nature genetics

Recurrent DNA copy number variation in the laboratory mouse

Chris M Egan[1], Srinath Sridhar[2], Michael Wigler[1] & Ira M Hall[1]

Different species, populations and individuals vary considerably in the copy number of discrete segments of their genomes. The manner and frequency with which these genetic differences arise over generational time is not well understood. Taking advantage of divergence among lineages sharing a recent common ancestry, we have conducted a genome-wide analysis of spontaneous copy number variation (CNV) in the laboratory mouse. We used high-resolution microarrays to identify 38 CNVs among 14 colonies of the C57BL/6 strain spanning ~967 generations of inbreeding, and we examined these loci in 12 additional strains. It is clear from our results that many CNVs arise through a highly nonrandom process: 18 of 38 were the product of recurrent mutation, and rates of change varied roughly four orders of magnitude across different loci. Recurrent CNVs are found throughout the genome, affect 43 genes and fluctuate in copy number over mere hundreds of generations, observations that raise questions about their contribution to natural variation.

Segmental changes in DNA copy number are particularly common in mammals. A substantial fraction of genomic DNA (~2–6%) is contained within segmental duplications[1], and variability in copy number is widespread among different human[2–9] and chimpanzee individuals[10], and among different inbred mouse strains[11,12]. Relatively little, however, is known about the underlying behavior of CNVs. As for any other class of mutation, of fundamental importance are the rate at which new variants arise, the rate at which variants revert to their former state, and the uniformity of these rates across the length of the genome. In mammals, rates of large-scale structural mutation have thus far been directly estimated at only a small number of select loci. Insertions and deletions within the 2.5-megabase (Mb) human *DMD* gene are found in the human population at frequencies of ~1 in 10^5 and 10^4 newborns, respectively[13], and sperm-based estimates at several loci have indicated rates of ~1×10^{-5}–10^{-7} per generation[14–16]. In contrast, several human disorders are associated with specific rearrangements that arise at a relatively high frequency (~1 in 10^4 newborns)[17], and high rates of structural mutation (~$2–4 \times 10^{-4}$ per generation) have been documented at four complex regions of the human Y chromosome[18]. It is not clear to what extent the above loci reflect genome-wide behavior. Linkage disequilibrium (LD) between CNVs and SNPs suggests that many CNVs arose only once in human history[5,6,8,9], but attempts to examine LD at sites of segmental duplication have been limited by poor SNP coverage in these regions[8,9]. A few indirect lines of evidence suggest the presence of hot spots where CNVs arise at elevated rates: segmental duplications and CNVs are nonrandomly distributed across the genome[1–4]; some CNVs are variable within diverse human populations[19]; and some loci are CNVs in both humans and chimpanzees[10]. However, it is difficult to distinguish between variable mutation rates and the effects of population structure and selection, and simple endpoint comparisons of distantly related genomes are blind to transient and reversible fluctuations that may occur over shorter time scales. The rate of copy number change and the prevalence of recurrent mutation have therefore remained open questions.

Here, we have directly examined spontaneous DNA copy number change across the entire genome of the laboratory mouse, a mammal with the unique features of an inbred genome and a known breeding history. Our preliminary experiments suggested that CNVs might arise at a high rate: individuals from the same inbred mouse strain often differed by new CNVs (**Supplementary Note** online). To explore this, we sought to systematically identify spontaneous mutations over a large number of reproductive generations. Within the widely used C57BL/6 inbred strain (hereafter referred to as B6), many different colonies, or substrains, are bred independently at different institutions, and, as a consequence of mutation and genetic isolation, their genomes are continually diverging from one another. Substrains are propagated by a single brother-sister mating each generation, and new substrains are founded by a single sibling pair. We were able to obtain pedigreed individuals and precise genealogical information for 13 B6 substrains spanning ~967 generations of divergence (± 29; **Supplementary Fig. 1** online). Because these substrains separated from one another after considerable inbreeding, they should differ only by spontaneous mutations[20].

We carried out representational oligonucleotide microarray analysis (ROMA)[21] for two different tissues from a single pedigreed mouse from each of the 13 different B6 substrains. We used a microarray containing 83,032 probes distributed across the entire genome[22]. For each substrain we carried out at least three independent experiments

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [2]Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. Correspondence should be addressed to I.M.H. (hall@cshl.edu).

Received 20 February; accepted 27 August; published online 28 October 2007; doi:10.1038/ng.2007.19

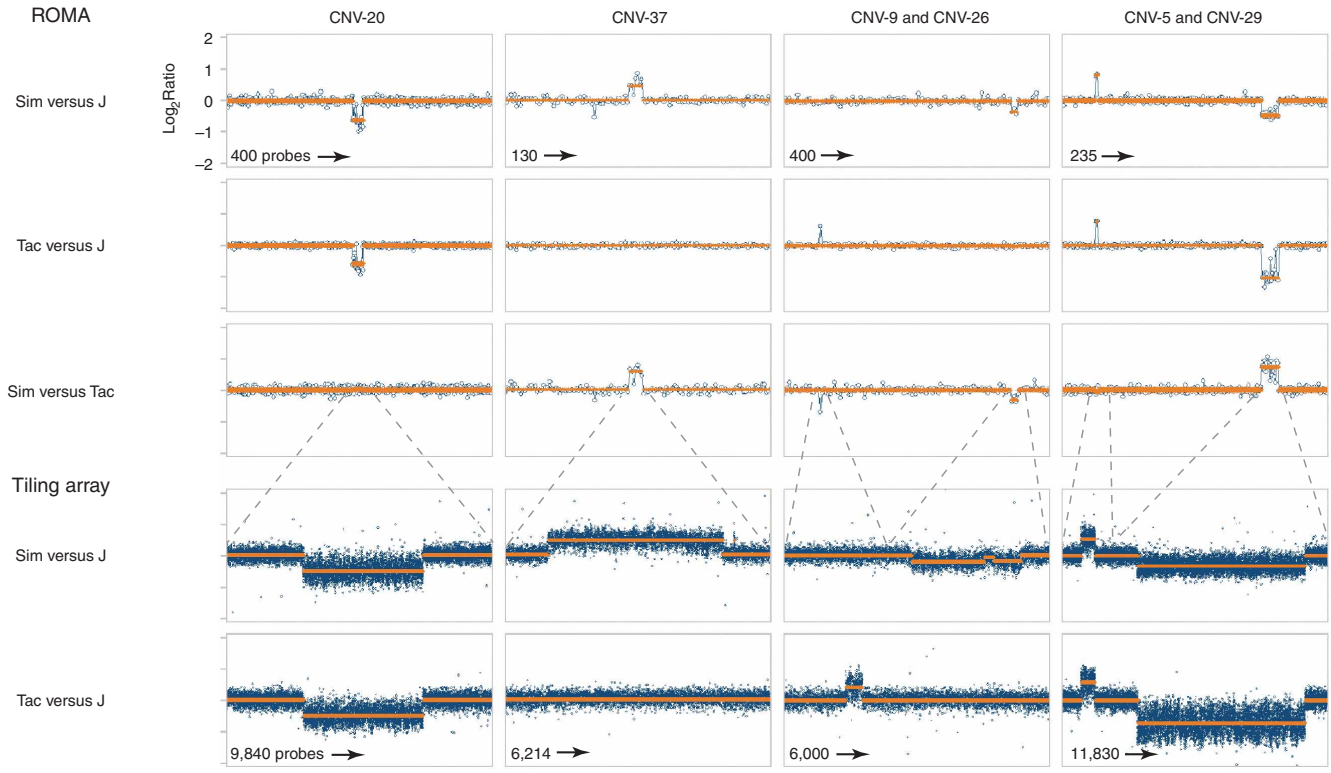

Figure 1 Examples of microarray data for three substrains at six loci, demonstrating representative data quality and experimental design. Substrains were compared to the B6J (J) reference substrain and to each other in triangular fashion using ROMA, and CNVs were verified and precisely mapped by tiling array CGH. Shown here are results for B66NSim (Sim) and B6NTac (Tac). Indicated at left is the type of microarray comparison done and the strains used. On the *y* axis is the mean $\log_2$ ratio from 1–3 experiments; on the *x* axis are the probes shown in genome order. Blue indicates ratios from primary data; orange indicates the mean ratios of the segmented dataset and of the identified CNV. Probes are shown once even if they map to multiple locations. The total number of probes shown in each graph is indicated at the lower left corner.

against the B6J reference strain, and at least one experiment against a third strain to identify loci with more than two alleles (**Fig. 1**). We identified 23 segmental CNVs affecting two or more consecutive probes, and 65 mutations affecting just a single probe (**Supplementary Methods** online). We did not observe any ROMA differences between the two tissues of the same individual. To verify and more precisely map these mutations, we carried out comparative genomic hybridization (CGH) for each substrain using high-resolution oligonucleotide microarrays (tiling arrays). These arrays contained 385,213 probes placed at high density (~1 per 46 bp) across each locus identified by ROMA. We verified 22 of the 23 segmental CNVs and discovered 16 additional CNVs (defined as >1 kb) among the 65 single-probe ROMA differences. We further verified 9 of 9 CNVs by quantitative PCR (**Supplementary Figs. 2** and **3** online). These 38 CNVs are dispersed throughout the genome, range in size from ~4 kb to 4 Mb, and affect 60 genes. These differences among varieties of the B6 strain present a confounding factor as well as an experimental resource for genetic researchers.

We assessed relative DNA copy number at the 38 CNV loci within a larger panel of inbred strains. These included four unpedigreed individuals from the B6Crl substrain, as well as two individuals from each of 12 inbred strains related to B6 in varying degrees: three are C57-family strains derived from the same original genetic cross in 1921 that gave rise to B6, one is a B6 substrain with ~16% genetic contamination by DBA (C57BLKS/J), another is derived from the same male founder individual as B6 (C58), and seven are classical inbred strains related to B6 only through their common genetic origins in European and Asian 'fancy' mice[23] (**Fig. 2a**). We used ROMA to compare multiple individuals with the B6J reference, and we analyzed at least one individual by tiling array CGH.

Notably, 18 of 38 CNVs arose multiple times within distinct lineages. On the basis of the most parsimonious explanation for the distribution of CNVs relative to the structure of the genealogy, 20 of 38 could be ascribed to a single mutational event (class I, **Fig. 2b**). Eight CNVs arose once within the B6 substrains but were also variable among the panel of non-B6 strains, indicating that there were a minimum of 2–3 independent occurrences (class II, **Fig. 2c**). Ten loci underwent a minimum of 2–6 copy number changes over just ~967 generations of inbreeding within the B6 substrains themselves (class III, **Fig. 2d**). Thus, large segmental differences in DNA copy number are often the product of recurrent mutation, and this process can occur over very short time scales.

The above results indicate that the rate of spontaneous copy number change varies considerably among loci, with class I and class III CNVs representing opposite ends of the spectrum. To calculate rates, we used only data from B6 substrains, and we assumed that spontaneous germline mutations segregated to descendants in a random fashion. It should be noted that this is a simplification given selection for health and fecundity in mouse colonies. Most spontaneous mutations are lost from an inbreeding population as a result of random drift, hence the 967 generations of inbreeding shown in **Figure 2a** are equivalent to ~278 parent-offspring trios,

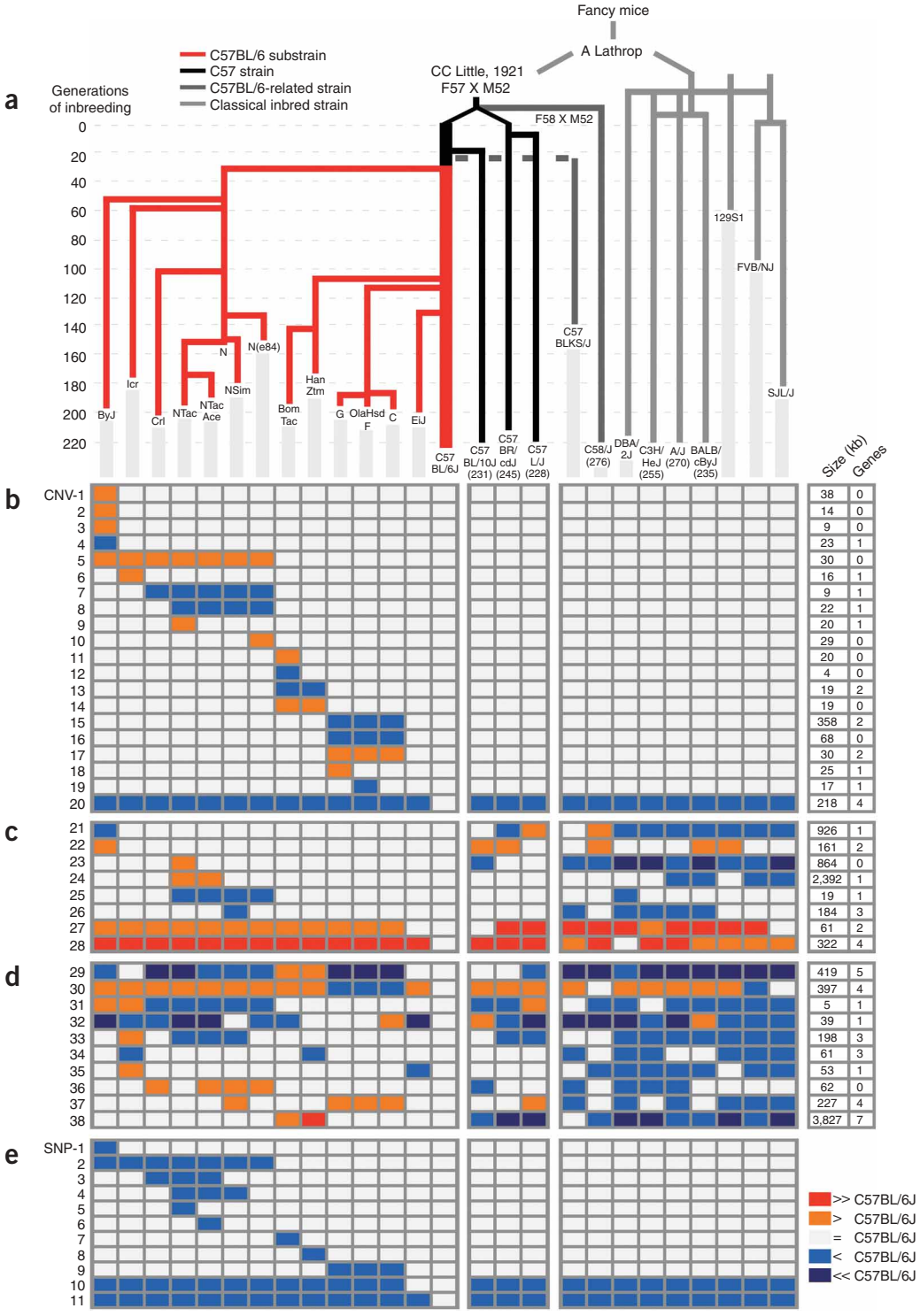

Figure 2 Distribution of CNVs relative to the strain genealogy. (**a**) Relationship of the inbred mouse strains analyzed in this study. C57BL/6 substrains are shown in red, C57 strains derived from the same genetic cross as B6 (female 57 by male 52) in black, strains moderately related to B6 in dark gray and strains relatively unrelated to B6 in light gray. The number of generations of inbreeding is shown at left and is indicated by dashed lines. (**b**–**d**) Copy number of the 38 CNVs across 26 strains, all relative to B6J. Each row in the grid is a locus, and each column is a strain. Locus identifiers are shown to the left, and the size of each CNV and the number of affected genes are shown to the right. Orange indicates that a segment has more copies than B6J, and blue indicates fewer copies. In cases where additional alleles have been identified by ROMA experiments directly comparing substrains to one another (as in CNV-29, **Fig. 1**), red indicates more copies than orange, and dark blue indicates fewer copies than blue. Loci are grouped into three classes: class I CNVs arose once within B6 substrains (**b**), class II CNVs arose once within B6 substrains but more than once when all strains are taken into account (**c**), and class III CNVs arose more than once among B6 substrains (**d**). (**e**) Genotypes of single-nucleotide mutations among the strains, shown in blue because they appear as 'losses' as a result of decreased hybridization efficiency.

or ~556 meiotic generations (we used the latter for all rates). Under these assumptions, the minimum rate of copy number change at the ten class III CNV loci ranged from 3.6×10^{-3} to 1.1×10^{-2} per generation (**Supplementary Table 1d** online). This corresponds to one spontaneous event per 46–139 newborns. To our knowledge, such high rates of large-scale DNA copy number change in the germline have not been observed in mammals. They are an order of magnitude higher than rates of structural mutation reported on the human Y chromosome[18] and 1–2 orders of magnitude higher than rates of changes underlying the most common human genomic disorders[17], and are rivaled only by rates for considerably smaller (1–100 bp) micro- and minisatellite repeats. Moreover, seven of the ten class III loci contain one or more entire genes. In general support of the rates calculated above, we observed additional mutations at three class III CNVs in entirely independent experiments: CNV-29 and CNV-32 were variable among seven B6J individuals; CNV-32 was also variable between two C57BL/10J individuals; and CNV-37 was variable in progeny from 2 of 6 genetic crosses between B6 individuals, most likely resulting from 2 independent mutations in 41 generations (**Supplementary Note** and **Supplementary Fig. 4** online).

In contrast, most of the genome seems to be relatively stable. Although there will be a range of mutation rates across the genome, we can calculate the rate of class I CNVs (those that arose once) under the assumption that they have an equal chance of occurring at any location. We observed 20 events involving 57 of 83,032 probes over ~556 generations, and therefore at any one of our probes we would expect a mutation rate of about $57/(83{,}032 \times 556) = 1.2 \times 10^{-6}$ per generation. This is in general agreement with rates estimated at several loci[13–17], but it may be an underestimate given imperfect CNV detection and selection against deleterious mutations. The above considerations indicate that the CNV mutation rate varies by roughly four orders of magnitude across the genome.

Given the large fraction of CNVs that recurred and the high rates of change at certain loci, we sought to rule out alternative explanations for recurrence. First, false recurrence might result from an incorrect genealogy; this can be tested by observing whether other classes of mutation show the expected pattern across the strains. We sequenced 45 single-probe ROMA differences that were not CNVs and identified 11 single-nucleotide mutations, all of which showed the expected distribution among the substrains (**Fig. 2e**). Second, some recurrent mutations could be explained if certain B6 substrains carried genomic segments of non-B6 origin, as could occur either by genetic contamination or by persistent heterozygosity of ancestral variation. This explanation is difficult to entirely rule out, but it does not seem likely. ROMA comparisons between B6 and the 12 other inbred strains revealed ~15,000 polymorphic probes, many due to small-scale sequence variation[22] (**Supplementary Note**). Genetic contamination would result in the simultaneous introgression of many genetic markers and should be readily apparent. Similarly, there exist ~3,000 polymorphic probes among the four C57 strains, the vast majority of which would have been segregating in the original genetic cross giving rise to B6. Persistent heterozygosity should affect more than one class of genetic marker at a time, and with the exception of CNVs, none were variable between B6 substrains. Arguing against all of the above scenarios, 15 of 18 recurrent CNVs have multiple alleles unlikely to have originated from the same mutational event. A final concern is that high mutation rates result because the B6 genealogy spans far more than the predicted number of generations, but, for this, ROMA presents a test: each experiment utilizes ~158,888 restriction sites, and we should be able to detect mutations at >80% of these, effectively 'sequencing' >762 kb of DNA. We found two mutations within restriction sites, resulting in a mutation rate ($\sim 5 \times 10^{-9}$) that is consistent with various estimates of $\sim 1 \times 10^{-8}$ per base per generation[24].

As defined by their structure in the fully sequenced genome of the B6J reference substrain and by the general structure of mutations identified by microarray experiments, recurrent CNVs have several notable characteristics. Seventeen of eighteen occur in multiple copies in B6J, whereas the same is true for only 1 of 20 of the class I CNVs. Three recurrent CNVs are large (2- to 4-Mb) arrays of tandem duplications, each containing assembly gaps, at which mutations can produce complex patterns (**Supplementary Fig. 5** online). CNV-32 is a hypervariable locus with alleles that vary widely in size and copy number and is found proximal to the highly unstable pseudoautosomal region[25]. CNV-22 is a single-copy locus at which one duplication resulted in distinct boundaries, and two others did not. However, most recurrent CNV loci (13 of 18) in B6J are composed of 2–4 copies of discrete 14- to 140-kb segments. At 12 of 13 loci, duplicate copies are found near each other (<1 Mb), and at 6, they are immediately adjacent. Unfortunately, locus structure is often unclear; five loci contain assembly gaps, and CNV-34 seems to be misassembled (there is one copy in the reference sequence, but we can amplify DNA from strains showing losses). At these loci, recurrent mutations did not give rise to distinguishable boundaries and seem to entail iterated expansion and contraction of discrete units. The correlation among segmental duplications in B6J, instability in B6 substrains, and CNV between less-related strains is consistent with the role of nonallelic homologous recombination in generating structural mutations, but we caution that a causal relationship between large-scale structure and recurrent mutation at these loci remains to be demonstrated. We further note that in four cases where informative alleles were sequenced in the B6J substrain and fully assembled, a repetitive element was found at the junction of duplications (CNV-27, CNV-29, CNV-35 and CNV-37).

We have demonstrated that recurrent structural mutation is a key force generating CNV in laboratory mice and that large genomic segments can fluctuate in copy number over exceptionally short time scales. We were able to observe this because, by contrast with previous studies, we examined genome-wide CNV in a context where ancestral and spontaneous variation could be clearly distinguished, and because we examined many independent lineages each separated by a small number of generations. Several considerations suggest that recurrent structural mutation may represent an important biological process. First, we have certainly underestimated the prevalence of recurrent CNV in mice. Genome coverage in this study is relatively sparse and uneven (mean spacing ~32 kb, median ~17 kb) and is strongly biased toward unique sequences as a consequence of the probe design algorithm. Although 4.7% of the mouse genome is contained within segmental duplications (X. She and E.E. Eichler, personal communication), this is the case for only 1.7% of our ROMA probes, and there exist substantial gaps in certain regions. Second, although their phenotypic consequences remain unclear, the 18 recurrent CNVs that we report affect 43 known genes, notable among which are factors involved in reproduction (CNV-27, CNV-28, CNV-29 and CNV-30), immunity (CNV-21, CNV-26 and CNV-33) and cognition (CNV-22 and CNV-38). Finally, it seems unlikely that this phenomenon is limited to mice. Less of the human genome is contained within tandem duplications than is true for the mouse genome (2.3% versus 4.3%)[26], but this is nevertheless a substantial fraction, and many of these loci are known to coincide with CNV. Care will be required to discern the contribution of these extraordinarily plastic genomic regions to natural variation, evolution and disease.

METHODS

Mouse strains. We collected B6 substrains with the requirements that (i) we could acquire a pedigreed individual of a known generation number, (ii) breeders could verify that it was maintained by strict brother-sister mating and (iii) the precise generation number at which the substrain diverged from its parent substrain was known, or could be closely approximated by a year date. Generation numbers for substrain divergence points were established through direct communication with breeders, with the exception of the B6ByJ strain, for which we consulted diagrams drawn by its creator, D.W. Bailey (**Supplementary Fig. 1**). To help establish parsimony, we included data from four unpedigreed B6Crl individuals obtained in January 2005, but we did not use this substrain for rate estimates, nor do we report CNVs unique to it. All non-B6 strains were obtained from the Jackson Laboratory between November 2004 and December 2005. All inbred individuals from the same strain were obtained on the same date except for the 7 B6J mice: B6J-1 and 2 were received in November 2004; B6J-3, 4, 5 and 6 in February 2005; and our pedigreed foundation-stock reference individual was received in January 2006. All mice were female except for male parents and progeny of B6 crosses. Experiments were approved by the Institutional Animal Care and Use Committee of Cold Spring Harbor Laboratory.

Representational oligonucleotide microarray analysis (ROMA). Genomic representations, dye incorporation, hybridization and data processing were carried out as described previously[2]. Representations compared to one another were always prepared in parallel. We defined a single experiment as replicate hybridizations done with reversal of Cy5 and Cy3 dyes relative to DNA samples. Replicate experiments used either separate tissues (liver and tail), separate individuals (liver or tail), or independently constructed genomic representations (**Supplementary Table 1b**). Experiments were carried out on a microarray synthesized by NimbleGen Systems containing 84,033 50-mer oligonucleotide probes, 1,001 of which are controls with no match in the mouse genome[22].

Tiling array CGH. Tiling-array experiments were carried out on a microarray containing 385,213 isothermal probes 45–70 bp in length, designed and synthesized by NimbleGen Systems. Probes were placed at maximal density (1 per 46 bp) across each locus identified by ROMA, requiring a minimum distance of two ROMA probes or 20 kb to either side, avoiding high-copy repeats. Two micrograms of DNA was sonicated to a size range of 200–2,000 bp, and dye incorporation, hybridization and data processing were carried out as for ROMA. A subset of experiments was done by NimbleGen Systems, as previously described[27].

Identification of mutations in B6 substrains. We identified CNVs using a hidden Markov model (HMM) algorithm loosely based on a model for human ROMA data that has been described previously[2]. Our model has three states: duplicated ('up'), equivalent ('ground') and deleted ('down'). We assume that the $\log_2$ ratio of each probe is generated from one of three Gaussian distributions representing up, ground and down. We model each state of the HMM as a mixture of these three Gaussians with different mixture proportions for each state. The parameters of the model, such as the three Gaussian distributions, state-specific mixture weights and transition probabilities, were estimated using heuristics explained in the **Supplementary Methods**. Our parameter estimation for ROMA differed from that of the tiling arrays mainly owing to differences in probe density and the amount of experimental noise. To obtain the most probable state path, we used the Viterbi algorithm on the HMM; this classified multi-probe segments as polymorphic (up or down) or not (ground).

We identified single-probe ROMA polymorphisms using thresholds. We required the mean $\log_2$ ratio of the probe to differ by >1 s.d. from the mean of the entire dataset for each of the three replicate experiments against the reference, and to satisfy one of the following criteria: (i) differs by >6 s.d. from the combined mean of all three experiments comparing a single substrain against the reference; (ii) differs by >5 s.d. from the mean of all three experiments against the reference, and does so for two different substrains; or (iii) differs by >4 s.d. from the mean of all three experiments against the reference, and does so for three different substrains.

Rules for scoring CNVs. We considered a CNV polymorphic in a given (sub)strain if either of the following two criteria were met: (i) a CNV identified in tiling array–CGH experiments encompassed one or more of the probes originally identified as polymorphic in ROMA experiments between B6 substrains or (ii) in ROMA experiments the median $\log_2$ ratio of the probes contained within a CNV differed from the median $\log_2$ ratio of the whole dataset by >5 median absolute deviations (MAD). The dataset MAD was calculated using a sliding window consisting of the same number of probes as the segment being scored. We curated CNV calls by inspection of primary data, and corrected 13 obvious errors out of 950 total calls (**Supplementary Table 1a**).

A locus was judged to have more than three possible alleles (that is, 'up,' 'down' or 'ground' relative to B6J) only if we obtained direct evidence for this in ROMA experiments comparing non-B6J substrains. Once the presence of additional alleles was established, we used thresholds to call CNVs in additional strains on the basis of the relative amplitudes of the variable segments.

Estimation of mutation rates. For rate estimates, we assumed that mutations have an equal probability of occurring at any generation within the B6 genealogy, that mutations are transmitted according to the laws of mendelian segregation, and that breeding pairs are randomly chosen from a colony. Although a new mutation has a 0.75 probability of being lost from an inbreeding lineage after an infinite number of generations, over measurable time scales this probability can be substantially lower. For each generation in the tree after the F32 split between B6J and B6N, we calculated the probability that a mutation occurring in that generation would be lost from the entire subtree below. We computed this probability exactly by using an efficient dynamic-programming algorithm. We let p denote the average of these probabilities. If k mutations arose, then the expected number of observed mutations would be $k(1-p)$. Therefore, if we observed r mutations over n loci over t generations of inbreeding, then the maximum likelihood estimate of the mutation rate would be $r/n(1-p)2t$. On the basis of this analysis, we expected 71.3% of new mutations to have been lost.

For class III loci, we estimated the most parsimonious scenario by minimizing the number of mutations required to explain the observed alleles. We assumed that a single mutation gives rise to exactly one new heterozygous allele, that a single mutation can alter copy number arbitrarily, and that different alleles did not segregate within a single colony for more than 50 generations (the probability of such a segregation event is $\sim 2.2 \times 10^{-5}$).

Identification and genotyping of single-nucleotide mutations. Any mutation that affects the PCR amplification efficiency of a *Bgl*II restriction fragment, or the efficiency with which a fragment hybridizes to a probe on the microarray, can cause a single-probe ROMA difference. For each such difference we attempted to PCR amplify the predicted *Bgl*II restriction fragment and to assay both the size of the fragment and the integrity of the probe target sequence. We designed PCR primers to span fragments in a manner such that, when the fragments were cleaved by *Bgl*II, three distinguishable electrophoretic bands would be produced. Bands from test and reference strains were compared and scored for visible mutations. To identify mutations within the probe target sequence, we subjected gel-purified PCR products to DNA sequencing. When SNP genotypes in additional strains were not clear by ROMA, we confirmed them by DNA sequencing (**Supplementary Fig. 2d**).

Quantitative PCR (qPCR). Quantitative PCR was carried out with an ABI PRISM 7900HT Sequence Detection System using SYBR Green PCR Master Mix (Applied Biosystems). Amplicon size range was typically 70–200 bp, and primers were tested for specificity by PCR and gel electrophoresis preceding qPCR. For each primer pair, four reactions were set up for the query DNA, the reference DNA and a control lacking DNA. For quantitative purposes, each qPCR plate also included a primer pair corresponding to a control locus known to be at equivalent copy number in the query and reference DNA.

Genome sequence analysis. Segmental duplications were identified by BLAT[28] of each CNV locus against the University of California Santa Cruz mm8 genome assembly. Dot plots were carried out on at least 2 Mb of DNA sequence surrounding each CNV using the LBDOT program[29] with window sizes of 100, 250 and 500 bp, and 0–10 allowed mismatches. For each locus, genes were identified by comparison of CNV coordinates with the University of Santa Cruz Known Genes track[30]. Redundant and overlapping genes were discarded for **Figure 2** but included in **Supplementary Table 2c**. Probes were mapped to

segmental duplications on the basis of data from the website of E.E. Eichler (http://eichlerlab.gs.washington.edu/database.html).

Accession codes. National Center for Biotechnology Information Gene Expression Omnibus: Microarray data have been deposited with GEO accession codes GSE8980 (ROMA) and GPL5777 (tiling array CGH).

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

ACKNOWLEDGMENTS

We thank J.C. Mell, L. Stein, B. Stillman and J.D. Watson for comments on the manuscript; L. Bianco for technical assistance; C. Ward for the B6N substrain and for advice; H. Hedrich and M. Dorsch for supplying the B6HanZtm substrain; A. Lerro for the B6Icr substrain; all strain suppliers for crucial genealogical details; N. Navin, V. Grubor, B. Lakshmi, A. Leotta and J. Kendall for computational contributions; and X. She and E.E. Eichler for sharing unpublished data. This work was supported by grants to I.M.H. from the Cold Spring Harbor President's Council and the Burroughs Wellcome Fund and to M.W. from The Simons Foundation. M.W. is an American Cancer Society Research Professor.

AUTHOR CONTRIBUTIONS

C.M.E. performed most of the microarray experiments, all of the PCR and DNA sequencing, and some of the data analysis. S.S. designed and implemented algorithms to identify CNVs and calculate mutation rates, and advised on all computational matters. M.W. contributed reagents and advice and helped calculate rates. I.M.H. conceived the study, obtained strains and genealogical information, performed many of the microarray experiments and most of the data analysis, and wrote the paper.

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

1. Bailey, J.A. & Eichler, E.E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
2. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
3. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
4. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
5. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
6. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X.Y. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
7. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
8. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
9. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
10. Perry, G.H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. USA* **103**, 8006–8011 (2006).
11. Li, J. *et al.* Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36**, 952–954 (2004).
12. Snijders, A.M. *et al.* Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res.* **15**, 302–311 (2005).
13. van Ommen, G.J. Frequency of new copy number variation in humans. *Nat. Genet.* **37**, 333–334 (2005).
14. Hollies, C.R., Monckton, D.G. & Jeffreys, A.J. Attempts to detect retrotransposition and de novo deletion of Alus and other dispersed repeats at specific loci in the human genome. *Eur. J. Hum. Genet.* **9**, 143–146 (2001).
15. Han, L.L., Keller, M.P., Navidi, W., Chance, P.F. & Arnheim, N. Unequal exchange at the Charcot-Marie-Tooth disease type 1A recombination hot-spot is not elevated above the genome average rate. *Hum. Mol. Genet.* **9**, 1881–1889 (2000).
16. Tusie-Luna, M.T. & White, P.C. Gene conversions and unequal crossovers between CYP21 (steroid 21-hydroxylase gene) and CYP21P involve different mechanisms. *Proc. Natl. Acad. Sci. USA* **92**, 10796–10800 (1995).
17. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
18. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
19. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
20. Haldane, J.B.S. The amount of heterozygosity to be expected in an approximately pure line. *J. Genet.* **32**, 375–391 (1936).
21. Lucito, R. *et al.* Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**, 2291–2305 (2003).
22. Lakshmi, B. *et al.* Mouse genomic representational oligonucleotide microarray analysis: detection of copy number variations in normal and tumor specimens. *Proc. Natl. Acad. Sci. USA* **103**, 11234–11239 (2006).
23. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nat. Genet.* **24**, 23–25 (2000).
24. Drake, J.W., Charlesworth, B., Charlesworth, D. & Crow, J.F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
25. Kipling, D., Salido, E.C., Shapiro, L.J. & Cooke, H.J. High frequency *de novo* alterations in the long-range genomic structure of the mouse pseudoautosomal region. *Nat. Genet.* **13**, 78–80 (1996).
26. She, X. *et al.* A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res.* **16**, 576–583 (2006).
27. Selzer, R.R. *et al.* Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosom. Cancer* **44**, 305–319 (2005).
28. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
29. Huang, Y. & Zhang, L. Rapid and sensitive dot-matrix methods for genome analysis. *Bioinformatics* **20**, 460–466 (2004).
30. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).