UNIL | Université de Lausanne
Faculté de biologie
et de médecine

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Mémoire de Maîtrise en médecine

Cholesterol Biosynthesis: A Systems Genetics Approach to Identify Novel Regulatory Genes

Etudiant
Peter Virginie

Tuteur
Prof. Auwerx Johan
Institut de Bioingénierie, EPFL

Co-tuteur
Dr Williams Evan
Institut de Bioingénierie, EPFL

Expert
Prof. Lluis Fajas Coll
Centre intégratif de génomique, UNIL

Lausanne, 08.12.2016

Abstract

Cholesterol is a lipid that is essential in membrane structure and function and that is a precursor for bile acid and steroid hormone synthesis. In humans, elevated plasma cholesterol levels are an important cardiovascular risk factor, as they predispose to the development of atherosclerosis. Understanding cholesterol metabolism is therefore the focus of extensive research. Yet, if there is evidence for substantial genetic contributions to variations in cholesterol levels, the underlying functional genetic architecture remains largely unknown and current approaches have yielded results of little predictive value regarding future disease occurrence. It is now becoming commonly accepted that complex traits, like plasma cholesterol levels, are the product of complex molecular networks, modulated by sets of genetic loci and environmental factors. To address these complexities, it is necessary to study cholesterol metabolism as a whole system and to get an understanding of the interacting effects of its individual parts. Here, we propose a systems genetics approach of the cholesterol biosynthetic pathway in genetically diverse BXD-type recombinant mouse lines. For this, we have combined large-scale genomic, transcriptomic and phenomic results and integrated it to reconstruct the co-expression network of cholesterol biosynthetic genes. We identified a set of new genes, previously not known to be involved in cholesterol metabolism. In particular, *Echdc1* is one of the promising candidates and it was shown that variations in its sequence and transcript abundance are significantly associated to plasma cholesterol levels. We believe that examining the role of these genes will provide new insights on the regulation of cholesterol biosynthesis and on the basis of cholesterol level variability among individuals in a genetically admixed population. Translation to humans may eventually have clinical implications in prediction of disease risk and response to treatment and drug development.

Acknowledgements

I would like to thank Prof. Johan Auwerx for giving me the opportunity to work in his laboratory and for supervising this master project, Dr. Evan Williams for introducing me to the project, for sharing his bioinformatic skills and scientific insight with me - It was a pleasure being your master student, the Auwerx laboratory for welcoming me, and for always being very helpful - it was great working with you.

Contents

Abstract 3

Acknowledgements 4

Contents 5

1. Introduction 7

1.1 ***The cholesterol molecule and its metabolism***
1.2 ***Cholesterol level variations in a natural population***
1.3 ***Research in humans and in mice***
1.4 ***Computational methods***
1.5 **Our study**

2. Results 10

1.1 ***Cholesterol biosynthesis genes cluster in a network of highly correlating genes***
1.2 ***Plasma cholesterol correlates with the cholesterol biosynthesis gene regulatory network***
1.3 ***Candidate regulatory genes can be mapped to the cholesterol biosynthesis network***
1.4 ***Six candidate genes also correlate to the cholesterol biosynthesis gene network in human data sets***
1.5 ***Variation at the Echdc1 locus is associated to differences in cholesterol levels***
1.6 ***Human ECHDC1 gene variants may be associated with breast cancer risk***

3. Discussion 13

4. Conclusion 14

5. Methods 15

Tables & Figures 16

Supplementary figures 25

References 27

1. Introduction

1.1 *The cholesterol molecule and its metabolism*

Cholesterol is well known for its central role in the pathogenesis of cardiac vascular disease. Yet, it is firstly an essential molecule in many animals, including humans, as a universal component of the cell membrane. It is also the precursor for steroid hormones, vitamin D, oxysterols and bile acids, and is thus a key player in animal metabolism. Moreover, recent research has shown the implication of cholesterol in inflammation and immunity [1, 2], cancer [3], and Alzheimer's disease [4]. In the body, cholesterol can be derived both from diet and *de novo* synthesis (Figure 1A). In humans, most of it is synthesized *de novo* and a small fraction originates from the diet [5], whereas other animals, such as *Caenorhabditis elegans* and *Drosophila melanogaster* are sterol auxtrotrophs. With its unique structure of 27 carbon atoms arranged in 4 fused rings, the cholesterol molecule is produced through a complex biosynthetic pathway of sequentially acting enzymes - the mevalonate pathway. This pathway leads to the formation of cholesterol by deriving all carbons from acetate and through the formation of four major intermediates: mevalonate, farnesyl pyrophosphate, squalene, and lanosterol. HMG-CoA-reductase is the rate-controlling enzyme of the pathway, which catalyzes the conversion of HMG-CoA to mevalonate, a step that is targeted by cholesterol lowering drugs called statins. Cholesterol obtained from food is first absorbed in the gut and transported to the liver, which then mediates its delivery to the whole body. As cholesterol and other lipids are insoluble, they are transported through the circulation as complexes with proteins: lipoproteins. Among them, low density lipoprotein (LDL) is the main lipoprotein that delivers cholesterol to peripheral cells, whereas high density lipoprotein (HDL) is responsible for reverse cholesterol transport, the process by which cholesterol is transported from peripheral tissues to the liver and secreted into the bile (Figure 1B). Bile cholesterol ends up in the small intestine where it can either be reabsorbed with bile salts - as part of the enterohepatic cycle - or excreted into faeces. The processes of cellular cholesterol synthesis, absorption and trafficking are tightly regulated at the cellular level [6]. There are two main transcriptional regulatory systems keeping cellular cholesterol homeostasis: sterol regulatory element-binding proteins (SREBPs) and liver X-activated receptors (LXRs). These transcription factors aim at keeping intracellular cholesterol levels steady by balancing contributions of endogenous cholesterol synthesis and endocytosis of LDL cholesterol, and by regulating the formation of cholesterol esters for storage of excess cholesterol. SREBPs are a family of transcription factors that activate the expression of gene products dedicated to the synthesis and uptake of cholesterol [7]. On the other hand, liver X-activated receptors (LXRs), are oxysterol-gated nuclear receptors that favor reverse cholesterol transport [8].

1.2 *Cholesterol level variations in a natural population*

Cholesterol levels vary continuously over a range of distribution in natural populations; a process that is controlled by several genetic and environmental factors. The prime exogenous factors well known to change cholesterol levels include diet, exercise and drugs, with potentially many more factors contributing. On the other hand, genetic makeup also changes an individuals' propensity in cholesterol synthesis, absorption and output, resulting in inter-individual variations, and hence also making people more or less susceptible to disease. Indeed, not all individuals exposed to the same exogenous factors/stressors will develop health issues, revealing the importance of genetic factors [9], which are estimated to contribute to roughly 50% of the variation in cholesterol level in humans [10]. However, if there is evidence for substantial genetic contributions to the variation of cholesterol and lipoprotein level, the underlying genetic machinery remains largely unsolved [11]. Due to the important implications of plasma lipoprotein levels in cardiovascular diseases, elucidating the genetic factors implicated has been and remains the focus of extensive research. Finding out these factors will undoubtly contribute to better prevention and management of disease linked with altered cholesterol levels [9, 12, 13].

1.3 *Research in humans and in mice*

In the research for the genetic factors underlying complex traits, large efforts have been invested to discover the individual genes accounting for inter-individual variability. Genome-wide association studies (GWAS) have made it possible to identify hundreds of genetic loci associated to common traits in humans. In the field of plasma lipids and lipoproteins, these have revealed more than 50 variants underlying dyslipidemias, including previously known and entirely novel SNPs [14-16]. Although, these results have provided valuable insights into the current understanding of complex traits through identifying new-involved genes and pathways, they also brought about disappointment and questioning, as the effects explained by these genes represent only a limited portion of the heritable component of any complex trait, thus providing little predictive value regarding future disease occurrence [17]. The question arose hence as to why and which aspects have been missed through this approach [18]. Complex traits, such as cholesterol levels, typically result from the action of many genes, as well as from gene-gene and gene-environment interactions [19]. Accordingly, effects of genetic variants depend on the genetic and environmental context in which they operate [20], and thus show differences in expressivity and penetrance by virtue of interaction effects. This concept implies that complex traits cannot be explained solely by the sum of the actions of single genes [21]. This could in part explain why approaches aiming to identify single genes, like through GWAS, have failed to capture the complete range of genetic variability, as context-dependent effects, such as interactions, were overlooked. It is now becoming generally accepted that many traits need to be studied as the product of a whole system of larger scale, in which all the components of this system should be analyzed simultaneously [22]. The systems approach is based on the concept that the whole is greater than the sum of its parts, and thus, looking at systems as a whole will give a new perspective to the understanding of biological phenomena [23]. This approach provides us with a vision that wouldn't be possible otherwise and becomes even more attractive when used in complementarity with the study of gene function at the level of individual pathways [24]. For the study of cholesterol metabolism, mice have remained a popular model organism, for practical reasons [25]. The ease with which their genome can be manipulated and their environment can be controlled is inconceivable in humans. Furthermore, the information resources available from mouse genetic resources are an additional benefit. Mouse models have already proven useful for identifying new genetic determinants of common traits, for the study in vivo of relevant pathways and for validating the data obtained from human genetic studies [26]. Yet, it is important to note that in the mouse circulating cholesterol is largely found in HDL, differentiating them from humans and rendering the animals resistant to the development of atherosclerosis [27]. In this respect, genetically engineered mouse models have been generated that are prone to develop atherosclerosis, such as apoE deficient (*Apoe*-/-) or LDL receptor deficient (*Ldlr*-/-) mice, making them more apt to study cardiovascular disease [28]. Despite that these mice strains developed atherosclerosis, they were not mimicking the full complexity of human populations in relation to cholesterol metabolism. Aiming a systems approach to understand complex traits, it was a challenge to find an experimental design that allows the analysis of many components simultaneously (genes, transcripts, phenotypes,…), and that comprises multifactorial perturbations, namely genetic variation and different environmental conditions [29]. Among the best models to this purpose are murine genetic reference populations (GRPs) and, more specifically, recombinant inbred (RI) strains [30]. GRPs are panels of strains with unique and fixed genomes that have been derived from multiple inbreeding of genetically diverse parental strains. As such, they model aspects of natural populations in showing genetic and phenotypic diversity, while keeping the advantages of model organisms. In particular, two features make them especially attractive for the study of complex traits [26]. First, RI strains are renewable and thus allow replicate studies: genotype once, phenotype many times. Second, the environment can be rigorously controlled in reference populations, enabling to discriminate between the effects of genes, environment and gene by environment interactions (GxE), much better than in a natural population. Moreover, as GRPs have fixed genotypes, publicly available databases from different centers can be shared, offering a richness of data and greatly increasing statistical robustness. The BXD family is one of the most important GRP with more than 160 strains derived from crosses between C57BL/6J and DBA/2J [31]. A great amount of data, ranging from genotypes over molecular (e.g. transcriptome) to clinical phenotypes has already been accumulated across the BXD strains by different independent groups and is publicly available on the GeneNetwork data repository (www.genenetwork.org) [31, 32].

1.4 *Computational methods*

As new experimental approaches now allow testing of multiple components simultaneously and across multiple genetic backgrounds, computational approaches evolved to assemble more complex and integrated systems. Among these, the network-based approach is of particular interest for the analysis of complex traits such as cholesterol levels [21]. Networks organize large-scale data by modeling molecules as nodes (for example genes or gene products) and the relationship between them as edges, thus enabling the representation of all components of a system simultaneously along with extensive interactions among them. Molecular regulatory networks and pathways underlying complex traits lend themselves to evaluation by such advanced approaches. Eventually, new causative genes and mutations can be implicated by mapping their involvement in a network with known susceptibility loci. In this way, networks provide a convenient framework for integrating a diversity of data and visualizing the complexities of a biological system [33]. Computational approaches enhance traditional molecular techniques, by moving beyond examining genes and their downstream products one at a time, and integrating the entire complexity of the data. This view can only be obtained by interrogating the system as a whole.

1.5 *Our study*

In the present study, we sought to identify new regulatory genes of the cholesterol biosynthesis pathway. To this purpose, we propose a systems genetics approach of the cholesterol biosynthesis pathway based on whole genome, transcriptome, proteome, metabolome, and phenome data from 84 cohorts of mice, derived from 42 different BXD strains on two dietary regimens – chow diet (CD) or high fat diet (HFD) – with 3 to 5 replicate animals for each cohort. We use computational and network analysis to identify new genes involved in this pathway, as well as the relations between them and the already-known factors.

2. Results

1.1 *Cholesterol biosynthesis genes cluster in a network of highly correlating genes*

A co-expression network was constructed based on the liver mRNA levels of a hundred cholesterol related genes. These were selected from the literature and previous correlation analyses, as any gene known to be involved in cholesterol metabolism [34]. From this broad, prior-based network appears a sub-network of strongly interconnected nodes, which is highly enriched in cholesterol biosynthetic genes (Supplementary Figure 1). As correlating genes are those which show a similar pattern of differential expression among the different BXD strains, it suggests that, to some extent, they are responding to the same genetic variability and thus to a common regulatory mechanism. It is therefore not surprising to find cholesterol biosynthetic genes grouped, as they are known and expected to be tightly regulated at the transcriptional level [35]. Importantly, it validates our method for finding sets of co-expressed genes. Accordingly, we thereafter have focused on this sub-network (Figure 2A), using it as a confirmed basis to which novel gene involvements could be mapped. We found also that some genes of the beta-oxidation pathway (*Acadvl, Cpt1, Cpt2, Ehhadh, Hadha, Hadhb, Acadm*) also add to the network, correlating strongly and negatively with cholesterol biosynthetic genes, whereas genes related to lipoprotein assembly and transport, like those encoding for the apolipoproteins stayed unrelated. Then, we repeated the analysis using the HFD cohort data and reproduced a similar sub-network (Figure 2B, Suppl. Figure 1B), further confirming the method and indicating the robustness of the cholesterol biosynthetic pathway to changes implied by the HFD. Interestingly, we observed an even higher count of significant correlations within the genes of the cholesterol subnetwork in the HFD, with an average of 3.1 significant ($p < 1E-10$) edges per node, against 1.2 in the CD, suggesting biosynthetic genes are more tightly regulated when dietary cholesterol is increased.

1.2 *Plasma cholesterol correlates with the cholesterol biosynthesis gene regulatory network*

As phenotypes were integrated to the cholesterol biosynthetic network, we found that cholesterol, glucose, liver weight, triglycerides, body weight loss at fasting, lactic acid, and iron levels are closely linked to the system (Figure 3). Most of these phenotypes are common manifestations of or are linked to the metabolic syndrome, but our data suggest that they are linked to cholesterol biosynthesis at the transcriptional level. Of note, LDL and HDL cholesterol levels did not manifest strong correlations. Notably, the link between iron and cholesterol levels is intriguing, and has also been described before in literature [36], where hepatic iron was shown to enhance cholesterol synthesis.

1.3 *Candidate regulatory genes can be mapped to the cholesterol biosynthesis network*

We proposed that if our approach is useful for re-discovering genes participating in cholesterol biosynthesis, then it should be valid for identifying novel gene involvements. Effectively, a few genes previously not known to be involved in cholesterol biosynthesis correlate strongly with the known set of genes and cluster along. Looking for novel regulators of the cholesterol biosynthetic pathway, we performed correlation analyses, using principal component analysis (PCA) to condense 6 core genes of the network into one, more robust, variable (Figure 4). We found a series of genes showing a similar pattern of expression as cholesterol biosynthetic genes in different strains throughout 6 independent datasets (Figure 4A,B). These genes demonstrate strong correlation to the first principal components across the different datasets, including the BXD CD and HFD cohorts [37] (Figure 4C), yet they have not been reported to be involved in or linked to this pathway previously. The main features of the 13 top candidate genes are summarized in Table 1. All of these genes being co-expressed with the cholesterol biosynthetic genes, it remains to be determined whether they are responding or acting upon the network regulation system. This question still has to be addressed experimentally, for example by looking at the situation where the gene in question has been knocked-down/-out in cells or in vivo. In this respect, we found that knock-outs (KO) already exist for some of the candidates. One of the genes, *Acsl5*, has been previously examined in *Acsl5*-KO mice, which showed a strong decrease in circulating cholesterol levels (p=2.55E-15) [38]. In humans, a variant in this gene was also reported to associate with better response to diet, lower total cholesterol, LDL and triglycerides [39]. *Pstpip2*-KO mice were shown to associate to multiple skeletal morphologic anomalies, as well as to changes in hematopoiesis, adipose tissue, and behavior. No significant phenotypic traits were found in *Paox*-KO mice. Regarding the other candidates, we can formulate hypotheses based on their response upon environmental changes such as addition of dietary cholesterol or statin treatment.

1.4 *Six candidate genes also correlate to the cholesterol biosynthesis gene network in human data sets*

In order to further validate our results, we tested the candidate genes for correlation to the cholesterol biosynthesis gene network in two independent human databases: The Human Liver Cohort (HLC) [40] (Figure 5A), and The Genotype-Tissue Expression (GTEx) project [41, 42] (Figure 5B). Analysis of the transcriptional data from these two sources revealed that six of the candidate genes do correlate strongly to the principal components computed from the same core cholesterol biosynthetic genes as described previously. *Echdc1*, *Rdh11*, *Mmab*, and *C14Orf* showed significant correlation values (p<0.001) in both datasets, while *Paox* and *Gck* had significant values in a single dataset.

1.5 *Variation at the Echdc1 locus is associated to differences in cholesterol levels*

We further examined the first candidate gene, *Echdc1*, which is coding for ethylmalonyl-CoA decarboxylase: a protein known to have a role in lipid metabolism but whose function is still unclear [43]. We found that sequence variants in the 5Mb region containing *Echdc1* (Figure 6A) are significantly linked to cholesterol levels in two independent datasets, a CD [37] and a female cohort [31], using a PheWAS approach recently developed (Li et al.) (Figure 6B). To better understand how this gene relates to cholesterol metabolism, we first searched for its conservation in species and found it is conserved in vertebrates (Figure 6C). We then analyzed open-access data on differential tissue expression and found from the GTEx project that *Echdc1* mRNA level is highest in adipose, lung, bladder, and breast tissue (Figure 7A). Human tissue protein staining is strongest in hepatocytes, adrenal cortex, renal tubules and Leydig cells based on data from the Human Protein Atlas [44]. A more in-depth analysis of the phenotypes linked to *Echdc1* in the BXD CD dataset revealed that blood glucose, liver weight in grams and as percentage of the total body weight, total plasma cholesterol, sub-cutaneous white adipose tissue (scWAT), and aspartate transaminase (ASAT) levels are the most correlated traits (Figure 7B). This implies that changes in *Echdc1* expression are associated, not only to cholesterol levels, but also to phenotypes such as liver weight and ASAT levels - signatures of liver disease - and could also potentially be inducing them. When looking at the gene's regulatory response to environmental changes, we found that it is up-regulated when mice are fed a HFD (Figure 7C). When mice are fed with a high fat diet depleted of cholesterol, *Echdc1* expression is induced, whereas its levels are decreased when cholesterol is added to the diet [45] (Figure 7D). Moreover, in human primary hepatocytes treatment by statins up-regulates *Echdc1 levels* [46].

1.6 *Human ECHDC1 gene variants may be associated with breast cancer risk*

In 2008, a genome-wide association study (GWAS) in an Ashkenazi Jew population provided evidence for a breast cancer risk locus at 6q22.33, a region including *ECHDC1* and RNF146 [47]. This study suggested that ~7% of breast cancers in this population could be attributable to this risk factor, due to a relatively low risk ratio (1.5) but a high frequency of carriers (23%).

However, this association was not reproduced in subsequent breast cancer GWAS, which include one study in Chinese Women [48] and another study in African-American Women [49]. Nonetheless, *ECHDC1* was mentioned in another study aiming at exploring the effects of the mucin 1 oncoprotein, which is aberrantly overexpressed in human breast cancer cells [50]. This protein was shown to induce the expression of a set of 38 genes involved in cholesterol and fatty acid metabolism, including *ECHDC1*, which shapes a network that converges to the SREBP-1 gene. This result supports our observations of a role of *ECHDC1* in lipid metabolism, and it strengthens our hypothesis of a possible participation of *ECHDC1* in breast cancer pathogenesis. In this context, there is also epidemiological evidence that statins can offer protection against breast cancer [51], but not against all cancer types [52]. Furthermore, in a recent study a metabolite of cholesterol, 27-hydroxycholesterol, was shown to link hypercholesterolemia and breast cancer pathophysiology [53].

3. Discussion

We set out to find new regulatory genes of the cholesterol biosynthesis pathway by means of network analysis. Genes of the cholesterol biosynthesis pathway were shown to associate into a co-expression network of highly correlating nodes. This was consistent with the prior knowledge that these genes are subject to fine regulation at the transcriptional level, validating our approach. We searched for phenotypes linked to the cholesterol biosynthesis pathway and found traits related to the metabolic syndrome. In fact, plasma glucose levels, triglycerides, cholesterol, liver weight, plasma iron and lactic acid levels all correlate to cholesterol biosynthesis gene transcription. With the same general approach, we found several new co-expressed genes showing strong similarity to the known cholesterol biosynthetic genes regarding their pattern of expression in genetically diverse strains. Different situations can lead to co-expression of genes: they can share a transcriptional regulator, directly regulate one another's expression, be in genomic proximity, respond to a product of one gene, and more. We identified a set of 13 novel genes co-expressed with the known cholesterol biosynthesis genes, among which some will eventually prove to have a regulatory action on cholesterol metabolism.. One of the candidate genes that came out of our analysis is *Acls5*. As we subsequently looked for open-access data on the candidate genes, we found that *Acsl5*-KO mice show changes in circulating cholesterol levels as most striking phenotype. This biologically confirms the phenomenon that we bioinformatically anticipated and thus strongly supports our results. *Echdc1* is another promising candidate coming out of this analysis. We have shown that variations in *Echdc1* sequence and transcript abundance are significantly associated to plasma cholesterol levels, reinforcing the idea that *Echdc1* is a novel causative gene in cholesterol level changes. Liver weight and ASAT levels - signatures of liver disease - also associated strongly to *Echdc1* mRNA levels implying *Echdc1* is linked to changes in hepatocyte physiology as well. In GWAS, this same gene has been identified as a significant risk factor of breast cancer in an Ashkenazi Jewish populations and linked with a lipid metabolic gene network involved in breast cancer. These findings further highlight the need for additional studies on *Echdc1* and it's involvement in cholesterol metabolism and breast cancer. The other 11 candidates also deserve further attention, and in particular, their validation through in-depth biological studies. As a first step, we propose to examine the transcriptional response of the genes to different perturbations (pathway inhibitors and knock-downs of regulatory genes) in hepatocyte cell lines. In a second step, phenotypic and metabolic characterization of model animals with loss-of-function of these genes could constitute a robust reference before further investigations.

4. Conclusion

Our analysis has led to the identification of several genes with potent regulatory functions on the rate of cholesterol biosynthesis. At this stage, our findings are just enough to formulate a hypothesis. However, we propose to pursue this research by a functional analysis and validation of the involvement for one or more candidate genes. We believe it may lead us one step further into understanding the genetic basis of cholesterol-related traits in a genetically diverse population. Moreover, the biological pathways that will be uncovered as a result of these studies might give rise to a better understanding of the basic mechanisms of cholesterol metabolism. Ultimately, elucidating these genetic factors will offer insights into both treatment and prevention of diseases ensuing high cholesterol levels.

5. Methods

GeneNetwork (www.genenetwork.org) was used to retrieve all the data, i.e. all BXD genotypic, transcriptomic and phenotypic data. Network Graphs were constructed in R from BXD liver gene expression and phenotype results, using the custom package imsbInfer on Github (https://github.com/wolski/imsbInfer). The connectivity between nodes was calculated as the spearman rank correlation, with a threshold at $p < 0.001$ (and $p < 1E\text{-}10$) for significant connections. Selection of phenotypes was done by plotting the cholesterol biosynthesis network with all phenotypes, then taking phenotypes with > 1 significant connection to the cholesterol network. The research for novel genes was done using GeneNetwork by PCA and correlation analyses in the following manner. We computed the values for the first principal component (factor 1), obtained from the PCA of 6 core genes of the cholesterol biosynthesis network (*Idi1, Nsdhl, Lss, Mvk, Fdps, Dhcr7*). Then we calculated spearman rank correlations between factor 1 and all liver genes and kept the 1000 first genes. We repeated the process in 6 different databases (liver mRNA: EPFL/LISP CD&HFD [37]; UCLA/BHF2 (Apoe Null) F&M [54]; UNC F&M [55], MDP JAX [56]), so that in total we used 2 female, 2 male, 1 CD, and 1 HFD, and 1 other independent cohorts from three different genetic mouse model systems (BXD, MDP, BHF2). Then we selected correlating genes with highest incidence among correlation results in all datasets. The same approach and the same genes were used to compute the first principal component in two human datasets (GTEx [41, 42] and HLC [40]). Student's t-test was used to calculate significance of the difference between two groups. Pearson correlation was used to calculate r and p correlation values for gene expression and phenotype. Geneious software was used to draw the Echdc1 phylogenetic tree [57]. Prism software was used to draw correlation and mRNA expression figures (GraphPad Prism 6.0, GraphPad Software, San Diego, CA, USA).

Tables & Figures

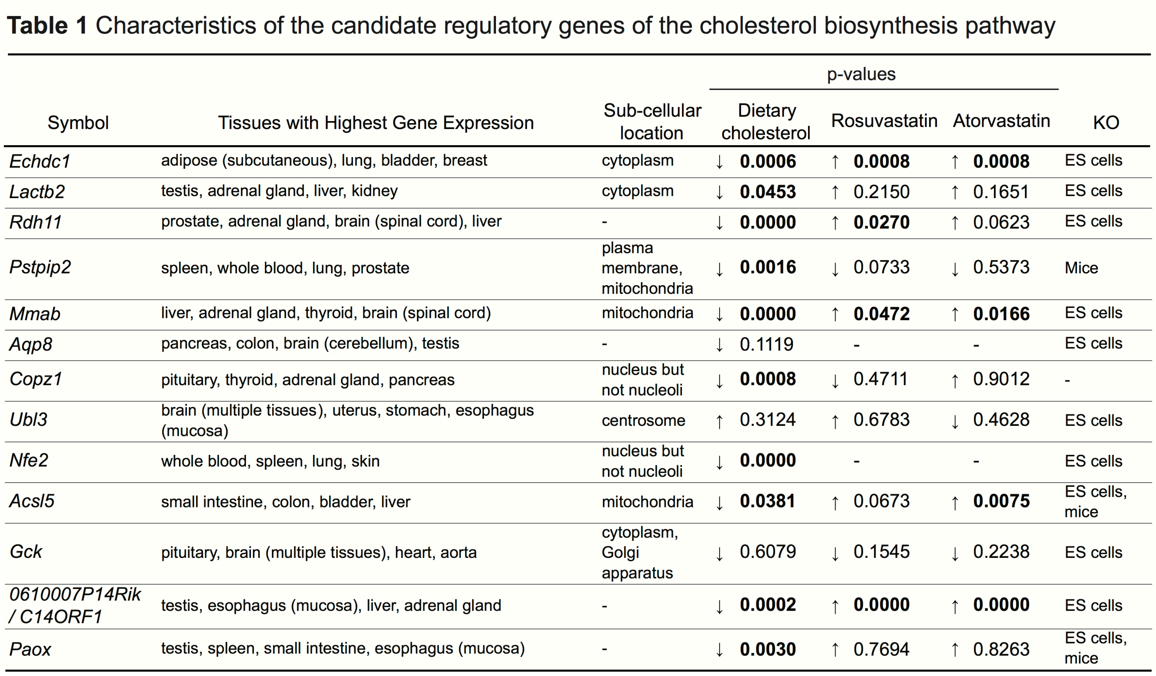


Table 1 Characteristics of the candidate regulatory genes of the cholesterol biosynthesis pathway

Symbol	Tissues with Highest Gene Expression	Sub-cellular location	p-values Dietary cholesterol	 Rosuvastatin	 Atorvastatin	KO
Echdc1	adipose (subcutaneous), lung, bladder, breast	cytoplasm	↓ **0.0006**	↑ **0.0008**	↑ **0.0008**	ES cells
Lactb2	testis, adrenal gland, liver, kidney	cytoplasm	↓ **0.0453**	↑ 0.2150	↑ 0.1651	ES cells
Rdh11	prostate, adrenal gland, brain (spinal cord), liver	-	↓ **0.0000**	↑ **0.0270**	↑ 0.0623	ES cells
Pstpip2	spleen, whole blood, lung, prostate	plasma membrane, mitochondria	↓ **0.0016**	↓ 0.0733	↓ 0.5373	Mice
Mmab	liver, adrenal gland, thyroid, brain (spinal cord)	mitochondria	↓ **0.0000**	↑ **0.0472**	↑ **0.0166**	ES cells
Aqp8	pancreas, colon, brain (cerebellum), testis	-	↓ 0.1119	-	-	ES cells
Copz1	pituitary, thyroid, adrenal gland, pancreas	nucleus but not nucleoli	↓ **0.0008**	↓ 0.4711	↑ 0.9012	-
Ubl3	brain (multiple tissues), uterus, stomach, esophagus (mucosa)	centrosome	↑ 0.3124	↑ 0.6783	↓ 0.4628	ES cells
Nfe2	whole blood, spleen, lung, skin	nucleus but not nucleoli	↓ **0.0000**	-	-	ES cells
Acsl5	small intestine, colon, bladder, liver	mitochondria	↓ **0.0381**	↑ 0.0673	↑ **0.0075**	ES cells, mice
Gck	pituitary, brain (multiple tissues), heart, aorta	cytoplasm, Golgi apparatus	↓ 0.6079	↓ 0.1545	↓ 0.2238	ES cells
0610007P14Rik / C14ORF1	testis, esophagus (mucosa), liver, adrenal gland	-	↓ **0.0002**	↑ **0.0000**	↑ **0.0000**	ES cells
Paox	testis, spleen, small intestine, esophagus (mucosa)	-	↓ **0.0030**	↑ 0.7694	↑ 0.8263	ES cells, mice

Table 1: Key characteristics of candidate regulatory genes of the cholesterol biosynthesis pathway. The table summarizes four tissues with highest gene expression as GTEx [41, 42] and sub-cellular location of the protein products for each candidate, as reported by the Human Protein Atlas [44]. Response to exogenous perturbations was examined by looking for transcriptional changes in mice hepatocytes upon addition of dietary cholesterol [45] and after human hepatocyte treatment by cholesterol lowering drugs [46]: Rosuvastatin and Atorvastatin (arrows: ↓ lower, ↑ higher mRNA levels, p-values < 0.05 are bold). Knock-out (KO) availability in mice is according to the IMPC [38]. Genes abbreviations, *Echdc1*: enoyl Coenzyme A hydratase domain containing 1, *Lactb2*: lactamase, beta 2, *Rdh11*: retinol dehydrogenase 11, *Pstpip2*: proline-serine-threonine phosphatase-interacting protein 2, *Mmab*: methylmalonic aciduria type B protein, *Aqp8*: aquaporin 8, *Copz1*: coatomer protein complex, subunit zeta 1, *Ubl3*: ubiquitin-like 3, *Nfe2*: nuclear factor, erythroid derived 2, *Acsl5*: acyl-CoA synthetase long-chain family member 5, *Gck*: glucokinase, *Paox*: polyamine oxidase (exo-N4-amino).

Figure 1

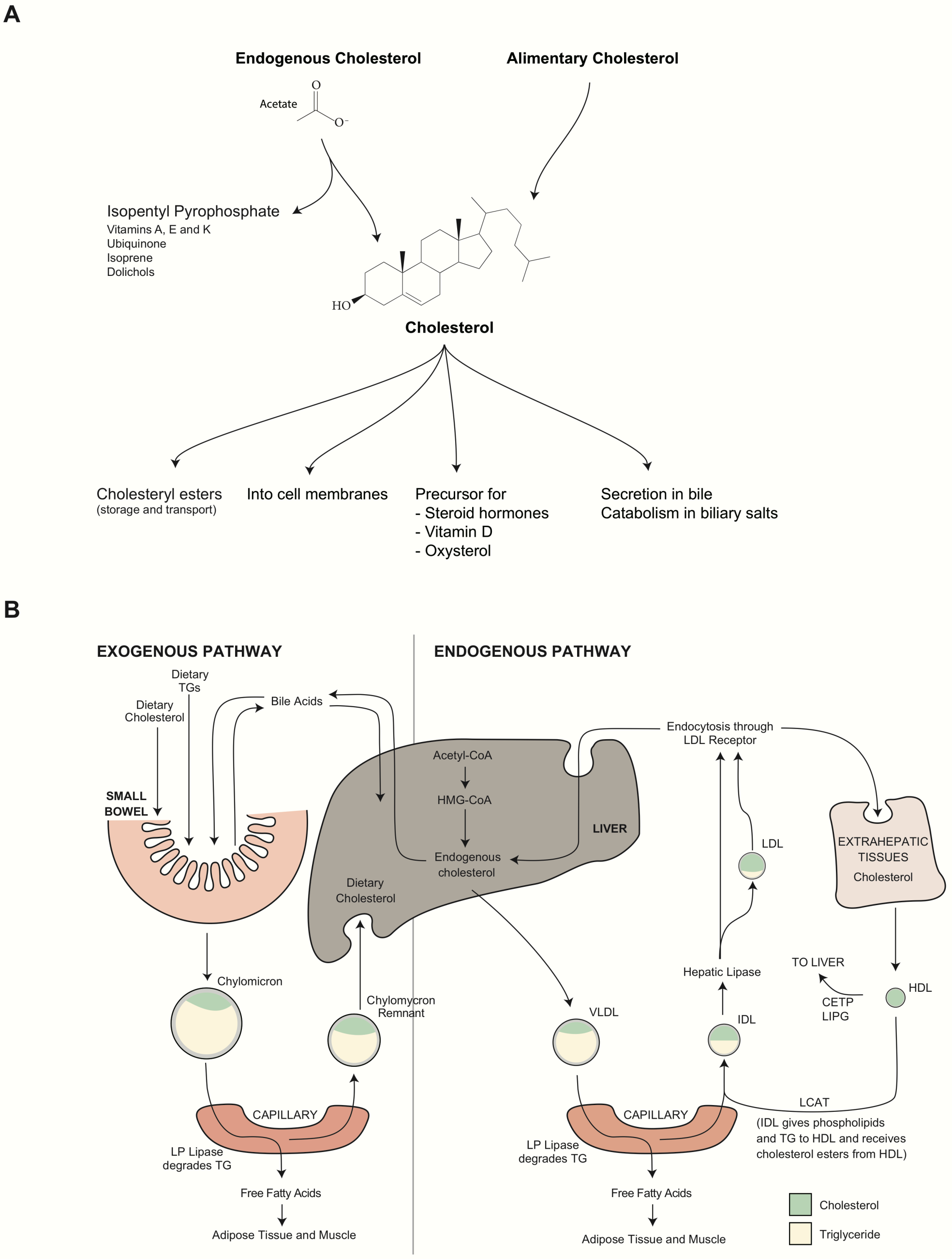

A
Endogenous Cholesterol
Alimentary Cholesterol
Acetate
O
O⁻
Isopentyl Pyrophosphate
Vitamins A, E and K
Ubiquinone
Isoprene
Dolichols
HO
Cholesterol
Cholesteryl esters
(storage and transport)
Into cell membranes
Precursor for
- Steroid hormones
- Vitamin D
- Oxysterol
Secretion in bile
Catabolism in biliary salts
B
EXOGENOUS PATHWAY
ENDOGENOUS PATHWAY
Dietary TGs
Dietary Cholesterol
Bile Acids
SMALL BOWEL
Acetyl-CoA
HMG-CoA
Endogenous cholesterol
LIVER
Dietary Cholesterol
Endocytosis through LDL Receptor
EXTRAHEPATIC TISSUES
Cholesterol
LDL
Chylomicron
Chylomycron Remnant
VLDL
Hepatic Lipase
IDL
TO LIVER
CETP
LIPG
HDL
CAPILLARY
LP Lipase degrades TG
Free Fatty Acids
Adipose Tissue and Muscle
CAPILLARY
LP Lipase degrades TG
Free Fatty Acids
Adipose Tissue and Muscle
LCAT
(IDL gives phospholipids and TG to HDL and receives cholesterol esters from HDL)
Cholesterol
Triglyceride

Figure 1: (A) Schematic view of cholesterol equilibrium. Body cholesterol is derived from both an exogenous (diet) and an endogenous (*de novo* synthesis) source. Endogenous cholesterol is produced through multiple reductive polymerizations from the simple building block acetate. The rate of *de novo* synthesis in hepatocytes is regulated according to alimentary, thus plasmatic, cholesterol levels and is the major homeostatic factor for maintaining stable plasmatic concentrations. Differences in sensitivity of this retro control mechanism contribute substantially to individuals' cholesterol statuses. The four-ring structure of the cholesterol molecule is shown, with both a polar head group and a nonpolar hydrocarbon body making up its unique biophysical properties. Intermediates in the cholesterol biosynthesis pathway have many alternative fates, with in particular isopentyl pyrophosphate as the precursor of a large array of biomolecules with diverse biological roles. Cholesterol itself is the precursor for steroid hormones, vitamin D and oxysterols. Also, cholesterol can be fatty acylated to form cholesteryl esters for storage and transport. Liver cells excrete cholesterol into the bile both as free sterol and after conversion to bile acids. (B) Overview of cholesterol metabolism. Exogenous and endogenous biosynthetic cholesterol pathways and the metabolism of low-density lipoprotein (LDL) and high-density lipoprotein (HDL) are shown. Dietary cholesterol is packaged with triglycerides (TGs) into chylomicrons in the intestinal mucosa. Chylomicrons, secreted via the lymphatic system, enter the blood and circulate until they are hydrolyzed by an endothelial lipoprotein (LP) lipase in the capillaries, releasing free fatty acids (FAs). The chylomicron remnants are taken up by hepatic LDL receptor (LDLR), or by LDLR-related protein-1 (LRP1). Some of the cholesterol enters the metabolic pool, and some is excreted as free cholesterol or bile acids into the biliary tract. The endogenous synthesis of cholesterol begins in the liver, with 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGCR) being rate-limiting. LDL is responsible for cholesterol transport from the liver to the periphery. Its synthesis begins as liver cells secrete cholesterol and triglycerides incorporated into very low-density lipoprotein (VLDL) into the circulation. In the capillaries of adipose tissue and muscle, the TG contained in VLDL is hydrolysed by LP lipase, releasing free FAs and IDL. Further metabolism of IDL includes direct uptake through the LDL receptor and hydrolysis by hepatic lipase, yielding cholesterol-rich LDL. LDL is endocytosed by peripheral cells and hepatocytes by LDLR. HDL, which also originates from hepatic synthesis, mediates reverse cholesterol transport, by interacting with ATP-binding cassette A1 (ABCA1) and ABCG1 transporters on non-hepatic cells. ABCA1 mediates the rate-limiting step in HDL particle formation and maintenance of plasma HDL levels [58]. Lecithin-cholesterol acyltransferase (LCAT) esterifies cholesterol and produces globular HDL particles. Cholesteryl esters are selectively taken up from HDL in the liver by a scavenger receptor B1 (SRB1) dependent process. Adapted from "Toronto Notes" by Zamir Merali & Jason D. Woodfine, 2016. Copyright 2016 by Toronto Notes for Medical Students, Inc.

Figure 2

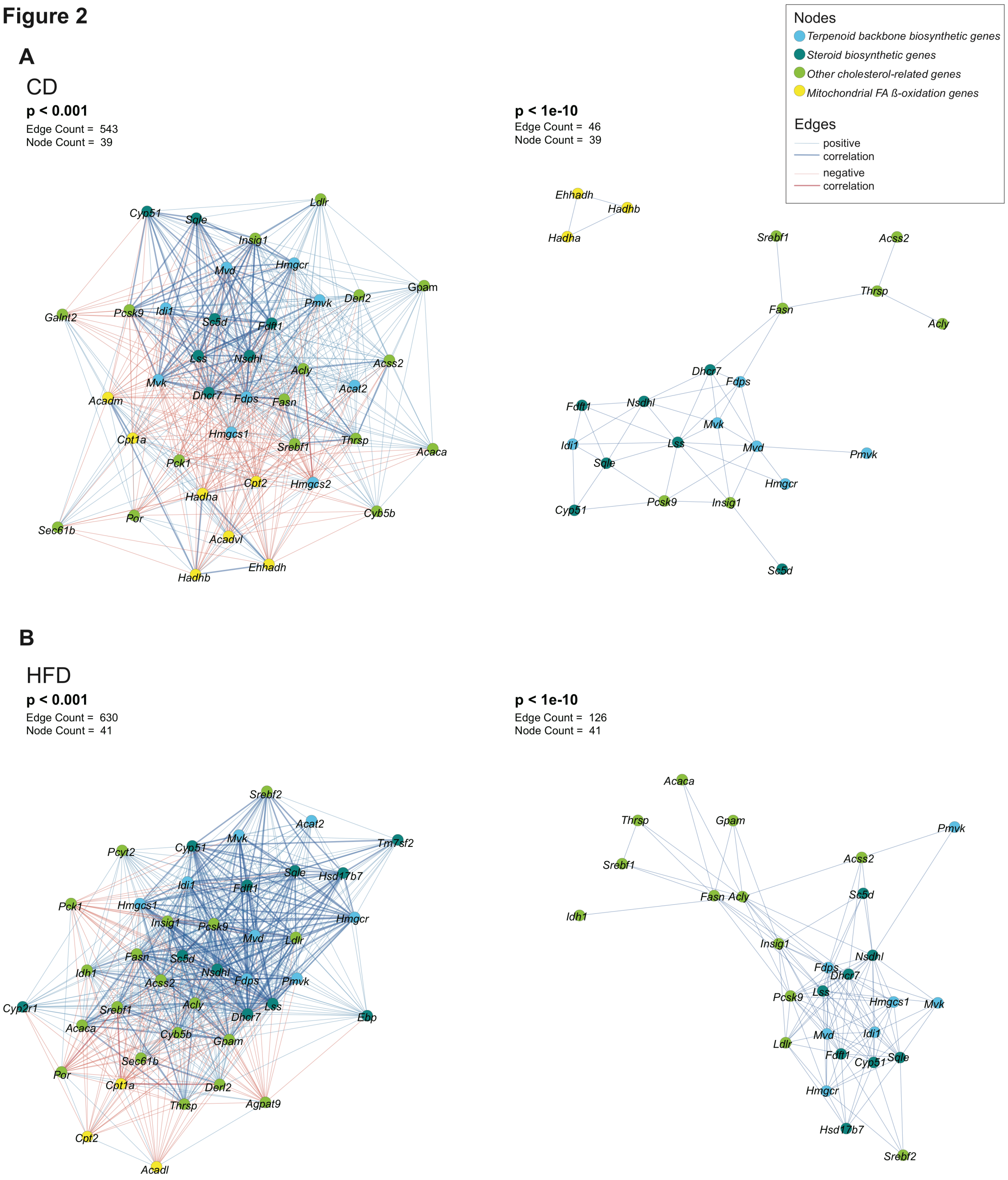


Figure 2: Regulatory network of cholesterol biosynthetic genes based on the transcriptome analysis in 42 different BXD strains. Nodes represent genes, and edges represent correlations between gene transcript levels among the different strains (co-expression). Genes were selected from literature and previous correlation analyses (see Suppl. Figure 1). Networks were computed for datasets on two dietary regimens: (A) chow diet (CD) and (B) high fat diet (HFD) with two levels of significance: p < 1e-3 and p < 1e-10.

Figure 3

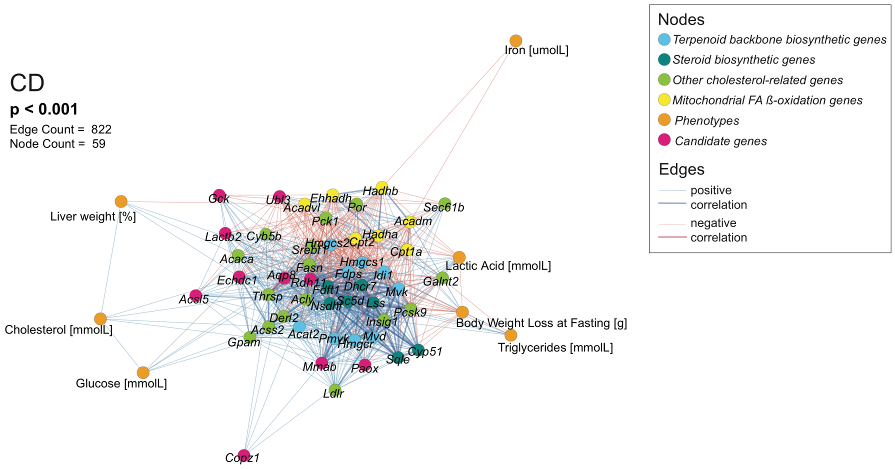


Figure 3: Transcriptional network of cholesterol biosynthetic genes with correlating phenotypes and mapping of some of the candidate regulatory genes.

Figure 4

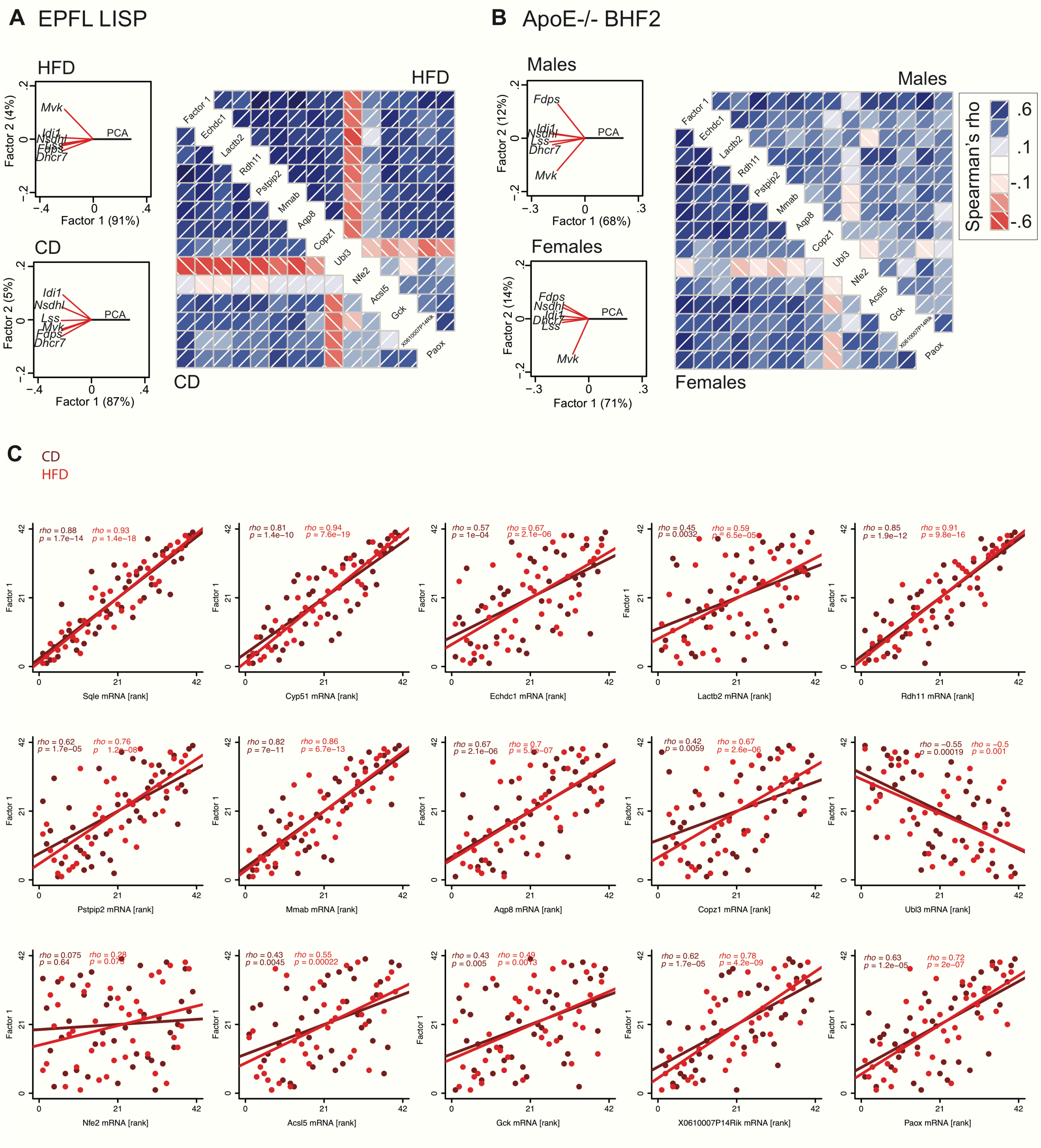


Figure 4: Identification of novel genes related to cholesterol biosynthesis by principal component analysis (PCA). (A) (Left) PCA plot for six cholesterol biosynthetic gene transcripts from the BXD liver mRNA dataset in both cohorts (CD, HFD). A single factor explains ~87% of the variance in their transcript levels in the CD and 91% in the HFD. (Right) 13 candidate genes are shown in a Spearman correlation matrix with the first PCA factor in CD (bottom left) and HFD (top right) conditions. (B) PCA was repeated in six independent datasets, including the BHF2 Apoe Null mouse population [54] shown here. (Left) PCA plot for the same six genes in both a male and a female cohort. The first principal component explains ~68% of the variance in transcript levels in the male cohort and 71% in the female cohort. (Right) A Spearman correlation matrix of factor 1 and the gene candidates is shown for the male (top right) and the female cohorts (bottom left). (C) Two known cholesterol genes and 13 candidates correlation plots to factor 1 in the PCA for both CD and HFD dataset.

Figure 5

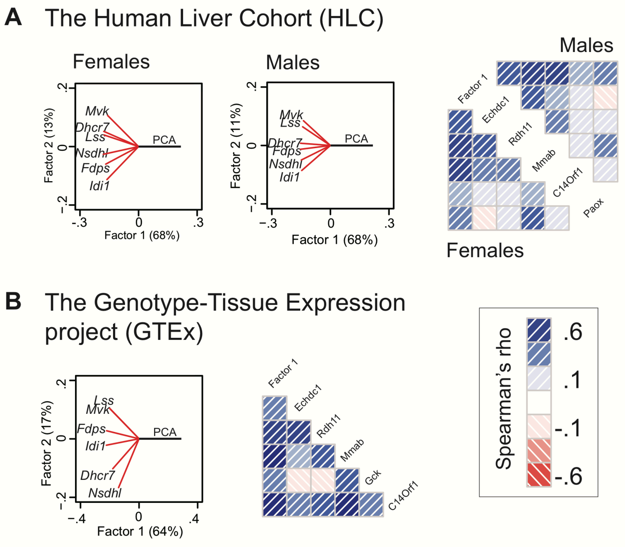
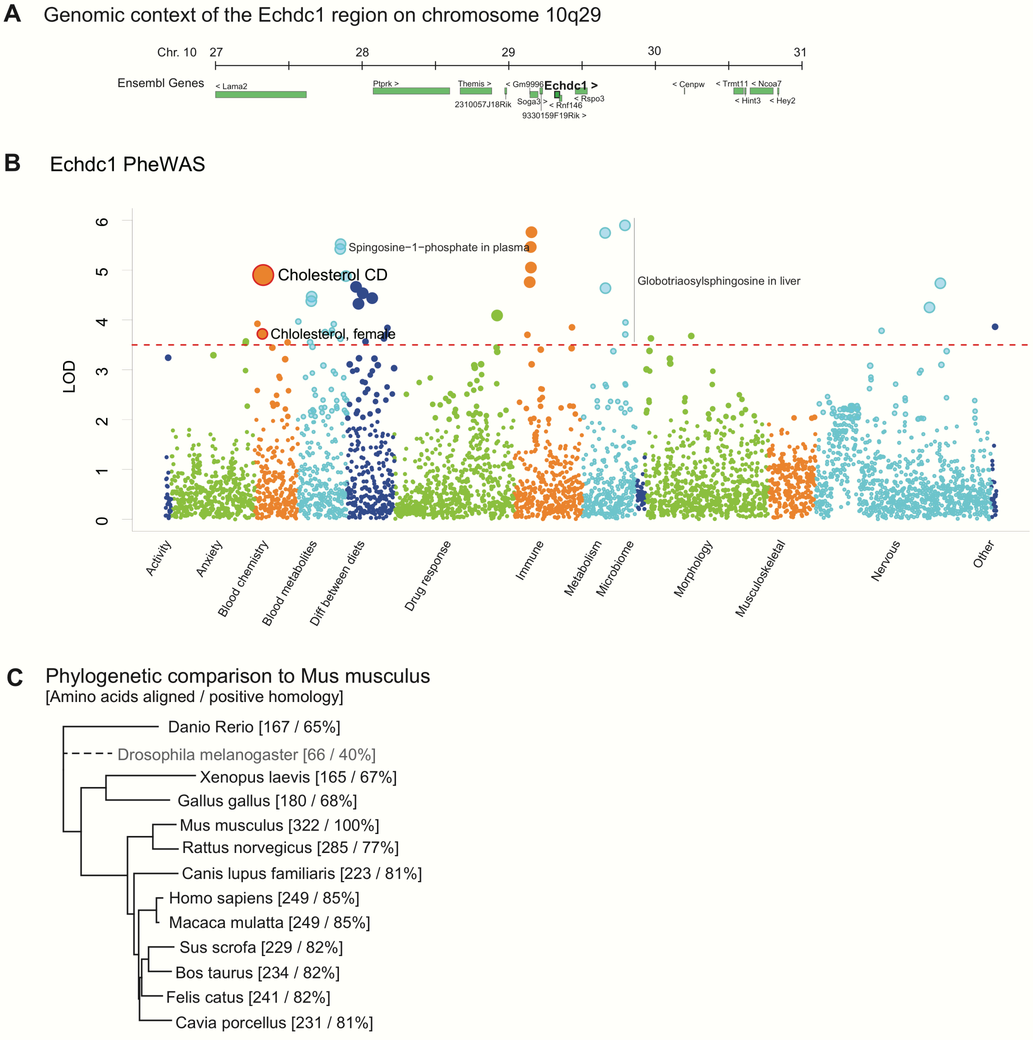


Figure 5: PCA validation of candidate genes in two human databases (HLC [40] and GTEx [41, 42]). Values of gene transcript abundance from the same core cholesterol biosynthetic genes as described previously were computed into principal components. (A) (Left) PCA plot from the HLC show that a single factor explains 68% of the variation in both the female and male cohorts. (Right) Five genes are significantly correlated to factor 1 in both cohorts. (B) PCA plot from the GTEx database show a single factor explain 64% of the variation in transcript levels. (Right) Five genes significantly correlate to factor 1. Genes that significantly ($p < 0.001$) correlate to factor 1 are shown in the corrgrams.

Figure 6


A Genomic context of the Echdc1 region on chromosome 10q29

B Echdc1 PheWAS

C Phylogenetic comparison to Mus musculus
[Amino acids aligned / positive homology]

Figure 6: *Echdc1* as a potent novel regulatory gene in cholesterol metabolism and biosynthesis. (A) Genomic context of the *Echdc1* region. (B) PheWAS Manhattan plot of molecular traits linked to the nearby region (±2.5Mb) of *Echdc1* on chromosome 10q29. (C) Phylogenetic comparison of *Echdc1* sequence of *Mus musculus*, which is conserved in vertebrates.

Figure 7

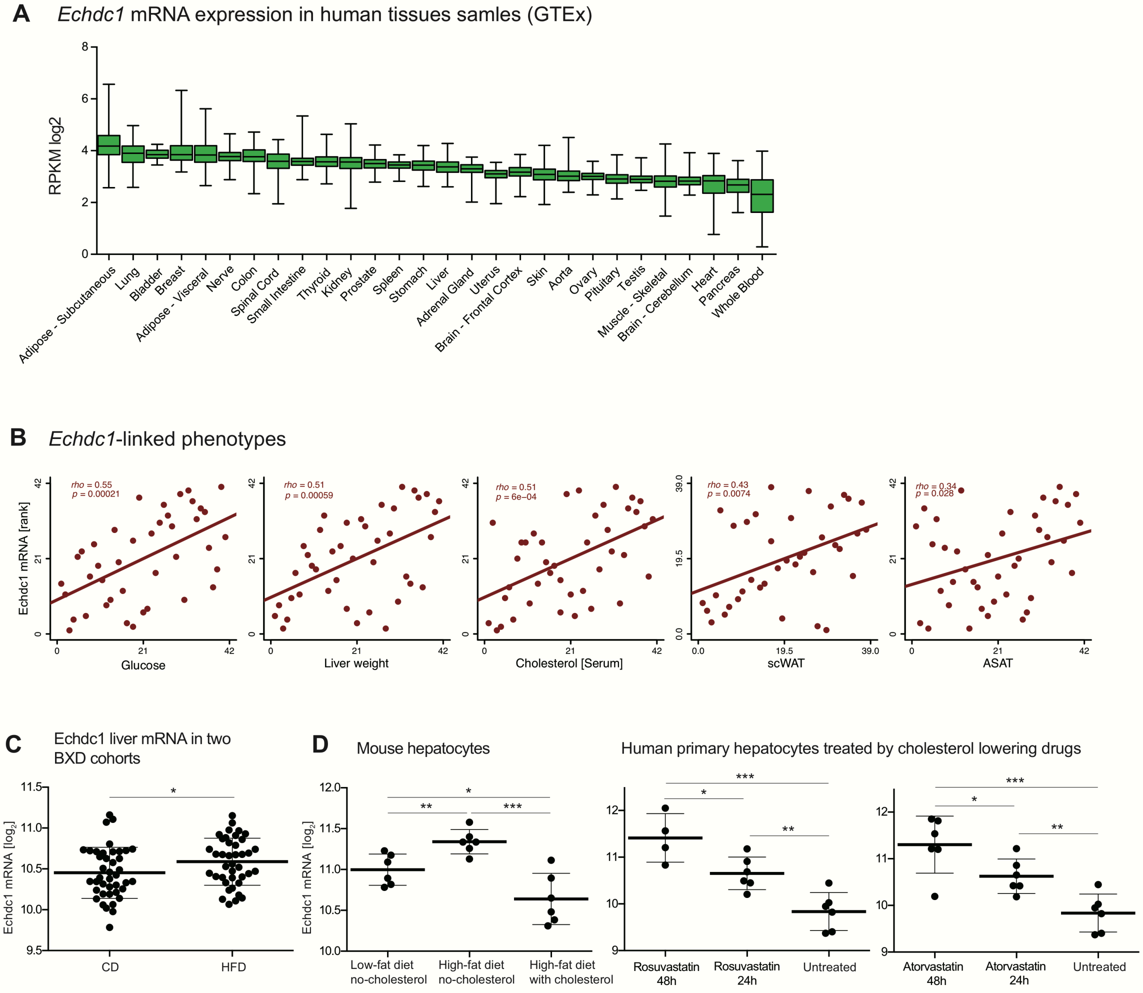
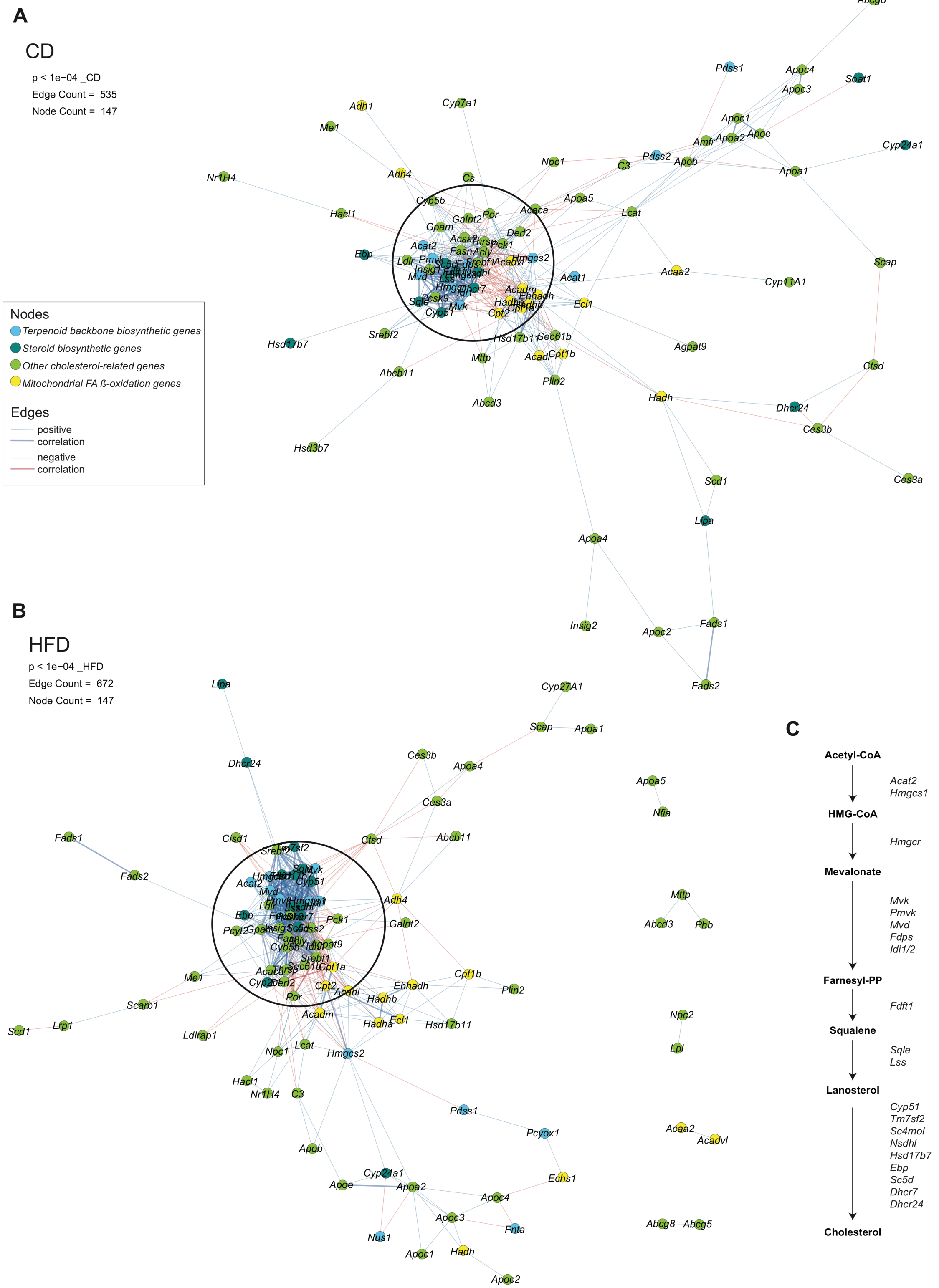


Figure 7: (A) *Echdc1* mRNA expression levels in different human tissues reported as RPKM (Reads Per Kilobase of transcript per Million mapped reads, corresponding to mean values of the different individual samples from each tissue type), from the GTEx project [41, 42] (B) *Echdc1* mRNA levels correlate to blood glucose, liver weight, plasma cholesterol, subcutaneous white adipose tissue (scWAT) and aspartate aminotransferase (ASAT) levels in the CD BXD mouse cohort. (C) *Echdc1* mRNA expression is up-regulated in the HFD BXD cohort (p = 0.0466). (D) Data retrieved from GEO [60] show that *Echdc1* is down-regulated in the hepatocytes of mice fed a HFD with additive cholesterol [45] whereas it is up-regulated in human hepatocytes treated by cholesterol lowering drugs [46].

Supplementary Figure 1



Supplementary figure 1: Transcriptional network of cholesterol related genes based on the transcriptome analysis in different BXD strains on two dietary regimens – chow diet (CD) or high fat diet (HFD). Genes related to cholesterol metabolism were selected from the literature. (A) CD co-expression network, where a core sub-network of high correlation density was defined manually (encircled). (B) HFD co-expression network. (C) Genes of the cholesterol biosynthesis pathway [61].

References

1. Tall, A.R. and L. Yvan-Charvet, *Cholesterol, inflammation and innate immunity.* Nat Rev Immunol, 2015. **15**(2): p. 104-16.
2. Simon, A., *Cholesterol metabolism and immunity.* N Engl J Med, 2014. **371**(20): p. 1933-5.
3. Silvente-Poirot, S. and M. Poirot, *Cancer. Cholesterol and cancer, in the balance.* Science, 2014. **343**(6178): p. 1445-6.
4. Wolozin, B., *Cholesterol, statins and dementia.* Curr Opin Lipidol, 2004. **15**(6): p. 667-72.
5. Grundy, S.M., *Absorption and metabolism of dietary cholesterol.* Annu Rev Nutr, 1983. **3**: p. 71-96.
6. Ikonen, E., *Cellular cholesterol trafficking and compartmentalization.* Nat Rev Mol Cell Biol, 2008. **9**(2): p. 125-38.
7. Horton, J.D., J.L. Goldstein, and M.S. Brown, *SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver.* J Clin Invest, 2002. **109**(9): p. 1125-31.
8. Kalaany, N.Y. and D.J. Mangelsdorf, *LXRS and FXR: the yin and yang of cholesterol and fat metabolism.* Annu Rev Physiol, 2006. **68**: p. 159-91.
9. Hegele, R.A., *Plasma lipoproteins: genetic influences and clinical implications.* Nat Rev Genet, 2009. **10**(2): p. 109-21.
10. Namboodiri, K.K., et al., *The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins.* Genet Epidemiol, 1985. **2**(3): p. 227-54.
11. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
12. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
13. Brown, M.S. and J.L. Goldstein, *Biomedicine. Lowering LDL--not only how low, but how long?* Science, 2006. **311**(5768): p. 1721-3.
14. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nat Genet, 2008. **40**(2): p. 161-9.
15. Kathiresan, S., et al., *Common variants at 30 loci contribute to polygenic dyslipidemia.* Nat Genet, 2009. **41**(1): p. 56-65.
16. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.
17. Kathiresan, S., et al., *Polymorphisms associated with cholesterol and risk of cardiovascular events.* N Engl J Med, 2008. **358**(12): p. 1240-9.
18. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
19. Flint, J. and T.F. Mackay, *Genetic architecture of quantitative traits in mice, flies, and humans.* Genome Res, 2009. **19**(5): p. 723-33.
20. Wei, W.H., G. Hemani, and C.S. Haley, *Detecting epistasis in human complex traits.* Nat Rev Genet, 2014. **15**(11): p. 722-33.
21. Schadt, E.E. and P.Y. Lum, *Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes.* J Lipid Res, 2006. **47**(12): p. 2601-13.
22. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* Nat Rev Genet, 2014. **15**(1): p. 34-48.
23. Balling, R., *From mouse genetics to systems biology.* Mamm Genome, 2007. **18**(6-7): p. 383-8.

24. Williams, E.G. and J. Auwerx, *The Convergence of Systems and Reductionist Approaches in Complex Trait Analysis.* Cell, 2015. **162**(1): p. 23-32.
25. Allayee, H., A. Ghazalpour, and A.J. Lusis, *Using mice to dissect genetic factors in atherosclerosis.* Arterioscler Thromb Vasc Biol, 2003. **23**(9): p. 1501-9.
26. Peters, L.L., et al., *The mouse as a model for human biology: a resource guide for complex trait analysis.* Nat Rev Genet, 2007. **8**(1): p. 58-69.
27. Getz, G.S. and C.A. Reardon, *Diet and murine atherosclerosis.* Arterioscler Thromb Vasc Biol, 2006. **26**(2): p. 242-9.
28. Whitman, S.C., *A practical approach to using mice in atherosclerosis research.* Clin Biochem Rev, 2004. **25**(1): p. 81-93.
29. Darvasi, A., *Experimental strategies for the genetic dissection of complex traits in animal models.* Nat Genet, 1998. **18**(1): p. 19-24.
30. Collaborative Cross, C., *The genome architecture of the Collaborative Cross mouse genetic reference population.* Genetics, 2012. **190**(2): p. 389-401.
31. 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.
32. Wang, J., R.W. Williams, and K.F. Manly, *WebQTL: web-based complex trait analysis.* Neuroinformatics, 2003. **1**(4): p. 299-308.
33. del Sol, A., et al., *Diseases as network perturbations.* Curr Opin Biotechnol, 2010. **21**(4): p. 566-71.
34. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.
35. Sharpe, L.J. and A.J. Brown, *Controlling cholesterol synthesis beyond 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR).* J Biol Chem, 2013. **288**(26): p. 18707-15.
36. Graham, R.M., et al., *Hepatic iron loading in mice increases cholesterol biosynthesis.* Hepatology, 2010. **52**(2): p. 462-71.
37. Wu, Y., et al., *Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population.* Cell, 2014. **158**(6): p. 1415-30.
38. Brown, S.D. and M.W. Moore, *The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping.* Mamm Genome, 2012. **23**(9-10): p. 632-40.
39. Adamo, K.B., et al., *Peroxisome proliferator-activated receptor gamma 2 and acyl-CoA synthetase 5 polymorphisms influence diet response.* Obesity, 2007. **15**(5): p. 1068-1075.
40. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.
41. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
42. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
43. Linster, C.L., et al., *Ethylmalonyl-CoA decarboxylase, a new enzyme involved in metabolite proofreading.* J Biol Chem, 2011. **286**(50): p. 42992-3003.
44. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
45. Lorbek, G., et al., *Lessons from hepatocyte-specific Cyp51 knockout mice: impaired cholesterol synthesis leads to oval cell-driven liver injury.* Sci Rep, 2015. **5**: p. 8777.
46. Hafner, M., et al., *The human primary hepatocyte transcriptome reveals novel insights into atorvastatin and rosuvastatin action.* Pharmacogenetics and Genomics, 2011. **21**(11): p. 741-750.
47. Gold, B., et al., *Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33.* Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4340-5.

48. Long, J., et al., *Evaluation of breast cancer susceptibility loci in Chinese women.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(9): p. 2357-65.

49. Zheng, W., et al., *Evaluation of 11 breast cancer susceptibility loci in African-American women.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(10): p. 2761-4.

50. Pitroda, S.P., et al., *MUC1-induced alterations in a lipid metabolic gene network predict response of human breast cancers to tamoxifen treatment.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5837-41.

51. Ahern, T.P., et al., *Statin prescriptions and breast cancer recurrence risk: a Danish nationwide prospective cohort study.* J Natl Cancer Inst, 2011. **103**(19): p. 1461-8.

52. Dale, K.M., et al., *Statins and cancer risk: a meta-analysis.* JAMA, 2006. **295**(1): p. 74-80.

53. Nelson, E.R., et al., *27-Hydroxycholesterol links hypercholesterolemia and breast cancer pathophysiology.* Science, 2013. **342**(6162): p. 1094-8.

54. Yang, X., et al., *Tissue-specific expression and regulation of sexually dimorphic genes in mice.* Genome Res, 2006. **16**(8): p. 995-1004.

55. Gatti, D., et al., *Genome-level analysis of genetic regulation of liver gene expression networks.* Hepatology, 2007. **46**(2): p. 548-57.

56. Shockley, K.R., et al., *Effects of atherogenic diet on hepatic gene expression across mouse strains.* Physiol Genomics, 2009. **39**(3): p. 172-82.

57. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-9.

58. Oram, J.F. and A.M. Vaughan, *ATP-Binding cassette cholesterol transporters and cardiovascular disease.* Circ Res, 2006. **99**(10): p. 1031-43.

59. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic Acids Res, 2005. **33**(Database issue): p. D428-32.

60. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.

61. Goldstein, J.L., R.A. DeBose-Boyd, and M.S. Brown, *Protein sensors for membrane sterols.* Cell, 2006. **124**(1): p. 35-46.