nature genetics

Mouse segmental duplication and copy number variation

Xinwei She[1], Ze Cheng[1], Sebastian Zöllner[2], Deanna M Church[3] & Evan E Eichler[1,4]

Detailed analyses of the clone-based genome assembly reveal that the recent duplication content of mouse (4.94%) is now comparable to that of human (5.5%), in contrast to previous estimates from the whole-genome shotgun sequence assembly. However, the architecture of mouse and human genomes differs markedly: most mouse duplications are organized into discrete clusters of tandem duplications that show depletion of genes and transcripts and enrichment of long interspersed nuclear element (LINE) and long terminal repeat (LTR) retroposons. We assessed copy number variation of the C57BL/6J duplicated regions within 15 mouse strains previously used for genetic association studies, sequencing and the Mouse Phenome Project. We determined that over 60% of these base pairs are polymorphic among the strains (on average, there was 20 Mb of copy-number-variable DNA between different mouse strains). Our data suggest that different mouse strains show comparable, if not greater, copy number polymorphism when compared to human; however, such variation is more locally restricted. We show large and complex patterns of interstrain copy number variation restricted to large gene families associated with spermatogenesis, pregnancy, viviparity, pheromone signaling and immune response.

Initial estimates suggested that 1–2% of the mouse genome[1–3] consisted of high-identity (>90%) duplications. These estimates, however, were complicated by the whole-genome shotgun sequence assembly (WGSA) method, which cannot resolve large, highly identical duplications. In particular, the largest (>15 kb) and most identical (>97%) duplicated segments[4] are often missing, collapsed, or mis-assigned as part of WGSA draft assemblies. Missing duplications, for example, are thought to result from difficulties in assembling regions of the genome where there is an excess of sequence mate-pair violations due to paralogous sequences. As the mouse genome assembly has progressed from WGSA to an ordered BAC-based assembly, the segmental duplication content identified has gradually increased[5,6]. Accurate resolution of the duplicated segments is particularly critical, as some of these regions have been shown to be highly variable in copy number between commonly related strains of mice[7–11], enriched in lineage-specific gene families undergoing positive selection[12,13], and preferential sites of large-scale rearrangement associated with chromosome evolution in the rodent lineage[6,14–16]. Here, we present a detailed analysis of the recent duplication content of the mouse clone-based finished genome assembly and assess copy number variation (CNV) of these regions in 15 different inbred strains of mice. The results suggest distinct properties of mouse segmental duplications when compared to those of human and reveal previously unrecognized complex patterns of structural variation.

A self-comparison of the current mouse assembly genome (Build36) identifies 141.4 Mb of segmental duplication (>1 kb in length and >90% identity; **Supplementary Note** online). We confirmed 96% (83.14/86.63 Mb) of the largest (>10 kb) and most identical (>94%) duplications using a previously described detection strategy that is independent from the assembly[2,17]. As a second measure of validation, we examined a total of 24 large-insert clones that had been shown to produce multisite signals by FISH on C57BL/6J metaphase chromosomes[2,8]. Of the corresponding sequences, 23/24 were confirmed as duplicated by at least one of our measures for duplication (**Supplementary Table 1** online). Using only the assembly-based comparison, we found that most (21/24) carried more than 40% duplicated base pairs, attesting to the high quality of the mouse assembly (**Supplementary Table 1**). In total, considering all pairwise alignments (<94% identity) and all those (>94% sequence identity) that are confirmed by two independent methods, we calculated the segmental duplication content of the mouse genome to be 4.94%. This value represents a two- to threefold increase from previous estimates[1–3].

The availability of the human and mouse genomes as clone-ordered BAC-based sequence assemblies provides the first opportunity to systematically compare segmental duplication sequence properties for two mammalian genomes (**Table 1**). Both genomes show similar levels of duplication (~5%) distributed in a highly nonrandom fashion (**Supplementary Fig. 1** online). We find that recent mouse duplications are restricted to fewer genomic locations, with a total of 149 mouse duplication blocks (**Table 1** and **Fig. 1**) >100 kb in length compared to 269 blocks within the human genome (Build36). Although fewer in number, murine duplication blocks are 50–80% larger in size. For example, there is a total of 19 mouse duplication blocks greater than 1 Mb (**Fig. 1**) compared to 11 mapped within the human genome (**Table 1** and **Supplementary Table 2** online). Intrachromosomal duplications are more abundant in both genomes (**Table 1**); however, in the mouse genome there is a mode at ~95% sequence identity, whereas in humans

[1]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, USA. [2]Department of Biostatistics and Department of Psychiatry, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [4]Howard Hughes Medical Institute, Seattle, Washington 98195, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 19 February; accepted 14 May; published online 22 May 2008; doi:10.1038/ng.172

Table 1 Segmental duplication features of mouse and human

	Mouse (Build36)	Human (Build36)
Nonredundant base pairs	126.0 Mb (4.94% genome)	159.2 Mb (5.52% genome)
Number of pairwise alignments (intrachromosomal)	39,168 (519.7 Mb)	10,384 (149.4 Mb)
Number of pairwise alignments (interchromosomal)	52,423 (130.5 Mb)	15,530 (149.7 Mb)
Duplication blocks (>100 kb)	149	269
Duplication blocks (>1 Mb)	19	11
Number of pairwise alignments per block	4–7,557 (median 87)	2–601 (median 34)
Proportion of tandem duplications (%)		
All duplications (>1 kb)	35.2	21.6
Duplications >10 kb	88.6	28.4
Duplications >20 kb	89.2	32.9
LINE enrichment (all duplication)	69% enriched ($P < 0.001$)	5% depleted ($P > 0.05$)
SINE enrichment (all duplication)	49% depleted ($P < 0.001$)	9.9% enriched ($P < 0.05$)
LTR enrichment (duplication >20 kb)	80% enrichment ($P < 0.001$)	21.8% enriched ($P < 0.05$)
Exon density (exon/Mb)		
RefSeq	32 (55.8% depleted, $P < 0.001$)	56 (14.7% depleted, $P < 0.05$)
EST (spliced)	230 (7.9% depleted, $P > 0.05$)	599 (62.3% enriched, $P < 0.001$)

Duplication blocks were defined as regions containing large, high-identity pairwise alignments (>10 kb, >95% identity) where the sum of nonredundant basepairs is >100 kb. The significance of the enrichment was determined by simulating the genomic features in a random sample ($n = 1,000$) of mouse and human genomic sequence.

the mode is shifted to >99%. This difference cannot be explained solely by differences in the effective substitution rate[18]. There remains the possibility that the largest and most identical duplications map to gaps in the current mouse genome assembly.

As noted previously[2,5,8], there are few examples of large interchromosomal duplication (**Table 1**), and most large (>10 kb) intrachromosomal duplications are tandemly organized with >89% of the pairwise alignments mapping in close proximity to one another (**Fig. 2**). Mouse duplicated sequences have three to four times as many paralogs when compared to those of human. This finding implies that structural variation of the mouse genome mediated by nonallelic homologous recombination may be more common but should be more locally restricted. We compared the exon density (RefGene annotation) between unique and duplicated regions of the mouse genome (**Table 1**) and found a greater depletion of exons in mouse segmental duplication when compared to that of human. To eliminate the possibility of incomplete gene annotation and potential processed pseudogenes, we examined the density of all ESTs that show evidence of splicing. Once again, we found that the proportion of spliced ESTs is reduced (7.9%) when compared to unique regions of the genome, although this difference is not significant by simulation (**Table 1**). In contrast, the human genome shows a strong ($P < 0.001$) enrichment of spliced ESTs within segmental duplications.

The enrichment of Alu-SINE repeat elements at the boundaries of new human segmental duplications has been taken as evidence that these elements had a role in the dispersal of segmental duplications in the ancestral primate genome[19,20]. We examined the repeat composition of mouse segmental duplications and found them significantly enriched for both LINE and repeat elements (1.5- to twofold enrichment) (**Fig. 3** and **Supplementary Table 3** online). In contrast, SINE elements were underrepresented (49%, **Table 1**) in segmental duplications when compared to unique regions of the mouse genome. An examination of the transition boundaries between larger (>20 kb) segmental duplication alignments showed the most dramatic enrichment. Approximately 32% of the base pairs at these boundaries consisted of LINE repeat sequences (**Fig. 3**), whereas 20% were LTR repeat elements. When we limited the analysis to the transition regions between unique and duplication sequences, we found the most significant enrichment for LTR sequences in the duplicated portion when compared to the flanking unique sequence (**Fig. 3c**). Either side of the transition region seems equally enriched in LINE repeat elements, although this enrichment is significant only for the youngest LINEs (<12% sequence divergence from consensus; **Supplementary Table 4** online).

Numerous studies in different organisms have shown that segmental duplications are enriched four- to tenfold for copy number variation[9,21–23], although such variation also occurs outside regions of segmental duplication. Using our duplication map of the mouse genome, we specifically focused on the design of a customized high-density oligonucleotide array (average 1 probe per 481 bp) targeted to C57BL/6J segmental duplications that were confirmed by both computational methods (**Supplementary Note**). As a control, we also selected 273 regions that had been predicted to be copy number variant on the basis of earlier BAC–array CGH experiments (**Supplementary Table 5** online). We selected 15 inbred strains of mice on the basis of their genealogical relationship to C57BL/6J or their use as National Institutes of Environmental Health Sciences (NIEHS) sequencing strains or as part of the Mouse Phenome Project. All array-CGH experiments were done using C57BL/6J as the reference strain.

On the basis of the raw $\log_2$ signal intensity data[24], we observed marked CNV between the C57BL/6J and the other inbred strains (**Fig. 4a**). Signal intensity differences as detected by array CGH were greater than a similar dataset generated for assessing human CNV over segmental duplications[21], possibly because of the high level of homozygosity

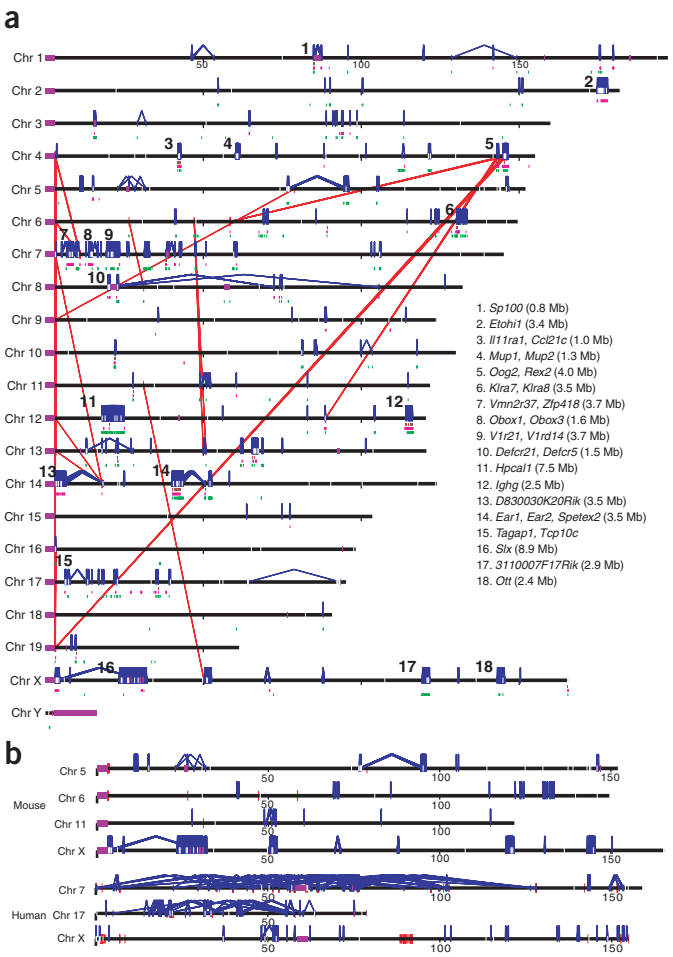

Figure 1 Mouse duplication patterns. (**a**) Mouse duplication and copy-number-variant genomic landscape. Interchromosomal (red) and intrachromosomal (blue) duplications (>20 kb and >94% sequence identity) are shown for the C57BL/6J mouse genome. Copy-number-polymorphic duplicated regions are flagged if two or more strains show a gain (green bars) or loss (pink bars) with respect to C57BL/6J. Brown bars highlight regions showing both gain and loss. Some of the largest duplicated and CNV regions are enumerated and labeled on the basis of representative RefSeq genes within the block. *Spetex2* is a gene family identified within rat but which does not have an MGI-approved gene family designation as of yet. Mouse chromosomes 7, 12, 14 and X show the greatest preponderance of large duplication blocks. In the case of chromosome 7, the duplication blocks account for 32% of the first 50 Mb of that chromosome. (**b**) Mouse versus human genome duplication pattern. Mouse and human intrachromosomal duplication patterns are compared for chromosomes 7, 17 and X. Note the human interspersed pattern of recent duplications when compared to the tandem clusters in mouse for the autosomes. A greater fraction of the mouse X chromosome is duplicated (12.8% in mouse vs. 7.8% in human). The X chromosome is syntenic between man and mouse. Human chromosome 17 is syntenic to mouse chromosome 11, and human chromosome 7 is syntenic to mouse chromosomes 6 and 5 based on UCSC Genome Browser Human Net track.

and the fixation of copy number variation within each inbred strain, facilitating their detection. We used a hidden Markov model (HMM) to identify significant transitions in $\log_2$ ratios corresponding to a likely copy number gain or loss. Our HMM requires at least 24 probes of unchanged state before calling a region as copy number variant, thereby limiting our detection to CNVs >12 kb in length. We validated our CNVs by comparing our results to 42 'high confidence' copy number variants for intervals that had been predicted previously[9] in five inbred strains that overlapped with our dataset. The comparison (**Supplementary Table 5**) showed that the HMM performed well, correctly identifying 95% (41/42) of these sites. As a control, we compared two different individuals from the C57BL/6J and identified 4/2,424 potential copy number differences (**Supplementary Table 6** online). Two of these positives corresponded to known sites of somatic variation (IgH), leaving two potential false positives or regions that are variable between C57BL/6J individuals.

When comparing all 15 strains against the C57BL/6J reference, we identified a total of 2,424 CNV sites (1,259 gains and 1,958 losses). Of these CNV events in each strain, 56% are predicted as high-confidence intervals ($P > 0.8$), and of these, 85–92% are newly identified (**Table 2, Supplementary Tables 6**, **7** online). Most of the variation in segmental duplications was not detected previously, as probes were underrepresented tenfold when compared to unique regions and 50-fold when compared to our C57BL/6J duplication-specific microarray (**Supplementary Table 8** online)[10]. Even among the confirmed sites of CNV, we observed significantly more substructure than previously reported, revealing a complex pattern of copy number gain and loss associated with mouse segmental duplications (**Figs. 1** and **4b**). We note that our HMM approach is particularly conservative on boundary definition and consequently overfragments genomic regions by an estimated factor of 2 (**Supplementary Note**). Nevertheless, we identified over 182 large intervals (>100 kb) of copy number loss and gain (**Figs. 1** and **4** and **Supplementary Fig. 2** online). Overall, on the basis of our survey,

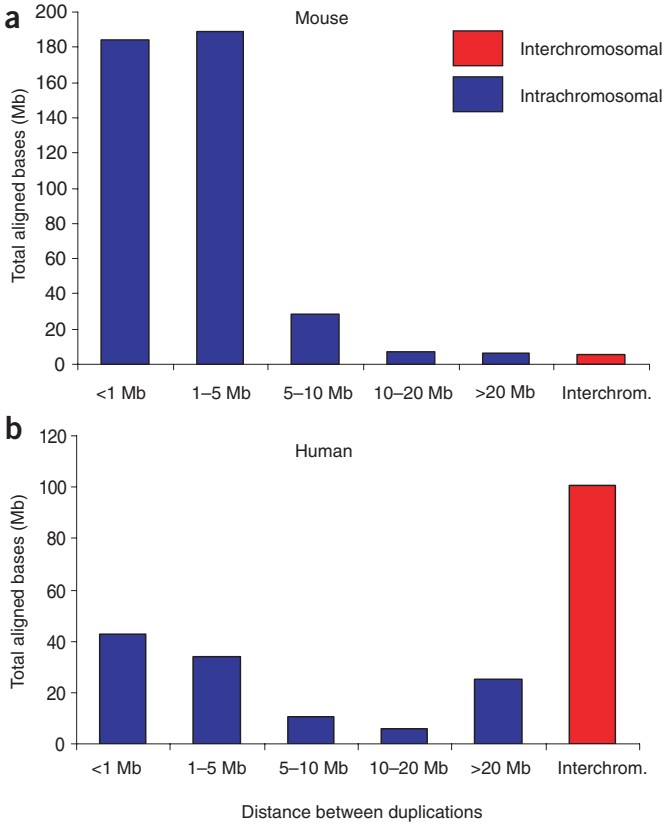

Figure 2 Distribution of mouse versus human duplication pairwise alignments. The distance between segmental duplications was computed for the mouse (Build36) and the human (Build36) genome. All pairwise alignments >10 kb in length were binned into various categories. Tandem duplications that map within 5 Mb of one another constitute the bulk of mouse segmental duplications.

Table 2 Mouse CNV regions mapping to segmental duplications

				CNV loss		CNV gain		All CNV (gain or loss)		CNV (gain and loss)
Regions	Genome content (Mb)	Probe count	Probe density (bp/probe)	NRS Mbp in all strains (%)	Avg. (Mbp per strain) NRS (%)	NRS Mbp in all strains (%)	Avg. (Mbp per strain) NRS (%)	NRS Mbp in all strains (%)	Avg. (Mbp per strain) NRS (%)	NRS Mbp in all strains (%)
SD	97.86	203,307	481	38.63 (39.5)	12.73 (13.0)	26.46 (27.0)	6.38 (6.5)	56.89 (58.1)	19.0 (20.5)	8.21 (8.4)
Unique flanking sequence	22.94	54,199	423	3.9 (17.0)	1.01 (4.4)	2.87 (12.5)	0.66 (2.9)	6.12 (26.7)	1.65 (7.9)	0.65 (2.8)
CNV in SD	49.71	127,520	390	8.54 (17.2)	3.39 (6.6)	5.93 (11.9)	2.04 (4.1)	12.79 (25.7)	5.31 (10.9)	1.68 (3.4)
CNV in unique	37.94	105,692	359	1.83 (4.8)	0.61 (1.6)	1.53 (4.0)	0.45 (1.2)	3.21 (8.5)	1.05 (3.0)	0.16 (0.4)
All probe regions	159.4	385,206	414	44.37 (27.8)	14.30 (9.0)	30.78 (19.3)	7.43 (4.7)	66.18 (41.5)	21.54 (14.4)	8.98 (5.6)

NRS, nonredundant sequence in Mbp; SD, segmental duplications (WGAC and WSSD combined); unique flanking sequence, 10 kb of unique sequence flanking SD. CNV gains and losses based on HMM analysis. The 'CNV (gain and loss)' column shows regions with evidence of both gains and losses on the basis of array CGH of 15 tests strains against C57BL/6J. CNV regions that were previously identified[7] within segmental duplications (CNV in SD) or unique regions that did not intersect SD (CNV in unique) are indicated.

we predict that 61.6% of the segmental duplications are variable in copy number with, on average, 20 Mb of duplication for any strain showing copy number difference when compared to C57BL/6J.

We identified 353 genes embedded within segmental duplications that showed either gain or loss (**Supplementary Table 6**). Of these, 194 CNV intervals are sufficiently large enough to affect the entire gene, including 31 genes showing both gains and losses in different strains with respect to C57BL/6J (**Supplementary Table 6**). Several of the copy-number-variant genes are associated with spermatogenesis, pregnancy, and viviparity (for example, *Slx* (also known as *Xmr*, *Tcte, Ott, Prl, Plf* and *Il11ra*). Other gene families associated with pheromone response show large-scale CNV between the strains (for example, vomeronasal receptor (*V2r* and *V1r*) and major urinary proteins (*Mup*) gene families). As in the human genome, immune response genes in the mouse genome show extensive copy number polymorphism. For example, the defensin genes (*Defcr21/22/23* and *Defcr5*), the neuronal apoptosis inhibitory protein (*Naip*) gene family and the killer cell lectin-like receptor family a (*Klra*) gene members are all part of CNV duplication blocks associated with strain variability to infection[25].

Although similar in proportion (~5%), recent mouse genomic duplications, in contrast to those of humans, are organized into discrete clusters of tandem duplications that are depleted for genes and transcripts and enriched for LINE and LTR retroposons. We hypothesize that the strong association with younger LINE elements, as opposed to primate Alu-SINE elements, might explain some of the key differences between human and mouse duplications. For example, LINE repeat sequences preferentially map to AT-rich, gene-poor regions as a result of the sequence preference of the RT-endonuclease[26]. Similar bias against genes has been observed for LTR elements[27]. If LINE and LTR sequences promote segmental duplication, it may explain why there is a deficiency of genes and transcripts in mice, whereas in humans the trend is in the opposite direction (that is, segmental duplications associate with SINE-rich, gene-rich regions of the genome)[26]. In addition, we find that mouse duplicated sequences have three to four times as many paralogs when compared to human duplicated sequences. We conservatively estimate that at least 20 Mb of segmental duplication is copy number variable between strains (**Table 2**). When compared to recent surveys of copy number variation in humans[28,29], we find that different strains of mice show as much, if not more, copy-number-variable DNA within the duplicated regions. We propose that the larger number of local pairwise alignments in tandem orientation within the mouse increases the potential for nonallelic homologous recombination and,

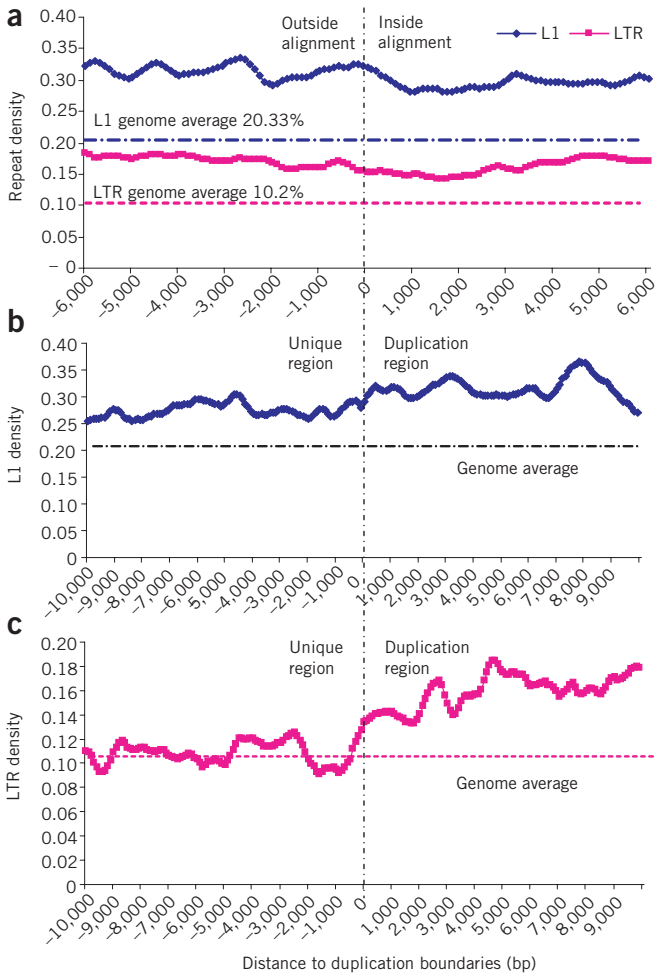

Figure 3 LINE and LTR enrichment within mouse segmental duplications. (**a**) We examined all large pairwise alignments (>20 kb) and computed the LINE and LTR content (in 500-bp windows; sliding increments of 100 bp) on either side of the alignment boundary as determined by the whole-genome analysis comparison method. Segmental duplications are significantly enriched for both LINE and LTR repeats. (**b**,**c**) We next examined all transition regions where there was at least 10 kb of unique sequence abutting segmental duplication (n = 5,325 alignments) and computed the LINE (**b**) and LTR (**c**) content on either side of the transition boundary between the unique and duplicated sequences. LTR repeat sequences show specific enrichment for segmental duplications when compared to unique transition regions, whereas both the flanking unique and duplicated regions were enriched for LINE repeats.

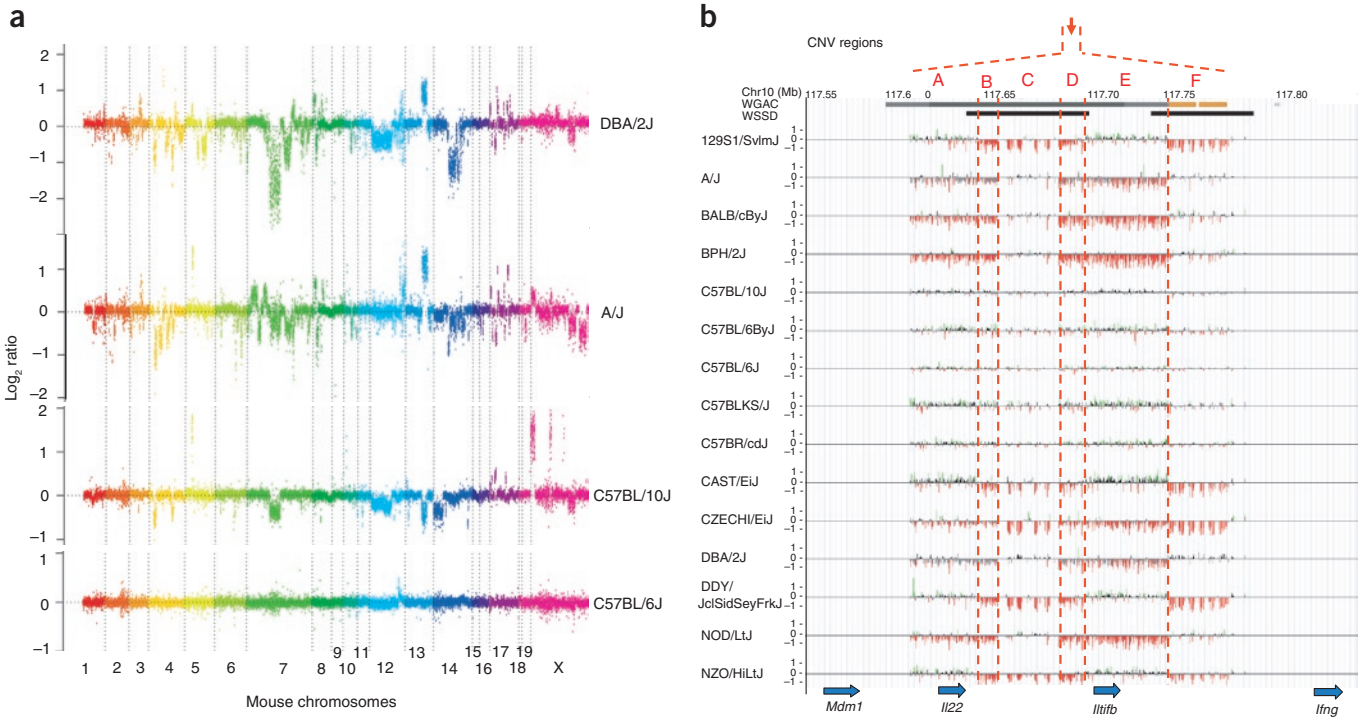

Figure 4 Copy-number variable mouse segmental duplications. (**a**) Underlying array comparative genomic hybridization data are shown for four strains compared to C57BL/6J. Segmental duplication and flanking regions (159 Mb) were ordered and collapsed according to chromosomal position (color). (**b**) A ~170-kb segmental duplication region on chromosome 10 shown from the browser (see URLs section in Methods) in more detail for 15 different mouse strains. Significant (>1.5 s.d.) decreases (red) and increases (green) are highlighted. At least six distinct regions (A–F) of copy number variation can be discerned within the duplication block (WGAC, whole-genome assembly comparisons; WSSD, whole-genome shotgun sequence detection). Regions A and E represent high-identity duplications of the interleukin 22 gene and, therefore, the array-CGH signal represents the average differential of both regions and the array-CGH patterns mirror one another. See **Supplementary Figure 2** for additional examples.

thus, the mutation frequency. In this regard, it is interesting that of the 15 CNVs that intersect with those found by a previous study, 14/15 were shown to occur recurrently within mouse strains[10]. Those with the highest frequency of new mutation (~1 spontaneous mutation per 100 newborns) are composed almost entirely (77–92%) of segmental duplications (**Supplementary Table 7**). Further studies of the normal pattern of copy number variation within wild outbred lines of mice and sequencing of additional murid genomes will be necessary to assess the generality of these findings.

METHODS

DNA samples. We obtained all spleen-derived DNA samples from male individuals representing 15 inbred strains of mice (Jackson Laboratory). These included C57BL/6J, DBA/2J, A/J, C57BL/10J, CZECHI/EiJ, CAST/EiJ, BPH/2J, BALB/cByJ, C57BLKS/J, 129S1/SvlmJ, DDY/JclSidSeyFrkJ, C57BR/cdJ, C57BL/6ByJ, NZO/HiLtJ and NOD/L5J. The reference sample in all these experiments was C57BL/6J (Prep#37347, a G227 male individual born 4 October 2005). As a control, we carried out an array-CGH experiment against a second C57BL/6J individual (Prep#37579, a G230 male individual born 27 September 2006). Inbred strains were selected in an effort to sample genetic diversity[30] and to include strains from the Mouse Phenome Project and NIEHS sequencing projects.

Segmental duplication characterization. We used two independent approaches to detect segmental duplications: WGAC (whole-genome assembly comparison), which is a BLAST-based analysis of all assembled sequence that detects self alignments (>90% and 1 kb); and WSSD (whole-genome shotgun sequence detection), which is an assembly-independent approach that examines the reference sequence for an increase in WGS read depth-of-coverage (WSSD-DOC) and/or increase in the divergence read ratio (WSSD-DRR). We mapped 40,782,208 sequence reads against the Build36 genome assembly as part of the mouse WSSD analysis. We estimated the duplication content of the mouse genome on the basis of the sum of low-identity WGAC (<94%) and high-identity WGAC (>10 kb, >94%) that were confirmed by the union of WSSD-DOC and WSSD-DRR estimates. Repeat content and subfamily designation was determined using RepeatMasker. Significance was determined by permutation (randomly sampling the genome and computing an enrichment greater or equal to that observed within regions classified as segmentally duplicated). All underlying segmental duplication analysis data are available online (see URLs section below) and have been placed as customized tracks on the University of California Santa Cruz browser and the NCBI MapViewer for Build36.

Array comparative genomic hybridization and CNV detection. We designed a customized oligonucleotide microarray platform for array comparative genomic hybridization (NimbleGen). We targeted 385,000 probes to 159.4-Mb regions of the mouse genome assembly (Build36) where segmental duplications and/or CNVs were previously identified, as indicated in **Table 2**. Probe design and the sample hybridization were done at NimbleGen using standard tiling array protocol. We identified copy-number-variant regions between mouse strains using a novel HMM (see **Supplementary Note** for detailed description and software availability).

URLs. Mouse Paralogy Server, http://mouseparalogy.gs.washington.edu/.

Accession codes. NCBI Gene Expression Omnibus: microarray data have been deposited under accession number GSE11369.

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

ACKNOWLEDGMENTS

We thank L. Rowe, C. Birkenmeier and G. Churchill for providing additional information regarding the relatedness of different inbred strains of mice used in this study. We thank A. Morrison for DNA sample preparation. We thank T. Brown, K. Augustyn and H. Mefford for assistance in preparation of this manuscript. This

work was supported by US National Institutes of Health grant HG002385 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

X.S., E.E.E. and Z.C. analyzed the duplication content and organization as well as the array comparative genomic hybridization data. D.M.C. provided access to the mouse genome assembly, annotated duplication/gene content and provided detailed quality control assessment regarding status of these regions and misassembly issues. S.Z. developed a novel HMM algorithm to identify significant transitions in $\log_2$ ratios corresponding to a likely copy number gain or loss. E.E.E. and X.S. conceived of the analyses 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. Cheung, J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol.* **4**, R47 (2003).
2. Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M. & Eichler, E.E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14**, 789–801 (2004).
3. Bailey, J.A. & Eichler, E.E. in *Proceedings of the 68th Cold Spring Harbor Symposium: Genome of Homo sapiens.* (ed. Ebert, J.) Genome-wide detection of segmental duplication within mammalian organisms (Cold Spring Harbor Press, New York, 2003).
4. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
5. 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).
6. Sainz, J. *et al.* Segmental duplication density decrease with distance to human-mouse breaks of synteny. *Eur. J. Hum. Genet.* **14**, 216–221 (2006).
7. Li, J. *et al.* Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36**, 952–954 (2004).
8. Snijders, A.M. *et al.* Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res.* **15**, 302–311 (2005).
9. Graubert, T.A. *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet.* **3**, e3 (2007).
10. Egan, C.M., Sridhar, S., Wigler, M. & Hall, I.M. Recurrent DNA copy number variation in the laboratory mouse. *Nat. Genet.* **39**, 1384–1389 (2007).
11. Watkins-Chow, D.E. & Pavan, W.J. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome Res.* **18**, 60–66 (2008).
12. Bailey, J.A. & Eichler, E.E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
13. Nguyen, D.Q., Webber, C. & Ponting, C.P. Bias of selection on human copy-number variants. *PLoS Genet.* **2**, e20 (2006).
14. Armengol, L., Pujana, M.A., Cheung, J., Scherer, S.W. & Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12**, 2201–2208 (2003).
15. Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D. & Eichler, E.E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
16. Armengol, L. *et al.* Murine segmental duplications are hot spots for chromosome and gene evolution. *Genomics* **86**, 692–700 (2005).
17. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
18. Waterston, R. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
19. Bailey, J.A., Giu, L. & Eichler, E.E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
20. Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V. & Jurka, M.V. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc. Natl. Acad. Sci. USA* **101**, 1268–1272 (2004).
21. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
22. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
23. Perry, G.H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. USA* **103**, 8006–8011 (2006).
24. 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).
25. Lee, S.H. *et al.* Susceptibility to mouse cytomegalovirus is associated with deletion of an activating natural killer cell receptor of the C-type lectin superfamily. *Nat. Genet.* **28**, 42–45 (2001).
26. IHGSC. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
27. Medstrand, P., van de Lagemaat, L.N. & Mager, D.L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* **12**, 1483–1495 (2002).
28. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
29. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
30. Beck, J.A. *et al.* Genealogies of mouse inbred strains. *Nat. Genet.* **24**, 23–25 (2000).