Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse

Kerstin Lindblad-Toh[1], Ellen Winchester[1], Mark J. Daly[1], David G. Wang[1,2], Joel N. Hirschhorn[1,3], Jean-Philippe Laviolette[1], Kristin Ardlie[1], David E. Reich[1], Elizabeth Robinson[1], Pamela Sklar[1,4], Nila Shah[5], Daryl Thomas[5], Jian-Bing Fan[5], Thomas Gingeras[5], Janet Warrington[5], Nila Patil[5], Thomas J. Hudson[1,6] & Eric S. Lander[1,7]

Single-nucleotide polymorphisms (SNPs) have been the focus of much attention in human genetics because they are extremely abundant and well-suited for automated large-scale genotyping. Human SNPs, however, are less informative than other types of genetic markers (such as simple-sequence length polymorphisms or microsatellites) and thus more loci are required for mapping traits. SNPs offer similar advantages for experimental genetic organisms such as the mouse, but they entail no loss of informativeness because bi-allelic markers are fully informative in analysing crosses between inbred strains. Here we report a large-scale analysis of SNPs in the mouse genome. We characterized the rate of nucleotide polymorphism in eight mouse strains and identified a collection of 2,848 SNPs located in 1,755 sequence-tagged sites (STSs) using high-density oligonucleotide arrays. Three-quarters of these SNPs have been mapped on the mouse genome, providing a first-generation SNP map of the mouse. We have also developed a multiplex genotyping procedure by which a genome scan can be performed with only six genotyping reactions per animal.

To identify SNPs, we screened 3,717 STSs that had been developed as part of our recent efforts to construct genetic and physical maps of the mouse genome[1,2]. The STSs were derived from random genomic sequences (2,948 STSs) and expressed sequence tags (769 ESTs), and comprised 442 kb of genomic sequence. The loci were amplified from eight commonly used inbred mouse strains: seven laboratory strains of *Mus musculus domesticus* (129/Sv, A/J, AKR/J, BALB/cByJ, C3H/HeJ, C57BL/6J, DBA/2J); and one strain from the distinct subspecies *M. m. castaneus* (CAST/Ei), which separated from *M. m. domesticus* approximately 1 million years ago[3]. The resulting products were then screened for polymorphisms by hybridizing them to GeneChip probe arrays (microarrays) specific for the sequences to be interrogated, using a similar strategy as described in our recent surveys for SNPs in the human genome[4,5](Fig. 1).

We identified 2,848 candidate SNP, distributed across 1,755 of the STSs (Tables 1 and 2, Fig. 2). The rate of polymorphism between *M. m. castaneus* and any of the *M. m. domesticus* strains was roughly 1 SNP per 200 bp, with an average of 2,183 informative SNPs in 1,419 STSs in any such pairwise combination. The rate of polymorphisms among *M. m. domesticus* strains was roughly 1 SNP per 1,060 bp, with an average of 430 informative SNPs in 323 STSs in any such pairwise combination. Thus, the number of useful markers added to the previously characterized 6,600 SSLPs (average of 6,300 polymorphic for *M. m. domesticus* strain versus *M. m. castaneus* and 3,300 polymorphic for *M. m. domesticus* strain combinations[1]) is approximately 1,400 for *M. m. domesticus* strain versus *M. m. castaneus* and 350 for typical *M. m. domesticus* strain combinations. The results of this screen, however, indicate that there are roughly three million SNPs available for fine-structure genetic mapping in crosses between inbred strains. The distribution of SNPs observed among the strains corresponded well with previous analysis of the genealogical relationships among inbred strains[6,7], confirming that A/J, C3H/HeJ and BALB/cByJ share a relatively recent history and are more distantly related to C57BL/6J and 129/Sv (Fig. 2*e*).

The SNPs were distributed in a ratio of roughly 2:1 between transitions and transversions, both in the intersubspecific and intraspecific comparisons. The proportion of SNPs occurring at

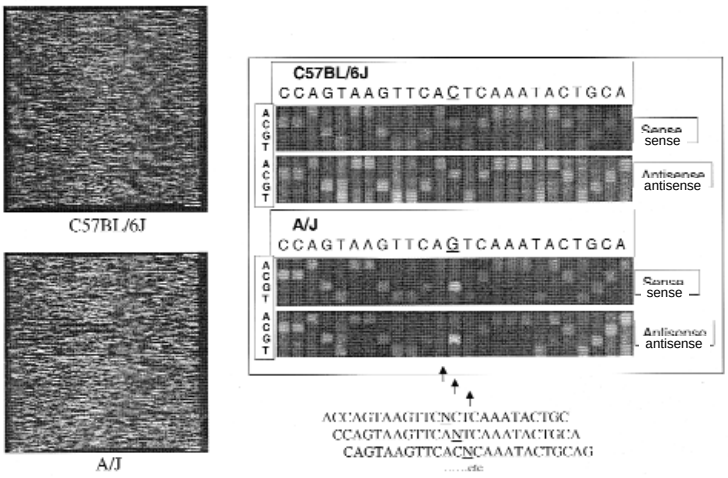

Fig. 1 A SNP in A/J and C57BL/6J strains detected by hybridization to a high-density oligonucleotide array. STSs were PCR amplified, pooled, labelled and hybridized to an array, which was then stained and scanned. Each sequence position was queried by a 25-nt oligonucleotide, where the thirteenth position ('N') was substituted with A, C, G or T in the four rows. Both complementary strands were scanned (arbitrarily marked as sense and antisense). In the example shown, the indicated base is C in C57BL/6J and G in A/J.

[1]*Whitehead Institute/MIT Center for Genome Research, Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA.* [2]*Bristol-Myers Squibb, P.O. Box 5400, Princeton, New Jersey, USA.* [3]*Division of Endocrinology, Children's Hospital, Boston, Massachusetts, USA.* [4]*Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts,USA.* [5]*Affymetrix, Inc., Santa Clara, California, USA.* [6]*Montreal Genome Centre, McGill University Health Centre, Montréal, Québec, Canada.* [7]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to K.L.-T.* ***[author: please provide email address]*** *or E.S.L.* ***[author: please provide email address]***

Table 1 • Results of SNP screen

Variable	All STSs	ESTs[a]	Randoms[b]
no. of STS screened	3,717	769 (20%)	2,948 (80%)
no. of bases screened	441,858	93,681	348,177
no. of candidate SNPs found	2,848	502	2,346
SNP frequency			
CAST/Ei versus lab strain (mean)	1/202	1/321	1/247
lab strain versus lab strain (mean)	1/1,058	1/1,236	1/984
no. of STSs containing SNPs	1,755	325	1,430
% SNPs involving transitions	69%	72%	68%
% SNPs occurring within CpG	19%	24%	17%

[a]ESTs are mostly 3´ UTR. [b]Random genomic loci.

CpG dinucleotides was 19%, representing a 22-fold enrichment for polymorphisms at such relatively rare (approximately 1%) but highly mutable positions. These results are consistent with prior observations for human SNPs (refs 4,5).

To assess the accuracy of the candidate SNPs identified in the DNA microarray–based survey, we randomly selected 180 SNPs and performed direct DNA sequencing. The SNP was confirmed in 173 of 180 cases (96%). To check for SNPs not detected by the chip, 44 STS (5,242 bp) were sequenced in two randomly chosen strains. Only one SNP was detected, corresponding to a false-negative rate of approximately 9%. These validation rates are considerably higher than has been observed in human SNP surveys using DNA microarrays[4,5]. The high accuracy is likely due to the fact that homozygous inbred strains were screened, rather than outbred humans; false positives in SNP surveys typically result from attempts to detect polymorphisms in the heterozygous state.

The chromosomal position of the SNPs must be known for them to be used in genetic linkage studies. Because the STSs were largely derived from our previous mapping efforts, most of the SNP-containing loci (71%) had been assigned positions in the mouse genome by virtue of genetic mapping[1], radiation hybrid mapping[8] or YAC-STS mapping[2]. We are currently mapping the remaining SNPs using RH mapping. The currently available information provides a map of 1,942 mouse SNPs in 1,199 STSs, with a mean spacing of 2.2 cM between SNPs (Table 3 and Fig. 2*f*). For crosses between *M. m. castaneus* and any of the inbred strains, the mean spacing between informative SNPs is 2.5 cM and there are no gaps greater than 20 cM. The map is thus already sufficient for genotyping such intrasubspecific crosses. For crosses among the inbred lab strains, the density of SNPs is somewhat lower. The mean spacing between informative SNPs is 7.1 cM and there is an average of 13 gaps exceeding 20 cM (Fig. 2*d*). This appears to be due primarily to clustering in the locations of the STSs screened, rather than to a deficit of polymorphism in specific regions. The SNPs can be used for rapidly genotyping crosses, followed by supplementary SSLP mapping for refining linkage and filling gaps. The SSLP map, YAC-STS map, RH map and SNP map are all based on the same framework of genetic markers, making it possible to integrate information from all sources. It would be valuable, however, to increase the density of SNPs by two- to threefold to obtain dense coverage throughout the genome.

SNPs were not randomly distributed across the STS loci. Specifically, we rejected the model that SNPs are randomly distributed over STSs by comparing our observed distribution of SNPs/STS with that expected assuming a random (Poisson) distribution given the lengths of the STSs and mean observed polymorphism rate ($P<10^{-35}$). There was an excess of loci with either no SNPs or multiple SNPs. This variation was not due to a difference in the SNP frequency between ESTs and random STSs. Instead, such local variation in polymorphism rate may arise because some loci are inherently more mutable than others. Alternatively, such variation might reflect differences in 'gene history' across loci—that is, the coalescent time to the most recent common ancestor for alleles at a locus can vary greatly across the genome, owing to either natural selection or demographic effects. Loci that trace to a more recent coalescent event have had less time to accumulate new mutations and will therefore show less variation. Recent coalescence may be due to selection in the wild or selection for unusual coat colours, occurring during the breeding of 'fancy' mice in Asia and Europe during the eighteenth and nineteenth centuries[9]. Alternatively, the actual inbreeding of laboratory strains, which are derived from the 'fancy' mice, may have lead to recent coalescence at some loci.

We searched for evidence of differences in mutation rates across loci by examining 16 STSs with no SNPs and 16 STSs with 5 or more SNPs. To test whether the difference in the number of SNPs seen in the two groups of loci reflected inherent differences in mutation rates, we assessed the rate of polymorphism in the two groups in an independent comparison involving unrelated mouse species. The two groups of STSs were resequenced in three distantly related mouse species (*M. spretus*, *M. caroli*, *M. cookii*); they showed no difference in the relative numbers of SNPs per base pair (Table 4). The loci were also resequenced in three wild *M. m. domesticus* and three wild *M. m. castaneus* animals. High levels of polymorphism between *M. m. domesticus* and *M. m. castaneus* were seen in both groups of STSs (Table 4), but the number of newly discovered SNPs was actually higher in the group of STSs that had previously had no SNPs ($P<0.013$). These results suggest that variation in inherent mutation rates across loci is not the major explanation for the variation in the number of SNPs seen at different loci. Instead, we favour the explanation that the deviation from a Poisson distribution is due to variation in gene history across loci, probably resulting from selective breeding in the history of the inbred strains.

Table 2 • Distribution of number of SNPs per STS

	STSs (fraction of base pairs, %)					
SNPs per STS	CAST/Ei only[a]		*M. m. domesticus*[b]		All strains	
0	2,547	(67.0)	2,932	(78.0)	1,962	(50.8)
1	771	(21.2)	565	(16.0)	1,043	(28.6)
2	261	(7.5)	153	(4.3)	451	(12.8)
3	99	(3.1)	49	(1.4)	170	(5.1)
4	30	(0.9)	16	(0.5)	68	(2.1)
5	8	(0.2)	2	(0.05)	18	(0.5)
6	1	(0.03)	0		4	(0.1)
7	0		0		1	(0.03)

[a]All SNPs found between CAST/Ei and any other strain. [b]All SNPs found between any pair of *M. m. domesticus* strains.

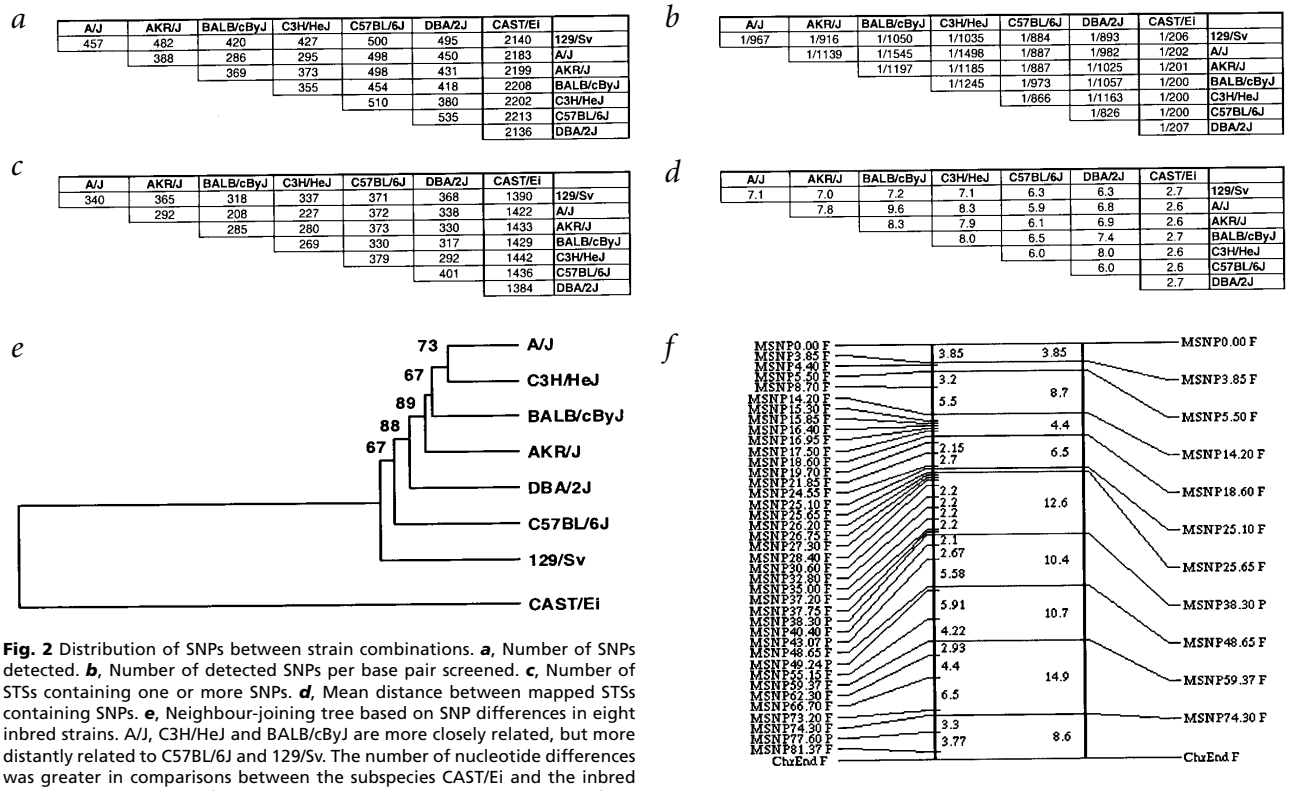

Fig. 2 Distribution of SNPs between strain combinations. ***a***, Number of SNPs detected. ***b***, Number of detected SNPs per base pair screened. ***c***, Number of STSs containing one or more SNPs. ***d***, Mean distance between mapped STSs containing SNPs. ***e***, Neighbour-joining tree based on SNP differences in eight inbred strains. A/J, C3H/HeJ and BALB/cByJ are more closely related, but more distantly related to C57BL/6J and 129/Sv. The number of nucleotide differences was greater in comparisons between the subspecies CAST/Ei and the inbred strains than among any of the inbred strains, and this can be clearly seen from its more distant relationship here. The numbers above the branch points represent the bootstrap confidence limits from 500 trials. ***f***, SNPs on chromosome 5. A map of the location of all SNPs found on chromosome 5 is shown on the left, with the subset of SNPs differing between A/J and C57BL/6J being shown on the right. Distances are shown in cM.

Genome scans in mice require the genotyping of approximately 100 well-spaced, informative markers. To use SNPs in mouse genetic studies, it is necessary to have a system for high-throughput genotyping; we developed such a system for approximately 100 SNPs, based on a procedure that we refer to as length-multiplexed single-base extension (LM–SBE; Fig. 3*a*). The procedure involves first amplifying the SNP-containing loci with flanking primers and then performing an SBE reaction[10] using a primer adjacent to the SNP to incorporate a fluorescently labelled dideoxynucleotide to indicate the SNP allele. To minimize reagent cost and labour time, reactions are performed in a multiplex format. SNPs are amplified in 2 multiplex PCR reactions consisting of 50 loci each. Each of the 2 reactions is then divided into 3 aliquots, which are subjected to multiplex SBE reactions involving approximately 18 loci each. Within each of these six multiplex SBE sets, the primers are designed with 'tails' of varying lengths at the 5′ end. The resulting multiplex SBE reaction can then be scored by electrophoresis on a fluorescence-based DNA sequencer. Because the SBE products are small, it is possible to reload a single gel on an ABI 377 sequencer every 30 minutes and thereby assay approximately 50 genotypes per lane in 2.5 hours (Fig. 3*b*,*c*). Using this procedure, it is possible to genotype a mouse cross within a few days.

In our initial experiments, we discovered a number of issues important to designing successful LM–SBE assays. First, a small proportion of SNPs produced apparently heterozygous genotypes even in inbred strains. This is likely due to concurrent amplification of duplicated but slightly diverged loci, a phenomenon that has also been observed with human SNPs. Second, about 15% of SNPs gave a weak or undetectable signal, suggesting that they competed poorly in the multiplex PCR reaction. When such loci were combined in a separate multiplex pool (to

Table 3 • Average SNP distribution among chromosomes for lab strain and *M. m. castaneus* versus lab strain combinations

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	All
Average for typical lab strain versus lab strain combination																					
No. of SNPs	28.6	22.3	16.9	13.6	17.1	25.1	31.2	11.7	14.6	10.4	30.0	15.4	15.3	11.0	13.9	8.7	8.3	13.8	5.7	2.2	314.9
Unique STS locations	16.7	14.4	9.8	9.1	11.4	13.0	13.0	6.6	7.5	6.3	19.0	7.9	8.9	5.7	8.3	5.4	4.7	5.9	4.3	2.2	179.6
Mean gap	6.8	6.4	6.7	8.3	6.9	4.7	4.8	12.0	9.7	12.2	4.4	8.1	6.4	11.7	7.4	9.9	9.7	6.2	11.9	24.2	7.1
No. of gaps >10 cM	2.9	3.0	1.7	2.6	2.4	1.8	1.9	2.2	1.8	2.2	1.7	1.7	1.7	2.9	2.4	2.0	2.1	0.6	2.9	2.4	42.5
Average for typical *M. m. castaneus* versus lab strain combination																					
Number of SNPs	106.0	111.9	73.7	111.7	91.9	117.3	88.0	68.1	70.9	64.9	111.6	75.3	85.0	62.6	62.6	73.1	35.0	80.7	31.6	24.3	1546.0
Unique STS locations	46.1	37.4	25.6	32.6	34.6	34.6	28.6	25.9	26.9	25.6	39.4	23.0	24.7	25.0	28.4	24.9	17.6	19.3	16.4	15.1	551.6
Mean gap	2.3	2.5	2.6	2.2	2.3	1.7	2.0	2.7	2.3	2.8	2.0	2.6	2.3	2.5	2.1	2.0	2.8	2.0	3.3	4.6	2.5
No. of gaps >10 cM	0.0	0.1	0.6	1.0	0.0	1.0	0.0	1.1	1.4	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	2.0	11.3

Table 4 • SNPs detected in distantly-related mouse species

	spretus-cookii-caroli		*cookii-caroli only*		Wild *domesticus-castaneus*	
	group 1	group 2	group 1	group 2	group 1*	group 2*
No. of SNPs observed	53	50	42	27	0 (22)	41 (5)
Base pairs surveyed	1,061	900	1,431	1,015	1,528	1,057
SNPs/bp observed	1/20	1/18	1/34	1/38	0 (1/69)	1/26 (1/211)

Distribution of SNPs detected in *M. spretus*, *M. cookii*, *M. caroli*, *M. m. domesticus* and *M. m. castaneus* in 16 STS with no SNPs (group 1) and 16 STS with at least 5 SNPs (group 2) in the inbred strains. *Parentheses indicate new SNPs beyond those in the original 8 strains surveyed.

eliminate competition with more strongly amplifying loci), roughly 80% performed well and could thus be used. Third, some SNPs yielded genotypes even in the absence of genomic DNA template. These data resulted from incorporation of fluorescent nucleotides due to 'fold-back' of the SBE primer upon itself (which can occur especially with the longer SBE primers having 5′ tails). Algorithms were developed to design SBE primers that avoid complementarity between the last four bases at the 3′ end and internal sequences within the primer. This eliminated most problems associated with folding back of the primer. It was also necessary to exclude the small number of loci that were false positives and to redesign those assays that failed to amplify in one strain due to the presence of an unknown polymorphism in the primer sequence.

Using these rules, we designed a set of LM–SBE assays for 98 candidate SNPs between A/J and C57BL/6J: 8 loci were excluded for technical reasons (4 proved to be false positives, 2 were heterozygous in both strains, 2 failed to amplify in one strain); 80 loci gave strong signals; and 10 loci amplified weakly. Initially, 13 loci amplified weakly, but 3 gave strong signals after transfer to a separate pool. The 80 loci were genotyped with LM–SBE in a previously analysed backcross of 48 offspring between A/J and C57BL/6J (approximately 3,700 genotypes). All 80 loci mapped to a unique location with a lod score of more than 3 and no double crossovers were observed. In addition, the eight inbred strains as well as an A/J×C57BL/6J F1 heterozygote were genotyped. The genotypes of the inbred strains corresponded with those from the microarray results in all cases and the heterozygote was scored as heterozygous at each locus (accuracy=100%, n=655). Thus, the error rate appears to be low.

Our results provide the foundation for constructing a dense SNP map of the mouse genome which would be suitable for studying arbitrary mouse crosses and for positional cloning of monogenic traits and genetic dissection of polygenic traits. All data (including the sequence and location of the STSs and SNPs, and the genotypes in the eight strains) are freely available on the web site of the Whitehead Institute/MIT Center for Genome Research (http://www.genome.wi.mit.edu/SNP/mouse/). The multiplex genotyping method (LM–SBE) should streamline genome scans in the mouse (as well as in other experimental organisms used in biomedical and agricultural research) and should be useful for genotyping collections of scores of SNPs in human genetics.

Methods

STSs. We chose STSs from the Whitehead Mouse STS database (http://www-genome.wi.mit.edu/cgi-bin/mouse/index). STS sequences containing repetitive sequences or unknown bases were excluded, as were STSs in which the sequence between the primers was shorter than 100 bp. DNA chips were designed corresponding to a final set of 3,884 STSs.

Template amplification for SNP screening. We separately amplified genomic DNA from eight inbred mouse strains (129/Sv, A/J, AKR/J, BALB/cByJ, C3H/HeJ, C57BL/6J, DBA/2J and CAST/Ei), using primers from each of 3,884 STSs. Reactions contained mouse DNA (4 ng), forward and reverse primer (1 μM), $MgCl_2$ (1.5 mM), AmpliTaq Gold (1.225 U; Perkin Elmer) and the buffer supplied by the manufacturer. Reactions were amplified as follows: denaturation 96 °C for 10 min, 35 cycles of 97 °C for

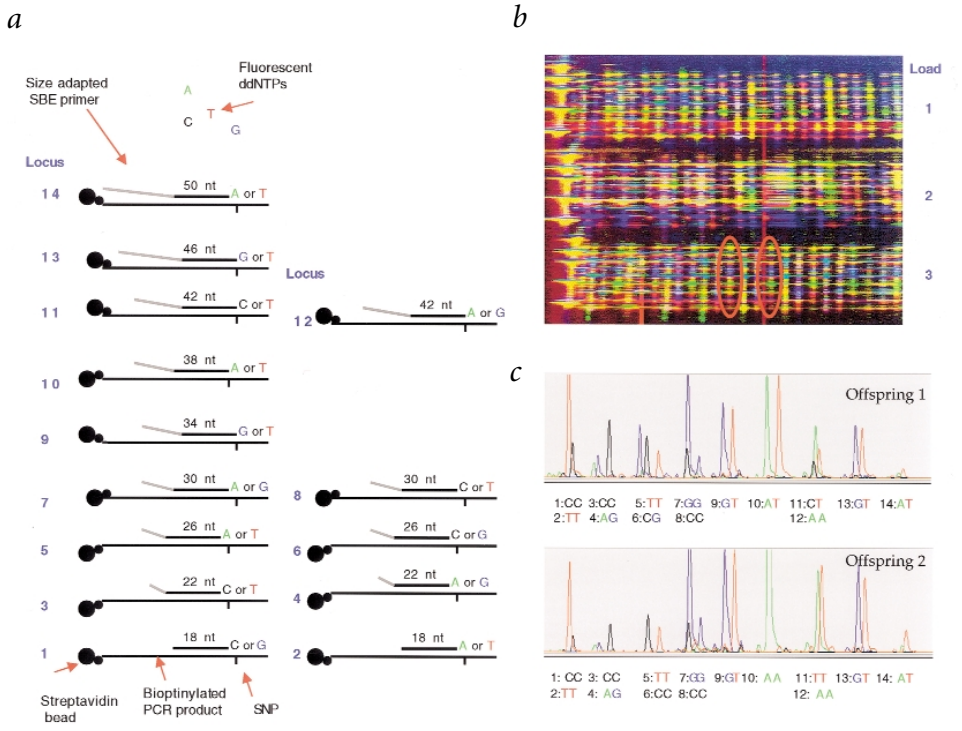

Fig. 3 Principle of gel-based length multiplex single base extension (LM–SBE). ***a***, To generate LM–SBE products, 50 loci are co-amplified in a single reaction volume using two rounds of PCR amplification, with the second round involving common universal primers with a biotin label that allows purification from dNTPs by binding to streptavidin-coated magnetic beads. Each PCR pool is then split into three pools for multiplex SBE reactions, with each pool including up to 18 loci. The SBE primers are chosen to have lengths between 18 and 50 nt at size intervals of 4 nt. The primer is adjusted to the desired length by addition of a piece of neutral sequence to the 5′ end. For each size, two loci with complementary alleles (for example, an A/T SNP and a C/G SNP) can be placed. ***b***, The LM–SBE products are separated and because the products are small, the gel can be loaded three successive times at 30-min intervals with detection completed in 2.5 h. **c**, An automatic genotyping software has been developed whereby the locus is identified by the size and position of the product on the gel and the genotype is determined by the colour.

30 s, 54 °C for 2 min and 72 °C for 2 min and a final extension of 72 °C for 5 min on a MJ Research 384 well Tetrad.

Hybridization and analysis. For loci represented on a single microarray design, the amplification products from a given mouse strain were pooled and hybridized. A total of 18 microarray designs were used. Arrays (Affymetrix) contained 300,000 features and represented 30 kb of sequence. Hybridization was performed at 44 ºC for 14 h, followed by washing in 6×SSPET at RT and stained with streptavidin-phycoerythrin (0.002 mg/ml). The protocol was as described[4]. The results were scored and SNPs identified by using the Ulysses software developed by Affymetrix as described[4,11]. Only 3,717 STSs, with at least 80% of bases called with high confidence, were included in the analyses.

Verification of SNP results. We randomly selected 180 STSs containing at least 1 SNP for verification. The STSs were amplified in a pair of strains that had been identified as containing different alleles. PCR products were generated as above, except that the forward primer was modified to contain the M13 –21 primer site (5′–TGTAAAACGACGGCCAGT–3′) at the 5′ end. Products were then sequenced using fluorescent dye-primer sequencing, using the M13 primer (Perkin Elmer), on an ABI 377 sequencer. We selected 44 STSs (5242 bp) containing no SNPs for verification of false negatives. Two random strains for each STS were amplified and sequenced as above. The sequences were compared for any sequence variation. The false-negative rate was computed from the observation of 1 novel SNP discovered in these 44 STSs using Bayes' theorem.

Phylogeny. To calculate the genealogical relationship among strains, a neighbour-joining tree was constructed from a genetic distance matrix in the eight inbred strains using the program MEGA (Molecular Evolutionary Analysis v.1.01, S. Kumar, K. Tamura and M. Nei). This analysis used the 983 SNPs with the highest confidence scores from the Ulysses program.

Map construction. STSs found to contain at least one high-confidence call SNP (Ulysses software confidence call of "certain" or "likely", call rate≥0.8), were used to prepare the overall SNP map as well as maps for the pair-wise strain combinations. Map distance for each STS was taken from the most recent WI mouse STS database (http://carbon.wi.mit.edu:8000/cgi-bin/mouse/yac_info). Genetic positions were inferred from our previously published maps involving an OB×CAST intercross and Spretus backcross data[1], when available (183 STSs). Genetic position for other markers was inferred by taking the average of the adjacent genetic markers co-localized to the same YAC contig from our YAC/STS-content map of the mouse genome[2]. An additional 44 STSs were positioned by using RH mapping[8]. Any multiple STSs with one or more SNPs falling at the same genetic distance position were regarded as one data point in calculations of marker distribution and density.

SNP distribution per STS. The hypothesis that SNPs are randomly distributed over STSs was tested by comparing the observed distribution of SNPs/STS with that expected assuming a random (Poisson) distribution given the lengths of the STSs and the mean observed polymorphism rate. The observed and expected distributions were compared by comparing the number of STSs with 0, 1, 2, 3, ≥4 STPs, using a χ^2 test with 4 d.f. No significant difference was seen between the distributions generated by SNPs in ESTs and random STSs. No difference was also seen between STSs with different different proportion of bases called by the microarray analysis (80% versus 95% versus 98%), indicating that sequence quality did not contribute to the deviation of our observed distribution from expectation. To check that the presence of a large number of SNPs in an STS did not cause higher inaccuracy on the microarray, all the SNPs in the group of 16 STSs with 5 or more SNPs was subjected to validation by fluorescent sequencing and all the SNPs were confirmed.

Test for variation in mutation rate. The 16 STSs with 5 or more SNPs and 16 STSs with 0 SNPs (matched for number of bp) were chosen and directly sequenced as described above in 3 distantly related mouse species (*M. spretus*, *M. caroli* and *M. cookii*). The number of SNPs found between the three species (*M. caroli*, *M. cookii* and *M. spretus)*, or in a comparison of only the two most distant species (*M. caroli* and *M. cookii*) was compared for the two groups of STSs using a χ^2 test. Only STSs where the amplification was successful for both or all three strains respectively were included in the analysis and the number of base pairs screened was reduced accordingly. Additionally, three wild *M. m. domesticus* and three wild *M. m. castaneus* were screened for SNPs using the same protocol. SNPs not detected in the original screen were used to compare the SNP frequency in the two groups.

Primer design for length-multiplexed single-base extension (SBE). PCR primers were designed using Primer3.0 (WI website software) (melting temperature of 57–63 °C; CG content, 20–80%; primer length of 15–30 nt). Primers were designed to hybridize as close as possible to the SNPs, amplifying a maximum product length of 150 bp. Forward primers had T7 tails at their 5′ ends and reverse primers had T3 tails at their 5′ ends. These T7 and T3 tails were used for secondary amplification. Primer pairs were checked for homology to all amplicons and sorted into pools consisting of up to 50 primer pairs. Pools were selected to avoid homologous loci being present in the same pool.

Primer design for SBE. Primers were designed to have a melting temperature of 50–60 °C, a length of 15–30 bp, to terminate on the base 5′ to the SNP and to avoid containing any neighbouring SNPs. The primers were sorted into pools consisting of up to nine complementary SNP pairs (for example, a C/G SNP was paired with an A/T SNP and a C/T SNP with an A/G SNP). For each such pair of SNPs, the length of the primers was adjusted to a distinct size (18, 22, 26, 30, 34, 38, 42, 46 and 50 nt, respectively), by addition of a 'neutral' sequence to the 5′ end. The suitable size of neutral sequence was taken from the 5′ end of either 5′–AACTGACTAAACTAGGTGCCACGTCGT GAAAGTCTGACAA–3′ or 5′–ATGCTCAGACACAATTAGCG CGACCCT TAATCCTTAGGTA–3′, both of which are random sequences that yielded no matches when compared with the NCBI nonredundant database. Primers in which four or more bases at the 3′ end were complementary to another part of the primer were discarded or redesigned to avoid artefacts due to primer 'foldback'. Primers were redesigned either by attaching the alternative segment of 'neutral' sequence or by using a primer on the opposite strand.

PCR amplification. Loci were subjected to two rounds of PCR amplification. First, mouse genomic DNA (5 ng), a pool of 50 primer pairs (0.1 μM), $MgCl_2$ (5 mM), dNTPs (0.5 mM), Amplitaq Gold (2.5 U; Perkin Elmer) and the supplied buffer were amplified in 12.5 μl. Samples were denatured for 9 min at 95 °C followed by 31 cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s and a final extension of 72 °C for 5 min. An aliquot consisting of 3 μl of the primary amplification product was transferred to a new plate and subjected to a second round of PCR amplification with biotinylated-T7 and biotinylated-T3 primers (0.8 μM), $MgCl_2$ (4 mM), dNTPs (0.4 mM), *Taq* (3 U) and the supplied buffer in 30 μl for 8 min at 95 °C followed by 32 cycles of 95 °C for 30 s, 55 °C for 1 min 30 s and 72 °C for 30 s and a final extension of 72 °C for 7 min.

SBE amplification. We purified 7 μl PCR product from unincorporated dNTPs by incubation with streptavidin-coated Dynabeads (20 μg; Dynal) in 0.5×TE, NaCl (1 M) and washed three times in 70% ethanol, Tris (3mM), EDTA (0.3 mM, pH 8.0). Beads with attached products were then subjected to multiplex SBE reactions with JOE-ddATP (0.12 μM), TAMRA-ddCTP (0.12 μM), FAM-ddGTP (0.12 μM) and ROX-ddTTP (0.60 μM; NEN DuPont), a pool of up to 18 LM–SBE primers at a concentration of 0.25 μM and Thermosequenase (0.5 U; Amersham) in Tris (50 mM, pH 9.5), $MgCl_2$ (2mM) in a volume of 20 μl. Thirty cycles of 96 °C for 30 s, 50 °C for 15 s and 60 °C for 1 min were performed.

To remove excess ddNTPs, SBE products were centrifuged in 96-well gel filtration blocks (Edge Biosystems). The eluates were dried and resuspended in 4 μl formamide and assayed by electrophoresis on an ABI377 sequencer, using a 10% denaturing polyacrylamide gel at 200 W for 2.5 h. Samples were loaded three consecutive times at 0, 0.5 and 1 h. A size standard, consisting of 1 μl of a 2 nM mixture JOE-, TAMRA-, FAM- and ROX-labelled fragments of 14, 21, 27, 37 and 48 nt, was loaded in the first and last lanes of the gel.

Genotyping. An algorithm was developed to identify loci based on fragment position on the gel and to score the alleles based on colour. The cross-sectional intensity of different segments of the gel was used to identify the position of each successive load. After background subtraction and colour separation, peaks were sorted into bins according to product size by com-

parison to the size standard. For each SNP marker, information about the expected alleles, product size, lane and DNA sample was used to assign the peak data to individual loci. The peak intensity ratio of the two expected alleles was used to determine the genotype. The peaks for a given marker were first analysed to determine the average intensity for each allele and the average intensity ratio for a heterozygote. An algorithm was then used to assign a final genotype of each sample, given these values.

Acknowledgements

We thank H. Nguyen and R. Steen for assistance in RH mapping; the HTS group at Affymetrix Inc. for assistance with chip hybridizations; and J. Singer for providing the backcross DNAs. This research was supported by a research contract from Bristol-Myers Squibb, Millennium Pharmaceuticals Inc. and Affymetrix (E.S.L. and T.J.H.) and the National Institute of Health (HG01806). K.L. is the recipient of a scholarship from the Swedish Institute (4478/1998). J.N.H. is the recipient of a Howard Hughes Medical Institute Postdoctoral Fellowship for Physicians. T.J.H. is a recipient of a Clinician-Scientist award from the Medical Research Council of Canada.

Received 8 December 1999; accepted 25 February 2000.

1. Dietrich, W.F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380**, 149–152 (1996).
2. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22**, 388–393 (1999).
3. Ferris, K.D., Sage, R.D., Prager, E.M., Titte, U. & Wilson, A.C. Mitochondrial DNA evolution in mice. *Genetics* **105**, 681–721 (1983).
4. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
5. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
6. Atchley, W.R. & Fitch, W. Genetic affinities of inbred mouse strains of uncertain origin. *Mol. Biol. Evol.* **10**, 1150–1169 (1993).
7. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nature Genet.* **24**, 23–25 (2000).
8. Van Etten, W.J. *et al.* Radiation hybrid map of the mouse genome. *Nature Genet.* **22**, 384–387 (1999).
9. Sage R.D. in *The Mouse in Biomedical Research* (eds Foster, H.L., Small, J.D. & Fox, J.G.) 40–90 (Academic, New York, 1981).
10. Syvanen, A.C. From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymoprhisms. *Hum. Mutat.* **13**, 1–10 (1999).
11. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).