High Throughput Computational Mouse Genetic Analysis

Ahmed Arslan[1+], Yuan Guan[1+], Xinyu Chen[1], Robin Donaldson[2&], Wan Zhu, Madeline Ford[1], Manhong Wu, Ming Zheng[1], David L. Dill[2*] and Gary Peltz[1*]

[1]Department of Anesthesia, Stanford University School of Medicine, Stanford, CA; and [2]Department of Computer Science, Stanford University, Stanford, CA

[+]These authors contributed equally to this paper

[&] Current Address: Ecree Durham, NC 27701

*Address correspondence to: gpeltz@stanford.edu

Abstract

Background: Genetic factors affecting multiple biomedical traits in mice have been identified when GWAS data, which measured responses in panels of inbred mouse strains, was analyzed using haplotype-based computational genetic mapping (HBCGM). Although this method was previously used to analyze one dataset at a time; but now, a vast amount of mouse phenotypic data is now publicly available, which could enable many more genetic discoveries.

Results: HBCGM and a whole genome SNP map covering 43 inbred strains was used to analyze 8300 publicly available datasets of biomedical responses (1.52M individual datapoints) measured in panels of inbred mouse strains. As proof of concept, causative genetic factors affecting susceptibility for eye, metabolic and infectious diseases were identified when structured automated methods were used to analyze the output. One analysis identified a novel genetic effector mechanism; allelic differences within the mitochondrial targeting sequence affected the subcellular localization of a protein. We also found allelic differences within the mitochondrial targeting sequences of many murine and human proteins, and these could affect a wide range of biomedical phenotypes.

Implications: These initial results indicate that genetic factors affecting biomedical responses could be identified through analysis of very large datasets, and they provide an early indication of how this type of '*augmented intelligence*' can facilitate genetic discovery.

Abbreviations: cSNP, codon-changing SNP; GWAS, genome-wide association study; HBCGM, haplotype-based computational genetic mapping; MTS, Mitochondrial targeting sequence.

Introduction

Mouse is the premier model organism for biomedical discovery, and mice were used for the discovery or development of many therapies that are now in clinical use. However, similar to the difficulties encountered in analyzing human GWAS results [1], it has been difficult to identify the genetic factors underlying biomedical trait response differences in GWAS using inbred mouse strains. Just as human subpopulations are descended from ancestral founders; inbred laboratory strains are derived from an estimated four ancestral founders [2, 3]. Because of their ancestral relatedness, GWAS results will identify a true causative variant along with multiple other false positive associations, which are caused by genomic regions with a correlated genetic pattern that is due to common inheritance (a property referred to as 'population structure'). Statistical methods have been developed to reduce the false discovery rate by correcting for the population structure that exists in humans [4, 5], plants [6], and mice [7]. While these correction methods have utility for analysis of human GWAS results, we have shown that they are less useful for analyzing murine GWAS results, and they also increase the chance that a true causative genetic factor will be discarded [*]. In brief, even though multiple genomic regions have a shared ancestral inheritance, one may be responsible for a phenotypic difference.

Haplotype-based computational genetic mapping (**HBCGM**) [8] is a method for analyzing mouse GWAS data, which has identified genetic factors underlying 22 biomedical traits in mice [8-30]. One finding generated a new treatment for preventing narcotic drug withdrawal [22] that is now being tested in a multi-center clinical trial [31]. In an HBCGM experiment, a property of interest is measured in available mouse strains whose genomes have been sequenced; and genetic factors are computationally predicted by identifying genomic regions (haplotype blocks) where the pattern of within-block genetic variation correlates with the distribution of phenotypic responses among the strains [9, 32, 33]. A next-generation version of HBCGM with a 30,000-fold improvement in computational efficiency was developed, and whole genome sequence data for 26 strains was analyzed to produce a whole genome map with 16M SNPs [14]. HBCGM was previously used to analyze the response data or one trait at a time. But now, a vastly increased amount of phenotypic data for inbred mouse strains has become available. The Mouse Phenome Database (**MPD**) [34] has 8300 phenotypic datasets (1.52M individual datapoints) that measure experimentally-induced responses in panels of inbred mouse strains. They have been shown to be useful for genetic discovery since a genetic susceptibility factor for haloperidol-induced toxicity was identified by analysis of one MPD dataset [14]. Many more genetic discoveries could be made if all 8300 of the MPD datasets could be analyzed. However, HBCGM analyses generate many false positive associations that appear along with the causative genomic region for the trait response difference in the list of correlated genes. This creates a significant hurdle that limits our ability to genetically analyze the information contained within a large database like the MPD. For example, if 50 correlated genomic regions were identified for each of the 8300 MPD datasets (i.e. 415K possible genetic leads), the output could not be carefully examined by a team of dedicated individuals (or even by a University's entire staff of scientists). This problem is compounded by the fact that the HBCGM program maintains a relatively low threshold for identification of correlated genetic regions to avoid false negatives. To overcome this problem, selection methods are used to identify the true causative factor from among the multiple correlated regions. Causative genetic candidates were selected from among the many genes with correlated allelic patterns by applying orthogonal criteria [9, 35], which include gene expression, metabolomic [21], or curated biologic data [36], or by examining candidates within previously identified genomic regions [23, 24]. This approach evaluates genetic candidates using multiple criteria; this can provide superior results to that of a typical GWAS that only uses a single highly stringent criterion to identify candidates. In order to more efficiently identify likely candidate genes, the logical paths that were used to select the previously identified genetic factors were used to develop structured computational methods for analyzing HBCGM output. We demonstrate the utility of this approach by identifying several murine genetic factors that were known to affect important biomedical phenotypes. This approach is then used to identify a novel genetic effector mechanism that alters the expression and subcellular localization of the encoded protein. We also find that this novel genetic effector mechanism could affect many mouse and human genes.

*Wang, M. and G. Peltz. The Effect of Population Structure on Murine Genome-Wide Association Studies Manuscript submitted. Supplied as a supporting manuscript.

Results

Bulk analysis of MPD datasets. Whole genome sequencing data was used to generate a database with 21.3M SNPs with alleles covering 43 inbred mouse strains (**Table S1**). The fold genome coverage averages 30x per strain (range 19x to 64x based upon a 3 Gb genome); which is similar to that of other recent studies [37, 38]. The high fold-genome coverage and the use of stringent tiered methods used for variant calling [14] ensured that variants were identified with high confidence.

Each of the 8,300 MPD datasets is categorized according to the type of biomedical trait response measured [34]. As examples, 472 datasets relate to body weight or body fat composition; 255 measure an immune system response; 96 relate to drug metabolism; and 233 datasets measure a neurologic response. In some cases, multiple datasets measure the same response at different times after a treatment (i.e. haloperidol toxicity [14]). Irrespective of their grouping, we identified 1363 MPD datasets that measured a response in 12 or more strains and had inter-strain differences that were likely to be heritable (ANOVA p value $<1x10^{-5}$). We found that the selected 1363 MPD datasets had an average of 24 inbred strains that were in our database. After these were bulk analyzed by HBCGM, 560 datasets had at least one gene with a genetic association p value $<1x10^{-5}$. The HBCGM output for all of these datasets is available-in a clickable format that is indexed relative to the MPD information-at http://peltz-app-02.stanford.edu/cgi-bin/haplomap/supplementary.html. Given this large number of datasets, we do not know how many of these results correctly identify causative genetic factors. However, we present the results from several analyses; and show how application of additional computational analysis methods enabled causative genetic factors to be identified.

Applying structured computational methods. One MPD data set (MPD:1501) characterized the susceptibility of bone marrow macrophages obtained from female mice of 23 inbred strains to the *Bacillus anthracis* lethal factor [39]. This toxin killed macrophages obtained from twelve strains, while macrophages from 11 other strains were resistant to the toxin. HBCGM analysis indicated the allelic pattern that was most highly correlated with susceptibility to this toxin was within the pyrin domain of the *NACHT, LRR and PYD domains containing protein 1b* (*Nlrp1b*) (**Fig 1**). The susceptible and resistant strains had distinct *Nlrp1b* haplotypes. Of note, all of the most highly correlated genes were located in the same region of mouse chromosome 11; and the second most highly correlated gene was an adjacent functional homologue (*Nlrp1a*) of Nalp1b. Nlrp1b has been shown to play a crucial role in the formation of a subset of inflammasomes, which is an important part of the response to pathogenic infections [40, 41]; and allelic variation in *Nlrp1b* was previously shown to regulate macrophage [39] and neutrophil-dependent responses [42] to anthrax infection among inbred strains.

The anthrax toxin response was binary (death or survival), it was measured in a large number of strains, and the different types of response was evenly distributed among the 23 strains analyzed. However, the responses in many MPD datasets were characterized in a smaller number of strains, and the different response types of were unevenly distributed among the strains. Because of this, HBCGM analysis of most MPD datasets identified a much larger number of genes with haplotypic patterns that were highly correlated with the phenotypic response pattern; and structured computational methods for filtering the output were needed to identify the true causative genetic factor. For the cases described here, automated methods were used to identify correlated genes that: (i) were expressed within the target organ for the trait; (ii) contained a codon-changing SNP (cSNP); and (iii) a search of the biomedical literature indicated that the gene was related to the phenotype. For example, one MPD data set (MPD: 26721) examined the retinas of 29 strains: 21 strains had normal retinas, and 8 strains had retinal degeneration (**Fig. 2**). Thirty-three genes had haplotype blocks with a perfect genotype-phenotype correlation: all strains with normal retinas had one haplotype, while those with retinal degeneration shared a second haplotype. The computational filtering process rapidly reduced the number of candidates to four genes, which were all located within a single chromosomal region. Of these four genes, only one gene had a SNP allele with a stop codon, and the published literature analysis indicated that it was associated with vision. *Phosphodiesterase 6b* (*Pde6b*) encodes a phosphohydrolase that plays key role in transducing light mediated retinal signals [43]. A SNP with a stop codon (*Tyr347X*) was located within a protein domain (GAF) that occurs in cGMP-regulated phosphodiesterases that is responsible for high affinity, non-catalytic binding of two cGMP molecules/holoenzyme. All eight strains with retinal degeneration had the stop codon, while the 29 strains with normal retinas had the *Tyr347* allele (Fig. 2). Although we did not know it when this analysis was performed, a form of retinal degeneration that occurred in some inbred mouse strains was first described 93 years ago [44]; and the causative mutation (*retinal degeneration 1*, *rd1*) was subsequently localized to this *Pde6b* SNP [45]. Although a previously known genetic factor was identified, this result demonstrates these structured computational analysis methods have utility for genetic discovery.

Another data set (MPD: 9904) measured plasma high-density lipoprotein (**HDL**) cholesterol levels in female mice of 30 different inbred strains on a high-fat diet for 17 weeks [46]. There was a large inter-strain variation in HDL cholesterol levels (range 40 to 125 mg/dL) (**Fig. 3**), which was highly heritable (ANOVA p value = 3 x 10^{-72}). Since this was a quantitative trait, many genes had haplotypic patterns that correlated with the HDL levels. Therefore, automated methods were used to evaluate the top 50 gene candidates identified by the HBCGM analysis to identify genes that were: (i) expressed in liver, and (ii) contained a codon-changing SNP. Application of these criteria reduced the number of candidates to three genes (*Tomm40l*, *Nr1i3* and *Apoa2*), which were all located within a single genomic region (Chr 1, 171 MB), and a literature search indicated that only one was related to cholesterol metabolism. *Apoa2* encodes apolipoprotein (Apo-) A-II, which is the second most abundant protein within HDL particles, and it is involved in the lipoprotein metabolism pathway. *Apoa2* alleles were previously shown to affect HDL size and composition in mice [47]; and HDL levels in *Apoa2* knockout mice were decreased by 70% [48]. ApoA-II also plays an important role in cholesterol efflux; it modulates the interaction of HDL with lipid transfer proteins and enzymes [49]. Although the roles of the Pde6b in retinal degeneration, ApoA-II in HDL metabolism, and of Nrlp1b in the anthrax toxin response were previously known, these results demonstrate that genetic factors underlying important biomedical traits can be identified by HBCGM analysis of a large number of phenotypic datasets. However, candidate gene filtering methods were required for successful analysis of these traits, and these structured methods enabled the analyses to be much more rapidly completed. An example of a potential novel genetic finding for another retinal trait is described in the supplement (**Fig. S1**).

A novel genetic effector mechanism for a metabolic phenotype. Another dataset (MPD: 50243) measured hepatic succinylcarnitine levels in 16-week-old male mice after a 16-hour fast [50]. The levels were highly variable and were highly heritable (ANOVA p value = 1x10^{-18}) across the 24 strains. Two (FVB, SJL) of the 24 strains had a very high hepatic level of this metabolite (**Fig. 4**). Our own analysis confirmed that SJL mice had dramatically increased hepatic succinylcarnitine levels (**Fig. S2**). HBCGM analysis identified 152 genes whose allelic pattern correlated with the hepatic succinylcarnitine level (i.e. the FVB and SJL haplotype differed from the 22 other strains). However, only 3 of these genes were expressed in liver and had a codon changing SNP; and a literature search revealed that only one was linked with succinylcarnitine metabolism. *Lactb* encodes a serine beta-lactamase-like protein that forms filaments that localize to the region between the inner and outer mitochondrial membranes, and it has been proposed that Lactb filaments could affect mitochondrial organization and mitochondrial metabolism [51]. Moreover, analysis of two large population-based European cohorts revealed that human *LACTB* alleles were associated with plasma succinylcarnitine levels (rs2652822, p=7x10^{-27}) [52], and the level of *Lactb* mRNA expression positively correlates with body mass index (*p*□=□1.19□×□10^{-8}) [53]. Expression of a *Lactb* transgene also was shown to affect fat mass in mice [54, 55].

Examination of the allelic pattern shared by the two strains (FVB, SJL) with elevated hepatic succinylcarnitine levels suggested a potential genetic effector mechanism (**Fig S3**). Nuclear DNA encodes all except 13 of the ~1158 proteins required for the assembly and function of mitochondria [56-58]. These proteins are synthesized in the cytoplasm and are then transported into the mitochondria. Although there are various targeting mechanisms [59], most have NH2-terminal mitochondrial targeting sequences (**MTS**) that are enriched in hydrophobic and positively charged amino acids. Comparative analyses of yeast, mouse and human MTS have indicated that their length, sequence and net charge (between +3 and +6) are highly conserved [60]. The sequence conservation is due to the fact that the MTS must: form amphiphilic α-helices; interact with subunits of a mitochondrial surface protein for mitochondrial import; and must then be removed from the mature protein by a cleavage reaction that is performed by a very limited set of proteases [59, 61-65]. FVB and SJ/L mice share unique alleles at five SNP sites: *Arg110Gly* and *Pro88Leu* are within the NH2-terminal MTS; while three SNPs (*Val217Ala*, *Ala266Thr* and *Ser237Pro*) are within the sequence of the mature Lactb protein. Since two SNPs (*Pro88Leu, Arg110Gly*) introduce significant amino acid changes within the MTS (**Fig. 5**), the mitochondrial localization of Lactb could be altered in FVB and SJL mice. To investigate this, cDNAs encoding the C57BL/6 and FVB allelic forms of *Lactb* were expressed as EGFP fusion proteins in 293T cells (**Fig. 6A**). The C57BL/6 Lactb-EGFP fusion protein was highly expressed; it had a punctate expression pattern that overlapped with cellular mitochondria; and it differed from that of a control (EGFP only) protein that was expressed throughout the cytoplasm. In contrast, the FVB Lactb-EGFP fusion protein was expressed at a much lower level than the C57BL/6 protein (**Fig. 6B**). Importantly, the mRNAs for the two allelic forms of these fusion proteins were expressed at the same level after transfection (**Fig 6C**). Since these cDNAs were transcribed at similar rates, the protein expression differences must result from an allelic effect on a post-transcriptional process. To more precisely determine the basis for this allelic effect, FVB alleles at two positions (88Leu, 110Gly) within the MTS were engineered into the C57BL/6 *Lactb* cDNA by site directed mutagenesis (Fig. 6A). Interestingly, the C57BL/6$^{88L\ 110Gly}$ Lactb-EGFP fusion protein was expressed at a reduced level, which was similar to that of the FVB allelic form (**Fig. 6C**). Similar differences in protein expression were also noted when the different allelic forms of the fusion proteins were expressed in HepG2 cells (**Fig. S4**). These results indicate that while the C57BL/6 protein was efficiently expressed and transported into mitochondria, FVB alleles at two sites within its MTS dramatically reduced its expression level within mitochondria.

SNPs are present in the MTS of many murine and human proteins. To determine if allelic variation within the MTS could affect other nuclear-encoded mitochondrial proteins, we used our mouse SNP database to investigate whether MTS SNPs were present in any of the 524 genes that were annotated as nuclear encoded mitochondrial proteins [66]. We found 188 SNPs within the MTS of 120 of these murine genes; and 109 SNP alleles caused a major amino acid change in the predicted MTS of 79 genes (blosum-62 matrix score <-1) (**Table S2**). We then examined the NCBI SNP database to determine whether SNP alleles altered the MTS in 544 annotated human proteins [67], and found 161 codon-changing SNPs within the MTS of 83 of these proteins. It is noteworthy that 78 of these genes have other mutations that are located outside of the MTS, which cause human genetic diseases with very severe phenotypic effects (**Table S3**). Also, 8 genes have SNP alleles within their MTS that introduce a stop codon, and 12 genes have a SNP allele affecting the initiator methionine (**Table 1**). Moreover, allelic changes in 55 of these disease-associated proteins are predicted to have a major effect on the MTS sequence (blossom-62 matrix score $\leq$ -1). For example, there are 13 SNPs within the 33 amino acid MTS of a *thymidine kinase 2* (*TK2*) (**Table S3**). Genetic mutations that inactivate *TK2* cause mitochondrial DNA depletion, which presents in early childhood with a progressive myopathy or encephalopathy [68-71]. As another example, there were four cSNPs within the 53 amino acid MTS of *Pyruvate Dehydrogenase Complex Component X (PDHX)* (Table S3), which encodes the E3 ubiquitin ligase binding protein of the pyruvate dehydrogenase complex that catalyzes the rate-limiting step in aerobic glucose oxidation. A genetic deficiency of PDHX produces a life-threatening condition that causes developmental retardation [72]. Interestingly, the *Arg23Cys and Arg24Gly* SNP alleles have a major effect on the charge (the minor alleles reduce the charge from +6.1 to +4.0) and isoelectric point (from pH 12.6 to pH 10.74) of the PDHX MTS. Hence, there are multiple examples of allelic effects that are likely to impact the mitochondrial localization, and possibly the function, of human proteins that are of importance for cellular and tissue function.

Discussion

We demonstrate how computational analysis of a large biomedical response database could accelerate the pace of genetic discovery. Our ability to analyze this large phenotypic database is dependent upon two factors: (i) an increased number of inbred strains whose genome has been sequenced; and (ii) the use of automated methods for analyzing HBCGM output. *Why is the breadth of strain coverage important?* The database analyses responses across many strains, and this differs from prior mouse genetic analysis methods. For many years, mouse genetic models were analyzed by characterizing intercross progeny generated from two parental strains. This required a large amount of time for the generation and analysis of intercross progeny, and causative genetic factors could not be precisely localized [73]. To improve genetic mapping precision [74, 75], investigators have produced large panels of recombinant inbred strains (i.e. increased depth), but the progeny are generated from only 6-8 founder strains [76][77]. While the increased depth can improve genetic mapping precision, its utility is limited by the lack of strain breadth. When a small number of strains are evaluated, the actual extent of phenotypic variation that is present in the mouse population is under-estimated [9, 33]. This is a critical, since a key factor for successful genetic discovery is analyzing strains that exhibit outlier responses. For example, the MTS allelic effect on mitochondrial metabolism could not have been uncovered using any of the available recombinant inbred strain panels [76, 77] since the two strains (SJL, FVB) with high hepatic succinylcarnitine levels were not among the founder strains used to generate the panel. In fact, our initial analysis of the 8300 MPD datasets indicated that inbred strains exhibiting outlier responses (i.e. those in the top or bottom 10%) were often not found among the 23 strains previously in our SNP database [14]. The number of evaluable datasets increased when the genome sequence of 43 strains became available. Each inbred strain has unique genetic variants, and possibly phenotypic responses, which could enable genetic discovery. Our projections [9] indicate that over ~100,000 new SNPs per strain will be found even after the genomes of 40 strains are sequenced. As the emphasis in 21st Century healthcare shifts from disease treatment to disease prevention [78], new murine genetic models will be needed for the new phenotypes that will be of interest 10 or more years from now. Since we cannot predict which strains will have outlier responses for phenotypes of future interest, obtaining the genomic sequence for an increasing number of the >450 available inbred strains [79] is of great importance for 21st Century genetic discovery.

Why are automated and structured methods needed for GWAS data analysis? Filtering methods are required for selecting a true causative factor from among the many genomic regions that correlate with a phenotypic response pattern. On multiple prior occasions [35], we have found that causative genetic factors could be identified when other types of data were used to filter the gene candidates output by HBCGM analysis [12, 21, 23, 24, 36]. Since those analyses were manually performed, and they examined one dataset at a time, this filtering process is far too cumbersome for analyzing the 8300 available MPD datasets. Therefore, to select the most likely gene candidates among those output by HBCGM analysis, we developed selection criteria that could be automated. In these initial studies, we utilize gene expression criteria, select candidate genes with cSNPs; and use available information in published literature to identify the most likely causative gene(s) output by the genetic mapping program. This resembles methods developed by others that use: transcriptome wide association results [80, 81] or functional information [82-84] to select causative loci from among the many SNP sites identified in a human GWAS; or those that identify SNPs near *a priori* identified trait-related gene candidates in plant GWAS analyses [85]. The automated analysis of 152 correlated genes in the HBCGM output for the hepatic succinylcarnitine data led to the identification of four likely candidate genes, which was quickly narrowed to one obvious candidate by the literature search. While the criteria used in our initial studies select for cSNPs, a recent analysis of developmental disorders [86] indicated that allelic variation within non-coding regions could impact many other traits. We anticipate that filtering process improvements will enable other types of SNPs to be identified. Improved methods for analyzing the impact of allelic changes in non-coding sequences [88] could subsequently be used. A method for automated identification of genetic factors underlying metabolomic differences [87] could enable metabolomic data to be incorporated into the analyses. Our rudimentary literature searches could be improved by using a deep neural network [89], which has already been used to identify mutations that cause rare diseases in human populations. Implementation of these methods could enable ‘augmented intelligence’ further improve genetic discovery capabilities. This could enable the computational power and the large number of available phenotypic datasets to be used to advance our understanding of how biomedical traits are genetically regulated.

These computational methods identified a novel genetic effector mechanism*:* allelic changes within the MTS of murine *Lactb* alter its expression and mitochondrial localization. Moreover, this genetic effector mechanism could be active in other murine and human nuclear encoded mitochondrial proteins, which suggests that it could be of broad importance for disease susceptibility. In addition to the many known genetic diseases that are associated with nuclear encoded mitochondrial proteins [90], mitochondrial dysfunction is fundamental to many commonly occurring disease [92, 93] and age-associated conditions [57], with AIDs progression [94], and with cancer susceptibility [91]. In one case, *glutathione peroxidase 1* (*Pro198Leu)* alleles were shown to differentially affect its relative expression level in mitochondria, and altered the cellular response to oxidative stress [95]. Detailed studies in mice [96] and fish [97] have demonstrated that polymorphisms within the mitochondrial and nuclear genomes interact, and these interactions affect physiologically important processes. However, little was known about the effect of MTS polymorphisms on phenotypic responses and/or disease susceptibility. Of particular importance, MTS SNPs are present in many human proteins, including those where mutations outside of their MTS have caused very genetic diseases with a very severe impact. Since mutations within these genes have such significant health consequences, it is likely that at least some of the MTS allelic changes will impact other biomedical traits and possibly disease susceptibility.

Conclusions: Implementation of the computational analysis methods described here could enable augmented intelligence to be used for genetic discovery. This could enable the computational power and the large number of phenotypic datasets that are now available to be used to advance our understanding of how biomedical traits are genetically regulated.

Methods are available in the online supplement.

Data availability: The data sets within the Mouse Phenome Database (**MPD**) that were analyzed in this study are available at (https://phenome.jax.org). All sequence data is available at http://www.ncbi.nlm.nih.gov/bioproject/593371 (Bioproject ID: PRJNA593371). The HBCGM output for all of these datasets is available-in a clickable format that is indexed relative to the MPD information-at http://peltz-app-02.stanford.edu/cgi-bin/haplomap/supplementary.html. The source code for candidate gene filtering is available at http://github.com/AhmedArslan/HbCGM_paper, and that used for the functional characterization of genes is available at [98].

Competing interest: The authors declare that they have no competing interests.

Acknowledgements. We thank Dr. Robert Lewis for helpful comments and for reviewing this manuscript.

Funding: This work was supported by a NIH/NIDA award (5U01DA04439902) to GP.

Author contributions. GP and AA wrote the paper; AA, MZ, XC, WZ, MF, and RD generated data; and AA, MZ, DD and GP analyzed the data.

References

1. Gallagher MD, Chen-Plotkin AS: **The Post-GWAS Era: From Association to Function.** *Am J Hum Genet* 2018, **102:**717-730.
2. Guenet JL, Bonhomme F: **Wild mice: an ever-increasing contribution to a popular mammalian model.** *Trends Genet* 2003, **19:**24-31.
3. Reuveni E, Birney E, Gross CT: **The consequence of natural selection on genetic variation in the mouse.** *Genomics* 2010, **95:**196-202.
4. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38:**203-208.
5. Reich DE, Goldstein DB: **Detecting association in a case-control study while correcting for population stratification.** *Genet Epidemiol* 2001, **20:**4-16.
6. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M: **An Arabidopsis example of association mapping in structured samples.** *PLoS Genet* 2007, **3:**e4.
7. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178:**1709-1723.
8. Liao G, Wang J, Guo J, Allard J, Cheng J, Ng A, Shafer S, Puech A, McPherson JD, Foernzler D, et al: **In Silico Genetics: Identification of a Functional Element Regulating H2-Ea Gene Expression.** *Science* 2004, **306:**690-695.
9. Zheng M, Dill D, Peltz G: **A better prognosis for genetic association studies in mice.** *Trends Genet* 2012, **28:**62-69.
10. Ren M, Kazemian M, Zheng M, He J, Li P, Oh J, Liao W, Li J, Rajaseelan J, Kelsall BL, et al: **Transcription factor p73 regulates Th1 differentiation.** *Nat Commun* 2020, **11:**1475.
11. Donaldson R, Sun Y, Liang D-Y, Zheng M, Sahbaie P, Dill DL, Peltz G, Buck KJ, Clark JD: **The multiple PDZ domain protein Mpdz/MUPP1 regulates opioid tolerance and opioid-induced hyperalgesia.** *BMC Genomics* 2016, **17**.
12. Zhang H, Zheng M, Wu M, Xu D, Nishimura T, Nishimura Y, Giffard RG, Xiaong X, Xu LJ, Clark JD, et al: **A Pharmacogenetic Discovery: Cystamine Protects against Haloperidol-Induced Toxicity and Ischemic Brain Injury.** *Genetics* 2016, **203:**599-609.
13. Liang DY, Zheng M, Sun Y, Sahbaie P, Low SA, Peltz G, Scherrer G, Flores C, Clark JD: **The Netrin-1 receptor DCC is a regulator of maladaptive responses to chronic morphine administration.** *BMC Genomics* 2014, **15:**345.
14. Zheng M, Zhang H, Dill DL, Clark JD, Tu S, Yablonovitch AL, Tan MH, Zhang R, Rujescu D, Wu M, et al: **The Role of Abcb5 Alleles in Susceptibility to Haloperidol-Induced Toxicity in Mice and Humans** *PLoS Medicine* 2015, **12:**e1001782.
15. Liu HH, Hu Y, Zheng M, Suhoski MM, Engleman EG, Dill DL, Hudnall M, Wang J, Spolski R, Leonard WJ, Peltz G: **Cd14 SNPs regulate the innate immune response.** *Mol Immunol* 2012, **51:**112-127.
16. Sorge RE, Trang T, Dorfman R, Smith SB, Beggs S, Ritchie J, Austin JS, Zaykin DV, Meulen HV, Costigan M, et al: **Genetically determined P2X7 receptor pore formation regulates variability in chronic pain sensitivity.** *Nat Med* 2012, **18:**595-599.
17. Peltz G, Zaas AK, Zheng M, Solis NV, Zhang MX, Liu H-H, Hu Y, Boxx GM, Phan QT, Dill D, Filler SG: **Next-Generation Computational Genetic Analysis: Multiple Complement Alleles Control Survival After Candida Albicans Infection** *Infection and Immunity* 2011, **79:**4472-4479.

18. Tregoning JS, Yamaguchi Y, Wang B, Mihm D, Harker JA, Bushell ESC, Zheng M, Liao G, Peltz G, Openshaw PJM: **Genetic Susceptibility to the Delayed Sequelae of RSV Infection is MHC-Dependent, but Modified by Other Genetic Loci.** *J Immunology* 2010, **185:**5384-5391.
19. Hu Y, Liang D, Li X, Liu H-H, Zhang X, Zheng M, Dill D, Shi X, Qiao Y, Yeomans D, et al: **The Role of IL-1 in Wound Biology Part II: In vivo and Human Translational Studies.** *Anesthesia & Analgesia* 2010, **111:**1534-1542.
20. Hu Y, Liang D, Li X, Liu H-H, Zhang X, Zheng M, Dill D, Shi X, Qiao Y, Yeomans D, et al: **The Role of IL-1 in Wound Biology Part I: Murine in Silico and In vitro Experimental Analysis.** *Anesthesia & Analgesia* 2010, **111:**1525-1533.
21. Liu H-H, Lu P, Guo Y, Farrell E, Zhang X, Zheng M, Bosano B, Zhang Z, Allard J, Liao G, et al: **An Integrative Genomic Analysis Identifies Bhmt2 As A Diet-Dependent Genetic Factor Protecting Against Acetaminophen-Induced Liver Toxicity** *Genome Research* 2010, **20:**28-35.
22. Chu LF, Liang D-Y, Li X, Sahbaie P, D'Arcy N, Liao G, Peltz G, Clark JD: **From Mouse to Man: The 5-HT3 Receptor Modulates Physical Dependence on Opioid Narcotics.** *Pharmacogenetics and Genomics* 2009, **19:**193-205.
23. LaCroix-Fralish ML, Mo G, Smith SB, Sotocinal SG, Ritchie JG, Austin JS, Melmed K, Schorscher-Petcu A, Laferriere AC, Lee TH, et al: **The β3 Subunit of the Na+,K+-ATPase Affects Pain Sensitivity.** *Pain* 2009, **144:**294-302.
24. Smith SB, Marker CL, Perry C, Liao G, Sotocinal SG, Austin JS, Melmed K, David Clark J, Peltz G, Wickman K, Mogil JS: **Quantitative trait locus and computational mapping identifies Kcnj9 (GIRK3) as a candidate gene affecting analgesia from multiple drug classes.** *Pharmacogenetics and Genomics* 2008, **18:**231-241.
25. Zaas AK, Liao G, Chein J, Usuka J, Weinberg C, Shore D, Giles D, Marr K, Burch L, Perara, et al: **Plasminogen Alleles Influence Susceptibility to Invasive Aspergillosis.** *PLoS genetics* 2008, **4:**e1000101.
26. Liang D, Liao G, Wang J, Usuka J, Guo YY, Peltz G, Clark JD: **A Genetic Analysis of Opioid-Induced Hyperalgesia in Mice** *Anesthesiology* 2006, **104:**1054-1062.
27. Guo YY, Weller PF, Farrell E, Cheung P, Fitch B, Clark D, Wu SY, Wang J, Liao G, Zhang Z, et al: **In Silico Pharmacogenetics: Warfarin Metabolism.** *Nature Biotechnology* 2006, **24:**531-536.
28. Rozzo SJ, Allard J, Choubey D, Vyse T, Izui S, Peltz G, Kotzin BL: **Evidence for an interferon-inducible gene, Ifi202, in the susceptibility to Systemic Lupus.** *Immunity* 2001, **15:**435-443.
29. Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G: **In silico mapping of complex disease-related traits in mice.** *Science* 2001, **292:**1915-1918.
30. Guo YY, Liu P, Zhang X, Weller PMM, Wang J, Liao G, Zhang Z, Hu J, Allard J, Shafer S, et al: **In vitro and In silico Pharmacogenetic Analysis in Mice.** *Proceedings of the National Academy of Sciences* 2007, **104:**17735-17740.
31. Peltz G, Sudhof TC: **The Neurobiology of Opioid Addiction and the Potential for Prevention Strategies.** *JAMA* 2018, **319:**2071-2072.
32. Liao G, Wang J, Guo J, Allard J, Chang J, Nguyen A, Shafer S, Puech A, McPherson JD, Foernzler D, et al: **In Silico Genetics: Identification of A Novel Functional Element Regulating H2-Ea Gene Expression** *Science* 2004, **306:**690-695.
33. Wang J, Liao G, Usuka J, Peltz G: **Computational Genetics: From Mouse to Man?** *Trends in Genetics* 2005, **21:**526-532.
34. Grubb SC, Bult CJ, Bogue MA: **Mouse phenome database.** *Nucleic Acids Res* 2014, **42:**D825-834.

35. Zheng M, Shafer SS, Liao G, Liu H-H, Peltz G: **Computational Genetic Mapping in Mice: 'The Ship has Sailed'.** *Science Translational Medicine* 2009, **1:**3ps4.

36. Zhang X, Liu H-H, Weller P, Tao W, Wang J, Liao G, Zheng M, Monshouwer M, Peltz G: **In Silico and In Vitro Pharmacogenetics: Aldehyde Oxidase Rapidly Metabolizes a p38 Kinase Inhibitor.** *The Pharmacogenomics Journal* 2011, **11:**15-24.

37. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, Chow W, Collins J, Collins S, Czechanski A, et al: **Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci.** *Nat Genet* 2018, **50:**1574-1583.

38. Doran AG, Wong K, Flint J, Adams DJ, Hunter KW, Keane TM: **Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations.** *Genome Biol* 2016, **17:**167.

39. Boyden ED, Dietrich WF: **Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin.** *Nat Genet* 2006, **38:**240-244.

40. Bruey JM, Bruey-Sedano N, Luciano F, Zhai D, Balpai R, Xu C, Kress CL, Bailly-Maitre B, Li X, Osterman A, et al: **Bcl-2 and Bcl-XL regulate proinflammatory caspase-1 activation by interaction with NALP1.** *Cell* 2007, **129:**45-56.

41. Finger JN, Lich JD, Dare LC, Cook MN, Brown KK, Duraiswami C, Bertin J, Gough PJ: **Autolytic proteolysis within the function to find domain (FIIND) is required for NLRP1 inflammasome activity.** *J Biol Chem* 2012, **287:**25030-25037.

42. Moayeri M, Crown D, Newman ZL, Okugawa S, Eckhaus M, Cataisson C, Liu S, Sastalla I, Leppla SH: **Inflammasome sensor Nlrp1b-dependent resistance to anthrax is mediated by caspase-1, IL-1 signaling and neutrophil recruitment.** *PLoS Pathog* 2010, **6:**e1001222.

43. Han J, Dinculescu A, Dai X, Du W, Smith WC, Pang J: **Review: the history and role of naturally occurring mouse models with Pde6b mutations.** *Mol Vis* 2013, **19:**2579-2589.

44. Keeler CE: **The Inheritance of a Retinal Abnormality in White Mice.** *Proc Natl Acad Sci U S A* 1924, **10:**329-333.

45. Pittler SJ, Keeler CE, Sidman RL, Baehr W: **PCR analysis of DNA from 70-year-old sections of rodless retina demonstrates identity with the mouse rd defect.** *Proc Natl Acad Sci U S A* 1993, **90:**9616-9619.

46. Svenson KL, Von Smith R, Magnani PA, Suetin HR, Paigen B, Naggert JK, Li R, Churchill GA, Peters LL: **Multiple trait measurements in 43 inbred mouse strains capture the phenotypic diversity characteristic of human populations.** *J Appl Physiol (1985)* 2007, **102:**2369-2378.

47. Doolittle MH, LeBoeuf RC, Warden CH, Bee LM, Lusis AJ: **A polymorphism affecting apolipoprotein A-II translational efficiency determines high density lipoprotein size and composition.** *J Biol Chem* 1990, **265:**16380-16388.

48. Weng W, Brandenburg NA, Zhong S, Halkias J, Wu L, Jiang XC, Tall A, Breslow JL: **ApoA-II maintains HDL levels in part by inhibition of hepatic lipase. Studies In apoA-II and hepatic lipase double knockout mice.** *J Lipid Res* 1999, **40:**1064-1070.

49. Blanco-Vaca F, Escola-Gil JC, Martin-Campos JM, Julve J: **Role of apoA-II in lipid metabolism and atherosclerosis: advances in the study of an enigmatic protein.** *J Lipid Res* 2001, **42:**1727-1739.

50. Ghazalpour A, Bennett BJ, Shih D, Che N, Orozco L, Pan C, Hagopian R, He A, Kayne P, Yang WP, et al: **Genetic regulation of mouse liver metabolite levels.** *Mol Syst Biol* 2014, **10:**730.

51. Polianskyte Z, Peitsaro N, Dapkunas A, Liobikas J, Soliymani R, Lalowski M, Speer O, Seitsonen J, Butcher S, Cereghetti GM, et al: **LACTB is a filament-forming protein localized in mitochondria.** *Proc Natl Acad Sci U S A* 2009, **106:**18960-18965.
52. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, CardioGram, Deloukas P, Erdmann J, et al: **Human metabolic individuality in biomedical and pharmaceutical research.** *Nature* 2011, **477:**54-60.
53. Pan DZ, Garske KM, Alvarez M, Bhagat YV, Boocock J, Nikkola E, Miao Z, Raulerson CK, Cantor RM, Civelek M, et al: **Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS.** *Nat Commun* 2018, **9:**1512.
54. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452:**429-435.
55. Yang X, Deignan JL, Qi H, Zhu J, Qian S, Zhong J, Torosyan G, Majid S, Falkard B, Kleinhanz RR, et al: **Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks.** *Nat Genet* 2009, **41:**415-423.
56. Calvo SE, Clauser KR, Mootha VK: **MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins.** *Nucleic Acids Res* 2016, **44:**D1251-1257.
57. Wallace DC: **A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine.** *Annu Rev Genet* 2005, **39:**359-407.
58. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, et al: **A mitochondrial protein compendium elucidates complex I disease biology.** *Cell* 2008, **134:**112-123.
59. Verechshagina NA, Konstantinov YM, Kamenski PA, Mazunin IO: **Import of Proteins and Nucleic Acids into Mitochondria.** *Biochemistry (Mosc)* 2018, **83:**643-661.
60. Calvo SE, Julien O, Clauser KR, Shen H, Kamer KJ, Wells JA, Mootha VK: **Comparative Analysis of Mitochondrial N-Termini from Mouse, Human, and Yeast.** *Mol Cell Proteomics* 2017, **16:**512-523.
61. Poveda-Huertes D, Mulica P, Vogtle FN: **The versatility of the mitochondrial presequence processing machinery: cleavage, quality control and turnover.** *Cell Tissue Res* 2017, **367:**73-81.
62. Garg SG, Gould SB: **The Role of Charge in Protein Targeting Evolution.** *Trends Cell Biol* 2016, **26:**894-905.
63. Omura T: **Mitochondria-targeting sequence, a multi-role sorting sequence recognized at all steps of protein import into mitochondria.** *J Biochem* 1998, **123:**1010-1016.
64. Hartmann C, Christen P, Jaussi R: **Mitochondrial protein charge.** *Nature* 1991, **352:**762-763.
65. Jaussi R: **Homologous nuclear-encoded mitochondrial and cytosolic isoproteins. A review of structure, biosynthesis and genes.** *Eur J Biochem* 1995, **228:**551-561.
66. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32:**D115-119.
67. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
68. Tyynismaa H, Suomalainen A: **Mouse models of mitochondrial DNA defects and their relevance for human disease.** *EMBO Rep* 2009, **10:**137-143.
69. Sun R, Wang L: **Thymidine kinase 2 enzyme kinetics elucidate the mechanism of thymidine-induced mitochondrial DNA depletion.** *Biochemistry* 2014, **53:**6142-6150.

70. Saada A, Shaag A, Mandel H, Nevo Y, Eriksson S, Elpeleg O: **Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy.** *Nat Genet* 2001, **29:**342-344.
71. Gotz A, Isohanni P, Pihko H, Paetau A, Herva R, Saarenpaa-Heikkila O, Valanne L, Marjavaara S, Suomalainen A: **Thymidine kinase 2 defects can cause multi-tissue mtDNA depletion syndrome.** *Brain* 2008, **131:**2841-2850.
72. Patel KP, O'Brien TW, Subramony SH, Shuster J, Stacpoole PW: **The spectrum of pyruvate dehydrogenase complex deficiency: clinical, biochemical and genetic features in 371 patients.** *Mol Genet Metab* 2012, **106:**385-394.
73. Darvasi A, Soller M: **A simple method to calculate resolving power and confidence interval of QTL map location.** *Behavior Genetics* 1997, **27:**125-132.
74. Chesler EJ: **Out of the bottleneck: the Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research.** *Mamm Genome* 2014, **25:**3-11.
75. Woods LC, Mott R: **Heterogeneous Stock Populations for Analysis of Complex Traits.** *Methods Mol Biol* 2017, **1488:**31-44.
76. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J: **Genome-wide genetic association of complex traits in heterogeneous stock mice.** *Nat Genet* 2006, **38:**879-887.
77. Valdar W, Flint J, Mott R: **Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice.** *Genetics* 2006, **172:**1783-1797.
78. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, McDonald DT, Kusebauch U, Moss CL, Zhou Y, et al: **A wellness study of 108 individuals using personal, dense, dynamic data clouds.** *Nat Biotechnol* 2017, **35:**747-756.
79. Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM: **Genealogies of mouse inbred strains.** *Nature Genetics* 2000, **24:**23-25.
80. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, Pasaniuc B: **Probabilistic fine-mapping of transcriptome-wide association studies.** *Nat Genet* 2019, **51:**675-682.
81. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al: **Opportunities and challenges for transcriptome-wide association studies.** *Nat Genet* 2019, **51:**592-599.
82. de Leeuw CA, Neale BM, Heskes T, Posthuma D: **The statistical properties of gene-set analysis.** *Nat Rev Genet* 2016, **17:**353-364.
83. Hammerschlag AR, de Leeuw CA, Middeldorp CM, Polderman TJC: **Synaptic and brain-expressed gene sets relate to the shared genetic risk across five psychiatric disorders.** *Psychol Med* 2019**:**1-11.
84. Watanabe K, Umicevic Mirkov M, de Leeuw CA, van den Heuvel MP, Posthuma D: **Genetic mapping of cell type specificity for complex traits.** *Nat Commun* 2019, **10:**3222.
85. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al: **Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines.** *Nature* 2010, **465:**627-631.
86. Martin HC, Jones WD, McIntyre R, Sanchez-Andrade G, Sanderson M, Stephenson JD, Jones CP, Handsaker J, Gallone G, Bruntraeger M, et al: **Quantifying the contribution of recessive coding variation to developmental disorders.** *Science* 2018.
87. Wu M, Zheng M, Zhang W, Suresh S, Schlecht U, Fitch WL, Aronova S, Baumann S, Davis R, St Onge R, et al: **Identification of drug targets by chemogenomic and metabolomic profiling in yeast.** *Pharmacogenet Genomics* 2012, **22:**877-886.
88. Madelaine R, Notwell JH, Skariah G, Halluin C, Chen CC, Bejerano G, Mourrain P: **A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2**

functional mutation associated to retinal vasculature defects in human. *Nucleic Acids Res* 2018, **46:**3517-3531.

89. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, et al: **Predicting the clinical impact of human mutation with deep neural networks.** *Nat Genet* 2018, **50:**1161-1170.
90. Jackson TD, Palmer CS, Stojanovski D: **Mitochondrial diseases caused by dysfunctional mitochondrial protein import.** *Biochem Soc Trans* 2018, **46:**1225-1238.
91. Wallace DC: **Mitochondria and cancer: Warburg addressed.** *Cold Spring Harb Symp Quant Biol* 2005, **70:**363-374.
92. Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF: **Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases.** *PLoS Genet* 2014, **10:**e1004369.
93. Fuku N, Park KS, Yamada Y, Nishigaki Y, Cho YM, Matsuo H, Segawa T, Watanabe S, Kato K, Yokoi K, et al: **Mitochondrial haplogroup N9a confers resistance against type 2 diabetes in Asians.** *Am J Hum Genet* 2007, **80:**407-415.
94. Hendrickson SL, Lautenberger JA, Chinn LW, Malasky M, Sezgin E, Kingsley LA, Goedert JJ, Kirk GD, Gomperts ED, Buchbinder SP, et al: **Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression.** *PLoS One* 2010, **5:**e12862.
95. Bera S, Weinberg F, Ekoue DN, Ansenberger-Fricano K, Mao M, Bonini MG, Diamond AM: **Natural allelic variations in glutathione peroxidase-1 affect its subcellular localization and function.** *Cancer Res* 2014, **74:**5118-5126.
96. Latorre-Pellicer A, Moreno-Loshuertos R, Lechuga-Vieco AV, Sanchez-Cabo F, Torroja C, Acin-Perez R, Calvo E, Aix E, Gonzalez-Guerra A, Logan A, et al: **Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing.** *Nature* 2016, **535:**561-565.
97. Baris TZ, Wagner DN, Dayan DI, Du X, Blier PU, Pichaud N, Oleksiak MF, Crawford DL: **Evolved genetic and phenotypic differences due to mitochondrial-nuclear interactions.** *PLoS Genet* 2017, **13:**e1006517.
98. Arslan A, van Noort V: **yMap: an automated method to map yeast variants to protein modifications and functional regions.** *Bioinformatics* 2017, **33:**571-573.
99. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y: **The I-TASSER Suite: protein structure and function prediction.** *Nat Methods* 2015, **12:**7-8.

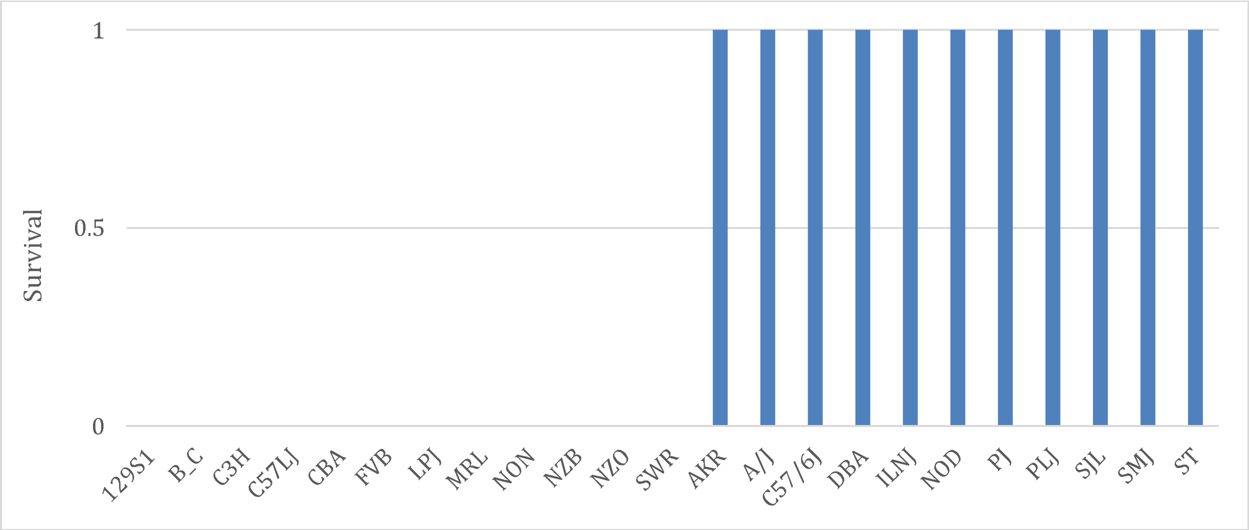
Table 1. We used the NCBI human SNP database to identify SNP alleles that altered the MTS of the 544 genes that encoded human proteins that were annotated as having a MTS. We found human SNP alleles that introduced a premature termination codon (PTC) within the MTS of eight of these genes (**A**), and human SNPs that altered the initiator methionine for 12 genes (**B**). This table shows the gene symbol, the predicted length of the amino-terminal MTS, the position of the altered amino acid within the MTS, and the identity of the reference and variant amino acid for each SNP. These human SNPs are of interest because they are located within genes that have one or more other mutations, which are located outside of the MTS, that have been shown to cause a human genetic disease. The name of the disease caused by other SNPs (outside of the MTS) within each indicated gene (obtained from the Mendelian Inheritance in Man database) is shown. While the MTS SNPs are not disease associated, they could be of interest. For example, seven of the 8 genes, which have a SNP that introduces a PTC allele in their MTS are associated with metabolic changes that cause severe diseases. Thus, an individual that expresses a truncated form of any of the proteins encoded by these genes could have a metabolic abnormality. Similarly, a polymorphism affecting the initiator methionine of any of the 12 genes shown could have an impact on several phenotypes.

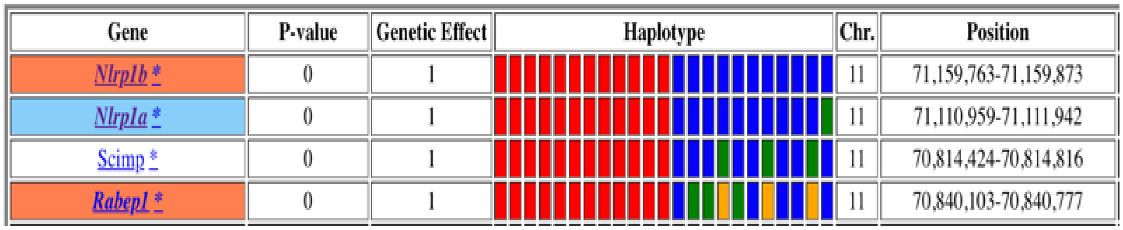
A

Symbol	***Length***	***SNP-location***	***Ref***	***Variant***	***Disease***
ACADVL	40	22	S	X	*Acyl-CoA dehydrogenase very long-chain deficiency (ACADVLD)*
COQ6	28	14	W	X	*Coenzyme Q10 deficiency, primary, 6 (COQ10D6)*
FH	44	3	R	X	*Fumarase deficiency (FMRD)*
GDF5OS	48	20	Q	X	*Cerebral creatine deficiency syndrome 3 (CCDS3)*
MUT	32	7	Q	X	*Methylmalonic aciduria type mut (MMAM)*
PCK2	32	23	S	X	*Mitochondrial phosphoenolpyruvate carboxykinase deficien (M-PEPCKD)*
SDHB	28	27	R	X	*Pheochromocytoma (PCC)*
TTC19	70	65	W	X	*Mitochondrial complex III deficiency, nuclear 2 (MC3DN2*

B

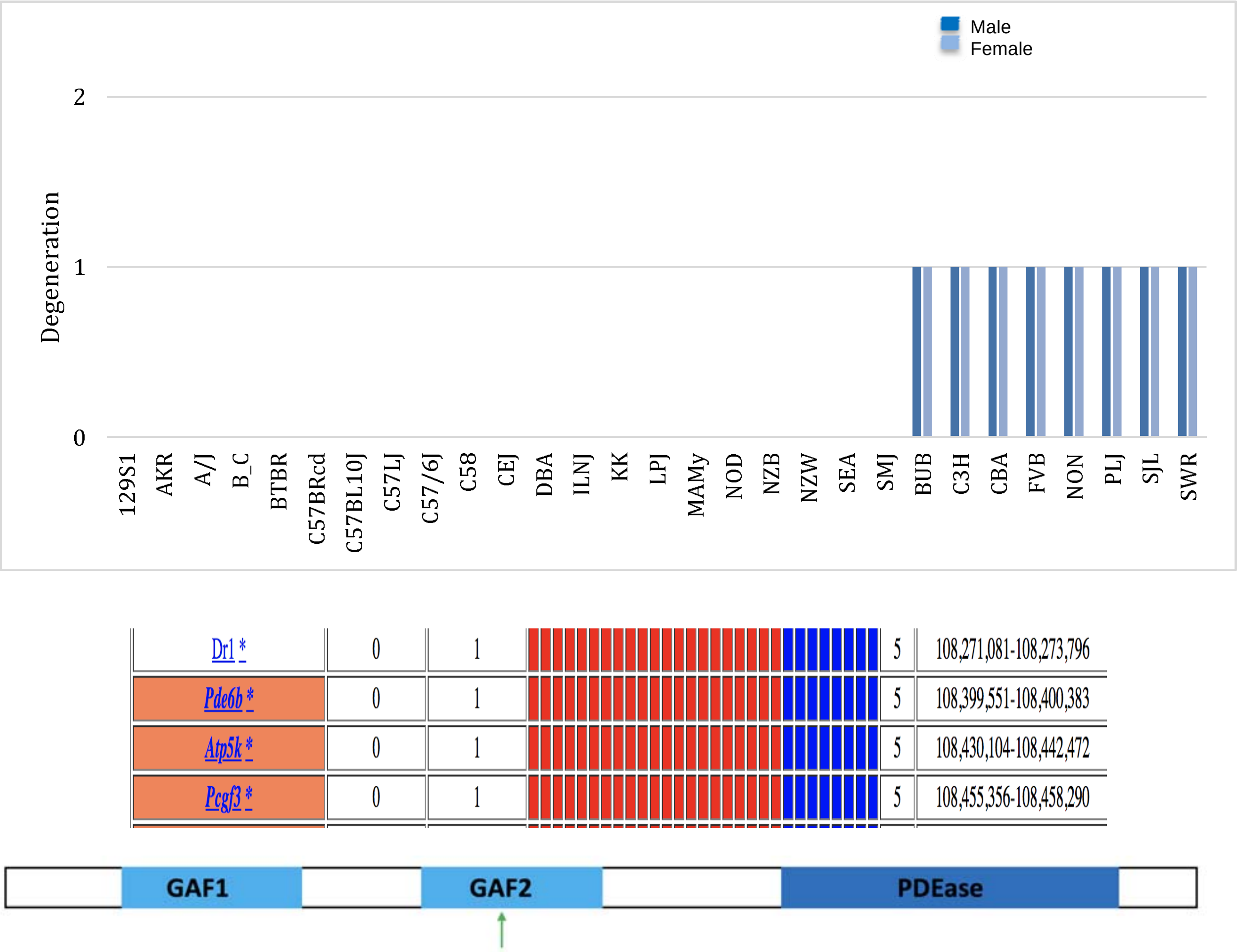
Symbol	***Length***	***SNP-location***	***Ref***	***Variant***	***Disease***
ACAT1	33	1	M	K	*3-ketothiolase deficiency (3KTD)*
CYP11B1	24	1	M	I	*Adrenal hyperplasia 4 (AH4)*
ETFDH	33	1	M	T	*Ethylmalonic encephalopathy (EE)*
ETHE1	7	1	M	I	*Ethylmalonic encephalopathy (EE)*
FXN	41	1	M	I	*Friedreich ataxia (FRDA)*
GLDC	35	1	M	T	*Non-ketotic hyperglycinemia (NKH)*
NFU1	9	1	M	K	*Multiple mitochondrial dysfunctions syndrome 1 (MMDS*
OAT	35	1	M	I	*Hyperornithinemia with gyrate atrophy of choroid and retina (HOGA)*
OTC	32	1	M	V	*Ornithine carbamoyltransferase deficiency (OTCD)*
PANK2	46	1	M	T	*Neurodegeneration with brain iron accumulation 1 (NBIA*
SDHA	42	1	M	L	*Mitochondrial complex II deficiency (MT-C2D)*
SDHD	56	1	M	I	*Paragangliomas 1 (PGL1)*





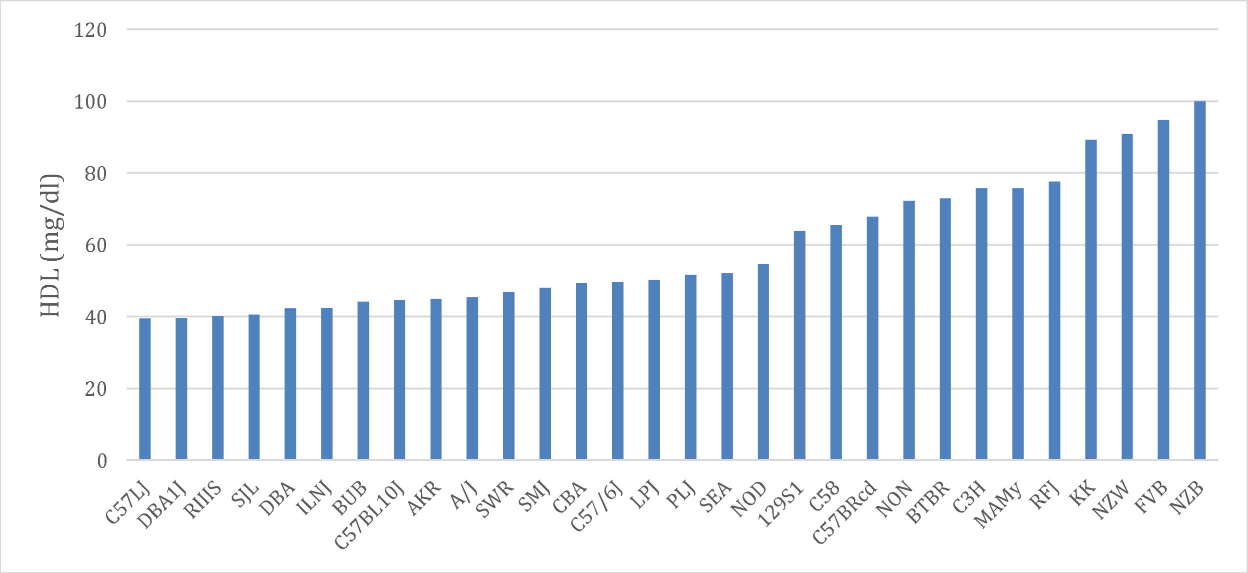
Gene	P-value	Genetic Effect	Haplotype	Chr.	Position
Nlrp1b *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦	11	71,159,763-71,159,873
Nlrp1a *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟩	11	71,110,959-71,111,942
Scimp *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟩🟦🟦🟩🟦🟦🟩🟦	11	70,814,424-70,814,816
Rabep1 *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟩🟩🟧🟩🟦🟧🟦🟦🟧🟦	11	70,840,103-70,840,777

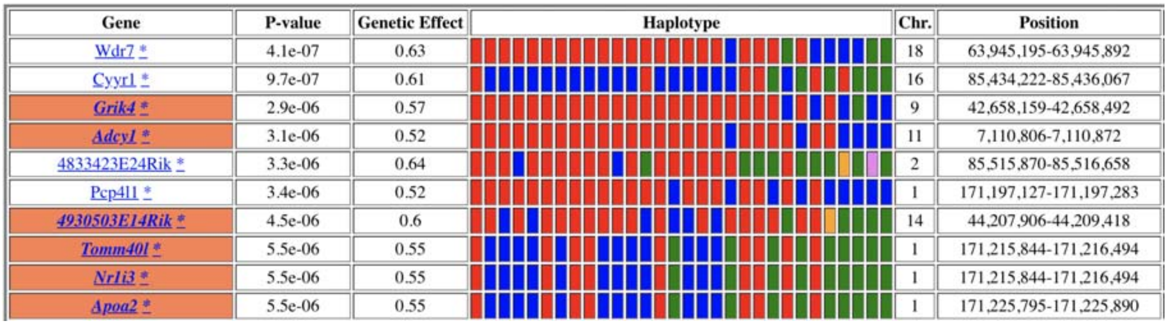
Figure 1. **Top**: The susceptibility of inbred strains to the lethal factor produced by *Bacillus anthracis*. Macrophages isolated from each indicated strain were incubated lethal toxin, and their survival was measured as described [39]. Strains with a "blue bar" (100% survival) are resistant to the toxin, while strains without a bar (100% lethality) were susceptible. **Bottom:** HBCGM analysis identifies the genes whose allelic patterns were most highly correlated with toxin susceptibility. The four co-linear genes with haplotype blocks that correlated with the phenotypic response pattern are indicated by their symbol; and an orange, white, or blue background indicates whether a SNP causes or does not cause an amino acid change, or if it affects a splice site, respectively. The haplotypic pattern is shown as colored rectangles that are arranged in the same order as the input data (shown in the top graph). Strains with the same colored rectangle have the same haplotype within the haplotype block within the indicated gene. The p-values and genetic effect size were calculated as previously described [32].



Dr1 *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦	5	108,271,081-108,273,796
Pde6b *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦	5	108,399,551-108,400,383
Atp5k *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦	5	108,430,104-108,442,472
Pcgf3 *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦	5	108,455,356-108,458,290

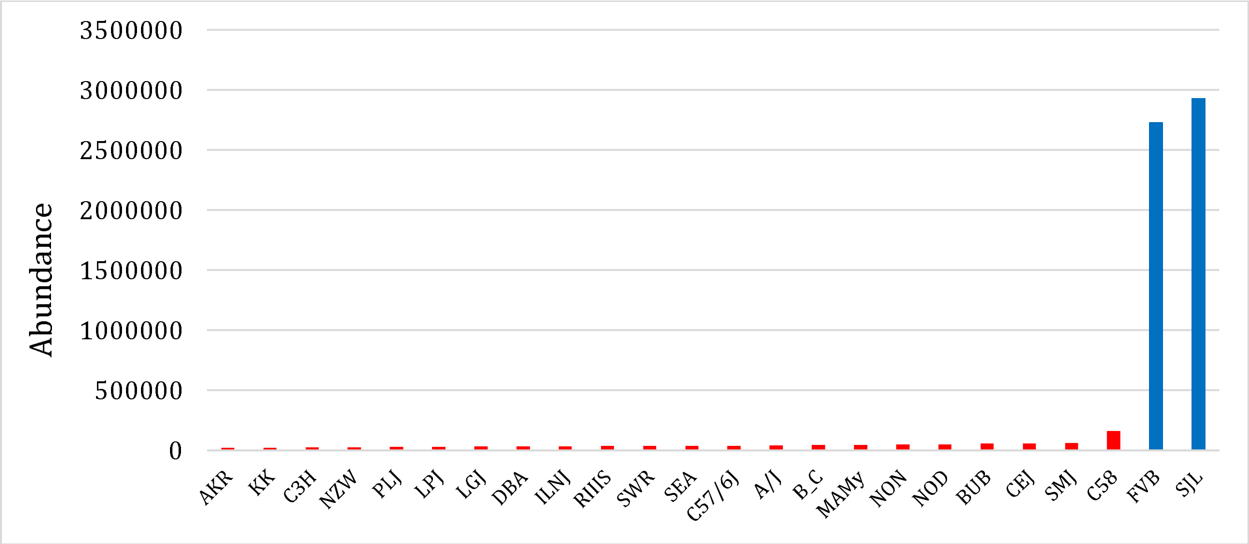
Figure 2. Top: The incidence of retinal degeneration in 29 inbred strains. Eight strains had significant retinal degeneration in all male and female mice (blue bars) examined, while 21 strains (indicated by the absence of a bar) had normal retinas. A bar with a value of 1 indicates all mice of that strain had retinal degeneration. **Middle:** HBCGM identified 33 genes with haplotype blocks whose allelic pattern was perfectly correlated with the retinal phenotype. However, only the four genes expressed in the retina are shown here. The genes within the correlated haplotype blocks are indicated by their symbol; and an orange, or white background indicates whether a SNP caused a significant or no amino acid change, respectively. The haplotypic pattern is shown as colored rectangles that are arranged in the same order as the input data shown above. Strains with the same colored rectangle have the same haplotype within the block. The p-values and genetic effect size were calculated as previously described [32]. **Bottom**: The domain structure of the Pde6b protein, and the location of its two GAF and the esterase domains are shown. The relative position of a SNP allele with a stop codon within the 2^{nd} GAF domain is indicated. The 8 strains with retinal degeneration all had the stop codon at position 347, while the 21 strains with normal retinas had the *Tyr347* allele.

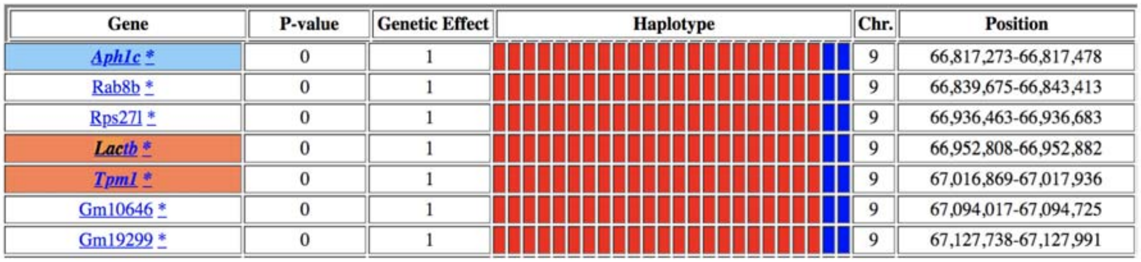




Gene	P-value	Genetic Effect	Haplotype	Chr.	Position
Wdr7 *	4.1e-07	0.63	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟥🟥🟥🟩🟥🟦🟦🟦🟦🟩🟩	18	63,945,195-63,945,892
Cyyr1 *	9.7e-07	0.61	🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟦🟦🟦🟦🟦🟦🟥🟥🟩🟦🟩🟥🟩🟥🟩🟩🟩	16	85,434,222-85,436,067
Grik4 *	2.9e-06	0.57	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟥🟦🟥🟦🟩🟦🟦	9	42,658,159-42,658,492
Adcy1 *	3.1e-06	0.52	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟥🟥🟥🟥🟦🟥🟥🟦🟦🟦🟦	11	7,110,806-7,110,872
4833423E24Rik *	3.3e-06	0.64	🟥🟥🟥🟦🟥🟥🟥🟥🟥🟥🟦🟥🟩🟥🟥🟥🟥🟥🟥🟩🟩🟩🟥🟩🟩🟩🟧🟩🟪🟩	2	85,515,870-85,516,658
Pcp4l1 *	3.4e-06	0.52	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟥🟥🟥🟦🟥🟥🟦🟥🟦🟥🟦🟦🟦🟦🟦	1	171,197,127-171,197,283
4930503E14Rik *	4.5e-06	0.6	🟥🟥🟦🟥🟦🟥🟥🟥🟥🟥🟥🟥🟦🟥🟦🟦🟥🟦🟥🟥🟥🟥🟩🟥🟥🟧🟩🟩🟩🟩	14	44,207,906-44,209,418
Tomm40l *	5.5e-06	0.55	🟥🟦🟦🟦🟦🟥🟦🟥🟥🟦🟦🟦🟦🟥🟩🟦🟦🟦🟩🟥🟥🟩🟥🟩🟥🟩🟩🟩🟩🟩	1	171,215,844-171,216,494
Nr1i3 *	5.5e-06	0.55	🟥🟦🟦🟦🟦🟥🟦🟥🟥🟦🟦🟦🟦🟥🟩🟦🟦🟦🟩🟥🟥🟩🟥🟩🟥🟩🟩🟩🟩🟩	1	171,215,844-171,216,494
Apoa2 *	5.5e-06	0.55	🟥🟦🟦🟦🟦🟥🟦🟥🟥🟦🟦🟦🟦🟥🟩🟦🟦🟦🟩🟥🟥🟩🟥🟩🟥🟩🟩🟩🟩🟩	1	171,225,795-171,225,890

Figure 3. **Top:** Plasma HDL levels were measured in female mice of 30 inbred strains maintained on a high-fat diet for 17 weeks. Each bar is the average of the plasma HDL ± SEM (mg/ml) measured in female mice (n=5-14 mice per group) of the indicated strain. **Bottom:** The HBCGM program identified the ten genes whose allelic pattern was most highly correlated with plasma HDL levels. The genes within the correlated haplotype blocks are indicated by their symbol; and an orange or white background indicates whether SNPs cause a significant or no amino acid change, respectively. The haplotypic pattern is shown as colored rectangles that are arranged in the same order as the input data (shown in the graph above). Strains with the same colored rectangle have the same haplotype within the block. The p-values and genetic effect size were calculated as previously described [32]. Of note, since the calculated genetic effect size for the *Apoa2* alleles was 0.55, other genetic factors could also affect the plasma HDL levels in murine strains.





Gene	P-value	Genetic Effect	Haplotype	Chr.	Position
***Aph1c* ***	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	66,817,273-66,817,478
Rab8b *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	66,839,675-66,843,413
Rps27l *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	66,936,463-66,936,683
***Lactb* ***	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	66,952,808-66,952,882
***Tpm1* ***	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	67,016,869-67,017,936
Gm10646 *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	67,094,017-67,094,725
Gm19299 *	0	1	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦	9	67,127,738-67,127,991

Figure 4. **Top:** Hepatic succinylcarnitine levels were measured in 16-week-old male mice after a 16-hour fast in 24 inbred strains. Each bar is the average succinylcarnitine level + SEM (shown as the abundance determined by mass spectroscopy) for each strain. **Bottom:** HBCGM analysis identified 152 genes whose allelic patterns were highly correlated with the succinylcarnitine level. However, only 3 of these genes (colored gene symbol background) were expressed in the liver and had a SNP causing change in the predicted amino acid sequence. The genes within correlated haplotype blocks are indicated by their symbol: an orange, white, or blue background indicates whether a SNP does or does not cause an amino acid change, or if it affects a splice site, respectively, is present. The haplotypic pattern is shown as colored rectangles that are arranged in the same order as the input data. Strains with the same colored rectangle have the same haplotype within the block. The p-values and genetic effect size were calculated as previously described [32].

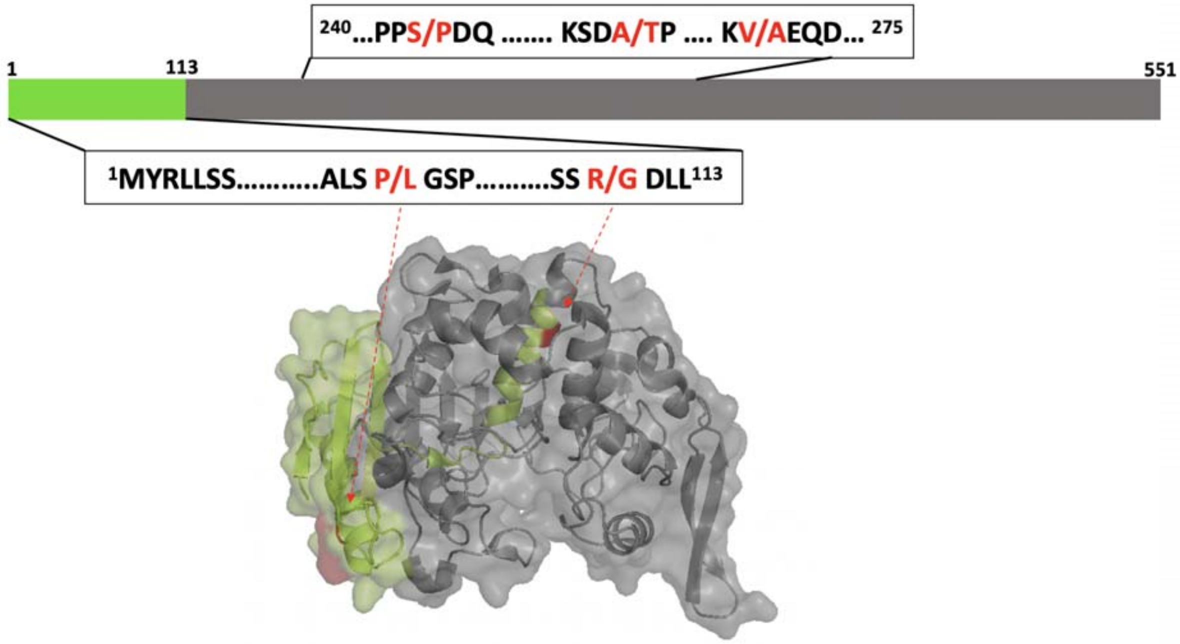

240...PPS/PDQ .... KSDA/TP .... KV/AEQD... 275
1
113
551
1MYRLLSS...........ALS P/L GSP..........SS R/G DLL113

Figure 5. Top: Lactb protein domains and the sites of the SNP alleles. The green region is the 113- amino acid mitochondrial targeting sequence (MTS), and the grey region is the mature protein with the Lactb domain. The box below shows the position of two murine SNPs (*Pro88Leu* and *Arg110Gly*) within the MTS where the two strains with high succinylcarnitine levels (SJL, FVB) share unique alleles that are not present in other strains. The box above shows the position of three SNPs (*Ser247Pro, Ala266Thr, Val217Ala*) within the mature Lactb protein where the two strains (SJL, FVB) with high succinylcarnitine levels share unique alleles (*247Pro, Thr266, Ala271*). **Bottom**: A protein structural model of the Lactb protein was produced using I-TASSER [99]. The green region is the NH2-terminal MTS, and the gray region shows the structure of the mature protein. The positions of the two SNP sites within the MTS are shown in red.

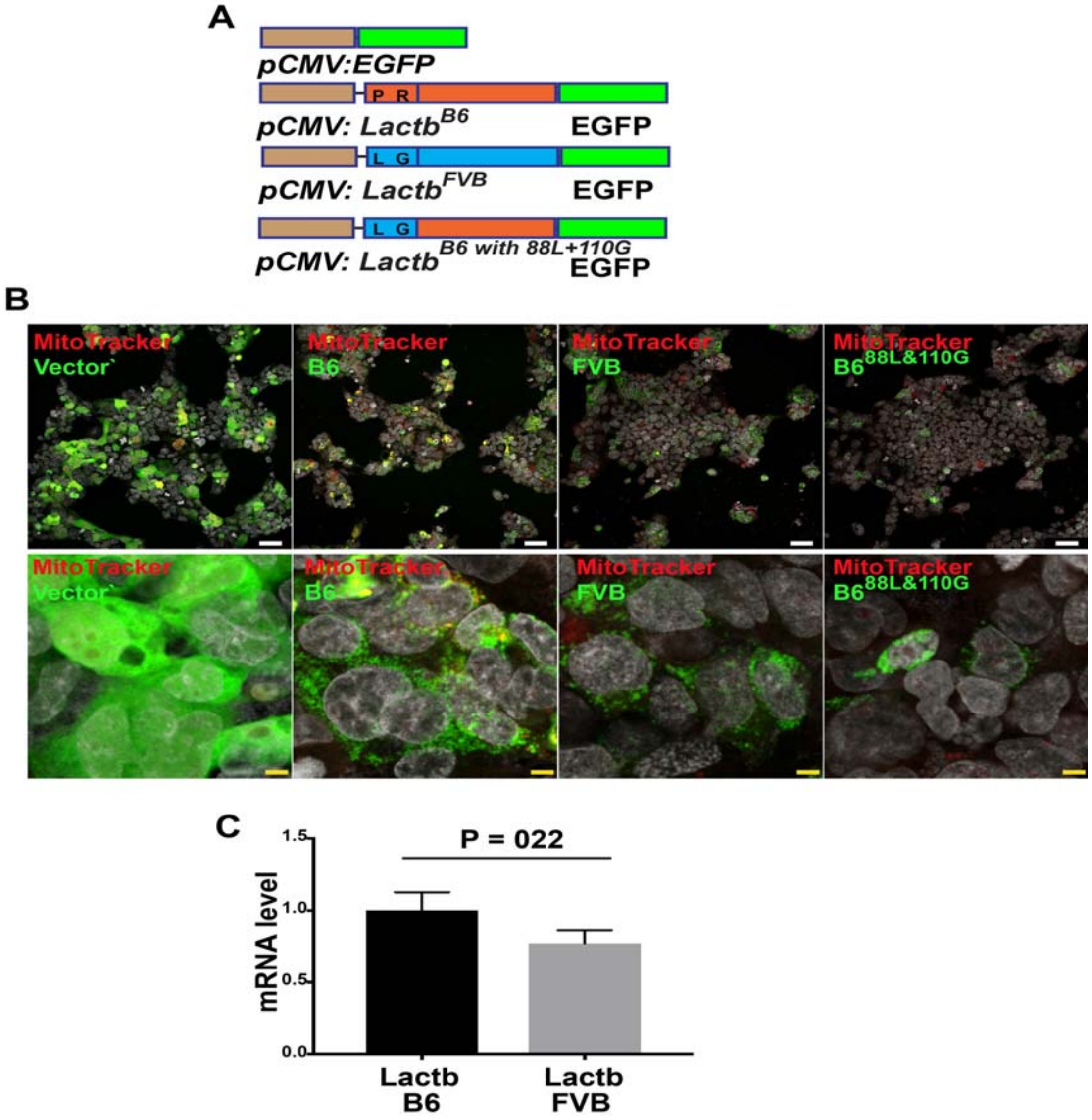


Figure 6. (**A**) Diagram of murine Lactb-EGFP fusion proteins. The control vector expresses an EGFP cDNA from a CMV promoter. The Lactb-EGFP fusion proteins were prepared from: (i) C57BL/6 or (ii) FBV *Lactb* mRNAs; or from a (iii) C57BL/6 *Lactb* mRNA with FVB alleles engineered at two positions (*88L*, *110G*) within its MTS. (**B**) Confocal images obtained 24 hours after 293T cells were transfected with the plasmids shown in (A) indicate the different levels of expression and sub-cellular localization of the fusion proteins. For each construct, low magnification images are shown in the upper row (scale bar 50 um), and higher magnification images of a region within the upper panel (single cell level) are shown in the lower row (scale bar 5 um). Mitochondria are stained red, and the green color indicates the Lactb fusion protein

expression. In contrast to the diffuse cytoplasmic expression pattern of the control EGFP protein, the C57BL/6 Lactb-EGFP fusion protein is expressed at a high level in a punctate pattern that overlapped with mitochondria. The FVB Lactb-EGFP protein was also expressed in a punctate pattern, but at a much lower level than the C57BL/6 Lactb-EGFP fusion protein. Also, the level and pattern of expression of the C57BL/6 $^{88L\ 110G}$ Lactb fusion protein resembled that of the FVB Lactb-EGFP fusion protein. This result is representative of 3 independently performed experiments. (**C**) RT-PCR measurement of the level of *Lactb-EGFP* mRNA expression in 293T cells 24 hrs after transfection with the plasmids shown in (A). Each bar is the average ± SE of 3 independent measurements. Despite the difference in the level of protein expression, there was no significant difference between the level of *C57BL/6* and *FVB Lactb-EGFP* mRNA expression (p=0.22).