Fine mapping in 94 inbred mouse strains using a high-density haplotype resource

Andrew Kirby\textsuperscript{1,2,*}, Hyun Min Kang\textsuperscript{3,*}, Claire M. Wade\textsuperscript{2,4}, Chris Cotsapas\textsuperscript{1,2,5}, Emrah Kostem\textsuperscript{6}, Buhm Han\textsuperscript{6}, Nick Furlotte\textsuperscript{6}, Eun Yong Kang\textsuperscript{6}, Manuel Rivas\textsuperscript{2,7}, Molly A. Bogue\textsuperscript{8}, Kelly A. Frazer\textsuperscript{9}, Frank M. Johnson\textsuperscript{10}, Erica J. Beilharz\textsuperscript{9}, David R. Cox\textsuperscript{9}, Eleazar Eskin\textsuperscript{6,11†}, Mark J. Daly\textsuperscript{1,2,5†}

\textsuperscript{1}Center for Human Genetics Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
\textsuperscript{2}Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.
\textsuperscript{3}Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
\textsuperscript{4}Faculty of Veterinary Science, The University of Sydney, New South Wales 2006, Australia.
\textsuperscript{5}Dept of Medicine, Harvard Medical School, Boston, Massachusetts 02114, USA.
\textsuperscript{6}Department of Computer Science, University of California, Los Angeles, Los Angeles, California 90095, USA.
\textsuperscript{7}Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
\textsuperscript{8}The Jackson Laboratory, Bar Harbor, Maine 04609, USA
\textsuperscript{9}Perlegen Sciences, Mountain View, California 94043, USA
\textsuperscript{10}Toxicology Operations Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA
\textsuperscript{11}Department of Human Genetics, University of California, Los Angeles, Los Angeles, California 90095, USA.

* These authors equally contributed to this work
† These authors equally contributed to this work.
The genetics of phenotypic variation in inbred mice has for nearly a century provided a primary weapon in the medical research arsenal. A catalogue of the genetic variation among inbred mouse strains, however, is required to enable powerful positional cloning and association techniques. A recent whole genome resequencing study of 15 inbred mouse strains captured a significant fraction of the genetic variation among a limited number of strains (Frazer et al., 2007) - yet the common use of hundreds of inbred strains in medical research motivates the need for a high-density variation map of a larger set of strains. Here we report a dense set of genotypes from 94 inbred mouse strains containing 10.77 million genotypes over 121,433 single nucleotide polymorphisms (SNPs), dispersed at 20 kb intervals on average across the genome, with an average concordance of 99.94% with previous SNP sets. Through pairwise comparisons of the strains, we identified an average of 4.70 distinct segments over 73 classical inbred strains in each region of the genome, suggesting limited genetic diversity between the strains. Combining these data with genotypes of 7,570 gap-filling SNPs, we further imputed the untyped or missing genotypes of 94 strains over 8.27 million Perlegen SNPs. The imputation accuracy among classical inbred strains is estimated at 99.7% for the genotypes imputed with high confidence. We demonstrated the utility of this data in high-resolution linkage mapping through power simulations and statistical power analysis and provide guidelines for developing such studies. We also provide a resource of in-silico association mapping between the complex traits deposited to Mouse Phenome Database (MPD) with our genotypes. We expect that these resources will facilitate effective designs of both human and mouse studies for dissecting genetic basis of complex traits.

Introduction
Phenotypic variation among inbred mouse strains exposed to a disease causing agent (be it genetic, infectious or environmental) provides potential insight into human disease processes that often cannot be practically achieved through direct human studies. Indeed hundreds of phenotype measurements related to human diseases are available for dozens of inbred strains in common use over the last 50-100 years (BOGUE et al., 2007; GRUBB et al., 2009). As with the direct study of chronic disease in humans, a key step towards determining the genetic underpinnings of this phenotypic variation is to develop a catalogue of the genetic variation among inbred mouse strains and interpreting the structure of variation patterns across the strains. Recent advances in high-throughput genotyping and DNA resequencing technologies are making it possible to rapidly uncover the genetic variation maps of many model organisms (FRAZER et al., 2007; LINDBLAD-TOH et al., 2005; MACKAY and ANHOLT, 2006; BOREVITZ et al., 2007; INTERNATIONAL HAPMAP CONSORTIUM, 2007; STAR CONSORTIUM, 2008). A recent whole genome resequencing study of 15 inbred mouse strains captured a significant fraction of the genetic variation among a limited number of strains allowing researchers to infer patterns of genetic variation and identify the ancestral origin of the genetic variation (FRAZER et al., 2007; YANG et al., 2007). Yet, the availability and common experimental employment of hundreds of inbred strains, including over 190 stocks available from the Jackson Laboratory, motivates the need of a high-density variation map for a larger set of strains. We have assembled the Mouse HapMap, a resource consisting of a dense set of genotypes for a total of 138,980 unique biallelic single nucleotide polymorphisms (SNPs) in 94 inbred mouse strains, at an average spacing of 20kb on chromosomes 1-19 and X.

This resource is ideal for performing high-resolution mapping studies under QTL peaks. We evaluate the feasibility and effectiveness of such studies by examining a typical study from the Mouse Phenome Database (MPD) (BOGUE et al., 2007; GRUBB et al., 2009) (http://www.jax.org/phenome) and measure the statistical power to detect genetic associations in regions of various sizes. We provide several resources to the
mouse genetics community for supporting such studies. We provide a webserver which can estimate the significance threshold, compute statistical power of a proposed study and perform in fine mapping of measured phenotypes. In addition, we provide a database of associations for the all phenotypes contained in the MPD. The web resources are available at http://mouse.cs.ucla.edu/.

Methods

Array Design

The mouse HapMap genotypes were obtained using two Affymetrix genotyping arrays with 20 or 36 PM/MM probe-pairs. SNPs were selected to as evenly spaced as possible across the NCBI build 33, and mapped to NCBI build 37. Genotypes were called with Affymetrix DM algorithm, and the genotypes with low confidence genotypes or with conflicting calls between replicated samples or any discovery strain were called as missing.

Analysis of shared segments

We assigned each segment in the 94 strains of the Mouse HapMap to a founder strain representing different ancestral origin as well as identified shared segments among strains using a hidden Markov model following the approach presented in Frazer el al. (2007). The mapping with four founder strains was performed with a hidden Markov model with four reference strains representing possible founders with additional state for unknown reference, learning the parameters from the genotype data using EM algorithm as described in the imputation method. A hidden Markov model with two states representing common and divergent regions was constructed for each pairwise comparison, with recombination parameter $\theta = 10^{-8}$ and mutational parameter $\mu = 0.03$, estimated from the distribution of maximum likelihood parameters using EM algorithm among all 4,371 comparisons. The fraction of genome with shared segments was
computed as the fraction of genome wide SNPs with the probability of shared segments greater than 0.9. The number of distinct ancestral segments at a genomic position was computed by taking all the pairwise probabilities of shared segments, and by performing hierarchical clustering with a median agglomeration method by taking the pairwise probabilities as elements of a similarity matrix.

Imputation of missing genotypes

We performed imputation using EMINIM (Kang et al., 2009) of the Perlegen/NIEHS data (Frazier et al., 2007) in the 94 strains to increase the number of genotypes available for the 94 strains. Briefly, a hidden Markov model was constructed, for each strain targeted for imputation, with 16+1 states per SNP representing each of 16 resequenced reference strains and a state representing equivocal reference strain. Unlike the previous methods (Marchini et al., 2007; Scheet and Stephens, 2006), the maximum-likelihood parameters of genome wide mutation and recombination parameters were learned from the data using EM algorithm and forward-backward algorithm, independently for each strain. For leave-one-out imputation for experimentally missing genotypes in the resequenced strains, 15+1 states were used excluding the target strain for imputation.

Threshold Estimation

In order to control the false positive rate of in silico mapping, significance thresholds guaranteeing a 5% false-positive rate were estimated for regions of size 10MB, 20MB and 30MB via simulation. For each size nMB region, the genome was split into non-overlapping bins of size n. For each of these bins a random phenotype, capturing background genetic effects, was generated, and EMMA (Kang et al., 2008) was used to perform association between the phenotype and all SNPs within the region. The most significant association was recorded for each bin. The threshold is then
determined by taking the p-value that is maximum amount the smallest 5% of all p-values within this set. A total of 10,000 simulations were performed.

**Power Simulation**

Statistical power for *in silico* mapping was estimated for a set of mouse strains by utilizing a simulation-based framework. First, we generated a random phenotype with a correlation structure that is consistent with the genetic background by using a kinship matrix derived from the relatedness between the strains. Second, we adjusted the phenotypic values based on a particular SNP having a genetic effect. That is, the phenotype values for strains that have one allele at the SNP were increased by a predetermined amount corresponding to the strength of the genetic effect. Finally, we used EMMA (Kang *et al.*, 2008) to detect the association of this SNP with this phenotype and recorded whether or not EMMA reports an association stronger than the significance threshold. This type of simulation was performed over 10,000 SNPs chosen uniformly at random from a set of over 100,000. The statistical power is then defined as the fraction of associations detected at the pre-determined significance threshold.

**Mapping Resolution Simulations**

Mapping resolution was estimated in each of the strain sets by utilizing a simulation-based framework. Phenotype data was generated in a similar manner as in the power simulations using a randomly chosen SNP as having a genetic effect. EMMA was used to detect the associations for all SNPs that are within 10 megabases of this SNP. For any simulation where any SNP in the region exceeds the significance threshold, we recorded the genomic distance between the causal SNP and the SNP with the most significant p-value in the region. This type of simulation was performed over 10,000 SNPs chosen uniformly at random from a set of over 100,000. The mapping resolution is then defined as the average of the distance between the most significant SNP and the SNP simulated to have a genetic effect.
Additional Strain Selection

Using the “Paigen2” mouse strain set (SVENSON et al., 2007) from the Mouse Phenome Database as a starting point, we determined a new set of 33 inbred strains that would provide an increase in power when compared with the Paigen2 set. Our strain set was selected by first removing both wild-type and some genetically similar strains from the original Paigen2 strain set of 42 strains (for HDL cholesterol). The resulting 27 strains were used as a template to build a new inbred panel. We used a subset of mouse HapMap strains as a candidate panel and iteratively selected mice that were genetically dissimilar to the template set. We selected candidate strains based on their maximum genetic correlation with all strains in the growing template panel. The strain that had the maximum correlation, which was minimum among all candidate strains, was selected and the procedure was repeated.

In-silico association mapping database

We downloaded the individual phenotype measurements of Mouse Phenome Database (MPD) from Jackson Laboratory, and selected 474 quantitative phenotypes containing phenotype measurements in at least 20 strains. We applied EMMA (Efficient Mixed Model Association) (KANG et al., 2008) as an implementation of linear mixed models to correct for population structure and genetic relatedness, using the kinship matrix generated as a genotype similarity matrix. The variance component was based on REML (Restricted Maximum Likelihood) estimate, and a standard F test was performed as previously suggested (YU et al., 2006; ZHAO et al., 2007). The FDR significance level was estimated using the q-value R package (STOREY and TIBSHIRANI, 2003). The males and females were mapped for association separately. The Genomic Control inflation factor was computed by taking the median p-value and computing the corresponding chi-square statistic divided by 0.455 (DEVLIN and ROEDER, 1999).
Results

The mouse HapMap resource

We have assembled a dense set of genotypes for a total of 138,980 unique biallelic single nucleotide polymorphisms (SNPs) in 94 inbred mouse strains, at an average spacing of 20kb on chromosomes 1-19 and X. We selected the most commonly used inbred laboratory strains – especially targeting priority strains from the Mouse Phenome Database (Bogue et al., 2007; Grubb et al., 2009) – and 19 wild-derived strains both as reference out-groups and to help identify ancestry of genomic segments (Table 1). Our dataset is a composite of 121,433 SNPs discovered and genotyped at the Broad Institute by comparing data from the two inbred mouse genome sequencing projects (Mouse Genome Sequencing Consortium, 2002; Mural et al., 2002), with additional discovery in a wild-derived strain in regions of low marker density. In addition, we include 7,570 SNPs covering physical gaps in the Broad Institute map revealed by examining data from the concurrent NIEHS/Perlegen effort to resequence 15 inbred strains (Frazer et al., 2007) and 13,094 SNPs discovered and genotyped at the Wellcome Trust Center for Human Genetics (WTCHG) that could be mapped to Build 37 of the mouse genome.

To evaluate the quality of these resources, we examined SNPs typed in common by Broad and WTCHG as well as compared each resource to the genotypes of strains produced from the NIEHS/Perlegen sequence data. SNPs overlapping between the Broad and WTCHG sets demonstrate a discordance rate of 0.00058, while SNPs overlapping WTCHG and NIEHS/Perlegen sequence-based genotypes demonstrate a discordance of 0.00688. The extremely high concordance of the Broad and WTCHG data and significantly higher accuracy than the array-based sequence genotypes are unsurprising; the Broad and WTCHG utilized established SNP genotyping techniques and need only distinguish between two homozygous genotype classes. An interesting disparity in
discordance rate is observed between Perlegen and WTCHG genotypes. When the WTCHG genotype is the reference strain allele (C57BL/6J) the disparity with Perlegen genotype is 0.00335 and is 0.0106 otherwise. This is consistent with the variant discovery strategy employed by Perlegen, which emphasized low false positive variant discovery at the expense of a higher false negative rate (Frazer et al., 2007). Figure 1 summarizes the genotype resources for each of the 94 strains.

**Haplotype structure among the strains**

Using these genotype resources, we are able to examine the fine-level haplotype structure among the strains. For example, a comparison of the six 129 strains shows that they share the vast majority of their genomic segments, but there are several noticeable differences. In particular, there is a large disparity between 129P2/OlaHsD and 129X1/SvJ from 35Mb to 100Mb on chromosome 7, and there are also differences specific to 129S6/SvEv on chromosomes 3, 5, and 12 (Supplementary Figure 1). Similarly, comparisons between the fifteen C57 strains revealed significant discrepancies between C57BL/6J and the other C57 strains (Supplementary Figure 2). We also identified that some strains appear to result from recent hybridizations between two or more strains. We observed that HTG/GoSfSnJ shares more than 99.9% of genome with either BALB/cByJ or C57BL/6J (Supplementary Figure 3) and that NOR/LtJ shares more than 99.9% segments with either NOD/LtJ or C57BLKS/J, confirming the annotated genealogical history (Beck et al., 2000) (Supplementary Figure 4). We also observed that two strains (RBA/DnJ and SOD/EiJ) are "hybrid" strains with genetic content from both classical inbred and wild-derived strains. (Supplementary Figure 5). When comparing the fraction of genome shared by any of the 12 classical inbred resequenced strains, there is a clear difference between rates of sharing with the wider set of classical inbred strains (97% of the genome on average and 81% minimum) and with the wild-derived strains (28% on average, 56% maximum) (Figure 2).

We allocated ancestry of local genomic regions to one of the four “founder” strains.
using the methods described previously for resequencing data (FRAZER et al., 2007). For
each of the remaining 90 strains, we identified the fractions of genomic regions
unequivocally close to *domesticus*, *musculus*, *castaneus*, and *molossinus* strains. On
average these ancestral strains contribute 32.3%, 9.19%, 4.52%, and 11.8%, respectively.
42.2% of observed total genomic regions are ambiguous for ancestry, meaning either that
the ancestry is not precisely represented by any of the four founder strains (37.3%), or
else that two or more ancestral sub-species share haplotypes in these regions (4.86%).
The fractions of regions identified as having *domesticus* or unknown ancestry differ from
previous studies (FRAZER et al., 2007) due to the sparser resolution of the SNP map, and
the SNP ascertainment bias inherent in both current and former datasets. We note that
these ancestry estimates make many assumptions, one of which is that the founder strains
represent the true ancestral populations of the strains. Other studies which make slightly
different assumptions such as Yang et al. (2007) differ in their ancestry estimates. All of
the classical inbred strains and hybrid strains share predominantly *domesticus* ancestry
(with Yang et al. (2007) having a higher estimate of the percentage of *domesticus*
ancestry compared to Frazer et al. (2007)), while the wild-derived strains are divided into
four groups corresponding to their respective ancestral subspecies: this is also reflected in
the phylogeny derived from the mouse HapMap data (Figure 3).

To investigate the average sizes of shared haplotype segments among strains, we
identified common (low SNP density) and divergent (high SNP density) ancestral
segments across the genome for each pair of inbred strains using a hidden Markov model
(FRAZER et al., 2007). Among the 4371 possible pair-wise comparisons of the 94 strains,
an average of 32.5% of the genomic regions are shared between any pair of strains
(Figure 4). The average number of shared ancestral segments genome-wide is 280 per
comparison, which is about one segment per 10Mb. On average, there are 176 segments
longer than 1Mb covering 28.8% of the genome, and 39 segments longer than 5Mb
covering 15.6% of the genome – reflective of the tight recent co-ancestry of these strains.
Given a cross between any of the two parental strains, it is possible to estimate the
genomic region excluded from mapping variations associated with phenotype traits due to the shared segments between them. For example, in mapping studies using BXD recombinant-inbred strains, 48.6% of genomic regions are shared between parental strains and loci in these regions will not be mapped to traits.

To ascertain whether intervening genotypes might be successfully imputed from the resequencing data, we counted how many distinct haplotypic segments exist for each genomic region and compared this with the numbers derived from the resequencing data by combining the shared segment analysis using hierarchical clustering. The average number of distinct segments within any region is estimated to be 4.70 over 73 classical inbred strains. This limited diversity likely reflects recent bottlenecks, where a limited number of chromosomes from the founder strains gave rise to the modern inbred strains (Frazer et al., 2007; Wade et al., 2002; Frazer et al., 2004). Among the 12 resequenced classical inbred strains, an average of 3.46 ancestral segments were identified. Like the analysis of shared segments, these results suggest that most of the genetic variation existing among the classical inbred strains can be explained by the variation present in the resequenced strains.

**Integrating NIEHS/Perlegen resequencing and HapMap data**

Now confident that we could identify segment ancestry by reference to the 16 resequenced strains, we proceeded to impute genotypes for the 8.27 million NIEHS/Perlegen SNPs on the 78 genotyped strains using a hidden Markov model that learns genome wide transition and mutation parameters using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Kang et al., 2009). A feature of the technique we used for imputation is that we can obtain confidence levels for each prediction (Kang et al., 2009). We were able to call the majority of SNPs (79.2%) with high confidence (posterior probability > 0.98), when genotypes were successfully called in all 16 resequenced strains (see Table 2, top panel for details). We found that confidence scores vary greatly, with 11 wild-derived strains having no high-confidence
imputed genotypes because their estimated mutation rates were very high. In contrast, all 9 strains with the C57BL/6 prefix have more than 99.7% of high-confidence call rate, due to their genetic proximity to the reference strain C57BL/6J. We were also able to impute genotypes missing in the 16 resequenced strains, but only 17.2% of these with high confidence due to poor probe quality resulting in unreliable data (Table 2, bottom panel).

We estimated the accuracy of our imputed genotypes in two different ways. First, we used a leave-one-out cross-validation approach to impute genotypes for each of the 16 resequenced strains using the remainder. When considering the SNPs with complete data in the resequenced strains, the average leave-one-out imputation error over the 12 classical inbred resequenced strains was 1.59%, dropping to 0.27% when only high confidence genotypes were used (Table 3, Supplementary Table S1). We found that these rates varied substantially between the 12 classical inbred strains (range 1.17% - 3.63%; high-confidence genotype error range 0.21% - 0.67%). These errors increase when considering the four wild-derived strains, with total imputation error ranging from 13.0% to 34.1% (Supplementary Table S2). These error differences likely reflect the divergent ancestry of the imputed strains, as the marker set remains biased towards the strains used for SNP discovery. Next, we estimated accuracy by comparing our imputed genotypes to data previously generated by the WTCHG on 47 of the 78 genotyped strains (Table 4), and found a total error rate of 4.86% (2.26% when excluding the 11 wild-derived and hybrid strains). Restricting to the 71.7% of the imputed genotypes called at high confidence genotypes reduces this error to 0.37%, more than ten times smaller than recently published results for this marker subset using a different method (SZATKIEWICZ et al., 2008). As in the previous error estimate, the imputation error again differs greatly by strain, ranging from 0.065% to 20.9% (0.019% to 4.41% for high confidence imputed genotypes) (Supplementary Table S3).

In summary, we were able to impute 657,300,201 genotypes across 8.27 million markers in 94 inbred strains, including 14,622,883 experimentally missing genotypes in the resequencing strains and 1,265,768 genotypes missing in the combined genotype sets
(Table 2). This creates a near-comprehensive snapshot of variation in commonly available mouse strains.

To estimate the cost-effectiveness of expanding this resource, we evaluated the potential imputation coverage made possible by either increasing the number of resequenced strains or the number of SNPs in the HapMap as discussed in the Supplementary materials.

**Trait mapping with the mouse HapMap resource**

This detailed picture of haplotype diversity in the mouse allows us to map traits in the inbred strains by correlating genomic ancestry to trait measurements, rather than generating *de novo* experimental crosses. This *in silico* association mapping has two advantages: it allows us to capture the full spectrum of diversity in the inbred strains rather than a subset used as progenitors of an experimental cross; and phenotypic noise can be minimized by performing replicates on genetically identical individuals. In particular, this approach should complement traditional QTL linkage mapping (often successful at locating large chromosomal segments) by providing a higher resolution, association-based component and indeed has already yielded several positive results (Grupe *et al.*, 2001; Pletcher *et al.*, 2004; Liao *et al.*, 2004; Liu *et al.*, 2006; Moran *et al.*, 2006; Liu *et al.*, 2007; Guo *et al.*, 2006; Guo *et al.*, 2007; Cervino *et al.*, 2007; McClurg *et al.*, 2007; Tang *et al.*, 2008).

The basic idea behind this type of study is that a region is first identified through a genetic cross or some other means resulting in a large QTL region typically in the tens of megabases in length which contain many genes. Several dozen inbred strains are then phenotyped and association analysis is then performed in this region. The association analysis requires two steps. The first is to determine an appropriate significance threshold which depends on the size of the QTL region. The second is to perform the association on each marker within this region. The key idea increasing the power of this approach is that since only the markers under the QTL are examined, the significance threshold will be much less conservative than a genome wide significance threshold. We
perform simulations to evaluate the statistical power and mapping resolution of such an approach as well as provide insights into the design of such studies.

**Statistical power of in silico association**

As has been previously shown by several studies (PAYSEUR and PLACE, 2007; MANENTI et al., 2009), datasets consisting of several dozen strains do not have enough statistical power to detect weak effects (5% variance explained) with genome wide significance. We evaluate the feasibility and effectiveness of mapping in a localized region the size of a typical QTL peak.

One complication in our analysis is that the high degree of relatedness between strains described above introduces a systematic bias in association mapping *in silico*: an inflation of test statistics leading to false positive associations, caused by population structure and genetic relatedness among the strains (KANG et al., 2008; ARANZANA et al., 2005; YU et al., 2006; ZHAO et al., 2007). For example, among the 180 phenotypes deposited in the Mouse Phenome Database at the Jackson Laboratory (MPD) with more than 30 distinct strains, 59% (106) of them have more than 50% of the inter-strain phenotypic variance explained by population structure and genetic relatedness measured using a variance component test. (Figure 5). At an FDR level of 0.05, 51% (91) of the phenotypes are significantly associated with population structure. We and others have shown that these issues can effectively be corrected using linear mixed models (KANG et al., 2008; YU et al., 2006; ZHAO et al., 2007).

To evaluate the effectiveness of fine mapping through *in silico* association, we use as a typical study representative of the types of studies in the MPD, the “Paigen2” study (SVENSON et al., 2007) which contains phenotype measurements for HDL cholesterol in 42 strains: 33 classical inbred strains and 9 wild derived strains and contained an average of 21 replicates per strain (Table 5). While this study is somewhat larger than most of the studies in the MPD, we chose this study to examine because it contained many phenotypes and included both wild derived and classical inbred strains which allowed us to explore study design choices in terms of the strains chosen.
We perform simulations to both obtain the significance threshold and statistical power. As described in Kang et al. (KANG et al., 2008), the statistical power depends on the background genetic effect or intuitively on how much the intra-strain relatedness explains the phenotypic variance. In our framework, this relatedness is modeled with a variance component defined by the genetic similarity between the strains. We perform our simulations varying the background genetic effect from 25% to 75% to capture the wide range of potential phenotypes. For actual phenotype values, the background genetic effect can be estimated from a phenotype by comparing the inter-strain variance to the intra-strain variance after fitting the variance component model. Using simulations (see Methods), we computed the 0.05 level of significance for 10, 20 and 30 megabase regions which is shown in Table 6. We found that the level of the background genetic effect did not affect the significance threshold (data not shown), but the choice of strains had a significant effect. We can observe this phenomenon by comparing the significance threshold of the full set of 43 strains to the significance threshold of just using 33 classical inbred strains.

Using these thresholds we performed an additional set of simulations to compute the power to detect various genetic effect sizes under different backgrounds and region sizes shown in Table 7. The results show that the statistical power of a study of this size is high for either phenotypes where the background genetic effect is low or for strong genetic effects. Since both the threshold and power depends on the strain set and are estimated using computationally intensive simulations, we provide a webserver resource (http://mouse.cs.ucla.edu/) for performing these simulations and threshold estimation.

We note that any set of strains is only able to map traits linked to variation that is polymorphic within the set of strains. Since the Mouse HapMap consists mostly of SNPs which are polymorphic among the laboratory strains (not wild-derived), there is little power to detect SNPs which are only polymorphic among the wild-derived strains. Our simulation framework essentially ignores this problem since the possible set of simulated genetic variants are by necessity within the set of data used to generate the simulation.
Alternative strategies for mapping traits such as genetic crosses which include wild-derived strains or inbred strains with significant proportions of wild-derived strain ancestry (such as the Collaborative Cross (Churchill et al., 2004)) will have higher power to map traits to loci polymorphic among wild-derived strains. Furthermore, the statistical power to map a trait to a specific locus greatly depends on the number of strains carrying the minor allele of the locus similar to the effect of minor allele frequency on the power of human association studies.

**Resolution of in silico association mapping**

The main advantage of the in silico approach is the increased resolution compared to traditional QTL approaches. To evaluate the resolution, we again perform simulations using the set of strains in the “Paigen2” study. We performed 10,000 simulations where in each simulation we generate phenotype data assuming a randomly selected causal variant and then perform association mapping over the generated phenotypes and record the difference between the genomic location of the most associated marker and the causal variant. The resolution experiments demonstrate that the median distance between the actual causal variant and the closest marker is around 3 megabases (Table 8), a significant improvement over a traditional cross.

**Design of in silico association studies**

Interestingly, as shown in the table, there was only a moderate power loss if we consider using only the 33 classical inbred strains from the Paigen2 study. Part of the reason for this is that for the SNPs which are unique to the wild-derived strains and are not polymorphic in the classical inbred strains, the power to detect associations at these SNPs is low since there are relatively few wild derived strains included in the study.

We further explore how the choice of strains included in the study affects statistical power. We notice that among the strains included in the “Paigen2” study, there are several strains from the C57 group which are very genetically similar. A total of 6 strains out of the 33 inbred strains are very similar to other strains in the set. We construct a new set of inbred strains removing these similar strains and replacing them with more distant
classical inbred strains (Table 9). Using power simulations, we observe that this set of 33 strains has more power than the set of classical inbred strains included in Paigen2 study and surprisingly has comparable power to the complete Paigen2 set (43) strains (Table 7). These simulations are computationally intensive and as a resource to the community, we provide a webserver (http://mouse.cs.ucla.edu) for performing these simulations.

**Mouse Phenome Association Database**

To enable the research community to have access to the population structure corrected associations, we have developed a corrected association database in conjunction with the MPD, in which we find 71/180 phenotypes collected in more than 30 strains have at least one significant association ($P < 1 \times 10^{-6}$). The database contains results for both the genotyped and imputed SNPs. Among the phenotypes, 11 (6.1%) phenotypes showed significant associations across more than 20 different genomic regions, which may indicate residual bias from other sources generating false positives. This may be compared to 24 (13%) phenotypes showing association without population structure correction to more than 20 different genomic regions, while the total number of phenotypes with significant associations is similar (Figure 6). When comparing the “inflation factor” suggested by Genomic Control between different statistical tests, t-test showed much higher overall inflation ($\lambda=2.08\pm1.29$) compared to the linear mixed model ($\lambda=1.15\pm0.18$) over the 180 MPD phenotypes, confirming the overly inflated false positive rates with the conventional t-test (Figure 7).

**Discussion**

We have described the high-density genotype resource for 94 inbred mouse strains and have demonstrated the viability of applying such a resource to fine mapping using *in silico* association. Our genotype data is available at http://mouse.cs.ucla.edu/mousehapmap/. In addition, we have established a website http://mouse.cs.ucla.edu/ at which researchers can download genotype data, and access a genome browser which allows the visualization of the haplotype and shared segment
analyses. The website also supports inbred association mapping allowing users to upload their collected phenotypes, as well as computes the significance thresholds and estimated statistical power. The website (http://mouse.cs.ucla.edu) include association results using the genotypes and all collected phenotype data in the Mouse Phenome Database.

A major concern for in silico association mapping has been the effect of population structure potentially causing false positives. We have shown previously that EMMA corrects for this population structure using a linear mixed model (KANG et al., 2008) with a variance component using a kinship matrix obtained from the genetic similarity between strains. However, even with the correction, there is a slight inflation of statistics observed for some phenotypes in the associations in the Mouse Phenome Database. This inflation may be caused by a variety of factors including different amounts of phenotypic variance in each strain, the kinship matrix not completely capturing the background genetic effects, and other confounding effects such as “cage effects” which are correlated with the strains. We report the genomic control $\lambda$ values (DEVLIN and ROEDER, 1999) as a means to quantify the amount of inflation and we urge users of the association database to take this inflation factor into account when interpreting the results of the phenotype associations.

Our study is one of several recent efforts for developing genetic and genomic resources for inbred mouse strains. Recently, Yang et al. (2009) developed a novel high density genotyping array which includes many SNPs chosen from both the Mouse HapMap as well as from the Perlegen resequencing data (FRAZER et al., 2007). In addition, the Wellcome Trust Sanger is currently sequencing 17 mouse genomes (SUDBERY et al., 2009) (http://www.sanger.ac.uk/resources/mouse/genomes/). Both of these efforts in combination with existing resources will lead to more dense and accurate genetic maps for laboratory strains.

Fine mapping of QTL loci by performing in silico association using inbred strains is just one of several approaches recently proposed to increase the mapping resolution of traditional QTL approaches. Alternate strategies include using the Collaborative Cross
which contains a large number of inbred strains derived from eight parental inbred strains using a breeding strategy to avoid population structure, using Hybrid Mouse Diversity Panel (Bennett et al., 2010) which combines classical inbred strains with recombinant inbred strains, and using outbred stock (Valdar et al., 2006). Each of these strategies has advantages and disadvantages.

The mouse community is just one of many communities developing genetic and genomic resources for mapping complex traits. Similar efforts are being undertaken for many model organisms including arabidopsis (Borevitz et al., 2007), drosophila (Mackay and Anholt, 2006), dog (Lindblad-Toh et al., 2005) and rat (Star Consortium, 2008).

Acknowledgements
A.K., C.M.W., C.C., M.R., and M.J.D. as well as the genotyping was supported by National Institutional of Health (NIH) grant P41-HG003056. H.M.K., E.K., B.H., N.F., E.Y.K, and E.E. are supported by National Science Foundation grants 0513612, 0731455 and 0729049, and NIH grants K25-HL080079 and U01-DA024417. H.M.K. and B.H. are supported by the Samsung Scholarship. H.M.K. is also supported by NIH grants HG00521401, NH084698 and GlaxoSmithKline. M.B. is supported by The Jackson Laboratory, NIH grants MH071984 and DA028420. This research was supported in part by the University of California, Los Angeles subcontract of contract N01-ES-45530 from the National Toxicology Program and National Institute of Environmental Health Sciences to Perlegen Sciences.

References
Aranzana, M. J., S. Kim, K. Zhao, E. Bakker, et al., 2005 Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1: e60.
Beck, J. A., S. Lloyd, M. Hafezparast, M. Lennon-Pierce, et al., 2000 Genealogies of mouse inbred strains. Nat Genet 24: 23-25.
Bennett, B. J., C. R. Farber, L. Orozco, H. Min Kang, et al., 2010 A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res.

Bogue, M. A., S. C. Grubb, T. P. Maddatu, and C. J. Bult, 2007 Mouse Phenome Database (MPD). Nucleic Acids Res 35: D643-D649.

Borevitz, J. O., S. P. Hazen, T. P. Michael, G. P. Morris, et al., 2007 Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. Proc Natl Acad Sci U S A 104: 12057-12062.

Cervino, A. C., A. Darvasi, M. Fallahi, C. C. Mader, and N. F. Tsinoremas, 2007 An integrated in silico gene mapping strategy in inbred mice. Genetics 175: 321-333.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, et al., 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36: 1133-1137.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39: 1-38.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics 55: 997-1004.

Frazer, K. A., C. M. Wade, D. A. Hinds, N. Patil, et al., 2004 Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. Genome Res 14: 1493-1500.

Frazer, K. A., E. Eskin, H. M. Kang, M. A. Bogue, et al., 2007 A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448: 1050-1053.

Grubb, S. C., T. P. Maddatu, C. J. Bult, and M. A. Bogue, 2009 Mouse phenome database. Nucleic Acids Res 37: D720-D730.

Grupe, A., S. Germer, J. Usuka, D. Aud, et al., 2001 In silico mapping of complex disease-related traits in mice. Science 292: 1915-1918.

Guo, Y., P. Lu, E. Farrell, X. Zhang, et al., 2007 In silico and in vitro pharmacogenetic analysis in mice. Proc Natl Acad Sci U S A 104: 17735-17740.

Guo, Y., P. Weller, E. Farrell, P. Cheung, et al., 2006 In silico pharmacogenetics of warfarin metabolism. Nat Biotechnol 24: 531-536.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, et al., 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Kang, H. M., N. A. Zaitlen, B. Han, and E. Eskin, 2009 An Adaptive and Memory Efficient Algorithm for Genotype Imputation, pp. 482-495 in Lecture Notes in Computer Science. Springer Berlin / Heidelberg.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, et al., 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Liao, G., J. Wang, J. Guo, J. Allard, et al., 2004 In silico genetics: identification of a functional element regulating H2-Ealpha gene expression. Science 306: 690-695.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, et al., 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

Liu, P., H. Vikis, Y. Lu, D. Wang, and M. You, 2007 Large-scale in silico mapping of complex quantitative traits in inbred mice. PLoS ONE 2: e651.
Liu, P., Y. Wang, H. Vikis, A. Maciag, et al., 2006 Candidate lung tumor susceptibility genes identified through whole-genome association analyses in inbred mice. Nat Genet 38: 888-895.

Mackay, T. F., and R. R. Anholt, 2006 Of flies and man: Drosophila as a model for human complex traits. Annu Rev Genomics Hum Genet 7: 339-367.

Manenti, G., A. Galvan, A. Pettinicchio, G. Trincucci, et al., 2009 Mouse genome-wide association mapping needs linkage analysis to avoid false-positive Loci. PLoS Genet 5: e1000331.

Marchini, J. B. Howie, S. Myers, G. McVean, and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-913.

McClurg, P., J. Janes, C. Wu, D. L. Delano, et al., 2007 Genomewide association analysis in diverse inbred mice: power and population structure. Genetics 176: 675-683.

Moran, J. L., A. D. Bolton, P. V. Tran, A. Brown, et al., 2006 Utilization of a whole genome SNP panel for efficient genetic mapping in the mouse. Genome Res 16: 436-440.

Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, et al., 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

Mural, R. J., M. D. Adams, E. W. Myers, H. O. Smith, et al., 2002 A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296: 1661-1671.

Payseur, B. A., and M. Place, 2007 Searching the genomes of inbred mouse strains for incompatibilities that reproductively isolate their wild relatives. J Hered 98: 115-122.

Pletcher, M. T., P. McClurg, S. Batalov, A. I. Su, et al., 2004 Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol 2: e393.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.

STAR Consortium, K. Saar, A. Beck, M. T. Bihoreau, et al., 2008 SNP and haplotype mapping for genetic analysis in the rat. Nat Genet 40: 560-566.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

Sudbery, I., J. Stalker, J. T. Simpson, T. Keane, et al., 2009 Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. Genome Biol 10: R112.

Svenson, K. L., R. Von Smith, P. A. Magnani, H. R. Suetin, et al., 2007 Multiple trait measurements in 43 inbred mouse strains capture the phenotypic diversity characteristic of human populations. J Appl Physiol 102: 2369-2378.

Szatkiewicz, J. P., G. L. Beane, Y. Ding, L. Hutchins, et al., 2008 An imputed genotype resource for the laboratory mouse. Mamm Genome 19: 199-208.

Tang, P. L., C. L. Cheung, P. C. Sham, P. McClurg, et al., 2008 Genome-Wide Haplotype Association Mapping in Mice Identifies a Genetic Variant in CER1
Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, et al., 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38: 879-887.

Wade, C. M., E. J. Kulbokas, A. W. Kirby, M. C. Zody, et al., 2002 The mosaic structure of variation in the laboratory mouse genome. Nature 420: 574-578.

Yang, H., T. A. Bell, G. A. Churchill, and F. Pardo-Manuel de Villena, 2007 On the subspecific origin of the laboratory mouse. Nat Genet 39: 1100-1107.

Yang, H., Y. Ding, L. N. Hutchins, J. Szatkiewicz, et al., 2009 A customized and versatile high-density genotyping array for the mouse. Nat Methods 6: 663-666.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, et al., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208.

Zhao, K., M. J. Aranzana, S. Kim, C. Lister, et al., 2007 An Arabidopsis example of association mapping in structured samples. PLoS Genet 3: e4.
Supplementary Text

Effects of larger resources for imputation

To determine the effect of using a larger number of resequenced strains on the accuracy of imputation, we assumed that the 62 WTCHG strains were all resequenced, and estimated the imputation accuracy of imputing the WTCHG genotypes from the mouse HapMap SNPs and the gap-filling SNPs, using the leave-one-out cross-validation for each of the 62 strains. Because each strain targeted for imputation now has 61 instead of 16 reference strains, the imputation accuracy is expected to be high. Overall, the errors are reduced from 4.86% to 2.45%, and the errors in the 36 classical inbred strains are reduced from 2.25% to 0.96%. In contrast, the accuracy in the high-confidence genotypes of the classical inbred strains is reduced from 0.35% to 0.16%. More importantly, high-confidence call rate was increased from 88.9% to 95.6% for the 36 classical inbred strains, and from 71.7% to 84.9% for all 47 strains. Several strains such as MRL/MpJ, C57L/J, C57BR/cdJ, PERA/EiJ and PWK/EiJ showed a substantial improvement in imputation accuracy when a larger set of reference strains was used, while many other wild-derived strains still retained imputation errors of greater than 10%.

Since the resequencing of more strains is expected to increase the imputation coverage significantly, we prioritized the strains that might be targeted for resequencing to improve the coverage, based on our analysis of the shared segments. To do this, we picked the strain that maximized the additional genomic coverage of shared segments with the other strains given the coverage by the resequenced reference strains. This procedure is repeated greedily to select the next target of reference strain given the previous set of reference strains. To increase the coverage including the wild-derived strains, many wild-derived strains are prioritized for resequencing. When considering only classical inbred strains, the strains with relatively higher imputation errors tend to be prioritized (Supplementary Table S4).

Next, we estimated the effectiveness of imputation when different numbers of mouse HapMap SNPs are collected. To do this, we selected a range of sparse subsets
(10,000 to 1,000,000 markers) of the NIEHS/Perlegen SNPs with complete data in the resequenced strains and estimated the imputation errors for each of 12 resequenced classical inbred strains using leave-one-out cross-validation. As expected, accuracy increased proportionally to subset size. Selecting a 100,000 SNP subset gave an overall imputation error of 1.36% (high-confidence genotype error 0.36% with 93.8% call rate). This is comparable to the imputation accuracy using the current mouse HapMap SNPs. We note that the current size of the HapMap SNP is well powered to capture the majority of variation at low error rates and high confidence. (Supplementary Figure S3). A several-fold increase in SNP map density to 1,000,000 markers further optimizes these rates, and as current genotyping platforms can accommodate this number of assays this would be a viable design for the next generation mouse HapMap.
Figure Legends

**Figure 1** Classification of 94 strains used in the mouse HapMap projects based on the availability in other resources, including 8.27 million NIEHS/Perlegen resequencing-based SNPs, WTCHG SNPs, and additional gap-filling SNPs. (C57BL/6J is not included in the 15 resequenced strains, but it is the reference strain that has been fully resequenced).

**Figure 2** A histogram of the fractions of genome covered by shared segments with one of the 12 classical inbred strains, over 78 non-resequenced mouse HapMap strains. The classical inbred strains are colored in blue, the hybrid strains in red, and the wild-derived strains in green.

**Figure 3** A phylogeny of the 94 mouse HapMap strains. The *domesticus* wild-derived strains are colored in blue, and the non-*domesticus* wild-derived strains are colored in red. The reference strain is colored in green. SOD1/EiJ and RBA/DnJ are hybrid strains.

**Figure 4** A histogram of the fractions of shared genomic segments between each of 4371 pairs between the 94 strains.

**Figure 5** Distribution of fraction of phenotypic variation explained by population structure among the strains over 180 quantitative phenotypes deposited in the Mouse Phenome Database (MPD) with 30 or more strains.

**Figure 6** Number of phenotypes with multiple genomic regions with significant associations illustrating the degree of inflated false positives, over 180 quantitative phenotypes deposited in the Mouse Phenome Database (MPD) with 30 or more strains.

**Figure 7** Comparison of genomic control “inflation factors” between t-test and linear mixed model across 180 MPD phenotypes.

**Supplementary Figure 1** A genome-wide visualization of shared segments among six strains of 129 variants, with respect to the resequenced strain, 129S1/SvImJ. A colored region represents a region that is identical by descent (IBD) with the resequenced strain 129S1/SvImJ. A black region represents a region where the strain differs from the resequenced strain. Gray regions represent ambiguous regions.

**Supplementary Figure 2** A genome-wide visualization of shared segments among ten strains of C57 variants, with respect to the resequenced variant, C57BL/6J.
Supplementary Figure 3 A genome-wide visualization of shared segments between HTG/GoSfSnJ and their predicted ancestors, BALB/cByJ and C57BL/6J, which together covers more than 99.9% of the genome by shared segments. Each colored region represents an IBD region between HTG/GoSfSnJ and either BALB/cByJ or C57BL/6J. A black region represents a region that is not IBD and gray regions represent ambiguous regions.

Supplementary Figure 4 A genome-wide visualization of shared segments between NOR/LtJ and their ancestor strains, NOD/LtJ and C57BLKS/J, as annotated in genealogical history.

Supplementary Figure 5 A genome-wide visualization of shared segments between C57BL/6J and two hybrid strains, SOD/EiJ and RBA/DnJ across the genome.

Supplementary Figure 6 Estimated imputation accuracy and coverage over fully-resequenced NIEHS/Perlegen SNPs across 12 classical inbred strains with various sizes of randomly selected SNP sets.
Table 1 List of strains used in mouse HapMap projects and the availability in other resources

*C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced. (O included, X excluded)

| Strain name | Perlegen resequenced | WTCGH genotyped | Additional gap-filling | Wild-derived or classical inbred |
|-------------|----------------------|-----------------|------------------------|---------------------------------|
| 129P2/OlaHsD | X                    | X               | X                      | IN                              |
| 129S1/SvImJ  | O                    | O               | X                      | IN                              |
| 129S2/SvHsd  | X                    | X               | X                      | IN                              |
| 129S4/SvJae  | X                    | X               | X                      | IN                              |
| 129S6/SvEv   | X                    | O               | X                      | IN                              |
| 129T2/SvEms  | X                    | X               | O                      | IN                              |
| 129X1/SvJ    | X                    | O               | O                      | IN                              |
| A/J          | O                    | O               | X                      | IN                              |
| AKR/J        | O                    | O               | X                      | IN                              |
| B6A6ESlineRegeneron | X                | X               | X                      | IN                              |
| BALB/cByJ    | O                    | O               | X                      | IN                              |
| BALB/cJ      | X                    | O               | X                      | IN                              |
| BPH/2J       | X                    | O               | O                      | IN                              |
| BPL/1J       | X                    | O               | O                      | IN                              |
| BPN/3J       | X                    | O               | O                      | IN                              |
| BTBRT<+>t/J  | O                    | O               | X                      | IN                              |
| BUB/BnJ      | X                    | O               | O                      | IN                              |
| C2T1ESlineNagy | X                | X               | X                      | IN                              |
| C3H/HeJ      | O                    | O               | X                      | IN                              |
| C3HeB/FcJ    | X                    | O               | X                      | IN                              |
| C57BL/10J    | X                    | O               | X                      | IN                              |
| C57BL/6ByJ   | X                    | X               | X                      | IN                              |
| C57BL/6J     | O                    | O               | X                      | IN                              |
| C57BL/6BomTac | X                | X               | X                      | IN                              |
| C57BL/6Cl    | X                    | X               | X                      | IN                              |
| C57BL/6JolaHsd | X               | X               | X                      | IN                              |
| C57BL/6Ncrl  | X                    | X               | X                      | IN                              |
| C57BL/6Nhsd  | X                    | X               | X                      | IN                              |
| C57BL/6Nj    | X                    | X               | X                      | IN                              |
| C57BL/6nnih  | X                    | X               | X                      | IN                              |
| C57BL/6ntac  | X                    | X               | X                      | IN                              |
| C57BLKS/J    | X                    | X               | O                      | IN                              |
| C57Br/edJ    | X                    | O               | O                      | IN                              |
| C57L/J       | X                    | O               | O                      | IN                              |
| C58/J        | X                    | O               | O                      | IN                              |
| Calb/RkJ     | X                    | O               | X                      | W1                              |
| CAST/EJ      | O                    | O               | X                      | W1                              |
| CBA/J        | X                    | O               | O                      | IN                              |
| CE/J         | X                    | O               | O                      | IN                              |
| Czech/I/eJ   | X                    | X               | O                      | W1                              |
| DBA/1J       | X                    | O               | O                      | IN                              |
| DBA/2J       | O                    | O               | X                      | IN                              |
| DDK/Pas      | X                    | X               | X                      | IN                              |
| DDY/JelSidSeyFkJ | X               | O               | O                      | IN                              |
| EL/SuzSeyFkJ | X                    | O               | X                      | IN                              |
| Strain     | FVB/NJ | Fline | HTG/GoSfSnJ | I/LnJ | ILS | IS/CamRkJ | ISS | JF1/Ms | KK/HIJ | LEWES/EiJ | LG/J | LP/J | Lline | MA/MyJ | MAI/Pas | MOLF/EiJ | MOLG/DnJ | MRL/MpJ | MSM/Ms | NOD/LtJ | NON/LtJ | NOR/LtJ | NZB/B1NJ | NZL/LtJ | NZO/HILtJ | NZW/LacJ | O20 | P/J | PERA/EiJ | PERC/EiJ | PL/J | PWD/PhJ | PWK/PhJ | Qsi5 | RBA/DnJ | RF/J | RIIIS/J | SEA/GnJ | SEG/Pas | SII/L/J | SKIVE/EiJ | SMJ | SOD1/EiJ | SPRET/EiJ | ST/hj | SWR/J | TALLYHO/JngJ | WSB/EiJ | ZALENDE/EiJ |
|------------|--------|-------|-------------|-------|-----|----------|-----|--------|--------|-----------|------|------|-------|--------|---------|---------|---------|--------|-------|--------|--------|-------|---------|--------|------|-------|----------|-------|---------|--------|-------|---------|--------|---------|--------|-------|---------|----------|-------|---------|----------|-------|---------|--------|-------|---------|----------|-------|
|            | O      | X     | X           | O     | X   | X        | X   | X      | O      | X         | O    | O    | X      | O      | O       | O       | X       | X      | X      | O      | X      | X       | O      | X      | X       | O      | O       | X      | X      | X       | O      | O       | X      | X      | X       | O       | X      | X      | X       | O      | X      | X       | X      | X      | X       | O      | X      | X       |
Table 2 Classification of imputed genotypes that are untyped or experimentally missing. The fraction of imputed genotypes in each category is shown within a parenthesis. The confidence level corresponds to the predicted posterior probability of the imputation method. The level of resequencing corresponds to the number of missing genotypes in the 16 resequenced strains.

| Category | SNP Quality | Genotype confidence | High confidence | Medium confidence | Low confidence | Total |
|----------|-------------|---------------------|-----------------|------------------|---------------|-------|
| Ungenotyped | Fully resequenced | 235,728,507 | 48,532,073 | 13,431,178 | 297,691,758 |
| 8.22 million NIEHS/Perlegen genotypes over 78 non-resequenced strains | Mostly resequenced | 137,628,908 | 34,464,866 | 21,237,494 | 193,331,268 |
| | Poorly resequenced | 72,753,547 | 25,350,239 | 52,284,738 | 150,388,524 |
| Total | | 446,110,962 | 108,347,178 | 86,953,410 | 641,411,550 |
| Experimentally missing NIEHS/Perlegen genotypes over 16 resequenced strains | Mostly resequenced | 1,109,113 | 958,986 | 1,316,561 | 3,384,660 |
| | Poorly resequenced | 1,407,303 | 1,753,637 | 8,077,233 | 11,238,223 |
| Total | | 2,516,416 | 2,712,673 | 9,393,794 | 14,622,883 |
| Missing genotypes in the combined set | | 744,725 | 263,196 | 257,847 | 1,265,768 |
| Total | | 449,372,103 | 111,323,047 | 96,605,051 | 657,300,201 |
| Grand total | | 68.4% | 16.9% | 14.7% | 100% |
Table 3 Leave-one-out imputation error rates of 12 resequenced classical inbred strains using mouse HapMap SNPs, WTCHG SNPs, and gap-filling Perlegen SNPs. The fraction of imputed genotypes in each category is shown within a parenthesis.

| Genotype confidence | High confidence | Medium confidence | Low confidence | Total |
|---------------------|-----------------|-------------------|---------------|-------|
| Fully resequenced   | 0.27% (46.1%)   | 6.40% (2.79%)     | 19.0% (2.73%) | 1.59% (51.7%) |
| Mostly resequenced  | 0.40% (25.3%)   | 3.94% (3.50%)     | 16.1% (2.98%) | 2.26% (31.8%) |
| Poorly resequenced  | 0.76% (9.59%)   | 4.05% (2.62%)     | 15.8% (4.29%) | 5.18% (16.5%) |
| Total               | 0.37% (81.1%)   | 4.74% (8.91%)     | 16.8% (10.0%) | 2.40% (100%) |
Table 4 Imputation error rates of 47 inbred strains genotyped only in WTCHG SNPs, using mouse HapMap SNPs, and gap-filling Perlegen SNPs. The fraction of imputed genotypes in each category is shown within a parenthesis.

| SNP Quality | Genotype confidence | High confidence | Medium confidence | low confidence | Total |
|-------------|---------------------|-----------------|-------------------|---------------|-------|
| 36 classical inbred strains | 0.35% | 9.63% | 29.7% | 2.25% | (88.9%) | (6.74%) | (4.37%) | (100%) |
| All 47 strains | 0.37% | 8.85% | 27.0% | 4.86% | (71.7%) | (16.7%) | (11.5%) | (100%) |
Table 5 Inbred strains included in Paigen2 study phenotyped for HDL cholesterol. Note that in the Paigen2 study, they report 43 strains because they are also for some phenotype measurements available for BALB/cJ which we ignore.

| Paigen2 Classical Inbred Strains | Paigen2 Wild-derived Strains |
|----------------------------------|-----------------------------|
| 129S1/SvImJ                      | C57L/J                      |
| A/J                              | C58/J                       |
| AKR/J                            | CBA/J                       |
| BALB/cByJ                        | CE/J                        |
| BTBR T+ tf/J                     | DBA/1J                      |
| BUB/BnJ                          | DBA/2J                      |
| C3H/HeJ                          | FVB/NJ                      |
| C57BL/10J                        | I/LnJ                       |
| C57BL/6J                         | KK/HJ                       |
| C57BLKS/J                        | LP/J                        |
| C57BR/cdJ                        | MA/MyJ                      |
|                                 | NOD/ShiLtJ                  |
|                                 | NON/ShiLtJ                  |
|                                 | NZB/BlNJ                    |
|                                 | NZW/LacJ                    |
|                                 | PL/J                        |
|                                 | RF/J                        |
|                                 | RIIIS/J                     |
|                                 | SEA/GnJ                     |
|                                 | SJL/J                       |
|                                 | SM/J                        |
|                                 | SWR/J                       |
|                                 | CAST/EiJ                    |
|                                 | CZECHII/EiJ                 |
|                                 | JF1/Ms                      |
|                                 | MOLF/EiJ                    |
|                                 | MSM/Ms                      |
|                                 | PERA/EiJ                    |
|                                 | PWK/PhJ                     |
|                                 | SPRET/EiJ                   |
|                                 | WSB/EiJ                     |


Table 6 $\alpha=0.05$ pointwise p-value significance thresholds. The Paigen2 Inbred set only consists of the 33 classical inbred strains contained in Paigen2.

| Strain set        | 10 Megabase region | 20 Megabase region | 30 Megabase region |
|-------------------|--------------------|--------------------|--------------------|
| Paigen2 All (42)  | 0.0001568          | 0.0000399          | 0.0000092          |
| Paigen2 Inbred (33)| 0.0001821          | 0.0000463          | 0.0000097          |
| Paigen2 Modified (33)| 0.0002380     | 0.0000716          | 0.0000201          |
Table 7 Statistical power of *in silico* association mapping.

| Region Size | Strain Set | Genomic Effect | 25% Genetic Background | 50% Genetic Background | 75% Genetic Background |
|-------------|------------|----------------|------------------------|------------------------|------------------------|
|             |            | Effect         | 5% | 10% | 20% | 5% | 10% | 20% | 5% | 10% | 20% |
| 10 MB Region| Paigen2 Full (43) |                | 0.8553 | 0.9951 | 1.0000 | 0.3632 | 0.7680 | 0.9777 | 0.1429 | 0.3855 | 0.7449 |
|             | Paigen2 Inbred (33) |                | 0.6185 | 0.9207 | 0.9942 | 0.1891 | 0.4379 | 0.7689 | 0.0754 | 0.1723 | 0.3795 |
|             | Paigen2 Modified (33) |               | 0.8667 | 0.9897 | 1.0000 | 0.4145 | 0.7162 | 0.9395 | 0.1910 | 0.3825 | 0.6505 |
| 20 MB Region| Paigen2 Full (43) |                | 0.8060 | 0.9918 | 1.0000 | 0.3032 | 0.7084 | 0.9639 | 0.1078 | 0.3232 | 0.6883 |
|             | Paigen2 Inbred (33) |                | 0.5196 | 0.8783 | 0.9885 | 0.1305 | 0.3449 | 0.6859 | 0.0484 | 0.1207 | 0.2910 |
|             | Paigen2 Modified (33) |               | 0.7919 | 0.9785 | 1.0000 | 0.3043 | 0.6072 | 0.8919 | 0.1218 | 0.2736 | 0.5346 |
| 30 MB Region| Paigen2 Full (43) |                | 0.7517 | 0.9881 | 1.0000 | 0.2506 | 0.6518 | 0.9481 | 0.0830 | 0.2663 | 0.6270 |
|             | Paigen2 Inbred (33) |                | 0.4909 | 0.8652 | 0.9867 | 0.1156 | 0.3195 | 0.6622 | 0.0428 | 0.1067 | 0.2707 |
|             | Paigen2 Modified (33) |               | 0.7649 | 0.9739 | 0.9998 | 0.2770 | 0.5709 | 0.8737 | 0.1045 | 0.2427 | 0.4966 |
Table 8 Resolution (in megabases) of in silico association mapping using the Paigen2 dataset.

| Strain set          | 1<sup>st</sup> Quartile | Median | Mean | 3<sup>rd</sup> Quartile |
|---------------------|--------------------------|--------|------|-------------------------|
| Paigen2 All (42)    | 0.1137                   | 1.3570 | 2.7170 | 4.6930                  |
| Paigen2 Inbred (33) | 0.4276                   | 2.6610 | 3.5420 | 6.0960                  |
| Paigen2 Modified (33)| 0.2224                   | 1.4350 | 2.7440 | 4.7950                  |
Table 9 Modified inbred strain set to increase statistical power. We removed the wild-derived strains and 6 inbred strains which are very genetically similar to other strains in the study and replaced these strains with more distant classical inbred strains. The number of strains in the new set, 33, is equal to the number of classical inbred strains phenotyped in Paigen2.

| Paigen2 Overlap Strains | Additional Strains | Paigen2 Removed Strains |
|-------------------------|--------------------|-------------------------|
| 129S1/SvImJ              | LP/J               | CAST/EiJ                |
| A/J                     | MA/MyJ             | CZECHII/EiJ             |
| AKR/J                   | NOD/ShiLtJ         | JF1/Ms                  |
| BALB/cByJ               | NON/ShiLtJ         | MOLF/EiJ                |
| BUB/BnJ                 | NZB/BlNJ           | MSM/Ms                  |
| C3H/HeJ                 | NZW/LacJ           | PERA/EiJ                |
| C57BL/6J                | PL/J               | PWK/PhJ                 |
| C58/J                   | RF/J               | SPRET/EiJ               |
| CBA/J                   | RIIS/J             | WSB/EiJ                 |
| CE/J                    | SEA/GnJ            | BTBR                    |
| DBA/2J                  | SJL/J              | C57BL/10J               |
| FVB/NJ                  | SM/J               | C57BLKS/J               |
| I/LnJ                   | SWR/J              | C57BR/cdJ               |
| KK/HIJ                  |                    | C57L/J                  |
|                         |                    | DBA/1J                  |
Figure 2

Coverage of shared segments with 12 resequenced classical strains

Genomic coverage of shared segments

- Classical Inbreds
- Hybrids
- Wild Deriveds
Figure 3
Figure 4

Fraction of shared segments for each pair of strains

Number of strain pairs

Fraction of shared segments

0.05-0.1
0.1-0.15
0.15-0.2
0.2-0.25
0.25-0.3
0.3-0.35
0.35-0.4
0.4-0.45
0.45-0.5
0.5-0.55
0.55-0.6
0.6-0.65
0.65-0.7
0.7-0.75
0.75-0.8
0.8-0.85
0.85-0.9
0.9-0.95
0.95-1
Figure 5

The bar graph illustrates the number of phenotypes across different fractions of inter-strain phenotypic variance explained by population structure. The x-axis represents the fraction of variance explained (ranging from 0 to 1), and the y-axis shows the number of phenotypes. The graph shows a peak in the number of phenotypes around 0.9 fraction of variance explained, indicating a significant increase in phenotypic diversity at this level of variance explained.
Figure 6
Figure 7

Inflation factor of mixed model vs. Inflation factor of t-test

$y = x$