Genetic and Haplotype Diversity Among Wild-Derived Mouse Inbred Strains

Folami Y. Ideraabdullah,[1,2,7] Elena de la Casa-Esperón,[5,7] Timothy A. Bell,[1,7] David A. Detwiler,[1] Terry Magnuson,[1,2,3,4] Carmen Sapienza,[5,6] and Fernando Pardo-Manuel de Villena[1,2,3,4,8]

[1]*Department of Genetics,* [2]*Curriculum in Genetics and Molecular Biology,* [3]*Lineberger Comprehensive Cancer Center,* [4]*Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, North Carolina 27599, USA;* [5]*Fels Institute for Cancer Research and Molecular Biology,* [6]*Department of Pathology and Laboratory Medicine, Temple University School of Medicine, Philadelphia, Pennsylvania 19140, USA*

With the completion of the mouse genome sequence, it is possible to define the amount, type, and organization of the genetic variation in this species. Recent reports have provided an overview of the structure of genetic variation among classical laboratory mice. On the other hand, little is known about the structure of genetic variation among wild-derived strains with the exception of the presence of higher levels of diversity. We have estimated the sequence diversity due to substitutions and insertions/deletions among 20 inbred strains of *Mus musculus*, chosen to enable interpretation of the molecular variation within a clear evolutionary framework. Here, we show that the level of sequence diversity present among these strains is one to two orders of magnitude higher than the level of sequence diversity observed in the human population, and only a minor fraction of the sequence differences observed is found among classical laboratory strains. Our analyses also demonstrate that deletions are significantly more frequent than insertions. We estimate that 50% of the total variation identified in *M. musculus* may be recovered in intrasubspecific crosses. Alleles at variants positions can be classified into 164 strain distribution patterns, a number exceeding those reported and predicted in panels of classical inbred strains. The number of strains, the analysis of multiple loci scattered across the genome, and the mosaic nature of the genome in hybrid and classical strains contribute to the observed diversity of strain distribution patterns. However, phylogenetic analyses demonstrate that ancient polymorphisms that segregate across species and subspecies play a major role in the generation of strain distribution patterns.

[Supplemental material is available online at www.genome.org.]

The mouse is a model organism widely used in genetic research and for which the genome sequence has been assembled. The fact that complete inbreeding is tolerated has allowed the establishment of hundreds of inbred strains (Beck et al. 2000). Mouse inbred strains may be divided into two groups, classical and wild-derived. The genome of classical inbred strains derives from a handful of progenitors (Ferris et al. 1982; Tucker et al. 1992; Beck et al. 2000) and represents a mosaic with unequal contributions of several *Mus musculus* subspecies (Bonhomme et al. 1987). The genome sequence of the classical inbred mouse strain C57BL/6J has been reported (Waterston et al. 2002) and sequence data are also available for three additional classical inbred strains at different coverage levels (www.celera.com). Recent reports have provided the first analyses of genome-wide patterns of genetic variation in inbred mouse strains (Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003, Yalcin et al. 2004). These studies indicate that classical inbred strains have limited levels of genetic variation when compared with humans, and that the pattern of genetic variation in the genome of classical inbred strains has a mosaic structure, with regions of low levels of polymorphism and regions of high levels of polymorphism (Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003, Yalcin et al. 2004). Although the fraction of the genome that belongs to each category and the size of these regions remain under discussion, these observations have major implications for mapping quantitative trait loci (Vogel 2003; Wiltshire et al. 2003; Yalcin et al. 2004). Accurate understanding of the haplotype structure of mouse inbred strains may provide powerful approaches in the identification of molecular variants underlying quantitative trait loci (Beck et al. 2000; Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003, Yalcin et al. 2004). One approach relies on associating phenotypic variation in inbred strains with their strain distribution patterns (SDP, the patterns of allelic similarities and differences among strains at a locus; Grupe et al. 2001; Yalcin et al. 2004). Because variants with different alleles may have the same SDP, the use of the latter simplifies the analysis of diallelic variation and is a staple in molecular phylogenetic studies. Therefore, the number, frequency, and spatial distribution of SDPs are critical parameters to define structure of sequence variation. A recent study has shown that only 13 SDPs account for almost 99% of 1465 variants identified in eight classical inbred strains over a 4.8-Mb region of mouse chromosome 1 (Yalcin et al. 2004). Importantly, the authors report that, despite the small number of SDPs observed, the haplotypes of inbred strains are complex because variants with the same SDPs are clustered together, but they do not generally occur in simple blocks (Yalcin et al. 2004). This study concludes that only a limited number of SDPs (on the order of the number of strains analyzed) will be present in regions of similar sizes. However, it also

[7]These authors contributed equally to this work.
[8]Corresponding author.
E-MAIL fernando@med.unc.edu; FAX (919) 966-3630.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2519704.

14:1880–1887 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

notes that it is not known how the number of SDPs across the genome may vary depending on the number of strains analyzed (Yalcin et al. 2004).

A second group of laboratory mice consists of several dozen inbred strains that have been derived from wild mice trapped, at different times and locations, from different populations. These include inbred strains established from mice belonging to several *Mus* species and inbred strains derived from different *Mus musculus* subspecies and their intersubspecific hybrids (Bonhomme and Guenet 1996). The phylogenetic relationships and the divergence times among these species and subspecies have been established on the basis of DNA hybridization and sequence comparison studies (She et al. 1990; Schalkwyk et al. 1999; Chevret et al. 2002; Lundrigan et al. 2002). Hybrids generated by crossing wild-derived and classical strains have high levels of genetic diversity. This feature has been used to generate the high-resolution linkage map of the mouse, in the study of genome imprinting, X-inactivation, and complex traits (for review, see Guenet and Bonhomme 2003). Although the usefulness of wild-derived strains stems from the high level of diversity, the structure and patterns of sequence variation in these strains are not well characterized. Previous studies analyzed a total of six wild-derived strains and concluded that the levels of variation observed between wild-derived and classical strains are significantly higher than among classical strains, and that the variation is distributed uniformly across the genome (Vogel 2003; Wiltshire et al. 2003; Yalcin et al. 2004). However, the level of intrasubspecific variation could not be determined; wild-derived strains were not included in the SDP analysis, and a maximum of four wild-derived strains was analyzed in each study.

We have determined the level and patterns of genetic diversity, including insertions/deletions, among 20 *M. musculus* inbred strains, with an emphasis on wild-derived strains, to provide a broader view of the genetic variation available in mouse inbred strains and to characterize the phylogenetic history of that variation. Inbred strains were chosen to include at least two representatives of each of four *M. musculus* subspecies. Our panel also includes individual strains of the closely related species *M. spretus* and *M. spicilegus* (Guenet and Bonhomme 2003).

RESULTS

Frequency and Distribution of Sequence Variants

We have determined the DNA sequence of 62 genomic segments located on 14 chromosomes. Those segments include a single exon (both coding and UTRs) of 19 genes, seven fragments spanning 6.5 kb of the *Il9r* gene (≈66% of the entire gene) and 36 fragments spanning ~650 kb of the *Cctb6-Ap2b1* region of chromosome 11 (see Supplemental Fig.1). Genes were selected on the basis of the following criteria: (1) candidate genes for schizophrenia and hypertension (*Agtr1a*, *Bdkrb2*, *Comt*, *Dao1*, *Diap1*, *Ncam1*, *Pparg*, and *Prodh*); (2) genes involved in DNA repair and cancer (*Mlh1* and *Pms2*); (3) imprinted genes (*Igf2*, *Grb10*, and *Ube3a*); and (4) map location, to provide a wide overview of different genomic regions (*Clasp1*, *Rgs4*, *Dutp*, *Mecp2*, *Npr1*, and *Tgm1*). We amplified exons to increase the probability of successful amplification of specific products in all strains analyzed. Finally, multiple fragments in the *Il9r* and *Cctb6-Ap2b1* regions were sequenced in order to characterize the spatial distribution of SDPs and the extent of haplotype sharing in wild-derived strains.

In total, 26,116 bp were sequenced from each strain. When all 22 strains are considered, 1007 sequence variants were identified, divided as follows: 89 microsatellite variants, 83 insertion/deletion variants, and 835 substitution variants (see Supplemental Table 1). Multiple alleles are found at most microsatellites, therefore, we have omitted microsatellites from all subsequent analyses presented in this study. Insertion/deletion and substitution variants are largely diallelic (98.7 ± 0.4%). Table 1 shows the classification of variants by type and sequence context among inbred strains grouped according to their phylogenetic relationships. As expected, the estimated rates of variants per kilobase in this panel are substantially higher than previously reported in classical inbred strains (Yalcin et al. 2004). Our estimates are also higher than the rates of variation observed in studies using fewer wild-derived strains (Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004). Overall, the highest density of variants is found in intron/intergenic regions and the lowest in coding sequences (Table 1). Insertions/deletions display the most pronounced difference in the rate of variation. The reversal

Table 1. Classification of Variants by Type, Context, and Phylogenetic Relationship

			Insertions/deletions			Substitutions		
Category	Context	bp	# of variants	Rate/kb	SE	# of variants	Rate/kb	SE
All *Mus* strains	Intron/Intergenic	13154	57	4.33	0.57	487	37.02	1.65
	UTR	6499	25	3.85	0.77	195	30.00	2.12
	Coding	6463	1	0.15	0.15	153	23.67	1.89
M. musculus	Intron/Intergenic	13154	35	2.66	0.45	314	23.87	1.33
	UTR	6499	11	1.69	0.51	117	18.00	1.65
	Coding	6463	0	0	0	105	16.25	1.57
M. m. domesticus	Intron/Intergenic	13154	17	1.29	0.31	162	12.32	0.96
	UTR	6499	6	0.92	0.38	42	6.46	0.99
	Coding	6463	0	0	0	50	7.74	1.09
Classical	Intron/Intergenic	13154	13	0.99	0.27	122	9.27	0.84
	UTR	6499	2	0.31	0.22	28	4.31	0.81
	Coding	6463	0	0	0	41	6.34	0.99

Analyses were performed among strains classified into the following four categories: (1) all inbred strains analyzed, (2) all *M. musculus* strains, (3) all wild-derived *M. m. domesticus* strains, and (4) classical strains. The table shows the length of high-quality sequence, the number of variants per kilobase, and the SE of the rate. Variants were divided into insertions/deletions and substitutions. Discrimination between microsatellites and insertions/deletions was performed on the basis of the presence/absence of more than four adjacent tandem repeats of the inserted/deleted sequence, respectively. This threshold was determined empirically by comparing the frequency of mononucleotide runs of different sizes in the sequenced region and the frequency of insertions/deletions observed in runs of each size. For runs of less than five nucleotides, the probability of observing insertions/deletions is roughly equal to their frequency. For runs of five or more nucleotides, the probability of an insertion/deletion is 10 to a 100 times greater than expected.

in the density of substitutions between coding sequences and UTRs in *M. m. domesticus* and classical inbred strains (Table 1) most likely reflects limited sampling of regions that may have different phylogenetic histories.

Classification of 918 variants according to the species in which the minor allele is present demonstrates that 35%–45% of the total variation identified in our panel would be observed in each one of the three possible types of interspecific crosses between *M. musculus*, *M. spretus*, and *M. spicilegus* inbred strains (Table 2; Supplemental Table 2). On the other hand, 63.4 ± 1.6% (582/918 in Table 2) of the genetic diversity identified in our study can be obtained in crosses between subspecies of *M. musculus* with few or none of the hybrid infertility problems encountered in interspecific crosses (Forejt 1996). Because the panel includes only one *M. spretus* (SPRET/EiJ) and one *M. spicilegus* strain (PANCEVO/EiJ), our study does not provide any information about intraspecific variation in these two species. Interestingly, our analysis identifies a subset of *M. musculus* diallelic variants, in which the minor allele is present also in either *M. spretus* or *M. spicilegus*, but not in both (Table 2). This class represents a sizable fraction of the variants (8.2 ± 0.9%, 75/918 in Table 2) and raises questions regarding the common assumption that variants identified in pairwise comparisons arise by a mutation event in one of the two branches emerging from the last common ancestor. This observation highlights the importance of comparisons using multiple samples of each taxon and the usefulness of wild-derived strains to interpret molecular variation within a clear evolutionary framework (Guenet and Bonhomme 2003; see below).

We have limited our analysis of intrasubspecific variation to the *M. m. domesticus* subspecies because the other three subspecies were represented in our panel by only two strains each. We have estimated that a maximum of 58.4 ± 2.0% of the 582 variants identified in *M. musculus* (Table 1) would be present in three-way comparisons using a single strain from each of *M. m. castaneus*, *M. m. musculus*, and *M. m. domesticus* subspecies. This fraction increases to 69.2 ± 1.9% with the inclusion of a *M. m. molossinus* strain. Table 1 indicates that >47% of the total *M. musculus* variation is present in the *M. m. domesticus* subspecies (277/582 in Table 1). Therefore, >50% of the total variation identified in *M. musculus* may be recovered in intrasubspecific crosses. On the other hand, only one-third of the *M. musculus* variation is present in the six classical inbred strains analyzed here (Table 2). The variation observed in classical strains is slightly lower than the variation found in *M. m. domesticus* (Table 1), despite the fact that classical strains appear to have haplotypes derived from two different *M. musculus* subspecies in five of the 62 fragments analyzed.

Table 2. Classification of Variants According to the Distribution of the Minor Allele

Minor allele present in	# of variants	% of total variants ± SD	% of *M. musculus* variants ± SD
SPRET/EiJ	144	15.7 ± 1.2	n. a.
PANCEVO/EiJ	121	13.1 ± 1.1	n. a.
SPRET/EiJ and PANCEVO/EiJ	71	7.7 ± 0.9	n. a.
M. musculus and SPRET/EiJ	34	3.7 ± 0.6	5.8 ± 1.0
M. musculus and PANCEVO/EiJ	41	4.3 ± 0.7	7.0 ± 1.1
M. musculus	507	55.5 ± 1.6	87.1 ± 1.4
Wild derived	301	32.9 ± 1.6	51.8 ± 2.1
Wild derived and Classical	185	20.2 ± 1.3	31.8 ± 1.9
Classical	21	2.9 ± 1.3	3.6 ± 0.8

The Table shows 918 insertion/deletion and substitution variants identified in this study classified according to the distribution of the minor allele among three *Mus* species into six mutually exclusive classes. Variants found exclusively in *M. musculus* are further subdivided depending on whether the minor allele is found in wild-derived strains only, in classical strains only, or in both types of strains.
(n. a.) Not applicable.

When all 20 *M. musculus* strains are considered, the density of variants across the fragments analyzed follows the expectations under a random (Poisson) distribution, suggesting that there is a uniform distribution of variants across the genome (Supplemental Fig. 2). However, when comparisons are restricted to *M. m. domesticus* or classical inbred strains, there is an excess of fragments with no variants, as also observed by Yalcin et al. (2004).

Analysis of Insertions/Deletions

In total, we have identified 83 insertions/deletions representing ~9% of the total variants (excluding microsatellites). The density of insertions/deletions varies in different types of sequence. In introns/intergenic regions and UTRs, insertions/deletions represent 12% of the total variants, although they are very rare (<1%) in coding regions (Table 1). The contribution of insertions/deletions to the diversity present in the 231 possible pairwise comparisons between the 22 strains analyzed here follows a normal distribution centered on the mean (data not shown). The size of the insertions/deletions ranges from 1 to 70 bp, with an average of 5.4 bp (Supplemental Fig. 3). However, the distribution of insertions/deletions is strongly skewed toward smaller sizes. One basepair insertions/deletions represent >40% of the total variants, and 80% of them are shorter than 6 bp (Supplemental Fig. 3). We were able to classify 81 of these 83 insertion/deletion variants as either insertions or deletions on the basis of predicted ancestral allele (identified using the SDP and the phylogenetic tree, see below). Deletions are significantly more frequent than insertions (52 vs. 29, respectively; H_0: equal number of deletions and insertions, $\chi^2 = 6.53$, 1 df, $P < 0.05$), and this trend is consistently observed in the three species analyzed here.

Strain Distribution Patterns

We have identified 164 SDPs in the 569 diallelic variants present among 20 *M. musculus* strains (variants found only in interspecific comparisons and triallelic variants were excluded in this analysis). Both the number and frequency distribution of SDPs we observed differ from that observed previously (Yalcin et al. 2004). Whereas similar numbers of SDPs (19) and strains (eight) were noted by Yalcin et al. (2004), we observe nearly an order of magnitude more SDPs than strains. In addition, 13 of the most common SDPs (those with a frequency >1%) described by Yalcin et al. (2004) accounted for 99% of the total variation. In our analysis, the 26 most common variants (those with frequency >1%) represent only 57% of the total variation. Only 10 SDPs have frequencies >2%, and more than half of the SDPs are defined on the basis of a single variant (Fig. 1). The high number and low frequency of SDPs suggest that there is very limited haplotype sharing among the panel of *M. musculus* strains analyzed here.

If alleles at a locus are identical by descent, gene flow has not occurred between the different branches of the phylogenetic tree, and there was no polymorphism at the branch points, the maximum number of SDPs that are consistent with any given phylogeny is $2n - 3$, where n is the number of strains. Therefore, for 20 strains, the maximum number of SDPs is 37 and the 127 SDPs in excess of this number (Table 3) must be due to alleles that are identical by state (IBS) rather than identical by descent (IBD), the presence of ancient variants, and/or gene flow between the branches. The presence of gene flow is expected, given the fact that 10 of the 20 strains included in our panel are intersubspecific

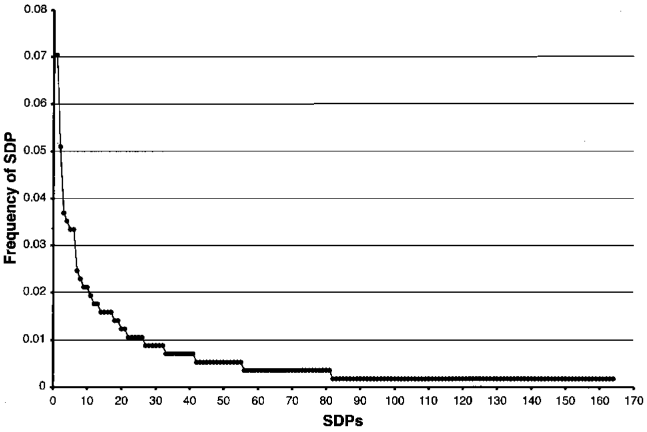


Figure 1 Frequency distribution of SDP. The frequency at which each one of 164 SDP is observed among *569 M. musculus* dialleic variants is shown in descending order.

hybrids and classical inbred strains (which also have a mixed phylogenetic history) (Bonhomme and Guenet 1996). Alleles at variant positions at different regions of the genome in these strains may originate from different subspecies. Therefore, exclusion of intersubspecific hybrids and classical strains should reduce the excess of SDPs, as should limiting the analysis to a single region of the genome. To fulfill these criteria, we compared the number SDPs present in both the total set of strains with the subset of 10 strains that are not known to be intersubspecific hybrids (CAST/EiJ; CASA/EiJ; CZECH1/EiJ; PWK/Ph; PERA/EiJ; PERC/EiJ; ZALENDE/EiJ; TIRANO/EiJ; LEWES/EiJ; and RBA/DnJ). This analysis was performed independently in the total data set, in the *Cct6b-Ap2b1* region (650 kb) and in the *Il9r* locus (6.5 kb). As expected, the number of excess SDPs in each group decreases significantly when only nonhybrid strains are analyzed (note that this decrease cannot be accounted for solely by the decrease in the number of variants; Table 3). Furthermore, the excess of SDPs decreases when the analysis is limited to smaller genomic regions. Therefore, the mosaic nature of the genome in hybrid strains does contribute significantly to SDP diversity. However, in all cases, there is still an excess of SDPs (Table 3), indicating that IBS and/or ancient polymorphisms are responsible. Inspection of SDPs in nonhybrid strains in each comparison demonstrates that alleles at 24%–43% of the SDPs segregate across *M. musculus* subspecies. In approximately half of these SDPs, variant alleles segregate simultaneously in two different subspecies, whereas in the other half, each one of the two alleles found among *M. m. domesticus* inbred strains is found in either *M. m. musculus* or *M. m. castaneus*, but not both. We conclude that the presence of a large fraction of variants that appear to segregate across subspecies contributes to the large number of SDPs found in our panel. The remaining excess is probably due to gene flow between subspecies rather than IBS (see below).

Table 3. **Variation in Number of SDP Observed at Different Genomic Regions and in Different Sets of Inbred Strains**

Genomic region	Strains (#)	Observed # of SDPs	Excess # of SDPs	# of SDPS with variants segregating across subspecies
Complete dataset	All (20)	164 (569)	127	n. a.
Complete dataset	Non hybrids (10)	60 (481)	43	26 (49)
Cct6b-Ap2b1	All (20)	83 (309)	46	n. a.
Cct6b-Ap2b1	Non hybrids (10)	44 (266)	27	18 (34)
Il9r	All (19)	41 (127)	6	n. a.
Il9r	Non hybrids (9)	17 (108)	2	4 (6)

Shown is the genomic region, the strains used in the analysis, the observed number of SDP, the excess number of SDP, and the number of SDP that segregate across *M. musculus* subspecies. Parentheses show the number of variants used to define the SDP. Excess number of SDP is the difference between the observed number of SDP and the maximum number of SDP that may be observed in a set of *n* strains if variants are identical by descent; there is no gene flow between branches, and all variation occurred after the divergence of the subspecies. (n. a.) Not applicable.

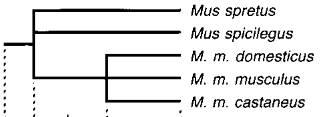
Phylogenetic History of the Genetic Variation Found in *Mus musculus*

Two of our previous analyses suggest that a considerable fraction of the total *M. musculus* variants segregate across species (Table 2) and across subspecies (Table 3). To formally address when the mutation events took place and which is the ancestral allele at each variant position, we determined the most parsimonious way to explain the SDPs that is also consistent with the true phylogeny (see Methods). The 569 diallelic variants, including insertions/deletions and substitutions, found among classical and wild-derived strains of *M. musculus* can be classified into three categories on the basis of whether the mutation event occurred before the divergence of the three specific lineages (a in Fig. 2), before the divergence of the *M. musculus* subspecies (b in Fig. 2) or after the divergence of the *M. musculus* subspecies (c in Fig. 2). Interestingly, 37.8% ± 2.0% of the total variants appear to predate the divergence of the *M. musculus* subspecies (a and b in Fig. 2). These ancient polymorphisms are distributed uniformly across the fragments sequenced. Importantly, on average, they represent 44.7% ± 6.7% of the sequence variants observed in pairwise comparisons between *M. musculus* strains (Fig. 3). In other words, the contribution of ancient polymorphisms to the sequence diversity is higher in pairwise comparisons than in the total data set, due to the higher frequency of the minor allele in this type of variant.

Because our ability to assign some *M. musculus* variants to the branch connecting the specific and subspecific divergence nodes (b in Fig. 2) depends on the correct identification of the ancestral allele, we tested whether our predictions are supported by the allele found at homologous positions in the rat. The predicted ancestral allele is supported by the rat sequence at 82.4% of variants at which rat has one of the two alleles found in *M. musculus*. On the other hand, we observe significant statistical evidence of an increase in the ratio of transitions to transversions at the remaining 17.6% of variants (H_0 = the ratio of transitions and transversions is independent on whether the predicted ancestral allele is consistent with the rat allele, 1 d.f., $\chi^2 = 4.74$; $P < 0.05$). This observation indicates that two substitution events (i.e., IBS), one in the *Mus* lineage and another in the rat lineage, are responsible for the inconsistencies. We conclude that errors in the determination of the ancestral allele are small and should not significantly affect our classification of variants segregating among *M. musculus* strains.

DISCUSSION

The level of genetic diversity reported here is higher than in previous reports (Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004). This holds true for the levels of genetic diversity observed in pairwise comparisons (Supplemental Table 2), the estimated rate of variants (Table 1)



	a	b	c	
# of variant events	75	140	354	
Total	215		354	
# of transitions	145		228	$\chi^2 = 0.74$
# of transversions	53		99	
# of insertions/deletions	17		27	

Figure 2 Phylogenetic history of *Mus musculus* variants. The figure shows a simplified phylogenetic tree of the *Mus* lineage (Guenet and Bohomme 2003). Insertion/deletion and substitution variants were assigned to one of three different branches of the tree (see Methods). Variants were further classified as transitions, transversions, and insertions/deletions. The χ-^{2} value was calculated on the basis of a test for independence between the origin of the variant and the type of substitution.

and the number and diversity in SDPs (Fig. 1). Regional differences in sequence variation, similar to those reported among genomic regions in comparisons between human and chimpanzee (Ebersberger et al. 2002), may be partly responsible for the differences. We suspect that sampling of a larger collection of strains with greater ancestral diversity account for most of the discrepancy between our results and those of previous reports (Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004).

To appreciate the full extent of the genetic diversity in *M. musculus*, it is instructive to compare our results with the sequence diversity observed among and between closely related mammalian species. Our estimate of the average frequency of SNPs in intersubspecific crosses in mouse (one SNP per 0.11 kb) is 10 times higher than in humans, and three times higher than in chimpanzees (one SNP per 1.08 kb and one SNP per 0.38 kb, respectively; Sachidanandam et al. 2001; Venter et al. 2001; Reich et al. 2003; Salisbury et al. 2003; Zietkiewicz et al. 2003). The estimate in chimpanzee has been corrected to account for the fact that the region analyzed is in the X chromosome and has low mutation rates (Kaessmann et al. 1999). In addition, the percentage of variant positions in the mouse genome, excluding microsatellites, (3.1%) is one order of magnitude higher than in humans (<1%) and between two and 30 times higher than in chimpanzees, bonobos, gorillas, and orangutans (1.7%, 0.1%, 0.8%, and 1.5%, respectively; Kaessmann et al. 2001). Although the existence of distinct subspecies contributes to the high level of diversity observed in the mouse, analyses of both chimpanzee and orangutan also include individuals from distinct subspecies (Kaessmann et al. 2001). Both the shorter generation time in mouse and the fact that a substantial fraction of the variation predates the divergence of the subspecies contribute significantly to the diversity. We conclude that the mouse is the mammalian species with the highest levels of genetic diversity yet described. In fact, the level of sequence diversity observed in *M. musculus* is more similar to that found between man and chimpanzee than between individual humans (Chen and Li 2001; Ebersbeger et al. 2002; Sakate et al. 2003).

Although our estimate for the sequence variants in *M. musculus* may appear high (Table 1), it represents, in all likelihood, an underestimate of the true value due to the limited sample size, the presence of sampling biases, and the type of variation detected. Mouse is a polytypic species with five to six recognized subspecies. Two models have been proposed to explain the origin and radiation of the commensal mice. The centrifugal model proposes that mice radiated outward from the central populations found on the Indian subcontinent (Auffray et al. 1990; Din et al. 1996). The linear model proposes that mice originated in West Eurasia and spread easterly to give rise to the progenitors of the different subspecies (Prager et al. 1998). Regardless of the model, representatives of the central populations in the centrifugal model (or representatives of *M. m. castaneus* in the linear model) harbor a significant fraction of the genetic variation of the whole species (Prager et al. 1998). The absence of strains from this group and from the *M. m. gentilulus* subspecies in our study would lead to underestimation of the true level of genetic diversity in the mouse. Furthermore, our analyses indicate that strain sampling in *M. m. castaneus* and *M. m. musculus* subspecies may itself be biased, because relatively little diversity has been captured among the strains of these two subspecies. Given the range of the distributions shown in other intersubspecific comparisons (Supplemental Table 2), this circumstance is most likely due to the limited geographical origins of these strains (www.jax.org) rather than an inherent lack of diversity within some subspecies. These data suggest that the fraction of variant positions that are present in mouse may be much higher than our estimate.

Here, we provide the first estimate of the level of variation among inbred strains of *M. musculus* subspecies (previous studies included only one wild-derived representative of four *M. musculus* subspecies and classical strains (Lindblad-Toh et al. 2000; Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004)). Our analyses of six wild-derived *M. m. domesticus* indicate that there is considerable variation within a subspecies (Table 1). In fact, in some fragments, the level of variation between *M. m. domesticus* inbred strains is similar to the average diversity found in intersubspecific pairwise comparisons (Supplemental Table 2). Although previous studies in classical inbred strains have equated the presence of segments with high frequency of polymorphisms with different subspecific origin (Wade et al. 2002; Wiltshire et al. 2003; Yalcin et al. 2004), the data presented here demonstrate that this is not a requirement to observe high levels of sequence variation (Supplemental Fig. 4). It also suggests that some wild-derived inbred strains descend from branches diverging early within subspecific lineages.

The analysis of SDPs in our panel of inbred strains complements the study of Yalcin et al. (2004), but also shows some striking contrasts, including the higher number and prevalence

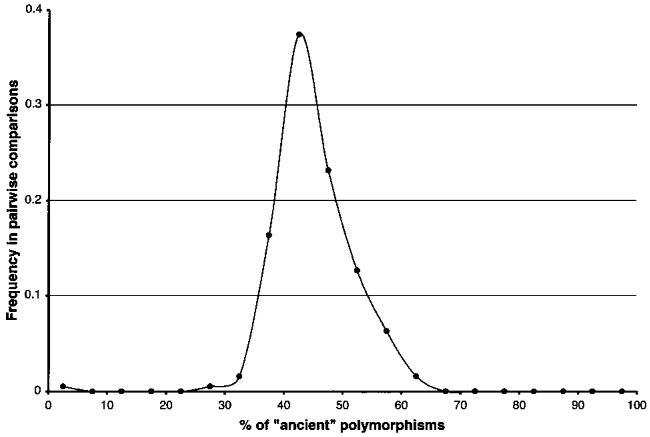


Figure 3 Frequency distribution of ancient polymorphisms in pairwise comparisons between *M. musculus* inbred strains. The percent of ancient polymorphisms was calculated in each of 190 pairwise comparisons between the 20 *M. musculus* inbred strains analyzed in our panel. Frequencies in pairwise comparisons were assigned to one of 20 equal percentile classes with respect to the percent of ancient polymorphisms.

of rare SDPs in our panel (Fig. 1). Although these studies differ in the number and type of strains and the location of the sequence analyzed, some useful comparisons can be made. When genome-wide SDP analysis in our panel is restricted to the six classical strains, the number and frequency distribution of SDPs are almost identical (data not shown) to those reported previously for eight strains in a 4.8-Mb region of chromosome 1 (Yalcin et al. 2004), suggesting that only a limited number of SDPs may be present in small sets of classical inbred strains. On the other hand, significant increases in SDP number may be achieved with the inclusion of wild-derived strains represented. Whether there is a predictable number of genome-wide SDPs in a panel remains unknown. However, it is evident that a very large number of SDPs may be found in panels of strains that include either natural or artificial hybrids. Furthermore, increase in SDP diversity depends not only on the number and type of strains, but also on the pervasive presence of ancient polymorphisms. This SDP diversity should provide higher mapping resolution but decrease the statistical power (Yalcin et al. 2004). Whether this trade off is suitable will depend on the type of genetic experiment to be conducted.

Our results confirm that the taxonomical classification of some wild-derived strains may need revision as sequence data accumulate. For example, although CALB/RkJ has been assigned to the *M. m. domesticus* subspecies (www.jax.org), it has haplotypes related to *M. m. castaneus* in at least five regions located in three different chromosomes, including the *Il9r* gene and the *Cct6b-Ap2b1* region. These results support the idea that CALB/RkJ is a hybrid of *M. m. domesticus* and *M. m. castaneus*. This raises the possibility that some inbred strains, derived from animals trapped in the periphery of the range of a subspecies or from recently colonized areas, may have a mosaic genome. This may be especially relevant in the case of the *M. m. domesticus* subspecies, as some of the more easily available and most commonly used wild-derived strains from this taxon are derived from mice trapped in the Americas (i.e., WSB/EiJ, PERA/EiJ, PERC/EiJ, LEWES/EiJ, etc.).

We also wish to note that ancient polymorphisms represent a substantial fraction of the total variants (Fig. 2). As stated in the previous section, three mechanisms may explain the presence of such polymorphisms in *M. musculus*. First, the variants may reflect gene flow among species and subspecies. Introgression of genetic variants between *M. spretus* and *M. musculus* has been reported (Greene-Till et al. 2000; Orth et al. 2002). However, the high frequency of such events in our data (13.2 ± 1.4%; Fig. 2), and the almost uniform distribution of such events across the regions analyzed (data not shown) does not support recent interspecific gene flow. In the same vein, introgression of genetic variants between subspecies is common in hybrid zones and in classical strains (Guenet and Bonhomme 2003), but our approach specifically excludes the use of strains derived from hybrid populations in the identification of ancient polymorphisms (Methods). Therefore, gene flow is unlikely to contribute significantly in explaining the level of the ancient polymorphisms reported here. On the other hand, these variants may be examples of polymorphisms that have been maintained through long evolutionary periods or, alternatively, IBS (i.e., the reoccurrence of the same mutation event in different lineages). The ratio of transitions to transversions may be used to discriminate between substitution variants IBD and IBS. In the case of IBS, the ratio of transitions to transversions should increase by twice the ratio of transitions to transversions in IBD. We found no statistical evidence for an excess of transitions among variants predating the divergence of the subspecies when compared with variants arising after it (Fig. 2).

Identification of ancient variants depends on the number and phylogenetic relationship of the samples analyzed. In this report, most ancient variants would have gone unrecognized without the inclusion of representatives of both *M. spretus* and *M. spicilegus* and of multiple inbred strains from each of three *M. musculus* subspecies. Our analysis relies on the use of the phylogenetic tree proposed by the centrifugal model (Auffray et al. 1990; Din et al. 1996). Some of our conclusions will require reevaluation if the linear model ultimately represents the true phylogeny of *M. musculus* subspecies. However, this reinforces the need for the characterization of genetic diversity in a wide and representative panel of wild-derived inbred strains. The persistence of ancient polymorphisms may help to explain how interfertile populations of the same species have maintained a greater degree of sequence diversity than that found between man and chimpanzee. Although the data presented here are based on inbred strains, our observations suggest that the effective population size in the *Mus* lineage has been relatively large and constant over a long evolutionary period. These variants contribute significantly to the genetic diversity present in *M. musculus*, but they may also, if unrecognized, affect the conclusions of evolutionary and haplotype studies.

Finally, our analysis of insertion/deletion variants provides an example of the value of wild-derived strains to interpret the molecular variation within a clear evolutionary framework (Guenet and Bonhomme 2003). The size of the mouse genome is significantly smaller than that of humans (Waterston et al. 2002). On the basis of genome-wide comparisons between the mouse and human genome, it has been proposed that the smaller size in the mouse is not simply the result of an increase in genome size in the human lineage (due to duplications and the addition of repetitive elements), but to the loss of ancestral sequences in the mouse lineage (Waterston et al. 2002). Our analysis of just over 80 insertion/deletion variants confirms this conclusion and suggests that it is an ongoing process in three different *Mus* species. Further studies are required to determine the relative contribution of small deletions, such as those reported here, to the overall decrease in size in the mouse genome.

METHODS

Mouse Inbred Strains

Mice or DNA samples from all strains used in this study were obtained from Jackson Laboratory, with the exception of JF1/Ms and DDK/Pas, which are maintained by Terry Magnuson and Fernando Pardo-Manuel de Villena, respectively. All mice described in this report were treated according to the IACUC of the University of North Carolina at Chapel Hill. DNA was prepared from small tissue biopsies using standard procedures. For all inbred strains, we use the taxonomic classification provided by Jackson Laboratory as follows: *M. spretus*, SPRET/EiJ; *M. spicilegus*, PANCEVO/EiJ; *M. m. castaneus*, CAST/EiJ, CASA/EiJ; *M. m. castaneus x M. m. domesticus*, CALB/RkJ (see Discussion); *M. m. molossinus*, JF1/Ms and MOLC/RkJ; *M. m. musculus*, CZECH1/EiJ and PWK/Ph; *M. m. musculus x M. m. domesticus* hybrid, SKIVE/EiJ; *M. m. domesticus*, PERA/EiJ, PERC/EiJ, ZALENDE/EiJ, TIRANO/EiJ, LEWES/EiJ, and RBA/DnJ; *M. musculus*, DDK/Pas, BALB/cJ, C57BL/6J, DBA/2J, A/J, and 129X1/SvJ.

Primer Design, PCR, and Sequencing

Cct6b-Ap2b1 Region

We selected 39 candidate segments for analysis, which were approximately evenly spaced over the 650-kb genomic region (www.ensembl.org). Sequences were selected to avoid duplicated and repeated regions in the primers, detected by BLAST and RepeatMasker software. At least one SNP was known to be present in the Celera database (www.celera.com) in 24 fragments. Primers were designed to amplify an average of 450 bp using PrimerQuest (www.idtdna.com, BioTools).

Il9r *Gene*

Nine sets of primers were designed to amplify an average of 800 bp spanning exons 2 to 9 and most introns.

Other Genes

Primers were designed to amplify one exon from each gene (www.ensembl.org) using PrimerQuest (www.idtdna.com, Bio Tools). Successful amplification was observed in 95% of PCRs. Failure was limited to a subset of strains and primer pairs and was most likely due to the presence of mismatches in the primer sequence. PCR reactions contained 1.5–2 mM $MgCl_2$, 0.2–0.25 mM dNTPs, 0.2–1.8 µM of each primer and 0.5–1 units of Taq polymerase (Promega) or Platinum Taq DNA Polymerase High Fidelity (Invitrogen) in a final volume of 10–50 µL. Cycling conditions were 94°C, 4 min, 35 cycles at 94°, 55° and 68–72°C for 30 sec each, with a final extension at 68–72°C, 7 min. PCR products were purified using the High Pure PCR Product Purification kit (Roche) or fragments were excised from the gel and purified using the Qiaquick gel extraction kit (QIAGEN). Sequencing was performed at the UNC-CH Automated DNA Sequencing Facility on an ABI Prism 3730 (Applied Biosystems) or at the University of Pennsylvania DNA Sequencing Facility using the Big Dye Terminator kit on an ABI 3730-48 capillary sequencer.

Sequence Alignment and SNP Identification and Validation

All sequences were initially aligned using the Sequencher (Gene Codes) software. More refined alignment was done by eye to minimize, first, the number of events, and second, the number of variant positions. Aligned sequences were trimmed to retain only high-quality sequences. One fragment was eliminated from the analysis because of the presence of intrastrain polymorphisms consistent with the existence of a duplication. Analysis for sequence diversity was performed in the regions between the first and last nucleotide of high-quality sequence characterized in all 22 strains. Validation of the genetic variants was performed using several approaches as follows: (1) most insertion/deletion polymorphisms were confirmed in standard denaturing polyacrylamide gels; (2) SNPs detected among C57BL/6J, A/J, DBA/2J, and 129X1/SvJ were validated using the Celera Discovery System database (www.celera.com); (3) some variants generated a restriction endonuclease cleavage site, and 27 of these variants were confirmed by restriction digestion and electrophoresis; and (4) two SNPs were confirmed using Lightyper Simple Probe technology (Roche) designed to discriminate between the two alleles.

SDP Analysis

Alleles at diallelic variants present in the 20 *M. musculus* strains were represented as a series of 0s and 1s in the same order shown in Supplemental Table 1. The allele present in the first strain, CAST/EiJ, is always 0. Strains with the same allele as CAST/EiJ were considered 0s and strains with a different allele were considered 1s. Missing data were treated to minimize SDP number and to maximize SDP frequency within fragments.

Phylogenetic Analysis

To date the insertion/deletion and substitution variants found among *M. musculus* inbred strains, we used the phylogenetic tree reported in Guenet and Bonhomme (2003). Diallelic variants were classified as arising before the divergence of the specific lineages and before or after the divergence of the *M. musculus* subspecies (a, b, and c, respectively, in Fig. 2) using the following steps: (1) variants at which SPRET/EiJ and PANCEVO/EiJ have different alleles were assigned to class a; (2) variants at which SPRET/EiJ and PANCEVO/EiJ share the same allele; this allele was considered ancestral; (3) within this class, variants were assigned to class b if the new allele was found in inbred strains belonging to more than one subspecies; (4) variants were assigned to class c if the new allele was found in inbred strains belonging to only one subspecies. To avoid biases due to the use of hybrid inbred strains derived from more than one subspecies, the following caveats were applied from step two onward. (1) Classical inbred strains were omitted. (2) For wild-derived hybrid strains, alleles present at each variant were grouped with the appropriate parental subspecies in order to minimize the number of subspecies in which the new allele is found. In other words, alleles at each variant in JF1/Ms and MOLC/RkJ were classified as *M. m. castaneus* or *M. m. musculus*; alleles in SKIVE/EiJ were classified as *M. m. musculus* or *M. m. domesticus*, and alleles in CALB/RkJ were classified as *M. m. castaneus* or *M. m. domesticus*. Lastly, if the allele present in either SPRET/EiJ or PANCEVO/EiJ was unknown, only variants in which both alleles were found in at least two subspecies each were considered to predate the divergence of the subspecies.

Discrimination between insertions and deletions was based on the predicted ancestral allele in the ancestor of the three *Mus* species analyzed here.

ACKNOWLEDGMENTS

We thank Scott Bultman, Rosann Farber, Jean-Louis Guénet, Patrick Sullivan, and David Threadgill for comments on the manuscript and Tammi Briscoe for technical assistance. This work was partly supported by grants from the National Institutes of Health (R01 HD34508 and GM62537, C.S. and T.M.), the National Science Foundation (MCB-0133526, F.P.-M.V.) and the Andrew Mellon Foundation (F.P.-M.V.).

REFERENCES

Auffray, J.-C., Vanlerberghe, F., and Britton-Davidian, J. 1990. The house mouse progression in Eurasia: A palaeontological and archaeozoological approach. *Biol. J. Linn. Soc.* **41:** 13–25.

Beck, J.A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J.T., Festing, M.F., and Fisher, E.M. 2000. Genealogies of mouse inbred strains. *Nat. Genet.* **24:** 23–25.

Bonhomme, F. and Guenet, J.L. 1996. The laboratory mouse and its wild relatives. In *Genetic variants and strains of the laboratory mouse*, 3rd ed. (ed. M.F. Lyon et al.), pp. 1577–1596. Oxford University Press, Oxford, New York.

Bonhomme, F., Guenet, J.L., Dod, B., Moriwaki, K., and Bulfield, G. 1987. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *Biol. J. Linn. Soc.* **30:** 51–58.

Chen, F.C. and Li, W.H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68:** 444–456.

Chevret, P., Jenkins, P., and Catzeflis, F. 2002. Evolutionary systematics of the Indian mouse *Mus famulus* Bonhote, 198: Molecular (DNA/DNA hybridization and 12S rRNA sequences) and morphological evidences. *Zool. J. Linn. Soc.* **137:** 385.

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

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70:** 1490–1497.

Ferris, S.D., Sage, R.D., and Wilson, A.C. 1982. Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. *Nature* **295:** 163–165.

Forejt, J. 1996. Hybrid sterility in the mouse. *Trends Genet.* **12:** 412–417.

Greene-Till, R., Zhao, Y., and Hardies, S.C. 2000. Gene flow of unique sequences between *Mus musculus domesticus* and *Mus spretus*. *Mamm. Genome.* **11:** 225–230.

Grupe, A., Germer S., Usuka, J., Aud, D., Belknap, J.K., Klein, R.F., Ahluwalia, M.K., Higuchi, R., and Peltz, G. 2001. In silico mapping of complex disease related traits in mice. *Science* **292:** 1915–1918.

Guenet, J.L. and Bonhomme, F. 2003. Wild mice: An ever-increasing contribution to a popular mammalian model. *Trends Genet.* **19:** 24–31.

Kaessmann, H., Wiebe, V., and Paabo, S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286:** 1159–1162.

Kaessmann, H., Wiebe, V., Weiss, G., and Paabo, S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* **27:** 155–156.

Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., et al. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24:** 381–386.

Lundrigan, B.L., Jansa, S.A., and Tucker, P.K. 2002. Phylogenetic relationships in the genus *Mus*, based on paternally, maternally, and biparentally inherited characters. *Syst. Biol.* **51:** 410–431.

Orth, A., Belkhir, K., Britton-Davidian, J., Boursot, P., Benazzou, T., and Bonhomme, F. 2002. Hybridation naturelle entre deuz expeces sympatriques de souris: *M. musculus domesticus* et *M. spretus Lataste*. *C.R. Biol.* **325:** 89–97.

Prager, E.M., Orrego, C., and Sage, R.D. 1998. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150:** 835–861.

Reich, D.E., Gabriel, S.B., and Altshuler, D. 2003. Quality and completeness of SNP databases *Nat. Genet.* **33:** 457–458.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Sakate, R., Osada, N., Hida, M., Sugano, S., Hayasaka, I., Shimohira, N., Yanagi, S., Suto, Y., Hashimoto, K., and Hirai, M. 2003. Analysis of 5′-end sequences of chimpanzee cDNAs. *Genome Res.* **13:** 1022–1026.

Salisbury, B.A., Pungliya, M., Choi, J.Y., Jiang, R., Sun, X.J., and Stephens, J.C. 2003. SNP and haplotype variation in the human genome. *Mutat. Res.* **526:** 53–61.

Schalkwyk, L.C., Jung, M., Daser, A., Weiher, M., Walter, J., Himmelbauer, N., and Lehrach, H. 1999. Panel of microsatellite markers for whole-genome scans and radiation hybrid mapping and a mouse family tree. *Genome Res.* **9:** 878–887.

She, J.X., Bonhomme, F., Boursot, P., Thaler, L., and Catzeflis, F. 1990. Molecular phylogenies in the genus *Mus*: Comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. *Biol. J. Linn. Soc.* **41:** 83–103.

Tucker, P.K., Lee, B.K., Lundrigan, B.L., and Eicher, E.M. 1992. Geographical origin of the Y chromosomes in "old" inbred strains of mice. *Mamm. Genome* **3:** 254–261.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Vogel, G. 2003. Scientists dream of 1001 complex mice. *Science* **301:** 456–457.

Wade, C.M., Kulbokas III, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., and Daly, M.J. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wiltshire, T., Pletcher, M.T., Batalov, S., Barnes, S.W., Tarantino, L.M., Cooke, M.P., Wu, H., Smylie, K., Santrosyan, A., Copeland, N.G., et al. 2003. Genome-wide single-nucleotidepolymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci.* **100:** 3380–3385.

Yalcin, B., Fullerton, J., Miller, S., Keays, D.A., Brady, S., Bhomra A., Jefferson, A., Volpi, E., Copley, R.R., Flint, J., et al. 2004. Unexpected complexity in the haplotypes of the commonly used inbred strains of laboratory mice. *Proc. Natl. Acad. Sci.* **26:** 9734–9739.

Zietkiewicz, E., Yotova, V., Gehl, D., Wambach, T., Arrieta, I., Batzer, M., Cole, D.E., Hechtman, P., Kaplan, F., Modiano, D., et al. 2003. Haplotypes in the dystrophin DNA segmen point to a mosaic origin of modern human diversity. *Am. J. Hum. Genet.* **73:** 994–1015.

WEB SITE REFERENCES

http://www.celera.com; Celera Discovery System Home page.
http://www.ensembl.org; Ensembl Genome Browser Home page.
http://www.idtdna.com; Integrated DNA Technologies Home page.
http://www.jax.org; The Jackson Laboratory Home page.

Received February 26, 2004; accepted in revised form August 4, 2004.