nature genetics

Subspecific origin and haplotype diversity in the laboratory mouse

Hyuna Yang[1], Jeremy R Wang[2], John P Didion[3–5], Ryan J Buus[3–5], Timothy A Bell[3–5], Catherine E Welsh[2], François Bonhomme[6], Alex Hon-Tsen Yu[7,8], Michael W Nachman[9], Jaroslav Pialek[10], Priscilla Tucker[11], Pierre Boursot[6], Leonard McMillan[2], Gary A Churchill[1] & Fernando Pardo-Manuel de Villena[3–5]

Here we provide a genome-wide, high-resolution map of the phylogenetic origin of the genome of most extant laboratory mouse inbred strains. Our analysis is based on the genotypes of wild-caught mice from three subspecies of *Mus musculus*. We show that classical laboratory strains are derived from a few fancy mice with limited haplotype diversity. Their genomes are overwhelmingly *Mus musculus domesticus* in origin, and the remainder is mostly of Japanese origin. We generated genome-wide haplotype maps based on identity by descent from fancy mice and show that classical inbred strains have limited and non-randomly distributed genetic diversity. In contrast, wild-derived laboratory strains represent a broad sampling of diversity within *M. musculus*. Intersubspecific introgression is pervasive in these strains, and contamination by laboratory stocks has played a role in this process. The subspecific origin, haplotype diversity and identity by descent maps can be visualized using the Mouse Phylogeny Viewer (see URLs).

Most mouse laboratory strains are derived from *M. musculus*, a species with multiple lineages that includes three major subspecies, *M. m. domesticus*, *Mus musculus musculus* and *Mus musculus castaneus*, with distinct geographical ranges[1]. In historical times, mice followed human migratory patterns and colonized new regions. In regions of secondary contact between subspecies, there is evidence of gene flow[1–3]. Hybridization between *M. m. musculus* and *M. m. castaneus* in Japan resulted in the *Mus musculus molossinus* subspecies[4].

Laboratory strains can be classified into two groups based on their origin. Classical inbred strains were derived during the twentieth century from fancy mice. These strains have been the preferred tools in biomedical research. Historical sources and genetic studies suggest that fancy mice had substantial inbreeding[5]. These sources indicate that three subspecies of *M. musculus* were represented in the genome of fancy mice, making classical strains artificial hybrids between multiple subspecies found in the wild. However, there is disagreement about the relative contribution of each subspecies to classical inbred strains[6,7]. Classical strains have substantial population structure because of the limited genetic diversity present in fancy mice and the complex schema used in their derivation.

Wild-derived laboratory strains are derived directly from wild-caught mice[8]. Each strain has been assigned to a subspecies or is a natural hybrid between subspecies. The population structure of wild-derived strains can be accounted for by their taxonomical classification.

The initial report of the genome sequence and annotation of the C57BL/6J classical inbred strain[9] was followed by an extensive SNP discovery effort in 15 laboratory strains[6] and the ongoing whole genome sequencing of 17 inbred strains[10]. These data will inform hundreds of projects that use the mouse as a model for biomedical research, including the International Knockout Mouse projects and the Collaborative Cross[11,12].

Despite this wealth of sequence data, our understanding of genetic diversity in mice is shallow and biased. SNP discovery has involved only a limited number of strains, resulting in SNP panels with substantial ascertainment bias[13]. Pedigree records continue to serve as the primary source of information about the origin and relationships among laboratory strains[5]. Although such records are valuable, genetic studies and the experience of mouse breeders indicate that contamination is common[7]. We have previously reported the presence of intersubspecific introgression in three commonly used wild-derived strains[7]. However, this conclusion has been controversial, and the lack of data from wild-caught mice has prevented consensus among the scientific community. Finally, the *M. musculus* subspecies are undergoing the early stages of speciation. There is shared variation among subspecies,

[1]The Jackson Laboratory, Bar Harbor, Maine, USA. [2]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [5]Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [6]Université Montpellier 2, CNRS UMR5554, Institut des Sciences de l'Evolution, Montpellier, France. [7]Institute of Zoology, National Taiwan University, Taipei, Taiwan. [8]Department of Life Science, National Taiwan University, Taipei, Taiwan. [9]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA. [10]Department of Population Biology, Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno and Studenec, Czech Republic. [11]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA. Correspondence should be addressed to F.P.-M.d.V. (fernando@med.unc.edu) or G.A.C. (garyc@jax.org).

Received 13 July 2010; accepted 5 May 2011; published online 29 May 2011; doi:10.1038/ng.847

mostly because of polymorphisms that have persisted from a common ancestor and introgression between subspecies in the wild. Thus, selection of a single reference genome for each subspecies cannot accurately reflect the population structure of these recently diverged taxa. Furthermore, the choice of a single inbred strain to represent an entire taxon is particularly problematic because laboratory strains were subject to many generations of selective mating in an artificial setting, where there is high potential for contamination[7].

Given the contradictory conclusions reached regarding the origin of the genome of classical and wild-derived laboratory mouse strains[6,7,14–16], it is crucial to select representative reference samples along with a platform that can address the limitations of previous studies. We collected a geographically diverse sample of mice from natural populations of the three major *M. musculus* subspecies to use as references and a large and diverse set of laboratory strains that can be effectively used to infer the genome of most remaining strains through imputation[13]. Our platform is a custom high-density genotyping array for mouse[17].

RESULTS

Sample and genotypes

We selected 198 samples for genotyping, including 36 wild-caught mice, 62 wild-derived laboratory strains and 100 classical strains (**Supplementary Table 1**). We used wild-caught mice, including representatives from *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*, as references to infer the phylogenetic origin of laboratory strains (**Supplementary Fig. 1**). Our laboratory samples included strains derived from different stocks and by different laboratories[5], as well as 13 sets of classical substrains that are thought to be closely related to each other.

Every sample was genotyped with the Mouse Diversity array[17]. We performed additional steps to improve the quality of the genotype calls and to detect residual heterozygosity and deletions larger than 100 kb. Our genotype dataset included SNPs and variable intensity oligonucleotides (VINOs). The latter represent previously unknown genetic variants that substantially alter the performance of SNP detection probes (Online Methods). We used 549,599 SNPs and 117,203 VINOs with six possible calls: homozygous for either allele, heterozygous, VINO, deletion and no call. In the analyses based on SNPs, we treated VINOs as no calls. In the analyses based on VINOs, we treated data as binary for the presence or absence of VINOs. SNPs and VINOs have complementary characteristics that can be used to strengthen phylogenetic analyses (see the discussion section).

Heterozygosity and deletions in laboratory strains

We used the local frequency of heterozygous calls to identify regions with two distinct haplotypes in a sample. We deemed such regions heterozygous. Wild-caught mice were predominantly heterozygous, and the variation in the heterozygosity rate (**Supplementary Table 1**) among subspecies was as expected from sequencing studies[2]. Wild-derived strains have wide variation in heterozygosity, and most classical strains are fully inbred. There are, however, some blocks of residual heterozygosity of variable size and distribution among lab strains (**Supplementary Table 2**). We detected the presence of deletions in 102 samples and determined their boundaries by visual inspection of probe intensity plots (**Supplementary Table 3**). We excluded these large deletions from our phylogenetic analysis. The analysis of structural variation in laboratory strains will be reported elsewhere.

Diagnostic alleles

We used the genotypes of the 36 wild-caught mice to determine the ability of each SNP or VINO to discriminate between subspecies, allowing for some misclassification caused by genotyping error, homoplasy or gene flow in the wild. Alleles found in only one subspecies were considered diagnostic. These include fully informative alleles, in which subspecies are fixed for different alleles, and partially informative alleles, in which an allele is restricted to one subspecies but not fixed. We identified 251,676 SNPs and 96,188 VINOs with diagnostic alleles distributed across every chromosome (**Supplementary Fig. 2**). SNPs and VINOs with nondiagnostic alleles are also distributed evenly across the genome but were not used to infer ancestry.

We found substantial differences between the number of SNPs and VINOs with diagnostic alleles for each the three subspecies detected. For example, 55% of all informative SNPs carry diagnostic alleles for *M. m. domesticus*, whereas only 27% and 18% carry diagnostic alleles for *M. m. musculus* and *M. m. castaneus*, respectively. This situation is reversed among VINOs, where 17%, 24% and 59% of diagnostic alleles identify *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*, respectively. These differences reflect two types of biases with compensatory effects. On one hand, the selection criteria for inclusion of SNPs in the array led to the over-representation of SNPs with *M. m. domesticus* diagnostic alleles and under-representation of *M. m. castaneus* SNPs[17]. On the other hand, our deeper knowledge of the genetic variation present in the *M. m. domesticus* subspecies allowed screening of candidate SNP probes with internal polymorphisms that could create VINOs, whereas our limited knowledge of the genetic variation present in the *M. m. castaneus* subspecies in particular results in an excess of *M. m. castaneus* diagnostic VINOs[2,7].

We confirmed the taxonomic classification of the 36 wild-caught samples by generating phylogenetic trees for the autosomes, sex chromosomes and mitochondria. All trees are consistent with the expected subspecific origin (**Supplementary Fig. 3**).

Subspecific origin of classical strains

We used informative SNPs and VINOs to infer the subspecific origin of every region of the genome of each sample. **Figure 1** shows the overall contribution of each subspecies to the autosomes; **Figure 2a** provides a map of the subspecific origin for chromosomes 6 and X (see URLs for a link to the complete data). The genome of classical inbred strains is predominantly derived from *M. m. domesticus* (94.3% ± 2.0% (s.d.)), with variable contribution from *M. m. musculus* (5.4% ± 1.9%) and a small contribution from *M. m. castaneus* (0.3% ± 0.1%). The contribution from subspecies other than *M. m. domesticus* is not distributed randomly across the genome or among strains (**Fig. 2**). In the combined 100 classical inbred strains, *M. m. musculus* haplotypes can be found in only 46.9% of the genome and *M. m. castaneus* haplotypes can be found in 2.8%. There is a strong bias toward multiple strains sharing the same *M. m. musculus* haplotype in some regions.

Notably, the *M. m. castaneus* and *M. m. musculus* contributions are not independent from each other, with the former frequently nested within or contiguous with the latter (**Fig. 2**). This association suggests an *M. m. molossinus* origin of the *M. m. musculus* contribution to the classical inbred strains[18,19]. We tested this hypothesis by comparing the *M. m. musculus* regions found in classical inbred strains to wild-caught *M. m. musculus* mice from Europe or Asia (**Supplementary Fig. 3**). Over 90% of the *M. m. musculus* haplotypes found in classical inbred strains cluster with Asian wild-caught mice.

Haplotype diversity and identity by descent in classical strains

The subspecific origin of classical inbred strains supports the hypothesis that these strains are derived from a small population of fancy mice that was itself subject to substantial inbreeding. To estimate the size of the fancy mice population from which classical inbred strains

are derived, we divided their genome into overlapping intervals that have no evidence for historical recombination (Online Methods). We identified 43,285 intervals (median size = 71 kb and median number of SNPs = 12). The distribution of the number of haplotypes in each interval (median and mode = 5) indicates that the original population harbored a limited number of distinct chromosomes (**Supplementary Fig. 4a**). Over 97% of the genome can be explained by fewer than ten haplotypes. In conclusion, classical strains can be partitioned locally into a small number of classes, within which all strains are identical by descent (IBD) with respect to their common origin. Intervals

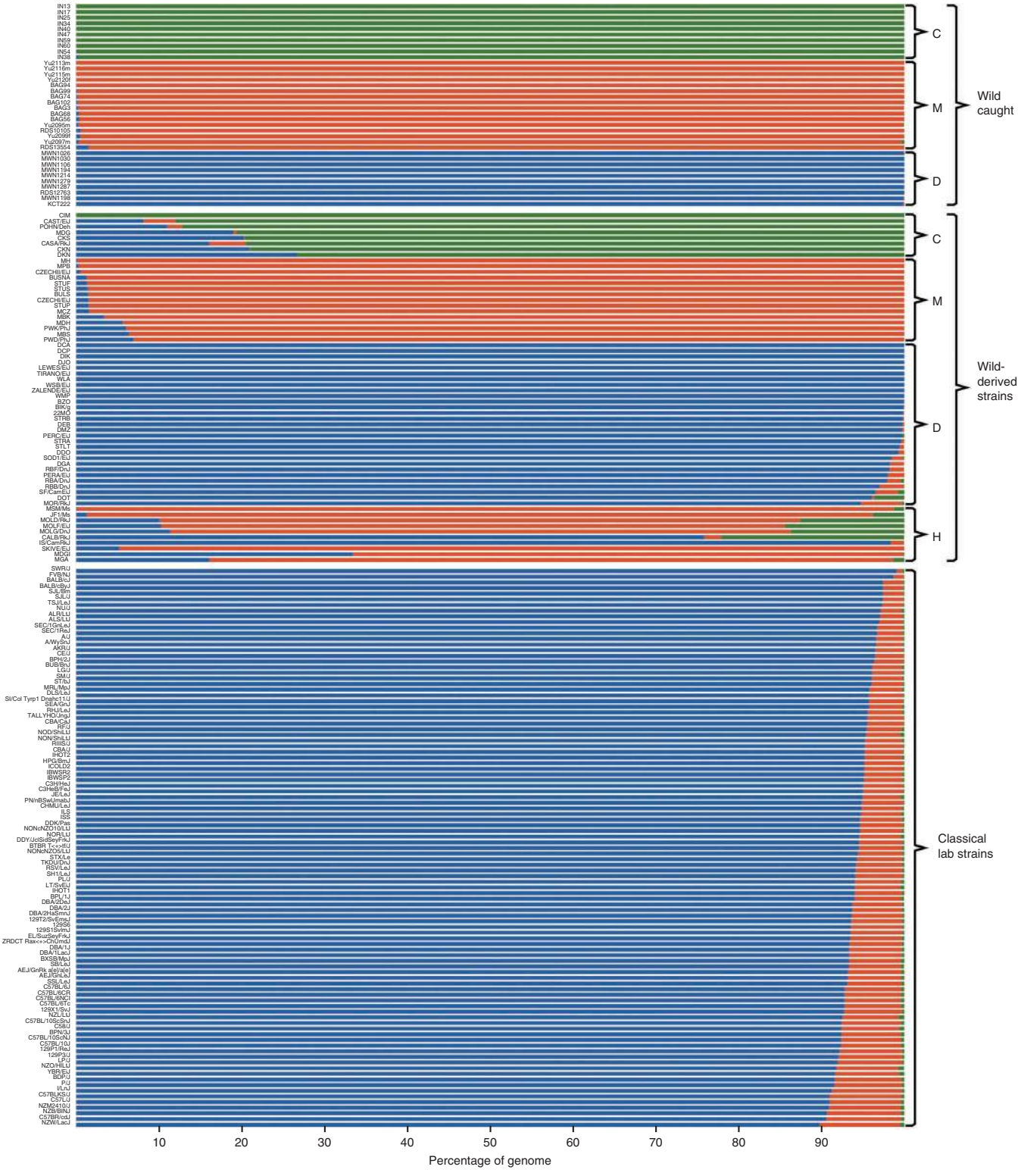

Figure 1 Overall contribution of each subspecies to the genome of wild and laboratory mice. For each sample, the figure depicts the cumulative contribution of *M. m. domesticus* (D, blue), *M. m. musculus* (M, red) and *M. m. castaneus* (C, green) subspecies for the autosomes. H, hybrid strains.

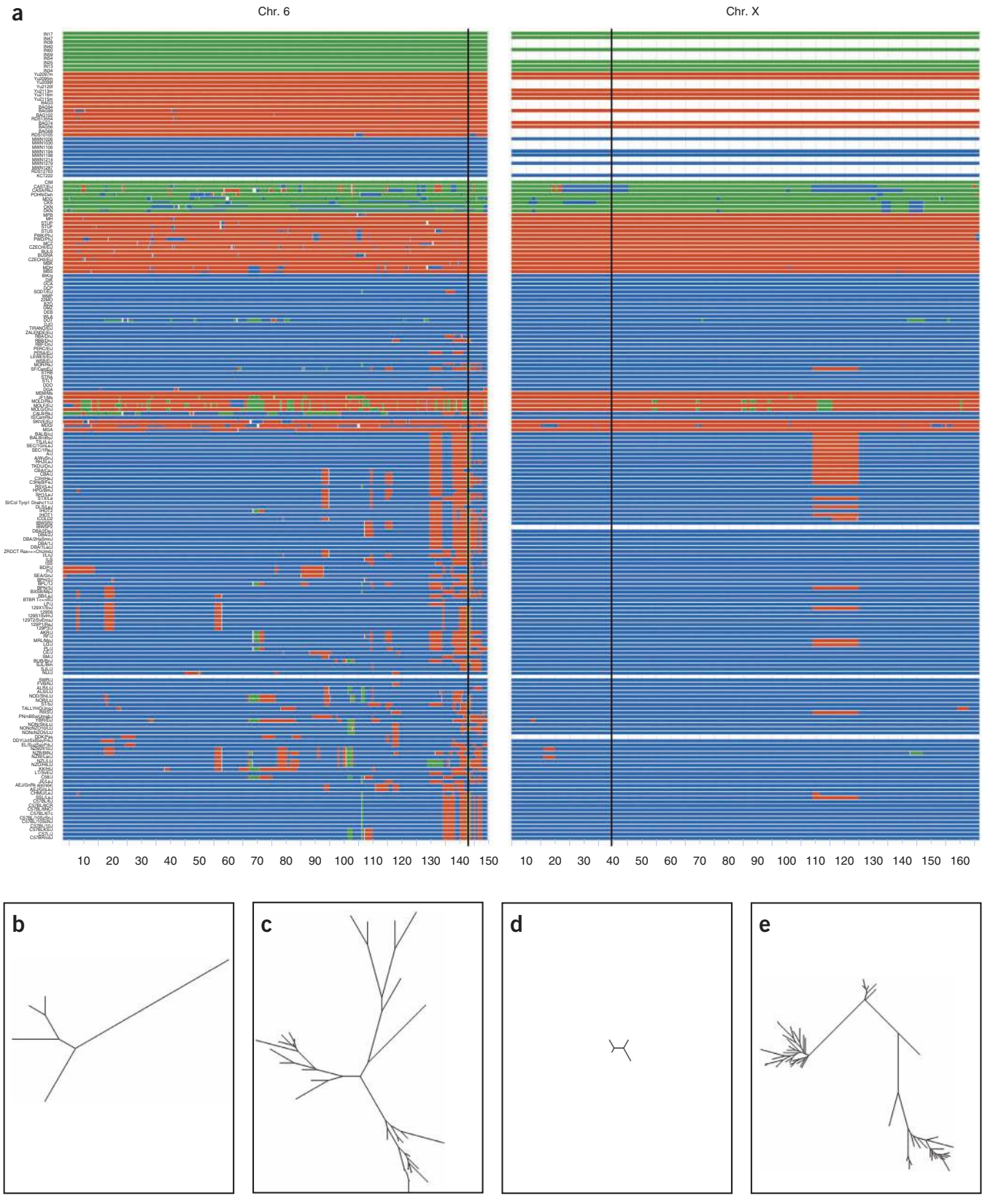

Figure 2 Subspecific origin and haplotype diversity of chromosomes 6 and X. (**a**) Subspecific origin of chromosome 6 (left) and X (right). Colors follow the same conventions as in **Figure 1**. (**b**–**e**) Phylogenetic trees for classical and wild-derived strains for two compatible intervals, one spanning positions 143,009,892–143,140,072 on chromosome 6 (**b**,**c**) and the other spanning positions 37,770,186–42,329,981 on chromosome X (**d**,**e**).

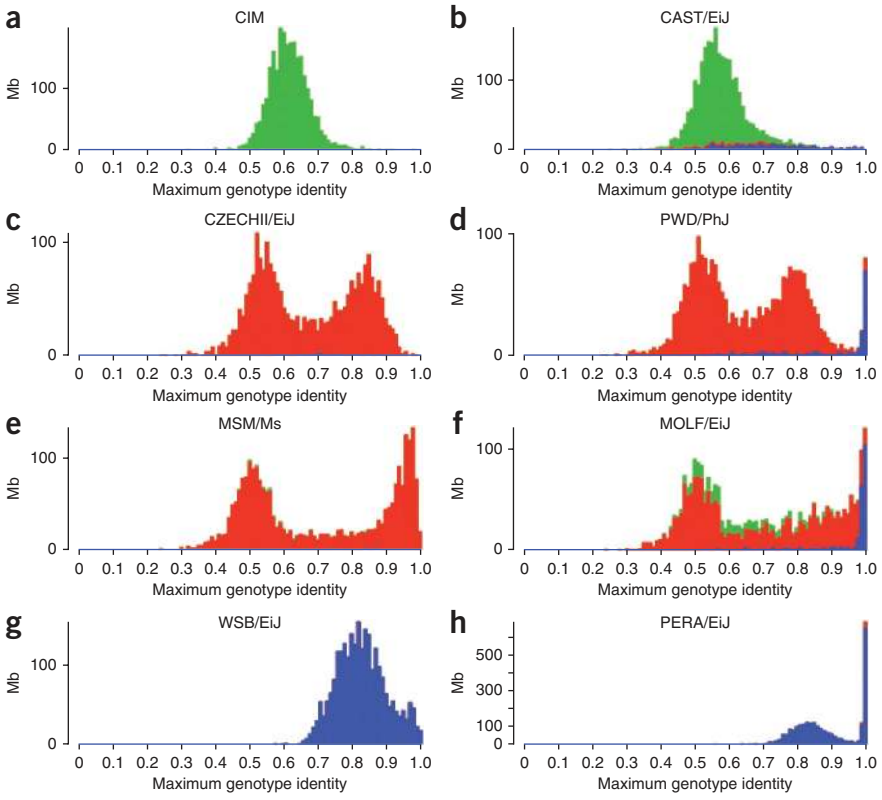

Figure 3 Intersubspecific introgression and contamination by classical strains in the wild-derived inbred strains. For each 1-Mb interval, we identified the classical inbred strain with maximum genotype similarity to a given wild-derived strain. (**a**–**h**) Frequency distribution of similarity for eight strains. Colors follow the same conventions as in the previous figures.

with larger numbers of haplotypes often reflect accumulation of new mutations in the past century, as shown by re-sequencing projects[6,7,10] and our analysis of substrains (**Supplementary Fig. 5**).

Recombination intervals provide a natural scaffold upon which to build genome-wide maps of haplotype diversity and IBD among classical strains. For each interval, we estimated the genotype identity among all pairs of strains and defined the minimum number and composition of cliques required to represent the haplotype variation. A critical step in this process was to determine a threshold of genotype identity that corresponds to IBD. This lower bound on genotype identity should be consistent with the accumulation of new mutations over several hundred generations and genotyping error. For this purpose, we carried out an analysis of local similarity among sister substrains. These closely related sets of strains, such as BALB/cJ and BALB/cByJ, did not show evidence of substantial genetic divergence or contamination (**Supplementary Fig. 5**). We established that 99.0% genotype identity is a suitable threshold for provisional assignment of local IBD status among strains. To further refine this assignment and to address the shortcoming of hard thresholding, we used clique completion to define sets of strains that are mutually IBD to each other and calculated the mean genotype identity within and between cliques. The distribution of the number of cliques is similar to the distribution of the number of haplotypes per interval (**Supplementary Fig. 4**). Using this approach, we generated a map of haplotype diversity in 100 classical inbred strains (see URLs).

Haplotypes can differ from each other just slightly more than our threshold to declare IBD (99%) or by as much as is typically observed between different subspecies (50%; see **Supplementary Fig. 6**). To estimate the local level of haplotype variation and to guide interpretation of the maps, we determined the quantitative similarity between haplotypes at each interval based on phylogenetic distance trees. **Figure 2c–e** shows two recombination intervals with obvious differences in the number of haplotypes and level of similarity among them. This illustrates the complex relationship between haplotype number and haplotype diversity among classical inbred strains.

Intersubspecific introgression in wild-derived laboratory strains

The recombination intervals computed for classical inbred strains cannot be easily extended to the wild-derived strains. Instead, we computed the frequency of diagnostic alleles in non-overlapping 1-Mb intervals and for each wild-derived strain. The majority of the genome of the 62 wild-derived laboratory strains originates from the expected subspecies or combination of subspecies (**Fig. 1**). However, only 9 strains have a genome derived entirely from a single subspecies, 18 have contributions from two subspecies and 35 have contribution from all three subspecies. The prevalence and extent of multi-subspecific origin is a defining characteristic of wild-derived laboratory strains as a group. Our set of wild-derived strains includes ten strains derived from natural intersubspecific hybrids (**Supplementary Table 1**), all of which have, unexpectedly, contributions from all three subspecies. The remarkable discordance in subspecific origin in several strains based on phylogenetic trees (**Supplementary Table 1** and **Supplementary Fig. 7**) provides further evidence for intersubspecific introgression. The sharing of patterns of subspecific origin between classical inbred strains and some wild-derived strains (**Fig. 2**) suggests that some of the intersubspecific introgressions in the latter group involved cross breeding with classical strains.

Relationship between classical and wild-derived laboratory strains

To characterize the relationship between the classical and wild-derived laboratory strains, we determined the maximum local level of genotype identity between each wild-derived strain and all classical inbred strains in non-overlapping 1-Mb windows and generated genome-wide similarity distributions (**Supplementary Fig. 6a**). The distributions of local similarity reveal the presence of distinct patterns for wild-derived strains of each of the three major subspecies. *M. m. domesticus* and *M. m. castaneus* wild-derived strains have typically unimodal distributions with distinct means (**Fig. 3**). In contrast, *M. m. musculus* and *M. m. molossinus* strains have a bimodal distribution of local genotype identity when compared to classical inbred strains.

This analysis provides insight into the origins of intersubspecific introgressions that occur in many of the wild-derived strains. Regions of near identity (>98%) with classical inbred strains indicate cross-breeding to extant classical strains or stocks descended from fancy mice. For example, 15 wild-derived strains (**Supplementary Table 1**) showed a distinct peak at levels of genotype identity (>98%) that are only consistent with recent IBD. The fraction of the genome involved

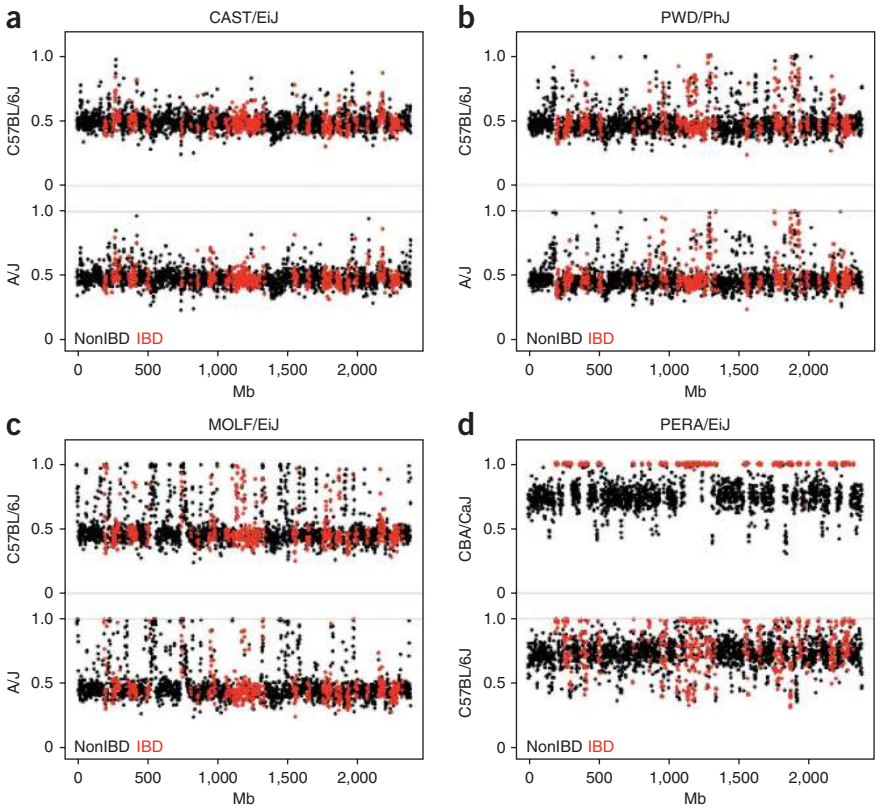

Figure 4 Identification of donor strains. (**a**–**d**) Examples of the approach used to identify the donor classical strain that contaminated a wild-derived strain. Red circles represent 1-Mb intervals in which a wild-derived strain is IBD to a haplotype present in classical inbred strains; black circles represent 1-Mb intervals that are not IBD.

ranges from 3.9% to 64.6%. Three wild-derived strains from three different subspecies (PWD/PhJ, MOLF/EiJ and PERA/EiJ) exemplify this pattern. In all three subspecies, regions of IBD to classical inbred strains are predominantly of *M. m. domesticus* origin but also include regions of *M. m. musculus* introgression (**Fig. 3**). This is particularly striking in the PERA/EiJ strain, providing further evidence of the role of classical laboratory strains in intersubspecific introgression in wild-derived laboratory stocks.

For each of the 15 wild-derived strains, we tested whether a single-donor classical strain can explain the overall pattern of IBD with all classical strains. Using this approach, we identified the donor of introgressed regions in six wild-derived strains (**Supplementary Table 1**), including PERA/EiJ. Contamination by CBA/CaJ explains all IBD regions in PERA/EiJ, whereas comparison with any of the other 99 classical inbred strains explains only a fraction of intervals of high local similarity (**Fig. 4**). Another six wild-derived strains appear to have been contaminated by classical laboratory mice that are not among our set of classical strains. The remaining 21 wild-derived strains that show evidence of intersubspecific introgression are not contaminated by classical laboratory strains.

The distribution of local similarity between wild-derived and classical inbred strains provides further insights into the origins of the non–*M. m. domesticus* regions in the genomes of classical inbred strains. When wild-derived *M. m. musculus* strains are compared to classical inbred strains (**Fig. 3e**,**f** and **Supplementary Fig. 6**), the peak with lower genotype similarity corresponds to genomic regions in which classical inbred strains completely lack *M. m. musculus* haplotypes. The peak with higher genotype similarity corresponds to regions in which at least one classical inbred strain carries a *M. m. musculus* haplotype and has an average SNP identity of 83%. When we make the same comparisons with *M. m. molossinus* wild-derived inbred strains, the high peak is shifted toward near complete identity (~98%). We conclude that the vast majority of *M. m. musculus* regions in classical strains are of *M. m. molossinus* origin.

DISCUSSION

There are two competing views on the origin and composition of the genome of classical inbred strains[6,7]. One study concluded that the genome of these strains is 68% *M. m. domesticus*, 10% *M. m. molossinus*, 6% *M. m. musculus*, 3% *M. m. castaneus* and 13% of unknown origin[6]. On the other hand, we previously concluded that 92% is of *M. m. domesticus*, 6% is of *M. m. musculus* and 1% is of *M. m. castaneus* origin[7]. Both studies were based on data from the National Institute of Environmental Health Sciences (NIEHS)[6], but they took different approaches to the use of wild-derived inbred strains as reference genomes to infer subspecific origin. Researchers from a previous study[6] assumed that the four wild-derived strains, WSB/EiJ, PWD/PhJ, CAST/EiJ and MOLF/EiJ, were faithful representatives of four subspecies, *M. m. domesticus*, *M. m. musculus*, *M. m. castaneus* and *M. m. molossinus*, respectively. We concluded, however, that three of these wild-derived strains, PWD/PhJ, CAST/EiJ and MOLF/EiJ, had introgressed haplotypes from other subspecies. In regions where a given wild-derived strain has undergone such intersubspecific introgression, the genotypes are not suitable as a reference for that subspecies. The results presented here conclusively show that classical inbred strains are overwhelmingly derived from *M. m. domesticus*, that the non–*M. m. domesticus* contribution to their genomes is largely of *M. m. molossinus* origin and that intersubspecific introgression is common in wild-derived laboratory strains.

The wild-caught mice used here represent a geographically diverse sample. The genomes of these mice are overwhelmingly derived from a single subspecies (mean = 99.84% and range = 98.42–100%). Half of wild-caught mice carry small regions with haplotypes from a second subspecies, mostly in heterozygous combinations. We acknowledge that a larger and more geographically diverse set of mice would be of great interest, but it would have little impact on our conclusions regarding the origin of the genome of the laboratory mouse. We also acknowledge that our definition of diagnostic alleles in SNPs and VINOs may change with the inclusion of more samples. However, this definition provides a simple and robust method to assign phylogenetic origin while preserving enough flexibility to account for genotyping error, homoplasy and gene flow among subspecies in the wild. Although our method works very well at a Mb genomic scale, it has limitations in providing subspecific assignments at finer scale (**Supplementary Fig. 8**).

Excluding hybrid strains, 28 wild-derived strains have intersubspecific introgressions covering between 1% and 27% of their genome

(**Fig. 1** and **Supplementary Table 1**). In CAST/EiJ and PWD/PhJ, the two strains that were used as references in previous studies, introgression covers 12% and 7% of their genome, respectively, confirming 96% of the regions that were declared introgressed in our previous study (**Supplementary Fig. 9**). We have been able to identify additional regions of introgression in CAST/EiJ and PWD/PhJ because of the better reference genotypes for each subspecies and the combined use of SNPs and VINOs. Subspecies, time since derivation and laboratory history appear to have a strong effect on the prevalence and extent of intersubspecific introgression, which could have occurred in the wild or in the laboratory. The limited extent of introgression in wild-caught samples suggests that breeding in the laboratory played a major role in shaping the genomes of wild-derived strains. Independent confirmation was obtained by comparing the genomes of wild-derived and classical inbred strains. Fifteen wild-derived strains have inherited haplotypes from classical inbred strains. Contamination by classical strains was expected, and likely intentional, in some cases (SOD1/EiJ and RBB/DnJ) but not in others (CASA/EiJ and CALB/RkJ). Introgression in the remaining wild-derived strains probably arose though a combination of gene flow in the wild (in samples captured close to hybrid zones and recently colonized regions) and breeding in the laboratory to non-classical mouse stocks (most likely other wild-derived mice). Wild-derived inbred strains have been used frequently as models in evolutionary studies[20]. Our results suggest that new information about the subspecific origin of the strains should be incorporated in the analyses.

A complementary strength of our study was the ability to account and correct for ascertainment biases in the SNPs included in the array. Most of these SNPs were selected on the basis of the local phylogeny among the NIEHS strains. This approach ensured that all major local branches were represented while ignoring minor branches. However, the approach also had limitations because locally all branches represented in the array were allocated the same number of SNPs, and therefore, long and short local branches would appear to be equal in length[17]. Furthermore, there are subspecies-specific false negative rates in SNP identification in the NIEHS study, and prior identification of a SNP is a necessary condition for its presence in the array[7]. Subspecies-specific false negative rates in SNP discovery should also negatively impact the rate at which selected SNPs are converted into successful genotyping assays[17]. For example, *M. m. castaneus* SNPs should be under-represented compared to the true level of diversity because of combined effects of our selection criteria and the higher assay failure rate. However, we were able to overcome the high failure rate by using VINOs. For the purpose of this study, VINOs have the critical advantage of being less subject to ascertainment biases within a given phylogenetic group. However, VINOs can only be reliably detected in homozygosity, resulting in a substantial undercounting of VINOs in some samples (**Supplementary Table 1**). We conclude that the combination of SNP and VINO genotype data in wild-caught mice has enormous value for population studies.

Among the most useful results from the present study are the maps of subspecific origin and haplotype diversity of the genomes of classical inbred strains (**Fig. 2**). These maps should allow researchers to combine information from multiple crosses to refine candidate intervals. It should also extend the advantages of the very high-density genotype data in the 15 NIEHS strains (and eventually whole genome sequence) to many additional classical strains[5,10]. Our maps will enable researchers to determine not only which strains share the same haplotype in a given region but also the sequence divergence among those strains that do not share them. We have also calculated the number of variants used to infer IBD and a score to guide interpretation of these trees by potential users. In particular, we have flagged haplotypes with weak support. Our data and tools should allow researchers to rapidly determine the number of haplotypes in a given region and the level of sequence divergence among them. Both are important considerations for association mapping. These data will also allow researchers to identify discrete regions of genetic divergence between substrains. Finally, they may be used to select strains with the desired level and type of genetic variation in any given region of the genome.

The spatial distribution of mean genetic variation observed in the 100 classical strains analyzed here is very similar to the one reported previously for a set of only 12 classical strains[7] (**Supplementary Fig. 10**). Although our approach of recombination intervals cannot directly be extended to wild-derived strains, we used a fixed-window approach to determine the level of haplotype diversity and IBD among these strains. This analysis shows that there is much more diversity in wild-derived strains than in classical strains (**Fig. 2b–e**), providing opportunities to optimize genetic research. Analysis of the frequency distribution of genotype identity in pairwise comparisons between wild-derived strains provides insight into the natural history of these strains and the populations from which they were derived. In contrast with comparison to classical inbred strains, these distributions are typically unimodal in intrasubspecific comparisons (**Supplementary Fig. 6b**). However, we also observed a strong signature of IBD in several pairwise comparisons. Some of the strongest instances involve pairs of strains derived from mice trapped in geographically close localities (**Supplementary Table 1**). Excess IBD can be explained by the presence of introgression from classical inbred strains that are themselves IBD for a substantial fraction of their genome (**Supplementary Fig. 6**). There are some strains that are connected to several cliques, creating a complex network. Finally, all *M. m. molossinus* wild-derived strains (**Supplementary Table 1**) have very high levels of IBD (~34%). This observation and the unusually high level of genotype identity between the *M. m. molossinus* haplotypes present in classical strains and the wild-derived *M. m. molossinus* strains strongly suggest a recent population bottleneck in this hybrid subspecies.

In summary, our observation of residual heterozygosity among inbred mouse strains, the striking local differences in the level of genetic similarity between substrains, the identification of large deletions of different ages and prevalence of contamination emphasizes the importance of deep, unbiased and frequent genetic characterization of laboratory stocks. Our genome browser provides access to the trees and links between recombination intervals, local trees and the maps for subspecific origin and haplotype diversity. Our analysis shows that classical inbred strains are in fact mosaics of a handful of haplotypes present in the founder fancy mice population. The genetic divergence among these haplotypes varies widely both locally and across the genome. Furthermore, the contribution of subspecies other than *M. m. domesticus* is limited, and its distribution highlights the complex population structure in these strains. On the other hand, wild-derived laboratory strains represent a deep reservoir of genetic diversity untapped in classical strains and are in many cases analogous to the three-way intersubspecific hybrids that classical inbred strains were thought to be. Our previous work[7,21] combined with the results of the deep survey of mouse resources presented here shows that the laboratory mouse is an unparalleled model for genetic studies in mammals.

URLs. MouseDivGeno, http://cgd.jax.org/tools/mousedivgeno; genotypes, http://cgd.jax.org/datasets/popgen.shtml; MPV, http://msub.csbio.unc.edu/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. All sequences have been submitted to GenBank under accession numbers GU992455–GU992863.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences (NIGMS) Centers of Excellence in Systems Biology program, grant GM-076468, by a US National Institutes of Health (NIH) grant to M.W.N. (R01 GM74245), by a grant to F.B. (ISEM 2010-141) and by a Czech Science Foundation grant to J.P. (206-08-0640). J.P.D. was partially supported by NIH Training Grant Number GM067553-04, University of North Carolina (UNC) Bioinformatics and Computational Biology Training Grant. J.P.D., R.J.B. and T.A.B. are partially supported by an NIH grant to F.P.-M.d.V. (P50 MH090338). We also thank F. Oyola for help annotating the samples genotyped in this study.

AUTHOR CONTRIBUTIONS

F.P.-M.d.V., G.A.C. and H.Y. conceived the study design and wrote the paper. H.Y., J.R.W., J.P.D., L.M. and C.E.W. carried out the bioinformatics analyses. J.P.D., T.A.B. and R.J.B. prepared the samples and conducted the targeted PCR amplification and sequencing. F.B., P.B., A.H.-T.Y., M.W.N., J.P. and P.T. provided biological samples. All authors contributed to the interpretation of the results and the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Boursot, P., Auffray, J.C., Britton-Davidian, J. & Bonhomme, F. The evolution of the house mice. *Annu. Rev. Ecol. Syst.* **24**, 119–152 (1993).
2. Geraldes, A. *et al.* Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol. Ecol.* **17**, 5349–5363 (2008).
3. Teeter, K.C. *et al.* Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76 (2008).
4. Yonekawa, H., Takahama, S., Gotoh, O., Miyashita, N. & Moriwaki, K. Genetic diversity and geographic distribution of *Mus musculus* subspecies based on the polymorphism of mitochondrial DNA. in *Genetics in Wild Mice. Its application to Biomedical Research* (eds Moriwaki, K., Shiroishi, T. and Yonekawa, H.) 25–40 (Japan Scientific Societies Press, Tokyo, Japan, 1994).
5. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nat. Genet.* **24**, 23–25 (2000).
6. Frazer, K.A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).
7. Yang, H., Bell, T.A., Churchill, G.A. & Pardo-Manuel de Villena, F. On the subspecific origin of the laboratory mouse. *Nat. Genet.* **39**, 1100–1107 (2007).
8. Guénet, J.L. & Bonhomme, F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet.* **19**, 24–31 (2003).
9. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
10. Sudbery, I. *et al.* Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. *Genome Biol.* **10**, R112 (2009).
11. Chesler, E.J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome* **19**, 382–389 (2008).
12. Guan, C., Ye, C., Yang, X. & Gao, J. A review of current large-scale mouse knockout efforts. *Genesis* **48**, 73–85 (2010).
13. Szatkiewicz, J.P. *et al.* An imputed genotype resource for the laboratory mouse. *Mamm. Genome* **19**, 199–208 (2008).
14. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).
15. Boursot, P. & Belkhir, K. Mouse SNPs for evolutionary biology: beware of ascertainment biases. *Genome Res.* **16**, 1191–1192 (2006).
16. White, M.A., Ané, C., Dewey, C.N., Larget, B.R. & Payseur, B.A. Fine-scale phylogenetic discordance across the house mouse genome. *PLoS Genet.* **5**, e1000729 (2009).
17. Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nat. Methods* **6**, 663–666 (2009).
18. Nagamine, C.M. *et al.* The musculus-type Y chromosome of the laboratory mouse is of Asian origin. *Mamm. Genome* **3**, 84–91 (1992).
19. Tucker, P.K., Lee, B.K., Lundrigan, B.L. & Eicher, E.M. Geographic origin of the Y chromosomes in "old" inbred strains of mice. *Mamm. Genome* **3**, 254–261 (1992).
20. Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J.C. & Forejt, J. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* **323**, 373–375 (2009).
21. Ideraabdullah, F.Y. *et al.* Genetic and haplotype diversity among wild-derived mouse inbred strains. *Genome Res.* **14**, 1880–1887 (2004).

ONLINE METHODS

Sample preparation and genotyping. Most DNA samples were prepared at the University of North Carolina and all were genotyped using the Mouse Diversity Array[17] at The Jackson Laboratory. The processed arrays were computationally genotyped using MouseDivGeno (see URLs), a genotyping software written in R language specifically designed for the Mouse Diversity array. Genotyping of the samples involved three steps: normalization of the intensity variation caused by restriction fragment lengths in the genome amplification step and the C+G content of probe sequences; genotype calling using a combined maximum likelihood and hierarchical clustering algorithm; and identification of VINOs, as described below. We excluded 73,525 SNPs out of a total of 623,124 based on poor performance among our samples. We identified thousands of previously unknown genetic variants using an algorithm designed for mutation discovery in the Affymetrix platform. VINOs are characterized by a distinct clustering of samples with low hybridization intensity and designated by the genotype 'V'. The genotype of the target SNP in a sample with a VINO call is missing. To confirm that VINOs do indeed represent previously unidentified genetic variation, we selected 15 SNP probes with VINO calls, and for each probe, we selected at least four samples of each genotype (homozygous for allele A, homozygous for B or VINO) for targeted sequencing. Strains for resequencing were selected to maximally sample across subspecies and strain type (classical or wild derived). Primers were designed approximately 200 bp proximal and distal to each probe using PrimerQuest (Integrated DNA Technologies). Probe regions were amplified by PCR and sequenced by automated Sanger sequencing at UNC. Sequences were aligned using Sequencher 4.9 (Gene Codes). **Supplementary Table 4** lists all probes, strains and primer sequences used. All homozygous SNP genotype calls were confirmed (211 out of 211) as were most of the VINO calls (14 out of 15). Unconfirmed VINO calls could be explained by polymorphisms outside of the sequenced region that, for example, alter the cut sites for the enzymes used for genome-wide amplification. Thus, 100% validation was not expected.

We mapped regions of heterozygosity in each laboratory strain by calculating the frequency of heterozygous calls in 500-kb windows with 250-kb overlaps and applied a Hidden Markov Model (HMM) with strain-specific noise level. We found that most heterozygous calls in inbred strains reflect genotype calling errors that are randomly distributed throughout the genome, whereas in truly heterozygous regions, heterozygous calls occur in clusters. Array probe design was based on the reference C57BL/6J genome, which is mainly *M. m. domesticus*. Thus, genotype error rates are higher in strains that do not share common subspecific origin with C57BL/6J. All heterozygous calls in laboratory strains outside of heterozygous regions were replaced by no calls.

We identified large deletions that resulted in hybridization failures (VINOs) in multiple consecutive probes by calculating the VINO frequency in 500-kb windows with 250-kb overlap. Using an HMM, we identified contiguous intervals in which VINO frequencies were higher than the strain-specific noise level. We visually mapped the start and end of deletions and designated genotypes in these regions as 'D'. We validated nine of the putative deletions using PCR to amplify markers within and flanking the deletions in DNA samples with or without the deletions. There was 100% concordance between our predictions and the results of this test. See URLs for all genotypes.

Identification of SNPs and VINOs with diagnostic alleles. We used 10 *M. m. domesticus*, 16 *M. m. musculus* and 10 *M. m. castaneus* wild-caught mice to identify informative SNPs and VINOs. For each subspecies, we identified SNPs and VINOs for which all mice from the remaining two subspecies shared the same allele and denoted the alternative allele as diagnostic. For instance, if all *M. m. domesticus* mice have an A allele and all *M. m. musculus* and all *M. m. castaneus* mice have a B allele at a SNP, then the A allele at that SNP is a fully informative and diagnostic for *M. m. domesticus*. We assigned fully informative SNPs a score of 1. In addition, there are cases where the A allele occurs in only one subspecies but is not fixed in that subspecies. These partially informative SNPs are assigned a score that is the fraction of mice with the homozygous A genotype over the total number of mice in the subspecies. We allowed for up to two misclassifications because of genotyping errors (typically homozygous calls), homoplasy or gene flow in the determination of diagnostic alleles and penalized the score by a factor of 0.5 (one genotype error) or 0.3 (two genotyping errors). No calls and VINOs were ignored in this procedure. We then applied the same rule to find fully and partially informative VINOs based on dichotomized genotypes (VINO or no VINO).

Assignment of subspecific origin. We assigned subspecific origin based on diagnostic alleles and scores from a given subspecies in each region of a sample. An HMM was used to identify the boundaries and subspecific origin based on the cumulative scores within these regions.

Recombination intervals and perfect phylogeny trees. The genome of classical inbred strains was partitioned into overlapping intervals that show no evidence of recombination using the four-gamete test. Maximal intervals were computed by a left-to-right scan, adding successive SNPs to an interval until one is not four-gamete compatible with any SNP in that interval. The starting point of the next interval was found by removing SNPs from the left side until all incompatibilities have been removed, and left-to-right scan resumed. All resulting intervals were maximal and could not be extended in either direction. A minimal subset of these intervals was found that covers the entire genome while maximizing their overlap. This is computed by finding the longest path in a k-partite graph[22]. For each such compatible interval, there exists a 'perfect' phylogenetic tree in which each node corresponds to a haplotype and each edge to SNPs with the same strain distribution.

Identity by descent. To identify IBD regions in classical strains, we first performed pairwise comparisons and then expanded the IBD strain set using a clique-finding algorithm. IBD regions were defined based on the compatible intervals framework described above. The sizes of the compatible intervals were often too small to calculate robust statistics; thus, we merged consecutive compatible intervals for pairs of strains sharing the same terminal leaf node of consecutive perfect trees. Based on the merged intervals, we calculated a pairwise genotype similarity score as the proportion of matching variants (SNPs and VINOs) in that interval. After we assigned the score to each pair in each compatible interval, we identified the cliques in each interval. We connected pairs of strains with similarity scores >0.99. To accommodate poorly performing samples and noise, we implemented a clique extension algorithm and generated a single clique if at least 80% of edges were connected and the mean average similarity was >0.99. Strains belonging to the same clique in an interval were considered IBD over that interval. The reliability of this IBD analysis depends on the number of variants used to calculate the similarity score. Thus, to estimate the degree of reliability in each clique, we calculated a clique penalty score. First, we calculated $P_{ij} = \log_{10}$ (number of variants used to calculate the similarity score) for every pair of strains, and we capped the number of variants per interval at 100. Then, the penalty score is calculated as a variance of P_{ij}. The logarithmic transformation inflates the variance from pairs with a small number of variants. If the number of variants from all pairs of strains is more than 100, the penalty is zero. We flagged cliques with less than 20 variants or less than 40 variants with high clique penalty score. We excluded regions with very low SNP density from the IBD analyses. Excluded regions are listed in **Supplementary Table 5**. Finally, we excluded a single region with a pattern consistent with structural variation (**Supplementary Table 6**).

To identify regions of IBD in comparisons involving wild-derived strains, we calculated the genotype similarity in pairwise comparisons using 1-Mb nonoverlapping intervals. We declared regions to be IBD based on a threshold of 0.98 identity, but we also considered the overall shape of the frequency distribution.

Distance trees. Each distance tree is based on the mean score of strains belonging to the same clique and provides a quantitative measure of difference among strains belong to different cliques. In each compatible interval, we generated a similarity clique score matrix M of size $N \times N$, where N is the number of cliques, and each element $M[i,j]$ was a mean similarity between strains belonging to clique i and clique j. We built a neighbor-joining tree based on this matrix.

Clique coloring. Using eight pastel colors, we assigned unique colors to each haplotype in an interval such that the total color change across all intervals was minimized. For the first interval, colors were assigned arbitrarily to each haplotype. If there were more than eight haplotypes in an interval, the least frequent were not assigned colors and remain white. For each subsequent

doi:10.1038/ng.847

interval, every haplotype was assigned a color such that the total number of color transitions in each interval was minimized. There were no constraints on the color differences among intervals that were not adjacent, so this method does not ensure that large blocks of identity, perhaps punctuated by a discordant interval, are of a consistent color.

Web browser. The Mouse Phylogeny Viewer (MPV, see URLs) is intended to provide visual summaries of the results of this study and to allow downloading of the relevant information for selected strains in selected regions of the genome. A tutorial and the LAMP capabilities and meaning of the different analysis are provided online. See URLs for the complete set of genotypes.

22. Wang, J., Moore, K.J., Zhang, Q., Pardo-Manuel de Villena, F., Wang, W. & McMillan, L. Genome-wide compatible SNP intervals and their properties. *Proceedings of ACM International Conference on Bioinformatics and Computational Biology* (Niagara Falls, New York, USA, 2010).

nature genetics

On the subspecific origin of the laboratory mouse

Hyuna Yang[1], Timothy A Bell[2], Gary A Churchill[1] & Fernando Pardo-Manuel de Villena[2]

The genome of the laboratory mouse is thought to be a mosaic of regions with distinct subspecific origins. We have developed a high-resolution map of the origin of the laboratory mouse by generating 25,400 phylogenetic trees at 100-kb intervals spanning the genome. On average, 92% of the genome is of *Mus musculus domesticus* origin, and the distribution of diversity is markedly nonrandom among the chromosomes. There are large regions of extremely low diversity, which represent blind spots for studies of natural variation and complex traits, and hot spots of diversity. In contrast with the mosaic model, we found that most of the genome has intermediate levels of variation of intrasubspecific origin. Finally, mouse strains derived from the wild that are supposed to represent different mouse subspecies show substantial intersubspecific introgression, which has strong implications for evolutionary studies that assume these are pure representatives of a given subspecies.

Laboratory mice, the most popular model organism in mammalian genetics[1,2], were derived from wild mice belonging to the *Mus musculus* species by an intricate process that included the generation of 'fancy' mice in both Asia and Europe and a complex web of relationships between inbred strains[3]. Early studies showed that the mitochondria and the Y chromosome present in many classical laboratory strains were derived from different subspecies, *M. m. domesticus* for the mitochondria and *M. m. musculus* for the Y chromosome[4,5]. Furthermore, the Y chromosome was introduced into the laboratory mouse through *M. m. molossinus* males[6]. Based on these findings, it was proposed that the genomes of inbred strains are a mosaic of regions that have different subspecific origins[7]. Recently, the fine structure of such mosaic variation was described[8]. This study reported that strain-to-strain comparisons reveal regions with extremely high variation that span one-third of the genome and regions with extremely low variation that cover the remaining two-thirds of the genome. This distinctively bimodal distribution was assumed to represent regions that have different, or the same, subspecific origins, respectively. This mosaic model has been the driving concept behind mouse association mapping studies and haplotype analysis[9–12]. However, the origin of a given region of a laboratory strain could not be directly assigned to a subspecies owing to the lack of reference sequences for the three main mouse subspecies. Subsequent studies raised questions about the haplotype structure[11,13], the effect of ascertainment biases in subspecific assignment[14–16] and the contributions of intersubspecific versus intrasubspecific variation[17]. Several studies reported the presence of substantial intrasubspecific variation, ancestral polymorphisms and secondary introgression after the divergence of the subspecies[17–20], further complicating the interpretation of the data.

In 2004, the National Institute of Environmental Health Sciences (NIEHS) contracted Perlegen Sciences to resequence 15 mouse inbred strains. This project has released more than 109 million genotypes for 8.3 million SNPs that span the nuclear and mitochondrial genomes[21]. The 15 strains were selected on the basis of their genetic diversity, ease of breeding, inclusion in the Mouse Phenome Project, widespread use in research and background information. This set includes 11 classical strains (129S1/SvImJ, A/J, AKR/J, BALB/cBy, C3H/HeJ, DBA/2J, FVB/NJ, NOD/LtJ, BTBR T^+ *tf*/J, KK/HlJ and NZW/LacJ) and four strains derived from the wild (hereafter 'wild-derived' strains) (WSB/EiJ, PWD/PhJ, CAST/EiJ and MOLF/EiJ), which represent the *M. m. domesticus*, *M. m. musculus*, *M. m. castaneus* and *M. m. molossinus* subspecies, in corresponding order[22–25] (http://www.jax.org). *M. m. molossinus* is a subspecies that arose by natural hybridization between *M. m. musculus* and *M. m. castaneus*. The data are hereafter referred to as the NIEHS data.

We set out to use this resource to examine the ancestral subspecific origin of classical strains, expecting to identify a mosaic of segments that could be assigned to one of three distinct lineages: *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*[8,26]. We planned to use the three wild-derived strains as a reference for each subspecies and then assign genomic segments from classical strains to a subspecies, based on the pattern of SNP similarity between the query strain and the reference strains.

RESULTS

Diagnostic SNPs

Evolutionary models suggest that the three main mouse subspecies diverged simultaneously from a common ancestor or, alternatively, that *M. m. musculus* and *M. m. castaneus* diverged from a common ancestor shortly after the divergence of *M. m. domesticus*[27–29]. This history should be reflected in the distribution of SNPs that are specific to each subspecies. SNPs that have arisen since the divergence of the three subspecies should be equal in number or alternatively, be slightly

[1]The Jackson Laboratory, Bar Harbor, Maine 04609, USA. [2]Department of Genetics, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA. Correspondence should be addressed to F.P.-M.d.V. (fernando@med.unc.edu).

Received 13 February; accepted 31 May; published online 29 July 2007; doi:10.1038/ng2087

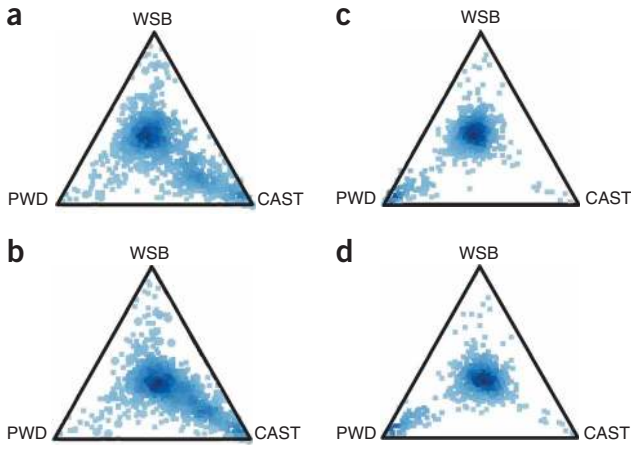

Figure 1 Frequency distribution of diagnostic subspecific SNPs. The relative frequency of diagnostic SNPs in 100-kb intervals is represented as a density plot over the simplex. In each plot, the three reference strains are indicated at the vertices of a triangle, and the relative proportions of diagnostic SNPs in each interval are represented as blue dots. Darker areas represent regions with a higher density of intervals. (**a**,**c**) Data for chromosome 14. (**b**,**d**) Data for chromosome X. **a** and **b** show the original data; **c** and **d** show the predicted proportions of diagnostic SNPs after correcting for the frequency-dependent SNP discovery rate.

enriched for *M. m. domesticus* SNPs, and these SNPs should be distributed evenly throughout the genome.

For the purpose of interpreting the NIEHS data, we define diagnostic SNPs as those that are completely genotyped and polymorphic among the three reference strains, WSB, PWD and CAST. Note that the diagnostic allele at some of these SNPs may not be shared by all individuals of that subspecies if it arose recently. Furthermore, because of incomplete sorting or homoplasy, the allele can also be present in individuals of other subspecies. Despite the limitations of using a single reference strain to define diagnostic SNPs, it remains the simplest method to test our expectations on the basis of phylogenetic history. We identified 4,373,427 diagnostic SNPs: 1,481,373 (33.9%) are *M. m. domesticus* SNPs that distinguish WSB from CAST and PWD; 1,280,328 (29.3%) are *M. m. castaneus* SNPs that distinguish CAST from WSB and PWD; and 1,611,726 (36.9%) are *M. m. musculus* SNPs that distinguish PWD from WSB and CAST.

We divided the genome into nonoverlapping 100-kb intervals and determined the proportion of diagnostic SNPs for each subspecies in each interval. These proportions can be represented in a simplex, a triangular region of three-dimensional space that represents the proportions of the three types of diagnostic SNP. In this representation, an interval that contains equal numbers of the three types of diagnostic SNP is located at the center, whereas an interval that contains only one type of diagnostic SNP is located at the corresponding vertex (**Fig. 1**). In contrast to our expectations, we found that most intervals are not located at the center but consistently deviate away from the CAST vertex (**Fig. 1a**,**b**). The degree of distortion varies among chromosomes, but this observation holds true for all autosomes and the X chromosome (**Supplementary Fig. 1** online). Although a slight deviation toward WSB (*M. m. domesticus*) is predicted by one evolutionary model[29], a genome-wide deficit of diagnostic *M. m. castaneus* SNPs can be explained only by either a differential mutation rate in that subspecies, or a systemic undercounting of diagnostic CAST SNPs across the genome.

We also found many intervals with extremely distorted frequencies of diagnostic SNPs (**Fig. 1**). The pattern of extreme distortion varies among chromosomes (**Supplementary Fig. 1**). The two most common patterns are intervals that have an excess of *M. m. castaneus* SNPs (intervals that are located close to the CAST vertex; **Fig. 1a**) and intervals that have an excess of *M. m. musculus* SNPs (intervals that are located close to the PWD vertex; **Fig. 1b**). Remarkably, the low-level distortion against *M. m. castaneus* SNPs in many intervals exists on the same chromosome with extreme distortion in favor of *M. m. castaneus* SNPs in other intervals. This inconsistency suggests that low-level distortion and extreme distortion have different origins.

SNP ascertainment bias

The basic properties of the NIEHS data are provided in a **Supplementary Note** online. We determined the false-positive rate (FPR) and false-negative rate (FNR) in the NIEHS data set by direct resequencing of selected genomic fragments in the 15 NIEHS strains and the C57BL/6 strain (Methods). The FPR is 1.3%, which is similar to previously reported results[11,30]. We found six discordant genotypes between our data and the NIEHS data (0.3%) among the 2,089 genotypes compared. Therefore, the FPR is low and should have little impact on the distorted frequency of diagnostic SNPs that are observed in the reference strains. By contrast, the FNR is significantly higher than previously reported in humans[30]. Because Perlegen's SNP discovery algorithm was designed to minimize the FPR, a high FNR is expected.

The FNR is strongly correlated with the minor allele frequency (MAF). The number of undetected SNPs decreases as the MAF increases from 76% for singletons (SNPs in which the minor allele is present in a single strain) to 42% for SNPs in which the minor allele is shared by seven strains (**Fig. 2a**). Among singletons, the FNR is constant with respect to the genomic position and strain (**Fig. 2b**,**c**), which suggests that the MAF is directly responsible for the pronounced differences in FNR. The local FNR varies across the genome, depending on the MAFs of SNPs that are present in a given region, which in turn depends on the phylogenetic relationships between the strains in that region. We estimate that the average genome-wide FNR in the 15 resequenced strains is 67%, based on the distribution of MAFs among the 3.8 million completely genotyped SNPs and the experimental FNR for each MAF. Based on that FNR and the genome

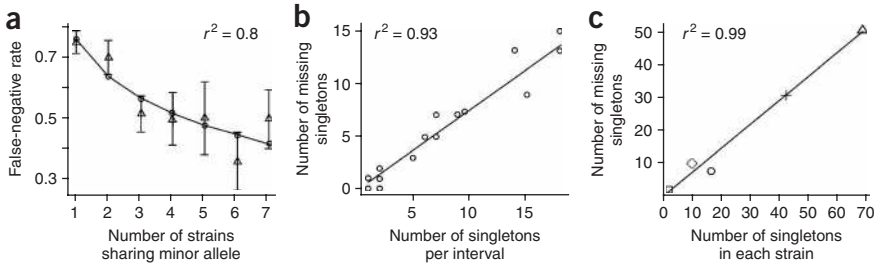

Figure 2 SNP discovery bias. (**a**) Effect of the minor allele frequency (MAF) on the false-negative rate (FNR). Triangles and vertical lines represent observed values that are ± 1 s.e.m. Circles represent the best fit of the data to the regression of log(FNR) on MAF. (**b**) The FNR for singletons in 24 resequenced intervals distributed across the genome. (**c**) The FNR for singletons from different strains (square, classical strains; open diamond, WSB; circle, PWD; cross, MOLF; triangle, CAST).

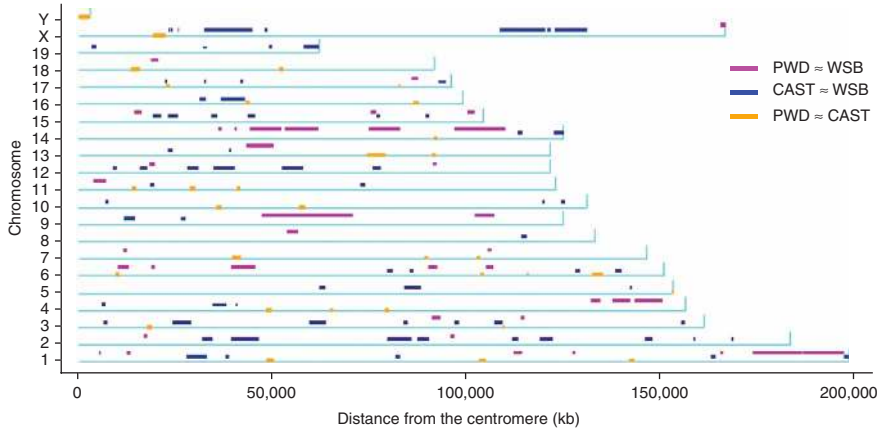

Figure 3 Regions of intersubspecific introgression in the reference strains. Inferred regions of intersubspecific introgression after smoothing the intervals with a hidden Markov model (HMM). Purple denotes regions with an excess of CAST diagnostic SNPs and a deficit of both WSB and PWD diagnostic SNPs. Blue denotes regions with an excess of PWD diagnostic SNPs and a deficit of both WSB and CAST diagnostic SNPs. Orange denotes regions with an excess of WSB diagnostic SNPs and a deficit of both CAST and PWD diagnostic SNPs.

coverage, we estimate that there are 45 million SNPs among the 15 resequenced strains. This number is within the range predicted by direct resequencing studies[18]. In conclusion, despite its exceptional size, density and quality, the NIEHS data capture only a fraction of the variation that is present in the laboratory mouse.

The finding that the FNR depends on the MAF implies that the probability of observing each type of diagnostic SNP depends on the local phylogenetic relationships between the 15 NIEHS strains. Furthermore, the MAFs of diagnostic SNPs vary among subspecies: singletons are predominantly CAST SNPs, doubletons are predominantly PWD and SNPs with higher frequencies are predominantly WSB (**Supplementary Fig. 2** online). To account for the allele-frequency-dependent FNR, we applied a branch-length correction (Methods) to phylogenetic trees. We then plotted the corrected length of the branches that represent each type of diagnostic SNP in each interval (the distance between the common node and the three reference strains CAST, PWD and WSB) and found that most intervals shifted toward the center of the simplex (**Fig. 1** and **Supplementary Fig. 1**). Therefore, the genome-wide low-grade distortion is due to a frequency-dependent SNP discovery rate that undercounts SNPs from lineages that are locally underrepresented among the NIEHS strains.

Intersubspecific introgression

Correcting for a frequency-dependent FNR had little effect on the intervals with extreme distortion (**Fig. 1** and **Supplementary Fig. 1**). These intervals are not SNP-poor and, therefore, are more prone to statistical fluctuations. Furthermore, intervals with an excess of diagnostic CAST SNPs or intervals with an excess of diagnostic PWD SNPs cluster in different megabase-long regions on particular chromosomes (**Fig. 3**). These features indicate that the distorted patterns are due to introgression of haplotypes from a different subspecies in one, or more, of the reference strains, which in turn indicates that the three wild-derived strains may not be pure representatives of each subspecies. Intersubspecific introgression has been reported in wild mice and in wild-derived strains[16,18–20]. Furthermore, MOLF, a wild-derived strain that is considered to be a representative of *M. m. molossinus*, carries an *M. m. domesticus* Y chromosome (**Supplementary Fig. 3** online). We conclude that MOLF also has introgressed haplotypes from a subspecies that is inconsistent with its phylogenetic history.

To delineate regions of the genome in which the reference strains accurately represent the three main subspecies, we first identified intervals with extreme distortion in favor of a single type of diagnostic SNP (**Fig. 1c,d** and **Supplementary Fig. 4** online) and applied a hidden Markov model (HMM) to consolidate larger regions that have a high concentration of unbalanced intervals of the same type. The HMM eliminates unbalanced intervals that lack local support. We left the status of these intervals undetermined. Balanced intervals within unbalanced regions are assigned by the HMM to an introgression class (**Fig. 3**). The remaining intervals with balanced frequencies of diagnostic SNPs and good local support span 72% of the mouse genome (**Supplementary Fig. 4**). These intervals probably represent regions in which the three reference strains are true representatives of the *domesticus*, *musculus* and *castaneus* mouse lineages. In summary, we partitioned the mouse genome into three classes: regions of potential introgression (13% of the genome; **Fig. 3**), regions with undetermined status (15%) and regions in which the three reference strains provide a balanced representation of the three main subspecies.

Approximately 5.7% of the genome has an excess of diagnostic CAST SNPs and a deficit of diagnostic SNPs for PWD and WSB (shown in purple in **Fig. 3**). In 5.9% of the genome there is an excess of diagnostic PWD SNPs and a deficit of CAST and WSB diagnostic SNPs (shown in blue in **Fig. 3**). The third pattern represented by an excess of WSB diagnostic SNPs (shown in orange in **Fig. 3**) is found in small regions spanning 1.3% of the genome, including the Y chromosome, and is consistent with the hypothesis that *M. m. musculus* and *M. m. castaneus* are sister subspecies[29].

We confirmed that, in the regions of potential introgression (**Fig. 3**), one of the three reference strains carries a haplotype from a different subspecies by sequencing short intervals in the three reference strains and in six additional wild-derived strains (**Supplementary Note**). These experiments confirm that most regions with extreme distortion in the frequency of diagnostic SNPs (**Fig. 1**) are due to introgression of *M. m. domesticus* haplotypes into PWD and CAST. Remarkably, some of the introgressed haplotypes span dozens of megabases and are unequally distributed along the genome. For example, there is an excess of *M. m. domesticus* introgression on chromosomes 14 and 9 in PWD and on the X chromosome in CAST (**Fig. 3**). Other wild-derived strains (CASA and PWK) may also have introgressed haplotypes from *M. m. domesticus*.

In regions of introgression we cannot directly determine the subspecific origin of classical strains. This shortcoming can be addressed by analyzing additional wild-derived strains. Previous conclusions in mouse on effective population sizes and on the rate of variants that are inconsistent with phylogeny owing to incomplete lineage sorting and homoplasy need to be re-evaluated[18,31]. Wild-derived inbred strains have previously been used to study the genetics of speciation[32–35]. Although our findings should not affect the hybrid sterility genes mapped using this approach, they may compromise the general conclusions that have been made about the genetic architecture of this critical process[14,32].

Ancestry of classical strains

To assign subspecific origins to genomic intervals of the 11 NIEHS classical strains, MOLF and C57BL/6, we examined the bias-corrected phylogenetic trees. In regions where diagnostic SNP frequencies are balanced, we assumed that the root of each tree was located at the node of common ancestry between PWD, CAST and WSB. Splitting the tree at this node partitions each of the remaining strains into one of three groups according to its local subspecific origin.

We first determined the matrilineal (mitochondria) and patrilineal (Y chromosome) inheritance patterns. We confirmed that the classical strains share an almost identical mitochondrial haplotype of *M. m. domesticus* origin (**Supplementary Fig. 3**), which supports the contention that laboratory strains descend from a very small pool of founders[4,36]. As expected from studies on mitochondrial variation in *M. m. molossinus*[23], MOLF has an *M. m. musculus* haplotype that is significantly different from the one carried by PWD.

Our analysis confirms the prevalence of the *M. m. musculus* (*molossinus*) Y chromosome among classical strains[6,37], and also indicates that many strains (FVB, NOD, BTBR and AKR) carry an *M. m. domesticus* Y chromosome. Interestingly, the *M. m. molossinus* strain MOLF carries an *M. m. domesticus* Y chromosome. Flow of mitochondrial DNA across subspecies boundaries[38] and discordant phylogenetic patterns between mitochondria and the Y chromosome have been reported in wild populations[19], which indicates that secondary introgression after radiation of the subspecies[20] might have contributed to the pattern of intersubspecific introgression observed in the wild-derived inbred strains, in addition to accidental 'contamination' in the laboratory.

We extended our assignment of the subspecific origin to the 72% of the autosomal and X-chromosomal genomic intervals for which the ancestry of the reference strains is unambiguous. **Figure 4** and **Supplementary Figure 5** online show the results for four representative chromosomes in which black denotes *M. m. domesticus* intervals, red denotes *M. m. musculus* intervals and green denotes *M. m. castaneus* intervals. In most regions, the subspecific assignments remain stable along a substantial length of the chromosome. This is particularly notable, given that the assignment was carried out automatically without any further attempt to smooth local fluctuations. Small, isolated segments of distinct subspecific origin do occur (**Fig. 4a**) and in some cases cluster in specific regions of the genome (**Supplementary Fig. 5**). The 100-kb interval size selected for our analysis may result in intervals that span transition zones between regions that have different subspecific origins, some intervals may contain smaller segments from a different subspecific origin embedded within them, and segmental duplications and copy number polymorphisms[39–41] may lead to unusual patterns in the assignment of ancestry. Although the subspecific assignment for each classical strain and MOLF is well supported by bootstrap replicates (85 out of 100 replicates in 94.5% of intervals and 99 out of 100 replicates in 86.3% of intervals), the ancestral origin may be

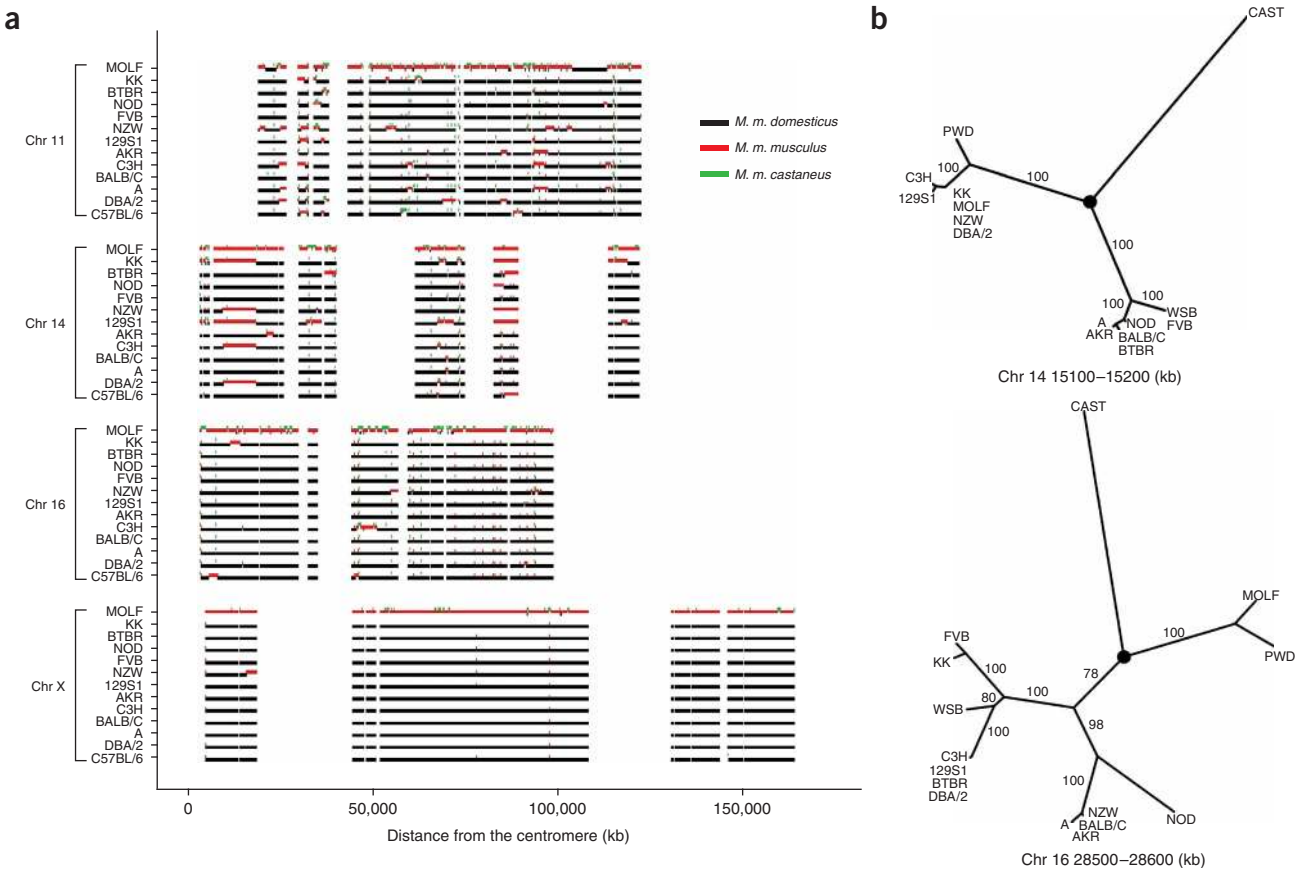

Figure 4 Subspecific origin of classical and hybrid strains. (**a**) Subspecific assignments for the 12 classical strains (including C57BL/6) for chromosomes 11, 14, 16 and X. Each 100-kb interval is shown as a vertical bar of a color that reflects its subspecific origin. Intervals without color are regions of intersubspecific introgression or of undetermined status. (**b**) Corrected phylogenetic trees for two 100-kb intervals in chromosomes 14 and 16. The circle denotes the assumed location of the root. Numbers represent bootstrap replicates that support each node.

Table 1 Contribution of the three main subspecific lineages to the genome of the laboratory strains

	B6	DBA/2	A	BALB/C	C3H	AKR	129S1	NZW	FVB	NOD	BTBR	KK	MOLF
M. m. domesticus	0.92	0.91	0.94	0.95	0.92	0.94	0.91	0.87	0.96	0.93	0.92	0.86	0.11
M. m. musculus	0.07	0.07	0.05	0.04	0.07	0.05	0.08	0.11	0.03	0.06	0.06	0.12	0.74
M. m. castaneus	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.15

Fractions of balanced 100-kb intervals assigned to each subspecies are shown.

incorrectly inferred in some intervals owing to limitations of the data and methodology.

The genomes of classical strains are overwhelmingly of *M. m. domesticus* origin (**Table 1**). Although a predominant contribution of that subspecies was predicted[8], the exceptionally high levels (ranging from 86% to 96%) observed in all strains were unexpected. The *M. m. musculus* subspecies has the second largest contribution to the genome of classical strains, whereas only 1–2% of their genome derives from *M. m. castaneus*. The spatial distribution of subspecific origins and the contribution of subspecies other than *M. m. domesticus* to the genomes of classical strains are not random. On chromosomes 16 and X, the classical strains are largely of *M. m. domesticus* origin, whereas MOLF is of *M. m. musculus* origin (**Fig. 4a**). By contrast, the classical strains have a more equilibrated contribution of *M. m. domesticus* and *M. m. musculus* on chromosome 14. Finally, chromosome 11 shows an intermediate situation. Many regions of *M. m. musculus* origin are shared among multiple classical strains, which suggests that the history and close relationships between strains has played a part in shaping the distribution of subspecific diversity.

Our analysis also confirms the presence of extensive introgression of *M. m. domesticus* haplotypes in MOLF (for example, distal chromosome 11; **Fig. 4** and **Supplementary Fig. 5**). The genome of this strain is a mosaic of three subspecies with 74% of *M. m. musculus* origin, 15% of *M. m. castaneus* origin and 11% of *M. m. domesticus* origin (**Table 1**).

Once the subspecific assignment of classical and hybrid strains was completed, we estimated the specificity and sensitivity of diagnostic SNPs. Remarkably, 88.7% of the 3,220,959 diagnostic SNPs that can be tested are completely specific; in other words, the diagnostic allele is not present among any of the NIEHS strains that have a different subspecific assignment. Conversely, the allele present at diagnostic SNPs is shared by all strains assigned to that subspecies in 59% of the 2,354,446 diagnostic SNPs in which this test can be carried out. Thus, despite the limited number of reference strains, diagnostic SNPs identified under our definition can be used collectively to assign subspecific origin.

Detailed images of regions with intersubspecific introgression, their subspecific ancestry and the supporting phylogenetic trees are available at the following website: http://www.genomedynamics.org.

Genetic variation in classical strains

In contrast with previous analyses[8–11,17], we have determined that, on average, 9% of the genome has a different subspecific origin between any given pair of classical strains, whereas 91% of the genome shares the same subspecific origin (**Fig. 5**). This indicates that many of the regions with high variation identified in previous studies might have the same subspecific origin. The deep branching of the *M. m. domesticus* lineage in many of the phylogenetic trees supports the model of high intrasubspecific variation (**Fig. 4b**).

To investigate the extent of genetic variation within versus between subspecies, we measured the pairwise distances along the bias-corrected trees in each 100-kb interval. These distances were then normalized to the average distance between pairs of strains from different subspecies in that interval. This normalized measure of variation allows direct comparisons between genomic regions that have different coverage, gene density and mutation rates, and makes it possible to calculate the normalized variation for all 25,400 100-kb intervals, regardless of the distribution of diagnostic SNPs. The distribution of within versus between subspecies variation (**Fig. 5**) demonstrates that our subspecific assignments that are based on tree topology are correct for the vast majority of intervals because pairs of classical strains thought to inherit segments from different species that are based on tree topology (**Fig. 4**) show a unimodal frequency

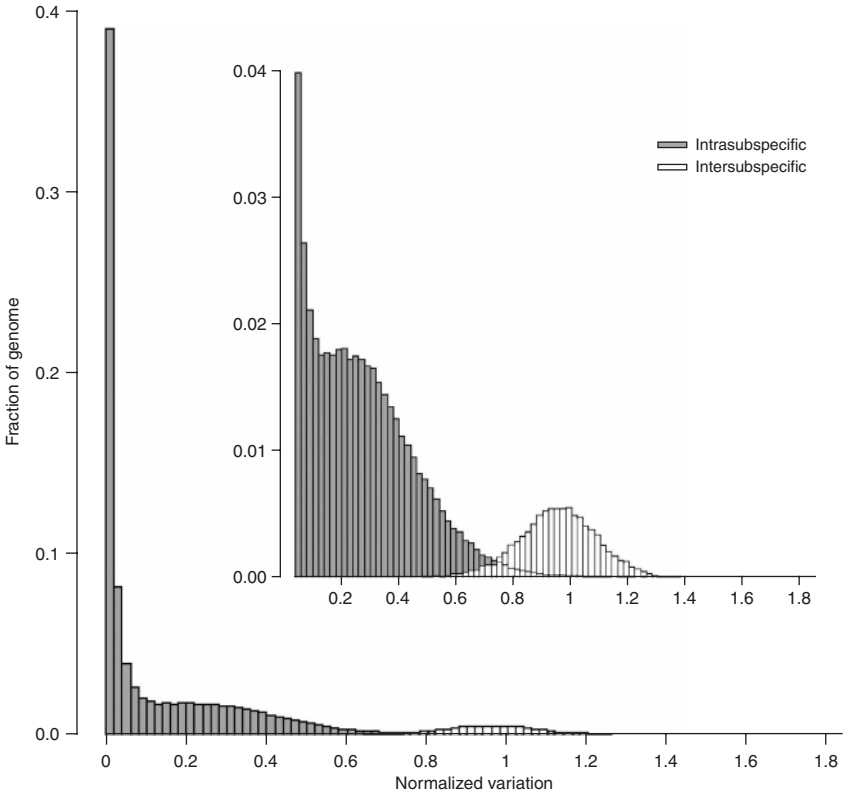

Figure 5 Frequency distribution of the normalized variation in pairwise comparisons between classical strains. The horizontal axis shows the normalized variation over 100-kb intervals for the 55 pairwise comparisons between 11 classical strains. The average variation in intersubspecific comparisons is set at one. White bars correspond to comparisons in which the two classical strains have haplotypes that are derived from different subspecies. Gray bars denote intrasubspecific comparisons. The inset expands the frequency distribution to emphasize the component of variation that is >0.4%.

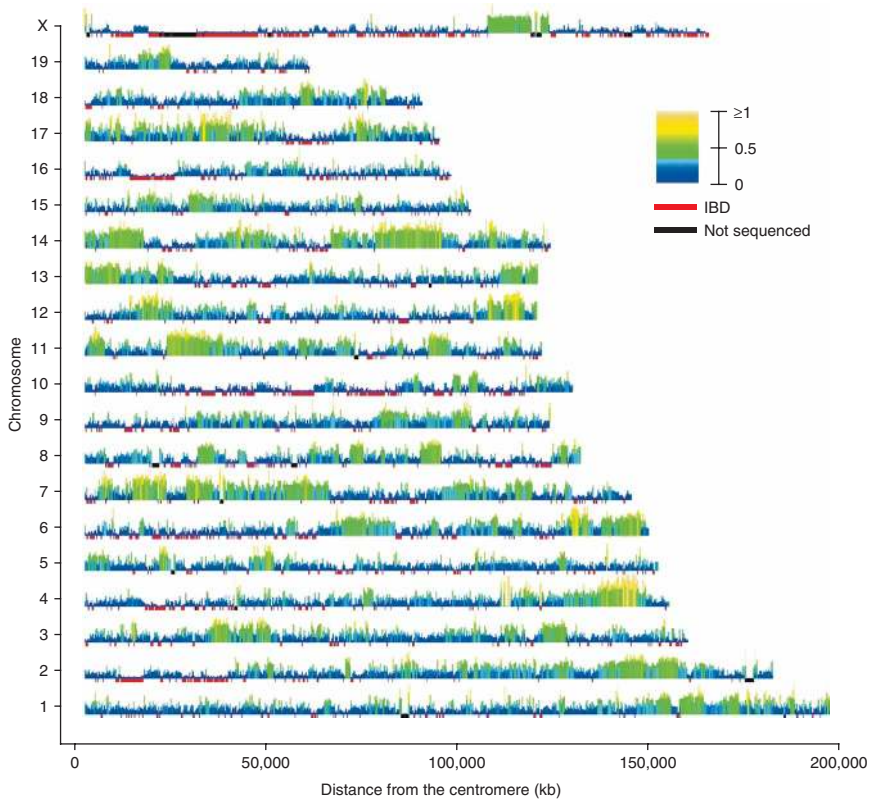

Figure 6 Frequency and spatial distributions of the mean normalized genetic variation observed among 11 resequenced strains. Spatial distribution of the mean normalized variation in the 11 resequenced classical strains is shown as vertical bars of different color and height for each 100-kb interval. IBD, identical by descent.

distribution that is centered on the average level of variation between subspecies (white bars in **Fig. 5**). On the other hand, more than 99.5% of intervals in which both strains were thought to have the same subspecific origin based on tree topology have normalized variation that is less than one, which would be expected if the subspecific assignments were correct.

The distribution of variation between pairs of inbred strains seems to be a composite of three distinct but overlapping distributions (**Fig. 5**). Two of these distributions contain the intrasubspecific variation, and the third encompasses the intersubspecific variation. This third distribution, as expected, has the highest level of variation. Direct resequencing indicates that, on average, there is one SNP every 151 bp in intersubspecific comparisons between wild-derived strains, which is similar to previous reports[18]. The more prominent of the two intrasubspecific components has variation that is less than 2% of the intersubspecific variation. These regions are expected to have, on average, one SNP every 20 kb, and they probably represent inherited regions that were identical by descent (IBD) within the recent derivation of the classical strains. The fraction of the genome in IBD regions ranges between 36% and 62% in 55 pairwise comparisons among the 11 NIEHS classical strains. These IBD regions represent blind spots in genetic studies that use crosses between classical strains. The remaining intrasubspecific variation encompasses 50% of the genome and has a broad distribution (**Fig. 5**). The distribution peaks at one-third of the typical level of intersubspecific variation. We propose that this variation is representative of the natural intrasubspecific variation found in every *M. musculus* subspecies.

We also determined the distribution of variation in comparisons between the three reference strains and MOLF and between the reference strains and each one of the classical strains independently (**Supplementary Fig. 6** online). The intrasubspecific variation in comparisons involving MOLF has a similar range and mode as the distribution between classical strains, which indicates that this feature is neither restricted to these strains nor restricted to one subspecies. The absence of a peak that corresponds to variation near zero in comparisons involving MOLF indicates that this strain shares little or no IBD regions with the three reference strains. On the other hand, the classical strains do share regions of IBD with WSB, although to a lesser extent than observed within classical strains. These results demonstrate that there is substantial intrasubspecific variation in the *M. m. domesticus* and *M. m. musculus* subspecies, and that the classical strains have captured a fraction of that variation.

In addition to pairwise analyses, we have determined the subspecific origin and variation level in comparisons that include all 11 classical strains. Our analyses indicate that, for approximately two-thirds of the fraction of the genome in which subspecific origin was assigned, all classical strains are derived from a single subspecies, primarily *M. m. domesticus* (**Supplementary Fig. 6**). Most of the remaining one-third has two subspecific origins, with a predominant contribution from the *M. m. domesticus* lineage. In those genomic regions with two subspecific origins, the minor subspecies is represented on average in 2 out of the 12 classical strains. These conclusions also hold true for the 28% of the genome for which we were not able to assign subspecific origins in the classical strains (**Supplementary Fig. 6**).

We determined the mean of the normalized variation among the 11 classical strains in every 100-kb interval (**Supplementary Fig. 6**). This analysis revealed that 11% of the genome is IBD when all strains are considered together. The level of identity is remarkable given that these strains include Castle, Swiss and Asian-derived strains[3]. This finding reinforces the conclusion that there is a very limited pool of founders and raises questions about whether, in addition to drift, selection for desirable traits in 'fancy' mice was involved in establishing the IBD regions. The addition of the C57BL/6 strain does not substantially reduce the fraction of the genome within IBD regions.

The spatial distribution of variation (**Fig. 6**) reveals substantial heterogeneity at the chromosomal, regional and local levels. For example, chromosomes X, 10 and 16 have low variation in most intervals, whereas chromosomes 14, 17 and 11 have high variation in most intervals. The most striking cases of regional variation are the island of high variation found on chromosome X and the distal regions of chromosomes 4 and 12. IBD regions are also clustered, spanning megabase-long regions in some chromosomes.

Assigning a subspecific origin in any given strain remains constant for extended regions, and consecutive trees with similar topologies are the norm. However, there are frequent minor changes in the topology of trees across consecutive 100-kb intervals that are due to historical

recombination events between haplotypes from within the same subspecies. High historical recombination rates should be beneficial for mapping complex traits. However, we also found that the most common strain distribution patterns in classical strains, representing 99.4% of all complete SNPs, are found on average in almost half of mouse chromosomes. This indicates that false positives will be a formidable obstacle in association mapping studies.

DISCUSSION

In summary, our analyses of the NIEHS data show that the genomes of classical inbred strains are largely derived from the *M. m. domesticus* subspecies. The distribution of genetic variation within the classical strains is nonrandomly distributed. Large regions of the genome are essentially IBD, whereas in other regions the level of diversity approaches that found in intersubspecific comparisons. More than half of the genome of the classical strains shows intermediate variation that is consistent with an intrasubspecific origin. We also found unexpected and frequent intersubspecific introgressions in the wild-derived strains. These features, and the limited amount of diversity that segregates among the classical strains (26% of the estimated total variation in the NIEHS data set), argue for the development of new mouse inbred lines that harbor greater allelic diversity and more complete randomization of ancestry. In particular, our results support the use of larger, heterogeneous populations[42] and the Collaborative Cross[43], a large panel of recombinant lines that randomizes the natural variation in inbred strains from the three main mouse subspecies.

METHODS

PCR-directed resequencing. DNA was obtained from The Jackson Laboratory, with the exception of CIM/Pas, which was a gift. To determine the FPR and FNR, we resequenced 70 fragments located in 14 chromosomes and spanning 28 kb (**Supplementary Table 1** online). The position of each base pair in these fragments was tiled in the Perlegen arrays and was completely sequenced in this study in the NIEHS strains and C57BL/6J. In the intersubspecific introgression studies, we resequenced 15 fragments located in ten chromosomes and spanning almost 12.5 kb (**Supplementary Table 1**) in the following strains: *M. m. castaneus*: CAST/EiJ, CASA/RkJ and CIM/Pas; *M. m. musculus*: CZECHII/EiJ, PWK/PhJ and PWD/PhJ; and *M. m. domesticus*: WSB/EiJ, PERC/EiJ and TIRANO/EiJ. Amplification and purification of PCR products were carried out as previously described[18]. Sequencing was carried out at the Automated DNA Sequencing Facility, University of North Carolina at Chapel Hill, on an ABI Prism 3730 (Applied Biosystems). All sequences were initially aligned using the Sequencher (GeneCodes) software. Aligned sequences were trimmed to retain only high-quality sequences. We determined the genomic positions of each SNP in Build 36 of Ensembl, and the region of complete overlap between our sequences and the NIEHS sequences. Overlapping regions were then compared, and shared and non-shared SNPs were identified. We considered the SNPs to be shared if both data sets had a SNP at the same position, with the same alternative alleles and the same strain distribution pattern.

Frequency of diagnostic SNPs using the raw data. We defined a SNP to be diagnostic if the genotypes in the three reference strains (CAST/EiJ, PWD/PhJ and WSB/EiJ) were complete and the SNP was polymorphic among these strains. There are three types of diagnostic SNP that correspond to the three strain distribution patterns among the reference strains. The number of diagnostic SNPs was determined in every 100-kb interval and their proportions were mapped to a simplex.

False-negative rate. Based on a comparison of the NIEHS SNPs with our resequencing data, we determined the FNR for the different classes of MAFs (**Fig. 2b**). Regression of a log-transformed MAF on the FNR provides a robust smoothed estimate of the MAF-specific FNR. The estimated FNRs that correspond to SNPs with the minor allele shared by 1 to 7 strains were 0.76, 0.64, 0.57, 0.52, 0.48, 0.45 and 0.42, respectively. The average genome-wide FNR was computed as a weighted average of the FNRs across the MAF classes. To estimate the proportion of SNPs that are variable within the classical strains, we calculated the proportion that are variable among classical strains within each MAF class, applied bias correction (see Branch-length correction) to each MAF class and calculated the weighted average.

Branch-length correction. The FNRs of SNPs that have different MAFs were used to correct the estimated branch lengths in the phylogenetic trees. The corrected length is proportional to the expected total number of SNPs (observed plus unobserved). To obtain this correction, we multiplied the estimated length of each branch by the factor 1/(1 – FNR), corresponding to the MAF of that branch. Terminal branches of the tree, with lower MAFs and correspondingly higher FNRs, expand more than the inner branches, which have higher MAFs and lower FNRs.

Phylogenetic analyses. Phylogenetic analyses were carried out using the PHYLIP version 3.6 phylogeny inference software package (Felsenstein, J., Department of Genome Sciences, University of Washington, Seattle, 2005). A tree was generated for each 100-kb interval using the SNP genotypes for the 15 NIEHS strains. Branch lengths were corrected as described above. We used Dnapars (DNA parsimony algorithm version 3.6) with default options as described online (see URLs section below), although using the search option to be 'Rearrange on one best tree'. Although we were aware that this search option is less thorough, our pilot study showed that, in most cases, trees found from searches that rearrange from all the possible most parsimonious trees were similar except some discordant results for the terminal branches. We determined the robustness of the tree by bootstrap analysis (Seqboot software (100 replicates)) using the Consense (majority rule) software program. The mitochondrial and Y-chromosomal analyses were carried out with the genotypes at 286 and 4935 SNPs, respectively. Similar results were obtained using both distance (neighbor joining) and maximum likelihood (Dnaml) approaches.

Frequency of diagnostic SNPs using corrected data. Using the corrected trees, we determined the distance from the common node to each of the three reference strains, WSB/EiJ, PWD/PhJ and CAST/EiJ, in each 100-kb interval. These distances were transformed into fractions representing the local contribution of diagnostic SNPs and were represented in the simplex as described above.

Discrimination between introgressed and balanced regions. The simplex was divided into five regions (**Supplementary Fig. 6**). Three regions located at the vertices of the triangles contain the three possible types of unbalanced interval. In these regions, the ratio between the length of the longest and shortest branch for the three reference strains in the corrected tree is $>4{:}1$. Geometrically, this corresponds to the inside of three circles centered at the vertices of the simplex, whose radius is 1/sqrt(12) (inside circle in **Supplementary Fig. 6**). Intervals were classified as potential intersubspecific introgression after running a HMM to fill isolated balanced intervals within large blocks of unbalanced intervals and to remove isolated unbalanced intervals. The HMM has four hidden states that correspond to three types of introgression pattern and a fourth balanced state. Here the 'true' introgression status of each 100-kb interval is considered a hidden state. The 'output' of the HMM is an indicator of which region of the simplex an interval was assigned. The HMM parameters were set to revisit the same state with a probability of 0.99 and to tolerate 1% of intervals that are inconsistent with the 'true' state. The HMM inference algorithm has been described previously[44]. We considered an interval balanced if it was not found to be in an introgression region by the HMM and if the ratio of the longest versus the shortest branch length for the three reference strains in the corrected phylogenetic tree was $<3{:}1$ (central circle in **Supplementary Fig. 6**). Unbalanced intervals excluded from the putative introgression regions by the HMM and intervals located in the periphery between the unbalanced and balanced regions of the simplex were considered undetermined regions.

Normalized variation in pairwise comparisons. In pairwise comparisons, we used the distance between a pair of strains in the corrected phylogenetic tree as an estimate of genetic variation. In each interval, we estimated the variation among the three reference strains. The normalized variation is the ratio between the distance separating a given pair of inbred strains and the average distance among all pairs of strains from different subspecific origins. For balanced

regions of the genome, the intersubspecific average includes seven pairwise comparisons (the variation between the three pairs of reference strains and the variation between the four possible combinations of the three reference strains and the two classical strains from different subspecies). For unbalanced regions of the genome, the intersubspecific average was determined using the two pairs of reference strains that are from distinct subspecies.

Subspecific origin in unbalanced intervals. For unbalanced regions, it was not possible to assign an ancestral subspecific origin to segments of each classical strain. We inferred the number of ancestral subspecies present among all classical strains in each interval, using the observed distribution of intersubspecific and intrasubspecific variation in the balanced regions of the genome (**Fig. 5**). Specifically, we calculated the ratio between each pair of classical inbred strains and the mean distance between the two pairs of wild-derived strains that have no evidence for introgression. We considered that a pair of classical strains belonged to the same subspecies if the ratio was <0.73 or had haplotypes derived from different subspecies if the ratio was >0.73. This threshold was derived from the distributions of the mean intrasubspecifc versus intersubspecific variation observed in classical strains (**Fig. 5**).

URLs. NIEHS Mouse Genome Resequencing and SNP Discovery Project: http://www.niehs.nih.gov/crg/cprc.htm; NIEHS/Perlegen mouse SNP and genotype data: http://mouse.perlegen.com/mouse/download.html; NIEHS/Perlegen strain selection criteria: http://mouse.perlegen.com/mouse/strain_selection.html. Jackson Laboratory: http://www.jax.org. Default options for Dnapars: http://evolution.genetics.washington.edu/phylip/doc/dnapars.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank S. Ahmed for technical assistance; K. Paigen, K. Broman and B. Payseur for helpful comments during the preparation of the manuscript; J. Felsenstein for advice and L. Wu for assistance with the phylogenetic tree computations and A. Smith for developing a genome browser format for displaying phylogenetic trees. CIM/Pas was provided by F. Bonhomme (University Mont Pellier II). We thank the NIEHS and Perlegen for making the SNP data set freely available prior to publication. This work was supported by the US National Institute of General Medical Sciences as part of the Center of Excellence in Systems Biology (1P50 GM076468).

AUTHOR CONTRIBUTIONS

This study was designed by F.P.-M.V. and G.A.C. The genome-wide analyses were carried out by H.Y. The sequence data used to determine the false-negative and false-positive rates and to confirm the presence and direction of intrasubspecific introgression were generated by T.A.B.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Paigen, K. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). *Genetics* **163**, 1–7 (2003).
2. Paigen, K. One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). *Genetics* **163**, 1227–1235 (2003).
3. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nat. Genet.* **24**, 23–25 (2000).
4. Ferris, S.D., Sage, R.D. & Wilson, A.C. Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. *Nature* **295**, 163–165 (1982).
5. Bishop, C.E., Boursot, P., Baron, B., Bonhomme, F. & Hatat, D. Most classical *Mus musculus domesticus* laboratory mouse strains carry a *Mus musculus musculus* Y chromosome. *Nature* **315**, 70–72 (1985).
6. Nagamine, C.M. *et al.* The musculus-type Y chromosome of the laboratory mouse is of Asian origin. *Mamm. Genome* **3**, 84–91 (1992).
7. Bonhomme, F., Guenet, J.-L., Dod, B., Moriwaki, K. & Bulfield, G. The polyphyletic of the laboratory inbred mice and their rate of evolution. *J. Linn. Soc.* **30**, 51–58 (1987).
8. Wade, C.M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574–578 (2002).
9. Wiltshire, T. *et al.* Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci. USA* **100**, 3380–3385 (2003).
10. Pletcher, M.T. *et al.* Use of a dense single nucleotide polymorphism map for *in silico* mapping in the mouse. *PLoS Biol. [online]* **2**, e393 (2004) (doi:10.1371/journal.pbio.0020393).
11. Frazer, K.A. *et al.* Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Res.* **14**, 1493–1500 (2004).
12. Petkov, P.M. *et al.* Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet. [online]* **1**, e33 (2005) (doi:10.1371/journal.pgen.0010033).
13. Yalcin, B. *et al.* Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci. USA* **101**, 9734–9739 (2004).
14. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).
15. Harr, B. Regions of high differentiation–worth a check. *Genome Res.* **16**, 1193–1194 (2006).
16. Boursot, P. & Belkhir, K. Mouse SNPs for evolutionary biology: beware of ascertainment biases. *Genome Res.* **16**, 1191–1192 (2006).
17. Zhang, J. *et al.* A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. *Genome Res.* **15**, 241–249 (2005).
18. Ideraabdullah, F.Y. *et al.* Genetic and haplotype diversity among wild derived mouse inbred strains. *Genome Res.* **14**, 1880–1887 (2004).
19. Boissinot, S. & Boursot, P. Discordant phylogeographic patterns between the Y chromosome and mitochondrial DNA in the house mouse: selection on the Y chromosome? *Genetics* **146**, 1019–1034 (1997).
20. Bonhomme, F. *et al.* Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among House Mouse subspecies. *Genome Biol. [online]* **8**, R80 (2007) (doi:10.1186/gb-2007-8-5-r80).
21. Frazer, K.A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature,* advance online publication 29 July 2007 (doi:10.1038/nature06067).
22. *Genetic Variants and Strains of the Laboratory Mouse* (Lyon, M.F., Rastan, S. & Brown, S.D.M., eds.) 3rd edn. (Oxford University Press, Oxford 1996).
23. Yonekawa, H. *et al.* Hybrid origin of Japanese mice "*Mus musculus molossinus*": evidence from restriction analysis of mitochondrial DNA. *Mol. Biol. Evol.* **5**, 63–78 (1988).
24. Sakai, T. *et al.* Origins of mouse inbred strains deduced from whole-genome scanning by polymorphic microsatellite loci. *Mamm. Genome* **16**, 11–19 (2005).
25. Abe, K. *et al.* Contribution of Asian mouse subspecies *Mus musculus molossinus* to genomic constitution of strain C57BL/6J, as defined by BAC-end sequence-SNP analysis. *Genome Res.* **14**, 2439–2447 (2004).
26. Wade, C.M. & Daly, M.J. Genetic variation in laboratory mice. *Nat. Genet.* **37**, 1175–1180 (2005).
27. Auffray, J.-C., Vanlerberghe, F. & Britton-Davidian, J. The house mouse progression in Eurasia: a palaeontological and archaeozoological approach. *Biol. J. Linn. Soc.* **41**, 13–25 (1990).
28. Din, W. *et al.* Origin and radiation of the house mouse: clues from nuclear genes. *J. Evol. Biol.* **9**, 519–539 (1996).
29. Prager, E.M., Orrego, C. & Sage, R.D. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150**, 835–861 (1998).
30. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
31. Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, 282–288 (2005).
32. Forejt, J. Hybrid sterility in the mouse. *Trends Genet.* **12**, 412–417 (1996).
33. Thrachtulec, Z. *et al.* Positional cloning of the hybrid sterility 1 gene: fine genetic mapping and evaluation of two candidate genes. *Biol. J. Linn. Soc.* **84**, 637–641 (2005).
34. Payseur, B.A. & Hoekstra, H.E. Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of mice. *Genetics* **171**, 1905–1916 (2005).
35. Oka, A. *et al.* Disruption of genetic interaction between two autosomal regions and the x chromosome causes reproductive isolation between mouse strains derived from different subspecies. *Genetics* **175**, 185–197 (2007).
36. Dai, J. *et al.* The absence of mitochondrial DNA diversity among common laboratory inbred mouse strains. *J. Exp. Biol.* **208**, 4445–4450 (2005).
37. Tucker, P.K., Lee, B.K., Lundrigan, B.L. & Eicher, E.M. Geographic origin of the Y chromosomes in "old" inbred strains of mice. *Mamm. Genome* **3**, 254–261 (1992).
38. Ferris, S.D. *et al.* Flow of mitochondrial DNA across a species boundary. *Proc. Natl. Acad. Sci. USA* **80**, 2290–2294 (1983).
39. Li, J. *et al.* Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36**, 952–954 (2004).
40. Snijders, A.M. *et al.* Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res.* **15**, 302–311 (2005).
41. Graubert, T.A. *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet. [online]* **3**, e3 (2007) (doi:10.1371/journal.pgen.0030003).
42. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38**, 879–887 (2006).
43. Churchill, G.A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36**, 1133–1137 (2004).
44. Churchill, G.A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94 (1989).