The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains

Toyoyuki Takada,1,2 Toshinobu Ebata,3,4 Hideki Noguchi,4 Thomas M. Keane,5 David J. Adams,5 Takanori Narita,3 Tadasu Shin-I,3,4 Hironori Fujisawa,2,6 Atsushi Toyoda,4 Kuniya Abe,7 Yuichi Obata,7 Yoshiyuki Sakaki,8,9 Kazuo Moriwaki,7 Asao Fujiyama,4 Yuji Kohara,3 and Toshihiko Shiroishi1,2,10

1Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; 2Transdisciplinary Research Integration Center, Research Organization of Information and Systems, Minato-ku, Tokyo 105-0001, Japan; 3Genome Biology Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; 4Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; 5The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom; 6The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; 7RIKEN BioResource Center, Tsukuba, Ibaraki 305-0074, Japan; 8Genome Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan

Commonly used classical inbred mouse strains have mosaic genomes with sequences from different subspecific origins. Their genomes are derived from the Japanese subspecies Mus musculus molossinus, with the remaining sequences mostly derived from the Western European subspecies M. m. domesticus. The resequencing of two M. m. molossinus–derived inbred strains, MSM/Ms and JF1/Ms, originating from Japanese wild mice, revealed a vast number of SNPs and indels between the two strains, and compared these sequences to the C57BL/6J reference sequence (henceforth B6) and the genome sequences of other classical inbred strains that were generated in the Mouse Genome Project (MGP). The results unequivocally show that genome introgression from M. m. molossinus into M. m. domesticus provided the primary framework for the mosaic genomes of classical inbred strains. Furthermore, the genomes of C57BL/6J and other classical inbred strains have long consecutive segments of “Mendelism,” which suggests that the ancestor of the extant JF1/Ms strain provided the origin of the M. m. molossinus genome in classical inbred strains and largely contributed to its intersubspecific genome diversity.

[Supplemental material is available for this article.]

Classical inbred strains of mice were established in America in the early 20th century from European-derived fancy mice that were reared as pets (Morse 1981; Beck et al. 2000). It is well established that extant classical inbred strains are hybrids between multiple subspecies of Mus musculus, and they have a mosaic genome architecture comprised of sequences originating from different subspecies (Wade et al. 2002; Frazer et al. 2007; Yang et al. 2007). Recent high-resolution single nucleotide polymorphism (SNP) genotyping of wild-caught mice and a comparison of these sequences to those of classical inbred strains revealed that classical inbred strains are derived from a relatively small pool of fancy mice with limited haplotype diversity; their genomes are overwhelmingly derived from the Western European subspecies M. m. domesticus, with the remaining sequences mostly derived from the Japanese subspecies M. m. molossinus (Yang et al. 2011). Although detailed genealogical information is of great help for discovering the genes responsible for specific phenotypes, very little is known about how the intersubspecific genome introgression from M. m. molossinus into M. m. domesticus occurred during the establishment of classical inbred strains. In this study, we resequenced the genomes of two inbred strains, MSM/Ms and JF1/Ms, and compared these sequences to the C57BL/6J reference sequence (henceforth B6) and the genome sequences of other inbred strains. We resequenced the genomes of two inbred strains, MSM/Ms (henceforth MSM) and JF1/Ms (Japanese fancy mouse 1; henceforth JF1), and compared these sequences to the C57BL/6J reference sequence (henceforth B6) and the genome sequences of other inbred strains. These findings confirmed that fragments of the M. m. molossinus genome...
are scattered in classical inbred strains and comprise less than one-tenth of their genomes. The information of the nucleotide sequence variants obtained from this study will facilitate cloning genes responsible for phenotypic difference in classical inbred strains, which are attributable to the intersubspecific genome divergence.

This study also demonstrated that many genomic segments of classical inbred strains have extremely high sequence similarity to JF1, suggesting that the ancestor of JF1 could be the origin of the M. m. molossinus genome in classical inbred strains. This notion is further supported by early literature reporting that JF1-like Japanese waltzing mice were often crossed with European fancy mice for studies of the Mendelian inheritance of coat color and behavioral traits (Darbishire 1902; Yerkes 1907; Gates 1925, 1926; Schwarz 1942). Collectively, the findings from this study have unveiled the history of how the mosaic genomes of classical inbred strains were formed.

Results

The ancestors of the MSM strain were wild mice that were captured in Mishima, Shizuoka, Japan, in 1978 and established as an inbred strain at the National Institute of Genetics (NIG) (Morikawa et al. 2009). The ancestors of the JF1 strain were purchased in a pet market in Denmark and transferred to the NIG in 1987, where they were established as an inbred strain in 1993. Morphological and genetic characterization suggests that the JF1 strain was derived from M. m. molossinus (Koide et al. 1998). Genomic DNA samples were prepared from MSM and JF1, as well as all other inbred strains used in this study (listed in Supplemental Table S1), and using the Illumina GAII, we resequenced the genomic DNA isolated from the MSM and JF1 strains. We also generated close to 10 million shotgun Sanger reads from the MSM whole-genome sample, and we used the high-quality reads obtained using this method to clarify ambiguous calls from the short-read genome sequence data. For comparative analysis, we used each MSM and JF1 genome sequence in the same form as that of the reference strain (C57BL/6J; MGSC37 assembly); those data are also downloadable through the NIG mouse genome database (MSMv2 for MSM and JF1v1 for JF1).

The sequencing data statistics are summarized in Supplemental Table S2. The candidate SNPs were detected by aligning the MSM and JF1 sequences to the reference B6 strain (MGSC37; NCBI m37.1/mm9), as summarized in Supplemental Figure S1. The genotypes of 17 MGP strains at those SNPs are summarized in Supplemental Table S3. The candidate SNPs were detected by aligning the MSM and JF1 genomes. With respect to candidate structural variants, we identified 204 and 219 SNPs and indels found to generate premature stop codons in MSM and JF1, respectively, and some of these were found in alternative splicing variants. Of these, 125 were SNPs and indels common to the two strains (Supplemental Table S5). In comparison to the B6 reference sequence, the MSM and JF1 sequences contain 38,182 and 38,124 changes, respectively, and some of these were found in alternative splicing variants. Of these, 125 were SNPs and indels common to the two strains (Supplemental Table S5). In comparison to the B6 reference sequence, the MSM and JF1 sequences contain 38,182 and 38,124 nonsynonymous SNPs in 11,489 and 11,313 genes, respectively.

Next, using ANNOVAR (Wang et al. 2010), we analyzed the SNPs that introduce a nonsynonymous substitution or premature stop codon or cause the loss of a stop codon relative to the B6 reference sequence (Table 1; Supplemental Table S4). Then, we performed functional annotations of these SNPs using DAVID Bioinformatics Resources 6.7 tools (Huang et al. 2009a,b). The MSM strain was found to contain 205 stop codon gains and 39 stop codon losses, while the JF1 strain contained 217 gains and 43 losses. These two strains share 112 stop codon gains and 32 stop codon losses. Excluding in silico predicted genes, pseudogenes, and noncoding sequences, 204 and 219 SNPs and indels were found to generate premature stop codons in MSM and JF1, respectively, and some of these were found in alternative splicing variants. Of these, 125 were SNPs and indels common to the two strains (Supplemental Table S5). In comparison to the B6 reference sequence, the MSM and JF1 sequences contain 38,182 and 38,124 nonsynonymous SNPs in 11,489 and 11,313 genes, respectively.

We next explored the potential human disease-related phenotypes of the genes with nonsynonymous SNPs by searching the Online Mendelian Inheritance in Man (OMIM) database (Table 2). A total of 28 genes appeared to be disease-associated genes, and 24 of these were common between MSM and JF1. To examine whether each amino acid substitution would lead to a change in protein function, we calculated a GRANTHAM matrix score (GMS) (Grantham 1974) for each SNP. The GMS reflects differences in physicochemical properties between different amino acids and was calculated using an option for ANNOVAR that was released on October 23, 2012 (Wang et al. 2010). For both MSM and JF1, we found that 4.8% of the nonsynonymous substitutions were radical (GMS > 150), whereas 10.5% were moderately radical (100 < GMS ≤ 150). We annotated the functions of 1530 genes (for MSM) and 1529 genes (for JF1) with radical substitutions using DAVID, and the results showed a significant enrichment of genes associated with the ‘G-protein–coupled receptor’ and ‘receptor’ PANTHER molecular function categories for MSM and JF1 and with the ‘H2 antigen’ for JF1 only (Supplemental Table S6).

We next investigated the phylogeny of inbred mouse strains, including MSM and JF1, based on the present sequence data and a publicly available sequence data set. Because the simplest and fastest way to define loci within a chromosome alignment is to consider fixed-length intervals (Ané 2011), we segmented the B6 reference sequence into 26,398 100-kb blocks and compared each block with the corresponding sequences from the MSM, JF1, and WSB/Eij (henceforth WSB) strains. The inbred strain WSB is derived from wild M. m. domesticus (Frazer et al. 2007; Yang et al. 2007), and its sequence was generated from MGP (Keane et al. 2011). First, we found that the B6 strain had a large number of blocks with high sequence similarity to WSB, consistent with previous reports showing that the genomes of classical inbred strains are derived overwhelmingly from M. m. domesticus (Fig. 1; Wade et al. 2002; Frazer et al. 2007; Yang et al. 2007, 2011). In addition, a sequence reported BAC sequences (MSMg01-122K02 [GenBank AP007207] and MSMg01-275M02 [GenBank AP007208]) of MSM (Abe et al. 2004). We then compared these SNPs with those detected between the B6 reference sequence and the manually aligned repeat-masked MSM sequence generated in this study. This analysis confirmed 1322 SNPs and 130 indels, and the results are summarized in Supplemental Table S3. The concordance rates were 98.9% (1308/1322) for the SNPs and 94.6% (123 of 130) for the indels, and the false-positive and false-negative rates were 0.30% and 0.76% for the SNPs and 10.9% and 12.6% for the short (1–6 bp) indels, respectively.

Next, using ANNOVAR (Wang et al. 2010), we analyzed the SNPs that introduce a nonsynonymous substitution or premature stop codon or cause the loss of a stop codon relative to the B6 reference sequence (Table 1; Supplemental Table S4). Then, we performed functional annotations of these SNPs using DAVID Bioinformatics Resources 6.7 tools (Huang et al. 2009a,b). The MSM strain was found to contain 205 stop codon gains and 39 stop codon losses, while the JF1 strain contained 217 gains and 43 losses. These two strains share 112 stop codon gains and 32 stop codon losses. Excluding in silico predicted genes, pseudogenes, and noncoding sequences, 204 and 219 SNPs and indels were found to generate premature stop codons in MSM and JF1, respectively, and some of these were found in alternative splicing variants. Of these, 125 were SNPs and indels common to the two strains (Supplemental Table S5). In comparison to the B6 reference sequence, the MSM and JF1 sequences contain 38,182 and 38,124 nonsynonymous SNPs in 11,489 and 11,313 genes, respectively.

We next explored the potential human disease-related phenotypes of the genes with nonsynonymous SNPs by searching the Online Mendelian Inheritance in Man (OMIM) database (Table 2). A total of 28 genes appeared to be disease-associated genes, and 24 of these were common between MSM and JF1. To examine whether each amino acid substitution would lead to a change in protein function, we calculated a GRANTHAM matrix score (GMS) (Grantham 1974) for each SNP. The GMS reflects differences in physicochemical properties between different amino acids and was calculated using an option for ANNOVAR that was released on October 23, 2012 (Wang et al. 2010). For both MSM and JF1, we found that 4.8% of the nonsynonymous substitutions were radical (GMS > 150), whereas 10.5% were moderately radical (100 < GMS ≤ 150). We annotated the functions of 1530 genes (for MSM) and 1529 genes (for JF1) with radical substitutions using DAVID, and the results showed a significant enrichment of genes associated with the ‘G-protein–coupled receptor’ and ‘receptor’ PANTHER molecular function categories for MSM and JF1 and with the ‘H2 antigen’ for JF1 only (Supplemental Table S6).

We next investigated the phylogeny of inbred mouse strains, including MSM and JF1, based on the present sequence data and a publicly available sequence data set. Because the simplest and fastest way to define loci within a chromosome alignment is to consider fixed-length intervals (Ané 2011), we segmented the B6 reference sequence into 26,398 100-kb blocks and compared each block with the corresponding sequences from the MSM, JF1, and WSB/Eij (henceforth WSB) strains. The inbred strain WSB is derived from wild M. m. domesticus (Frazer et al. 2007; Yang et al. 2007), and its sequence was generated from MGP (Keane et al. 2011). First, we found that the B6 strain had a large number of blocks with high sequence similarity to WSB, consistent with previous reports showing that the genomes of classical inbred strains are derived overwhelmingly from M. m. domesticus (Fig. 1; Wade et al. 2002; Frazer et al. 2007; Yang et al. 2007, 2011). In addition, a sequence
Sliding window analysis. We found that the B6 sequence has long similarity of B6 and MSM or JF1 along each chromosome using Bayesian concordance analysis (BCA) (Ane´ et al. 2007; Ane´ 2011). The minimum rate was 3.32%, found in the A/J strain. The whole genome was 7.03%, which was found in the LP/J strain, and the maximum rate of ratio for the other classical inbred strains in the MGP data set (Supplemental Data S1, 2). The maximum rate of ratio for the genome introgression of the ancestors of classical inbred strains into JF1 occurred, as denoted by the horizontal red line in Figure 2A. We defined the genomic regions for which a single B6/MSM topology was supported with a higher posterior probability according to the BCA as molossinus-derived regions (MDRs) (Supplemental Fig. S6; Supplemental Table S10).

To further clarify the origins of the MDRs, we carried out PCR-based genome sequencing of various wild mouse-derived inbred strains at 67 selected regions residing in MDRs. A molecular phylogenetic tree constructed from the sequence data showed that the JF1 strain belongs to the same clade as B6 (Fig. 2B) and that the inbred strains PWD/Ph, PWK/Ph, and BLG/Ms, which are derived from Eastern European populations of M. m. musculus, belong to different clades than the B6 and M. m. molossinus–derived JF1 and MSM strains. These results clearly indicate that the MDRs are indeed derived from M. m. molossinus. The results of the genome partitioning and the genomic partitioning ratio of the phylogenetic history of the whole chromosomes, which were obtained by BCA, are shown in Supplemental Figures S6 and S7, respectively. We next calculated the nucleotide sequence similarity between B6 and MSM or JF1 in the MDRs defined by BCA. The average nucleotide sequence similarity (99.698%) between B6 and JF1 was higher than (99.535%) between B6 and MSM, and this result was further supported by the comparison of distribution plots of the B6-MSM and B6-JF1 similarities in MDRs (Supplemental Fig. S8).

Our finding that the JF1 strain tended to show higher sequence similarity to the reference B6 sequence than the MSM strain prompted us to analyze the B6 sequence blocks demonstrating an extremely high similarity to MSM and JF1, which likely reflects a recent introgression from M. m. molossinus into the founders of classical inbred strains. Then, we compared the frequency distributions of the sequence blocks with extremely high similarity to MSM and JF1. The number of B6 blocks with >99.998% sequence similarity to JF1 was significantly larger than 99.998% sequence similarity to MSM and JF1, as shown in PP in Figure 2A. The exceptions included genomic regions where intersubspecific genome introgression of the ancestors of classical inbred strains into JF1 occurred, as denoted by the horizontal red line in Figure 2A. We defined the genomic regions for which a single B6/MSM topology was supported with a higher posterior probability according to the BCA as molossinus-derived regions (MDRs) (Supplemental Fig. S6; Supplemental Table S10).

Table 1. The number of nonsynonymous (NS), synonymous (S), and premature termination (PMT) variants observed in the MSM and JF1 genomes relative to the reference B6 genome

| Chromosome | C57BL/6J Total Gene no. | NS | NS total | S | S total | Stop_gain | Stop_loss |
|------------|-------------------------|----|----------|---|---------|-----------|-----------|
| 1          | 1298                    | 632| 625      | 2488| 2486    | 867       | 879       |
| 2          | 1984                    | 950| 923      | 2849| 2829    | 1296      | 1312      |
| 3          | 1097                    | 533| 537      | 1794| 1831    | 756       | 775       |
| 4          | 1350                    | 661| 659      | 2207| 2244    | 902       | 920       |
| 5          | 1313                    | 704| 703      | 2430| 2421    | 997       | 998       |
| 6          | 1235                    | 637| 625      | 2198| 2186    | 860       | 875       |
| 7          | 2111                    | 1183| 1215 | 4202| 4399    | 1545      | 1593      |
| 8          | 1103                    | 440| 331      | 1369| 1076    | 622       | 467       |
| 9          | 1310                    | 695| 693      | 2097| 2118    | 935       | 960       |
| 10         | 1083                    | 536| 540      | 1708| 1661    | 789       | 800       |
| 11         | 1745                    | 867| 836      | 2630| 2496    | 1257      | 1252      |
| 12         | 791                     | 291| 300      | 976 | 997     | 443       | 450       |
| 13         | 886                     | 403| 393      | 1443| 1473    | 573       | 578       |
| 14         | 835                     | 389| 393      | 1137| 1186    | 501       | 512       |
| 15         | 866                     | 450| 448      | 1534| 1535    | 635       | 640       |
| 16         | 726                     | 388| 376      | 1465| 1418    | 510       | 521       |
| 17         | 1129                    | 565| 541      | 1982| 1977    | 735       | 726       |
| 18         | 570                     | 299| 297      | 1227| 1227    | 426       | 430       |
| 19         | 758                     | 417| 417      | 1258| 1288    | 589       | 598       |
| X          | 1027                    | 448| 461      | 1208| 1276    | 509       | 604       |
| Y          | 13                      | —  | —        | —  | —       | —         | —         |
| Un         | 25                      | —  | —        | —  | —       | —         | —         |
| Total      | 23,255                  | 11,489| 11,313 | 38,182| 38,124 | 15,807 | 15,890 |

...
Discussion

We sequenced the whole genomes of two *M. m. molossinus*-derived inbred mouse strains, MSM and JF1, using next-generation sequencing technology. We also sequenced the MSM genome using the capillary sequencing method. Upon comparing these sequences to the B6 reference sequence, we identified about 15 million high-confidence SNPs (Supplemental Fig. S1). A large number of intersubspecific SNPs and indels detected in this study underlie the large phenotypic differences between *M. m. domesticus* and *M. m. molossinus* (Takada et al. 2008; Takahashi et al. 2008; Koide et al. 2011; Takada and Shiroishi 2012). In addition, we identified a much larger number of nonsynonymous SNPs leading to radical amino acid substitutions (GMS > 150) than previously detected between the FVB/NJ and B6 laboratory strains (Wong et al. 2012), and this is likely due to the intersubspecific genome divergence between *M. m. molossinus* and *M. m. domesticus*, from which the genomes of the classical laboratory strains are predominantly derived. We also detected MSM- or JF1-type variants that contribute to the phenotypic differences between these two strains. For example, the MSM, but not JF1, genome contains a SNP that causes a premature stop codon in the C8a gene, which encodes the eighth component of serum complement (Supplemental Table S5). Indeed, the MSM strain was reported to carry this mutation and to have a deficiency in C8 activity (Tanaka et al. 1991).

Yang et al. (2011) reported that the origins and compositions of the genomes of classical inbred strains depend on the use of wild-derived inbred strains as reference genomes to infer subspecific origin. This is because some wild-derived inbred strains suffered intersubspecific genome introgression from classical inbred strains, which likely occurred in the laboratory. Thus, such strains are not suitable as reference strains for the subspecies.

Table 2. Human disease-associated nonsynonymous single nucleotide variants in the MSM and JF1 genomes (DAVID analysis)

| Chromosome | Gene name | Descriptions | GeneBank_ID | OMIM_DISEASE |
|------------|-----------|--------------|-------------|--------------|
| 1          | Gpr161    | G protein–coupled receptor 161 | NM_001081126 | Vacuolated lens |
| 1          | Rd3       | Retinal degeneration 3 | NM_023727 | Retinal degeneration 3 |
| 1          | Sgk3      | Serum/glucoorticoid regulated kinase 3 | NM_001037759 | Fuzzy |
| 1          | Freem2    | Fras1 related extracellular matrix protein 2 | NM_172862 | Myelencephalic blebs |
| 1          | Chd7      | Chromodomain helicase DNA binding protein 7 | NM_001081417 | Cyclone, dizzy, eddy, leda, orbito, tornado, whirligig |
| 2          | Cdkn2a    | Cyclin-dependent kinase inhibitor 2A | NM_001040654 | Plasmatypoma resistance 1 |
| 2          | Mtf1      | Metal response element binding transcription factor 1 | NM_008636 | Thymic lymphoma susceptibility |
| 2          | Coor1a    | Coronin, actin binding protein 1A | NM_009898 | Lupus in MRL and B6 F2 cross, QTL 3 |
| 2          | Odz4      | Similar to DOC4; odd Oz/ten-m homolog 4 (Drosophila) | NM_011858 | Lethal, Chr 7, Rinchik 3 |
| 3          | Lgi4a     | Leucine-rich repeat LGI family, member 4 | NM_144556 | Claw paw |
| 3          | Afg3l2    | a Similar to AFG3 (ATPase family gene 3)-like 2 (yeast); FGS1-like 1 (S. cerevisiae) | NM_001093749 | Rough coat |
| 3          | Lmf1      | Lipase maturation factor 1 | NM_001168289 | Kidney disease |
| 3          | Acp1      | Adrenocortical dysplasia | NM_175407 | Jackson cicler |
| 4          | Smpd3b    | Sphingomyelin phosphodiesterase 3, neutral | NM_139061 | Wobbler |
| 4          | Rd3       | Retinal degeneration 3 | NM_144788 | Open mind |
| 4          | Dock5     | Dedicator of cytokinesis 5 | NM_177780 | Rupture of lens cataract |
| 4          | Fndc3a    | Fibronectin type III domain containing 3A | NM_207636 | Symplastic spermatids |
| 4          | Mtap7     | Microtubule-associated protein 7 | NM_009699 | Congenital progressive hydronephrosis |
| 5          | Vps54     | Vacuolar protein sorting 54 (yeast) | NM_153533 | Nephrosis |
| 5          | Lpap1     | Lipase, member H | NM_153404 | Lipid defect 1 |
| 5          | Lmp1      | Lipase maturation factor 1 | NM_029624 | Combined lipase deficiency |
| 6          | Afg3l2a   | Similar to AFG3 (ATPase family gene 3)-like 2 (yeast); AFG3(3-ATPase family gene 3)-like 2 (yeast) | NM_027130 | Paralyze |
| 6          | Scy1      | SCY1-like 1 (S. cerevisiae) | NM_024392 | Muscle deficient |
| 6          | Exoc6     | Exocyst complex component 6 | NM_175353 | Hemoglobin deficient |
| X          | Nhs       | Nance-Horan syndrome (human) | NM_001081052 | X-linked cataract (model for Nance-Horan syndrome) |

*Observed only in JF1/Ms data.

*Observed only in MSM/Ms data.
In this study, we used the MSM strain as a reference for *M. m. molossinus*. Importantly, we obtained a complete pedigree record of past MSM inbreeding generations because the founders were captured from a wild population. Therefore, it is highly unlikely that the MSM strain contains introgressed genomic regions from other strains.

Our study clearly showed that the genomes of classical inbred strains are overwhelmingly composed of sequences from *M. m. domesticus*, with the remaining sequences mostly derived from *M. m. molossinus*, which supports the recent SNP-based high-resolution genotyping of wild-caught mice (Yang et al. 2011). However, the majority of the genome of highly domesticated fancy mouse–derived JF1 mice originated from *M. m. molossinus*, but a small fraction of its genome was introgressed from the ancestors of classical inbred strains. Although the original sequence of *M. m. molossinus* composes <10% of the genomes of classical inbred strains, intersubspecific divergence led to a disproportionately large contribution from the *M. m. molossinus* genome to the total genome diversity in classical inbred strains. We estimate that ~30%–40% of the SNPs detected in pairwise comparisons of classical inbred strains are attributable to *M. m. molossinus* genome introgression (Supplemental Fig. S10). Thus, the SNPs and structural variants we detected should facilitate the discovery of genes underlying specific phenotypes in classical inbred strains. Furthermore, the JF1 and MSM strains are frequently used for studies of genomic imprinting and epigenetics (Tasi et al. 2002; Hirasawa et al. 2008) because the large genetic distance between these strains and classical inbred strains allows researchers to mark alleles at almost any locus of interest. Thus, the vast number of nucleotide variants detected in this study could facilitate studies in many relevant fields.

Collectively, our data demonstrate that the genome introgression from *M. m. molossinus* into *M. m. domesticus* constitutes the primary framework for the mosaic genomes of classical inbred strains. Furthermore, our data unequivocally show that the ancestors of the JF1 strain introduced the *M. m. molossinus* genome into classical inbred strains. JF1 has a recessive piebald (*s*) allele for the endothelin receptor type B gene (*Ednrb*), which is responsible for the spotted coat color also found in the classical inbred strain SSL/Le (Hosoda et al. 1994). Because literature published in Japan in 1787 described a small mouse with a piebald-like coat color (Tokuda 1935), the origin of the JF1 strain is likely the early Japanese fancy mouse. Our previous study also indicated that the JF1 strain displays *molossinus*-specific polymorphisms in its mitochondrial DNA and MHC class I gene (Koide et al. 1998), consistent with the SNP-based genotyping results in this study. Thus, the JF1 strain has a genome derived from *M. m. molossinus*, which supports the notion that it was likely reared as a pet in Japan in the 18th century before being transported to Europe in the middle to late 19th century, where its genome was introduced into European fancy mice for early studies of the Mendelism of coat color and waltzing behavior. The descendants of these mice were then transported to America (Keeler 1931; Morse 1981) and established as the classical inbred strains by the pioneers of mouse genetics, such as W. E. Castle and C. C. Little (Fig. 5; Morse 1981).

However, we could not detect B6 genomic regions with extremely high sequence similarity (>99.998%) to JF1 in a few regions derived from the *M. m. molossinus* genome (Fig. 3B), which
may be because the JF1 ancestor had heterozygous haplotypes or because some of the sequences that included a mutation responsible for waltzing behavior are extinct in the present JF1 genome.

This study showed that a vast amount of nucleotide sequence variants scattered in the genomes of classical inbred strains are concentrated in genomes of a single strain JF1 or MSM, indicating that a single genetic cross between a classical inbred strain and the MSM or JF1 strain provides a parsimonious platform for genetic and epigenetic analyses of a wide-range of complex traits, which would otherwise require many different crosses between classical inbred strains.

Methods

All Supplemental Material (Supplemental Data 1–3; Supplemental Tables S7–S12) is available on the FTP site of the NIG Mouse Genome Database (ftp://molossinus.lab.nig.ac.jp/pub/msmdb/Takada_et_al_2013).

Samples of genomic DNA

We resequenced genomic DNA isolated from the MSM and JF1 strains, which were maintained as pedigreed breeding stocks at the NIG. We also obtained complete pedigree records of past inbreeding generations for both strains. The wild mouse-derived strains MSM, JF1, KJR/Ms, CHD/Ms, BLG2/Ms, PGN2/Ms, and HMI/Ms were established and maintained at the NIG (http://www.shigen.nig.ac.jp/mouse/strain/). Samples of genomic DNA from the PWD/Ph and PWK/Ph strains were kind gifts from Prof. J. Forejt of the Institute of Molecular Genetics, ASCR, Czech Republic. Genomic DNA samples from the AIZ/Stm, KOR1/Stm, KOR5/Stm, and KOR7/Stm strains were kind gifts from Dr. Y. Matsushima of the Saitama Cancer Center, Saitama, Japan. Genomic DNA samples from the Mae/Stm, STM1/Stm, STM2/Stm, and MOM/Nga strains were obtained from the RIKEN BioResource Center (Tsukuba, Japan). Genomic DNA samples from the B6 and C57BL/10J strains were obtained from Jackson Laboratory. The genomic DNA samples used in this study are listed in Supplemental Table S1. All animal experiments were approved by the Animal Care and Use Committee of the NIG.

Sequence data generated using next-generation sequencing technology

For resequencing, we used an Illumina Genome Analyzer II (Illumina) according to the manufacturer’s protocols. Sequence lane data were mapped individually using BWA (v0.5.5) and SAMtools (v0.1.7a) for MSM reads and BWA (v0.5.7) and SAMtools (v0.1.7a) for JF1 reads. The MSM reads were also mapped individually onto the B6 reference sequence using maq-0.7.1 with the following parameters: −a 250 (maximum distance between
two paired reads) and \(-m 0.01\) (rate of the difference between reads and references).

Capillary sequencing of the MSM genome

Genomic DNA was fragmented mechanically, and size-fractionated fragments were used to generate a shotgun-sequencing library set. Approximately 10 million DNA sequences encompassing 6.4 billion bp of the mouse genome were generated using enzymatic dideoxy chain termination chemistry with automated ABI3700 or ABI3730 sequencers (Life Technologies). Seventy-five percent of the reads were generated at the NIG DNA Sequencing Center, and 23% were produced at the RIKEN Genome Science Center. The remainder of the sequence was derived from the paired-ends of the MSM/Ms_BAC-end sequences (Abe et al. 2004). Sequence base calling using \textit{phred} v0.020425 (http://www.phrap.org/index.html), quality clipping, and screening for paired-ends of vectors were performed using Crossmatch. Sequences of \(>300\) bp with average \textit{phred} scores >20 were subjected to repeat detection using RepeatMasker (version 3.1.6; http://www.repeatmasker.org/). Non-repeat sequences >100 bp were used for SNP detection.

Detection of SNPs and indels between MSM/JF1 and B6

SNPs and indels were detected in the Illumina sequencing data. Sequence reads with a read coverage of greater than three and less than 30 were used for the MSM strain, and sequence reads with more than 15 and less than 59 reads were used for the JF1 strain.

Figure 3. The 100-kb B6 blocks with extremely high sequence similarity to the MSM and JF1 strains. (A) The number of 100-kb B6 blocks with extremely high sequence similarity to the MSM and JF1 strains. The B6 blocks were compared to their counterparts in the MSM and JF1 strains for each 0.001% block from 99.990% to 100% similarity. (B) The chromosomal locations of the 100-kb B6 blocks with sequence similarity \(>99.998\%\) to the MSM and JF1 chromosomes. Horizontal black boxes depict the regions with \(>99.85\%\) sequence similarity to the MSM strain. Gray boxes indicate gaps in the B6 reference sequence.
For the MSM data, SNPs and short indels (in the 1- to 6-bp range) were detected using BWA (v0.5.5) and SAMtools (v0.1.7a) software for the MSM/Ms reads and BWA (v0.5.7) and SAMtools (v0.1.7a) software for the JF1 reads. For MSM reads, capillary sequence data with phred scores greater than 30 were used to make high-quality nucleotide sequence calls. For JF1 reads, only SNPs with a heterozygote allele balance between 30% and 80% with a QV > 20 for the reference sequence were included in the final call set for the subsequent analysis.

Quality control

The false-positive rates of SNP discovery for the MSM and JF1 sequences were estimated by genotyping 186 randomly selected SNPs from the MSM, JF1 and B6 data using the Sequenom MassARRAY iPLEX Gold Assay (Sequenom). The information used for genotyping are listed in Supplemental Table S7. PCR products of up to 300 bp were analyzed using MALDI-TOF mass spectrometry with a 384-spot format SpectroCHIP. The data were recorded and interpreted using MassARRAY software (Sequenom). To estimate the false-positive and false-negative rates of nucleotide variants (SNPs and 1- to 6-bp indels), we compared our sequence data to the previously reported BAC sequences (MSMg01-122K03 and MSMg01-275M02) of the MSM strain (Abe et al. 2004). The nucleotide variant calls made by comparing the manually aligned repeat-masked multiple sequences of the MSM BACs and the MGSC37 reference B6 sequence of the corresponding regions were considered high-confidence SNPs and indels (Supplemental Table S3). The sequences of the PCR primers and other information pertaining to the validation of randomly selected indels in the range of 7–210 bp are shown in Supplemental Table S8. A total of 232 candidate indels (100 insertions and 132 deletions) were confirmed as true indels.

Functional annotation of SNPs

ANNOVAR (Wang et al. 2010) and DAVID Bioinformatics Resources 6.7 tools (Huang et al. 2009a,b) were used to characterize single nucleotide variants detected between the B6 and MSM or JF1 strains (Supplemental Data 3). For the analysis of genes with radical amino acid substitutions, a list of GenBank identifications (1530 genes for MSM and 1529 genes for JF1) was submitted to the DAVID website. We eliminated SNPs, in silico predicted genes, pseudogenes, and noncoding genes from the functional annotation.

SNPs detected by comparing the MSM and JF1 sequences to the MGP data set

Sequence reads for the MSM and JF1 strains were compared to the MGP data set. The sequences of 17 inbred strains, including 129P2, 129S1/SvImJ, 129S5, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CAST/Eij, CBA/J, DBA/2J, LP/J, NOD/ShiLtJ, NZO/HiLtJ, PWK/PhJ,
The following wild mouse-derived inbred strains were used for genome introgression from M. m. molossinus into classical inbred strains. European fancy mice originated from M. m. domesticus. In the late 18th century, a Japanese publication entitled “Chingan-sodategusa,” which means “How to breed fancy mice” (Tokuda 1935), reported small and spotted (piebald) mice reared by Japanese fanciers (lower right; courtesy of Kouwa-shypunyan, Tokyo, Japan). In the middle to late 19th century, British traders likely introduced Japanese waltzing mice carrying the “piebald” (EdnrB) mutant allele to Europe. The ancestor of JF1, which was referred to as the Japanese waltzing mouse, was used for early studies of the Mendelism of its coat color and waltzing behavior. The mouse with the piebald phenotype (“a” in the top right photo) resembles the JF1 mouse (photograph courtesy of Carnegie Institution for Science, Washington DC, USA). Experimental crosses of the JF1 ancestor and European fancy mice conveyed the M. m. molossinus genome into the M. m. domesticus genetic background. Later, their descendants were transported to America (Keeler 1931; Morse 1981), where they were established as classical inbred strains.

Calculation of percentage of sequence similarity
We divided the repeat-masked B6 reference sequence (MGSC37) into 26,398 100-kb blocks. To avoid misidentifying SNPs due to the incorrect assembly of regions with copy number variation (CNV), we used the cnD program (Simpson et al. 2010) to detect candidate regions of CNV (>1 kb) (for the results of the cnD candidate regions, see Supplemental Table S9), and the SNPs in these CNV regions were omitted from the SNP detection. After eliminating the CNV regions (>1 kb using cnD), we calculated the percentage similarity of each block of the B6 reference sequence to the sequences of MSM and JF1, as well as the entire MGP data set individually. The distribution of the number of blocks with a given sequence similarity is shown in Figure 1 and in Supplemental Data 1 and 2. The profiles illustrating the percentage of sequence similarity along representative chromosomal regions are shown in Figure 3.

Phylogenetic analysis
A previously reported BCA method (White et al. 2009; Keane et al. 2011) was used, with slight modifications. Consensus sequences from the MSM, WSB, and SPRET/Eij strains were mapped to the alignment, and gaps were filled with Ns. Collinear blocks were partitioned into 124,174 loci using a minimum description length algorithm with a default maximum cost (shown in Supplemental Table S10).

Construction of a phylogenetic tree of MDRs
The following wild mouse-derived inbred strains were used for PCR-based resequencing: the M. m. musculus-derived PWD/Phj, PWK/PhJ, BLG2/Ms, CHD/Ms, and KJR/Ms strains; the M. m. domesticus-derived PGN2/Ms strain; the M. m. castaneus-derived HMI/Ms strain; and the Japanese fancy mouse-derived JF1 strain. For this analysis, we selected 67 MDRs with an average size of 683 bp and a total size of 45,761 bp. The sequences of the PCR primers used for amplification of the 67 regions are listed in Supplemental Table S11. DNA fragments representing the selected regions were obtained by PCR amplification of genomic DNA samples from two individuals of each strain.

The PCR products were sequenced according to the standard method described above using an ABI3700 or ABI3730 capillary sequencer. SNPs identified in the sequence data with a QV < 30 were excluded from analysis. The nucleotide sequence of each strain’s genome was searched against the B6 sequence using BL2seq5 to determine its similarity. In this analysis, we used nucleotide sequence data with a QV ≥ 30, and indel data were omitted. Subsequently, we constructed a molecular phylogenetic tree of these strains based on the sequence data for the B6 and MSM regions with a high sequence similarity. All sequences for each strain were merged, and a neighbor-joining phylogenetic tree was constructed from overlapping alignments using the program MEGA4 (Tamura et al. 2007).

SNP-based genotyping using the MassARRAY system
SNP genotyping was carried out for the 102 nucleotide sites listed in Figure 4 using the MassARRAY iPLEX Gold Assay (Sequenom). The information used for genotyping are listed in Supplemental Table S12.

Data access
The sequence data from this study have been submitted to the DDBJ Sequence Read Archive (DRA) (http://trace.ddbj.nig.ac.jp/dra/index_e.shtml) under accession numbers DRA000194 for MSM and DRA000323 for JF1. The BAM files containing the sequences of the two strains are downloadable through the NIG mouse genome database (ftp://molossinus.lab.nig.ac.jp/pub/msmdb/Takada_et_al_2013). Other sequence data by Sanger reads have been submitted to the DDBJ under accession numbers BAAG010000001–BAAG011237600 for MSM and DE993413–DE995782 for the phylogenetic studies. The sequence data of MSM and JF1 are also available through the NIG mouse genome database (http://molossinus.lab.nig.ac.jp/msmdb/), with side-by-side comparison to the B6 reference sequence MGSC37.

Acknowledgments
We thank J. Forejt and Y. Matsushima for providing genomic DNA samples of wild mice–derived strains. We thank M. White, C. Ané, B. Paysere, H. Fujimiya, N. Osada, N. Niinaya, and Y. Nakamura for valuable suggestions regarding statistical analysis and bioinformatics; Y. Yamazaki for the management of the NIG mouse genome database server; and the Information Technology Unit for
use of the supercomputer system of the NIG. We also thank A. Mita for careful long-term maintenance of the MSJM and JF1 strains at the NIG, and K. Yamabe, N. Harima, M. Naganuma, H. Watanabe, K. Nishimoto, K. Masuyama, and A. Matsuse for technical assistance. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas “Comparative Genomics” and “the Genome Information Upgrading Program” in the National BioResource Projects (NBRP) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. This work was also supported in part by the Biodiversity Research Project of the Transdisciplinary Research Integration Center, Research Organization of Information and Systems. This study is contribution no. 2515 from the NIG.

References

Abe K, Noguchi H, Tagawa K, Yuzuriha M, Toyoda A, Kojima T, Ezawa K, Saitou N, Hattori M, Sakaki Y, et al. 2004. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57Bl/6, as defined by BAC-end sequence-SNP analysis. Genome Res 14: 2439–2447.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. GBE 3: 246-258.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of phylogenetic breakpoints and discordance from the NIG. This work was also supported in part by the Biodiversity Research Project of the Transdisciplinary Research Integration Center, Research Organization of Information and Systems. This study is contribution no. 2515 from the NIG.

References

Abe K, Noguchi H, Tagawa K, Yuzuriha M, Toyoda A, Kojima T, Ezawa K, Saitou N, Hattori M, Sakaki Y, et al. 2004. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57Bl/6, as defined by BAC-end sequence-SNP analysis. Genome Res 14: 2439–2447.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. GBE 3: 246-258.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of discordance among gene trees. Mol Biol Evol 24: 412-426.

Beck JA, Lloyd S, Halpern M, Lemoine-Pierce M, Eppig JT, Festing MF, Fisher EM. 2000. Genealogies of mouse inbred strains. Nat Genet 24: 23–25.

Darbishire AD. 1902. Note on crossing Japanese waltzing mice with European albino races. Biometrika 2: 101–104.

Frazer KA, Lander EA,咔哇 K, Baum DA, Smith SD, Rokas A. 2007. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57Bl/6, as defined by BAC-end sequence-SNP analysis. Genome Res 14: 2439–2447.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. GBE 3: 246-258.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of discordance among gene trees. Mol Biol Evol 24: 412-426.

Beck JA, Lloyd S, Halpern M, Lemoine-Pierce M, Eppig JT, Festing MF, Fisher EM. 2000. Genealogies of mouse inbred strains. Nat Genet 24: 23–25.

Darbishire AD. 1902. Note on crossing Japanese waltzing mice with European albino races. Biometrika 2: 101–104.

Frazer KA, Lander EA,咔哇 K, Baum DA, Smith SD, Rokas A. 2007. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57Bl/6, as defined by BAC-end sequence-SNP analysis. Genome Res 14: 2439–2447.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. GBE 3: 246-258.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of discordance among gene trees. Mol Biol Evol 24: 412-426.

Beck JA, Lloyd S, Halpern M, Lemoine-Pierce M, Eppig JT, Festing MF, Fisher EM. 2000. Genealogies of mouse inbred strains. Nat Genet 24: 23–25.

Darbishire AD. 1902. Note on crossing Japanese waltzing mice with European albino races. Biometrika 2: 101–104.

Frazer KA, Lander EA,咔哇 K, Baum DA, Smith SD, Rokas A. 2007. Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57Bl/6, as defined by BAC-end sequence-SNP analysis. Genome Res 14: 2439–2447.