The mosaic structure of variation in the laboratory mouse genome

Claire M. Wade*†, Edward J. Kulbokas III*, Andrew W. Kirby*, Michael C. Zody*, James C. Mullikin‡, Eric S. Lander*, Kerstin Lindblad-Toh*§ & Mark J. Daly*§

* *Whitehead Institute for Biomedical Research and Whitehead/MIT Center for Genome Research, 9 Cambridge Center, Cambridge, Massachusetts 02139, USA*
† *School of Veterinary Science, The University of Queensland, Queensland 4072, Australia*
‡ *The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1RQ, UK*
§ *These authors contributed equally to this work*

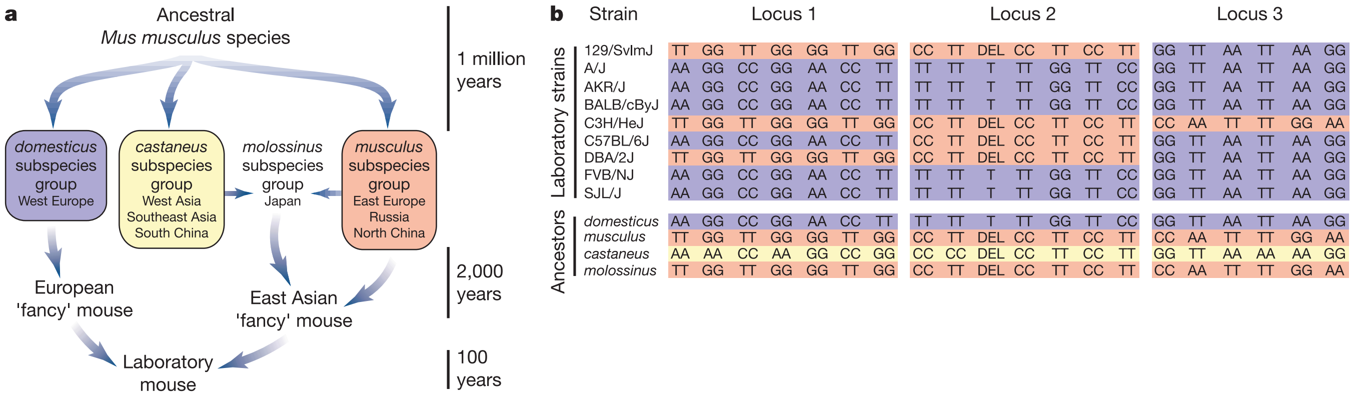
Most inbred laboratory mouse strains are known to have originated from a mixed but limited founder population in a few laboratories[1,2]. However, the effect of this breeding history on patterns of genetic variation among these strains and the implications for their use are not well understood. Here we present an analysis of the fine structure of variation in the mouse genome, using single nucleotide polymorphisms (SNPs). When the recently assembled genome sequence from the C57BL/6J strain[3] is aligned with sample sequence from other strains, we observe long segments of either extremely high (~40 SNPs per 10 kb) or extremely low (~0.5 SNPs per 10 kb) polymorphism rates. In all strain-to-strain comparisons examined, only one-third of the genome falls into long regions (averaging >1 Mb) of a high SNP rate, consistent with estimated divergence rates between *Mus musculus domesticus* and either *M. m. musculus* or *M. m. castaneus*. These data suggest that the genomes of these inbred strains are mosaics with the vast majority of segments derived from *domesticus* and *musculus* sources. These observations have important implications for the design and interpretation of positional cloning experiments.

Patterns of genetic variation provide insight into the evolutionary history of a species and define the complexity of mapping phenotypes in that organism. The commonly used inbred laboratory strains of mice constitute the primary mammalian model system and are an integral component of medical genetic research. These inbred laboratory strains were predominantly derived in the early twentieth century from mouse breeders who originally bred 'fancy' mice (for unusual coat colours and behaviours) as a hobby. Many of the most commonly used strains trace their origins to W. Castle's laboratory at Harvard University and even more strains originate from his supplier A. Lathrop of Granby, Massachusetts. Although these mice are generally thought to reflect predominantly the *M. m. domesticus* subspecies, there are some historical contributions from 'fancy' mice bred in Japan and China[1,2] (Fig. 1a). As a result, we would expect to see in these strains recognizable contributions from several other subspecies such as *M. m. musculus* (and possibly *M. m. castaneus* through the hybrid *M. m. molossinus*). Indeed, most of these inbred laboratory strains carry a *M. m. musculus* Y chromosome[4] (previous work had shown that most carry *M. m. domesticus* mitochondrial DNA[5,6]).

The genomes of these strains were predicted to be a 'mosaic' of regions with origins in the different subspecies[7], but a clear description of this variation has remained elusive, largely owing to a lack of high-resolution data across the genome. Here, we describe the fine structure of variation among inbred laboratory strains, first through an analysis of available finished sequence and then through more comprehensive analysis of genome-wide shotgun SNP discovery data produced by the Mouse Genome Sequencing Consortium[3].

To study the nature of genetic variation among strains, we wished to compare the recent C57BL/6J (henceforth B6) draft genome assembly, MGSCv3 (ref. 3), with available finished sequence from other strains. We first, however, performed a control comparison of 205 finished B6 sequences (obtained from GenBank) totalling 49 megabases (Mb) with the draft genome assembly. Sequences were compared using SSAHA-SNP software[8] to identify differences in high-quality matches. We detected 0.9 candidate SNPs per 10,000 base pairs (bp). The occurrence of candidate SNPs in this intrastrain (B6–B6) comparison suggests that these SNPs result from sequencing errors (rather than an unexpectedly high mutation rate within an ostensibly inbred strain). To clarify this, we re-sequenced 15 such candidate SNPs in mice from seven distinct B6 founder lines from the Jackson Laboratories as well as the DNA stock used in creating the draft genome sequence at the Whitehead Institute. None of the 15 apparent polymorphisms was confirmed—indicating that they are indeed sequencing artefacts. All 15 cases seem to result from errors in the MGSCv3 sequence, indicating a low error rate (0.9 per 10 kilobases, kb) that is actually below the target accuracy rate for 'finished sequence' of less than one error per 10 kb.

We then set out to perform the same comparison with finished



	Strain	Locus 1							Locus 2							Locus 3					
Laboratory strains	129/SvImJ	TT	GG	TT	GG	GG	TT	GG	CC	TT	DEL	CC	TT	CC	TT	GG	TT	AA	TT	AA	GG
	A/J	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
	AKR/J	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
	BALB/cByJ	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
	C3H/HeJ	TT	GG	TT	GG	GG	TT	GG	CC	TT	DEL	CC	TT	CC	TT	CC	AA	TT	TT	GG	AA
	C57BL/6J	AA	GG	CC	GG	AA	CC	TT	CC	TT	DEL	CC	TT	CC	TT	GG	TT	AA	TT	AA	GG
	DBA/2J	TT	GG	TT	GG	GG	TT	GG	CC	TT	DEL	CC	TT	CC	TT	GG	TT	AA	TT	AA	GG
	FVB/NJ	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
	SJL/J	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
Ancestors	*domesticus*	AA	GG	CC	GG	AA	CC	TT	TT	TT	T	TT	GG	TT	CC	GG	TT	AA	TT	AA	GG
	musculus	TT	GG	TT	GG	GG	TT	GG	CC	TT	DEL	CC	TT	CC	TT	CC	AA	TT	TT	GG	AA
	castaneus	AA	AA	CC	AA	GG	CC	GG	CC	CC	DEL	CC	TT	CC	TT	GG	TT	AA	AA	AA	GG
	molossinus	TT	GG	TT	GG	GG	TT	GG	CC	TT	DEL	CC	TT	CC	TT	CC	AA	TT	TT	GG	AA

Figure 1 History of the inbred laboratory mouse and the resulting patterns of genetic variation. **a**, Genetic contributions are expected from both European (*domesticus*) and Asian (*musculus* and *castaneus*) subspecies because the mice were derived from Asian or European 'fancy' stocks collected by a few mouse suppliers. **b**, Most loci show two haplotypes and match wild-derived representatives of *domesticus* (WSB/Ei) (purple), *musculus* (CZECHII/Ei) (orange) or *molossinus* (MOLF/Ei) but not *castaneus* (CAST/Ei) (yellow). Chromosome 14:121 Mb (locus 1), chromosome 6:139 Mb (locus 2) and chromosome 2:107 Mb (locus 3) are represented by polymorphisms from 1-kb windows re-sequenced in nine strains.

sequence from strain 129 (the strain with the most finished sequence available after B6). A total of 59 finished sequence contigs from strain 129 (various substrains and clone lines including 129, 129/Sv, 129SvEvTac, 129Ola, RPCI_21, RPCI_22, MGS1 and CITB_CJ7) ranging in size from 27 kb to 1.7 Mb and totalling 17 Mb, were obtained from GenBank and compared with the B6 genome assembly as described above. Considering only the first 50 kb of each sequence (to normalize the sequence contributions of different genomic regions in the comparison), a distinct bimodality of polymorphism rate is revealed (Fig. 2). The observed distribution of SNP rates in 50-kb segments is strikingly nonrandom when compared with the expected distribution, assuming a constant polymorphism rate (14 SNPs per 10 kb, the overall observed rate), consistent with the observations of nonrandomness made in a previous small-scale study[9]. Of the regions examined, 39% (±6.3%) fall into the higher mode of this distribution (more than 10 SNPs per 10 kb) and have a mean of 36 SNPs per 10 kb. Those in the lower mode (less than 5 SNPs per 10 kb) have an average polymorphism rate of 1.6 SNPs per 10 kb, suggesting that these 58% (±6.4%) of genomic segments share a much more recent co-ancestry between B6 and 129. Because of existing evidence of considerable diversity caused by contamination of various sub-strains of 129 (refs 10–12) we separated individual substrains and clone libraries (CITB-CJ7-129Sv, RPCI21-129S6/SvEvTac, RPCI22-129S6/SvEvTac and MGS1-129/Ola) but see the same general bimodality in each when compared with B6.

As in the B6–B6 comparison, we selected and re-sequenced 15 putative 129–B6 SNPs from distinct genomic regions with very low SNP rates. Only four of these 15 were determined to be true polymorphisms. This suggests that there are a few true polymorphisms in these low-SNP-rate regions, but that the actual rate is probably closer to one per 20,000 bp (~0.5 SNPs per 10 kb). By contrast, 79 of 80 candidate B6–129 SNPs selected from regions with a high SNP rate were found to be true SNPs by validation through re-sequencing. Because nearly all SNPs (>95%) detected in this study come from regions of high SNP rate, this data indicates an overall high validation rate for SNPs discovered in this survey.

Low-SNP-rate regions in this two-strain comparison often appear to persist over considerable distances in the finished sequence (most often the full length of the sequence). By examining those finished sequences that have a low or absent SNP rate in the first 50 kb, we find only three of 32 (9%) with a high SNP rate (>10 SNPs per 10 kb) in the second 50 kb of the sequence and four of 28 (14%) with a high SNP rate in the third 50-kb segment (both $P < 0.01$ when compared with the overall average of 37% of 50-kb segments that have a high SNP rate). Although the number of finished sequences longer than 150 kb was limited, these data suggest that segments of high or low polymorphism rate between strains extend over long genomic distances. In a few cases, long finished sequences contain an overall intermediate SNP rate, which reflects a sharp transition between a low and high SNP rate region (Fig. 3); this transition emphasizes the bimodal nature of polymorphism rate in these genome comparisons. Only a handful of finished sequences showed such transitions, precluding genome-wide generalizations, but qualitative observation of the transitions suggests that they are sharp, rather than diffuse, in character.

Having discovered a strong bimodality distinguishing regions of low and high polymorphism rate over long genomic segments, we then extended the observations to the entire genome and to several strains. We used whole genome shotgun (WGS) reads for three strains of inbred mice to examine patterns of variation on a genome-wide scale; the strains were: 129S1/SvImJ (Jackson Laboratories stock 002488, henceforth 129) with 119,232 reads, C3H/HeJ (stock 000659, henceforth C3H) with 68,160 reads, and BALB/cByJ (stock 001026, henceforth BALB) with 38,400 reads. SNP detection was carried out once more with SSAHA-SNP[8], using the WGS reads that could be uniquely placed on the B6 assembly[3]. A total of 79,269 candidate SNPs were found across all strains. The observed distribution in each strain follows the general pattern seen when using the finished sequences, with many long regions having either no SNPs (for example, a 26-Mb region on chromosome X contained 131 consecutively placed reads with no SNPs) or many SNPs (for example, a 5-Mb segment of chromosome 12 contained 36 consecutively placed reads with one or more candidate SNPs).

We would not expect regions of low-SNP rate to contain

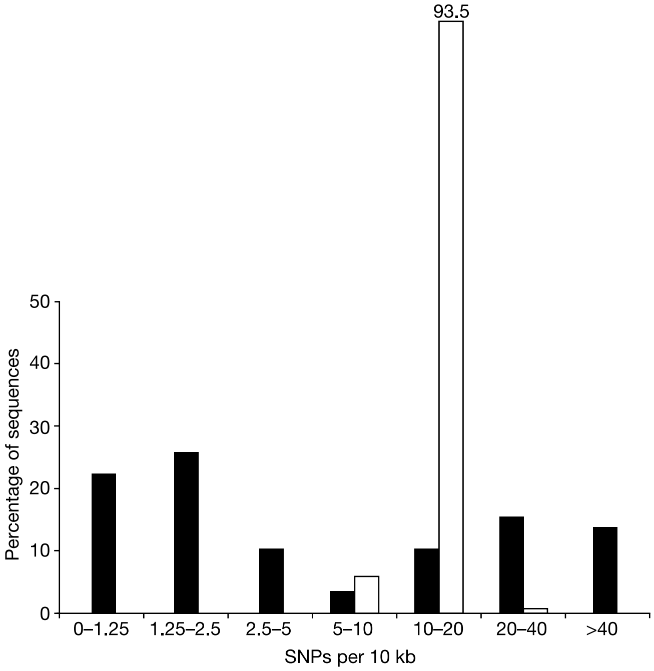


Figure 2 SNP frequency in the first 50 kb of finished sequences for strains 129 as compared to the B6 assembly. We note the bimodal distribution of observed SNPs (black) and the difference from the expected Poisson distribution (white) with an average of 14 SNPs per 10 kb (the observed rate across the entire data set). This indicates a very non-uniform distribution of SNPs in the genome with the presence of regions with a high SNP rate (~45 SNPs per 10 kb) and regions with a low SNP rate (~0.5 SNPs per 10 kb).

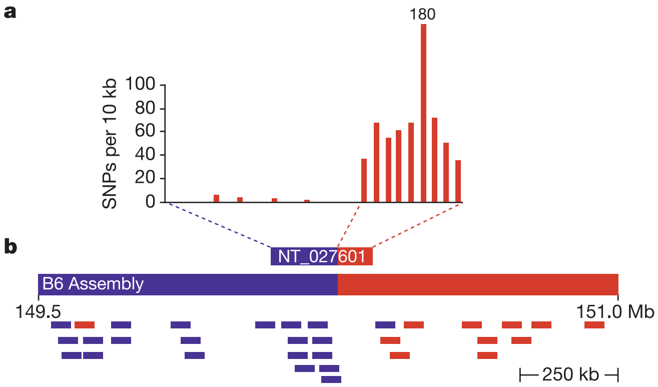


Figure 3 A clear transition from low to high SNP rate between strains 129 and B6 in a 1-Mb region at the telomeric end of chromosome 5. **a**, Finished sequence NT_027601 when aligned against the B6 whole genome assembly demonstrates a low-SNP-rate region (the first 161 kb) followed by a high-SNP-rate region (the last 101 kb) and an abrupt transition between SNP rate regions. **b**, Whole genome shotgun reads from the same genomic region provide evidence of the same well-defined transition point. Bars correspond to 500-bp reads containing one or more SNPs (red), or no SNPs (blue) when compared with the B6 assembly.

exclusively reads with no candidate SNPs, if only because of the inherent sequencing error rate. Similarly, regions of high SNP rate will certainly contain individual reads with no SNPs because the average rate in such regions is only about two SNPs per 500-bp read. We therefore developed a hidden Markov model (HMM)[13] to use the WGS read data to parse the genome. The approach assumes that there are two types of genomic regions (of low SNP rate and high SNP rate, as suggested by the finished sequence analysis) and uses all reads on a chromosome to identify those genomic segments that have an unambiguously high or low SNP rate. In the 129–B6 comparison, more than 90% of the genome is contained in such unambiguous segments (high or low) and 50% of the genome is contained in defined segments of greater than 2.0 Mb in length (the N50 value). Although shotgun sampling may occasionally miss short segments and will not capture the discrete nature of transitions (as in Fig. 3), the available 70,000 placed reads allow an accurate estimate of the N50 segment size (Supplementary Fig. 1). With a random distribution of transitions, an N50 value of 2.0 Mb suggests an average block size of 1.2 Mb. SNP-poor ('low-SNP') regions account for 67% of the genome in the comparison between strains 129 and B6. The overall SNP rate found in high-SNP regions is 45 per 10 kb, and the rate in low-SNP regions is 1.0 per 10 kb. The fraction of genome in low- and high-SNP-rate regions and the SNP rate within these regions are very similar for the C3H and BALB comparisons to B6 and attests to the common origins of all four strains (Supplementary Fig. 1).

Such long regions of either recent or distant co-ancestry should, once identified, be clearly distinguishable from one another using data from existing SSLP maps. Regions identified as 'high' or 'low' in the HMM analysis were cross-referenced with SSLP data from the MIT genetic map[14] (using 2,605 SSLPs that were uniquely mapped onto the genome assembly[3] and for which SSLP allele size information was available for B6, C3H and BALB). In pairwise comparisons between B6 and either C3H or BALB, SSLP polymorphism rates correlate strongly with the low- and high-SNP-rate regions assigned by the HMM: 31% (low) versus 75% (high). The fact that SSLPs are polymorphic in SNP-poor regions attests to their high mutability even in regions of recent co-ancestry and to the difficulty of recovering ancestral patterns directly from SSLP data.

To examine haplotype variation across several inbred laboratory strains, we selected 27 genomic segments (500–1,000 bp in length); nine from each of the three strains (C3H, 129 and BALB) in which the shotgun SNP discovery had revealed 4–15 candidate SNPs between the sequenced strain and B6. These segments were then completely sequenced in a total of nine inbred strains: B6, 129, C3H, BALB, A/J, AKR/J, DBA/2J, FVB/NJ and SJL/J (examples in Fig. 1b). In 22 of the 27 regions, only two different haplotypes were observed—emphasizing that the founder population from which these strains have been derived was extremely limited. The ancestry of these haplotypes was classified by sequencing representative wild-derived inbred lines of *M. m. domesticus* (WSB/Ei), *M. m. musculus* (CZECHII/Ei), *M. m. molossinus* (MOLF/Ei), and *M. m. castaneus* (CAST/Ei). Even in this extremely limited sampling of the ancestral populations, 67% of the haplotypes can be readily ascribed to the *domesticus* line and another 21% appears to derive from the Asian mice (primarily *musculus* and *molossinus,* which are usually identical). Despite comparing the inbred strains to only a single isolate from each heterogeneous ancestral population, this straightforward experiment appears to support the historical expectation (Fig. 1) completely. To examine the extent of these ancestral haplotypes, we sequenced these same strains at eight pairs of 1-kb windows separated by 100–200 kb that occurred in long high-SNP-rate segments identified in the original B6–129 finished sequence comparison. In 49 of 56 cases, identity in the seven other inbred strains (excluding B6 and 129) to either B6 or 129 persisted over the entire segment, providing an estimate (through linkage disequilibrium) consistent with 1-Mb blocks. In the other cases, the identity to B6 or 129 did not persist between segments, suggesting an historical recombination had taken place during the creation of the inbred strains.

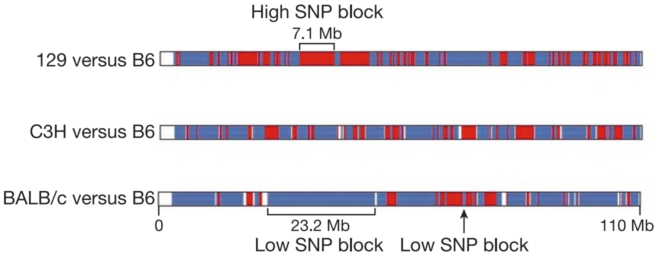
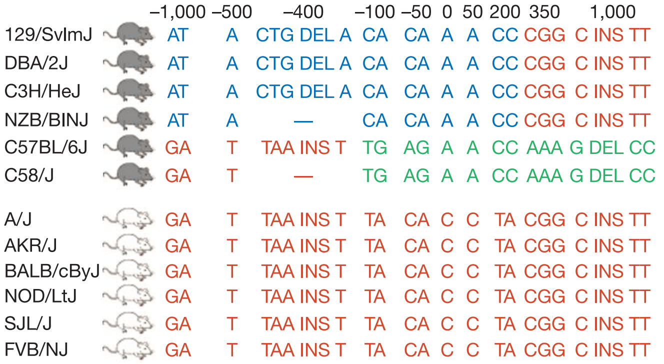


Figure 4 Mosaic structure of the mouse genome. Chromosome 15 comparisons between 129S1/SvImJ, C3H/HeJ and BALB/cByJ with C57BL/6J reveal regions of low SNP rate (∼0.5 SNPs per 10 kb, blue) and regions of high SNP rate (∼45 SNPs per 10 kb, red). Half of the genome is contained in segments (low or high) of greater than 2 Mb in length. Low-SNP-rate regions define roughly two-thirds of each genome comparison.

The present observations of juxtaposed regions of consistently high and low diversity reveal the impact of breeding history on diversity in the mouse genome. In comparisons of any inbred laboratory strain with B6, we find two-thirds of the genome to have a very low polymorphism rate, indicating genomic regions in which the two strains share a very recent common subspecies origin (usually both *domesticus* or both *musculus*) (Fig. 4). By contrast, the remaining one-third of the genome in these comparisons shows considerable divergence (1 SNP every 200 bp), indicating distant ancestry consistent with a previous measure of genome-wide diversity between the laboratory strains and *M. m. castaneus*[9] (*M. m. musculus* and *M. m. castaneus* are roughly equidistant from *M. m. domesticus*[2,15]), suggesting that these represent regions in which the two strains inherited the region from different subspecies (most often, one *domesticus* and the other *musculus*). Finally, a more detailed look at individual regions in these inbred laboratory strains suggests that these patterns represent the effects of recent breeding practice that has created an admixture of two major ancestral sources (Fig. 1).

The segmented nature of these genomes and the lack of heterogeneity within the segments provides important insights into the interpretation of quantitative trait mapping experiments using inbred strains and offers the potential for the acceleration of



	–1,000	–500	–400	–100	–50	0	50	200	350	1,000
129/SvImJ	AT	A	CTG DEL A	CA	CA	A	A	CC	CGG	C INS TT
DBA/2J	AT	A	CTG DEL A	CA	CA	A	A	CC	CGG	C INS TT
C3H/HeJ	AT	A	CTG DEL A	CA	CA	A	A	CC	CGG	C INS TT
NZB/BINJ	AT	A	—	CA	CA	A	A	CC	CGG	C INS TT
C57BL/6J	GA	T	TAA INS T	TG	AG	A	A	CC	AAA	G DEL CC
C58/J	GA	T	—	TG	AG	A	A	CC	AAA	G DEL CC
A/J	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT
AKR/J	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT
BALB/cByJ	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT
NOD/LtJ	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT
SJL/J	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT
FVB/NJ	GA	T	TAA INS T	TA	CA	C	C	TA	CGG	C INS TT

Figure 5 Association between a single haplotype and the albinism phenotype caused by a mutation at the tyrosinase locus[20]. Columns show SNPs discovered in ten 500-bp assays with positions (kb) relative to the centre of the genomic segment containing the gene (GenBank accession GI:12852585). The causal mutation (Cys103Ser) is located at +32.6 kb. The association between phenotype and ancestral haplotype for 12 strains would be sufficient to identify a haplotype background and 'critical region' of ∼500 kb (including the assays from −100 kb before Tyr to 200 kb after Tyr) likely to contain the albinism mutation.

subsequent positional cloning. For many phenotypes, positional cloning after initial quantitative trait locus (QTL) mapping should now focus on subregions where the parental strains are recognized to differ, as more than 95% of the total genetic variation will be found in that one-third of the genome. This can be much more powerful in cases where a locus has been mapped in several strain combinations. Several parental strains could then be simultaneously compared to discover an even smaller critical region. In addition, the observation of the same QTL in several strain combinations makes it even more likely that the QTL is due to ancestral differences common to many strains rather than a rare mutation specific to a single strain. For phenotypes not subject to strong selection (for example, otherwise phenotypically benign modifiers of knockouts, transgenes and spontaneously arising mutations), we should be able to use the parental phenotype data of many inbred strains to correlate ancestral patterns directly with phenotype.

As an example, we constructed haplotypes from 12 inbred strains across 2 Mb surrounding the *tyrosinase* gene on chromosome 7. Figure 5 shows the clear correlation of ancestral haplotype and the phenotype (albinism) due to a mutation in this gene. Without knowing the gene or functional polymorphism, the pattern of ancestral haplotypes in only 12 strains would suffice to identify the short genomic segment (~500 kb) surrounding the gene as the region likely to contain the albinism locus had we initially identified the chromosome 7 region in a traditional mapping cross. A much larger panel of strains could reduce the critical region further and would provide much stronger statistical evidence for association of genomic segment to phenotype. In effect, existing laboratory strains serve as a natural set of recombinant inbred lines—although with more than ten times as many breaks as in conventional recombinant inbred strains. A detailed understanding of the ancestral origins of haplotypes should allow researchers to use the entire panel of inbred laboratory strains in this fashion.

A genome-wide characterization of the segmented genetic variation of the 50–100 commonly used inbred mouse strains would thus enrich the utility of large-scale phenotyping efforts currently underway[16] (such as the Mouse Phenome Project[17], http://www.jax.org/phenome). As shown, it can in some cases allow *de novo* mapping of major loci to small candidate regions simply on the basis of correlation between ancestral segment identity and phenotype. In addition, it will significantly improve our ability to recognize appropriate strain combinations for confirmatory mapping of QTLs (those where genetic ancestry as well as phenotype differs) and for mapping modifiers (those with the greatest phenotypic differences among strains known to share ancestral identity at major loci). A haplotype map offering a precise picture of the history of each individual genomic region is likely to become a valuable complement to traditional approaches in genetic mapping. □

Methods

Finished sequences

Single or merged finished bacterial or phage artificial chromosome sequences for *Mus musculus* were retrieved from the National Center for Biotechnology Information web site (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/NTStrainList). Those of either solely strain 129 and its derivatives, or solely strain B6 were extracted from the Entrez database at NCBI.

SNP discovery

The MGSCv3 C57BL/6J whole genome shotgun assembly, generated from over 40 million paired-end reads and assembled using Arachne[18], was compared with finished sequences (incorporating quality scores) in 1,000-bp windows using SSAHA-SNP[8]. Uniquely placed sequences with a minimum base quality score of Phred 23 (ref. 19) and a maximum of twenty SNPs per 1,000-bp window were included.

SNP rate

The SNP rate for each 50-kb segment of finished sequence was assessed by taking the total of SNPs divided by the total of high-quality bases examined. Individual finished sequences were grouped by SNP rate (0–1.25, 1.25–2.5, 2.5–5, 5–10, 10–20, 20–40, >40 SNPs per 10 kb) within each strain and compared to a Poisson expectation based on the overall observed SNP rate.

SNP validation

Fifteen putative SNPs from the C57BL/6J assembly versus finished C57BL/6J sequence discovery were re-sequenced in 17 DNAs. The DNAs represented males and females from seven different C57BL/6J founderlines kept at the Jackson Laboratories and the C57BL/6J DNA remaining at the Whitehead Institute after construction of the WGS libraries. The SNP loci were amplified and sequenced using fluorescent M13 dye-primer sequencing on an ABI 377 Sequencer (Applied Biosystems; for details see Supplementary Information).

Validation as above was carried out for putative SNPs from low-rate (15 SNPs) and high-rate (80 SNPs) regions detected in the B6–129 finished sequence comparison.

Whole-genome shotgun read analysis

119,232, 68,160 and 38,400 random shotgun reads (paired-ends from 4-kb plasmids) for strains 129S1/SvImJ, C3H/HeJ and BALB/cByJ respectively were found as part of the Mouse Genome Sequencing Consortium[3] SNP discovery. Vector- and quality-trimmed reads were assessed for 20-base windows having an average quality score below Phred 20, implying an error rate of lower than 1 base miscalled per 100 bases[19] . The longest block of sequence between any two windows was passed onto SNP discovery if that block exceeded 250 bp. SNP detection was by SSAHA-SNP[8] using 500-bp windows and a maximum of ten SNPs per window. The number of reads passing our quality thresholds, SSAHA-SNP, and post-SSAHA filtering for repetitive sequence were 67,974, 34,949 and 19,686 for strains 129S1/SvImJ, C3H/HeJ and BALB/cByJ respectively.

Determination of SNP low- and high-rate regions

A hidden Markov model (HMM) fitted a two-state model of low polymorphism rate (low) and high polymorphism rate (high) using the forward–backward algorithm[13]. Prior estimates for the observational and transitional probabilities used in the HMM were obtained by computational observation of frames of three consecutive reads on either side of the read to be evaluated.

Regions of high and low polymorphism were defined as continuous segments estimated to have greater than 0.95 or less than 0.05 probability of being 'high' according to the HMM, respectively. Block sizes were the distances between the first base for the first read and the last base for the final read where consecutive reads met 'low' or 'high' criteria.

Re-sequencing of multiple strains for haplotype studies

To determine the number and ancestry of haplotypes among inbred strains a panel of nine inbred strains (C57BL/6J, 129S1/SvImJ, C3H/HeJ, BALB/cByJ, A/J, AKR/J, DBA/2J, FVB/NJ and SJL/J) and four wild-derived inbred ancestral strains (*M. m. domesticus* WSB/Ei, *M. m. musculus* CZECHII/Ei, *M. m. molossinus* MOLF/Ei and *M. m. castaneus* CAST/Ei) were re-sequenced as above (see 'SNP validation' section). A total of 27 random 1-kb regions (nine from the 129 strain, nine from C3H and nine from BALB/c discovery) containing 4–11 SNPs were re-sequenced and searched for SNPs. Haplotypes were inferred by inspection at each locus. Haplotypes varying by only one SNP between inbred strains were counted as the same haplotype. The number of haplotypes at each locus was tallied and compared with the haplotypes in the ancestral mice.

Eight paired 1-kb regions derived from the finished 129 sequence comparison were re-sequenced in the same strains. The paired regions all resided 100–200 kb apart in high-SNP regions. Haplotypes were scored and counted as above. The observation of 87% of paired loci separated by an average of 150 kb with concordant haplotype suggests (by Poisson) an average block size of 1.1 Mb.

Raw data from SNP discovery and re-sequencing is available at http://www-genome.wi.mit.edu/~mjdaly.

Received 12 September; accepted 28 October 2002; doi:10.1038/nature01252.

1. Silver, L. M. *Mouse Genetics* (Oxford Univ. Press, New York, 1995).
2. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nature Genet.* **24,** 23–25 (2000).
3. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* (this issue).
4. Bishop, C. E., Boursot, P., Baron, B., Bonhomme, F. & Hatat, D. Most classical *Mus musculus domesticus* laboratory mouse strains carry a *Mus musculus musculus* Y chromosome. *Nature* **325,** 70–72 (1985).
5. Yonekawa, H. *et al.* Relationship between laboratory mice and subspecies *Mus musculus domesticus* based on restriction endonuclease cleavage patterns of mitochondrial DNA. *Jpn. J. Genet.* **55,** 289–296 (1980).
6. Ferris, S. D., Sage, R. D. & Wilson, A. C. Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. *Nature* **295,** 163–165 (1982).
7. Bonhomme, F., Guénet, J.-L., Dod, B., Moriwaki, K. & Bulfield, G. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *J. Linn. Soc.* **30,** 51–58 (1987).
8. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
9. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24,** 381–386 (2000).
10. Simpson, E. M. *et al.* Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nature Genet.* **16,** 19–27 (1997).
11. Schalkwyk, L. C. *et al.* Panel of microsatellite markers for whole-genome scans and radiation hybrid mapping and a mouse family tree. *Genome Res.* **9,** 878–887 (1999).
12. Threadgill, D. W., Yee, D., Matin, A., Nadeau, J. H. & Magnuson, T. Genealogy of the 129 inbred strains: 129/SvJ is a contaminated inbred strain. *Mamm. Genome* **8,** 390–393 (1997).
13. Rabiner, L. A. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–286 (1989).
14. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380,** 149–152 (1996).
15. Prager, E. M., Orrego, C. & Sage, R. D. Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* **150,** 835–861 (1998).

16. Threadgill, D. W., Hunter, K. R. & Williams, R. W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* **13,** 175–178 (2002).
17. Paigen, K. & Eppig, J. T. A mouse phenome project. *Mamm. Genome* **11,** 715–717 (2000).
18. Batzoglou, S. *et al.* ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
19. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. 2. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
20. Yokoyama, T. *et al.* Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. *Nucleic Acids Res.* **18,** 7293–7298 (1990).

Supplementary Information accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

Acknowledgements We thank the Mouse Genome Sequencing Consortium for sequencing and SNP discovery efforts and the University of Queensland for providing study leave to C.M.W. We thank W. Frankel and K. Scott for providing DNA from seven B6 founder stocks from the Jackson Laboratories. We also thank R. Jaenisch, D. Page, A. Chess, W. Dietrich, J. Hirschhorn, D. Altshuler, J. Barrett and M.-P. Reeve for comments on the manuscript; B. Gilman, S. Schaffner and E. Karlsson for computational assistance; and L. Gaffney for assistance with figures. M.J.D. is supported through a computational biology fellowship funded by Pfizer, Inc.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.D. (e-mail: mjdaly@genome.wi.mit.edu) or K.L.T. (e-mail: kersli@genome.wi.mit.edu).

Numerous potentially functional but non-genic conserved sequences on human chromosome 21

Emmanouil T. Dermitzakis*, Alexandre Reymond*, Robert Lyle*, Nathalie Scamuffa*, Catherine Ucla*, Samuel Deutsch*, Brian J. Stevenson†‡, Volker Flegel†‡, Philipp Bucher†§, C. Victor Jongeneel†‡ & Stylianos E. Antonarakis*

* *Division of Medical Genetics, 1 Rue Michel-Servet, University of Geneva Medical School and University Hospitals of Geneva, CH-1211 Geneva, Switzerland*
† *Swiss Institute of Bioinformatics,* § *Swiss Institute for Experimental Cancer Research, and* ‡ *Office of Information Technology, Ludwig Institute for Cancer Research, 155 Chemin des Boveresses, CH-1066 Epalinges, Switzerland*

The use of comparative genomics to infer genome function relies on the understanding of how different components of the genome change over evolutionary time[1–3]. The aim of such comparative analysis is to identify conserved, functionally transcribed sequences such as protein-coding genes and non-coding RNA genes, and other functional sequences such as regulatory regions[4,5], as well as other genomic features. Here, we have compared the entire human chromosome 21 with syntenic regions of the mouse genome, and have identified a large number of conserved blocks of unknown function. Although previous studies have made similar observations[6,7], it is unknown whether these conserved sequences are genes or not. Here we present an extensive experimental and computational analysis of human chromosome 21 in an effort to assign function to sequences conserved between human chromosome 21 (ref. 8) and the syntenic mouse regions. Our data support the presence of a large number of potentially functional non-genic sequences, probably regulatory and structural. The integration of the properties of the conserved components of human chromosome 21 to the rapidly accumulating functional data for this chromosome[9,10] will improve considerably our understanding of the role of sequence conservation in mammalian genomes.

The sequence of human chromosome 21 (ref. 8) was obtained from the National Center for Biotechnology Information (NCBI) and aligned with PipMaker[11] to the mouse orthologous sequences (both sequences were hard-masked with Repeatmasker). Details and parameters of the alignment are described in the Methods. Briefly, 33.5 megabases (Mb) of human chromosome 21 sequence was compared to approximately 21 Mb of mouse sequence from chromosomes 16, 17 and 10 (refs 7, 12). A large number of ungapped conserved sequences were identified and their

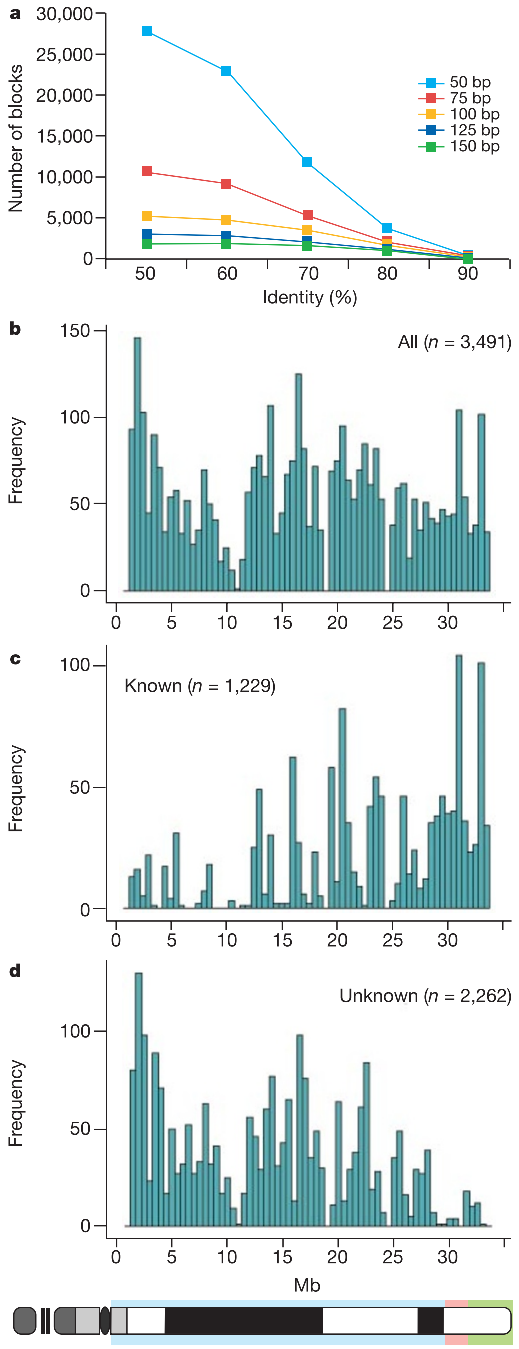


Figure 1 Distribution of conserved blocks on the 33.5 Mb of human chromosome 21 (long arm). **a**, Number of conserved sequence blocks identified (known and unknown) as a function of the threshold criteria for the size and percentage identity of the ungapped conserved blocks. **b**–**d**, The distribution of conserved blocks on the 33.5 Mb of human chromosome 21 at ≥100 bp and ≥70% identity is shown. Histograms showing all 3,491 conserved blocks (**b**), 1,229 conserved blocks corresponding to known exonic sequences (**c**), and 2,262 conserved blocks of unknown function (**d**). Below the histograms, human chromosome 21 with its banding pattern, and with the corresponding mouse syntenic regions is shown (mouse chromosomes 16, 17 and 10 are indicated in blue, pink and green, respectively).