying and Localizing the Mitochondrial Proteome Across Five Tissues in A Mouse Population.* Mol Cell Proteomics, 2018.
192. Stroud, D.A., et al., *Accessory subunits are integral for assembly and function of human mitochondrial complex I.* Nature, 2016. **538**(7623): p. 123-126.
193. Arroyo, J.D., et al., *A Genome-wide CRISPR Death Screen Identifies Genes Essential for Oxidative Phosphorylation.* Cell Metab, 2016. **24**(6): p. 875-885.
194. Floyd, B.J., et al., *Mitochondrial Protein Interaction Mapping Identifies Regulators of Respiratory Chain Function.* Mol Cell, 2016. **63**(4): p. 621-632.
195. Lefebvre-Legendre, L., et al., *Identification of a nuclear gene (FMC1) required for the assembly/stability of yeast mitochondrial F(1)-ATPase in heat stress conditions.* J Biol Chem, 2001. **276**(9): p. 6789-96.
196. Cameron, J.M., et al., *Mutations in iron-sulfur cluster scaffold genes NFU1 and BOLA3 cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes.* Am J Hum Genet, 2011. **89**(4): p. 486-95.
197. Lissanu Deribe, Y., et al., *Mutations in the SWI/SNF complex induce a targetable dependence on oxidative phosphorylation in lung cancer.* Nat Med, 2018. **24**(7): p. 1047-1057.
198. Barabasi, A.L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nat Rev Genet, 2011. **12**(1): p. 56-68.
199. Li, Y., P. Agarwal, and D. Rajagopalan, *A global pathway crosstalk network.* Bioinformatics, 2008. **24**(12): p. 1442-7.
200. Ortega, Z. and J.J. Lucas, *Ubiquitin-proteasome system involvement in Huntington's disease.* Front Mol Neurosci, 2014. **7**: p. 77.
201. Ross, J.M., L. Olson, and G. Coppotelli, *Mitochondrial and Ubiquitin Proteasome System Dysfunction in Ageing and Disease: Two Sides of the Same Coin?* Int J Mol Sci, 2015. **16**(8): p. 19458-76.
202. D'Amico, D., V. Sorrentino, and J. Auwerx, *Cytosolic Proteostasis Networks of the Mitochondrial Stress Response.* Trends Biochem Sci, 2017. **42**(9): p. 712-725.
203. Harrigan, J.A., et al., *Deubiquitylating enzymes and drug discovery: emerging opportunities.* Nat Rev Drug Discov, 2017.
204. Schroeder, E.A., N. Raimundo, and G.S. Shadel, *Epigenetic silencing mediates mitochondria stress-induced longevity.* Cell Metab, 2013. **17**(6): p. 954-64.
205. Tian, Y., et al., *Mitochondrial Stress Induces Chromatin Reorganization to Promote Longevity and UPR(mt).* Cell, 2016. **165**(5): p. 1197-1208.
206. Merkwirth, C., et al., *Two Conserved Histone Demethylases Regulate Mitochondrial Stress-Induced Longevity.* Cell, 2016. **165**(5): p. 1209-1223.
207. Schriml, L.M., et al., *Human Disease Ontology 2018 update: classification, content and workflow expansion.* Nucleic Acids Res, 2019. **47**(D1): p. D955-D962.
208. Pinero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic Acids Res, 2017. **45**(D1): p. D833-D839.
209. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
210. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
211. Evangelou, E. and J.P. Ioannidis, *Meta-analysis methods for genome-wide association studies and beyond.* Nat Rev Genet, 2013. **14**(6): p. 379-89.

212. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nat Genet, 2016. **48**(3): p. 245-252.
213. Gandal, M.J., et al., *Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder.* Science, 2018. **362**(6420): p. eaat8127-17.
214. Gusev, A., et al., *Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights.* Nat Genet, 2018. **50**(4): p. 1-15.
215. Lu, Y., et al., *A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk.* Cancer Res, 2018. **78**(18): p. 5419-5430.
216. Mancuso, N., et al., *Large-scale transcriptome-wide association study identifies new prostate cancer risk regions.* Nat Commun, 2018. **9**(1): p. 1-11.
217. Raj, T., et al., *Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility.* Nat Genet, 2018: p. 1-14.
218. Theriault, S., et al., *A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis.* Nat Commun, 2018. **9**(1).
219. Wu, L., et al., *A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer.* Nat Genet, 2018. **50**(7): p. 968-+.
220. Atkins, I., et al., *Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma.* Cancer Res, 2019. **79**(8): p. 2065-2071.
221. Gamazon, E.R., et al., *Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits.* Nat Genet, 2019. **51**(6): p. 933-+.
222. Giri, A., et al., *Trans-ethnic association study of blood pressure determinants in over 750,000 individuals.* Nat Genet, 2019. **51**(1): p. 51-+.
223. Gusev, A., et al., *A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants.* Nat Genet, 2019. **51**(5): p. 815-+.
224. Hu, Y., et al., *A statistical framework for cross-tissue transcriptome-wide association analysis.* Nat Genet, 2019. **51**(3): p. 568-576.
225. Li, Y.I., et al., *Prioritizing Parkinson's disease genes using population-scale transcriptomic data.* Nat Commun, 2019. **10**(1): p. 994.
226. Mancuso, N., et al., *Probabilistic fine-mapping of transcriptome-wide association studies.* Nat Genet, 2019. **51**(4): p. 675-+.
227. Ratnapriya, R., et al., *Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration.* Nat Genet, 2019. **51**(4): p. 606-+.
228. Wainberg, M., et al., *Opportunities and challenges for transcriptome-wide association studies.* Nat Genet, 2019. **51**(4): p. 592-599.
229. Wu, L., et al., *Identification of Novel Susceptibility Loci and Genes for Prostate Cancer Risk: A Transcriptome-Wide Association Study in over 140,000 European Descendants.* Cancer Res, 2019. **79**(13): p. 3192-3204.
230. Sandoval-Sierra, J.V., et al., *Influence of body weight at young adulthood on the epigenetic clock and lifespan in the BXD murine family.* bioRxiv, 2019: p. 791582.
231. Komorowski, M., et al., *The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care.* Nat Med, 2018. **24**(11): p. 1716-1720.
232. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nat Med, 2019. **25**(1): p. 44-56.
233. Tomasev, N., et al., *A clinically applicable approach to continuous prediction of future acute kidney injury.* Nature, 2019. **572**(7767): p. 116-119.
234. Zullo, J.M., et al., *Regulation of lifespan by neural excitation and REST.* Nature, 2019. **574**(7778): p. 359-364.
235. Nelson, B.R., et al., *A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.* Science, 2016. **351**(6270): p. 271-5.
236. Makarewich, C.A., et al., *The DWORF micropeptide enhances contractility and prevents heart failure in a mouse model of dilated cardiomyopathy.* Elife, 2018. **7**.
237. D'Amico, D., et al., *The RNA-Binding Protein PUM2 Impairs Mitochondrial Dynamics and Mitophagy During Aging.* Mol Cell, 2019. **73**(4): p. 775-787 e10.
238. Lapointe, C.P., et al., *Multi-omics Reveal Specific Targets of the RNA-Binding Protein Puf3p and Its Orchestration of Mitochondrial Biogenesis.* Cell Syst, 2018. **6**(1): p. 125-135 e6.
239. Makarewich, C.A., et al., *MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid beta-Oxidation.* Cell Rep, 2018. **23**(13): p. 3701-3709.
240. Stein, C.S., et al., *Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency.* Cell Rep, 2018. **23**(13): p. 3710-3720 e8.
241. Chugunova, A., et al., *LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism.* Proc Natl Acad Sci U S A, 2019. **116**(11): p. 4940-4945.
242. Busch, J.D., et al., *MitoRibo-Tag Mice Provide a Tool for In Vivo Studies of Mitoribosome Composition.* Cell Rep, 2019. **29**(6): p. 1728-1738 e9.
243. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nat Rev Mol Cell Biol, 2017. **18**(7): p. 437-451.
244. Lee, Y. and D.C. Rio, *Mechanisms and Regulation of Alternative Pre-mRNA Splicing.* Annu Rev Biochem, 2015. **84**: p. 291-323.
245. Romero, J.P., et al., *Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm.* BMC Genomics, 2018. **19**(1): p. 703.

246. Cole, M.B., et al., *Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq.* Cell Syst, 2019. **8**(4): p. 315-328 e8.
247. Vallejos, C.A., et al., *Normalizing single-cell RNA sequencing data: challenges and opportunities.* Nat Methods, 2017. **14**(6): p. 565-571.
248. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nat Rev Genet, 2015. **16**(3): p. 133-45.
249. Wang, J., et al., *Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction.* Mol Cell Proteomics, 2017. **16**(1): p. 121-134.
250. Lapek, J.D., Jr., et al., *Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities.* Nat Biotechnol, 2017. **35**(10): p. 983-989.
251. Kustatscher, G., et al., *The human proteome co-regulation map reveals functional relationships between proteins.* bioRxiv, 2019: p. 582247.
252. Kustatscher, G., et al., *Co-regulation map of the human proteome enables identification of protein functions.* Nat Biotechnol, 2019. **37**(11): p. 1361-1371.
253. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.
254. Franz, M., et al., *GeneMANIA update 2018.* Nucleic Acids Res, 2018. **46**(W1): p. W60-W64.
255. Pagliarini, D.J., et al., *A mitochondrial protein compendium elucidates complex I disease biology.* Cell, 2008. **134**(1): p. 112-23.
256. GTEx Consortium, *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
257. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
258. Joyner, M.J. and N. Paneth, *Promises, promises, and precision medicine.* J Clin Invest, 2019. **129**(3): p. 946-948.

List of abbreviations

AAF	Alternative allele frequency
AD	Alzheimer's disease
AUC	Area under the curve
BCKDH	Branched-chain alpha-keto acid dehydrogenase
BXD	Recombinant inbred mouse cross derived from C57**B**L/6J x **D**BA/2J
CAMERA	Correlation Adjusted MEan RAnk gene set test
CC	Collaborative cross
CD	Chow diet
DO	Diversity outbred
ePheWAS	Expression-based phenome-wide association study
ETC	Electron transport chain
FDR	False discovery rate
G-MAD	Gene-module association determination
G/LOF	Gain- or loss-of-function
G2P	Gene-to-phenotype
GeneRIF	Gene reference into function
GMAS	Gene-module association score
GO	Gene ontology
GRP	Genetic reference population
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
GXE	Gene-by-environmental interaction
HFD	High fat diet
HMDP	Hybrid mouse diversity panel
KEGG	Kyoto encyclopedia of genes and genomes
M-MAD	Module-module association determination
MMAS	Module-module association score
NES	Normalized enrichment score
OCR	Oxygen consumption rate
PEER	Probabilistic estimation of expression residual
PheWAS	Phenome-wide association study
QTL	Quantitative trait locus
RNAi	RNA interference
ROC	Receiver operating characteristic
SNP	Single nucleotide polymorphism
T/PWAS	Transcriptome-/proteome-wide association study
TF	Transcription factor

Curriculum Vitae

Hao Li

EPFL SV IBI-SV LISP
AI 1146 (Bâtiment AI)
Station 15
CH-1015, Lausanne, Switzerland

Office : +41 21 69 31848
Mobile: +41 78 68 29409
Email: hao.li@epfl.ch,
lihaone@gmail.com

ORCID: 0000-0001-5677-3377. **ResearcherID**: K-7001-2017.

Google Scholar: https://scholar.google.com/citations?user=2p0AOZ8AAAAJ

Education:

2015 — 2019 **PhD candidate** in Computational and Quantitative Biology

Laboratory of Integrative Systems Physiology, École Polytechnique Fédérale in Lausanne (EPFL), Switzerland.

Thesis supervisor: Prof. Johan Auwerx, M.D., Ph.D.

2011 — 2014 **Master** in Biochemistry and Molecular Biology

Center for Mitochondrial Biology and Medicine, Xi'an Jiaotong University, China.

Thesis supervisor: Prof. Jiankang Liu, Ph.D.

2007 — 2011 **Bachelor** in Biology

Program of Life Science and Biotechnology, Xi'an Jiaotong University, China.

Research Experiences:

Doctoral Assistant, 2015 — Present

Laboratory of Integrative Systems Physiology, École Polytechnique Fédérale in Lausanne (EPFL), Switzerland.

Research Assistant, 2011 — 2014

Center for Mitochondrial Biology and Medicine, Xi'an Jiaotong University, China.

Lab Assistant, 2008 — 2011

Department of Biological Engineering, School of Life Science and Technology, Xi'an Jiaotong University, China.

Publications:

1. **Li H**, Rukina D, David F, Li TY, Oh C-M, Gao AW, Katsyuba E, Bou Sleiman M, Komljenovic A, Huang Q, Williams RW, Robinson-Rechavi M, Schoonjans K, Morgenthaler S, Auwerx J. Identifying gene function and module connections by the integration of multi-species expression compendia. ***Genome Res***. 2019. doi:10.1101/gr.251983.119.
2. **Li H***, Wang X*, Rukina D, Huang Q, Lin T, Sorrentino V, Zhang H, Bou Sleiman M, Arends D, McDaid A, Luan P, Ziari N, Velázquez-Villegas LA, Gariani K, Kutalik Z, Schoonjans K, Radcliffe RA, Prins P, Morgenthaler S, Williams RW, Auwerx J. An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function. ***Cell Syst***. 2018 Jan 24;6(1):90-102.e4. doi: 10.1016/j.cels.2017.10.016. (* Equal contribution)
3. **Li H**, Auwerx J. Mouse systems genetics as a prelude to precision medicine. ***Trends Genet***. In revision.
4. Zhao L*, **Li H***, Wang Y, Zheng A, Cao L, and Liu J. Autophagy deficiency leads to impaired antioxidant defense via p62-FOXO1/3 axis. ***Oxid Med Cell Longev***. 2019. In press. (* Equal contribution)

5. Yoon H, Spinelli JB, Zaganjor E, Wong SJ, German NJ, Randall EC, Dean A, Clermont A, Paulo JA, Garcia D, **Li H**, Agar NY, Goodyear LJ, Shaw RJ, Gygi SP, Auwerx J, Haigis MC. PHD3 controls energy homeostasis and exercise capacity. ***bioRxiv.*** 2019:781765. doi: 10.1101/781765.

6. Komljenovic A, **Li H**, Sorrentino V, Kutalik Z, Auwerx J, Robinson-Rechavi M. Cross-species functional modules link proteostasis to human normal aging. ***PLoS Comput Biol***. 2019 Jul 3;15(7):e1007162. doi: 10.1371/journal.pcbi.1007162.

7. D'Amico D, Mottis A, Potenza F, Sorrentino V, **Li H**, Romani M, Lemos V, Schoonjans K, Zamboni N, Knott G, Schneider BL, Auwerx J. The RNA-Binding Protein PUM2 Impairs Mitochondrial Dynamics and Mitophagy During Aging. ***Mol Cell***. 2019 Feb 21;73(4):775-787.e10. doi: 10.1016/j.molcel.2018.11.034.

8. Besprozvannaya M, Dickson E, **Li H**, Ginburg KS, Bers DM, Auwerx J, Nunnari J. GRAM domain proteins specialize functionally distinct ER-PM contact sites in human cells. ***Elife***. 2018 Feb 22;7. pii: e31019. doi: 10.7554/eLife.31019.

9. McDaid AF, Joshi PK, Porcu E, Komljenovic A, **Li H**, Sorrentino V, Litovchenko M, Bevers RPJ, Rüeger S, Reymond A, Bochud M, Deplancke B, Williams RW, Robinson-Rechavi M, Paccaud F, Rousson V, Auwerx J, Wilson JF, Kutalik Z. Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. ***Nat Commun***. 2017 Jul 27;8:15842. doi: 10.1038/ncomms15842.

10. Fan W, Waizenegger W, Lin CS, Sorrentino V, He MX, Wall CE, **Li H**, Liddle C, Yu RT, Atkins AR, Auwerx J, Downes M, Evans RM. PPARδ Promotes Running Endurance by Preserving Glucose. ***Cell Metab***. 2017 May 2;25(5):1186-1193.e4. doi: 10.1016/j.cmet.2017.04.006.

11. Zheng A, **Li H**, Xu J, Cao K, Li H, Pu W, Yang Z, Peng Y, Long J, Liu J, Feng Z. Hydroxytyrosol improves mitochondrial function and reduces oxidative stress in the brain of db/db mice: role of AMP-activated protein kinase activation. ***Br J Nutr***. 2015 Jun 14;113(11):1667-76. doi: 10.1017/S0007114515000884.

12. Zheng A, **Li H**, Cao K, Xu J, Zou X, Li Y, Chen C, Liu J, Feng Z. Maternal hydroxytyrosol administration improves neurogenesis and cognitive function in prenatally stressed offspring. ***J Nutr Biochem***. 2015 Feb;26(2):190-9. doi: 10.1016/j.jnutbio.2014.10.006.

13. Wang X, **Li H**, Zheng A, Yang L, Liu J, Chen C, Tang Y, Zou X, Li Y, Long J, Liu J, Zhang Y, Feng Z. Mitochondrial dysfunction-associated OPA1 cleavage contributes to muscle degeneration: preventative effect of hydroxytyrosol acetate. ***Cell Death Dis***. 2014 Nov 13;5:e1521. doi: 10.1038/cddis.2014.473.

14. Cao K, Zheng A, Xu J, **Li H**, Liu J, Peng Y, Long J, Zou X, Li Y, Chen C, Liu J, Feng Z. AMPK activation prevents prenatal stress-induced cognitive impairment: Modulation of mitochondrial content and oxidative stress. ***Free Radic Biol Med***. 2014 Oct;75:156-66. doi: 10.1016/j.freeradbiomed.2014.07.029.

15. Zou X, Yan C, Shi Y, Cao K, Xu J, Wang X, Chen C, Luo C, Li Y, Gao J, Pang W, Zhao J, Zhao F, **Li H**, Zheng A, Sun W, Long J, Szeto IM, Zhao Y, Dong Z, Zhang P, Wang J, Lu W, Zhang Y, Liu J, Feng Z. Mitochondrial dysfunction in obesity-associated nonalcoholic fatty liver disease: the protective effects of pomegranate with its active component punicalagin. ***Antioxid Redox Signal***. 2014 Oct 10;21(11):1557-70. doi: 10.1089/ars.2013.5538.

16. Zhao L, Zou X, Feng Z, Luo C, Liu J, **Li H**, Chang L, Wang H, Li Y, Long J, Gao F, Liu J. Evidence for association of mitochondrial metabolism alteration with lipid accumulation in aging rats. ***Exp Gerontol***. 2014 Aug;56:3-12. doi: 10.1016/j.exger.2014.02.001.

17. Cao K, Xu J, Zou X, Li Y, Chen C, Zheng A, **Li H**, Li H, Szeto IM, Shi Y, Long J, Liu J, Feng Z. Hydroxytyrosol prevents diet-induced metabolic syndrome and attenuates mitochondrial abnormalities in obese mice. ***Free Radic Biol Med***. 2014 Feb;67:396-407. doi: 10.1016/j.freeradbiomed.2013.11.029.

18. Zheng A, **Li H**, Wang X, Feng Z, Xu J, Cao K, Zhou B, Wu J, Liu J. Anticancer effect of a curcumin derivative B63: ROS production and mitochondrial dysfunction. ***Curr Cancer Drug Targets***. 2014;14(2):156-66.