Published in final edited form as:
Mamm Genome. 2013 February ; 24(0): 1–20. doi:10.1007/s00335-012-9441-z.

Deconstructing *Mus gemischus*: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse

John P. Didion and
Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Fernando Pardo-Manuel de Villena
Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

John P. Didion: jdidion@email.unc.edu; Fernando Pardo-Manuel de Villena: fernando@med.unc.edu

Abstract

The laboratory mouse is an artificial construct with a complex relationship to its natural ancestors. In 2002, the mouse became the first mammalian model organism with a reference genome. Importantly, the mouse genome sequence was assembled from data on a single inbred laboratory strain, C57BL/6. Several large-scale genetic variant discovery efforts have been conducted, resulting in a catalog of tens of millions of SNPs and structural variants. High-density genotyping arrays covering a subset of those variants have been used to produce hundreds of millions of genotypes in laboratory stocks and a small number of wild mice. These landmark resources now enable us to determine relationships among laboratory mice, assign local ancestry at fine scale, resolve important controversies, and identify a new set of challenges—most importantly, the troubling scarcity of genetic data on the very natural populations from which the laboratory mouse was derived. Our aim with this review is to provide the reader with an historical context for the mouse as a model organism and to explain how practical decisions made in the past have influenced both the architecture of the laboratory mouse genome and the design and execution of

Correspondence to: John P. Didion, jdidion@email.unc.edu; Fernando Pardo-Manuel de Villena, fernando@med.unc.edu.

Dedicated to Kenneth and Beverly Paigen and François Bonhomme for their outstanding contributions to the field of mouse genetics.

Electronic supplementary material The online version of this article (doi:10.1007/s00335-012-9441-z) contains supplementary material, which is available to authorized users.

current large-scale resources. We also provide examples on how the accomplishments of the past decade can be used by researchers to streamline the use of mice in their experiments and correctly interpret results. Finally, we propose future steps that will enable the mouse community to extend its successes in the decade to come.

Introduction

The laboratory mouse is an artificial construct, absent from nature and shaped by human selection and chance. Mice, like all laboratory organisms, are experimental creatures, and as such:

> …are a special kind of technology in that they are altered environmentally or physically to do things that humans value but that they might not have done in nature" (Kohler 1994).

For more than a century, laboratory mice derived from *Mus musculus* have been one of the most widely used animal models. The mouse's status as the most popular mammalian model organism owes as much to serendipity as to a favorable combination of small size, short generation time, ease of manipulation, and a wide range of phenotypes with relevance to human physiology, behavior, and pathology. Here we review the breathtaking advances in our understanding of the genome of the laboratory mouse achieved in the past decade, with emphasis on determining how past decisions may have shaped current consensus and may be partly responsible for current controversies (Fig. 1). For example, this year marks the tenth anniversary of the public release of the mouse reference genome, which was based largely on the sequence of the C57BL/6 (B6) inbred strain (Waterston et al. 2002). The decision to sequence a single inbred strain (rather than create a consensus from multiple distinct individuals, as has been done for other species including humans (Li et al. 2009a)) has had long-term and wide-ranging implications. Among them are the efforts to catalog sequence variants present in laboratory mice, to create platforms for high-density genotyping, and to analyze and interpret the next-generation sequencing data now flooding the mouse community. The choice of B6 reflected a growing momentum toward the use of that strain in large-scale biological resources and also influenced the development of future resources. Those resources included genetic reference populations, engineered mutations, knockouts, and stem cell lines. In this review we discuss both the benefits and the drawbacks that have stemmed from those decisions. However, we do not question the ultimate wisdom of those decisions; rather we wish to make researchers aware of the context in which key resources were generated. For example, the majority of effort to identify mouse genetic variation has been focused on classical inbred laboratory strains, which represent only a minor fraction of the total genetic variation in the mouse (Ideraabdullah et al. 2004; Keane et al. 2011; Mural et al. 2002; Salcedo et al. 2007). As a consequence, attempts to study natural populations using currently available single-nucleotide polymorphism (SNP) marker panels and genotyping arrays must properly account for the biases inherent in such platforms (Boursot and Belkhir 2006; Didion et al. 2012; Harr 2006).

Another focus of this review is to connect laboratory stocks, which were shaped by artificial selection and breeding programs that are exclusive to the laboratory environment, with

natural mouse populations. We subscribe to Theodosius Dobzhansky's motto that "Nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973); thus, to properly interpret experimental data and to take full advantage of the exceptional research tool that is the laboratory mouse requires a correct understanding of the diverse and complex relationships between and within laboratory strains and their wild relatives.

The natural origin of an unnatural creature

There are approximately 5,400 mammalian species, of which about 560 are Murine rodents (Fig. 2a) (Wilson and Reeder 2005). Out of this menagerie, how did the mouse become the most common laboratory mammal and one of the most widely used model organisms in science? The small footprint of the mouse lends itself to living and breeding in close quarters, but neither this nor any other feature of its morphology is particularly remarkable compared to other rodents. In fact, the uniformity of genus *Mus* has caused great difficulty in interpreting and organizing the fossil record into a consistent and cohesive framework (Boursot et al. 1993). This has led to a dichotomy in taxonomic classification, with some treating mice as a single, highly polytypic taxon while others treat each new variant as a separate taxon (Schwarz and Schwarz 1943). Molecular analysis has mostly eliminated taxonomic redundancy and resolved the phylogeny of genus *Mus*; however, it has been inconclusive as to the branching order of the four discrete and largely monophyletic subgenera (Fig. 2b). This is due to the roughly concurrent divergence of the *Mus* subgenera from each other approximately 5 million years ago (Suzuki et al. 2004).

The biology of the mouse was important to its ascendancy as a model organism, but factors such as a short generation time (10–12 weeks) and high degree of genetic relatedness and physiological and pathological relevance to humans are common among rodents and other mammals. One aspect of *M. musculus* that is exceptional is its geographic range (Supplementary Fig. 1). House mice are found on all inhabited continents and islands, and in nearly every type of terrestrial ecosystem (Gabriel et al. 2010). Even more notable is that this dispersal has occurred over a relatively short evolutionary time scale. Fossil and molecular analyses place the ancestral range of the house mouse in the Himalayan foothills of Northern India and Pakistan (Bonhomme et al. 1994; Boursot et al. 1993, 1996; Din et al. 1996) and Central Asia (Prager et al. 1998; Rajabi-Maham et al. 2008) (Supplementary Fig. 1). *M. musculus* began to diverge approximately 500,000 years ago (Geraldes et al. 2008; Salcedo et al. 2007; Suzuki et al. 2004) into three distinct lineages (Fig. 2c; Supplementary Fig. 1) that we refer to here as subspecies (Boursot et al. 1993; Yonekawa and Takahama 1994), but which some have argued are species in their own right (Geraldes et al. 2008; Sage et al. 1993). Once established, the three subspecies expanded their ranges into Asia and the Middle East, but remained largely isolated over much of their history (Duvaux et al. 2011). Well before its deliberate introduction into the laboratory, the domestication of the mouse began as it formed a commensal relationship with humans approximately 15,000 years ago (Boursot et al. 1993). The development of long-distance modes of travel within the past ~1,000 years has accounted for the bulk of the house mouse's geographic dispersal.

This evolutionary history provides the context for why *M. musculus* became a favorite model to address broad scientific questions, including (1) genetics of adaptation, due to the

rapidity and extent of its dispersal and the readiness with which it has flourished in a diverse array of environments (Gabriel et al. 2010; Hardouin et al. 2010); (2) speciation, due to the ongoing divergence of the three subspecies and their secondary contact at several hybrid zones around the world (Mihola et al. 2009); and (3) as a means of tracking the historical movements of human populations (Gabriel et al. 2010, 2011; Jones et al. 2011, 2012). In addition, the house mouse is an agricultural pest and a vector for disease, and there is ongoing research focused on combating its economic and public health impact (Brown and Singleton 2002; Meerburg et al. 2009; Stenseth et al. 2003). Finally, the house mouse is also an invasive species that poses a significant threat to sensitive native flora and fauna. This is especially true for isolated and frequently unique ecosystems such as islands, and there are efforts to combat this problem (http://www.islandconservation.org/).

All of these scientific applications of the house mouse pale in comparison to its laboratory use as a model organism, but its dominance in the laboratory is to some degree the result of happenstance. Although they are morphologically uniform, mice display extensive variation within a few outward traits such as coat color and tail length. Mendel began his studies of inheritance using coat color traits of mice. Unfortunately, the objections of a shortsighted local bishop robbed the mouse of its place of honor in genetics textbooks and forced Mendel to turn to pea plants instead (Paigen 2003a). Those same traits also made house mice attractive as pets that could be crossed and selected for appearances and behaviors that were deemed favorable by mouse fanciers. As with other pets, selective breeding eventually led to inbreeding (Boyko et al. 2010). It was from one of those fancy mouse stocks, maintained by breeder Abbie Lathrop, that William Castle and his student Clarence Little developed the first laboratory strains in the early 20th century as a genetically uniform and reproducible model organism for studying the heritability of disease (Fig. 3) (Beck et al. 2000; Festing 1997). The relationship between scientists and mouse breeders, and the momentum that followed the success of early mouse experiments, accounted in large part for the mouse's dominance in the laboratory today.

Expanding genetic and phenotypic diversity in the laboratory

Following the creation of the first mouse strains, additional stocks were initiated in Europe, China, and Japan as well as America (Beck et al. 2000). Mice from those colonies were intercrossed and interesting spontaneous mutations were selected and fixed to generate a large catalog of so-called "classical" strains (Fig. 4a). Classical strains are related to standard outbred stocks (Yalcin et al. 2010), whose genome is derived in large part from fancy mice brought from Switzerland to the US by Clara Lynch in 1926. The number and diversity of classical inbred strains and outbred stocks expanded dramatically during the second half of the 20th century.

In the past 30 years, new types of mouse resources have been developed to expedite the search for genetic causes of both simple (Mendelian) and complex traits. On one hand, genetic reference panels have been created using updated versions of traditional breeding methods. Those include gene- and chromosome-replacement strains (congenics and consomics, respectively) (Gregorová et al. 2008; Singer et al. 2004; Takahashi et al. 2008), recombinant inbred panels (Collaborative Cross Consortium 2012; Hrbek et al. 2006; Peirce

et al. 2004), advanced intercross lines, and outbred stocks (Cheverud et al. 2001; Mott et al. 2000; Svenson et al. 2012). Those panels often prove more powerful for genetic mapping than classical strains or traditional intercrosses because of their greater genetic and phenotypic diversity. For example, lines of the emerging Collaborative Cross (CC) population display a much wider variation in traits, such as gene expression, body weight, wheel running, response to allergens, and infectious diseases, than do the CC founder strains (Aylor et al. 2011; Bottomly et al. 2012; Collaborative Cross Consortium 2012; Kelada et al. 2012; Mathes et al. 2011). On the other hand, mutagenesis and genetic engineering techniques have enabled the introduction of new variation, up to complete human genes, into the well-characterized genetic background of inbred strains (Austin et al. 2004; Paigen 2003b). Beginning in the 1980 s, several staff members at The Jackson Laboratory began to introgress many mutations into the B6 background. This effort had a role in the choice of B6 as the index mouse genome.

Additionally, strains derived from wild-caught mice have dramatically expanded the available pool of genetic variation. The term “wild-derived strain” is loosely used to refer to any laboratory mouse that is not derived from the same common genetic pool as the classical strains. The depth of the wild-derived strain resources that are available (Fig. 4b) is not widely recognized because of the reliance on just a few popular wild-derived strains in high-profile efforts (discussed later). Wild-derived strains have expanded the phenotypes and behaviors available to researchers and are used as models of disease, speciation, chromosomal evolution, and behavioral genetics (Blanchet et al. 2011; Guénet and Bonhomme 2003). There are also several wild-derived strains that have been created from other *Mus* species (Fig. 2b). *M. spretus* has been the longest and most widely used non-*M. musculus* mouse model, in part due to the ability to create fertile interspecific crosses despite ~1.5 million years of evolutionary divergence between the two species (Bonhomme and Selander 1978; Orth et al. 2002). Those crosses were invaluable for the generation of the first complete mouse linkage map (Dietrich et al. 1996).

Although wild-derived inbred strains are behaviorally more “wild” than their classical cousins, researchers should not assume that they are wild in the sense of being “natural.” In fact, wild-derived inbred strains have been shaped by many of the same artificial processes as classical strains, such as selection, inbreeding, and intercrossing with both classical and other wild-derived strains.

The mouse genome: all our eggs in one B(6)asket

The post-genome era in the house mouse started with the complete-genome sequencing of a single laboratory mouse strain (Waterston et al. 2002). After careful consideration, a committee of experts decided that B6 should be used to construct the bacterial artificial chromosome (BAC) library that would serve as the index mouse genome, citing “confidence of strain derivation, widespread use among the research community and favorable breeding characteristics” (Battey et al. 1999). Concurrent with the sequencing of B6, a private company (Celera) sequenced three additional strains (Mural et al. 2002), although until 2005 that data was kept private and made available on a subscription basis. The main contribution of that project was to assemble the first catalog of sequence variants in the mouse.

It is difficult to overestimate the significance of choosing a single classical inbred strain for the mouse reference genome. First, the use of a single inbred strain precluded the identification of genetic variation because those strains are almost completely homozygous. In contrast, 36 wild mice, genotyped at high density using the mouse diversity array (MDA) (Yang et al. 2009), had an average of 8.4 % (±2.9 %) heterozygous SNPs (Yang et al. 2011). Second, a laboratory mouse strain is the product of both artificial selection and domestication. Those processes favor certain phenotypes (e.g., docility) that may be rare in nature, and the associated alleles are now incorporated into the reference genome. Third, for all practical purposes, the reference genome assembly is the result of sequencing a single chromosome from a single cell from a single mouse. Therefore, variation that is exclusive to that specific DNA molecule and to B6 in general is now part of the reference sequence. Furthermore, structural variation (insertions, deletions, inversions, duplications) and transposable element variation that is private to B6 or rare in the house mouse is invisible or hard to interpret in sequences from other strains (Nellåker et al. 2012; Yalcin et al. 2011). For example, deep sequencing of 17 laboratory strains revealed 30 Mb of sequence that was conserved across multiple strains but was absent from the reference sequence (Keane et al. 2011).

The selection of a single reference strain has also served to amplify the inertia toward B6 as the default choice in any new scientific endeavor, even when another inbred strain (or wild mice) may be more appropriate. PubMed searches comparing the century preceding the decision to sequence B6 with the 14 years that have followed showed a 3-fold increase in papers containing the term “C57BL/6” versus only a 1.3-fold increase in papers containing the terms “mouse inbred strain” (after controlling for total number of papers published). B6 is a major (and in many cases exclusive) component in nearly every large project or resource involving the mouse genetics community, including the International Knock-Out Mouse Project (Austin et al. 2004), the CC (Collaborative Cross Consortium 2012), the Diversity Outcross (Svenson et al. 2012), and the expansion of the BXD recombinant inbred panel (Peirce et al. 2004).

We expect that advances in genome assembly methods will largely address the shortcomings associated with the use of a single reference strain. The Genome Reference Consortium (Church et al. 2011) recently released an update to the mouse genome assembly that included 83 Mb of additional sequence (mostly on chromosome Y) and closed 200 gaps (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/mouse/index.shtml). The availability of complete genome assemblies from multiple genetic backgrounds (Keane et al. 2011; Wong et al. 2012) has enabled the construction of pseudo-genome references to improve read mapping and variant calling for new sequences. A pseudo-genome is created by stitching together parts of existing whole-genome sequences based on their local genotype similarity to the new genome being assembled (Welsh et al. 2012). In addition, de novo assembly (i.e., assembly in the absence of a reference) of sequences that are very different from the standard reference (such as those of *M. m. musculus* and *M. m. castaneus* origin) may improve coverage of sequence that is absent from the standard reference (Yalcin et al. 2012a). With the accumulation of whole-genome sequences in inbred strains (and

eventually wild mice), it may be beneficial to create a mouse “pan-genome” reference (Yalcin et al. 2012a).

Expanding the catalog of mouse genetic variation

It was immediately recognized that the strategy of sequencing a single inbred strain required complementary efforts to identify genetic variation. A large catalog of variants was developed using three additional inbred strains (Mural et al. 2002), followed by more extensive projects conducted by Perlegen Sciences, on behalf of the NIEHS (Frazer et al. 2007), and the Mouse Genomes Project at the Sanger Institute (Keane et al. 2011; Yalcin et al. 2011, 2012a, b; Wong et al. 2012). Although the NIEHS and Sanger projects were conducted only a few years apart, advances in sequencing technology enabled a fourfold improvement by the later group upon the initial NIEHS catalog of 8.27M SNPs (Table 1). In addition, the Sanger project extended the catalog to include structural and regulatory variation (Nellåker et al. 2012; Shen et al. 2012; Yalcin et al. 2011, 2012b). Those projects were instrumental in the exceptional productivity of the mouse genetics community in recent years, including the construction of improved genetic maps (Brunschwig et al. 2012; Cox et al. 2009), the discovery and fine-mapping of causative variants for a number of traits and the development of new mouse reference panels and better use of existing mouse resources (Collaborative Cross Consortium 2012, Box 1).

Due to cost constraints, only a select group of strains could be surveyed in each of the sequencing projects. The strains were chosen using criteria such as popularity, expected genetic dissimilarity (based on pedigree and small-scale genotyping and sequencing), and “taxonomical” classification. Additionally, some strains were selected for their relevance to large-scale community efforts or to the study of specific diseases. For example, the Sanger project included four strains that represent the genetic background of embryonic stem (ES) cell lines used in knockout experiments: C57BL/6N (a substrain of B6) and three related 129 strains (Guan et al. 2010). Of the 15 strains selected by NIEHS and 17 selected by Sanger, 9 were common between the two projects, 17 were classical strains, and 22 were of *M. musculus* origin (*M. spretus* was represented in the Sanger project by SPRET/EiJ). It is important to note that despite the overrepresentation of classical strains in those projects, the vast majority of all SNPs were identified in the handful of wild-derived lines analyzed (Table 1). For example, of the 53.7M SNPs that were discovered in the Sanger project, only 3M were found to be private to classical strains, and of those, only 0.7M were private to a single inbred strain (Keane et al. 2011). It is evident that the return-on-investment of future sequencing efforts (i.e., SNPs discovered per dollar) will be much greater for wild mice and wild-derived strains than for classical strains.

The NIEHS and Sanger projects represent the vast majority of genomic variants that have been discovered in the house mouse. However, the 23 strains surveyed in those projects are only a small subset of all the strains that are actively used in laboratories. To minimize the expense of extending the benefits of whole-genome sequencing to additional strains, the MDA was designed from maximally informative subsets of the SNP catalog (Yang et al. 2009). The MDA was used to genotype hundreds of the most commonly used classical and wild-derived strains, as well as wild-caught mice (Didion et al. 2012; Yang et al. 2011).

Among other applications, the MDA genotypes were used to create a high-resolution map of the variation present in classical strains (Yang et al. 2011).

Genotype imputation: the thousand-dollar genome today

Although the dense genotype data that are available for most commonly used laboratory strains have enabled considerable improvement in the resolution and accuracy of studies that use those strains, several applications, such as the identification of causal variants, still require single-base resolution. Fortunately, recent advances in genotype imputation methods (Wang 2012a, Box 1) have made a large fraction of the complete catalog of genetic variants available to any strain at the relatively affordable cost of array genotyping.

Genotype imputation in humans and other outbred populations with large effective sizes requires a substantial number of sequences to achieve a low error rate (Li et al. 2009b). An exception to this is related individuals; for example, the genomes from a mother and father can be used to impute variants in their children. The fact that the classical strains are essentially a large family with inbred progeny that can trace their lineage from a small number of ancestors has proved instrumental in the ability to perform genotype imputation. The high degree of population structure in classical strains means that in most regions of the mouse genome, a local phylogenetic tree based on dense genotypes has a small number of branches, and the strains located at each leaf of the tree are essentially identical to each other (excluding private mutations that have arisen in the 100 years since their derivation). In 88 classical strains for which only MDA genotypes were available, Wang et al. (2012a, b) were able to identify an appropriate reference among 12 strains sequenced by the Sanger project across an average of 92 % of the genome. In those regions, they were able to impute 977M new genotypes (11.1M per strain on average) with an error rate of ~0.4 %. This was a significant improvement over previous genotype imputation efforts (Kirby et al. 2010; Szatkiewicz et al. 2008). In the regions where no appropriate Sanger reference existed for a strain, genotypes could not be imputed. The distribution of those regions was bimodal, with a minority of strains accounting for the majority of missing genotypes. The missing haplotypes in those regions probably originated from the small number of Asian (*M. m. molossinus*) or Swiss fancy mice that have contributed to the classical strain genome (Nagamine et al. 1992). Those regions have highlighted gaps in our knowledge of the genome of the laboratory mouse and identified the strains for which whole-genome sequencing would provide the most information (Wang et al. 2012a).

Unfortunately, the improvements achieved in classical strains have not extended to wild-derived strains and wild mice. In contrast to the family-like structure of classical strains, wild mice are more diverse than the human species. The effective sizes (and thus the haplotype diversity) of natural house mouse populations are between 6- and 20-fold greater than humans (Geraldes et al. 2008) and thousands of times greater than classical strains (Yang et al. 2011), meaning that a large number of reference sequences will be required for high-quality imputation.

Haplotype blocks: building a picture of genomic diversity

The known pedigrees of laboratory strains (Beck et al. 2000) indicate that all extant classical strains were derived from a small founder population. The founders had sufficient genetic diversity such that combining them to create the progenitors of the classical strains resulted in substantial heterozygosity (Bonhomme et al. 1987). Each strain experienced several generations of haplotype shuffling due to recombination before it became inbred. For studies in which multiple strains are compared, such as genetic mapping, it is important to know the level of genetic diversity that existed within the founder population and how that diversity is organized in the genome of classical strains.

A useful unit for the analysis of genome organization is a haplotype block, a contiguous interval in which the number of unique sequences (haplotypes) is much smaller than the total number of sequences due to a high degree of genetic similarity (approaching identity) within subsets of strains. A natural criterion to define haplotype blocks in classical strains is to identify regions of shared ancestry among multiple strains which have not recombined (compatible intervals) (Wang et al. 2010; Yang et al. 2011) using the 4-gamete rule (Hudson and Kaplan 1985), although other approaches have been tried (Frazer et al. 2007; Yalcin et al. 2004). Yang et al. (2011) used the 4-gamete rule to identify 43,285 haplotype blocks with a median size of 71 kb in 100 classical strains. The majority of blocks contained between four and six haplotypes, and there were fewer than ten haplotypes across 97 % of the genome. Those findings confirmed the small size of the classical strain founder population. The larger numbers of haplotypes in the remaining 3 % of the genome were due to a combination of new mutations in the past century and contributions from outside of the founder population. Blocks with large numbers of different haplotypes should be further investigated to understand their origins.

There is a significant risk of introducing bias into a study when local differences in haplotype diversity are not accounted for (Boursot and Belkhir 2006; Harr 2006). For example, a high LOD score in a genome-wide association study (GWAS) may be due to the presence of a causal allele, or it may be an artifact of local population structure (Flint and Eskin 2012). Such local structure is often overlooked when a pedigree or global phylogeny is used to select strains for an intercross (Pamilo and Nei 1988). Supplementary Fig. 2 provides an example of a local phylogenetic tree within a region containing the *Pgk2* gene, which is essential for male fertility (Danshina et al. 2010). The local tree (Supplementary Fig. 2a) is substantially different from the global phylogenetic tree (Fig. 4a); some strains that are closely related in the global tree are discordant in the local tree and vice versa. For example, an intercross between A/J and B6 (Groups L3 and L2, respectively), selected on the basis of their diversity in the global tree, would severely limit the ability to identify functional genetic variation in the *Pgk2* region because of their local similarity. The local haplotype map (Supplementary Fig. 2b) can guide the selection of strains for modifier screens.

Haplotype structure also has important implications for the ability to conduct genetic mapping because it can significantly affect the level and rate of decay of linkage disequilibrium (LD) (Collaborative Cross Consortium 2012; Laurie et al. 2007). Gametic

disequilibrium (GD), which is also known as long-range LD, is problematic because it can introduce false genotype–phenotype associations (Burgess-Herbert et al. 2009; Kang et al. 2008). An analysis of LD decay in a panel of 88 classical strains revealed widespread GD (Collaborative Cross Consortium 2012), suggesting caution when interpreting the results of mapping experiments in those strains. The effect of population structure can be reduced by using a genetic reference population (such as the CC) or wild-caught mice (Flint and Eskin 2012).

The local ancestry of classical strains

Thirty years ago it was discovered that laboratory mice do not belong to a single taxa but rather represent a mosaic between multiple *M. musculus* subspecies (Bishop et al. 1985; Bonhomme et al. 1987; Moriwaki et al. 1982; Paigen 2003a). Some have even suggested that the laboratory mouse be given it own taxonomic designation, *Mus laboratorius* (Artzt et al. 1991; Guénet and Bonhomme 2003) or *Mus gemischus* (gemisch is a Yiddish word meaning “mixture”) (Paigen and Eppig 2000). There is no doubt that the laboratory mouse genome is a mosaic, but the quantity and distribution of the contribution from each subspecies has been fiercely debated. A popular model was that the ancestry of the laboratory mouse was a roughly equal mixture of European *M. m. domesticus* and Japanese *M. m. molossinus* (a natural hybrid between*M. m. musculus* and*M. m. castaneus*, see Supplementary Fig. 1) (Yonekawa et al. 1988). This view had a pervasive influence in the planning and interpretation of SNP discovery efforts.

Several groups have recently presented results on the subspecific origin of laboratory mice using the newly available genotyping (Frazer et al. 2007; Yang et al. 2007, 2011) and sequencing (Keane et al. 2011) platforms. The sets of strains used in those studies were different but highly overlapping (Supplementary Table 1). In each study, the authors chose one or more samples to serve as a reference for each *M. musculus* subspecies. They then examined the local phylogenetic relationships among strains (called strain distribution patterns (SDPs)) in small regions spanning the genome. Within each region, they attempted to assign a subspecific origin to each group of related strains based on the reference sample(s) that clustered with the group. Remarkably, the local concordance between SDPs was high across all studies despite the use of distinct genotype data sets that differed in density by several orders of magnitude. However, in spite of the local agreement between phylogenetic relationships, the studies drew opposite conclusions about the ancestral origin of the laboratory mouse genome.

In 2007, Frazer et al. analyzed the NIEHS data and concluded that the ratio of *M. m. domesticus* to non-*domesticus* (or unknown) ancestry in the classical strains was about 2:1, a finding that supported the traditional mosaic model (Wade et al. 2002). Using the same data set, Yang et al. (2007) determined that classical strains are primarily of *M. m. domesticus* origin (92 %), with only a minor contribution from *M. m. musculus* and *M. m. castaneus* (6–7 and 1–2 %, respectively). In 2011, Yang and colleagues conducted a similar analysis using dense genotyping in a larger sampling of strains and their results confirmed their earlier conclusions. That study was later improved using genotypes imputed from whole-genome sequence to further refine subspecific assignments (Box 1). Concurrently, Keane et al.

(2011) challenged the rationale of conducting local subspecific origin assignment across the genomes of the classical strains on the basis of widespread phylogenetic discordance.

The conflicts between those studies stemmed from the choices of reference samples and in the interpretation of the relationship between those references and the classical strains. First, the studies disagreed over the number of reference samples for each taxon that were necessary for accurate local ancestry assignment. In the studies based on the NIEHS and Sanger data (Frazer et al. 2007; Keane et al. 2011; Yang et al. 2007), only a single inbred (homozygous) reference sample was available for each taxon, whereas the study based on MDA data (Yang et al. 2011) used 10 *M. m. domesticus*, 16 *M. m. musculus*, and 10 *M. m. castaneus* wild (heterozygous) reference samples. The number of available reference samples is important because probabilistic methods for local ancestry assignment depend on the marker allele frequencies within each population (or taxon). Variants for which allele frequencies differ markedly in one reference population compared to the others (diagnostic alleles) are the most important for the correct assignment of local ancestry, while those with allele frequencies that are similar across taxa (ancestral variants) or private to a single individual or subpopulation (private variants) are less useful. Furthermore, the genome of each inbred strain represents a single ancestral chromosome at each position because it has been forced into homozygosity, whereas wild-caught samples are capable of representing two different chromosomes. That means as many as 72 haplotypes were available to Yang and colleagues for subspecific origin assignment in each genomic region. When only a single reference was available from each taxon, correctly establishing allele frequencies was difficult and error-prone, and the resulting set of diagnostic alleles contained significant gaps (Yang et al. 2007). In contrast, Yang et al. (2011) identified 347,864 diagnostic alleles at 60 % of available markers (for some markers, multiple alleles were diagnostic), which enabled them to assign local ancestry across the entire genome of classical strains. While a marked improvement over previous studies, the sampling by Yang and colleagues fell short of capturing the full extent of variation present in natural populations. Most importantly, the ten *M. m. castaneus* samples were all captured within a ~200-mile radius in northern India. Recent evidence that *M. m. castaneus* is polytypic and harbors several quite divergent subgroups (Rajabi-Maham et al. 2012) indicates that broader sampling will be required to fully characterize the ancestry of laboratory strains.

The second disagreement between the studies was over whether wild-derived laboratory strains were appropriate to serve as reference samples, or if it was instead necessary to sample mice from nature (see below).

Barking up the wrong tree: introgression in wild-derived reference strains

The ideal reference for each *M. musculus* taxon would comprise samples from the natural populations found within its native range, similar to the HapMap catalog of population diversity that has been developed for humans (The International HapMap 3 Consortium 2010). However, developing such a reference has not been a priority in the mouse genetics community. Instead, researchers have used wild-derived laboratory strains under the assumption that those strains are faithful proxies for their wild ancestors. That assumption guided the design and interpretation of the Frazer et al. (2007) and Keane et al. (2011)

studies, as well as other projects (Dumont and Payseur 2011; Mihola et al. 2009; White et al. 2009). However, that assumption fails to account for the widespread problem of introgression in the wild-derived strains.

Introgression is the movement of variants from one population into the gene pool of another population by the repeated backcrossing of a hybrid to one of its parent populations. Introgression can occur between individuals of the same subspecies that belong to different subpopulations (e.g., two *M. m. musculus* mice, one from Poland and one from China, crossed in the laboratory) or between individuals of different subspecies. While all members of the same subspecies are generally able to interbreed, there are several barriers to intersubspecific mating (Macholán et al. 2007; Mihola et al. 2009; Tucker et al. 1992; White et al. 2011). The general effect of intersubspecific incompatibilities is to limit the extent of gene flow between subspecies by reducing the fertility and/or fecundity of hybrid animals. This means that in wild mice, introgression typically exists on a small scale and is difficult to observe, even with high-density genotype data. However, exceptions occur in places with a high rate of mixing between individuals of divergent genetic backgrounds (Staubach et al. 2012). Those regions are known as hybrid zones, and they may be natural or man-made.

The derivation of new wild-derived strains has in large part been driven by a few fields of study, such as hybrid zone biology, that dictate the local populations from which founders are drawn. For example, the PWK/PhJ and PWD/PhJ strains were established from mice trapped near the European zone where *M. m. musculus* and *M. m. domesticus* hybridize (Supplementary Fig. 1) (Gregorová and Forejt 2000). It is therefore not surprising that those strains, while primarily *M. m. musculus*, harbor a sizable genetic contribution from *M. m. domesticus* (6.1 and 7.0 %, respectively) (Fig. 4b). Other strains have been derived from mice trapped in or near transportation hubs. Mice that are introduced into cities and seaports from widely dispersed geographic (and thus subspecific) origins by passive transport on human vessels will inevitably interbreed to create mosaics. For example, CAST/EiJ was established from mice trapped in Bangkok, Thailand, a large city within the natural range of *M. m. castaneus* but also with close proximity to a seaport. Approximately 12 % of the CAST/EiJ genome was derived from non-*M. m. castaneus* origin (7.9 % *M. m. domesticus*, 3.8 % *M. m. musculus*) (Fig. 4b). Regions of introgression in PWK/PhJ and CAST/EiJ are reliably detected in data generated by different methods at different densities (Fig. 5). The nonlinear relationship between genetic and geographic distance in wild-derived strains can clearly be observed using pairwise comparison (Supplementary Fig. 3). Remarkably, of 63 wild-derived strains analyzed in a recent study (Yang et al. 2011), only a minority was derived from a single genetic background (Fig. 4b).

The laboratory environment also brings together in close proximity populations that are very distant in the natural environment and provides abundant opportunity for interbreeding. The same challenges to intersubspecific mating exist in the lab as in the wild. For example, at least 80 % of CC lines have gone or will go extinct, due in large part to intersubspecific incompatibilities (Chessler et al. 2008; D. Aylor and F. Pardo-Manuel de Villena, unpublished). However, the impact of a successful interbreeding in the lab is much more extreme than that in the wild because introgression can quickly become fixed in a small inbreeding colony. While most breeding in the lab is directed, unintentional mating does

occur and can lead to contamination of the entire inbred strain. In contrast to introgressions in wild mice, contaminations are readily detected in pairwise comparisons of genotype data by a high rate of identity to another strain (Yang et al. 2011).

While both Frazer et al. (2007) and Keane et al. (2011) conducted their analyses under the assumption that each *M. musculus* subspecies (or pseudospecies in the case of *M. m. molossinus*) could be accurately represented by a single wild-derived strain, Yang et al. (2007, 2011) allowed that wild-derived strains may themselves be mosaics due to introgression. In their analysis of the NIEHS data (Yang et al. 2007), they used SNPs that segregated among the three wild-derived strains (diagnostic variants) to identify a number of large and nonrandomly distributed regions in which the nucleotide diversity between two or more of the wild-derived strains was drastically reduced (Fig. 5; Supplementary Fig. 4). Such a reduction in diversity indicated a region of intersubspecific introgression, which had profound implications for subspecific origin assignment of classical strains in that region. They found that the wild-derived strains could be used only to confidently infer subspecific origin in 72 % of the genome, while up to 13 % of the individual wild-derived strain genomes exhibited introgression. Those introgressions were largely confirmed in the later study (Yang et al. 2011). Some of those introgressions were clearly shared among multiple and otherwise genetically and geographically distinct wild-derived strains (Fig. 2 of Yang et al. 2011). Shared introgressions were evidence of cross breeding between classical strains and wild-derived strains, and in many cases a single classical inbred strain donor could be identified.

In order to overcome the shortcomings of the NIEHS data set and to analyze the ancestry of the entire genome, Yang et al. (2011) genotyped at high density 36 wild mice trapped in each of the native ranges of the three *M. musculus* subspecies (Supplementary Fig. 1). The wild-mouse genotypes enabled them to ascertain a much more robust set of diagnostic alleles, with which they were able to assign local ancestry at high confidence across the entire genome of all laboratory strains. The results of that study both confirmed the existence of introgression in most wild-derived strains and supported the conclusion that while classical strains are mosaics of the three *M. musculus* subspecies, *M. m. domesticus* is by far the dominant “color” in that mosaic.

In contrast to the studies based on genotyping array data, Keane et al. (2011) used whole-genome sequence and a sophisticated analysis based on Bayesian concordance of local phylogenetic trees determined in regions of orthology between *M. musculus* (represented by three inbred strains), *M. spretus* (represented by a single inbred strain), and *Rattus*, represented by the whole-genome sequence of the rat (Rat Genome Sequencing Project Consortium 2004). The goal of that experiment was to determine the history of speciation in *M. musculus*; however, they used nearly the same wild-derived strains as did Frazier and colleagues (substituting PWK/PhJ for the closely related PWD/PhJ) and also made the same assumption as the earlier study that those strains were accurate representatives of the three *M. musculus* subspecies. Keane and colleagues concluded that of the three possible subspecies histories, the most likely one placed *M. m. domesticus* as an outgroup to *M. m. musculus* and *M. m. castaneus* (i.e., *M. m. domesticus* was the first subspecies to diverge from the ancestral *M. musculus* species). This agrees with the maximum-parsimony tree that

we created based on MDA genotypes (Fig. 2c) and with previous studies (Boursot et al. 1996; Tucker et al. 2005). However, Keane et al. found that only about 39 % of the genome supported the primary history, as opposed to about 30 % for each of the other two possible histories. They inferred that ancestral polymorphism is common, that ancestral polymorphism was the cause of the observed phylogenetic discordance, and that phylogenetic discordance precludes the ability to assign subspecific origin to genomic regions.

Although ancestral polymorphism certainly exists within *M. musculus* (Bonhomme et al. 1994; Ideraabdullah et al. 2004; Keane et al. 2011; Salcedo et al. 2007), the fact remains that each subspecies harbors a combination of ancestrally segregating alleles and alleles that are private to that subspecies (Supplementary Fig. 5). Yang et al. (2007, 2011) demonstrated that diagnostic alleles are common and evenly distributed in the genome of *M. musculus*. Their method accounted for local introgression by allowing a small amount of discordance in identifying diagnostic alleles. For example, if nine *M. m. domesticus* samples were homozygous for allele A at a particular marker and the tenth sample was heterozygous, whereas all *M. m. musculus* and *M. m. castaneus* samples were homozygous for allele B, it was most likely that allele A was private to *M. m. domesticus* and the single B allele was the result of introgression, recent mutation, or genotyping error. When applied to a larger and nonoverlapping set of samples from additional *M. m. domesticus* and *M. m. musculus* populations, their method yielded diagnostic SNPs with 92 % agreement with those identified in the original study (J. P. Didion, J. B. Searle, and F. Pardo-Manuel de Villena, unpublished). Furthermore, the introgressions identified in wild-derived strains based on MDA data are clearly present in the NIEHS and the Sanger data sets as well (Fig. 5; Supplementary Fig. 4), indicating that the method used by Yang and colleagues should yield similar results when applied to whole-genome sequences as better reference samples become available.

It is important to note that the subspecific origin assignment in Yang et al. (2011) pertained to Mb-long regions. Although the local phylogenies created from the MDA and Sanger data sets were concordant over relatively large regions, there were fine-scale inconsistencies that represented either genotyping errors in the MDA genotypes for the wild mouse reference samples or errors in the interpretation of the genotype data. The robustness of the diagnostic alleles argued against genotyping error. Instead, it was most likely that the resolution of the MDA data and the hidden Markov model used to assign subspecific origin were not sensitive enough to recognize small-scale haplotype switching (e.g., see the highlighted region of Fig. 5a).

We believe that the current lack of data on the relative abundance of ancestral and private SNPs in house mouse populations and the difference in resolutions of the MDA and Sanger data mostly account for the apparent controversy between the two studies. We agree with Keane et al. (2011) that as whole-genome sequencing becomes more affordable and we are able to sample a greater number of natural populations at a finer scale, a clearer picture of the evolutionary history of the house mouse subspecies, and thus of the local ancestral origin of laboratory mice, will emerge. A recent example of this is the improvement of the

subspecific assignments from Yang et al. (2011) for several strains by making use of the whole-genome sequences from the Sanger project (Box 1).

In light of the discordance between past assumptions and current knowledge of wild-derived strain ancestry, studies that have made conclusions based on an incorrect or incomplete knowledge of that ancestry may benefit from reanalysis. For example, previous estimates of population size and mutation rate in *M. musculus* that relied on wild-derived strains may have underestimated the level of variation present in ancestral populations and thus overestimated ancestral population sizes (Phifer-Rixey et al. 2012). It is also apparent that genetic reference populations derived in part or wholly from partially introgressed wild-derived strains, such as the CC and the Diversity Outcross (DO), harbor less genetic variation locally than was originally predicted. Now that the genetic makeup and subspecific origin of many wild-derived strains are known, researchers are in a position to create new mouse reference panels are better optimized for genetic and phenotypic variation.

Additionally, caution must be used in conducting association mapping with wild-derived strains because differences in subspecific origin can introduce GD. To demonstrate this, we used MDA genotypes from 62 wild-derived strains (Yang et al. 2011) to create genome-wide maps of LD. Pervasive high GD due to population structure was expected when strains from multiple taxa were examined together (Supplementary Fig. 6a). However, GD was also observed when only strains from a primarily *M. m. domesticus* background were considered (Supplementary Fig. 6b). Although it is possible that this GD is due to population structure, there is strong evidence against significant intrasubspecific population structure in wild mice (Salcedo et al. 2007). Rather, we attribute the existence of GD in *M. m. domesticus* wild-derived strains to both intersubspecific introgression and contamination from laboratory strains (Yang et al. 2011).

Wild mice: ending at the beginning

Before the decade is out, at least 1,000 humans will have had their genomes sequenced (1000 Genomes Project Consortium 2010). Unfortunately, there is no similar plan in place for the house mouse. Despite the shortcomings of association studies in laboratory mice (Eichler et al. 2010) and evidence that wild mice present a deep reservoir of genetic diversity (Salcedo et al. 2007), there has never been whole-genome sequence produced for a wild mouse. This is a pressing need in mouse genetics and can be addressed by the establishment of a small consortium to prioritize and sequence a select group of wild mice.

The relative lack of genetic variation in classical strains limits their utility in at least two respects. First, it constrains the phenotypic variation that exists in classical strains. Second, use of classical strains is inappropriate to study evolutionary processes for many reasons, including that they may be invariant for many of the genes involved in speciation (Piálek et al. 2008). The extent of additional variation present in natural populations of *M. musculus* is hinted at by limited studies in wild mice (Salcedo et al. 2007) and by the recent whole-genome sequencing of three wild-derived strains (Table 1) (Keane et al. 2011), but it is not known for certain. To examine sequence variation at the genome scale, we determined the nucleotide diversity in classical, wild-derived, and wild-caught mice and MDA genotype

data from Yang et al. (2011). We used the method of Nei and Li (1979) to compute the average pairwise genetic distance between individuals within a population (π). Overall, we found greater diversity in wild-derived and wild mice than in classical strains (π = 0.298, 0.282, and 0.203, respectively). The contrast is even more striking, however, when comparing diversity between regions with different subspecific origin (Fig. 6). In classical strains, intervals derived from European fancy mice founders (*M. m. domesticus*), which represent the majority of classical strain founder haplotypes, had 6 and 17 times greater diversity than intervals derived from the minority Asian fancy mouse founders (*M. m. musculus* and *M. m. castaneus*, respectively) (Fig. 6a), whereas in wild-derived lines (Fig. 6b) and wild mice (Fig. 6c), variation is similar among regions of different ancestry. This confirms an earlier finding that across most of the genome there is at most one non-*M. m. domesticus* haplotype (Yang et al. 2011).

A practical answer to the question of why so little attention has been paid to natural populations of mice is that mouse genetics has historically been driven by biomedical research, which relies on an endless supply of genetically identical animals in order to prove reproducible results under controlled conditions. Laboratory strains are easier to produce, house, and handle than their truly wild counterparts. In this sense, it may seem that wild mice have no place in the laboratory. However, wild mice are a potentially powerful tool for fine-scale association mapping (Guénet and Bonhomme 2003; Laurie et al. 2007). Wild *M. m. domesticus* mice display incredible karyotypic variability, with over 100 “races” having fixed different combinations of Robertsonian translocations (a common type of chromosomal fusion), while such fixations are absent in laboratory mice (Nachman et al. 1994; Piálek et al. 2005). Robertsonian translocations are the most common structural aberration in humans (Nielsen and Wohlert 1991) and are implicated in segregation disorders (Chiang et al. 2012), male sterility (Daniel 2002), and an increased rate of trisomy in offspring (Down, Edward, and Patau syndromes). Understanding the mechanism driving this karyotypic variability might lead to improved prenatal testing, prevention, and treatments. Wild mice have also been used with success in functional and behavioral genomic studies (Guénet and Bonhomme 2003). Finally, wild mice can be used to determine the ancestral origin of variants that are present in laboratory mice. That information will facilitate the correct interpretation of results from biomedical studies.

It is also important to remember that the house mouse is a household and agricultural pest as well as a vector for disease. Mice damage crops, stored grain and fodder, farm infrastructure, and equipment. Mice can spoil food with feces and urine and can transmit diseases and parasites to humans and livestock (e.g., salmonella). A study of the impact of house mice on crops in Australia estimated annual losses as high as AU$60 M (Brown and Singleton 2002). Although rodenticides are currently used to combat mouse infestations, they are controversial due to potential collateral damage to native species. Better characterization of wild mice may enable the development of purely genetic eradication strategies, as have been effective for other pests (Wise de Valdez et al. 2011).

The technological developments of the past decade have advanced the field of mammalian genetics faster than ever before in its history. For the continued success of mouse genetics,

research must leverage the existence of two connected and equally important resources: laboratory strains and wild mice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Ping Fu for preparing Supplementary Fig. 6 and to Jeremy Wang and Leonard McMillan for contributing to Box 1. We also thank François Bonhomme and an anonymous reviewer for their comments on an earlier version of the manuscript. JPD is supported by grants from the National Institutes of Health (P50MH090338 and P50HG006582 (to FPMV)). National Institute of General Medical Sciences Centers of Excellence supported critical work reviewed here in the Systems Biology program (Grant GM-076468).

Glossary of terms

Ancestral polymorphism	A polymorphic locus known to be segregating in the most recent common ancestor of multiple lineages rather than having arisen following their divergence. Note that gene flow may lead to the appearance of ancestral polymorphism
Ascertainment bias	Systematic deviations from an expected theoretical result attributable to the sampling processes used to find (ascertain) SNPs and estimate their population-specific allele frequencies
Classical strain	An inbred laboratory strain derived from a small population of “fancy” mice beginning in the early 20th century
Commensal	A form of symbiosis in which one organism derives a benefit while the other is unaffected. House mice are traditionally called a commensal of humans however their status as economic pest and carriers of disease arguably classifies them as parasites
De novo assembly	Assembling a new genome without using an existing reference genome as a guide. De novo assembly is computationally difficult, but it has the benefit of assembling sequences that reference-guided alignment may miss due to their absence from the reference sequence
Diagnostic marker	A polymorphic marker for which one allele is present in only one population. In human studies, these are often referred to as ancestry-informative markers (AIMs). Alleles that are diagnostic for a single house mouse subspecies are used to assign subspecific origin to regions of the genome
Effective population size	The minimum size of a population that would be required to observe the same dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration

Fancy mice	Mice bred as pets. The breeding of fancy mice selects for traits that are attractive to enthusiasts rather than researchers, such as interesting coat colors and behaviors. Fancy mice comprised the founder population of classical inbred strains
Haplotype	A collection of co-occurring, contiguous alleles. May be used to refer to the alleles at a specific locus or across the entire genome
House mouse	Common name for *Mus musculus* species. Includes three distinct subspecies: *domesticus* (Western Europe), *musculus* (Eastern Europe and North Asia), and *castaneus* (Southeast Asia). Laboratory mice are of house mouse origin but are a mixture of multiple subspecific origins
Hybrid zone	A boundary between two distinct interbreeding populations. The best-known example is the ~2,500-km-long European transect where *M.m. domesticus* and *M. m. musculus* meet
Imputation	A statistical method of deriving the complete sequence of a large number of samples using genotype information to identify matching haplotypes in a small number of reference sequences
Inbred strain	A mouse strain created by successive generations of sibling–sibling or parent–offspring mating, which results in a completely homozygous genome. Until the availability of high-density arrays, the consensus was to declare a strain homozygous after 20 generations of inbreeding
Introgression	The transmission of a novel allele into a population by hybridization followed by backcrossing to one of the parental populations
Monophyly	When a taxon forms a single clade in a phylogeny, meaning that it contains all descendants of the most recent common ancestor of all members of the group
Mosaic genome	A genome derived from multiple distinct ancestries. Mosaicism is identified by haplotype blocks that contain diagnostic alleles. The house mouse is a mosaic of the three *M. musculus* subspecies, although most of its genome is *M. m. domesticus* in origin
Mouse	In both colloquial and taxonomic usage the name “mouse” is applied to many different species of small mammals. We use the term strictly to refer to an animal belonging to genus *Mus*
Nucleotide diversity	The degree of polymorphism within a population. Calculated as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population

Outbred stock	A heterogeneous strain typically maintained in a colony and allowed to random-mate. Outbred strains are primarily derived from intercrossing classical strains, including fancy mice of Swiss origin
Pan-genome	A consensus reference sequence created by aligning sequences from multiple individuals. Polymorphic sites may be annotated as heterozygous, or the most common allele may be assigned to that position in the consensus sequence
Pseudo-genome	A synthetic genome created by imputation. A new sample is genotyped and compared to multiple, fully assembled reference genomes. In each region, the most similar reference genome is identified, and those regions are concatenated
Reference genome	A whole-genome sequence that is agreed upon as the index for a species. The genome must be fully assembled and given positional annotations. This enables researchers to associate a physical position with genes and other genomic features. A reference genome may be created from a single individual or it may be a consensus of multiple individuals
Single nucleotide polymorphism (SNP)	A site in the genome that is polymorphic within a population
SNP discovery	Comparison of sequence from multiple individuals to identify polymorphic loci
Structural variation	A polymorphism that alters the structure rather than just the content, of the ancestral genome. Insertion and deletion (indels) of bases, ranging from single bases to entire genes, are the most common structural variation. Copy number variation is a special class of indel in which the number of copies of a short tandem repeat increases or decreases between generations
VINO	A type of genotypic marker that represents previously uncharacterized variation. Useful for counteracting ascertainment bias in phylogenetic studies (Didion et al. 2012)
Wild-derived strain	Generally, any laboratory strain that is not descended from the same common genetic pool as classical strains. Most wild-derived strains have been derived from wild-caught mice
Wild mouse	A mouse trapped in nature and not a product of any genetic manipulation or selective breeding

References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

Artzt K, Barlow D, Dove WF, Lindahl KF, Klein J, Lyon MF, Silver LM. Maps of mouse chromosome 17: first report. Mamm Genome. 1991; 1:5–29. [PubMed: 1794045]

Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, Collins FS, Dove WF, Duyk G, Dymecki S, Eppig JT, et al. The knockout mouse project. Nat Genet. 2004; 36:921–924. [PubMed: 15340423]

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, Ferris MT, Frelinger JA, Heise M, Frieman MB, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 2011; 21:1213–1222. [PubMed: 21406540]

Battey J, Jordan E, Cox D, Dove W. An action plan for mouse genomics. Nat Genet. 1999; 21:73–75. [PubMed: 9916794]

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM. Genealogies of mouse inbred strains. Nat Genet. 2000; 24:23–25. [PubMed: 10615122]

Bishop CE, Boursot P, Baron B, Bonhomme F, Hatat D. Most classical *Mus musculus* domesticus laboratory mouse strains carry a *Mus musculus musculus* Y chromosome. Nature. 1985; 315:70–72. [PubMed: 2986012]

Blanchet C, Jaubert J, Carniel E, Fayolle C, Milon G, Szatanik M, Panthier JJ, Montagutelli X. *Mus spretus* SEG/Pas mice resist virulent *Yersinia pestis*, under multigenic control. Genes Immun. 2011; 12:23–30. [PubMed: 20861861]

Bonhomme F, Selander RK. Estimating total genic diversity in the house mouse. Biochem Genet. 1978; 16:287–297. [PubMed: 678296]

Bonhomme F, Guénet JL, Dod B, Moriwaki K, Bulfield G. The polyphyletic origin of laboratory inbred mice and their rate of evolution. Bio J Linn Soc. 1987; 30:51–58.

Bonhomme, F.; Anand, R.; Darviche, D.; Din, W. The house mouse as a ring species. In: Moriwaki, K.; Shiroishi, T.; Yonekawa, H., editors. Genetics in Wild Mice: Its Application to Biomedical Research. Tokyo: Japanese Scientific Societies Press; 1994. p. 13-23.

Bottomly D, Ferris MT, Aicher LD, Rosenzweig E, Whitmore A, Aylor DL, Haagmans BL, Gralinski LE, Bradel-Tretheway BG, Bryan JT, et al. Expression quantitative trait loci for extreme host response to influenza a in pre-collaborative cross mice. G3. 2012; 2:213–221. [PubMed: 22384400]

Boursot P, Belkhir K. Mouse SNPs for evolutionary biology: beware of ascertainment biases. Genome Res. 2006; 16:1191–1192. [PubMed: 17018517]

Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. The evolution of house mice. Annu Rev Ecol Syst. 1993; 24:119–152.

Boursot P, Din W, Anand R, Darviche D, Dod B, Von Deimling F, Talwar GP, Bonhomme F. Origin and radiation of the house mouse: mitochondrial DNA phylogeny. J Evol Biol. 1996; 9:391–415.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, et al. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 2010; 8:e1000451. [PubMed: 20711490]

Brown, PR.; Singleton, GR. Impacts of house mice on crops in Australia: costs and damage. In: Clark, L.; Hone, J.; Shivik, JA.; Watkins, RA.; VerCauteren, KC.; Yoder, JK., editors. Human conflicts with wildlife: economic considerations. Fort Collins: National Wildlife Research Center; 2002. p. 48-58.

Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. Fine-scale maps of recombination rates and hotspots in the mouse genome. Genetics. 2012; 191:757–764. [PubMed: 22562932]

Burgess-Herbert SL, Tsaih SW, Stylianou IM, Walsh K, Cox AJ, Paigen B. An experimental assessment of in silico haplotype association mapping in laboratory mice. BMC Genet. 2009; 10:81. [PubMed: 20003225]

Chessler EJ, Miller DR, Branstetter LR, Galloway LD, Jackson BD, Philip VM, Voy BH, Culiat CT, Threadgill DW, Williams RW, et al. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. Mamm Genome. 2008; 19:382–389. [PubMed: 18716833]

Cheverud JM, Vaughn TT, Pletscher LS, Peripato AC, Adams ES, Erikson CF, King-Ellison KJ. Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. Mamm Genome. 2001; 12:3–12. [PubMed: 11178736]

Chiang T, Schultz RM, Lampson MA. Meiotic origins of maternal age-related aneuploidy. Biol Reprod. 2012; 86:1–7. [PubMed: 21957193]

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, et al. Modernizing reference genome assemblies. PLoS Biol. 2011; 9:e1001091. [PubMed: 21750661]

Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 2012; 190:389–401. [PubMed: 22345608]

Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J, et al. A new standard genetic map for the laboratory mouse. Genetics. 2009; 182:1335–1344. [PubMed: 19535546]

Daniel A. Distortion of female meiotic segregation and reduced male fertility in human Robertsonian translocations: consistent with the centromere model of co-evolving centromere DNA/centromeric histone (CENP-A). Am J Med Genet. 2002; 111:450–452. [PubMed: 12210311]

Danshina PV, Geyer CB, Dai Q, Goulding EH, Willis WD, Kitto GB, McCarrey JR, Eddy EM, O'Brien DA. Phosphoglycerate kinase 2 (PGK2) is essential for sperm function and male fertility in mice. Biol Reprod. 2010; 82:136–145. [PubMed: 19759366]

Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, Pardo-Manuel de Villena F, Churchill GA. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics. 2012; 13:34. [PubMed: 22260749]

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ, et al. A comprehensive genetic map of the mouse genome. Nature. 1996; 380:149–152. [PubMed: 8600386]

Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E, Orth A, Talwar GP, Cazenave PA, Bonhomme F. Origin and radiation of the house mouse: clues from nuclear genes. J Evol Biol. 1996; 9:519–539.

Dobzhansky T. Nothing in biology makes sense except in the light of evolution. Am Biol Teach. 1973; 35:125–129.

Dumont BL, Payseur BA. Genetic analysis of genome-scale recombination rate evolution in house mice. PLoS Genet. 2011; 7:e1002116. [PubMed: 21695226]

Duvaux L, Belkhir K, Boulesteix M, Boursot P. Isolation and gene flow: inferring the speciation history of European house mice. Mol Ecol. 2011; 20:5248–5264. [PubMed: 22066696]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]

Festing M. Inbred strains of mice: a vital resource for biomedical research. Mouse Genome. 1997; 95:845–855.

Flint J, Eskin E. Genome-wide association studies in mice. Nat Rev Genet. 2012; 13:807–817. [PubMed: 23044826]

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 2007; 448:1050–1053. [PubMed: 17660834]

Gabriel SI, Jóhannesdóttir F, Jones EP, Searle JB. Colonization, mouse-style. BMC Biol. 2010; 8:131. [PubMed: 20977781]

Gabriel SI, Stevens MI, Mathias MDL, Searle JB. Of mice and "convicts": origin of the Australian house mouse *Mus musculus*. PLoS One. 2011; 6:e28622. [PubMed: 22174847]

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu AH-T, Bulatova N, Ziv Y, Nachman MW. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol. 2008; 17:5349–5363. [PubMed: 19121002]

Gregorová S, Forejt J. PWD/Ph and PWK/Ph inbred mouse strains of *Mus m. musculus* subspecies: a valuable resource of phenotypic variations and genomic polymorphisms. Folia Biol (Praha). 2000; 46:31–41. [PubMed: 10730880]

Gregorová S, Divina P, Storchova R, Trachtulec Z, Fotopulosova V, Svenson KL, Donahue LR, Paigen B, Forejt J. Mouse consomic strains: exploiting genetic divergence between *Mus m.*

musculus and *Mus m. domesticus* subspecies. Genome Res. 2008; 18:509–515. [PubMed: 18256238]

Guan C, Ye C, Yang X, Gao J. A review of current large-scale mouse knockout efforts. Genesis. 2010; 48:73–85. [PubMed: 20095055]

Guénet J-L, Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet. 2003; 19:24–31. [PubMed: 12493245]

Hardouin EA, Chapuis JL, Stevens MI, van Vuuren JB, Quillfeldt P, Scavetta RJ, Teschke M, Tautz D. House mouse colonization patterns on the sub-Antarctic Kerguelen Archipelago suggest singular primary invasions and resilience against reinvasion. BMC Evol Biol. 2010; 10:325. [PubMed: 20977744]

Harr B. Genomic islands of differentiation between house mouse subspecies. Genome Res. 2006; 16:730–737. [PubMed: 16687734]

Hrbek T, de Brito RA, Wang B, Pletscher LS, Cheverud JM. Genetic characterization of a new set of recombinant inbred lines (LGXSM) formed from the intercross of SM/J and LG/J inbred mouse strains. Mamm Genome. 2006; 17:417–429. [PubMed: 16688532]

Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985; 111:147–164. [PubMed: 4029609]

Ideraabdullah FY, la Casa-Esper00F3;n de E, Bell TA, Detwiler DA, Magnuson TR, Sapienza C, Pardo-Manuel de Villena F. Genetic and haplotype diversity among wild-derived mouse inbred strains. Genome Res. 2004; 14:1880–1887. [PubMed: 15466288]

Jones EP, Jensen JK, Magnussen E, Gregersen N, Hansen HS, Searle JB. A molecular characterization of the charismatic Faroe house mouse. Bio J Linn Soc. 2011; 102:471–482.

Jones EP, Skirnisson K, McGovern TH, Gilbert M, Willerslev E, Searle JB. Fellow travellers: a concordance of colonization patterns between mice and men in the North Atlantic region. BMC Evol Biol. 2012; 12:35. [PubMed: 22429664]

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477:289–294. [PubMed: 21921910]

Kelada SNP, Aylor DL, Peck BCE, Ryan JF, Tavarez U, Buus RJ, Miller DR, Chesler EJ, Threadgill DW, Churchill GA, et al. Genetic analysis of hematological parameters in incipient lines of the Collaborative Cross. G3. 2012; 2:157–165. [PubMed: 22384394]

Kirby A, Kang HM, Wade CM, Cotsapas C, Kostem E, Han B, Furlotte N, Kang EY, Rivas M, Bogue MA, et al. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics. 2010; 185:1081–1095. [PubMed: 20439770]

Kohler, RE. Lords of the Fly. Chicago: University of Chicago Press; 1994.

Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, Nachman MW. Linkage disequilibrium in wild mice. PLoS Genet. 2007; 3:e144. [PubMed: 17722986]

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. Building the sequence map of the human pangenome. Nat Biotechnol. 2009a; 28:57–63. [PubMed: 19997067]

Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009b; 10:387–406. [PubMed: 19715440]

Macholán M, Munclinger P, Sugerková M, Dufková P, Bímová B, Bozíková E, Zima J, Piálek J. Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution. 2007; 61:746–771. [PubMed: 17439609]

Mathes WF, Aylor DL, Miller DR, Churchill GA, Chesler EJ, Pardo-Manuel de Villena F, Threadgill DW, Pomp D. Architecture of energy balance traits in emerging lines of the Collaborative Cross. Am J Physiol Endocrinol Metab. 2011; 300:E1124–E1134. [PubMed: 21427413]

Meerburg BG, Singleton GR, Kijlstra A. Rodent-borne diseases and their risks for public health. Crit Rev Microbiol. 2009; 35:221–270. [PubMed: 19548807]

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A mouse speciation gene encodes a meiotic histone H3 methyl-transferase. Science. 2009; 323:373–375. [PubMed: 19074312]

Moriwaki, K.; Shiroishi, T.; Yonekawa, H.; Miyashita, N.; Sagai, Y. Genetic status of Japanese wild mice and immunological characters of their h-2 antigens. In: Muramatsu, T.; Gachelin, G.; Monscona, AA.; Ikawa, Y., editors. Teratocarcinoma and embryonic cell interactions. Tokyo: Japan Scientific Society Press; 1982. p. 41-56.

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci USA. 2000; 97:12649–12654. [PubMed: 11050180]

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GLG, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science. 2002; 296:1661–1671. [PubMed: 12040188]

Nachman MW, Boyer SN, Searle JB, Aquadro CF. Mitochondrial DNA variation and the evolution of Robertsonian chromosomal races of house mice, *Mus domesticus*. Genetics. 1994; 136:1105–1120. [PubMed: 8005418]

Nagamine CM, Nishioka Y, Moriwaki K, Boursot P, Bonhomme F, Lau YF. The *musculus*-type Y chromosome of the laboratory mouse is of Asian origin. Mamm Genome. 1992; 3:84–91. [PubMed: 1352158]

Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA. 1979; 76:5269–5273. [PubMed: 291943]

Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 2012; 13:R45. [PubMed: 22703977]

Nielsen J, Wohlert M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. Hum Genet. 1991; 87:81–83. [PubMed: 2037286]

Orth A, Belkhir K, Britton-Davidian J, Boursot P, Benazzou T, Bonhomme F. Natural hybridization between two sympatric species of mice *Mus musculus domesticus* L *Mus spretus Lataste*. C R Biol. 2002; 325:89–97. [PubMed: 11980180]

Paigen K. One hundred years of mouse genetics: an intellectual history I The classical period (1902–1980). Genetics. 2003a; 163:1–7. [PubMed: 12586691]

Paigen K. One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). Genetics. 2003b; 163:1227–1235. [PubMed: 12702670]

Paigen K, Eppig JT. A mouse phenome project. Mamm Genome. 2000; 11:715–717. [PubMed: 10967127]

Pamilo P, Nei M. Relationships between gene trees and species trees. Mol Biol Evol. 1988; 5:568–583. [PubMed: 3193878]

Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet. 2004; 5:7. [PubMed: 15117419]

Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, Sargent EE, Asquith S, Crew V, Johnson KA, Robinson P, et al. An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res. 2004; 14:1806–1811. [PubMed: 15342563]

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW. Adaptive evolution and effective population size in wild house mice. Mol Biol Evol. 2012; 29(10): 2949–2955. [PubMed: 22490822]

Piálek J, Hauffe HC, Searle JB. Chromosomal variation in the house mouse. Bio J Linn Soc. 2005; 84:535–563.

Piálek J, Vyskocilová M, Bímová B, Havelková D, Piálková J, Dufková P, Bencová V, Dureje L, Albrecht T, Hauffe HC, et al. Development of unique house mouse resources suitable for evolutionary studies of speciation. J Hered. 2008; 99:34–44. [PubMed: 17965200]

Prager EM, Orrego C, Sage RD. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. Genetics. 1998; 150:835–861. [PubMed: 9755213]

Rajabi-Maham H, Orth A, Bonhomme F. Phylogeography and postglacial expansion of *Mus musculus domesticus* inferred from mitochondrial DNA coalescent, from Iran to Europe. Mol Ecol. 2008; 17:627–641. [PubMed: 18179435]

Rajabi-Maham H, Orth A, Siahsarvie R, Boursot P, Darvish J, Bonhomme F. The south-eastern house mouse *Mus musculus castaneus* (Rodentia: Muridae) is a polytypic subspecies. Bio J Linn Soc. 2012; 107:295–306.

Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004; 428:493–521. [PubMed: 15057822]

Sage RD, Atchley WR, Capanna E. House mice as models in systematic biology. Syst Biol. 1993; 42:523–561.

Salcedo T, Geraldes A, Nachman MW. Nucleotide variation in wild and inbred mice. Genetics. 2007; 177:2277–2291. [PubMed: 18073432]

Schwarz E, Schwarz HK. The wild and commensal stocks of the house mouse *Mus musculus* Linnaeus. J Mammal. 1943; 24:59.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. A map of the cisregulatory sequences in the mouse genome. Nature. 2012; 488(7409):116–120. [PubMed: 22763441]

Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, O'Brien WE, Conti DV, Witte JS, Lander ES, et al. Genetic dissection of complex traits with chromosome substitution strains of mice. Science. 2004; 304:445–448. [PubMed: 15031436]

Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005; 21:456–463. [PubMed: 15608047]

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). PLoS Genet. 2012; 8:e1002891. [PubMed: 22956910]

Stenseth NC, Leirs H, Skonhoft A, Davis SA, Pech RP, Andreassen HP, Singleton GR, Lima M, Machang'u RS, Makundi RH, et al. Mice, rats, and people: the bio-economics of agricultural rodent pests. Front Ecol Environ. 2003; 1:367–375.

Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. Mol Phylogenet Evol. 2004; 33:626–646. [PubMed: 15522792]

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA. High-resolution genetic mapping using the mouse diversity outbred population. Genetics. 2012; 190:437–447. [PubMed: 22345611]

Szatkiewicz JP, Beane GL, Ding Y, Hutchins L, Pardo-Manuel de Villena F, Churchill GA. An imputed genotype resource for the laboratory mouse. Mamm Genome. 2008; 19:199–208. [PubMed: 18301946]

Takahashi A, Nishi A, Ishii A, Shiroishi T, Koide T. Systematic analysis of emotionality in consomic mouse strains established from C57BL/6J and wild-derived MSM/Ms. Genes Brain Behav. 2008; 7:849–858. [PubMed: 18616609]

The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]

Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. Evolution. 1992; 46:1146.

Tucker PK, Sandstedt SA, Lundrigan BL. Phylogenetic relationships in the subgenus *Mus* (genus *Mus* family Muridae, subfamily Murinae): examining gene trees and species trees. Bio J Linn Soc. 2005; 84:653–662.

Wade CM, Kulbokas EJ, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ. The mosaic structure of variation in the laboratory mouse genome. Nature. 2002; 420:574–578. [PubMed: 12466852]

Wang, J.; Moore, KJ.; Zhang, Q.; Pardo-Manuel de Villena, F.; Wang, W.; McMillan, L. Genome-wide compatible SNP intervals and their properties. Proceedings of the first ACM international conference on bioinformatics and computational biology, Niagara Falls; 2–4 August 2010; New York. New York: Association for Computing Machinery; 2010. p. 43-52.

Wang JR, Pardo-Manuel de Villena F, Lawson HA, Cheverud JM, Churchill GA, McMillan L. Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny. Genetics. 2012a; 190:449–458. [PubMed: 22345612]

Wang JR, Pardo-Manuel de Villena F, McMillan L. Comparative analysis and visualization of multiple collinear genomes. BMC Bioinformatics. 2012b; 13(Suppl 3):S13. [PubMed: 22536897]

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. [PubMed: 12466850]

Welsh CE, Miller DR, Manly KF, Wang J, McMillan L, Morahan G, Mott R, Iraqi F, Threadgill DW, Pardo-Manuel de Villena F. Status and access to the Collaborative Cross population. Mamm Genome. 2012; 23:706–712. [PubMed: 22847377]

White MA, Ané C, Dewey CN, Larget BR, Payseur BA. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 2009; 5:e1000729. [PubMed: 19936022]

White MA, Steffy B, Wiltshire T, Payseur BA. Genetic dissection of a key reproductive barrier between nascent species of house mice. Genetics. 2011; 189:289–304. [PubMed: 21750261]

Wilson, DE.; Reeder, DM. Mammal species of the world. Baltimore: Johns Hopkins University Press; 2005.

Wise de Valdez MR, Nimmo D, Betz J, Gong HF, James AA, Alphey L. Black WC4, Genetic elimination of dengue vector mosquitoes. Proc Natl Acad Sci U S A. 2011; 108:4772–4775. [PubMed: 21383140]

Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DJ, Keane TM. Sequencing and characterization of the FVB/NJ mouse genome. Genome Biol. 2012; 13:R72. [PubMed: 22916792]

Yalcin B, Fullerton J, Miller S, Keays DA, Brady S, Bhomra A, Jefferson A, Volpi E, Copley RR, Flint J, et al. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc Natl Acad Sci USA. 2004; 101:9734–9739. [PubMed: 15210992]

Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Østerås M, Whitley A, Yuan W, Gan X, et al. Commercially available outbred mice for genome-wide association studies. PLoS Genet. 2010; 6:e1001085. [PubMed: 20838427]

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, et al. Sequence-based characterization of structural variation in the mouse genome. Nature. 2011; 477:326–329. [PubMed: 21921916]

Yalcin B, Adams DJ, Flint J, Keane TM. Next-generation sequencing of experimental mouse strains. Mamm Genome. 2012a; 23(9–10):490–498. [PubMed: 22772437]

Yalcin B, Wong K, Bhomra A, Goodson M, Keane TM, Adams DJ, Flint J. The fine-scale architecture of structural variants in 17 mouse genomes. Genome Biol. 2012b; 13:R18. [PubMed: 22439878]

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nat Genet. 2007; 39:1100–1107. [PubMed: 17660819]

Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, Pardo-Manuel de Villena F, Churchill GA. A customized and versatile high-density genotyping array for the mouse. Nat Methods. 2009; 6:663–666. [PubMed: 19668205]

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH-T, Nachman MW, Piálek J, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 2011; 43:648–655. [PubMed: 21623374]

Yonekawa, H.; Takahama, S. Genetic diversity and geographic distribution of *Mus musculus* subspecies based on the polymorphism of mitochondrial DNA. In: Moriwaki, K.; Shiroishi, T.; Yonekawa, H., editors. Genetics in Wild Mice. Tokyo: Japanese Scientific Societies Press; 1994. p. 25-40.

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima Y, Shi LM, Cho WS, Zhen XL, Tagashira Y. Hybrid origin of Japanese mice "*Mus musculus molossinus*": evidence from restriction analysis of mitochondrial DNA. Mol Biol Evol. 1988; 5:63–78. [PubMed: 2833677]

Box 1 Using imputation to improve subspecific origin assignment

Recently, Wang and colleagues (2012) imputed the ~12M SNPs discovered in whole-genome sequences of 12 classical strains (Keane et al. 2011) in an additional 88 strains with low error rates. Using those imputed genotypes, Wang and colleagues were able to assign the ancestral origin of 15 classical strains in “gap” regions (regions where Yang et al. (2011) were not able to assign origin due to low marker density on the MDA). Those gap regions were always located between regions of different subspecific origin. For each gap in each target strain, they first identified the guide strain (the most similar Sanger strain) on either side of the gap. They then extended the proximal boundary of the gap for as long as the target strain was more than twice as similar to the proximal guide strain as it was to the distal guide strain. Finally, they repeated the process for the distal boundary. Overall, they reduced the size of the gap regions by 68.5 % (from 200 to 63 Mb). Of the refined subspecies regions, 55 % were determined to be *M. m. domesticus*, 39 % *M. m. musculus*, and 6 % *M. m. castaneus* (J. R. Wang and L. McMillan, personal communication). The updated assignments are publicly available in the Mouse Phylogeny Browser software, available at http://msub.csbio.unc.edu.

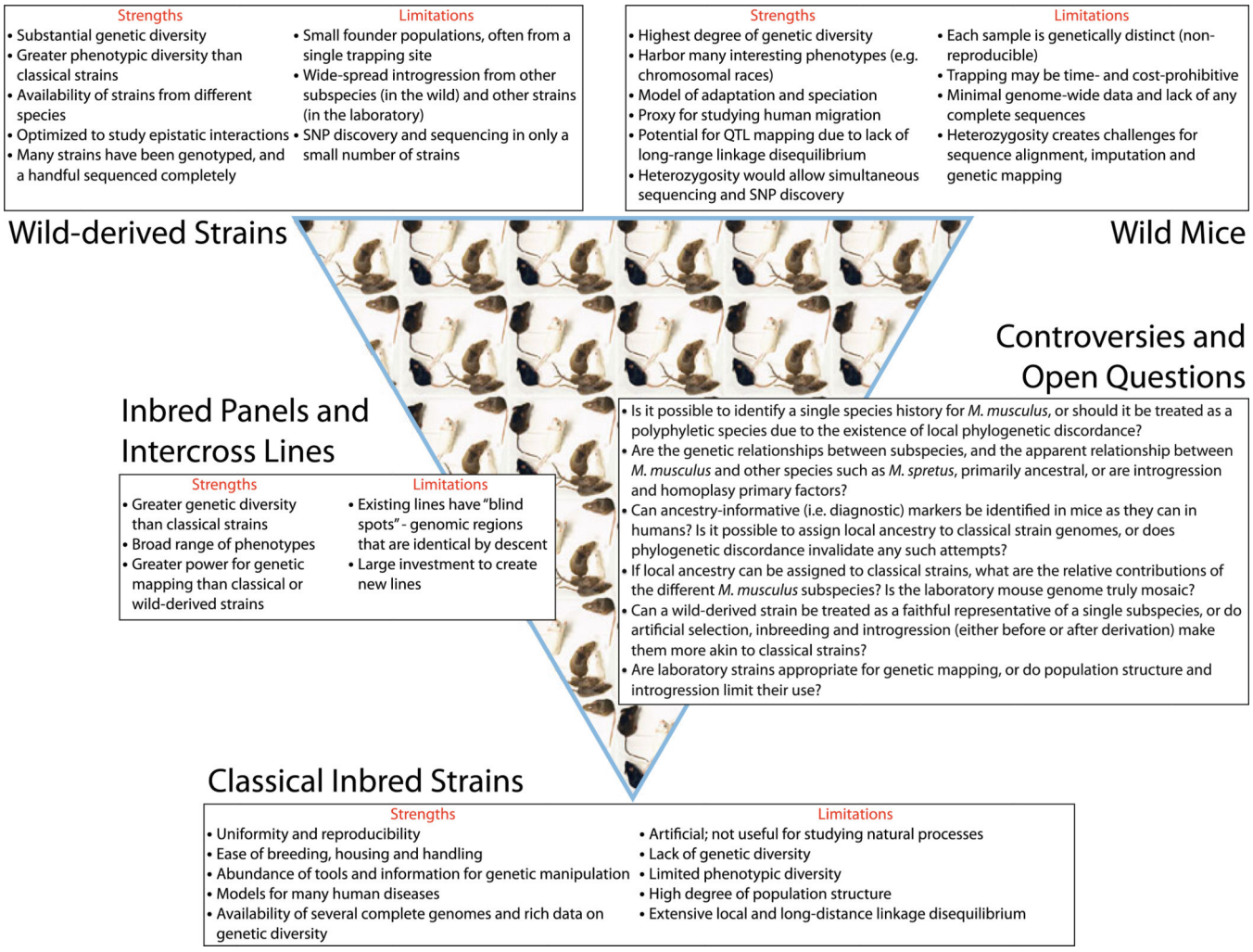

Fig. 1.

The mouse in research: types, uses, and controversies

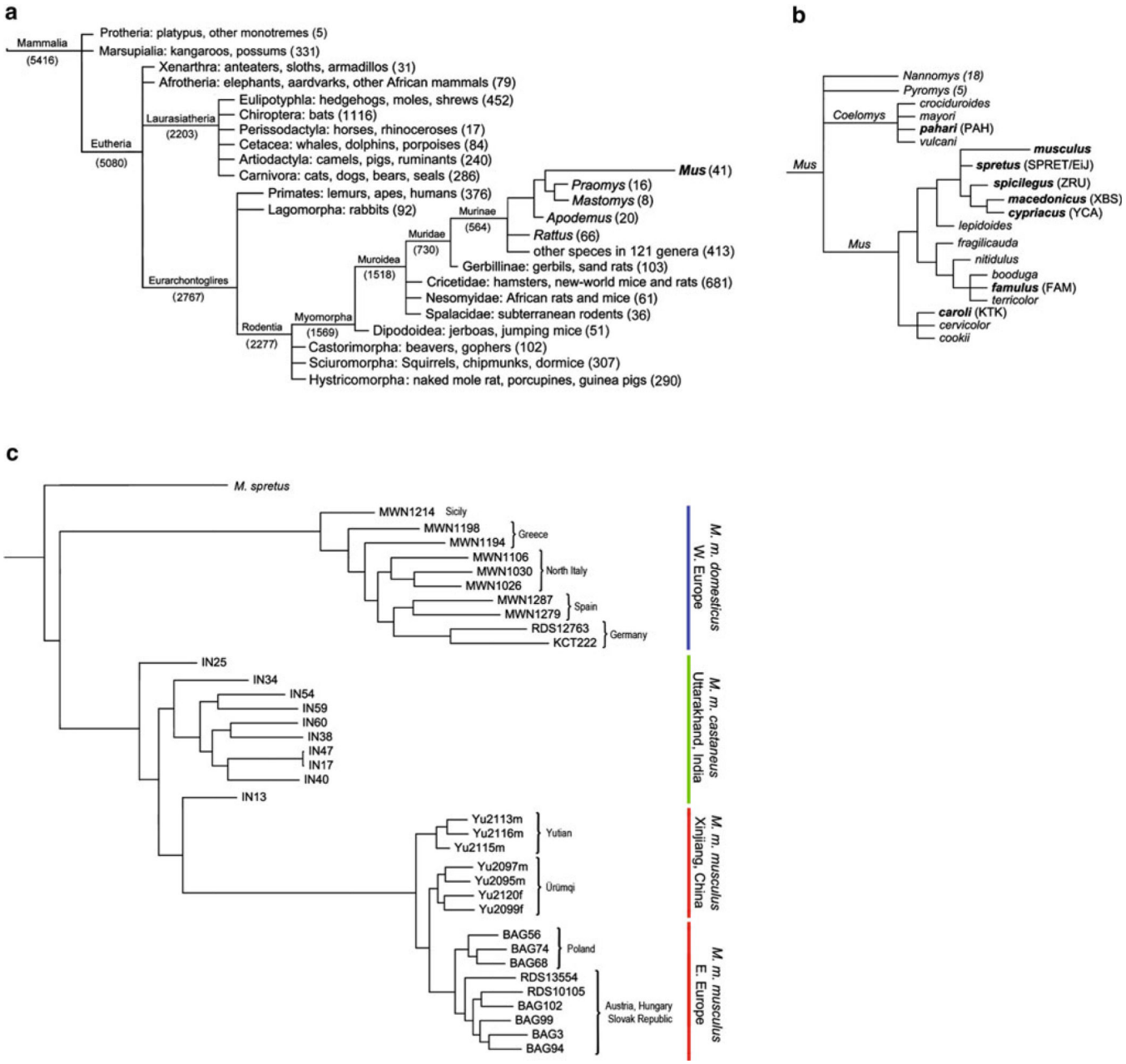

Fig. 2.

The taxonomic position and phylogeny of *M. musculus*. **a** Taxonomy of class Mammalia. The name of each taxon is followed by a list of example species and the total number of species in parentheses. All relationships and species counts are from Wilson and Reeder (2005). Genus *Mus* is shown in bold. **b** Phylogeny of genus *Mus*. Species in *bold* indicate that laboratory strains have been derived from wild-caught animals, and an example of such an inbred strain is given in *parentheses*. Approximate species relationships are shown for subgenus *Mus* based on a meta-analysis of several phylogenetic studies. **c** The single best maximum-likelihood tree for the phylogeny of *M. musculus*. We used RAxML (Stamatakis et al. 2005) to analyze genotypes for 547,406 SNP markers and 118,733 VINO markers

from 36 wild-caught *M. musculus* samples (10 *M. m. domesticus*, 16 *M. m. musculus*. and 10 *M. m. castaneus*) (Yang et al. 2011) and a single sample of the wild-derived *M. spretus* strain SPRET/EiJ (JPD and FPMV). Colored lines denote subspecific clades. *Blue*: *M. m. domesticus*; *green*: *M. m. castaneus*; *red*: *M. m. musculus*. Geographic origin of samples is given for *M. m. domesticus* and *M. m. musculus*; all *M. m. castaneus* samples are from the state of Uttarakhand, India. There is high support for monophyly in all clades and also a clear division between eastern European and Asian *M. m. musculus*

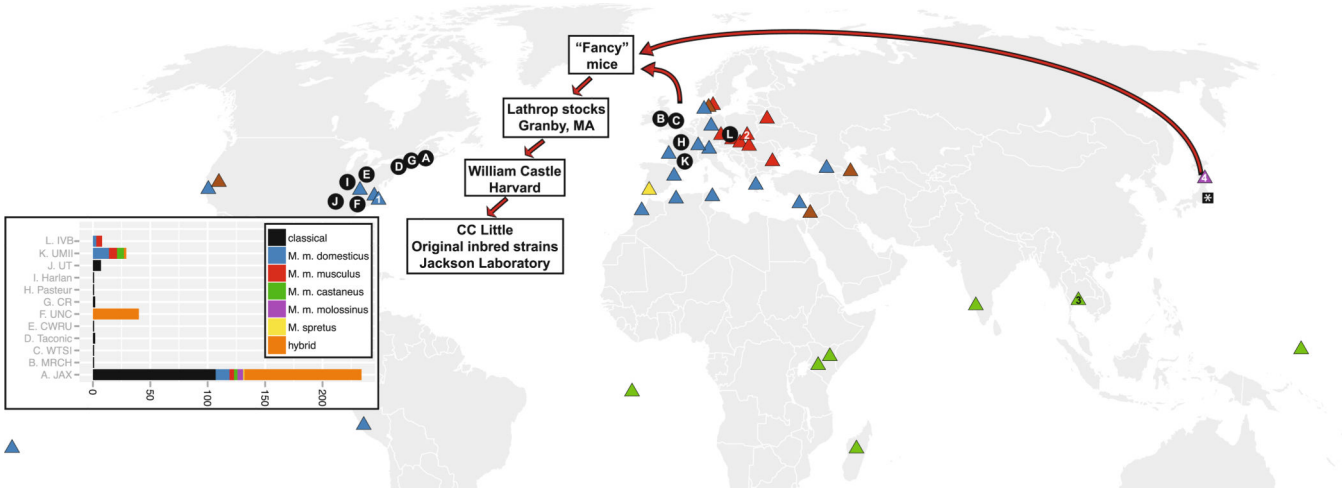

Fig. 3.

The origin of laboratory strains that have been sequenced or genotyped at high-density. *Circles*: providers of laboratory strains. The *inset bar chart* shows the number and types of laboratory strains from each provider. *A*, The Jackson Laboratory, Bar Harbor, ME, USA; *B*, MRC Harwell, Oxfordshire, UK; *C*, Wellcome Trust Sanger Institute, London, UK; *D*, Taconic Farms, Hudson, NY, USA; *E*, Case Western Reserve University, Cleveland, OH, USA; *F*, University of North Carolina, Chapel Hill, NC, USA; *G*, Charles River, Wilmington, MA, USA; *H*, Pasteur Institute, Paris, France; *I*, Harlan Laboratories, Indianapolis, IN, USA; *J*, University of Tennessee, Knoxville, TN, USA; *K*, Université Montpellier 2, Montpellier, France; *L*, Institute of Vertebrate Biology, Studenec, Czech Republic. *Triangles*: trapping sites of wild-derived strains. Numbered triangles represent the origins of widely used wild-derived strains. *1*, WSB/EiJ, Eastern Shore, MD, USA; *2*, PWK/PhJ and PWD/PhJ, Lhotka, Czech Republic; *3*, CAST/EiJ, Bangkok, Thailand; *4*, MOLF/EiJ, Fukuoka, Japan. The *asterisk* marks the Riken Institute, which originated several of the genotyped wild-derived strains before transferring them to other institutes and still provides a large repository of additional classical and wild-derived strains

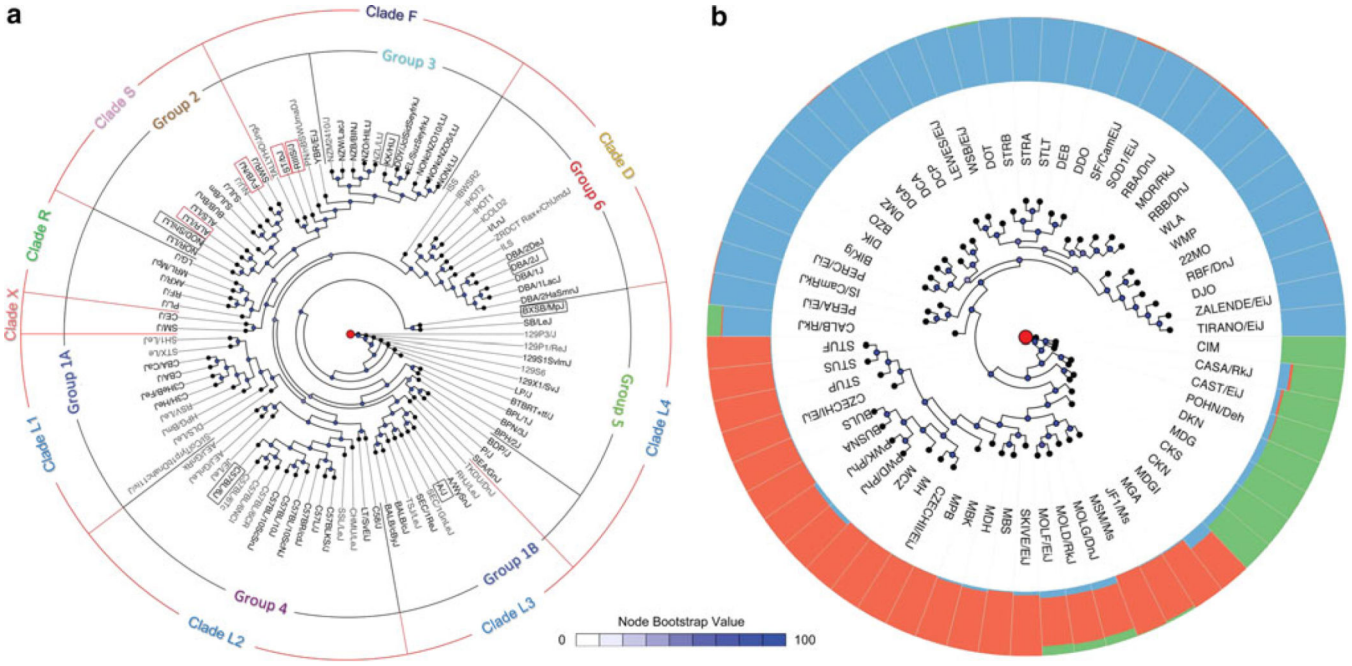

Fig. 4.

The phylogeny of laboratory strains. Neighbor-joining phylogeny of **a** 97 classical laboratory strains and **b** 62 wild-derived strains based on SNP and VINO genotypes (Yang et al. 2011). *Node colors* represent bootstrap values. **a** *Black inner circle* groupings and labels are based on Fig. 3 of Petkov et al. (2004). *Red outer circle* groupings are based on a partitioning of the current tree that creates monophyletic clades as similar as possible to the earlier tree, and labels are assigned based on genealogy (Beck et al. 2000) as follows: *Clade D*: DBA-related strains, including lines derived from the UC Berkley eight-way cross that have a disproportionate contribution from DBA/2; *Clade F*: lines derived from European, New Zealand, Japanese, and other non-Lathrop fancy mouse stocks; *Clades L1–L4*: lines descended primarily from the Lathrop stocks and lines created by C.C. Little and William Castle; *Clade R*: lines created at the Rockefeller Institute, and also LG/J (which is the primary background of MRL/MpJ but is not closely related to any other strain); *Clade S*: Swiss mice; *Clade X*: CE/J and SM/J are the results of multiway crosses and thus not substantially related to any single strain. CE/J also retains a large contribution from a wild-caught mouse. **b** The outer ring of colors shows the fraction of the genome of each strain that is derived from *M. m. domesticus* (*blue*), *M. m. musculus* (*red*), and *M. m. castaneus* (*green*)

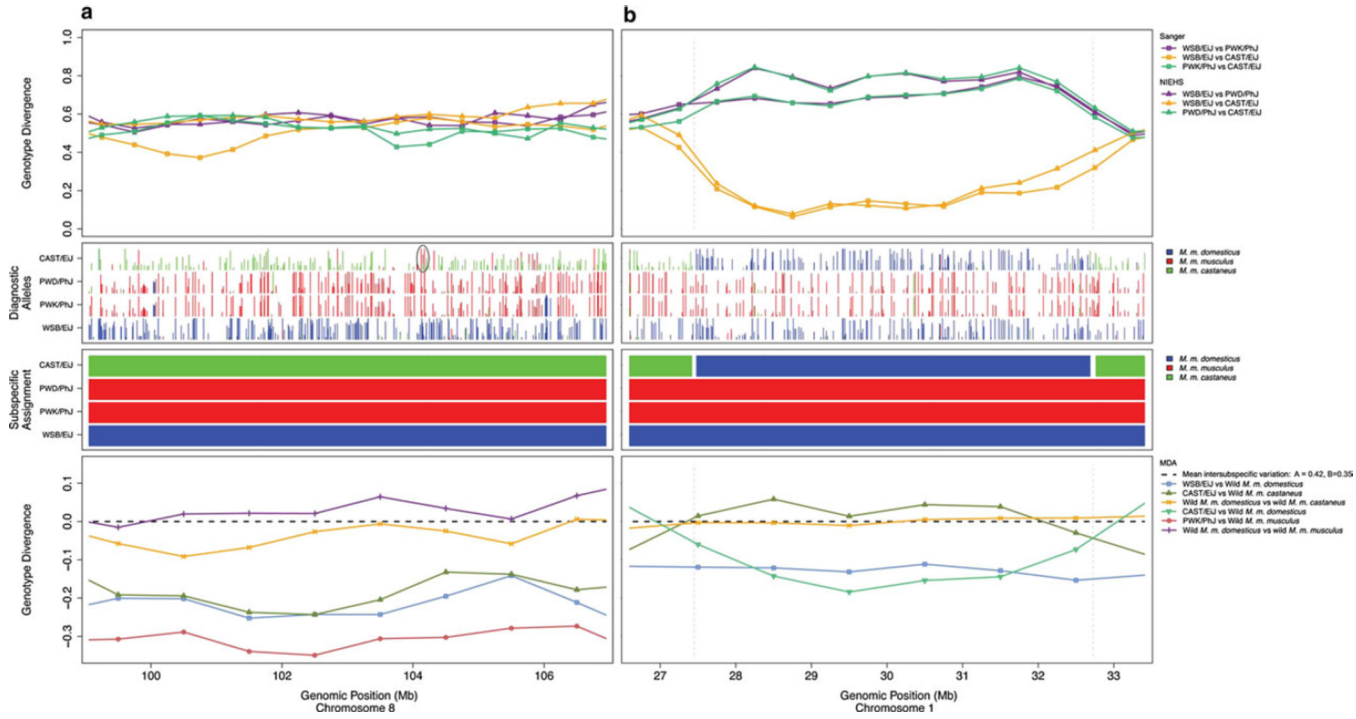

Fig. 5.

NIEHS and MDA genotypes and Sanger sequence data confirm the presence of introgression in wild-derived strains. Agreement between genotype and sequence data and subspecific origin assignment in **a** a region in which all wild-derived strains have the expected subspecific origin (Chromosome 8, 98–108 Mb), and **b** a region where CAST/EiJ has introgression from *M. m. domesticus* (Chromosome 1, 26–34 Mb). Top panels: for each pair of wild-derived strains, genotype divergence is measured as the fraction of informative SNPs for which the strains have different genotype calls. SNPs discovered in the NIEHS strains are consistent with those discovered in the Sanger strains. The low divergence between WSB/EiJ and CAST/EiJ in (**b**) suggests that they are actually of the same subspecific origin in that interval. Second panels: the location, subspecies, and diagnostic value of diagnostic alleles (Yang et al. 2011) in each strain. *Blue lines* indicate SNPs that are diagnostic for *M. m. domesticus*, *red lines* for *M. m. musculus*, and *green lines* for *M. m. castaneus*. The height of the line (values are between 0 and 1) indicates the frequency of the diagnostic allele in that subspecies. The *circled region* in **a** is referred to in the main text. The abundance of *M. m. domesticus* diagnostic SNPs in CAST/EiJ in (**b**) further supports a *M. m. domesticus* origin of the region. Third panels: the subspecific origin assigned by a Hidden Markov Model (HMM) (Yang et al. 2011). Bottom panels: comparisons between wild-derived strains and wild mice using MDA genotypes (Yang et al. 2011). Genotype divergence is given as the increase or decrease compared to the mean variation between different subspecies of wild mice. In **b**, the low divergence between WSB/EiJ and wild *M. m. domesticus* indicates that WSB/EiJ is of *M. m. domesticus* origin in that interval. The high divergence between CAST/EiJ and wild *M. m. castaneus* and the low divergence between CAST/EiJ and wild *M. m. domesticus* suggest that CAST/EiJ is also of *M. m. domesticus* in that interval

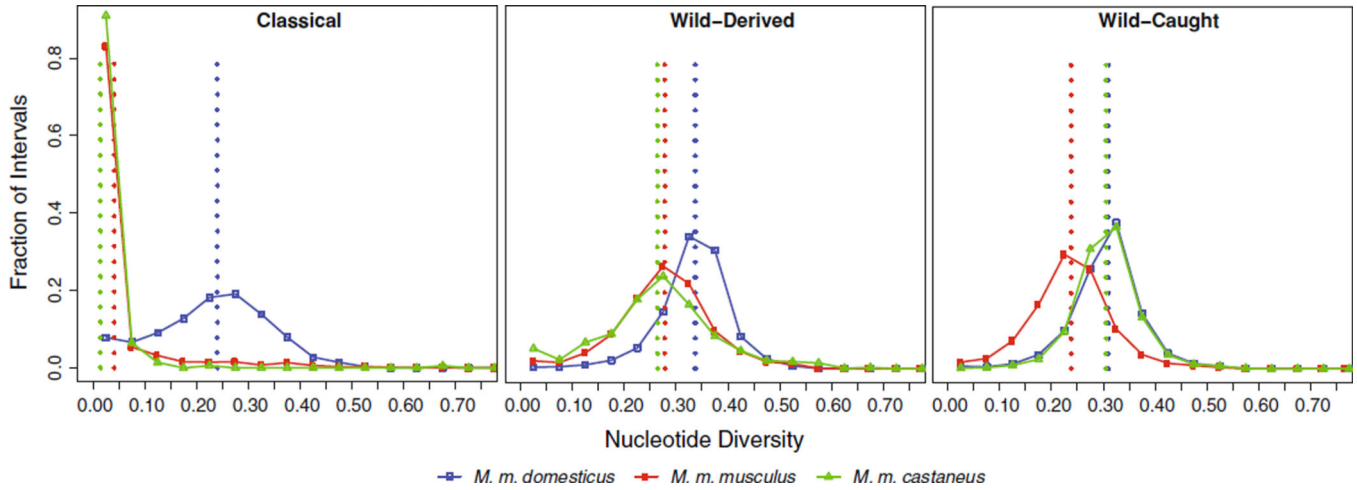

Fig. 6.

Nucleotide diversity is greater in wild mice than classical strains. We divided the genome into 16,331 intervals with no historical evidence of recombination in classical strains (Yang et al. 2011) and measured nucleotide diversity (π) at diagnostic SNPs in each interval for classical strains, wild-derived strains, and wild-caught mice. The *x axis* shows π for each subspecies in bins of 0.01. The *y axis* shows the fraction of intervals with the given subspecific origin that is in each bin. *Vertical dotted lines* show mean values of π for each subspecies. *Color* indicates subspecific origin. *Blue*: *M. m. domesticus*; *red*: *M. m. musculus*; *green*: *M. m. castaneus*

Table 1

Wild-derived strains contain extensive variation compared to classical strains

		Wild-derived		
Classical	**Project**	**Fixed (Ref)**	**Fixed (Var)**	**Segregating**
Fixed (Ref)	NIEHS	N/A	0.4	4.4
	Sanger	N/A	0.5	20.1
Segregating	NIEHS	0.9	0.5	2.0
	Sanger	3.4	1.4	7.1

Millions of SNPs discovered in the 11 classical strains and 4 wild-derived strains of the NIEHS project, and the 13 classical strains and 3 wild-derived *M. musculus* strains of the Sanger project. An additional 24.1 M SNPs are fixed for the reference in *M. musculus* Sanger strains but have an alternate allele in SPRET/EiJ

Ref reference allele, *Var* variant allele