Article

The Effect of Population Structure on Murine Genome-Wide Association Studies

Meiyue Wang and Gary Peltz*

Department of Anesthesia Stanford University School of Medicine, Stanford CA 94305

*Address Correspondence to: gpeltz@stanford.edu 300 Pasteur Dr. Room L232 Stanford, CA 94305.

Abstract

Population structure (PS) has been shown to cause false positive signals in genome-wide association studies (GWAS). Since PS correction is routinely used in human GWAS, it was assumed that it should be utilized for murine GWAS. Nevertheless, there are fundamental differences between murine and human GWAS, and the impact of PS on murine GWAS results has not been thoroughly investigated. We examined 8223 datasets characterizing biomedical responses in panels of inbred mouse strains to assess the impact of PS on murine GWAS. Surprisingly, we found that PS had a minimal impact on datasets characterizing responses in $\leq$20 strains; and relatively little impact on the majority of datasets characterizing >20 strains. Moreover, there were examples where association signals within known causative genes could be rejected if PS correction methods were utilized. PS assessment should be carefully used, and considered in conjunction with other criteria, for assessing the candidate genes that are identified in murine GWAS.

Abbreviations: GWAS, genome-wide association study; HBCGM, haplotype-based computational genetic mapping; PCA, principal component analysis; PS, population structure.

Introduction

Because of ancestral relatedness among the individuals within an analyzed population, a GWAS will identify a true causative genetic variant along with multiple other false positive associations, which arise because of genetic regions that were commonly inherited within a sub-population. This property, which is referred to as 'population structure' (PS) has been shown to exist in populations ranging from plants to humans ; and it inflates the number of false positive results obtained in a GWAS. Since PS was identified as a significant confounding factor for human GWAS, methods were developed to distinguish the false positive PS-based associations from the true causative genetic factors for a studied trait. Yu et al proposed a unified mixed model method to control the PS by a matrix (Pritchard, et al. 2000), where Bayesian clustering is used to infer the number of subpopulations and to estimate the effect of population structure. Improved methods were subsequently developed over the next decade. Zhao et al replaced the matrix with the use of principal components that summarized the genome-wide patterns of relatedness . Principal components analysis (PCA) has two advantages over the population structure matrix: (i) the subpopulations do not have to be specified prior to the analysis, which can be an arbitrary process that introduces errors; and (ii) it is far more computationally efficient, which is important when a large number of individuals with many SNPs are evaluated.

Although the methodology has improved, we do not know whether PS has a significant impact on GWAS results using inbred mouse strains. Mouse is the premier model organism for biomedical discovery, and many therapies were initially discovered using mice. Since the inbred laboratory strains are derived from what is estimated to be four ancestral founders that diverged ~1 million years ago (Guenet and Bonhomme 2003; Reuveni, et al. 2010), PS could certainly impact murine GWAS results. Others have advocated that PS correction should be used in murine GWAS . However, since murine and human GWAS differ in several fundamental ways, PS correction methods that are helpful for analyzing human GWAS results may not be as useful for murine GWAS studies. In murine GWAS, the inbred strains are homozygous and do not inter-breed; the number of inbred strains characterized are orders of magnitude lower than the many thousands of human subjects that are now routinely evaluated in a human GWAS; and genetic effect sizes are often much larger in a murine GWAS due to the fact that environmental and other variables are tightly controlled. Therefore, we used a large database of phenotypic responses that were measured in panels of inbred strains to examine the impact of PS on GWAS outcome. Haplotype-based computational genetic mapping (HBCGM) is a method for performing GWAS in mice, but it has important differences from conventional SNP-based studies . 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 (Liao, Wang, Guo, Allard, Chang, et al. 2004; Wang, et al. 2005; Zheng, et al. 2012). HBCGM analyses use a SNP database that now has 25M SNPs with alleles covering 49 inbred mouse strains, which was generated from analysis of whole genome sequence data . HBCGM has successfully identified genetic factors for >22 biomedical traits in mice , a; Zheng, et al. 2012). However, as with other GWAS methods, HBCGM analyses identify many genomic regions with allelic patterns that correlate with a phenotypic response pattern; but only one (or a few) contains a causative genetic factor. Therefore, we investigated the effect of PS on murine GWAS results, and the utility of applying a PS association test for eliminating false positives from the list of candidate genes identified by HBCGM.

Results

The Mouse Phenome Database (**MPD**) (https://phenome.jax.org) (Grubb, et al. 2014) has 8223 datasets that characterize basal, age-related, and experimentally-induced responses in panels of inbred mouse strains. A total of 1.52M different measured responses were within this database. We previously demonstrated that MPD datasets have utility for genetic discovery; a genetic susceptibility factor for a drug-induced CNS toxicity was identified by HBCGM analysis of one MPD dataset (Zheng, Zhang, Dill, Clark, Tu, Yablonovitch, Tan, Zhang, Rujescu, Wu, Tessarollo, Vieira, Gottesman, Deng, Eberlin, Zare, Billard, Gillet, L, et al. 2015). Therefore, we initially selected all MPD datasets that measured a response in 10 or more strains whose genomic sequence was available (2435 datasets) For each of these datasets, haplotype blocks with allelic patterns that correlated with the measured strain response pattern were identified by HBCGM. The average number of correlated blocks ($p_{HBCGM} < 0.01$) for each dataset was 3966. We then used a multi-variate association test (MANOVA) to determine whether the haplotypic strain groupings within the correlated blocks were related to PS among the analyzed strains. To do this, the number of PCs has to be specified in advance to perform the PS association test. Therefore, we first examined the percentage of the variance that was explained when a variable number of principal components (PCs), which ranged from 1 to 33 because ≤33 inbred strains were analyzed in any dataset, were used for the principal component analysis (PCA). The results indicated that four PCs, each of which explained >5% of the variance, were optimal for PS assessment (**Fig. S1**).

A pairwise IBS matrix divided the 49 sequenced inbred strains into four sub-populations (**Table 1, Fig. 1**). Sub-populations 2 and 3 contain the majority of the inbred strains, which are closely related. Sub-population 1 strains are derived from a C57BL ancestor; and the five strains in sub-population 4 are genetically distinct from the other groups. The spatial relationship of the 49 strains (plotted using the first two PCs for each strain) is concordant with the hierarchical clustering (**Fig 1)**. Populations 2 and 3 are quite similar and can be merged if the small amount of genetic variation is ignored. The number of inbred strains analyzed in each of the 2435 MPD datasets are summarized in **Table S1**. Our analyses indicated that we could not assess population structure in MPD datasets that analyzed $\leq$ 20 strains because: the sub-population structure was extremely variable, the strain groupings within these datasets often contained strains from different global sub-groups, and the sub-structure could be significantly altered by the addition or deletion of even a single strain (**Fig. S2**).

Therefore, we examined population sub-structure in 1750 MPD datasets that examined responses in > 20 inbred strains. During our analysis, we noted that many different MPD datasets used the same panel of inbred strains, which is probably because multiple phenotypes were evaluated by the same investigator and certain strains are commonly used by different laboratories. To illustrate the general properties that emerged from our analyses, we show 967 MPD datasets that repeatedly analyzed responses in the same sets of (n=23-32) inbred strains. The strains used in 432 of these datasets clearly lack population sub-structure (**Fig. S3**). PS was present in 535 other MPD datasets (**Fig. S4**), where the group 1 strains (C57BL related) are clearly separated from the other strains. However, the global group 2 and group 3 strains are broadly distributed in the graphs of these datasets, without an explicit boundary that separated them from the other strains; and the groups 2 and 3 strains are intermixed with group 1 strains in many of these graphs. Hence, even in datasets that examine responses in strains that appear to have PS, population structure would only have an effect if phenotypic response pattern completely mirrored that of the global strain sub-populations. If this response pattern does not occur, which appears to be the case for the majority of the measured responses (see below), PS would have a limited effect on genetic analysis results.

To more directly assess PS impact on the haplotype blocks generated by HBCGM analysis of the 2435 MPD datasets, a population structure association test was performed on each correlated haplotype block. An adjusted p-value for the PS association test for each block was generated using MANOVA. Blocks with a $p_{adj} \leq 0.05$ have a significant association with population structure (i.e. PS$^+$), and could be removed from further consideration, while those with a $p_{adj} \geq 0.05$ are viewed as viable candidate genes for further evaluation (PS$^-$). For 68% of the datasets (1,660 of 2435 analyzed), >50% of the correlated blocks were not associated with population structure (PS$^-$); and 39% of the datasets (949 of 2435) had 75 to 100% PS$^-$ blocks (**Fig. 2**). Only 32% of the datasets (n=775) had >50% PS$^+$ correlated blocks; and most of these (23%, 565 datasets) have between 25 and 49% PS$^-$ blocks. Only 9% of the MPD datasets (n= 210) have >75% PS$^+$ blocks. Overall, our results indicate that for most MPD datasets, the vast majority of the haplotype blocks identified by HBCGM are not affected by PS. We also investigated whether the magnitude of the PS impact is affected by the number of strains analyzed (i.e. the sample size). As the strain number increased, the number of correlated candidate blocks identified by HBCGM analysis increased (**Fig. 3A**). This result is consistent with prior studies indicating that genetic analyses, which are performed on large populations, will identify additional genetic variants with a small effect size (Visscher, et al. 2017). However, while the number of PS$^-$ blocks plateaued after 15 strains were analyzed, the number of PS$^+$ blocks increased as the number of analyzed strains increased (**Figs. 3B-C**). These results indicate that when an increased number of inbred strains are analyzed, the number of correlated haplotype blocks increases, as does percentage of PS$^+$ blocks. The results are completely consistent with the sample size effects that were previously noted in human-case control studies .

When considering whether PS correction should be utilized for assessing mouse genetic association results, it is important to determine whether this could lead to rejection of a true causative association. Therefore, we investigated whether PS was present in haplotype blocks within genes whose allelic patterns are known to be causal for phenotypic response differences in 6 MPD datasets (**Table 2)**. We first examined the haplotype blocks identified by HBCGM from analysis of data on strain susceptibility to anthrax toxin (MPD 1501), which is known to be caused by allelic variation within the *Nalp1a* and *Nalp1b* genes (Boyden and Dietrich 2006). Both of the correlated haplotype blocks within these genes were PS$^-$. Similarly, the identified haplotype block within an experimentally validated causative gene (*Abcb5*) (Zheng, Zhang, Dill, Clark, Tu, Yablonovitch, Tan, Zhang, Rujescu, Wu, Tessarollo, Vieira, Gottesman, Deng, Eberlin, Zare, Billard, Gillet, L, et al. 2015) affecting susceptibility to a drug (haloperidol)-induced CNS toxicity (MPD 39410) also was PS$^-$. The albino skin type (MPD 22001) that appears in some inbred strains is determined by a *Cys103Ser* SNP within the tyrosinase (*Tyr*) gene (Yokoyama, et al. 1990), and the correlated haplotype block identified by HBCGM analysis within *Tyr* was also PS$^-$.

However, the results of PS analyses for three other MPD datasets raised concerns. *Apoa2* encodes the second most abundant protein within high density lipoprotein (HDL) particles, and it is involved in lipoprotein metabolism. *Apoa2* alleles were previously associated with differences in plasma HDL cholesterol levels in mice (Doolittle, et al. 1990); and HDL levels were 70% decreased in *Apoa2* knockout mice (Weng, et al. 1999). HBCGM analysis of two datasets measuring HDL cholesterol levels (MPD 9904 and 9907) indicated that 3 of 4 correlated haplotype blocks within *Apoa2* are PS$^+$ blocks (MANOVA $p_{adj} < 0.05$). Retinal degeneration in inbred strains is known to be caused by a stop codon (*Tyr347X*) within *phosphodiesterase 6b* (*Pde6b*) (Pittler, et al. 1993). One MPD dataset (MPD 26721) examined the retinas of 29 inbred strains: 21 strains had normal retinas, and 8 strains had retinal degeneration. HBCGM analysis identified two *Pde6b* haplotype blocks that completely correlated with retinal degeneration in male and female mice ($p_{HBCGM} = 0$). However, the strain groupings within these blocks had PS; the PS association test p-values for these blocks were 0.02 ($p_{adj} = 0.049$) (Table 2). The blocks had PS because all 8 strains with retinal degeneration were from population group 3, and all population group 1 and 2 strains had normal retinas. However, several group 3 strains had normal retinas and *Pde6b Try347* alleles (**Fig. 4**). These examples demonstrate that some true positives, if the usual FDR control rate ($q = 0.05$) was applied, could have been falsely rejected based upon their association with PS.

Discussion

PS correction is commonly performed when analyzing GWAS results involving human or other species (cattle, maize, etc.), and PS correction has also been advocated for use in murine GWAS (Sul, et al. 2018). While PS correction helps to eliminate false positives, our analyses indicate that PS makes a smaller than expected contribution to most murine GWAS studies. Moreover, we found that PS correction can even generate a false negative result, i.e. it can lead to rejection of an experimentally confirmed true causative genetic factor. *Why is the utility of PS correction in murine HBCGM analyses different from that of association studies performed using* *different methods or involving other species?* We identify four factors that account for this difference. (i) A very limited number of inbred strains are examined in a murine GWAS, which usually analyze <20 (and never >40 inbred strains). This is orders of magnitude less than the number of subjects in human GWAS (now ranging from thousands to hundreds of thousands). Since the PS effect increases as the number of inbred strains analyzed are increased, PS has a more limited effect on most murine GWAS. (ii) We found that the vast majority of murine GWAS studies utilize strains with limited PS. Most (37 of 49 or 75%) of the commonly used inbred strains are derived from closely related populations, which have limited or no population structure. Among 25M SNPs that were analyzed, pairwise comparisons revealed that the level of allelic similarity among the classical inbred strains is >70%. The limited amount of genetic variation among these strains precludes their separation into distinct sub-populations. (iii) Human (and other species) GWAS identify trait associations using SNP markers, and the association signals depend upon the existence of linkage disequilibrium (LD) between SNP markers and causal genetic variants. The dependence upon LD, which extends over a region of indeterminate size, increases the effect that regional PS could have on an outcome. In contrast, HBCGM does not rely on LD between marker and causative SNPs. HBCGM uses a combination of adjacent SNPs to produce haplotype blocks, which are the composite genetic variants that are analyzed. Since haplotype blocks are assembled from analysis of whole genome sequence, the block boundaries are precisely determined, and the analyzed variants contain the causative genetic factors. (iv) The impact of a false negative result (excluding a true positive due to PS) is much greater for a murine GWAS. Genetic association studies involving large populations usually identify many genetic variants, with each having a small genetic effect size. In those situations, the loss of a few true positive associations does not create a large problem since many others remain. However, murine GWAS analyze a small number of inbred strains; and the heritability and genetic effect size of the identified candidate genes is relatively large (usually >0.3); since the mouse genome is homozygous and environmental and other confounding factors are minimized. Thus, unlike its small effect on human GWAS results, the elimination of a true positive by PS correction can be disastrous for a murine GWAS. Of note, only one factor is specific to HBCGM, the other three factors are relevant to all forms of murine GWAS.

We have shown in several situations that a true causative factor could be associated with strain phylogenetic background. In two examples, *Apoa2* (MPD 9904) and *Pde6b* (MPD 1501), PS correction could have removed true causative blocks from further consideration. However, in one case (retinal degeneration and *Pde6b)*, the identified haplotype block was much more strongly associated with the phenotypic response pattern (genetic association p-value=0) than with population sub-structure (p-value=0.49). In another case (HDL levels and *ApoA2*), the p-values for the genetic and the population structure association tests for the causative haplotype block were of a similar magnitude, but published information indicated that the gene candidate was very strongly associated with the analyzed phenotype. As suggested by others after examining GWAS results for multiple traits in plants , it is not easy to distinguish between a true and a spurious association due to genetic background, even after correcting for PS. However, when GWAS are performed under conditions with true genome wide coverage, a true association is expected to exhibit the strongest association . Allele sharing within a localized candidate genomic region should be greater than one based upon genome wide allelic correlations. Thus, examining the ratio of the p-values obtained from the GWAS and PS association tests could provide a more informative way to eliminate spurious positives while retaining the true positive associations. Nevertheless, as was previously observed in plants , there are situations (as with HDL and *ApoA2*) where a shared stain background can be responsible for trait response differences. In these situations, the strength of the functional evidence that a candidate gene could be responsible for a trait difference could override PS considerations.

Lastly, other methods can be used to eliminate false positive associations in GWAS. We have shown that true positives can be identified by the use of orthogonal criteria for analyzing HBCGM output. Causative genetic factors were selected from among the many genes with correlated genetic patterns using gene expression and metabolomic data (Liu, et al. 2010), curated biologic information (Zhang, et al. 2011), or the genomic regions delimited by prior QTL analyses (Smith, et al. 2008; LaCroix-Fralish, et al. 2009). This integrated approach evaluates genetic candidates using multiple criteria, and it can produce results that are superior to that of using a single highly stringent genetic criterion to identify gene candidates. Recent efforts to utilize transcriptome wide association studies or functional information to select causative loci from among the many SNP sites identified in a human GWAS, or to identify SNPs near *a priori* identified gene candidates in plant GWAS resemble our methods for analyzing HBCGM output. In summary, PS assessment may be one factor that should be used along with multiple other factors to assess a candidate gene, which include the relative strength of the GWAS and PS association results, tissue-specific gene expression criteria, and gene-phenotype relationship based upon information contained within the published literature.

Methods

Selection of Mouse Phenome Database datasets. MPD datasets (n=8223) were downloaded on March 24, 2020. We analyzed MPD datasets where the mean phenotypic measurement of each strain was obtained from > 5 mice of each strain. An ANOVA test was also performed to determine if the inter-strain variance was significantly greater than intra-strain variances; and a p-value < $1x10^{-10}$ was used as the cutoff for dataset selection. Datasets with categorical measurements were excluded from bulk analysis of MPD datasets.

Haplotype block construction and genetic mapping in mice. The sequence of 49 inbred mouse strains were analyzed as previously described (Zheng, Zhang, Dill, Clark, Tu, Yablonovitch, Tan, Zhang, Rujescu, Wu, Tessarollo, Vieira, Gottesman, Deng, Eberlin, Zare, Billard, Gillet, Li, et al. 2015). SNPs were dynamically organized into haplotype blocks for each dataset, which only used alleles for the strains contained within the dataset, according to the “maximal” block construction method (Peltz, et al. 2011b). In brief, this method produces haplotype blocks with a minimum of 3 SNPs; and each block is only allowed to a predetermined number of haplotypes, which ranges from 2 to 5. Since the “maximal” method enables blocks to overlap, blocks are assembled that cover all possible allelic combinations within a specific genomic region. If a smaller block was nested inside of a larger block and it contained the same haplotypes, it was removed and the larger block was used to cover that region (Peltz, et al. 2011b). This ensures that additional SNPs are only included within a block if additional haplotypes are added to the block. The relationship between the phenotypic response pattern and haplotype blocks was evaluated by HBCGM as described (Liao, Wang, et al. 2004b). Genes with correlated haplotype blocks were sorted based upon the ANOVA p-value. A cut-off of $p = 0.01$ was used to select haplotype blocks with a correlated allelic pattern. If a gene had multiple correlated blocks, the haplotype block with the smallest p-value was used.

Population structure association test. We use principal component analysis (PCA) to determine whether a haplotypic strain grouping was associated with PS. Principal components (PC) has been used assess population stratification (Price, et al. 2006b; Zhao, et al. 2007a); it is a major component of the linear mixed model (LMM) that is used to control PS-induced spurious associations in GWAS results. In the LMM, PS is treated as a covariate that influences the phenotypic values in addition to the effect of the genetic markers. However, we treat PS as a dependent variable, which is determined by a comprehensive analysis of genome-wide allelic similarity. For this analysis, the PS of the inbred strains (y) is determined by the equation

$$y = \mu + X\beta + e$$

where y is an $n \times p$ matrix that is derived from a PCA of sample size of *n* with p principal components; μ is an $n \times 1$ vector that contains the grand mean for each of the p variables ($\mu = \frac{1}{n}\sum_{i=1}^{n} y_i$); X is an $n \times 1$ vector of haplotype indicators for n strains; β is the effect of the haplotype, and e is an $n \times 1$ vector of the residual error. p is a hyperparameter to determine the number of PCs used in analysis, where it guarantees each PC can explain certain amount (say >5%) of the variance of the original genetic relationship. Alternatively, p can be arbitrarily selected based upon analysis on a Scree plot (to find the “elbow”), which ranks PCs based on the percentage of variance explained by each PC. If the elbow is observed at *p*-th PC; most of the true signals are captured in the first *p* PCs. By using PC to represent population structure, pre-determination of the number of sub-populations is not required. A multivariate analysis of variance (MANOVA) can be then used to assess the association between strain groupings within a haplotype block and PS, since the strain grouping within a block becomes a single variable that affects the first *p* PCs.

Generation of genetic relationship and identity-by-state similarity matrices. The genetic relationship matrix (GRM) for inbred mouse strains was generated using genome-wide SNP alleles and GCTA software (Yang, et al. 2011). The GRM is also known as the variance-covariance standardized relationship matrix, and the eigenvectors of this matrix were used as PC. The GRM eigenvalues for the inbred strains of each PC were used to estimate the amount of GRM variance that PC explains. Since we analyze 49 inbred strains and mice are homozygous, SNPs were not filtered based upon a minor allele frequency threshold. To verify that the PCs effectively represent the PS among the strains, we clustered individual strains using a pairwise identity-by-state (IBS) similarity matrix, which was also derived using whole genome SNP data. The IBS similarity matrix is a square, symmetric matrix that reflects the IBS distance between all pairs of inbred mouse strains. PLINK (Purcell, et al. 2007) was used to calculate the IBS similarity matrix, and it contains values that range from 0 to 1. The hierarchical clustering of n strains was determined using the hcut() function within the factoextra/R package.

Multiple test correction for the PS association test. Since the population structure association test was performed on 2435 datasets, the MANOVA test p-value for each block generated by the HBCGM program is adjusted by controlling for the false discovery rate (FDR at $q = 0.05$)

using Benjamini-Hochberg method (Benjamini and Hochberg 1995). The adjusted p-value for i-th block is $p_{adj} = p_i \times m/i$, where p_i is the MANOVA test p-value, *m* is the number of blocks (multiple tests), and i is the order of p_i in a series of p-values that satisfies $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. If a block has $p_{adj} \geq 0.05$, it is not considered as having a significant amount of PS (i.e. the null hypothesis, which is that the tested block does not have population structure, cannot be rejected).

Data availability: The data sets within the Mouse Phenome Database (**MPD**) analyzed in this study are available at (https://phenome.jax.org). All data generated or analyzed during this study are included in this published article [and its supplementary information files].

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

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

Author contributions: G.P. and M.W. wrote the paper. M.W. performed the analyses of the data.

References

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627-631.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological) 57:289-300.
Boyden ED, Dietrich WF. 2006. Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin. Nat Genet 38:240-244.
Chu LF, Liang D-Y, Li X, Sahbaie P, D'Arcy N, Liao G, Peltz G, Clark JD. 2009. From Mouse to Man: The 5-HT3 Receptor Modulates Physical Dependence on Opioid Narcotics. Pharmacogenetics and Genomics 19:193-205.
de Leeuw CA, Neale BM, Heskes T, Posthuma D. 2016. The statistical properties of gene-set analysis. Nat Rev Genet 17:353-364.
Donaldson R, Sun Y, Liang D-Y, Zheng M, Sahbaie P, Dill DL, Peltz G, Buck KJ, Clark JD. 2016. The multiple PDZ domain protein Mpdz/MUPP1 regulates opioid tolerance and opioid-induced hyperalgesia. BMC Genomics 17.
Doolittle MH, LeBoeuf RC, Warden CH, Bee LM, Lusis AJ. 1990. A polymorphism affecting apolipoprotein A-II translational efficiency determines high density lipoprotein size and composition. J Biol Chem 265:16380-16388.
Grubb SC, Bult CJ, Bogue MA. 2014. Mouse phenome database. Nucleic Acids Res 42:D825-834.
Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. 2001. In silico mapping of complex disease-related traits in mice. Science 292:1915-1918.

Guenet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet 19:24-31.
Guo YY, Liu P, Zhang X, Weller PMM, Wang J, Liao G, Zhang Z, Hu J, Allard J, Shafer S, et al. 2007. In vitro and In silico Pharmacogenetic Analysis in Mice. Proceedings of the National Academy of Sciences 104:17735-17740.
Guo YY, Weller PF, Farrell E, Cheung P, Fitch B, Clark D, Wu SY, Wang J, Liao G, Zhang Z, et al. 2006. In Silico Pharmacogenetics: Warfarin Metabolism. Nature Biotechnology 24:531-536.
Hammerschlag AR, de Leeuw CA, Middeldorp CM, Polderman TJC. 2019. Synaptic and brain-expressed gene sets relate to the shared genetic risk across five psychiatric disorders. Psychol Med:1-11.
Hu Y, Liang D, Li X, Liu H-H, Zhang X, Zheng M, Dill D, Shi X, Qiao Y, Yeomans D, et al. 2010a. The Role of IL-1 in Wound Biology Part I: Murine in Silico and In vitro Experimental Analysis. Anesthesia & Analgesia 111:1525-1533.
Hu Y, Liang D, Li X, Liu H-H, Zhang X, Zheng M, Dill D, Shi X, Qiao Y, Yeomans D, et al. 2010b. The Role of IL-1 in Wound Biology Part II: In vivo and Human Translational Studies. Anesthesia & Analgesia 111:1534-1542.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723.
LaCroix-Fralish ML, Mo G, Smith SB, Sotocinal SG, Ritchie JG, Austin JS, Melmed K, Schorscher-Petcu A, Laferriere AC, Lee TH, et al. 2009. The β3 Subunit of the Na+,K+-ATPase Affects Pain Sensitivity. Pain 144:294-302.
Liang D, Liao G, Wang J, Usuka J, Guo YY, Peltz G, Clark JD. 2006. A Genetic Analysis of Opioid-Induced Hyperalgesia in Mice Anesthesiology 104:1054-1062.
Liang DY, Zheng M, Sun Y, Sahbaie P, Low SA, Peltz G, Scherrer G, Flores C, Clark JD. 2014. The Netrin-1 receptor DCC is a regulator of maladaptive responses to chronic morphine administration. BMC Genomics 15:345.
Liao G, Wang J, Guo J, Allard J, Chang J, Nguyen A, Shafer S, Puech A, McPherson JD, Foernzler D, et al. 2004. In Silico Genetics: Identification of A Novel Functional Element Regulating H2-Ea Gene Expression Science 306:690-695.
Liao G, Wang J, Guo J, Allard J, Cheng J, Ng A, Shafer S, Puech A, McPherson JD, Foernzler D, et al. 2004a. In Silico Genetics: Identification of a Functional Element Regulating H2-Ea Gene Expression. Science 306:690-695.
Liao G, Wang J, Guo J, Allard J, Cheng J, Ng A, Shafer S, Puech A, McPherson JD, Foernzler D, et al. 2004b. In silico genetics: identification of a functional element regulating H2-Eα gene expression. Science 306:690-695.
Liu H-H, Lu P, Guo Y, Farrell E, Zhang X, Zheng M, Bosano B, Zhang Z, Allard J, Liao G, et al. 2010. An Integrative Genomic Analysis Identifies Bhmt2 As A Diet-Dependent Genetic Factor Protecting Against Acetaminophen-Induced Liver Toxicity Genome Research 20:28-35.
Liu HH, Hu Y, Zheng M, Suhoski MM, Engleman EG, Dill DL, Hudnall M, Wang J, Spolski R, Leonard WJ, et al. 2012. Cd14 SNPs regulate the innate immune response. Mol Immunol 51:112-127.
Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, Pasaniuc B. 2019. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet 51:675-682.
Peltz G, Zaas AK, Zheng M, Solis NV, Zhang MX, Liu H-H, Hu Y, Boxx GM, Phan QT, Dill D, et al. 2011a. Next-Generation Computational Genetic Analysis: Multiple Complement Alleles Control Survival After Candida Albicans Infection Infection and Immunity 79:4472-4479.
Peltz G, Zaas AK, Zheng M, Solis NV, Zhang MX, Liu H-H, Hu Y, Boxx GM, Phan QT, Dill D, et al. 2011b. Next-generation computational genetic analysis: multiple complement alleles control survival after Candida albicans infection. Infection and immunity 79:4472-4479.

Pittler SJ, Keeler CE, Sidman RL, Baehr W. 1993. 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 90:9616-9619.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006a. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006b. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38:904-909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. The American Journal of Human Genetics 67:170-181.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81:559-575.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4-16.

Ren M, Kazemian M, Zheng M, He J, Li P, Oh J, Liao W, Li J, Rajaseelan J, Kelsall BL, et al. 2020. Transcription factor p73 regulates Th1 differentiation. Nat Commun 11:1475.

Reuveni E, Birney E, Gross CT. 2010. The consequence of natural selection on genetic variation in the mouse. Genomics 95:196-202.

Rozzo SJ, Allard J, Choubey D, Vyse T, Izui S, Peltz G, Kotzin BL. 2001. Evidence for an interferon-inducible gene, Ifi202, in the susceptibility to Systemic Lupus. Immunity 15:435-443.

Smith SB, Marker CL, Perry C, Liao G, Sotocinal SG, Austin JS, Melmed K, David Clark J, Peltz G, Wickman K, et al. 2008. Quantitative trait locus and computational mapping identifies Kcnj9 (GIRK3) as a candidate gene affecting analgesia from multiple drug classes. Pharmacogenetics and Genomics 18:231-241.

Sorge RE, Trang T, Dorfman R, Smith SB, Beggs S, Ritchie J, Austin JS, Zaykin DV, Meulen HV, Costigan M, et al. 2012. Genetically determined P2X7 receptor pore formation regulates variability in chronic pain sensitivity. Nat Med 18:595-599.

Sul JH, Martin LS, Eskin E. 2018. Population structure in genetic studies: Confounding factors and mixed models. PLoS genetics 14:e1007309.

Tregoning JS, Yamaguchi Y, Wang B, Mihm D, Harker JA, Bushell ESC, Zheng M, Liao G, Peltz G, Openshaw PJM. 2010. Genetic Susceptibility to the Delayed Sequelae of RSV Infection is MHC-Dependent, but Modified by Other Genetic Loci. J. Immunology 185:5384-5391.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics 101:5-22.

Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. 2019. Opportunities and challenges for transcriptome-wide association studies. Nat Genet 51:592-599.

Wang J, Liao G, Usuka J, Peltz G. 2005. Computational Genetics: From Mouse to Man? Trends in Genetics 21:526-532.

Watanabe K, Umicevic Mirkov M, de Leeuw CA, van den Heuvel MP, Posthuma D. 2019. Genetic mapping of cell type specificity for complex traits. Nat Commun 10:3222.

Weng W, Brandenburg NA, Zhong S, Halkias J, Wu L, Jiang XC, Tall A, Breslow JL. 1999. 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 40:1064-1070.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88:76-82.

Yokoyama T, Silversides DW, Waymire KG, Kwon BS, Takeuchi T, Overbeek PA. 1990. Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. Nucleic Acids Res 18:7293-7298.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208.

Zaas AK, Liao G, Chein J, Usuka J, Weinberg C, Shore D, Giles D, Marr K, Burch L, Perara, et al. 2008. Plasminogen Alleles Influence Susceptibility to Invasive Aspergillosis. PLoS genetics 4:e1000101.

Zhang H, Zheng M, Wu M, Xu D, Nishimura T, Nishimura Y, Giffard RG, Xiaong X, Xu LJ, Clark JD, et al. 2016. A Pharmacogenetic Discovery: Cystamine Protects against Haloperidol-Induced Toxicity and Ischemic Brain Injury. Genetics 203:599-609.

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

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, et al. 2007a. An Arabidopsis example of association mapping in structured samples. PLoS genetics 3:71-82.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, et al. 2007b. An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4.

Zheng M, Dill D, Peltz G. 2012. A better prognosis for genetic association studies in mice. Trends Genet 28:62-69.

Zheng M, Zhang H, Dill DL, Clark JD, Tu S, Yablonovitch AL, Tan MH, Zhang R, Rujescu D, Wu M, et al. 2015. The role of Abcb5 alleles in susceptibility to haloperidol-induced toxicity in mice and humans. PLoS medicine 12:e1001782.

Zheng M, Zhang H, Dill DL, Clark JD, Tu S, Yablonovitch AL, Tan MH, Zhang R, Rujescu D, Wu M, et al. 2015. The Role of Abcb5 Alleles in Susceptibility to Haloperidol-Induced Toxicity in Mice and Humans PLOS Medicine 12:e1001782.

Table 1. The 49 inbred strains can be divided into the four groups shown in this table based on their pattern of genome-wide allelic sharing.

Group	Number of Strains	Strain List
1	7	C57BL/6J, B10, C57BL10J, C57BL6NJ, C57BRcd, C57LJ, C58
2	14	BTBR, CEJ, KK, NZB, NZW, 129P2, 129S1, 129S5, ILNJ, LPJ, NZO, PJ, SMJ, WSB
3	23	BUB, DBA1J, FVB, NON, NUJ, RFJ, RHJ, RIIIS, SJL, A/J, AKR, BALB, C3H, CBA, DBA, LGJ, MAMy, MRL, NOD, PLJ, SEA, ST, SWR
4	5	CAST, MOLF, PWD, PWK, SPRET

Table 2. Results of population structure analysis performed on haplotype blocks within the known causative genes for 6 MPD datasets. The MPD dataset number, the sex of the mice, and a description of the measured response are shown. The gene symbol for the causative gene, the chromosome and position of the identified haplotype block, and the p-value and adjusted p-value for the population structure association test for that block are shown.

MPD Dataset	Gene	Block position	HBCGM p-val	PS p-val	PS adj p val
1501 F susceptibility to Bacillus anthracis	*Nlrp1a*	Chr11: 71110959-71111942	0	0.9537	0.9702
1501 M susceptibility to Bacillus anthracis	*Nlrp1b*	Chr11: 71159763-71159873	0	0.5785	0.6782
26721 F retinal degeneration	*Pde6b*	Chr5: 108399551-108400383	0	0.0244	0.0491
26721 M retinal degeneration	*Pde6b*	Chr5: 108399551-108400383	0	0.0244	0.0491
9904 F HDL cholesterol baseline	*Apoa2*	Chr1: 171225795-171225890	5.5e-6	0.0005	0.0010
9904 M HDL cholesterol baseline	*Apoa2*	Chr1: 171225644-171225697	3.14e-5	0.1537	0.2448
9907 F HDL cholesterol after 17 wks on diet	*Apoa2*	Chr1: 171227457-171227593	0.0066	0.0039	0.0106
9907 M HDL cholesterol after 17 wks on diet	*Apoa2*	Chr1: 171227457-171227593	0.0008	0.0020	0.0044
22001 F coat color	*Tyr*	Chr7: 87446687-87446831	3.68e-6	0.1071	1
22001 M coat color	*Tyr*	Chr7: 87446687-87446831	3.68e-6	0.1071	1
39410 M Haloperidol induced latency day 30	*Abcb5*	Chr12: 118885164-118916966	4.19e-10	0.1507	0.8047

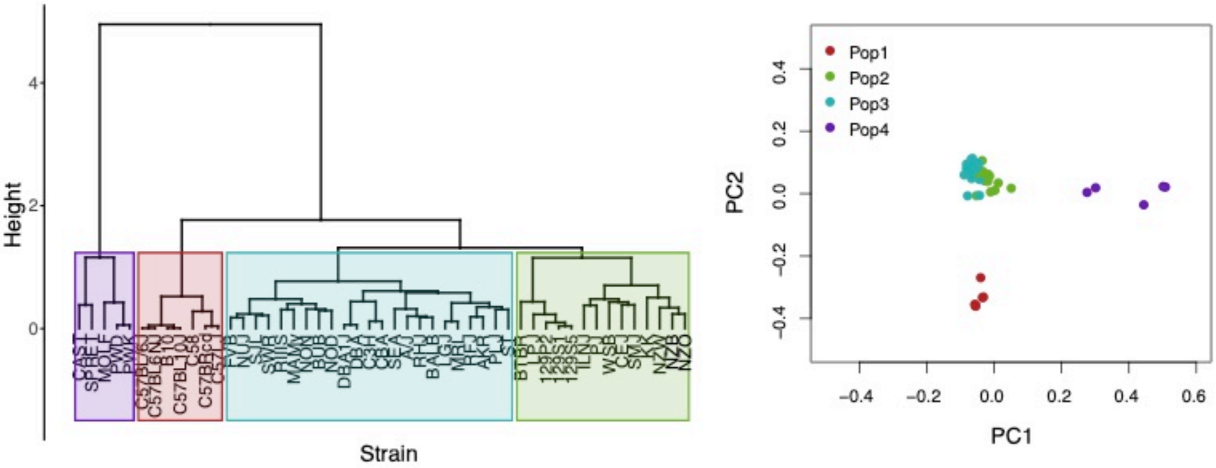

Figure 1. An analysis of population structure among 49 inbred mouse strains, which is based upon whole genome sequence analysis, identifies four sub-populations. The relatedness of the 49 inbred strains based upon hierarchical clustering using an identity-by-state similarity matrix (*Left*); or a scatter plot generated by PCA using the first two PCs for each strain are shown (*Right*). The sub-populations identified by the two methods are completely concordant. Sub-populations 1 and 4 are distinct from the majority of the inbred strains that contained in two closely related sub-populations (2 and 3).

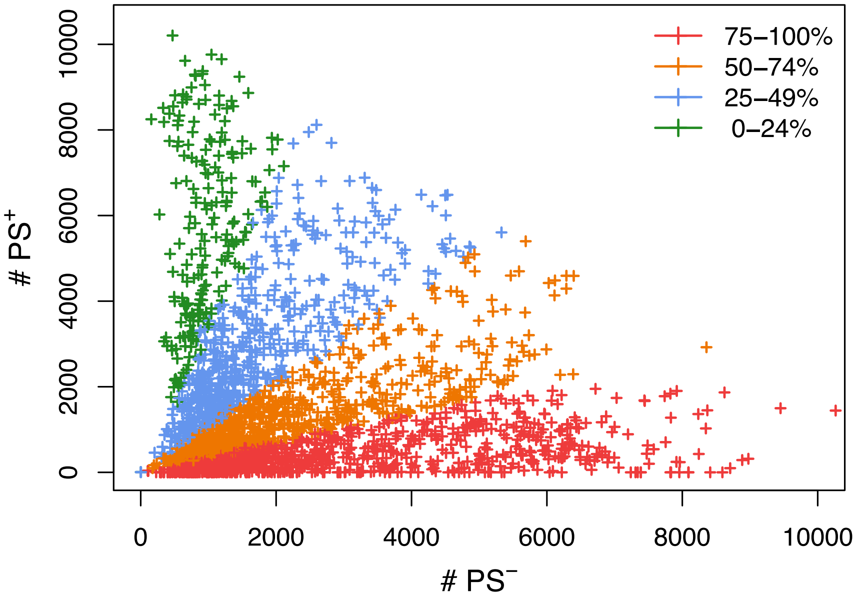

Figure 2. A scatter plot showing the number of candidate haplotype blocks associated with population structure (PS$^+$) relative to PS$^-$ candidate blocks. After 2435 MPD datasets were analyzed by HBCGM, candidate blocks ($p_{HBCGM} < 0.01$) were analyzed by an association test to determine whether they were related to population structure among the inbred strains that were analyzed. Each datapoint (+) indicates the number of PS$^+$ (y-axis) and PS$^-$ (x-axis) blocks identified for one MPD dataset. MPD datasets where 75% to 100% of the blocs are PS$^-$ are shown in red; orange datasets have 51-74% PS$^-$ blocks; blue datasets have 25-49% PS$^-$ haplotype blocks; and those with 0-24% PS$^-$ blocks are shown in green.

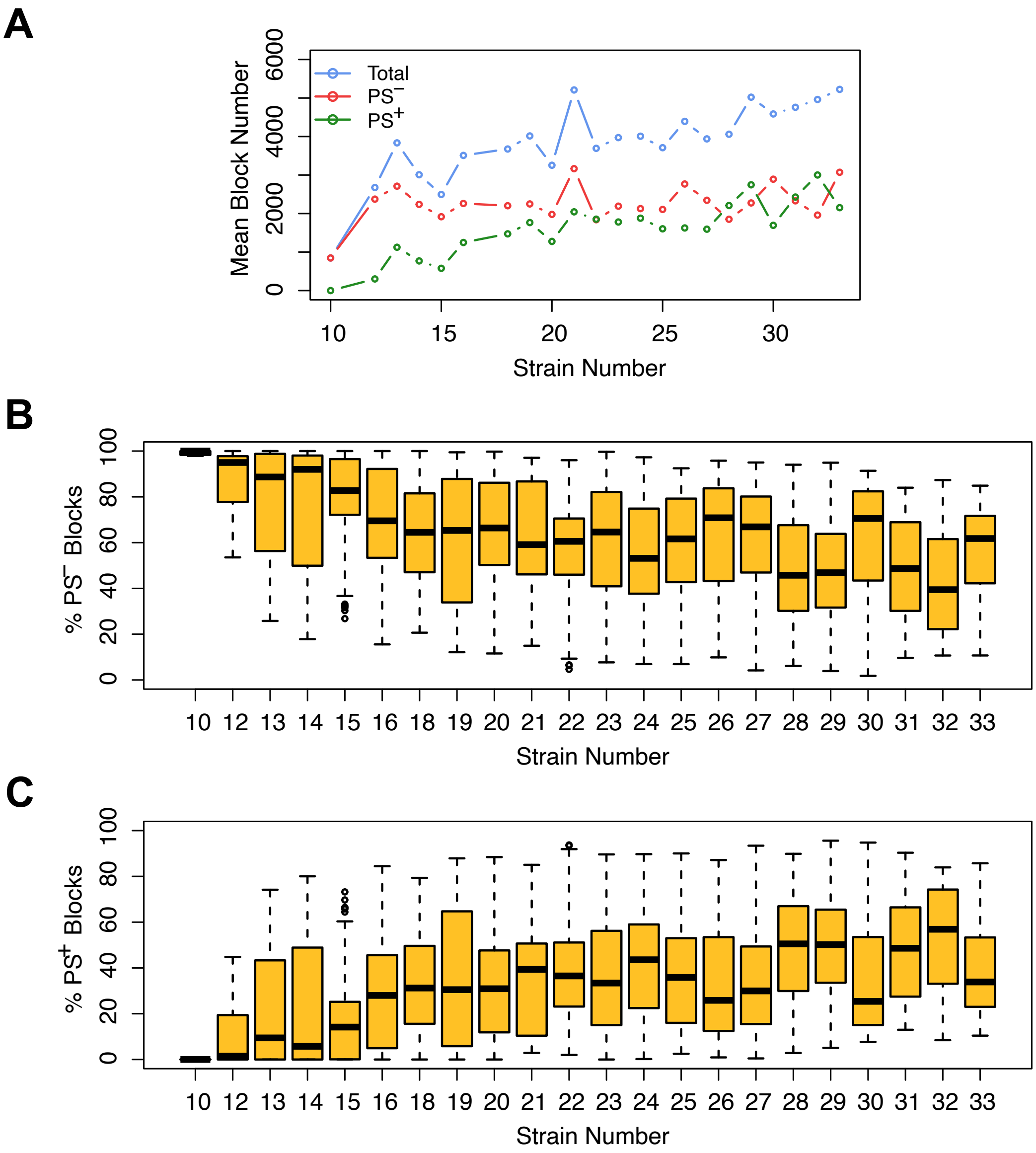

Figure 3. The effect of population structure increases with the number of analyzed strains. Analysis of the total number of candidate haplotype blocks, the number of blocks with population structure (PS$^+$), and the number of PS-independent (PS$^-$) blocks are shown as a function of the number of analyzed strains. After 2435 MPD datasets were analyzed by HBCGM, the correlated blocks ($p_{HBCGM} < 0.01$) were analyzed by an association test to determine whether population structure had a significant influence on the strain groupings within the blocks. (**A**) The results were then graphed as a function of the number of mouse strains within each dataset (range 10 – 33). A blue circle represents the average of the total number of candidate blocks, and the mean number of PS$^-$ (red) and PS$^+$ blocks (green) are also shown in this graph. (**B, C**) The percentage of (**B**) PS$^-$ and (**C**) PS$^+$ blocks was then assessed for each

dataset. The box plots indicate the 25th and 75th percentile, and the black bar indicates the median value. While the number of PS$^-$ blocks plateaued after 15 strains were analyzed, the number of PS$^+$ blocks increased in the datasets that analyzed an increased number strains.

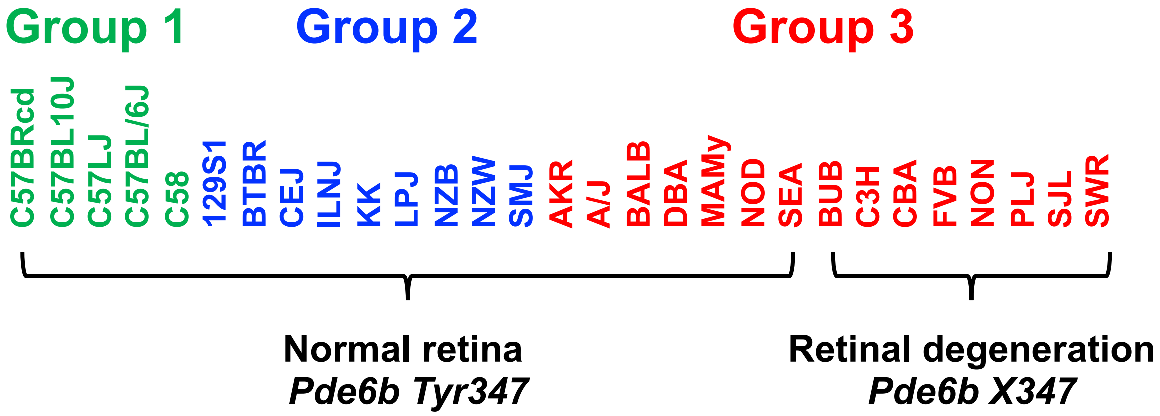

Figure 4. The haplotype block with a causative mutation is associated with population structure. MPD 26721 examined the retinas of 29 inbred strains: 21 strains had normal retinas and 8 strains had retinal degeneration. A haplotype block within *Pde6b* contained the causative SNP (*Tyr347X*) for this type of retinal degeneration. All strains with retinal degeneration had the *Pde6b 347X* allele, while those with normal retinas had the *Tyr347* allele. The haplotype block had PS, because all group 1 and 2 strains (based upon hierarchical clustering of whole genome sequence data from 49 inbred strains (**Table 1**)) had normal retinas; while all strains with retinal degeneration were group 3 strains. However, several group 3 strains (AKR, A/J, BALB, DBA, MaMy, NOD, SEA) had normal retinas and the *Tyr347* allele. Thus, while the strain groupings within the block have PS based upon their global allele sharing pattern, the allelic pattern within the haplotype block had a stronger association with retinal degeneration.