Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse

Tim Wiltshire*†‡, Mathew T. Pletcher†§, Serge Batalov*†, S. Whitney Barnes*, Lisa M. Tarantino*, Michael P. Cooke*, Hua Wu¶, Kevin Smylie‖, Andrey Santrosyan*, Neal G. Copeland**, Nancy A. Jenkins**, Francis Kalush††, Richard J. Mural††, Richard J. Glynne¶, Steve A. Kay*§¶, Mark D. Adams††, and Colin F. Fletcher*§

*Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121; §The Scripps Research Institute, La Jolla, CA 92121; ¶Phenomix, San Diego, CA 92121; ‖Sequenom, Inc., San Diego, CA 92121; **National Cancer Institute, Frederick, MD 21702; and ††Celera Genomics, Rockville, MD 20850

Communicated by Peter G. Schultz, The Scripps Research Institute, La Jolla, CA, January 7, 2003 (received for review November 19, 2002)

The nature and organization of polymorphisms, or differences, between genomes of individuals are of great interest, because these variations can be associated with or even underlie phenotypic traits, including disease susceptibility. To gain insight into the genetic and evolutionary factors influencing such biological variation, we have examined the arrangement (haplotype) of single-nucleotide polymorphisms across the genomes of eight inbred strains of mice. These analyses define blocks of high or low diversity, often extending across tens of megabases that are delineated by abrupt transitions. These observations provide a striking contrast to the haplotype structure of the human genome.

The availability of assembled genomic sequences has greatly facilitated the study of variation by providing a framework to map polymorphisms at a resolution that can't be achieved by recombination analysis. The most abundant genetic variants are single-nucleotide polymorphisms (SNPs), which provide the most comprehensive resource for ascertaining genetic diversity. Importantly, physically linked SNPs are coinherited as a series of alleles in a pattern known as a haplotype, but the regulation of haplotype structure is poorly understood. A detailed understanding of haplotype structure is a prerequisite for the design of association studies that are used to map trait and disease loci. The evolutionary factors that influence the structure of these blocks include recombination, mutation, population history, and selection. Recent studies in human, encompassing tens of megabases, suggest that haplotype blocks are largely defined by recombination hotspots and extend no more than tens of kilobases (1–5). Mice provide an experimental resource for investigating the mechanisms that regulate haplotype structure in mammalian genomes, in part because of well defined strain genealogies, standardized mapping tools, and control of breeding and selection. Importantly, inbred mouse strains are homozygous at all loci, and thus haplotypes can be observed directly. We have undertaken SNP discovery across eight inbred strains of mice and used multiplexed or pooled SNP assays in comprehensive genome scans and for SNP validation. Analysis of the distribution of alleles among strains reveals limited haplotype diversity, with strain pairs sharing common haplotypes across 30–60% of the genome. Simple sequence-length polymorphisms (SSLPs, dinucleotide repeats) show similar distribution patterns. These observations have important implications for quantitative trait loci (QTL) analysis as well as positional cloning of monogenic loci, because it is apparent that shared haplotype blocks reduce the percentage of the genome that can be queried for biological variation

Materials and Methods

SNP Discovery. The Jackson Laboratory T31 Mouse Radiation Hybrid database provided the data set to select markers mapped based on their radiation hybrid order. Sequence was retrieved for all radiation hybrid markers from NCBI Entrez, and corresponding genomic sequence was identified from Celera mouse sequence vR6 by using BLAST. We selected 2,600 loci evenly spaced across all autosomes and a reduced set (30) from the X chromosome for SNP discovery. In addition, we selected 554 SNP loci from the Whitehead Institute SNP set (www-genome.wi.mit.edu/SNP/mouse/) for SNP discovery in the *Mus spretus* strain. SNPs were identified by direct sequencing of PCR products generated for each of the SNP loci from eight inbred mouse strains (C57BL/6J, 129SvIm/J, C3H/HeJ, DBA/2J, A/J, BALB/cByJ, CAST/Ei, and SPRET/Ei). All mouse DNAs were procured from The Jackson Laboratory.

For two strains with 1,112 randomly distributed SNPs (A/J vs. BALB/cByJ), the theoretical distribution of the largest block lengths is the extreme value distribution with parameters determined by simulation; the mean expected block length is $L\_genome/(N\_SNP+N\_chr) = 2.3$ megabases (Mb), median = 1.6 Mb, mode of the largest block length, 16 Mb, and the probability of the largest block being 40, 60, 80, 100, and 120 Mb is 2.8×10^{-5}, 4.2×10^{-9}, 6.4×10^{-13}, 9.6×10^{-17}, and 1.4×10^{-20}, respectively. The hypothesis of randomness of the distribution of SNPs can be rejected with $P < 10^{-6}$ even if the top 10 haplotype blocks are dismissed as possibly split.

SNP Database and SNPview. Sequence data were assembled and annotated by using LASERGENE (DNASTAR, Madison, WI) sequence analysis programs. Publicly available SNPs useful in a C57BL/6J strain combination were incorporated from Whitehead Institute SNP data (www-genome.wi.mit.edu/SNP/mouse) and the Roche database (http://mouseSNP.roche.com). Dinucleotide CA repeats were incorporated from Whitehead Institute data (6–9). These data then were made available for interactive viewing through SNPview (www.gnf.org/SNP) where data are visualized with the DerBrowser (10) applet.

Shared haplotypes were defined (*i*) by identifying the longest regions of contiguous strain-pair identity, (*ii*) by taking those two strain pairs, and all other strains bearing the same alleles, and coloring them as shared haplotypes within that region. Where alleles diverged or any different allele pairings occurred within a longest block, they were identified with different colors, the same allele groups bearing the same colors. Sequenced loci were considered as a single unit, and breaks in strain-pair sharing were defined between and not within a locus. Thus, any number of SNPs identified within a locus established a new haplotype and a change in color pattern.

The dendrogram for 29 mouse strains was produced by using the TREE EXPLORER program from the multiple alignments of the sequences comprising 938 SNP alleles. The multiple align-

Abbreviations: SNP, single-nucleotide polymorphism; SSLP, simple sequence-length polymorphism; QTL, quantitative trait loci; Mb, megabase(s).

Data deposition: The SNP data have been deposited in the dbSNP, www.ncbi.nlm.nih.gov/SNP/(dbSNP ID nos. 6313904-6395786) and Mouse Genome Informatics database, www.jax.org/phenome (ID no. MPD:133).

†T.W., M.T.P., and S.B. contributed equally to this work.

‡To whom correspondence should be addressed. E-mail: timw@gnf.org.

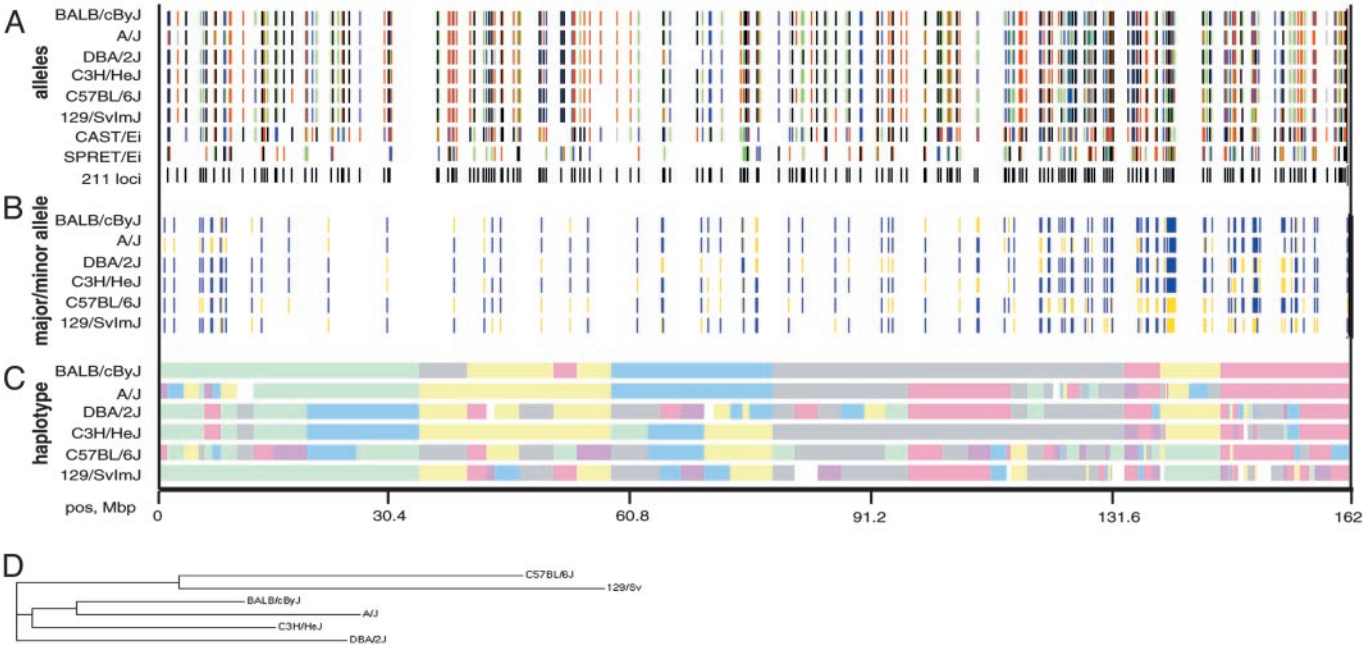


Fig. 1. Distribution of SNPs and haplotypes in SNPview on chromosome 4. (*A*) Polymorphic nucleotides for each strain plotted by physical position and colored by nucleotide. G, blue; A, green; T, black; C, red; insertions/deletions, orange; multiple nucleotide insertion/deletion, gray. All SNP-containing loci (sequence-tagged sites) are indicated by black stripes (row 9). (*B*) Polymorphic nucleotides (six laboratory strains) shown colored as major (blue) and minor (yellow) alleles. (*C*) Haplotypes, defined as regions in which two or more strains share the same SNP alleles across multiple loci, are indicated by colored blocks. Colors are reused for visual simplicity. Colors should be interpreted only to indicate similarity among strains at a particular position in the vertical axis, and no relationship is implied by similarly colored blocks that are separated in the horizontal axis. No color indicates an absence of data. (*D*) Dendrogram of strain relatedness based on CLUSTALW analysis of SNP alleles among six inbred strains for chromosome 4.

ment was calculated by using bootstrapped CLUSTALW 1.8 with high gap penalties.

Genetic and Physical Mapping of SNPs. The efficiency of conversion of an SNP into a useful assay can be gauged from an initial set of 510 sequences typed against 159 (B6 × 129) F_2 progeny. Ten (2%) could not have satisfactory primers designed, 48 (9.8%) were discarded because they only detected one genotyping allele, 38 (7.5%) were not robust assays because they did not consistently amplify both alleles, and 9 (2.6%) never worked because of a lack of PCR amplification. This represents an overall 78.1% success rate; no attempt was made to recover any failed assays. These data were combined with genotype data from 15 (B6 × BALB/c) F_2 mice and 13 CXB and 31 BXD recombinant inbred strains to produce an integrated genetic map containing 914 independent SNP markers. Markers polymorphic in all three data sets (132) were used to connect the data between the three strain combinations (Table 1, which is published as supporting information on the PNAS web site, www.pnas.org). A subset of B6/SPRET SNPs were genetically mapped to the Copeland–Jenkins (C57BL/6J × *M. spretus*) F_1 × C57BL/6J backcross BSB mapping panel (SNPview; Table 2, which is published as supporting information on the PNAS web site). DNAs (188) from the panel were genotyped for 480 SNPs by using a 5-fold multiplex protocol. SNP genotype data were integrated into the 1,858 assays currently available in the mapping panel and analyzed by using MAPMANAGER QTX (11). Three hundred and thirty-eight assays were included in the anchor SNP set with an average logarithm of odds score of 38 and a lowest logarithm of odds score of 6, and >80% of assays had an allele called for >80% of DNAs. All SNP loci sequences were mapped to both Celera R13 and the MGSC v3 assemblies by using a BLAT mapping program (12).

SNP Assay System. SNP assays were performed by using the Sequenom MassARRAY system. Primers for PCR and single base extension were designed by using the SPECTRODESIGNER software package (Sequenom). For SNP genotyping the DNA sample to be queried was diluted to 2.5 ng/μl, and 1 μl of DNA was combined with 3.04 μl of water/0.04 μl of 25 mM dNTPs (Invitrogen)/0.02 μl of 5 units/μl HotStar *Taq* (Qiagen)/0.5 μl of 10× HotStar PCR buffer containing 15 mM $MgCl_2$, 0.2 μl PCR primers mixed together at a concentration of 5 μM (1.25 μM if the primers are part of a multiplexed reaction set), and 0.2 μl of 25 mM $MgCl_2$. Reactions were heated at 95°C for 15 min followed by 45 cycles at 95°C for 20 s, 56°C for 30 s, and 72°C for 1 min and a final incubation at 72°C for 3 min. After PCR amplification, remaining dNTPs were dephosphorylated by adding 1.53 μl of water, 0.17 μl of homogeneous mass extend reaction buffer (Sequenom), and 0.3 units of shrimp alkaline phosphatase (Sequenom). The reaction was placed at 37°C for 20 min, and the enzyme was deactivated by incubating at 85°C for 5 min. After shrimp alkaline phosphatase treatment, the genotyping reaction was combined with 1.242 μl of water/0.2 μl of 10× Termination mix (Sequenom)/0.018 μl of 0.063 units/μl Thermosequenase (Sequenom)/0.54 μl of 10 μM extension primer. The MassEXTEND reaction was carried out at 94°C for 2 min and then 55 cycles of 94°C for 5 s, 52°C for 5 s, and 72°C for 5 s. The reaction mix was desalted by adding 3 mg of a cationic resin, SpectroCLEAN (Sequenom), and resuspended in 16 μl of water. Completed genotyping reactions were spotted in nanoliter volumes onto a matrix arrayed into 384 elements on a silicon chip (Sequenom SpectroCHIP), and the allele-specific mass of the extension product was determined by matrix-assisted laser desorption ionization/time-of-flight MS. Analysis of data by SPECTROTYPER software generates automated allele calling.

Concentrations of DNA from the BSB mapping panel were determined by fluorimetry using pico green dye (Molecular

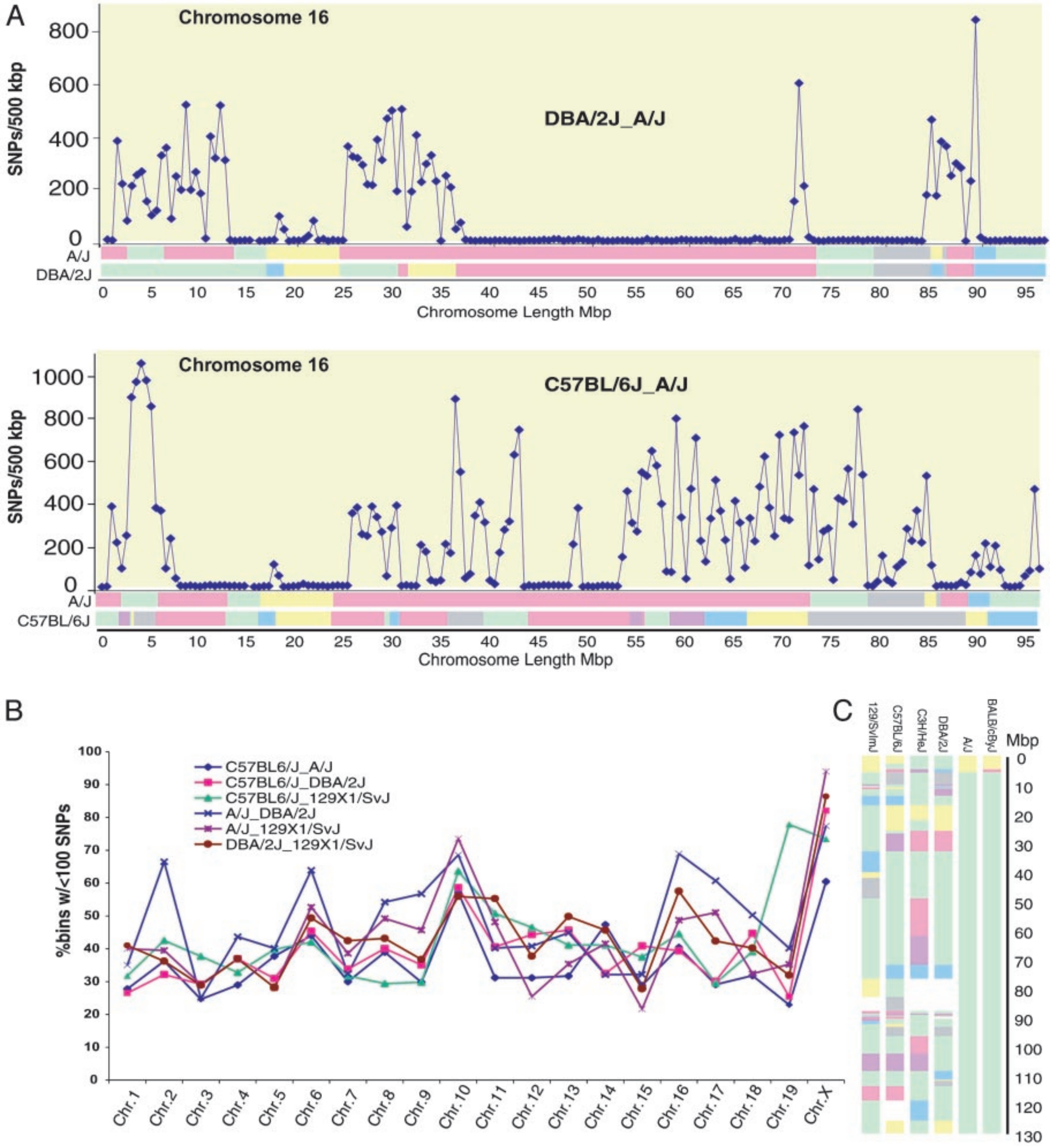


Fig. 2. SNP density varies greatly by strain combination, chromosome, and chromosomal region. (*A*) SNP density (SNPs per 500 kb) on chromosome 16 is plotted by physical position for specific strain pairs. Common haplotype blocks largely correspond to areas of low SNP density. In some cases, haplotype analysis missed small regions of high diversity due to undersampling, e.g., 71–74 Mb (*Upper*). In other cases a single data point may define a haplotype block in the six-way strain comparison, but that region may contain few SNPs, e.g., 93–98 Mb (*Lower*). (*B*) Percentages of 500-kb windows containing <100 SNPs are plotted by chromosome for all combinations of 129/Sv, A/J, DBA/2J, and C57BL/6J sequence data. Generally, low-density blocks span 30–60% of each chromosome for any strain comparison. (*C*) Consistent with *B*, analysis of chromosome 10 indicates very limited haplotype diversity, with >90% of the chromosome having no more than two haplotypes.

Probes) according to manufacturer recommendations. DNA concentrations were adjusted to 25 ng/μl. BSB DNA samples were divided into two groups for assignment to specific pools based on their genotype at chromosome 12 marker D38416, which could be either homozygote B6 or a heterozygote with both a SPRET and B6 allele. Out of each group of DNAs, a pool of 10, 15, 20, 30, and 50 samples was constructed for a total of 10 pools. Equal volumes of the diluted DNAs were combined to form each pool such that the final concentration of the mixture remained at 25 ng/μl. The pools were allelotyped with 13 SNP assays from mouse chromosome 12 by using the protocol described above. High-quality SNP assays for general genotyping do not always work well for allelotyping, and an additional round of quality control must be performed. Assays that gave a skewed F_1 allele frequency >70:30% were excluded from analysis, and results from replicate sampling with a standard deviation >5% were also excluded. Analysis of the data by SPECTROTYPER software includes an estimation of the area under the peak of each extension product so the frequency of each SNP allele can be calculated. Final values were expressed in the percentage of B6 alleles found in a given pool of DNA for a given SNP marker (Fig. 4). Determining a map position for a mutant locus is then

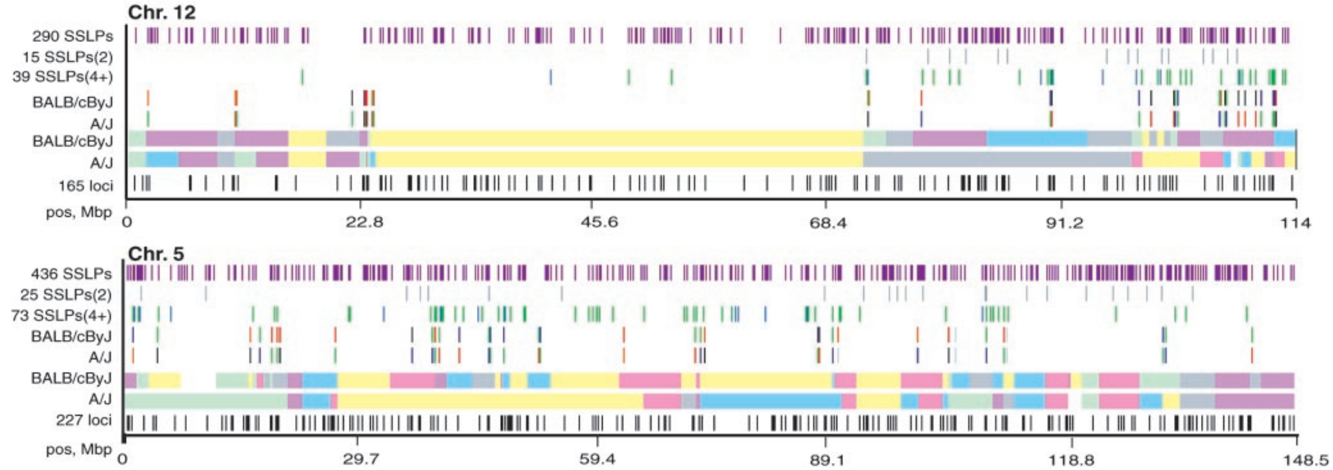


Fig. 3. Comparative analysis of SSLP and SNP distribution. A/J and BALB/cByJ SSLPs and SNPs are plotted by physical position for chromosomes 12 and 5. For each panel, the top row indicates the position of all loci containing polymorphic CA repeats. SSLPs polymorphic between A/J and DBA/2J (rows 2 and 3), SNPs colored by allele (rows 4 and 5), haplotypes defined by sample sequencing (rows 6 and 7), and sequenced loci (row 8) are shown. SNPs and SSLPs display a degree of codistribution even on chromosome 5, where many short, shared, and disparate haplotypes are defined.

defined easily by marked divergence of the mutant and wild-type allele frequencies.

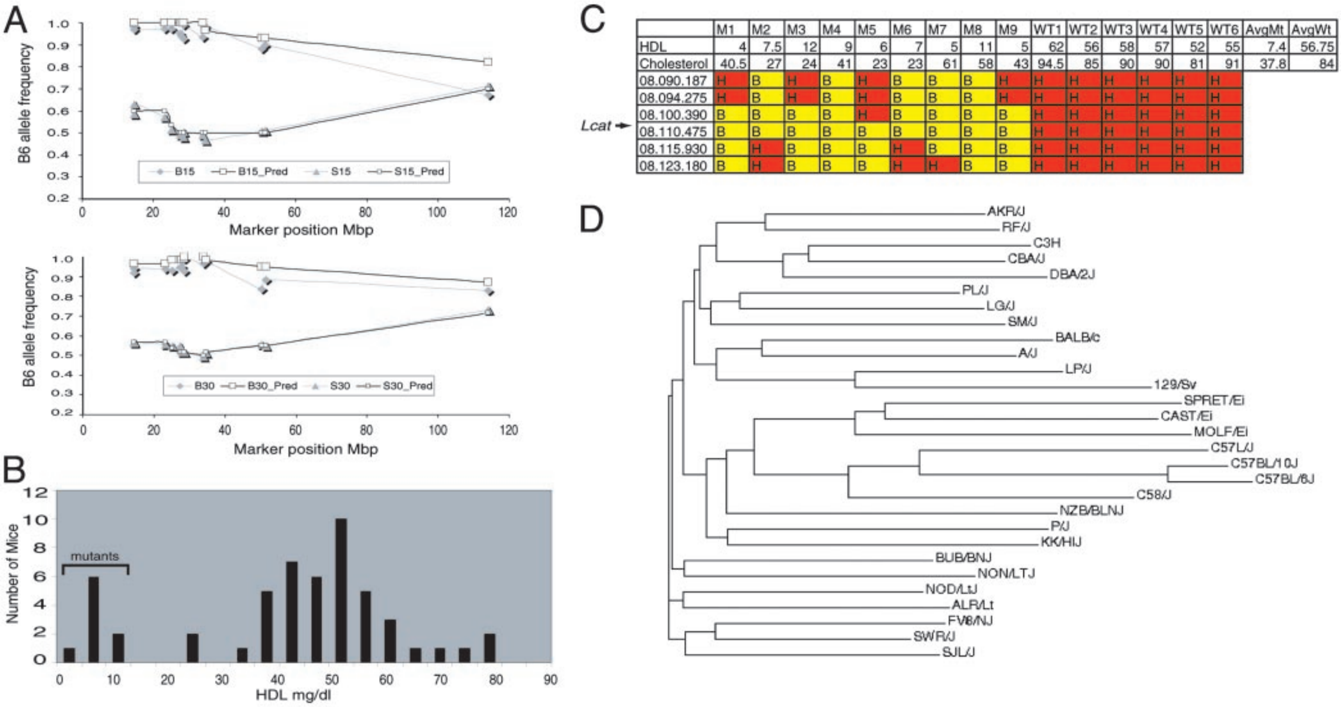
Celera SNP Discovery and Validation. The Celera mouse genome assembly was constructed by using data collected at Celera from strains 129×1/SvJ, 129S1/SvImJ (129S1/SvImJ was only sequenced to 0.3-fold coverage), DBA/2J, and A/J and from sequence reads from C57BL/6J obtained by the Mouse Genome Sequence Consortium. SNPs were identified by using an algorithm that evaluated the fragment coverage and base calls at each position and eliminated sequence differences likely to occur because of low-quality primary sequence, repetitive elements, or assembly of paralogous sequences. Because the genome sequence coverage of each strain is incomplete, data are not present from all strains at each SNP location. The resulting set of 70,957 SNPs contains 23,976 SNPs predicted on the basis of a single fragment representing the minor allele, and 33,168 SNPs have fragment coverage for all four major strains. Predicted SNPs were selected for validation with a range of characteristics, but particular emphasis was placed on SNPs with single fragment coverage and SNPs that were "block-breakers," exceptions to the broad haplotypes shared between a pair of strains, and 300 bp of flanking sequence on both sides of each SNP were imported into the TaqMan assay-design program. Primer Express (Applied Biosystems, Foster City, CA) was used to design both the PCR primers and the MGB TaqMan probes. One allelic probe was labeled with the fluorescent 6-carboxyfluorescein dye and the other with the fluorescent VIC dye. PCRs were run in TaqMan Universal Master mix without UNG (Applied Biosystems) with PCR primer concentrations of 900 nM and TaqMan MGB-probe concentrations of 200 nM. Reactions were performed in 384-well format in a total reaction volume of 5 ml using 1.0 ng of genomic DNA. The plates then were placed in a thermal cycler (PE 9700, Applied Biosystems) and heated at 95°C for 10 min followed by 50 cycles of 95°C for 15 s and 60°C for 1 min with a final soak at 25°C. The TaqMan assay plates were transferred from the thermal cyclers to the Prism 7900HT instruments (Applied Biosystems) where the fluorescence intensity in each well of the plate was read. Fluorescence data files from each plate were analyzed by automated allele-calling software (unpublished data).

Mapping of Low High-Density Lipoprotein Phenotype. An SNP mapping panel of 510 SNPs between B6 and 129SvIm/J strains was originally selected and used as input for the Sequenom SPECTRODESIGNER software. The multiplexing option was set to combine up to five assays into a single reaction. SPECTRODESIGNER was unable to design primers for 10 of the SNP sequences. The remaining sequences were condensed into 104 multiplexed groups. Four assays remained uniplexed, one remained duplex, and one remained quadplex. Genotyping assays were carried out as described above on nine mutant mice and six wild-type littermates generated from *N*-ethyl-*N*-nitrosourea-induced mutant mice. Results were imported to MAPMANAGER QTX. Sequencing of the *Lcat* gene was carried out for three wild-type and three mutant mice, and a single T > C transition was identified in exon 6 resulting in a leucine-to-proline amino acid change in mutant mice. A significant reduction in the average catabolic rate of Lcat was observed in mutants by using an Lcat assay (Roar Biomedical, New York).

Results and Discussion

To identify a genome-wide panel of SNPs, we selected 2,600 evenly distributed loci (sequence-tagged sites) for PCR amplification and sequencing from six common and well characterized inbred ("laboratory") mouse strains (C57BL/6J, 129S1/SvImJ, C3H/HeJ, DBA/2J, A/J, and BALB/cByJ) and two wild-derived inbred strains *Mus musculus castaneus* (CAST/Ei) and *M. spretus* (SPRET/Ei). The density, distribution across sequence-tagged site loci, strain distribution, and transversion/transition rate of SNPs in this set correspond well with previous SNP analysis (Table 3, which is published as supporting information on the PNAS web site; refs. 13 and 14). These data were supplemented with publicly available data and visualized with an interactive web-based browser (Fig. 1 *A–C*, SNPview, www.gnf.org/SNP/).

After mapping these SNPs to the assembled mouse genome (15), we examined the distribution of polymorphisms by pairwise comparison of C57BL/6J alleles with each of the other laboratory strains in this data set. This revealed 23 nonpolymorphic regions >20 Mb, with the largest being 61 Mb on chromosome 10. Two regions averaging >20 Mb (chromosomes 10 and 13) are completely deficient in polymorphisms in all five strain combinations. We have extended this analysis to all strain comparisons to identify regions of shared haplotype. Thus, regions containing a series of identical alleles between two or more strains are indicated as blocks of identical color in those strains (Fig. 1). Regions of haplotype sharing were in some cases surprisingly



	M1	M2	M3	M4	M5	M6	M7	M8	M9	WT1	WT2	WT3	WT4	WT5	WT6	AvgMt	AvgWt
HDL	4	7.5	12	9	6	7	5	11	5	62	56	58	57	52	55	7.4	56.75
Cholesterol	40.5	27	24	41	23	23	61	58	43	94.5	85	90	90	81	91	37.8	84
08.090.187	H	B	H	B	H	B	B	B	H	H	H	H	H	H	H		
08.094.275	H	B	H	B	H	B	B	B	H	H	H	H	H	H	H		
08.100.390	B	B	B	B	H	B	B	B	B	H	H	H	H	H	H		
08.110.475	B	B	B	B	B	B	B	B	B	H	H	H	H	H	H		
08.115.930	B	H	B	B	B	H	B	B	B	H	H	H	H	H	H		
08.123.180	B	H	B	B	B	H	H	B	B	H	H	H	H	H	H		

Fig. 4. Validation and utility of SNP discovery by mapping and genotyping. (*A*) Experimentally determined allele frequencies within pooled DNA samples were compared with known combined individual genotypes (_Pred). Pools of DNA were constructed from 15 or 30 samples from the *M. spretus* backcross panel. Individual DNAs were assigned to a specific pool based on their known genotype at marker D38418 (located at 33.9 Mb). B pools (B15 and B30) had an allele frequency of 100% for the C57BL/6J allele of D38418. S pools (S15 and S30) contain both the *M. spretus* and C57BL/6J allele at a 1:1 ratio. Thus, the expected C57BL/6J allele frequencies (100% and 50%) were detected at loci tightly linked to D38418, and the expected shift in frequency (to ≈75%), due to recombination, is seen in distal markers. (*B* and *C*) SNP assays can be used by reaction multiplexing. The histogram indicates mutant mice having abnormally low plasma high-density lipoprotein levels (nine, 17%). A genome scan of nine mutant mice (M1–9) and six normal littermates (WT1–6) identified a 15-Mb region on MMU8 that cosegregated with the phenotype. Names of individual markers (the first two digits indicate chromosome, and the next six digits indicate megabase and kilobase position) are listed in the first column. Genotypes are indicated by either a B (homozygous C57BL/6J) at that locus or H (heterozygote). (*D*) Dendrogram of genetic relationship among 29 mouse strains.

large, up to 120 Mb (see Fig. 5, which is published as supporting information on the PNAS web site, for haplotype size distribution), where no SNPs were identified between strain pairs by our sampling methods. The model of uniform distribution of SNPs can be rejected with high confidence.

To determine the density of polymorphisms in regions of apparently common haplotype accurately we analyzed genomic sequence from Celera, because that assembly includes much greater sequence coverage of the DBA/2J, A/J, C57BL/6J, and 129×1/SvJ strains (16). A subset of SNPs (70,957) from chromosome 16 (96 Mbp) was selected for analysis here based on the following criteria: (*i*) only ACTG polymorphisms were considered, not insertion/deletion polymorphisms, and (*ii*) SNPs were excluded as likely artifacts if they comprised more than one predicted allele from the same inbred strain based on previous SNP validation studies (data not shown). Pairwise comparison of strains across chromosome 16 reveals large intervals of very low SNP density (≈10 SNPs per 500 kbp) separated by regions of "normal" density (>100 SNPs per 500 kbp) that are demarcated by abrupt transitions (Fig. 2*A*). The regions of low SNP density correspond well with the regions of shared haplotype defined by sample sequencing. Small regions of discrepancy between the haplotype pattern inferred from whole-genome sequence data and from sample sequencing (e.g., 71–73 Mb, A/J-to-DBA2/J comparison) result from the higher resolution achieved by whole-genome sequencing. The resolution of haplotype regions and boundaries seems to be <2 Mb, because all regions of that size and greater were discovered by this sampling method. Across all chromosomes, strain pairs share common haplotypes across 30–60% of the genome (Fig. 2*B*). This observation has important implications for studies of strain traits, because it is apparent that shared identity blocks significantly reduce the percentage of the genome that can be queried easily for biological variation. At the same time, the identification of haplotypes also presents an opportunity for association studies with QTL. Specifically, the haplotypes identified in a QTL interval in a given mapping cross can be tested for association if that QTL has been mapped in additional strains. This approach can be effective in reducing QTL intervals to sizes amenable to candidate gene analysis.

If the blocks of common haplotypes represent regions that were similar or identical when the strains were derived, we presume that the less-abundant polymorphisms have arisen recently. If that is true, the polymorphism density should be similar to that between sublines derived soon after strains were inbred. As a test case, we compared C57BL/10J and C57BL/6J because the B6 and B10 sublines were derived in the 1930s from the inbred C57BL strain. We resequenced 2,384 sequence-tagged site loci from C57BL/B10J and identified 99 SNPs in 58 loci (≈1 SNP per 20 kb). This SNP density is similar to that described for shared haplotypes between more divergent strains, implying that the common haplotype blocks between strains did not diverge significantly earlier than B6 and B10. This is relevant to haplotypes because local differences in mutation rate can influence haplotype structure. However, the number of B6/B10 SNPs isn't sufficient to determine whether there are regional differences in the mutation rates in the mouse genome. It should be noted that the regions of low SNP density on chromosome 16 show a tight distribution around an average of 10 SNPs per 500 kb, suggesting a consistent mutation rate across at least this 65 Mb of DNA (Fig. 2*A*).

To investigate whether SNPs behave differently from polymorphic markers that arise by a different molecular mechanism, we placed the known polymorphic CA repeats (6,200 loci) (6) on the assembled sequence and examined the distribution of alleles among strains. A scarcity of SSLPs was identified in the SNP "deserts" (Fig. 3; Table 4, which is published as supporting information on the PNAS web site; SNPview). For example, for A/J and BALB/cByJ, chromosome 12 SSLPs and SNPs are both concentrated in the telomeric region where there are few common haplotypes. Chromosome 5, which contains many short blocks of common and disparate haplotypes, also shows a codistribution of SSLPs and SNPs. To test this observation, markers where counted in 5-, 2-, and 0.5-Mb bins, and codistribution of bins was evaluated by χ^2 test ($1\text{-}P = 10^{-27}$, 10^{-24}, and 7.4×10^{-11}, respectively). These observations indicate that regions of low diversity are not a peculiarity of SNPs.

SNPs were validated by designing assays that were typed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (Sequenom MassArray) (17). The efficiency of conversion of an SNP into a useful assay was 78%. For the 1,240 SNP assays that were typed and integrated into two genetic maps (SNPview; Tables 1 and 2) the genetic localization and physical position showed a high degree of agreement for all chromosomes. We also tested two methods to facilitate rapid whole-genome scans; pooling DNA samples and determining allele frequencies (18) and multiplexing of assays. We evaluated pooled-DNA methods by using 13 SNPs that consistently detected changes of allele frequency of 3.5 ± 3.4% (Fig. 4*A*). Approximately 400 multiplexed SNP assays (4-fold multiplex) were used to genotype progeny from an F_2 intercross segregating a phenotype of reduced high-density lipoprotein, defining a region associated with the mutant phenotype (Fig. 4 *B* and *C*). A candidate gene, Lecithin:cholesterol acyltransferase (*Lcat*), was identified in this region because it is known to be mutated in a human disease and a mouse knockout strain characterized by low high-density lipoprotein (19, 20).

The limited number of haplotype blocks found in the laboratory strains may reflect the fact that these strains were derived from a small pool of founders, which themselves may have been somewhat inbred at the time the strains were derived (21, 22). In addition, selection during inbreeding may have acted to reduce diversity further. Alternatively, limited diversity may be a general feature of the mouse genome, consistent with the fact that the majority of human chromosomes contain only 3–5 haplotypes. To examine this question, we used assays to type 21 additional strains derived from a variety of sources with a set of 934 SNP assays. A dendrogram identifies strain relatedness across all strains (Fig. 4*D*), and this generally recapitulates population history. It is evident from the typings that additional haplotypes are present. This is consistent with the fact that, in contrast to the situation among the laboratory strains, the *M. musculus castaneus* and *M. spretus* polymorphisms were distributed evenly around the genome (Fig. 1; SNPview). This indicates that the laboratory-strain haplotypes are not a common feature of all mouse species but more likely result from the recent population history of these strains. If true, then population history may play a significant role in shaping the haplotype structure of many experimentally and commercially important species, both plant and animal, which have been derived in a similar manner. However, the recent common ancestry of related strains does not preclude their use as experimental resources to study the evolutionary forces that determine haplotype. For example, sublines that arose from a common ancestor can be used for genome-wide studies of mutation rate. Additionally, controlled breeding can be used to examine the effects of selection and recombination. Thus, these studies lay the groundwork for a variety of experiments that can improve our ability to predict and interpret haplotypes.

We thank F. Lippincott, N. Ziaee, and D. Stradley for invaluable work in sequence annotation; P. Merritt and C. Motta for genotyping analysis (Genomic Institute of the Novartis Research Foundation); and D. Nusskern for analysis and preparation of SNP data from chromosome 16 (Celera). This study was supported by Novartis Grant SFP-1407.

1. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002) *Science* **296,** 2225–2229.
2. Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001) *Nat. Genet.* **29,** 217–222.
3. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. (2001) *Nat. Genet.* **29,** 229–232.
4. Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P., Richter, D., Lavery, T., Kouyoumjian, R., Farhadian, S., Ward, R., *et al.* (2001) *Nature* **411,** 199–204.
5. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., *et al.* (2001) *Science* **294,** 1719–1723.
6. Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M. J., Ingalls, K. A., O'Connor, T. J., *et al.* (1996) *Nature* **380,** 149–152.
7. Dietrich, W. F., Miller, J. C., Steen, R. G., Merchant, M., Damron, D., Nahf, R., Gross, A., Joyce, D. C., Wessel, M., Dredge, R. D., *et al.* (1994) *Nat. Genet.* **7,** 220–245.
8. Copeland, N. G., Jenkins, N. A., Gilbert, D. J., Eppig, J. T., Maltais, L. J., Miller, J. C., Dietrich, W. F., Weaver, A., Lincoln, S. E., Steen, R. G., *et al.* (1993) *Science* **262,** 57–66.
9. Nusbaum, C., Slonim, D. K., Harris, K. L., Birren, B. W., Steen, R. G., Stein, L. D., Miller, J., Dietrich, W. F., Nahf, R., Wang, V., *et al.* (1999) *Nat. Genet.* **22,** 388–393.
10. Grigoriev, A. (1997) *Trends Genet.* **13,** 499.
11. Manly, K. F., Cudmore, R. H., Jr., & Meer, J. M. (2001) *Mamm. Genome* **12,** 930–932.
12. Kent, W. J. (2002) *Genome Res.* **12,** 656–664.
13. Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J. K., Klein, R. F., Ahluwalia, M. K., Higuchi, R. & Peltz, G. (2001) *Science* **292,** 1915–1918.
14. 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) *Nat. Genet.* **24,** 381–386.
15. Kerlavage, A., Bonazzi, V., di Tommaso, M., Lawrence, C., Li, P., Mayberry, F., Mural, R., Nodell, M., Yandell, M., Zhang, J., *et al.* (2002) *Nucleic Acids Res.* **30,** 129–136.
16. Mural, R. J., Adams, M. D., Myers, E. W., Smith, H. O., Miklos, G. L., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J., *et al.* (2002) *Science* **296,** 1661–1671.
17. Jurinke, C., van den Boom, D., Cantor, C. R. & Koster, H. (2001) *Methods Mol. Biol.* **170,** 103–116.
18. Taylor, B. A., Navin, A. & Phillips, S. J. (1994) *Genomics* **21,** 626–632.
19. Sakai, N., Vaisman, B. L., Koch, C. A., Hoyt, R. F., Jr., Meyn, S. M., Talley, G. D., Paiz, J. A., Brewer, H. B., Jr., & Santamarina-Fojo, S. (1997) *J. Biol. Chem.* **272,** 7506–7510.
20. Funke, H., von Eckardstein, A., Pritchard, P. H., Albers, J. J., Kastelein, J. J., Droste, C. & Assmann, G. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4855–4859.
21. Bailey, D. (1978) in *Origins of Inbred Mice*, ed. Morse, H. (Academic, New York), pp. 423–438.
22. Beck, J. A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. F. & Fisher, E. M. (2000) *Nat. Genet.* **24,** 23–25.