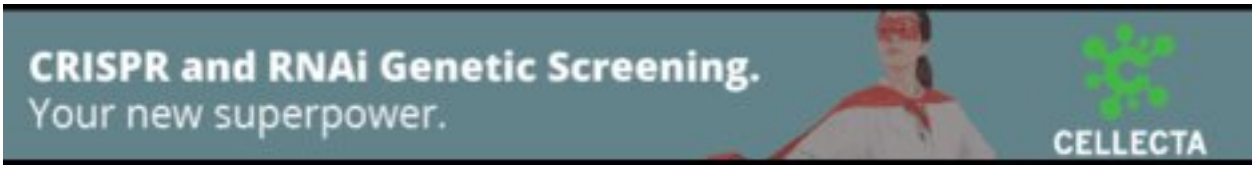
The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains

Toyoyuki Takada, Toshinobu Ebata, Hideki Noguchi, et al.

Genome Res. published online April 19, 2013
Access the most recent version at doi:10.1101/gr.156497.113



Revised

Resources

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

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

[1] Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[2] Genome Biology Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[3] Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[4] The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

[5] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

[6] BioResource Center, RIKEN Tsukuba Institute, Tsukuba, Ibaraki 305-0074, Japan

[7] Genome Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan; present address, Toyohashi University of Technology, Hibarigaoka, Tempaku, Toyohashi, Aichi 441-8580, Japan

[8] Transdisciplinary Research Integration Center, Research Organization of Information and Systems, Minato-ku, Tokyo 105-0001, Japan

*Corresponding author. E-mail: tshirois@lab.nig.ac.jp

Mammalian Genetics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

TEL: +81-55-981-6818**,** FAX: +81-55-981-6817

Keywords: mouse genome, MSM/Ms, JF1/Ms, inter-subspecific genome difference

Running title: *Mus musculus molossinus* genome

Abstract

Commonly used classical inbred mouse strains have mosaic genomes with sequences from different subspecific origins. Their genomes are derived predominantly from the Western European subspecies *Mus musculus domesticus*, with the remaining sequences derived mostly from the Japanese subspecies *M. m. molossinus*. However, it remains unknown how this intersubspecific genome introgression occurred during the establishment of classical inbred strains. In this study, we resequenced the genomes of *M. m. molossinus*-derived two inbred strains, MSM/Ms and JF1/Ms. MSM/Ms originated from Japanese wild mice, and ancestry of JF1/Ms was originally found in Europe and then transferred to Japan. We compared the characteristics of these sequences to those of the C57BL/6J reference sequence and the recent datasets from the resequencing of 17 inbred strains in the Mouse Genome Project (MGP), and 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 with extremely high similarity (>99.998%) to the JF1/Ms strain. In the early 20th century, Japanese waltzing mice with a morphological phenotype resembling that of JF1/Ms mice were often crossed with European fancy mice for early studies 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.

Introduction

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 *Mus musculus 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 traced the ancestries of classical inbred strains, focusing on the contribution of *M. m. molossinus*. This subspecies is known to be a hybrid of two subspecies, primarily *M. m. musculus* and, to a lesser degree, *M. m. castaneus* (Yonekawa *et al*. 1980; Sakai *et al*. 2005). 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 classical inbred strains that were generated in the Mouse Genome Project (MGP) (Keane *et al*. 2011). The results of this study revealed a vast number of SNPs and indels between the two strains, MSM and JF1, and classical inbred strains due to the large genetic distance between *M. m. molossinus* and *M. m. domesticus*. 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 would 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; Gates 1926; Schwarz *et al*. 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) (Moriwaki *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 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 MSM 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 Table S2. The candidate SNPs were detected by aligning the MSM and JF1 sequences to the reference B6 sequence (MGSC37; NCBI m37.1/mm9), as summarized in Figure S1. The genotypes of 17 MGP strains at those SNPs are summarized in Figure S2. We found 15,280,406 SNPs between the MSM and JF1 sequences and the B6 reference sequence, and 13,941,537 of these SNPs were found to be novel upon searching the SNP database (dbSNP) (Build 128). Furthermore, 2,302,645 of the SNPs were not found in the SNP calls of the MGP dataset. We also found 6,474,403 candidate SNPs between the MSM and JF1 genomes. With respect to candidate structural variants, we identified 439,922 and 617,551 short

insertions (1–6 bp), and 538,570 and 734,466 short deletions (1–6 bp) in the MSM and JF1 strains, respectively, relative to the B6 strain (Fig. S3).

The nucleotide variants calls were validated using two different procedures. First, we randomly selected SNPs found in MSM and JF1 for B6 and compared these SNPs with the genotype calls generated from the Sequenom MassARRAY iPLEX Gold Assay (Sequenom Inc., San Diego, CA, USA). Genotype calls were successfully made for 186 out of the selected 202 SNPs. The concordance rates were 98.9% (184/186) for MSM and 98.4% (183/186) for JF1. Second, we identified the SNPs between the B6 reference sequence and the previously reported BAC sequences (MSMg01-122K03 (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 1,322 SNPs and 130 indels, and the results are summarized in Table S3. The concordance rates were 98.9% (1,308/1,322) for the SNPs and 94.6% (123/130) 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 non-synonymous substitution or premature stop codon or cause the loss of a stop codon relative to the B6 reference sequence (Table 1, Table S4). Then, we performed functional annotations of these SNPs using DAVID Bioinformatics Resources 6.7 tools (Huang *et al*. 2009a; Huang *et al*. 2009b). 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 non-coding 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 (Table S5). In comparison to the B6 reference sequence, the MSM and JF1 sequences contain 38,182 and 38,124 non-synonymous SNPs in 11,489 and 11,313 genes, respectively.

We next explored the potential human disease-related phenotypes of the genes with non-synonymous 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 non-synonymous substitutions were radical (GMS >150), whereas 10.5% were moderately radical ($100 < GMS \leq 150$). We annotated the functions of 1,530 genes (for MSM) and 1,529 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 (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 dataset. Because the simplest and fastest way to define loci within a chromosome alignment is to consider fixed-length intervals (Ane, 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* (Wade *et al*. 2002; Frazer *et al*. 2007; Yang *et al*. 2007; Yang *et al*. 2011) (Fig. 1). In addition, a sequence comparison of B6 with MSM and JF1 revealed a bimodal distribution of blocks with varying sequence similarities (Fig. 1). The main population of blocks, with a peak at 99.00-99.05% similarity, represents the B6 sequence with intersubspecific genome divergence from *M. m. molossinus*, consistent with our previous report (Abe *et al*. 2004). The smaller population of blocks, with greater than 99.85% sequence similarity to MSM and JF1, likely represents *M. m. molossinus* genome introgression into the B6 genome.

To confirm this, we compared the percent sequence similarity of B6 and MSM or JF1 along each chromosome using sliding window analysis. We found that the B6 sequence has long consecutive regions that are highly similar (>99.85%) to MSM and JF1, with sharp boundaries between regions of high and low similarity (Fig. 2A).

However, we found that 0.75% of the regions in the JF1 genome show high similarity to B6 but not to MSM, as indicated on the distal portion of chromosome 8 and denoted as a red line in Figure 2A (Supplemental Fig. S4). These regions likely originated from reverse introgression from the ancestors of classical inbred strains into the JF1 genome. Regions with greater than 99.85% sequence similarity to MSM and JF1 were also found in the genomes of other classical inbred strains in the MGP dataset (Supplemental Fig. S5; Supplemental Data 1 and 2). The maximum rate of ratio for the whole genome was 7.03%, which was found in the LP/J strain, and the minimum rate was 3.32%, found in the A/J strain.

We also performed a genome-wide discordance survey using Bayesian concordance analysis (BCA) (Ane *et al*. 2007; Ane, 2011). In this analysis, we used the MSM and WSB sequences as references for *M. m. molossinus* and *M. m. domesticus*, respectively. We also used the sequence of SPRET/EiJ as a reference for *M. spretus* and used the rat sequence as an outgroup. The WSB and SPRET/EiJ sequences were obtained from the MGP dataset. The genomic regions for which a single B6/MSM topology was supported with a higher posterior probability according to the BCA mostly overlapped with those that are highly conserved (>99.85% similarity) between the B6 and MSM genomes, 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, 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 Figs. 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 that (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 (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 greater

than 99.998% sequence similarity to JF1 was significantly larger than the number of blocks with the same degree of similarity to MSM (Fig. 3A). In the most extreme case, we found a 717-kb unique and non-repetitive consecutive sequence in the region between the SNPs at position 91,144,048 and 91,861,518 on chromosome 14 that lacked any SNP or short indel between B6 and JF1. To determine whether these blocks are widely distributed across the B6 genome, we assigned the locations of the blocks with sequence similarity greater than 99.998% to MSM and JF1 on the mouse chromosomes (Fig. 3B). Consecutive blocks with extremely high similarity to JF1 were found widely distributed across most mouse chromosomes, except for chromosomes 15, 18 and the X chromosome. The preferentially high sequence similarity to JF1 was also observed in the genomes of other classical inbred strains in the MGP dataset (Supplemental Fig. S9), indicating that this is a general feature of the genome composition of classical inbred strains.

To examine whether JF1 is representative of *M. m. molossinus*, we conducted SNP-based genotyping of other *M. m. molossinus*-derived inbred strains as well as B6 and its related strain C57BL/10J at 102 randomly selected nucleotide sites that reside in MDRs but are polymorphic between MSM and JF1. The results clearly showed that the B6 and C57BL/10J strains contain the JF1-type SNP at those sites, and JF1-type SNPs were commonly observed in the other *M. m. molossinus*-derived inbred strains (Fig. 4).

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 approximately 15 million high-confidence SNPs (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 non-synonymous 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 (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.* 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 (Yang *et al.*, 2011). 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. 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 less than 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, leading to the variety of phenotypes observed today among classical inbred strains. We estimate that roughly 30 to 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 (Tsai *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 *et al*. 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 (Morse 1981; Keeler *et al*. 1931) and established as the classical inbred strains by the pioneers of mouse genetics, such as W. E. Castle and C. C. Little (Morse 1981) (Fig. 5).

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 vast amount nucleotide sequence variants scattered in the genomes of classical inbred strains are concentrated in genome 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 supplementary material (Supplemental data 1-3 and Table 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 (Bar Harbor, ME, USA). The genomic DNA samples used in this study are listed in 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-0.5.5 and samtool-0.1.7a for MSM reads and bwa ver0.5.7 and samtools ver0.1.7a 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 (max. 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 approximately 6.4 billion base pairs 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 Phred2 v0.020425.c (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 Phred scores >20 were subjected to repeat detection using RepeatMasker (version3.1.6: http://www.repeatmasker.org/). Non-repeat sequences larger than 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 >3 and <30 were used for the MSM strain, and sequence reads with >15 and <59 reads were used for the JF1 strain. For the MSM data, SNPs and short indels (in the 1-6 bp range) were detected using bwa-0.5.5 and samtool-0.1.7a software for the MSM/Ms reads and bwa-0.5.7 and samtools ver0.1.7a software for the JF1 reads. For MSM reads, capillary sequence data with Phred scores >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 Inc., San Diego, CA, USA). The information used for genotyping are listed in 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 Inc., San Diego, CA, USA). To estimate the false-positive and false-negative rates of nucleotide variants (SNPs and 1-6 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 variants 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 (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 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 and 2009b) 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 IDs (1,530 genes for MSM and 1,529 genes for JF1) was submitted to the DAVID web site. We eliminated SNPs, *in silico* predicted genes, pseudogenes and non-coding genes from the functional annotation.

SNPs detected by comparing the MSM and JF1 sequences to the MGP dataset

Sequence reads for the MSM and JF1 strains were compared to the MGP dataset. 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, WSB/EiJ and Spretus/EiJ, were obtained in BAM file format from the FTP site of the MGP at the Sanger Institute (http://www.sanger.ac.uk/resources/mouse/genomes/). The repeat-masked reference B6 sequence was used for sequence comparisons.

Calculation of percent 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) (the results of the cnD analysis are shown in 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 percent similarity of each block of the B6 reference sequence to the sequences of MSM and JF1, as well as the entire MGP dataset individually. The distribution of the number of blocks with a given sequence similarity is shown in Figure 1, and Supplemental data 1 and 2. The profiles illustrating percent 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 N's. Collinear blocks were partitioned into 124,174 loci using a minimum description length algorithm with a default maximum cost (shown in 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, Inc., San Diego, CA, USA). 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.

Acknowledgements

We thank J. Forejt and Y. Matsushima for providing genomic DNA samples of wild mice-derived strains. We are grateful to M. White, C. Ané, B. Payseur, H. Fujimiya, N. Osada, N. Niinaya and Y. Nakamura for valuable suggestions regarding statistical analysis and bioinformatics, to Y. Yamazaki for the management of the NIG mouse genome database server, to the Information Technology Unit for use of super computer system of the NIG. We also thank A. Mita for careful long-term maintenance of the MSM and JF1 strains at the NIG, and K. Yamabe, N. Harima, M. Naganuma, K. Masuyama and A. Matsusue 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. xxxxx from the NIG.

Figure Legends

Figure 1. Distribution of the number of 100-kb blocks in the reference B6 genome with various sequence similarities (%) to the corresponding blocks in the MSM, JF1 and WSB genomes.

Figure 2. Sequence similarity between the B6 and MSM and JF1 genomes. A) Sliding window analysis of the discordance across chromosome 8 between the B6 and MSM or JF1 sequences. The reference B6 sequence was used for comparison with 500-kb windows and 100-kb sliding intervals. The horizontal blue line indicates a 99.85% sequence similarity level. Fine-scale phylogenetic discordance of chromosome 8 is shown below (PP; posterior probability). B) Phylogenetic tree of the MDR sequences of wild-derived inbred strains. A neighbor-joining tree was generated for 67 concatenated MDR regions using MEGA4 software (Tamura *et al*. 2007). The 67 MDR regions show a single topology for B6/MSM, supported by a high posterior probability by BCA. The numbers adjacent to the branches indicate bootstrap values >50 (1,000 replicates). Subspecies names and the locations at which ancestors of the strains were collected are shown in parentheses. For more details, see Supplemental Table S1.

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 greater than 99.998% to the MSM and JF1 chromosomes. Horizontal black boxes depict the regions with greater than 99.85% sequence similarity to the MSM strain. Gray boxes indicate gaps in the B6 reference sequence.

Figure 4. SNP-based genotyping of *M. m. molossinus*-derived inbred strains, the B6 strain and its related strain C57BL/10J at nucleotide sites that reside in MDRs but are polymorphic between MSM and JF1. A total of 102 randomly selected nucleotide sites from MDRs were genotyped using the MassARRAY system. M = MSM genotype (green); J = JF1 genotype (coral); Blank = not determined.

Figure 5. 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-shyuppan, Tokyo, Japan). In the middle to late 19th century, British traders likely introduced Japanese waltzing mice carrying the “piebald” (*Ednrb*^{s}) 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 left 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 (Morse 1981; Keeler *et al*. 1931), where they were established as classical inbred strains.

Supplemental Figure S1. Venn diagrams showing the total numbers of SNPs and short indels detected by comparing the sequences of the MSM and JF1 strains to the reference B6 sequence (MGSC37).

Supplemental Figure S2. Venn diagrams showing the total numbers of novel SNPs detected as a result of comparison to the MGP dataset and dbSNP128, excluding data from the SPRET/EiJ strain of the MGP.

Supplemental Figure S3. The size distribution of short indels (1-6 bp) detected by comparing the MSM and JF1 sequences to the reference B6 sequence. The number of indels was plotted using BWA prediction for indels of varying size (≤6 bp).

Supplemental Figure S4. We calculated the ratio of the number of 100-kb blocks with high (>99.85%) similarity in the MSM and JF1 genomes to the total number of 100-kb blocks in the reference B6 genome (MGSC37). We also calculated the ratio of cumulative non-repeat nucleotide sequences in the 100-kb blocks with high (>99.85%) similarity to the MSM and JF1 genomes to the total non-repeat sequence of the reference B6 genome. The rate of reverse genome introgression from *M. m. domesticus*

into the JF1 genome was calculated to be 0.75%. This value was estimated by subtracting the ratio of 100-kb blocks with high (>99.85%) sequence similarity to MSM from the ratio of that to JF1.

Supplemental Figure S5. The distribution of 100-kb blocks with high (>99.85%) similarity to the MSM and JF1 genomes in the MGP dataset (Keane *et al.* 2012) of 17 inbred strains.

Supplemental Figure S6. Fine-scale phylogenetic discordance determined using whole-genome sequence information from MSM, WSB SPRET mice and the rat. The posterior probability of each topology is mapped onto chromosomes 1 to 19 and X to characterize the fine-scale patterns of discordance among the 124,174 loci. The colors correspond to the 3 topologies (shown in the box).

Supplemental Figure S7. Genomic partitioning of the phylogenetic history. Bayesian concordance factors were estimated from 124,174 individual locus trees. For 94.6% of the loci, SPRET/EiJ (*M. spretus*) mice and rats are placed as outgroups of the subspecies *M. musculus*. Within *M. musculus*, 10.1% of loci were supported with higher posterior probability of a single B6/MSM topology, and 85.2% of loci supported a single WSB/B6 topology (higher posterior probability).

Supplemental Figure S8. Distribution plots of the nucleotide sequence similarity

between B6 and MSM or JF1 with respect to MDRs defined by BCA (Fig. S7). Each plot shows the sequence similarity between the B6 and MSM strains (horizontal axis) and between the B6 and JF1 strains (vertical axis).

Supplemental Figure S9. Distribution of sequence blocks with high (>99.85% per 100 kb) similarity to the MSM and JF1 sequences for each of the 17 MGP strains (Keane *et al.* 2012). The data from wild-derived inbred strains were eliminated. Reference data for the comparisons between the B6 (MGSC37) and MSM or JF1 strains, which were derived from Figure 3B, are denoted above each chromosome. Gray boxes depict gaps in the B6 reference sequence.

Supplemental Figure S10. The ratio of the number of *M. m. molossinus*-derived SNPs to the total number of SNPs in pairs of classical inbred strains. We calculated the number of SNPs between the classical inbred strains that reside in regions with high sequence similarity to the MSM genome.

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/6J, as defined by BAC-end sequence-SNP analysis. *Genome Res* **14**: 2439–2447.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Mol Biol Evol* **24:** 412–26.

Ané, C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *GBE* **3:** 246–58.

Beck JA, Lloyd S, Hafezparast M, Lennon-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, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM. Nilsen GB, *et al*. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050-1053.

Gates WH. 1925. The Japanese waltzing mouse, its origin and genetics. *Proc Natl Acad Sci USA* **11:** 651.

Gates WH. 1926. The Japanese waltzing mouse: its origin, heredity and relation to the genetic characters of other varieties of mice. In Contributions to a Knowledge of Inheritance in Mammals, (eds Castle WE, Feldman HW, Gates WH), pp 83-138,

Carnegie Inst, Washington DC.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185:** 862–864.

Hirasawa R, Chiba H, Kaneda M, Tajima S, Li E, Jaenisch R, Sasaki H. 2008. Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev* **22:** 1607–1616.

Hosoda K, Hammer R E, Richardson J A, Baynash A G, Cheung J C, Glald A, Yanagisawa M. 1994. Targeted and natural (piebald-lethal) mutations of endothelin-B receptor gene produce megacolon associated with spotted coat color in mice. *Cell* **79:** 1267-1276.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* **4:** 44-57.

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1-13.

Keane TH, Goodstadt L, Danecek P, Payseur B, White MA, Yalcin B, Heger A, Agam A, Slater G, Goodson M, *et al*. 2011. Sequence variants among 17 mouse genomes: Effect on phenotypes and gene regulation. *Nature* **477:** 289-294.

Keeler CE. 1931. *The laboratory mouse: its origin, heredity, and culture.* Harvard Univ Press, Cambridge.

Koide T, Moriwaki K, Uchida K, Mita A, Sagai T, Yonekawa H, Katoh H, Miyashita N,

Tsuchiya K, Nielsen TJ, *et al*. 1998. A new inbred strain JF1 established from Japanese fancy mouse carrying the classic piebald allele. *Mamm Genome* **9:** 15-19.

Koide, T., Ikeda, K., Ogasawara, M., Shiroishi, T., Moriwaki, K., & Takahashi, A. (2011). A new twist on behavioral genetics by incorporating wild-derived mouse strains. *Exp Anim* **60:** 347–54.

Moriwaki K, Miyashita N, Mita A, Gotoh H, Tsuchiya K, Kato H, Mekada K, Noro C, Oota S, Yoshiki A, *et al*. 2009. Unique inbred strain MSM/Ms established from the Japanese wild mouse. *Exp Anim* **58:** 123-134.

Morse HC, III. 1981. The laboratory mouse-A historical perspective. In The mouse in biomedical research (eds Foster HL, Small JD, Fox JG), pp 1-16, Academic Press, Inc, San Diego.

Sakai T, Kikkawa Y, Miura I, Inoue T, Moriwaki K, Shiroishi T, Satta Y, Takahata N, Yonekawa H. 2005. Origins of mouse inbred strains deduced from whole-genome scanning by polymorphic microsatellite loci. *Mamm Genome* **16:** 11-19.

Schwarz E. 1942. Origin of the Japanese waltzing mouse. *Science* **95:** 2454.

Simpson JT, McIntyre RE, Adams DJ, Durbin R. 2010. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* **26**: 565–567.

Takada T, Mita A, Maeno A, Sakai T, Shitara H, Kikkawa Y, Moriwaki K, Yonekawa H, Shiroishi T. 2008. Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits. *Genome Res* **18:** 500–508.

Takada T, Shiroishi T. 2012. Complex quantitative traits cracked by the mouse inter-subspecific consomic strains. *Exp Anim* **61:** 375–88.

Takahashi A, Nishi A, Ishii A, Shiroishi, T, Koide, T. 2008. Systematic analysis of emotionality in consomic mouse strains established from C57BL/6J and wild-derived MSM/Ms. *Genes Brain Behav* **7:** 849–58.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24:** 1596–1599.

Tanaka S, Suzuki T, Sakaizumi M, Harada Y, Matsushima Y, Miyashita N, Fukumori Y, Inai S, Moriwaki K, Yonekawa H. 1991. Gene responsible for deficient activity of the beta subunit of C8, the eighth component of complement, is located on mouse chromosome 4. *Immunogenetics* **33:** 18-23.

Tokuda M. 1935. An eighteenth century Japanese guide-book on mouse-breeding. *J Hered* **26:** 481-484.

Tsai C, Lin S, Ito M, Takagi N, Takada S, Ferguson-Smith A. 2002. Genomic imprinting contributes to thyroid hormone metabolism in the mouse embryo. *Curr Biol* **12:** 1221–1226.

Wade CM, Kulbokas EJ, III, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data *Nucleic Acids Res* **38:** e164.

White MA, Ané C, Dewey CN, Larget BR, Payseur BA. 2009. Fine-scale phylogenetic

discordance across the house mouse genome. *PLoS genet* **5:** e1000729.

Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DA, Keane TM. 2012. Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biol* **13:** R72.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. *Nat Genet* **39:** 1100–1107.

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH, Nachman NW, Pialek J, *et al.* 2011. Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* **43:** 648-655.

Yerkes RM. 1907. The dancing mouse: A study in animal behavior. In The Animal Behavior Series. Vol.1, The Macmillan Comp, New York.

Yonekawa H, Moriwaki K, Gotoh O, Watanabe J, Hayashi J-I, Miyashita N, Petras ML, Tagashira Y. 1980. Relationship between laboratory mice and the subspecies *Mus musculus domesticus* based on restriction endonuclease cleavage patterns of mitochondrial DNA. *Japan J Genet* **55:** 289-296.

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

	C57BL/6J	NonSynonymous (NS)		NS Total		Synonymous (S)		S Total		Stop_gain		Stop_loss	
Chr	Total Gene #	MSM	JF1	MSM	JF1	MSM	JF1	MSM	JF1	MSM	JF1	MSM	JF1
1	1298	632	625	2488	2486	867	879	5164	5121	9	16	2	3
2	1984	950	923	2849	2829	1296	1312	6416	6446	11	10	0	1
3	1097	534	537	1794	1831	756	775	3874	3919	11	12	2	3
4	1350	661	659	2207	2244	902	920	4366	4399	27	19	1	2
5	1313	704	703	2430	2421	997	998	5870	5852	12	12	2	2
6	1235	637	625	2198	2186	860	875	4160	4129	14	16	2	2
7	2111	1183	1215	4202	4399	1545	1593	7384	7626	29	30	6	3
8	1103	440	331	1369	1076	622	467	3355	2525	6	8	5	4
9	1310	695	693	2097	2118	935	960	4804	4833	15	12	3	3
10	1083	536	540	1708	1661	789	800	4342	4367	7	4	2	4
11	1745	867	836	2630	2496	1257	1252	6737	6464	15	9	2	3
12	791	291	300	976	997	443	450	2448	2467	0	2	1	0
13	886	403	393	1443	1473	573	578	2993	3040	5	6	0	1
14	835	389	393	1137	1186	501	512	2541	2560	4	12	2	1
15	866	450	448	1514	1535	635	640	3701	3680	6	7	1	1
16	726	388	376	1465	1418	510	521	2659	2698	6	4	3	3
17	1129	565	541	1982	1977	735	726	3689	3642	12	13	1	3
18	570	299	297	1227	1227	426	430	2850	2844	2	3	0	0
19	758	417	417	1258	1288	589	598	3066	3079	3	4	1	1
X	1027	448	461	1208	1276	569	604	1821	1868	11	18	3	3
Y	13	-	-	-	-	-	-	-	-	-	-	-	-
Un	25	-	-	-	-	-	-	-	-	-	-	-	-
Total	23255	11489	11313	38182	38124	15807	15890	82240	81559	205	217	39	43

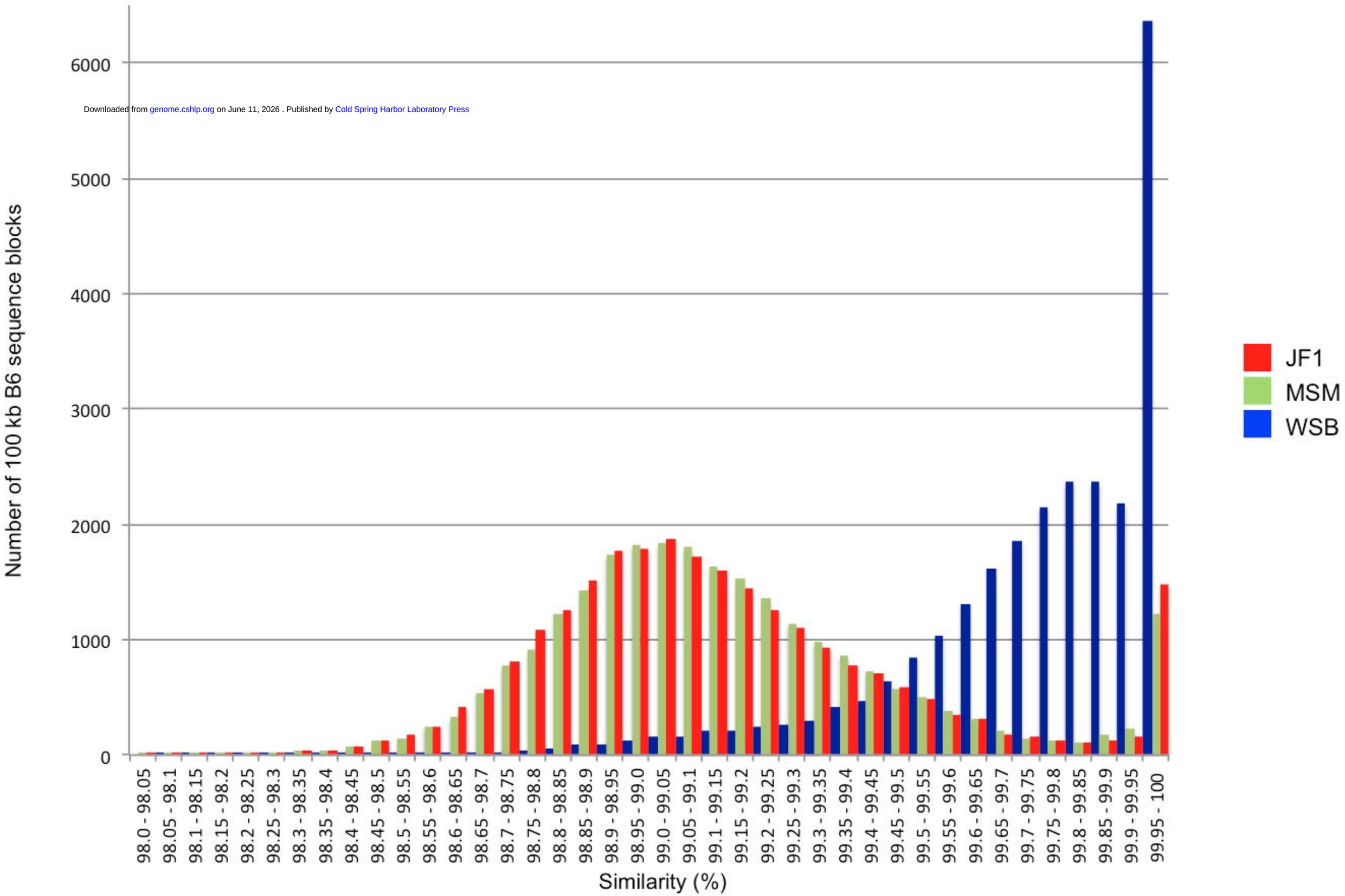
Table 2. Human disease-associated non-synonymous single nucleotide variants in the MSM and JF1 genomes (DAVID analysis).

Chr	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/glucocorticoid regulated kinase 3	NM_001037759	fuzzy
3	Frem2	Fras1 related extracellular matrix protein 2	NM_172862	myelencephalic blebs
4	Chd7	chromodomain helicase DNA binding protein 7	NM_001081417	cyclone,dizzy,eddy,leda, orbitor,tornado,whirligig
4	Cdkn2a	cyclin-dependent kinase inhibitor 2A	NM_001040654	plasmacytoma resistance 1
4	Mtf1	metal response element binding transcription factor 1	NM_008636	thymic lymphoma susceptibility
7	Coro1a	coronin, actin binding protein 1A	NM_009898	lupus in MRL and B6 F2 cross, QTL 3
7	Odz4	similar to DOC4; odd Oz/ten-m homolog 4 (Drosophila)	NM_011858	lethal, Chr 7, Rinchik 3
7	Lgi4 **	leucine-rich repeat LGI family, member 4	NM_144556	claw paw
8	Acd *	adrenocortical dysplasia	NM_001012638	adrenocortical dysplasia
8	Smpd3 *	sphingomyelin phosphodiesterase 3, neutral	NM_021491	fragilitas ossium
9	Mpzl3	myelin protein zero-like 3	NM_001093749, NM_176993	rough coat
10	Mtap7	microtubule-associated protein 7	NM_008635	male sterility and histoincompatibility
10	Pdss2	prenyl (solanesyl) diphosphate synthase, subunit 2	NM_001168289	kidney disease
10	Sobp	sine oculis-binding protein homolog (Drosophila)	NM_175407	Jackson circler
11	Vps54	vacuolar protein sorting 54 (yeast)	NM_139061	wobbler
12	Hectd1	HECT domain containing 1	NM_144788	open mind
14	Dock5	dedicator of cytokinesis 5	NM_177780	rupture of lens cataract
14	Fndc3a	fibronectin type III domain containing 3A	NM_207636	symplastic spermatids
15	Aqp2	aquaporin 2	NM_009699	congenital progressive hydronephrosis
15	Tenc1	tensin like C1 domain-containing phosphatase	NM_153533	nephrosis
16	Liph	lipase, member H	NM_153404	lipid defect 1
17	Lmf1	lipase maturation factor 1	NM_029624	combined lipase deficiency
18	Afg3l2 **	similar to AFG3(ATPase family gene 3)-like 2 (yeast); AFG3(ATPase family gene 3)-like 2 (yeast)	NM_027130	paralyse
19	Scyl1	SCY1-like 1 (S. cerevisiae)	NM_023912	muscle deficient
19	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 MSM/Ms data.

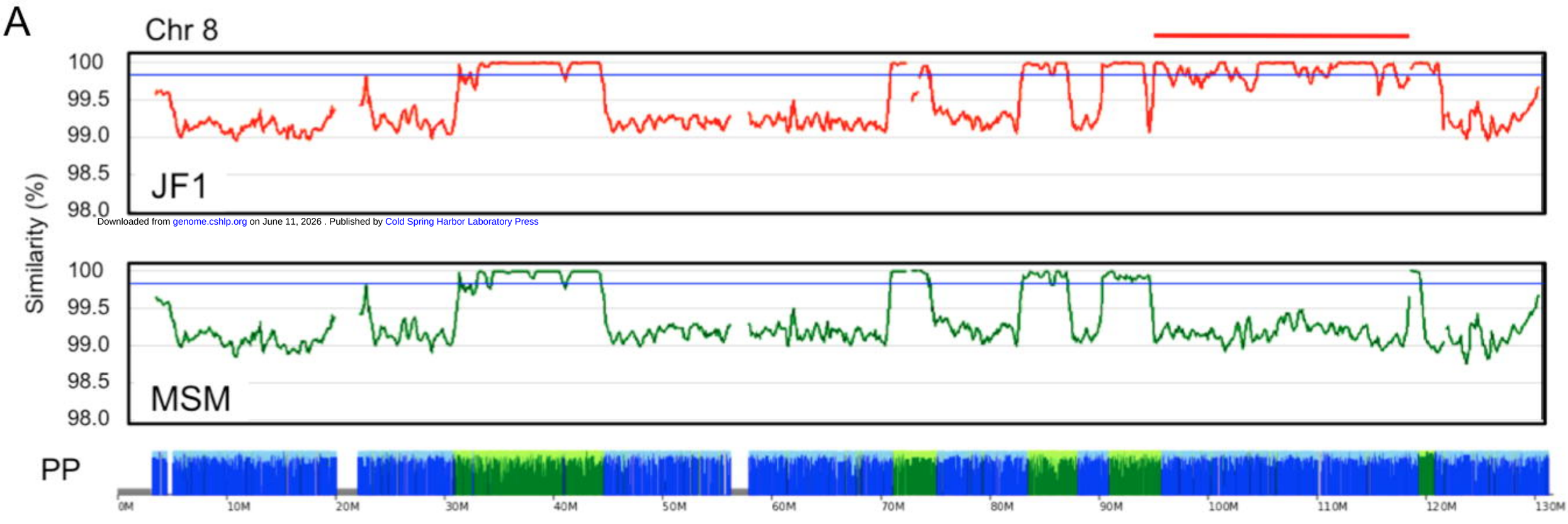
** Observed only in JF1/Ms data.

Takada_Fig1

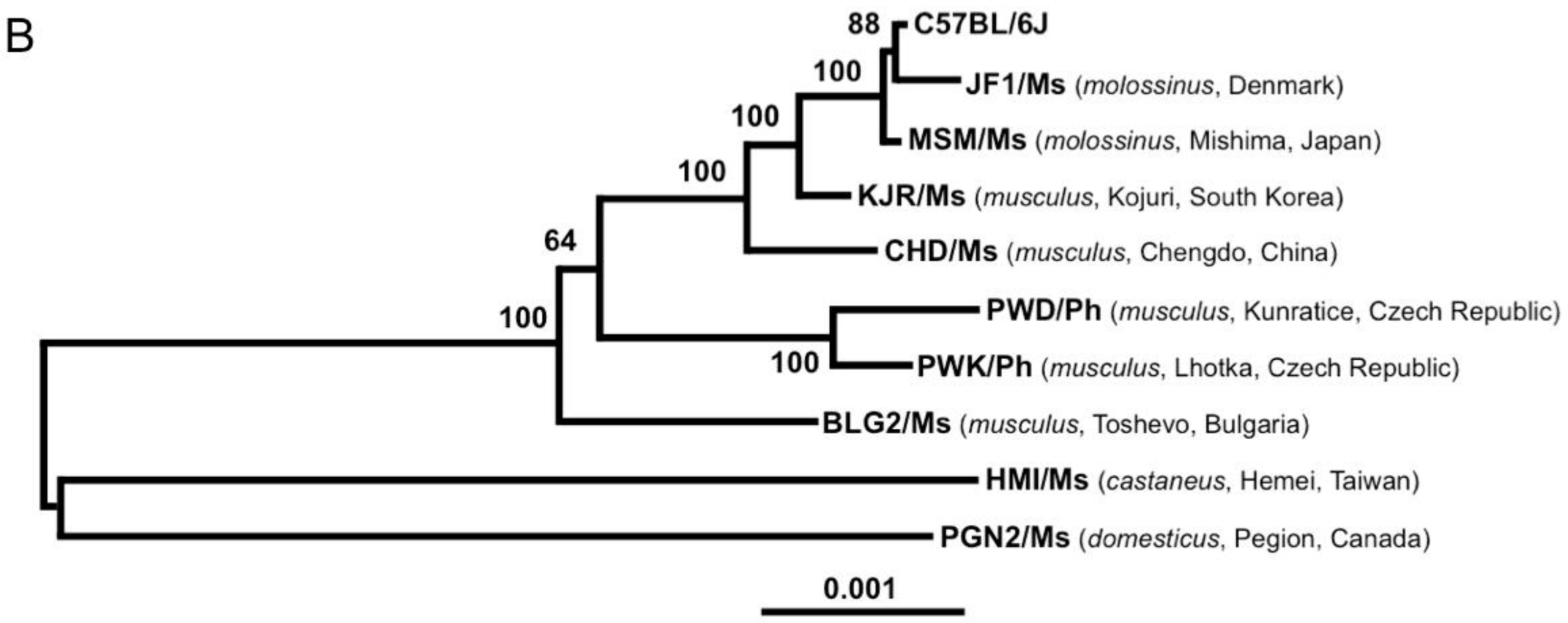


Takada_Fig2

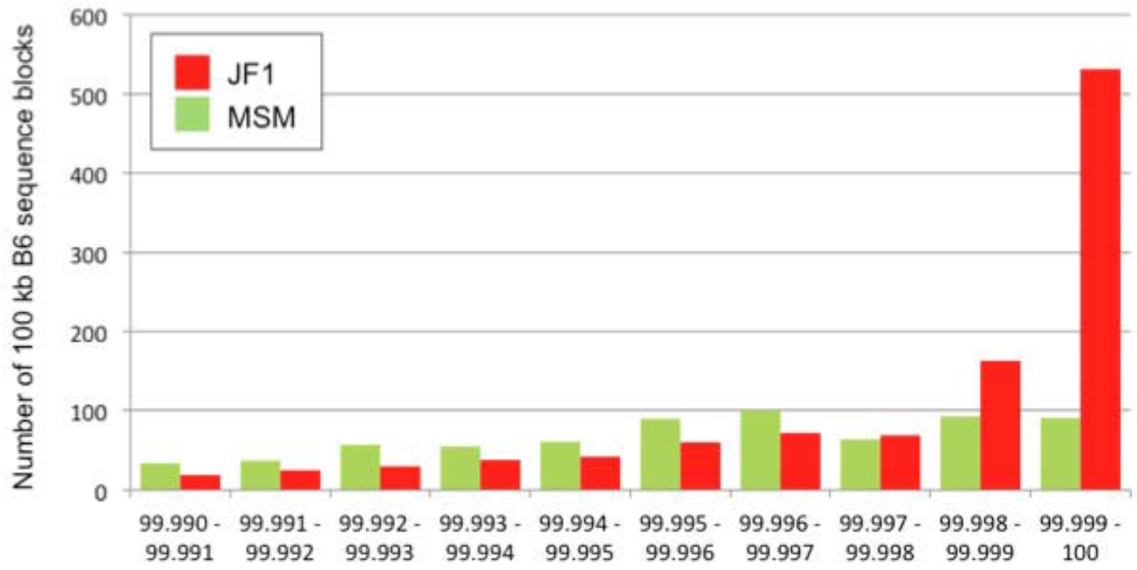
A



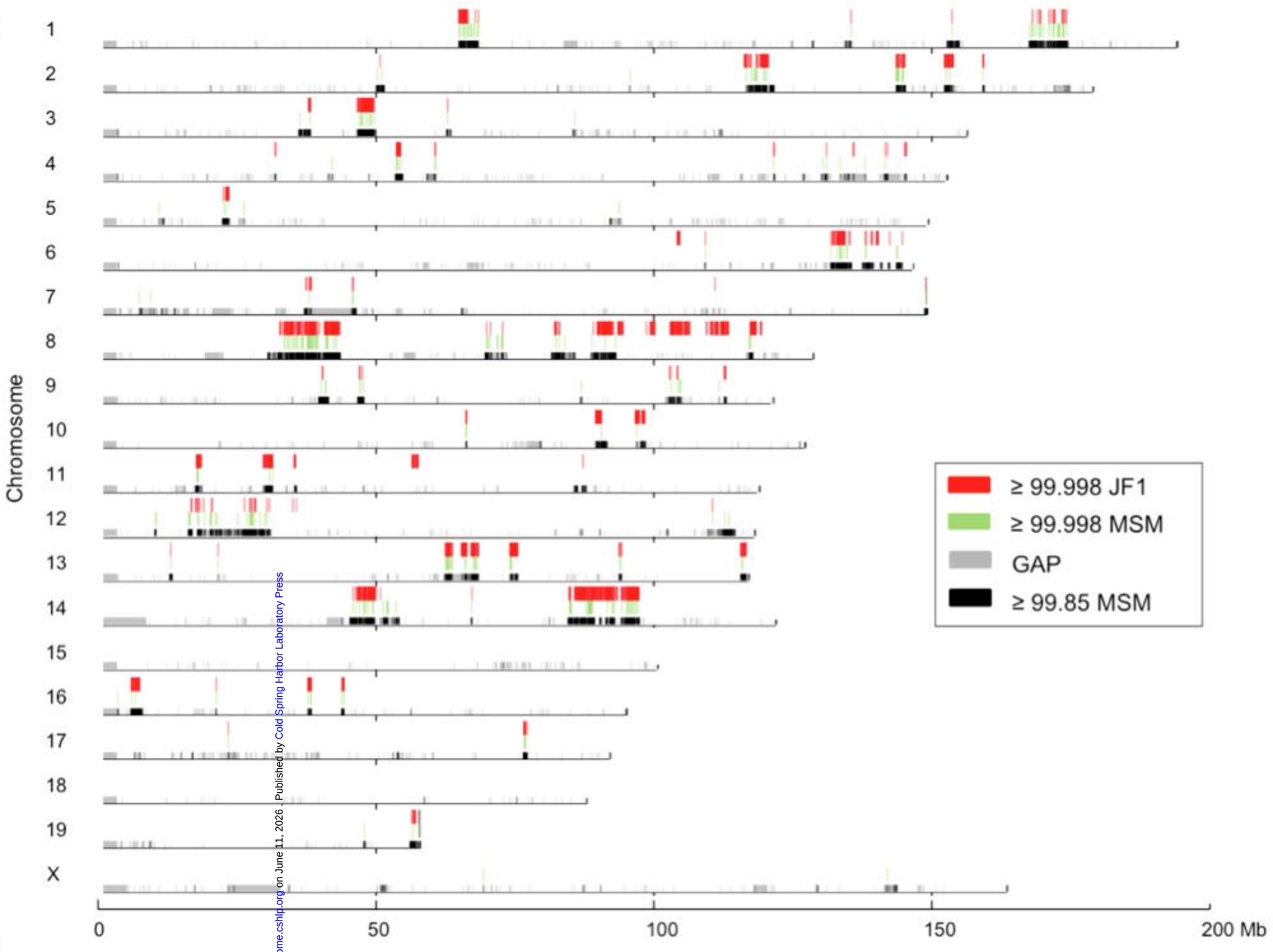
B



A

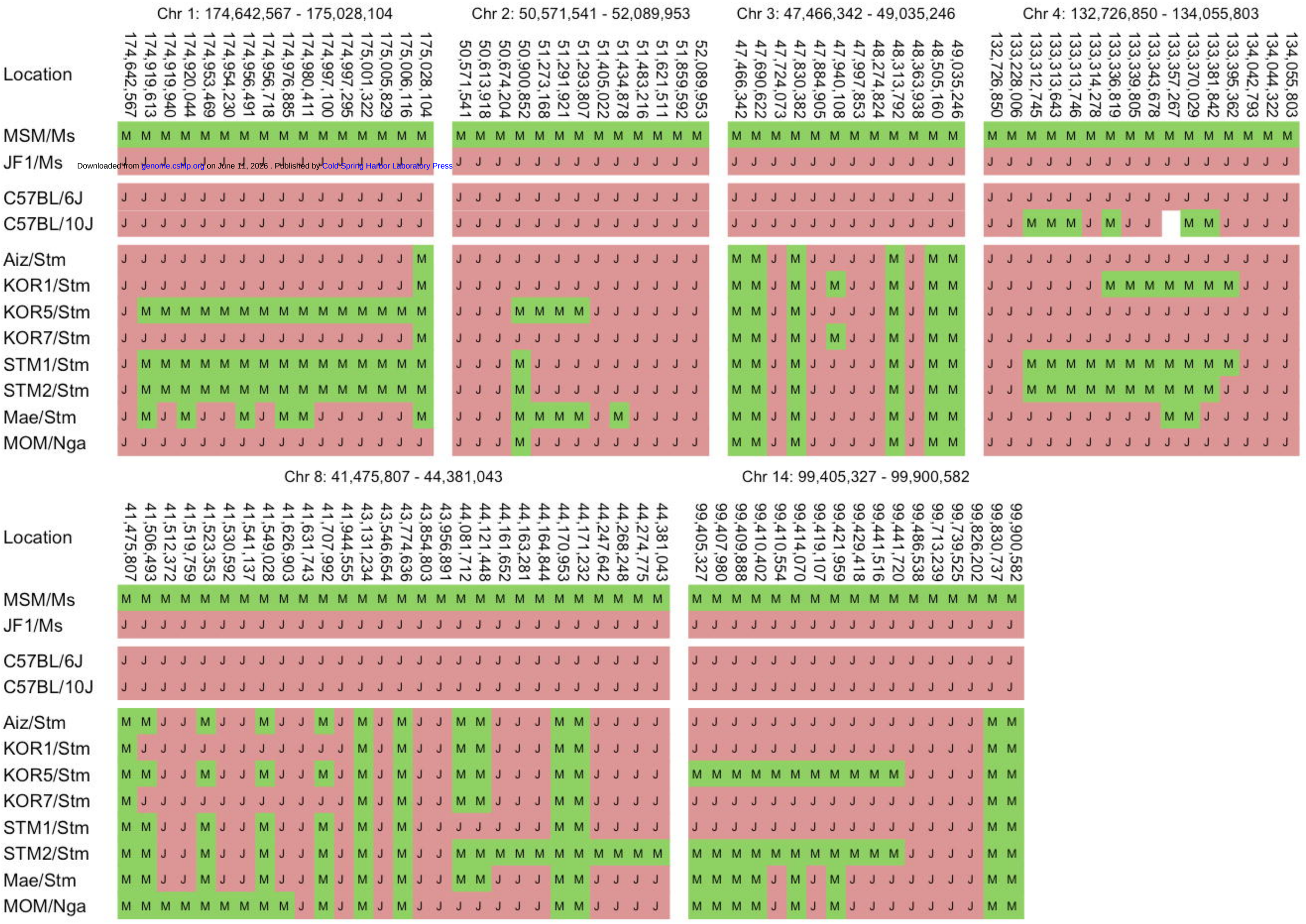


B

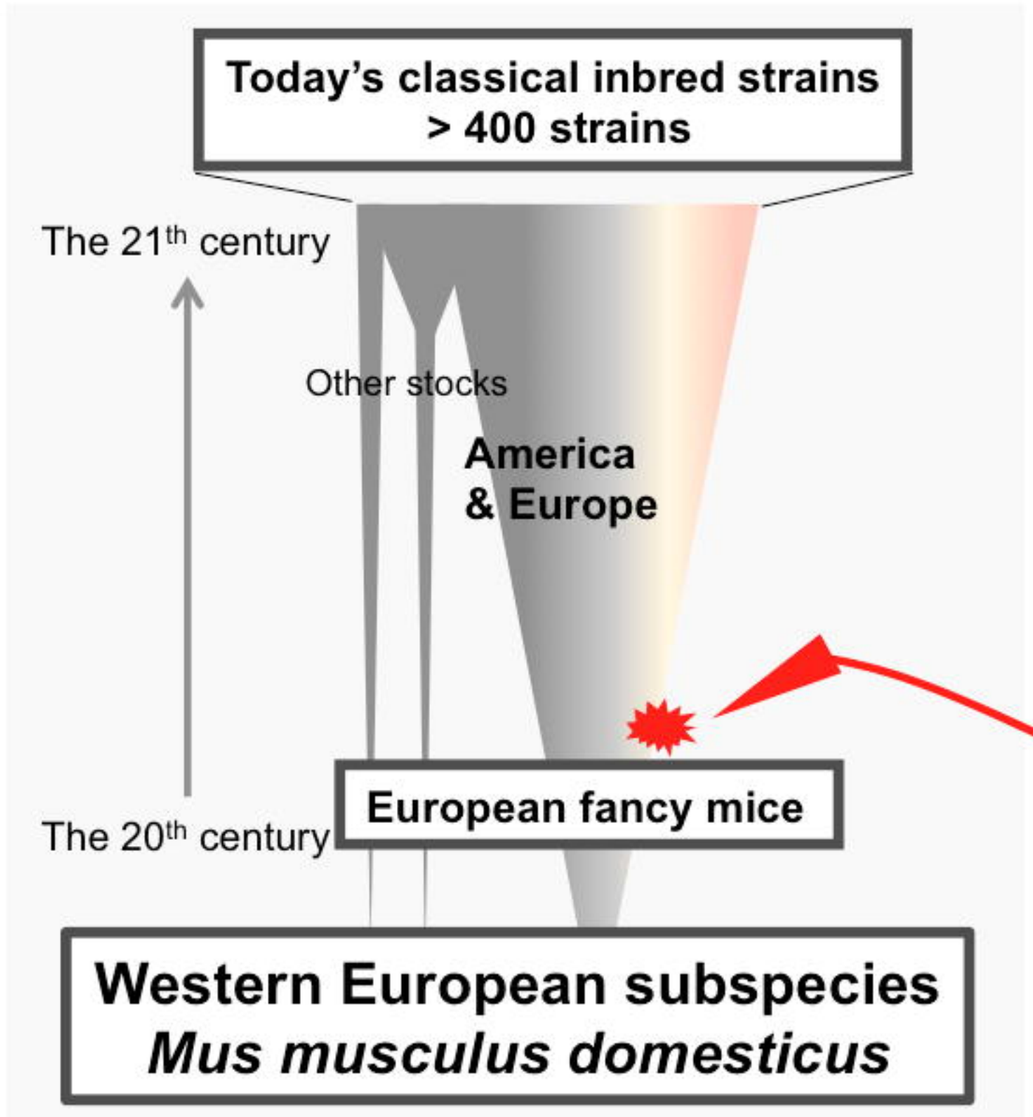


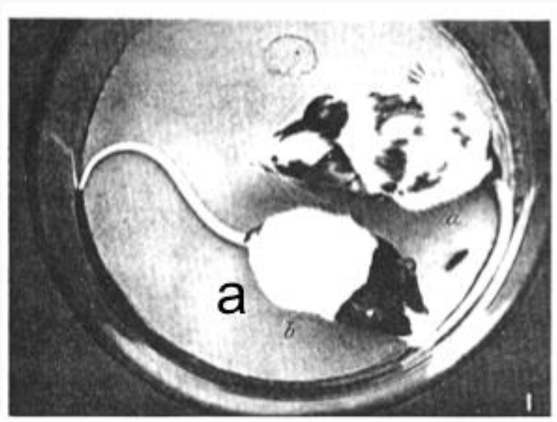
Takada_Fig3

Takada_Fig4

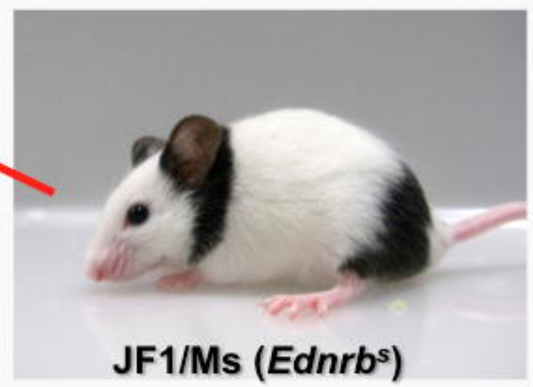


Takada_Fig5





"The Japanese waltzing mouse"
(Gates W. H., 1926)



JF1/Ms (*Ednrb*^{s})

"Ancestry of JF1/Ms"
(late 19th century)

"Chingan-sodategusa"
(Zeniya-chyobei, 1787)