Mamm Genome
DOI 10.1007/s00335-017-9704-9

# Whole exome sequencing of wild-derived inbred strains of mice improves power to link phenotype and genotype

**Peter L. Chang[1] · Emily Kopania[1,2] · Sara Keeble[1,2] · Brice A. J. Sarver[2] · Erica Larson[2,3] · Annie Orth[4] · Khalid Belkhir[4] · Pierre Boursot[4] · François Bonhomme[4] · Jeffrey M. Good[2] · Matthew D. Dean[1]**

Received: 23 March 2017 / Accepted: 23 June 2017
© Springer Science+Business Media, LLC 2017

**Abstract** The house mouse is a powerful model to dissect the genetic basis of phenotypic variation, and serves as a model to study human diseases. Despite a wealth of discoveries, most classical laboratory strains have captured only a small fraction of genetic variation known to segregate in their wild progenitors, and existing strains are often related to each other in complex ways. Inbred strains of mice independently derived from natural populations have the potential to increase power in genetic studies with the addition of novel genetic variation. Here, we perform exome-enrichment and high-throughput sequencing (~8× coverage) of 26 wild-derived strains known in the mouse research community as the "Montpellier strains." We identified 1.46 million SNPs in our dataset, approximately 19% of which have not been detected from other inbred strains. This novel genetic variation is expected to contribute to phenotypic variation, as they include 18,496 nonsynonymous variants and 262 early stop codons. Simulations demonstrate that the higher density of genetic variation in the Montpellier strains provides increased power for quantitative genetic studies. Inasmuch as the power to connect genotype to phenotype depends on genetic variation, it is important to incorporate these additional genetic strains into future research programs.

Illumina sequencing data are available in NCBI under the BioProject PRJNA326865.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00335-017-9704-9) contains supplementary material, which is available to authorized users.

✉ Matthew D. Dean
matthew.dean@usc.edu

[1] Molecular and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

[2] Division of Biological Sciences, University of Montana, Missoula, MT, USA

[3] Department of Biological Sciences, University of Denver, Denver, CO 80210, USA

[4] Institut des Sciences de l'Evolution, CNRS UMR554, Université de Montpellier, Montpellier, France

## Introduction

For more than 100 years, the house mouse (*Mus musculus*) has been a useful model for genetic research (Paigen 2003a, b). Several important features contribute to their utility, including a high-quality reference genome with more than a decade's worth of improved assembly and annotation (Church et al. 2009; Waterston et al. 2002), multiple complete genomes from distinct genetic strains (Keane et al. 2011; Nikolskiy et al. 2015; Srivastava et al. 2017; Wang et al. 2016; Waterston et al. 2002; Wong et al. 2012), and wild individuals (Harr et al. 2016), and dense genotyping of commonly used laboratory strains (Laurie et al. 2007; Lindblad-Toh et al. 2000; Petkov et al. 2004; Wade et al. 2002; Yang et al. 2007, 2009, 2011). Thousands of phenotypes have been gathered from hundreds of inbred mouse strains (Grubb et al. 2004; Wang et al. 2016; White et al. 2013), many of which are commercially available through institutions like The Jackson Laboratory.

Although the impact of the house mouse on biological research cannot be overstated, many existing inbred strains of mice are related to each other in complex ways, and capture only a small amount of genetic variation known to segregate in their wild progenitors (Beck et al. 2000; Keane et al. 2011; Salcedo et al. 2007; Wade et al. 2002; Yang et al. 2011). Inasmuch as the power to connect genotype to phenotype depends on genetic variation, it

Published online: 17 August 2017

is important to incorporate additional genetic strains into future research programs. Inbred mouse strains are generally classified as "wild-derived inbred strains" or "classical inbred strains." Wild-derived inbred strains represent independent derivations from particular geographic areas. In contrast, classical inbred strains derive from a small pool of founders whose origins trace to Japanese and European mouse fanciers (Beck et al. 2000; Frazer et al. 2007; Moriwaki 1994; Morse 1978, 2007; Silver 1995; Wade and Daly 2005; Wade et al. 2002), who likely crossed strains from different geographic or species origins prior to inbreeding (Didion and Pardo-Manuel de Villena 2013; Keane et al. 2011; Yang et al. 2007, 2011). As a result, 97% of the genome of classical inbred strains can be traced to ten different haplotypes (Yang et al. 2011).

Several groups have established wild-derived inbred strains in an attempt to increase the amount of genetic variation available to researchers. Bonhomme and colleagues at the University of Montpellier have set up one of the largest collections. The Montpellier collection includes 29 strains representing subspecies of *M. musculus*, 4 of *Mus spretus*, and one each of *Mus spicilegus, Mus macedonicus*, and *Mus caroli* (Supplementary Material 6), all of which are available for nominal fees through the Montpellier Stock Center (http://www.isem.univ-montp2.fr/recherche/les-plate-formes/conservatoire-genetique-de-souris-sauvages/). Thirty-two strains have undergone at least 20 generations of brother–sister mating, allowing for the same level of pseudo-replication that can be achieved with classical inbred strains. Even though the Montpellier strains could greatly increase the amount of known genetic variation, they have only been the topic of approximately 60 publications (Supplementary Material 7), which pales in comparison to the thousands of publications on just a few classical inbred strains such as the reference genome C57BL/6J.

To characterize genomic variation of the Montpellier strains, we generated and analyzed exome sequences from 26 strains derived from natural populations of *M. musculus* and *M. spretus*. Our goal was to increase the total amount of genetic variation known among inbred mouse strains. With very conservative methods, we identified 1.46 million SNPs, nearly 19% of which were not known from existing inbred strains. More than 77,000 were nonsynonymous or nonsense mutations that may introduce novel phenotypic variation, and we identified a few hundred genes that carry early termination codons and may provide alternatives to traditional knockouts. Using simulations, we show that inclusion of this new genetic variation would significantly improve the power of genetic mapping experiments. Our study demonstrates that the Montpellier strains represent a powerful yet under-utilized resource in genetic research.

## Materials and methods

### Mouse strains employed

All husbandry and experimental methods, as well as all personnel involved, were approved by the University of Southern California's Institute for Animal Care and Use Committee, protocol #11394. We chose 26 wild-derived inbred strains (WDIS) from the Montpellier genetic repository to perform high-throughput sequencing of enriched exomes. Fourteen of these strains were considered *Mus musculus domesticus*: 22MO (originally isolated from Monastir, Tunisia), BIK (Kefar Galim, Israel), BZO (Oran, Algeria), DCA (Akrotiri, Cyprus), DCP (Paphos, Cyprus), DDO (Odis, Denmark), DEB (Barcelona, Spain), DGA (Adjaria, Georgia), DIK (Keshet, Israel), DJO (Orcetto, Italy), DMZ (Azemmour, Morocco), DOT (Tahiti, French Polynesia), WLA (Toulouse, France), and WMP (Monastir, Tunisia). Four were considered *M. m. musculus*: MAM (Megri, Armenia), MBS (Sokolovo, Bulgaria), MGA (Alazani, Georgia), and MPB (Bialowieza, Poland). One was considered *M. m. castaneus*: CIM (Masinagudi, India). Four additional strains, BID (Birdjand, Iran), KAK (Khakhk, Iran), MPR (Rawalpindi, Pakistan), and TEH (Tehran, Iran), originated from regions that probably harbor multiple subspecies and were not assigned to any one subspecies (Hardouin et al. 2015). Lastly, we included three strains of *M. spretus*: SEG (Granada, Spain), SFM (Montpellier, France), and STF (Fondouk Djedid, Tunisia), a more distantly related species but one that can still interbreed with *M. m. musculus* (Bonhomme et al. 1978; Burgio et al. 2007; Dejager et al. 2009). Our sampling included all Montpellier strains considered to be *M. m. domesticus*; most classical inbred strains are most closely related to this subspecies (Yang et al. 2007).

With the exception of 22MO and WMP, all strains were collected at least several kilometers from every other strain, so they should not be close relatives. This type of sampling is expected to maximize the total amount of genetic variation captured; however, it is inappropriate for population genetic analyses since the samples are not derived from a single population. All strains were initially maintained under a moderate inbreeding scheme, then under brother–sister mating for at least 20 generations and are thus highly inbred.

### Exome sequencing

DNA was extracted from spleen collected from female mice, when possible, using the Qiagen MasterPure™ Complete DNA and RNA Purification Kit and protocol from Epicentre (Madison, WI). DNA was sheared using a Bioruptor UCD-200 with 7 rounds of sonication (7 min per round on high, 30 s on 30 s off) and genomic DNA libraries were constructed and individually barcoded using a previously

described protocol designed to facilitate multiplexed exome capture (Rohland and Reich 2012). To reduce molecular interference during enrichment, we used truncated adaptors containing unique "internal" barcodes on the P5 end of genomic fragments (Rohland and Reich 2012). PCR primers were designed according to Rohland and Reich (2012).

In-solution sequence capture was performed using Nimblegen SeqCap EZ Mouse Exome probes as described in Nimblegen's SeqCap EZ Library User's Guide. Libraries were pooled equally to obtain 1 µg total DNA for each hybridization experiment. Libraries were then enriched using two separate capture reactions with eight libraries each, including blocking oligonucleotides specific to our custom adapters (Rohland and Reich 2012) and mouse COT-1 DNA (Invitrogen) to reduce nonspecific hybridization. The capture reactions were hybridized for 68 h at 47 °C in an Eppendorf Mastercycler Pro, and then washed, eluted, and PCR-enriched. Capture enrichment success was verified using qPCR analysis of three targeted regions on pre- and post-capture library pools. Sequencing was performed using 76-bp paired-end reads on the Illumina Hi-seq 2500 platform provided by the Epigenome Center at the University of Southern California.

Illumina reads were mapped to different pseudoreference genomes dependent upon their species of origin (Sarver et al. 2017). A pseudoreference contains the backbone of the mm10 reference mouse genome (strain = C57BL/6 J), but allelic states taken from representative strains of *M. m. musculus, M. m. domesticus, M. m. castaneus*, or *M. spretus* are inserted into the mm10 reference genome. All pseudoreferences were taken from Sarver et al. (2017). Preliminary analyses demonstrated that the four strains of unknown origin (BID, KAK, MPR, and TEH) fell within the *M. m. musculus* clade, and so were mapped to the *M. m. musculus* pseudoreference. The advantage of this technique is that species-specific variation can be incorporated into the reference, thus improving mapping accuracy and recovery, while preserving the coordinates of the mm10 build so that genome annotation can be used. Sequences were mapped with BWA MEM (v0.7.9a, Li and Durbin 2009), using default mapping parameters. Alignments to their respective pseudoreferences were used to identify variants using GATK HAPLOTYPECALLER (McKenna et al. 2010), following PCR duplicate removal and indel realignment. Variants from dbSNP (version 142) were used as the training set during SNP recalibration and subjected to standard hard filtering parameters according to the GATK Best Practices recommendation (Auwera et al. 2013; DePristo et al. 2011): MQ > 56, QD > 24, FS < 12, MQRankSum < 8, ReadPosRankSum < 3, DP > 3. To assess confidence, we compared SNPs identified in the Montpellier strains that also overlapped with dbSNP and quantified how many calls agreed with known allelic states at those sites. SNPs were classified by functional categories using GRCm38.73 as annotated by EMBL-EBI Ensembl (http://www.ensembl.org).

### Genetic relationships

We placed the 26 Montpellier exomes in the context of an additional 36 inbred strains commonly used in genetic research and compiled in dbSNP (v142) (Sherry et al. 1999, 2001). Most of these are "classical inbred strains," which share a common ancestry predominantly originating from a limited *M. m. domesticus* stock, with some human-mediated contribution from other species (Keane et al. 2011; Yang et al. 2007, 2011). Seven strains from dbSNP are considered wild-derived strains: three *M. m. domesticus* (LEWES/EiJ, WSB/EiJ, and ZALENDE/EiJ), one *M. m. musculus* (PWD/PhJ), one *M. m. molossinus* (MOLF/EiJ), one *M. m. castaneus* (CAST/EiJ), and one *M. spretus* (SPRET/EiJ). *M. m. molossinus* is a subspecies that originated through natural hybridization between *M. m. musculus* and *M. m. castaneus* (Yonekawa et al. 1986, 1988).

In addition to placing the Montpellier strains in the context of known variation among other inbred strains, we repeated our analyses after also combining our data with the genetic variants called by Harr et al. (2016), specifically their file named AllMouse.vcf_90_recalibrated_snps_raw_indels_reheader_PopSorted.PASS.vcf available from http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/. We only used sites covered in our Montpellier strains and dbSNP. Harr et al. (2016) performed whole genome sequencing from 27 *M. m. domesticus*, 22 *M. m. musculus*, and 8 *M. spretus*, to an average depth of 20.9X autosomal coverage, and their file includes 10 *M. m. castaneus* genomes from Halligan et al. (2010). The Harr et al. (2016) data are from wild-caught animals, and provide valuable context through which to view genetic relationships of the inbred strains (Montpellier + dbSNP strains).

SNPs that were found in our Montpellier panel as well as dbSNP were used to generate genealogies with the SNPHYLO program (Lee et al. 2014), which uses the maximum likelihood framework of the DNAML program in the PHYLIP package (Felsenstein 1993) to construct trees. One thousand bootstrap analyses were performed using all variants with the PHANGORN package (Schliep 2011) in R (http://www.r-project.org). Trees were visualized using the APE package (Paradis 2012; Paradis et al. 2004) in R. Resulting trees will be strongly affected by introgression known to occur in the history of inbreeding and therefore serve only as a rough approximation of genetic relationships.

We assessed genetic structure in our sample using STRUCTURE (Falush et al. 2007; Hubisz et al. 2009; Pritchard et al. 2000) and Principal Components Analyses (PCA). STRUCTURE was run under the admixture and correlated allele frequency model with ten independent runs of 10,000 burn-in

MCMC iterations followed by 50,000 iterations for 2 to 8 clusters (k = 2–8). Results were inspected with STRUCTURE HARVESTER (Earl 2012; Evanno et al. 2005). PCA was performed using the SNPRELATE package (Zheng et al. 2012). Unlike the above genealogical analyses, STRUCTURE and PCA require sites called in all 62 strains (i.e., no missing data across the 26 Montpellier exomes plus the 36 additional dbSNP strains). Multiallelic sites were excluded. To reduce the size of the dataset, we chose sites that were at least 100 kb apart, roughly the extent of linkage disequilibrium in wild mice (Laurie et al. 2007), resulting in 26,991 sites used in STRUCTURE and PCA analyses.

### Protein-coding variation

In addition to classifying SNPs into basic categories such as nonsynonymous and synonymous, we estimated codon usage bias across strains, using ENCPRIME (Novembre 2002), which tests the null hypothesis that codons within an amino acid class are used at equal frequency, after accounting for background base composition. Estimates of codon usage theoretically range from 20 (every amino acid coded by a single codon, representing maximal bias) to 61 (each amino acid coded by each of its synonymous codons at equal frequency, representing minimal bias). Background base composition was estimated from flanking nonexonic sequence, which is inevitably sequenced even when performing exome enrichment. Because base composition varies across the genome, we analyzed codon usage bias in 5 Mb windows. Lastly, we annotated early stop codons, which could represent alternatives to traditional gene knockouts if they disrupt normal gene function. To infer their functional impact, we determined the percentage of the protein truncated by the early stop codons, and whether they occurred on constitutive or facultative exons.

### Gene flow between lineages and strains

Previous studies have shown that many laboratory strains have a mosaic genome that contains genetic material from multiple species (Dai et al. 2005; Ferris et al. 1982; Frazer et al. 2007; Ideraabdullah et al. 2004; Nagamine et al. 1992; Tucker et al. 1992; Yalcin et al. 2004; Yang et al. 2007, 2011). We tested for interspecific introgression using the ABBA-BABA test (Green et al. 2010). In this test, "A" and "B" indicate the distribution of a biallelic state across a rooted four-taxon genealogy. We tested the genealogy ([(*M. m. domesticus* strain #1, *M. m. domesticus* strain #2), *M. m. musculus*], *M. spretus*) represented by ([(DOM*, LEWES/EiJ), CzechII/EiJ], SPRET/EiJ), where DOM* represents each Montpellier strain of *M. m. domesticus* tested individually. LEWES/EiJ and CzechII/EiJ were chosen to represent "pure" *M. m. domesticus* and *M. m. musculus*, respectively, as neither shows high levels of introgression (Yang et al. 2011, and unpublished data). We have sequenced the genome of CzechII/EiJ as part of unrelated work, only using variant calls here for the ABBA-BABA implementation. These results should be treated with caution as *M. m. musculus* and *M. m. domesticus* are closely related, can still interbreed in nature, and their genomes are not fully differentiated. An excess of either ABBA or BABA sites is best explained by introgression, in our case from *M. m. musculus* (represented by CzechII/EiJ) into *M. m. domesticus*.

We divided the genome into 489 nonoverlapping 5 Mb windows, discarding any that had fewer than 20 ABBA-BABA sites with genotypes called with a minimum genotype quality (GQ) of 10. GQ is the Phred-scaled confidence in a called genotype and is strongly correlated to the number of reads that map to a particular site, the quality of the base calls from those reads, and sequencing error. For any DOM* strain that contained at least 20 windows with at least 20 ABBA-BABA sites, the null hypothesis of no introgression was evaluated using the block Jackknife procedure as implemented in the script JACKKNIFE.R from the ANGSD package (Korneliussen et al. 2014).

Several recent studies have raised the possibility of gene flow between some wild-derived strains that may have unintentionally occurred after their establishment in the laboratory (Yang et al. 2007, 2009, 2011). To test for more recent introgression, we inspected the distribution of pairwise divergence between every possible pair of Montpellier strains having shared the same lab environment. Recent introgression should result in long stretches of the genome that are identical by descent. We used the B-SMUCE segmentation algorithm of Futschik et al. (2014) implemented in the SMUCER function of the R package STEPR with default settings. This multiscale segmentation algorithm looks for compositionally homogeneous segments. It estimates the best fit to a series of a minimal number of steps and provides estimates for the number of segments and their boundaries at the same time. Originally employed to find regional variation in GC (1) versus AT (0) content, we applied this algorithm to a simple literal distance whereby two homologous SNPs yield 0 if they were identical or identically heterozygous, 1 if different, and 0.5 if one of them was heterozygous. The program was run with default options to estimate the chromosomal distribution of haplotype sharing between every pair of strains.

### Simulating mapping power

To quantify the effect of additional genetic variation on the power to map quantitative trait loci, we simulated two different backcross experiments, started by two different pairs of parental strains: C57BL/6J and DBA/2J, which are the parental strains to the BXD family, a classic recombinant

inbred family from which over 5000 phenotypes have been collected (Wang et al. 2016), and DGA and DJO, two Montpellier strains sequenced in this study. We confined the simulation to sites that were covered in all four strains. We did not include the X chromosome. Using the R package QTL (Broman and Sen 2009; Broman et al. 2003), we simulated a backcross design with heritability of 0.5 and 100 individuals sampled. We systematically simulated QTL along the genome at intervals of 10 cM, and then performed Haley–Knott regression (Haley and Knott 1992) using the SCANONE function of QTL (Broman and Sen 2009; Broman et al. 2003). We quantified the differences in maximal LOD scores, as well as the average length of the 95% confidence interval in QTL, quantified using the LODINT function in QTL, in these two hypothetical backcross designs.

## Results and discussion

### SNP discovery

An average of 19.9 million reads were generated per Montpellier strain, with an average of 73.5% uniquely mapping to the genome (Supplementary Material 8). This amounted to an average coverage of 8.1× at called SNPs (Supplementary Material 8). All raw sequencing data are deposited as NCBI BioProject PRJNA326865. All called SNPs are provided as a Variant Call Format file (Supplementary Material 9).

Using a conservative SNP-calling pipeline and alleles in the mm10 genome as reference, we identified 1,460,057 SNPs (77,439 nonsynonymous, 166,048 synonymous, 178,088 untranslated regions, 740,923 intronic, and 297,559 intergenic) among 26 Montpellier strains (Table 1). Of these, 1,184,277 (81%) occurred in dbSNP. The Montpellier strains thus contribute 275,780 novel SNPs not previously known from inbred strains (18,496 nonsynonymous, 35,833 synonymous, 40,325 untranslated regions, 127,641 intronic, and 53,485 intergenic), with roughly one novel SNP every 12,000 bp. We did not experimentally validate called SNPs through further sequencing; however, among the SNPs identified in the Montpellier strains that also overlapped with dbSNP, 99.8% agreed with the known allelic states at those sites, confirming our pipeline yielded high-quality SNP calls. In addition to novel SNPs, we identified 152,342 insertion/deletion mutations (Supplementary Material 10).

We repeated our analyses after including only a single subspecies, showing that novel variants were not confined to a subset of strains sampled (Table 1). We reached a similar conclusion after systematically excluding one subspecies at a time (Table 1).

### Genetic relationships

The four *M. spretus* strains (Fig. 1, left panel, strains labeled in black) formed a distinct group that was distantly related to the other strains, consistent with previous phylogenetic hypotheses (Lundrigan et al. 2002; Sarver et al. 2017). *M. m. castaneus* and *M. m. musculus* formed a group (Fig. 1, left panel, strains labeled in green) that then grouped with *M. m. domesticus* strains (Fig. 1, left panel, strains labeled in red), also consistent with previous phylogenetic inference (Keane et al. 2011; Phifer-Rixey et al. 2012; White et al. 2009). *M. m. domesticus, M. m. musculus*, and *M. m. castaneus* are closely related lineages that diverged less than 350 thousand

**Table 1** Number of SNPs observed (in parentheses: number not observed in dbSNP version 142) across the Montpellier strains

| | Include one (sub) species | | | | Exclude one (sub) species | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Type | All genotypes | *M. m. castaneus* | *M. m. domesticus* | *M. m. musculus* | *M. spretus* | M. m. castaneus | M. m. domesticus | M. m. musculus | M. spretus |
| All SNPs | 1,460,057 (275,780) | 423,044 (50,233) | 673,858 (72,241) | 713,103 (100,122) | 957,550 (104,101) | 1,404,809 (241,355) | 1,374,654 (236,018) | 1,296,270 (203,381) | 1,043,512 (192,527) |
| Nonsynonymous | 77,439 (18,496) | 22,230 (2947) | 39,558 (6068) | 33,688 (6204) | 52,373 (6196) | 74,073 (16,292) | 71,556 (14,505) | 70,305 (13,779) | 54,783 (13,601) |
| Synonymous | 166,048 (35,833) | 52,873 (6708) | 88,900 (10,993) | 76,383 (11,432) | 114,176 (12,887) | 158,386 (30,904) | 155,812 (28,991) | 153,004 (27,667) | 120,723 (25,575) |
| UTR_5_PRIME | 22,412 (5210) | 6966 (914) | 11,579 (1447) | 10,217 (1694) | 15,566 (2109) | 21,425 (4565) | 21,119 (4401) | 20,438 (4006) | 16,040 (3513) |
| UTR_3_PRIME | 155,676 (35,115) | 46,086 (6455) | 77,481 (9577) | 71,698 (12,568) | 104,339 (12,898) | 148,707 (30,520) | 146,639 (29,816) | 140,354 (25,927) | 112,167 (24,907) |
| Intronic | 740,923 (127,641) | 214,331 (23,643) | 317,763 (29,848) | 372,879 (48,261) | 486,745 (49,911) | 715,231 (112,058) | 701,972 (112,323) | 654,260 (92,930) | 521,351 (87,140) |
| Intergenic | 297,559 (53,485) | 80,558 (9566) | 138,577 (14,308) | 148,238 (19,963) | 184,351 (20,100) | 286,987 (47,016) | 277,556 (45,982) | 257,909 (39,072) | 218,448 (37,791) |

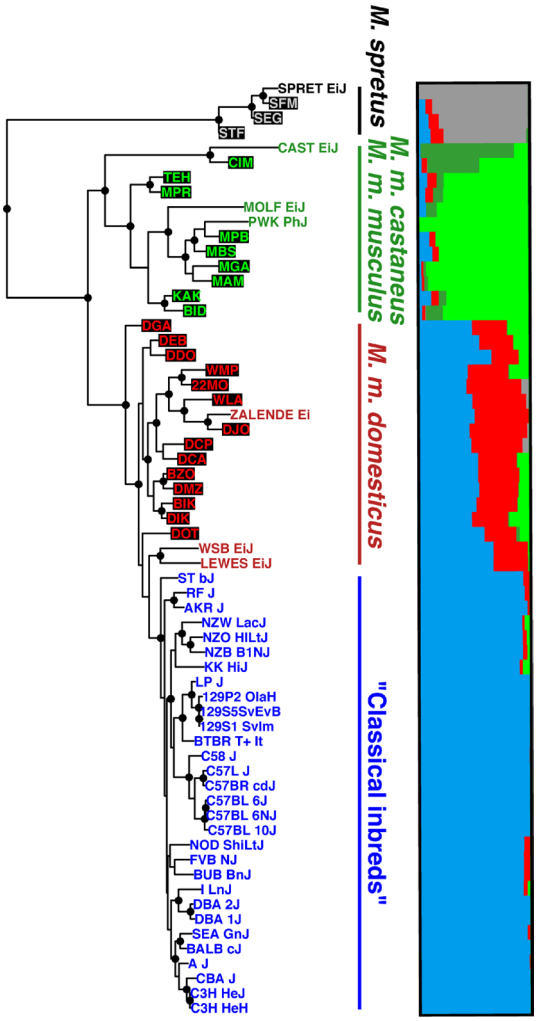


**Fig. 1** Genealogical relationships among 26 Montpellier strains and 36 strains from dbSNP, determined by **a** maximum likelihood tree, or **b** STRUCTURE analysis considering k = 5. Strain names on a *black box* indicate novel Montpellier exomes sequenced in this study. Nodes labeled with *small black circles* were supported with at least 95% bootstrap support

years ago (Boursot et al. 1996; Geraldes et al. 2008; Salcedo et al. 2007; She et al. 1990; Suzuki et al. 2004). Many genomic regions are still unsorted and lack species-specific substitutions (Geraldes et al. 2011; Salcedo et al. 2007). MOLF/EiJ falls within the other *M. m. musculus* strains, suggesting that even though it is a natural hybrid species between *M. m. castaneus* and *M. m. musculus* (Yonekawa et al. 1986, 1988), a majority of its genome is derived from *M. m. musculus*.

The classical inbred strains (Fig. 1, left panel, strains labeled in blue) fell within the *M. m. domesticus* group, consistent with previous studies suggesting they are mostly derived from this subspecies (Yang et al. 2007, 2009, 2011). Interestingly, the classical inbred strains formed their own group within wild-derived *M. m. domesticus*, and even other wild-derived inbred strains from The Jackson Laboratory, like LEWES/EiJ and WSB/EiJ, fall outside of this group. The distinct grouping of the classical inbred strains reinforces what is known about their history that they derived from a small set of founders during the early mouse fancy trade (Beck et al. 2000; Frazer et al. 2007; Moriwaki 1994; Morse 1978, 2007; Silver 1995; Wade and Daly 2005; Wade et al. 2002). These general groups remained unchanged after repeating the analysis with wild-caught *M. m. domesticus, M. m. castaneus, M. m. musculus*, and *M. spretus* of Harr et al. (2016) (Supplementary Fig. 1).

The four strains that were not assigned to a subspecies *a priori* (BID, KAK, MPR, and TEH) all fell within the *M. m. musculus* clade, but in a basal position within that group (Fig. 1, left panel). These strains originated from individuals collected near the center of mouse diversity in the Middle East (Pakistan and Iran) (Hardouin et al. 2015). Some of these populations may constitute independent (sub-) species that have yet to be described (Rajabi-Maham et al. 2012), or they may have captured ancestral polymorphism or secondary admixture that occurred in nature.

STRUCTURE analyses largely supported the genealogical tree, with log-likelihood values reaching a plateau at K = 5 groups. The Evanno method was used to characterize the clustering, with delta K peaking at K = 3 and falling to 0 at K = 5. The two *M. m. castaneus* strains (dark green, Fig. 1, right panel) shared a large proportion of genetic variation with *M. m. musculus* (light green, Fig. 1, right panel). Interestingly, the allele frequencies among wild-derived *M. m. domesticus* strains (red, Fig. 1, right panel) differed from those of classical inbred strains (blue, Fig. 1, right panel). There is no evidence for genetic structure within classical inbred strains, which capture approximately half of the structure seen in wild-derived *M. m. domesticus* (blue in Fig. 1, right panel). At K = 3, *M. spretus, M. m. castaneus* + *M. m. musculus*, and *M. m. domesticus* form distinct clusters. At K = 4, *M. m. castaneus* separates from *M. m. musculus*. At K = 5, the classical inbreds separate from the wild-derived *M. m. domesticus*. K = 6 does not separate out any samples from the rest.

PCA also supported the groupings seen in the genealogy (Supplementary Fig. 2). The first principal component explained 45.2% of the variation and separated the four *M. spretus* strains from the rest of the panel. The second principle component explained 16.7% of the genetic variation and separated the *M. m. musculus* and *M. m. castaneus* strains from the rest of the panel.

## Protein-coding variation

*M. spretus* showed low codon usage bias (corresponding to high ENCprime estimates), *M. m. castaneus* and *M. m.*

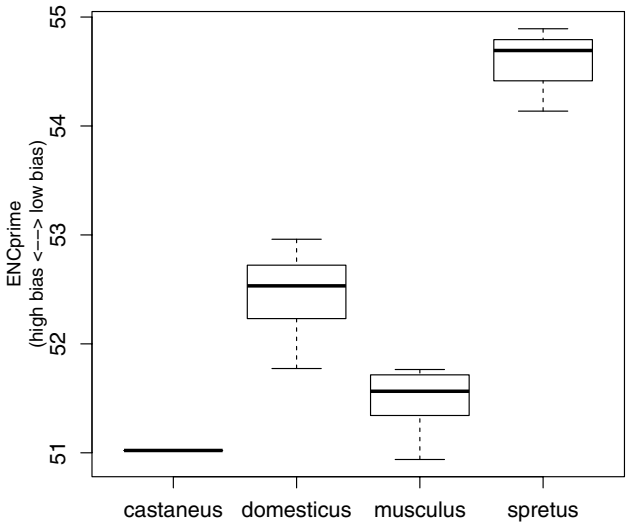


**Fig. 2** Codon usage bias among Montpellier strains

*musculus* strains had high codon usage bias, and *M. m. domesticus* was intermediate (Fig. 2). Natural selection shapes codon usage bias across mammals, but shows no consistent correlation to proxies of effective population size (Kessler and Dean 2014). The effective population sizes of *M. m. castaneus, M. m. domesticus*, and *M. m. musculus* have been genetically estimated at 220, 100, and 60 K, respectively (Geraldes et al. 2008, 2011; Phifer-Rixey et al. 2012). We might hypothesize that *M. spretus* has a smaller effective population size than *M. musculus* subspecies because it occupies a smaller geographic area and maintains lower density than *M. musculus* subspecies (Boursot et al. 1985; Britton and Thaler 1978; Dejager et al. 2009; Orsini et al. 1982). As predicted by a model of weak selection, *M. m. castaneus* has the strongest, and *M. spretus* the weakest, codon usage bias. However, the two intermediate taxa do not follow the predictions of weak selection. Thus, there is no consistent relationship between inferred effective population size and codon usage bias estimated from our exomes, as observed by Kessler and Dean (2014).

Across all 26 exomes of wild-derived inbred strains, we identified 262 genes with at least one early stop codon segregating in at least one isoform, many of which segregate in multiple strains (Supplementary Material 11). Of these, 193 truncate less than 10% of the protein and therefore may not have a strong functional impact (Supplementary Fig. 3). However, 117 genes segregate an early stop codon that truncates more than 50% of its wild-type length (Supplementary Fig. 3), and could represent a novel source of effective knockouts for future functional studies. Surprisingly, the length of wild-type protein truncated by early stop codons did not differ between early stop codons that occurred in constitutively vs. facultatively spliced exons ($t = 0.98$, df = 260, $P = 0.33$) (Supplementary Fig. 4), suggesting that both types of stop codons have similar functional impacts. In fact, more stop codons affected constitutively ($N = 143$) versus facultatively ($N = 119$) expressed exons.

Here, we highlight one gene, *Bard1*, with an early stop codon that truncates 50.6% of the wild-type protein in two Montpellier strains, BIK and DDO (Supplementary Material 11). BARD1 forms a heterodimer with BRCA1, which together plays an important role in chromosomal stability and tumor suppression. A truncated BARD1 protein results in defective homologous DNA repair (Westermark et al. 2003), and knockouts for *Bard1* die at an early embryonic stage (McCarthy et al. 2003). Whether or not the two Montpellier strains with truncated BARD1 have altered chromosomal stability represents one of many potentially interesting follow-up studies.

## Gene flow between lineages and strains

Eight wild-derived strains of *M. m. domesticus* (BIK, DCP, DDO, DEB, DIK, DJO, DOT, WLA) showed a significant ABBA-BABA result (|Z-score| > 3, Supplementary Material 12), suggesting past introgression from *M. m. musculus* (represented by CzechII/EiJ). Four strains (22MO, BZO, DMZ, DCA) did not show a significant ABBA-BABA result. Among these 12 strains, we analyzed an average of 2961.5 ABBA-BABA sites from an average of 79.3 windows. The remaining two strains (DGA and WMP) lacked enough data to apply an ABBA-BABA test (fewer than 20 windows with at least 20 ABBA-BABA sites).

Using the B-SMUCE segmentation algorithm (Futschik et al. 2014), two *M. m. domesticus* strains DJO and 22MO show "typical" segment lengths of estimated genetic distances between two independently derived strains (Supplementary Fig. 5, Supplementary Material 13). In contrast, the two *M. m. domesticus* strains DEB and BZO show a marked shift towards long segments with no genetic distance (Supplementary Fig. 5), a possible signature of recent introgression followed by limited recombination and inbreeding. However, the fact that all regions have pairwise genetic distances greater than zero (Supplementary Fig. 5) argues against recent introgression occurring after laboratory establishment.

## Simulating mapping power

The power to detect quantitative trait loci will depend in part on the number of markers across the genome. DJO and DGA have ample variants between them where

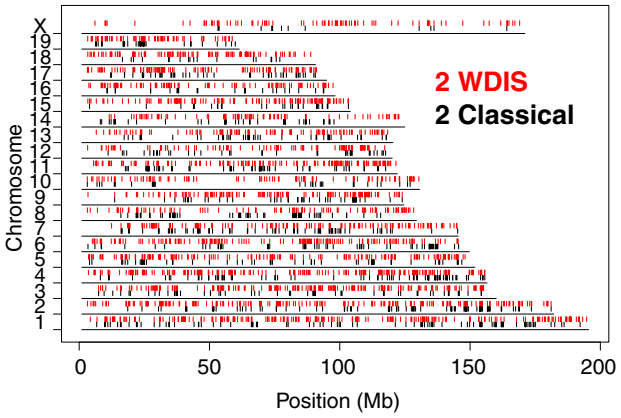


**Fig. 3** SNP density between DJO and DGA or between C57BL/6J and DBA/2J

well-known "SNP deserts" occur in classical inbred strains, including on chromosomes 10, 16 and X (Yang et al. 2007) (Fig. 3). The power to detect quantitative trait loci was higher, and the confidence intervals narrower, in the simulated DJO x DGA cross compared to the C57 × DBA cross. With a sample of 100 individuals, and a single QTL with heritability = 0.5, the median LOD score of DJO x DGA F2 descendants was 0.13 higher than C57 x DBA descendants, which amounts to a 25% reduction in a traditional p-value. Fourteen regions were detected as a significant QTL in the DJO × DGA cross that were not significant in C57 × DBA (Fig. 4a). Furthermore, the average confidence intervals were 3.19 cM shorter in DJO × DGA compared to C57 × DBA cross (Fig. 4b). The narrower confidence interval translates to roughly 3 Mb of genome, or roughly 20 genes.

## Conclusions

Mouse genetics relies heavily on classical inbred strains (the blue strains of Fig. 1). Given that most classical strains are highly related to each other and capture a small amount of genetic variation from their wild progenitors, such research is inherently underpowered to link genomic and phenotypic variation. Our study identified several thousand variants that were previously unknown from inbred strains. These include many that are predicted to have functional impact, including nonsynonymous mutations and early stop codons. Our data demonstrate that the Montpellier strains would increase the power of mouse genetics.

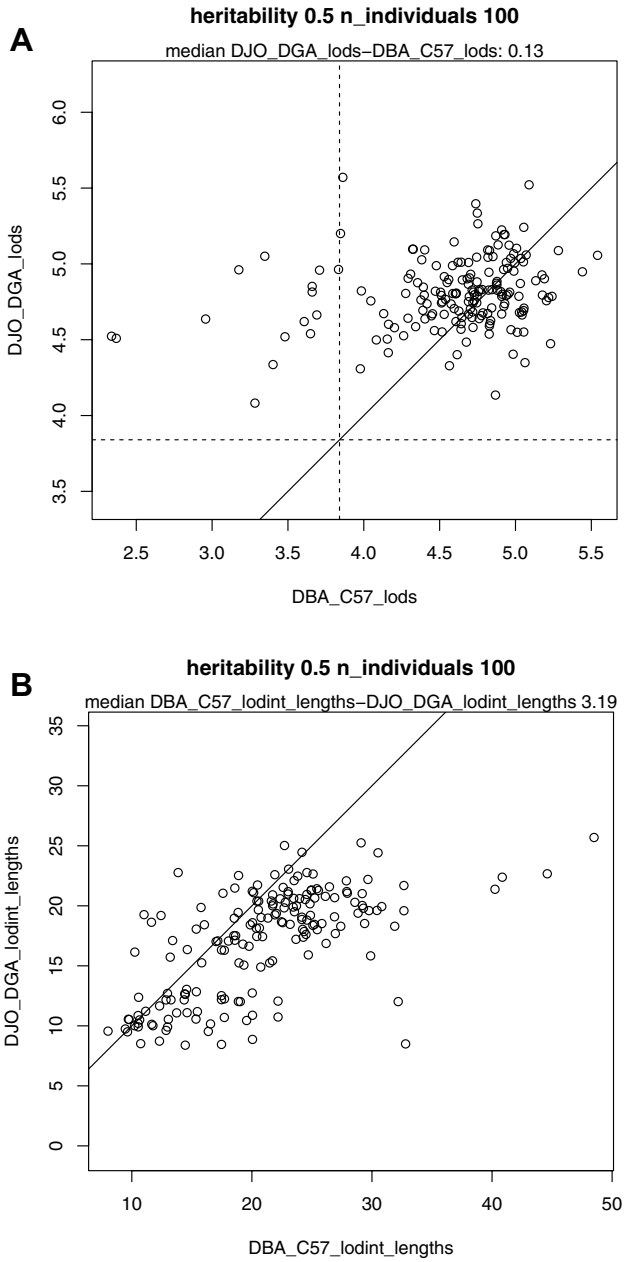


**Fig. 4** Simulated F2 cross between DJO and DGO resulted in **a** higher LOD scores, and **b** narrower confidence intervals compared to a F2 cross between C57BL/6J and DBA/2J

**Acknowledgements** We thank Charlie Nicolet and Selene Tyndale from the Epigenome Center at USC. Brent Young, Rachel Mangels, and Lorraine Provencio helped with molecular work. Matt Salomon and Rob Williams gave many helpful suggestions. Jean-Jacques Duquesne maintained the wild mouse repository in Montpellier. Funding was provided by the National Institutes of Health Grant #GM098536 (MDD), National Science Foundation Grant #1146525 (MDD), the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health Grant #HD073439 (JMG), and the University of Montana Genomics Core, supported by a grant from the M.J. Murdock Charitable Trust.

**Compliance with ethical standards**

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References


Auwera GA, Carneiro MO, Hartl C et al (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinform 43:11.10.11–11.10.33

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

Bonhomme F, Martin S, Thaler L (1978) Hybridation en laboratoire de *Mus musculus* L. et *Mus spretus* Lataste. Experientia 34:1140–1141

Boursot P, Jacquart T, Bonhomme F, Britton-Davidian J, Thaler L (1985) Differenciation geographique du genome mitochondrial chez *Mus spretus* Lataste. Comptes rendus de l'Academie des sciences 301:161–166

Boursot P, Din W, Anand R, Darviche D, Dod B, Von Deimling F, Talwar GP, Bonhomme F (1996) Origin and radiation of the house mouse: mitochondrial DNA phylogeny. J Evol Biol 9:391–415

Britton J, Thaler L (1978) Evidence for the presence of two sympatric species of mice (genus Mus L.) in southern France based on biochemical genetics. Biochem Genet 16:213–225

Broman KW, Sen S (2009) A guide to QTL mapping with R/qtl. Springer, New York

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Burgio G, Szatanik M, Guenet J-L, Arnau M-R, Panthier J-J, Montagutelli X (2007) Interspecific recombinant congenic strains between C57BL/6 and mice of the *Mus spretus* species: a powerful tool to dissect genetic control of complex traits. Genetics 177:2321–2333

Church DM, Goodstadt L, Hillier LW et al (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7:e1000112

Dai J-g, Min J-x, Xiao Y-b, Lei X, Shen W-h, Wei H (2005) The absence of mitochondrial DNA diversity among common laboratory inbred mouse strains. J Exp Biol 208:4445–4450

Dejager L, Libert C, Montagutelli X (2009) Thirty years of Mus spretus: a promising future. Trends Genet 25:234–241

DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

Didion J, Pardo-Manuel de Villena F (2013) Deconstructing Mus gemischus: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. Mamm Genome 24:1–20

Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resour 4:359–361

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578

Felsenstein J (1993) {PHYLIP}: phylogenetic inference package, version 3.5 c.

Ferris SD, Sage RD, Wilson AC (1982) Evidence from mtDNA sequences that common laboratory strains of inbred mice are descended from a single female. Nature 295:163–165

Frazer KA, Eskin E, Kang HM et al (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448:1050–1053

Futschik A, Hotz T, Munk A, Sieling H (2014) Multiscale DNA partitioning: statistical evidence for segments. Bioinformatics 30:2255–2262

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol 17:5349–5363

Geraldes A, Basset P, Smith KL, Nachman MW (2011) Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Mol Ecol 20:4722–4736

Green RE, Krause J, Briggs AW et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Grubb SC, Churchill GA, Bogue MA (2004) A collaborative database of inbred mouse strain characteristics. Bioinformatics 20:2857–2859

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD, Nachman MW (2010) Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet 6(1):e1000825

Hardouin EA, Orth A, Teschke M, Darvish J, Tautz D, Bonhomme F (2015) Eurasian house mouse (*Mus musculus* L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. BMC Evol Biol 15:26

Harr B, Karakoc E, Neme R et al. (2016) Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data 3:160075

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Molecular ecology resources 9:1322–1332

Ideraabdullah FY, de la Casa-Esperon E, Bell TA, Detwiler DA, Magnuson T, Sapienza C, Pardo-Manuel de Villena F (2004) Genetic and haplotype diversity among wild-derived mouse inbred strains. Genome Res 14:1880–1887

Keane TM, Goodstadt L, Danecek P et al (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294

Kessler MD, Dean MD (2014) Effective population size does not predict codon usage bias in mammals. Ecol Evol 4:3887–3900

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinform 15:1

Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, Nachman MW (2007) Linkage disequilibrium in wild mice. PLos Genet 3:e144

Lee TH, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Lindblad-Toh K, Winchester E, Daly MJ et al (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat Genet 24:381–386

Lundrigan BL, Jansa SA, Tucker PK (2002) Phylogenetic relationships in the genus *Mus*, based on paternally, maternally, and biparentally inherited characters. Syst Biol 51:410–431

McCarthy EE, Celebi JT, Baer R, Ludwig T (2003) Loss of Bard1, the heterodimeric partner of the Brca1 tumor suppressor, results in early embryonic lethality and chromosomal instability. Mol Cell Biol 23:5056–5063

McKenna A, Hanna M, Banks E et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Moriwaki K (1994) Wild mouse from a geneticist's viewpoint. In Genetics in wild mice Tokyo. Japan Scientific Societies Press, Japan

Morse HC (1978) Origins of inbred mice. Academic Press, Cambridge

Morse HCI (2007) Building a better mouse: one hundred years of genetics and biology. In: Fox JG, Barthold SW, Davisson MT, Newcomer CE, Quimby FW, Smith AL (eds) The mouse in biomedical research. Elsevier, Waltham

Nagamine CM, Nishioka Y, Moriwaki K, Boursot P, Bonhomme F, Lau YFC (1992) The musculus-type Y chromosome of the laboratory mouse is of Asian origin. Mamm Genome 3:84–91

Nikolskiy I, Conrad DF, Chun S, Fay JC, Cheverud JM, Lawson HA (2015) Using whole-genome sequences of the LG/J and SM/J inbred mouse strains to prioritize quantitative trait genes and nucleotides. BMC Genom 16:415

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394

Orsini P, Cassaing J, Duplantier J, Croset H (1982) Premieres donnees sur l'ecologie des populations naturelles de souris, Mus spretus Lataste et Mus musculus domesticus Rutty dans le Midi de la France

Paigen K (2003a) One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). Genetics 163:1–7

Paigen K (2003b) One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). Genetics 163:1227–1235

Paradis E (2012) Analysis of phylogenetics and evolution with R. Springer, New York

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Petkov PM, Ding Y, Cassell MA et al (2004) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14:1806–1811

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW (2012) Adaptive evolution and effective population size in wild house mice. Mol Biol Evol 29:2949–2955

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rajabi-Maham H, Orth A, Siahsarvie R, Boursot P, Darvish J, Bonhomme F (2012) The south-eastern house mouse Mus musculus castaneus (Rodentia: Muridae) is a polytypic subspecies. Biol J Linn Soc 107:295–306

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946

Salcedo T, Geraldes A, Nachman MW (2007) Nucleotide variation in wild and inbred mice. Genetics 177:2277–2291

Sarver B, Keeble S, Cosart T, Tucker P, Dean MD, Good JM (2017) Phylogenomic insights into mouse evolution using a pseudoreference approach. Genome Biol Evol 9:726–739

Schliep KP (2011) Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593

She JX, Bonhomme F, Boursot P, Thaler L, Catzeflis F (1990) Molecular phylogenies in the genus *Mus* - comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. Biol J Linn Soc 41:83–103

Sherry ST, Ward M, Sirotkin K (1999) dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9:677–679

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Silver L (1995) Mouse genetics: concepts and applications. Oxford University Press, New York

Srivastava A, Morgan AP, Najarian ML et al (2017) Genomes of the mouse collaborative cross. Genetics 206:537–556

Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K (2004) Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. Mol Phylogenet Evol 33:626–646

Tucker PK, Lee BK, Lundrigan BL, Eicher EM (1992) Geographic origin of the Y chromosomes in "old" inbred strains of mice. Mammalian genome 3:254–261

Wade CM, Daly MJ (2005) Genetic variation in laboratory mice. Nat Genet 37:1175–1180

Wade CM, Kulbokas EJ 3rd, 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 X, Pandey AK, Mulligan MK et al. (2016) Joint mouse–human phenome-wide association to test gene function and disease risk. Nat Commun. doi:10.1038/ncomms10464

Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Westermark UK, Reyngold M, Olshen AB, Baer R, Jasin M, Moynahan ME (2003) BARD1 participates with BRCA1 in homology-directed repair of chromosome breaks. Mol Cell Biol 23:7926–7936

White MA, Ané C, Dewey CN, Larget BR, Payseur BA (2009) Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet 5:e1000729

White JK, Gerdin A-K, Karp NA et al (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell 154:452–464

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

Yalcin B, Fullerton J, Miller S et al (2004) Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. Proc Natl Acad Sci USA 101:9734–9739

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, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, Paigen BJ, Graber JH, de Villena FP-M, Churchill GA (2009) A customized and versatile high-density genotyping array for the mouse. Nat Meth 6:663–666

Yang H, Wang JR, Didion JP et al (2011) Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43:648–655

Yonekawa H, Gotoh O, Tagashira Y, Matsushima Y, Shi LI, Cho WS, Miyashita N, Moriwaki K (1986) A hybrid origin of Japanese mice "*Mus musculus* molossinus". Curr Top Microbiol Immunol 127:62–67

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima Y, Shi LM, Cho WS, Zhen XL, Tagashira Y (1988) Hybrid origin of Japanese mice "Mus musculus molossinus": evidence from restriction analysis of mitochondrial DNA. Mol Biol Evol 5:63–78

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328