

1 *The recombination landscape in wild house mice inferred using population genomic*

2 *data*

3

4 Tom R. Booker¹, Rob W. Ness², Peter D. Keightley¹

5

6 ¹*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3FL,*

7 *United Kingdom*

8 ²*Department of Biology, University of Toronto Mississauga, Mississauga, ON,*

9 *Canada*

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26 **Abstract**

27

28 Characterizing variation in the rate of recombination across the genome is
29 important for understanding many evolutionary processes. The landscape of
30 recombination has been studied previously in the house mouse, *Mus musculus*, and
31 it is known that the different subspecies exhibit different suites of recombination
32 hotspots. However, it is not established whether broad-scale variation in the rate of
33 recombination is conserved between the subspecies. In this study, we construct a
34 fine-scale recombination map for the Eastern house mouse subspecies, *M. m.*
35 *castaneus*, using 10 individuals sampled from its ancestral range. After inferring
36 phase, we use LDhelmet to construct recombination maps for each autosome. We
37 find that the spatial distribution of recombination rate is strongly positively between
38 our *castaneus* map and a map constructed using inbred lines of mice derived
39 predominantly from *M. m. domesticus*. We also find that levels of genetic diversity in
40 *M. m. castaneus* are positively correlated with the rate of recombination, consistent
41 with pervasive natural selection acting in the genome. Our study suggests that
42 recombination rate variation is conserved at broad scales between *M. musculus*
43 subspecies.

44

45

46

47

48

49

50

51 **Introduction**

52

53 In many species, rates of crossing-over are not uniformly distributed across
54 chromosomes, and understanding this variation and its causes is important for many
55 aspects of molecular evolution. Experiments in laboratory strains or managed
56 populations examining the inheritance of markers through pedigrees have allowed
57 direct estimation of rates of crossing over in different regions of the genome. Studies
58 of this kind are impractical for many wild populations, where pedigree structures are
59 largely unknown (but see Johnston *et al.* 2016). In mice, there have been multiple
60 genetic maps published (e.g. Jensen-Seaman *et al.* 2004; Paigen *et al.* 2008; Cox *et al.*
61 *et al.* 2009; Liu *et al.* 2014), typically using the classical inbred laboratory strains, which
62 are predominantly derived from the Western European house mouse subspecies,
63 *Mus musculus domesticus* (Yang *et al.* 2011). Recombination rate variation in
64 laboratory strains may not, therefore, reflect natural rates and patterns in wild mice of
65 different subspecies. In addition, recombination rate modifiers may have become
66 fixed in the process of laboratory strain management. On the other hand, directly
67 estimating recombination rates in wild house mice is not feasible without both a
68 population's pedigree and many genotyped individuals (but see Wang *et al.* 2017).

69

70 To understand variation in recombination rates, patterns of linkage
71 disequilibrium (LD) in a sample of individuals drawn from a population can be used.
72 Coalescent-based methods have been developed that use such data to indirectly
73 estimate recombination rates at very fine scales (Hudson 2001; Mcvean *et al.* 2002;
74 Mcvean *et al.* 2004; Auton and Mcvean 2007; Chan *et al.* 2012). The recombination
75 rates estimated in this way reflect variation in crossing over rates in populations

76 ancestral to the extant population, and are averages between the sexes. Methods
77 using LD have been applied to explore variation in recombination rates among
78 mammals and other eukaryotes, and have demonstrated that recombination
79 hotspots are associated with specific genomic features (Myers *et al.* 2010; Paigen
80 and Petkov 2010; Singhal *et al.* 2015).

81

82 The underlying mechanisms explaining the locations of recombination events
83 have been the focus of much research. In house mice and in most other mammals,
84 the PRDM9 zinc-finger protein binds to specific DNA motifs, resulting in an increased
85 probability of double-strand breaks, which can then be resolved by reciprocal
86 crossing-over (Grey *et al.* 2011; Baudat *et al.* 2013). Accordingly, it has been shown
87 that recombination hotspots are enriched for PRDM9 binding sites (Myers *et al.*
88 2010; Brunschwig *et al.* 2012). PRDM9-knockout mice still exhibit hotspots, but in
89 dramatically different genomic regions (Brick *et al.* 2012). Variation in PRDM9,
90 specifically in the exon encoding the zinc-finger array, results in different binding
91 motifs (Baudat *et al.* 2010). Davies *et al.* (2016) generated a line of mice in which the
92 exon encoding the portion of the PRDM9 protein specifying the DNA binding motif
93 was replaced with the orthologous human sequence. The recombination hotspots
94 they observed in this 'humanized' line of mice were enriched for the PRDM9 binding
95 motif observed in humans.

96

97 Great ape species have different alleles of the PRDM9 gene (Schwartz *et al.*
98 2014) and relatively little hotspot sharing (Winckler *et al.* 2005; Stevison *et al.* 2015).
99 Correlations between the broad-scale recombination landscapes of the great apes
100 are, however, relatively strongly positive (Stevison *et al.* 2011; Stevison *et al.* 2015).

101 This suggests that, while hotspots evolve rapidly, the overall genetic map changes
102 more slowly. Indeed, multiple closely related species pairs with different hotspot
103 locations show correlations between recombination rates at broad scales
104 (Smukowski and Noor 2011), as do species that share hotspots or lack them
105 altogether (Singhal *et al.* 2015; Smukowski Heil *et al.* 2015).

106

107 It has been suggested that a population ancestral to the *M. musculus* species
108 complex began to split into the present day subspecies around 500,000 years ago
109 (Geraldes *et al.* 2008). In this time, functionally distinct alleles of the PRDM9 gene
110 and different suites of hotspots have evolved in the subspecies (Smagulova *et al.*
111 2016). In addition, between members of the *M. musculus* subspecies complex, there
112 is also variation in recombination rates at relatively broad scales for multiple regions
113 of the genome (Dumont *et al.* 2011), and there is genetic variation in recombination
114 rate is polymorphic between *M. m. domesticus* individuals (Wang *et al.* 2017).
115 Brunschwig *et al.* (2012) analyzed single nucleotide polymorphism (SNP) data for
116 classical laboratory strains of mice, and used an LD-based approach to estimate the
117 sex-averaged recombination landscape for the 19 mouse autosomes. The
118 recombination rate map they constructed is similar to a genetic map generated using
119 crosses by Cox *et al.* (2009). Both studies were conducted using the classical inbred
120 lines (whose ancestry is largely *M. m. domesticus*), and their estimated
121 recombination rate landscapes may therefore reflect that of *M. m. domesticus* more
122 than other members of the *M. musculus* species complex.

123

124 In this study, we construct a recombination map for the house mouse
125 subspecies *M. m. castaneus*. We used the genome sequences of 10 wild-caught

126 individuals of *M. m. castaneus* from the species' expected ancestral range, originally
127 reported by Halligan *et al.* (2013). In our analysis, we first phased SNPs and
128 estimated rates of error in phasing. Secondly, we simulated data to assess the
129 power of estimating recombination rates based on 10 individuals and the extent by
130 which phase errors lead to biased estimates of the rate of recombination. Finally,
131 using an LD-based approach, we inferred a sex-averaged map of recombination
132 rates and compared this to previously published genetic maps for *M. musculus*. We
133 show that variation in recombination rates in *M. m. castaneus* is very similar to rate
134 variation estimated in the classical inbred strains. This suggests that, at broad
135 scales, recombination rates have been relatively highly conserved since the
136 subspecies began to diverge.

137

138 **Materials and Methods**

139

140 *Polymorphism data for Mus musculus castaneus*

141

142 We analyzed the genomes of 10 wild-caught *M. m. castaneus* individuals
143 sequenced by Halligan *et al.* (2013). Samples were from North-West India, a region
144 that is believed to be within the ancestral range of the house mouse. Mice from this
145 region have among the highest levels of genetic diversity among the *M. musculus*
146 subspecies (Baines and Harr 2007). In addition, the individuals sequenced represent
147 a single population cluster and showed little evidence for substantial inbreeding
148 (Halligan *et al.* 2010). Halligan *et al.* (2013) sequenced individual genomes to high
149 coverage using multiple libraries of Illumina paired-end reads, which were mapped to
150 the mm9 reference genome using BWA (Li and Durbin 2009). Mean coverage was

151 >20x and the proportion of the genome with >10x coverage was more than 80% for
152 all individuals sampled (Halligan *et al.* 2013). Variants were called with the Samtools
153 *mpileup* function (Li *et al.* 2009) using an allele frequency spectrum (AFS) prior. The
154 AFS was obtained by iteratively calling variants until the spectrum converged. After
155 the first iteration, all SNPs at frequencies >0.5 were swapped into the mm9 genome
156 to construct a reference genome for *M. m. castaneus*, which was used for
157 subsequent variant calling (for further details see Halligan *et al.* 2013). The variant
158 call format files generated by Halligan *et al.* (2013) were used in this study. In
159 addition, alignments of *Mus famulus* and *Rattus norvegicus* to the mm9 genome,
160 also generated by Halligan *et al.* (2013), were used as outgroups.

161

162 For the purposes of estimating recombination rates, variable sites were
163 filtered on the basis of several conditions: Insertion/deletion polymorphisms were
164 excluded, because the method used to phase variants (see *below*) cannot process
165 these sites. We also excluded sites with more than two alleles and those that failed
166 the Samtools Hardy-Weinberg equilibrium test ($p < 0.002$).

167

168 *Inferring phase and estimating switch error rates*

169

170 LDhelmet estimates recombination rates from a sample of phased
171 chromosomes or haplotypes drawn from a population. To estimate haplotypes,
172 heterozygous SNPs called in *M. m. castaneus* were phased using read-aware
173 phasing in Shapelt2 (Delaneau *et al.* 2013). Shapelt2 uses sequencing reads that
174 span multiple heterozygous variants, phase-informative reads (PIRs), and LD to
175 phase variants at the level of whole chromosomes. Incorrectly phased heterozygous

176 SNPs, termed switch errors, may upwardly bias estimates of the recombination rate,
177 because they appear identical to legitimate crossing over events. To assess the
178 impact of incorrect phasing on our recombination rate inferences, we quantified the
179 switch error rate as follows. The population sample of *M. m. castaneus* comprised of
180 seven females and three males. The X-chromosome variants in males therefore
181 represent perfectly phased haplotypes. We merged the BAM alignments of short
182 reads for the X-chromosome of the three males (samples H12, H28 and H34 from
183 Halligan *et al.* (2013)) to make three datasets of pseudo-females, which are female-
184 like, but in which the true haplotypes are known (H12+H28 = H40; H12+H34 = H46;
185 H28 + H34 = H62). We then jointly re-called variants in the seven female samples
186 plus the three pseudo-females using an identical pipeline as used by Halligan *et al.*
187 (2013), as outlined above, using the same AFS prior.

188

189 Switch error rates in Shapeit2 are sensitive both to coverage and quality (per
190 genotype and per variant) (Delaneau *et al.* 2013). We explored the effects of
191 different filter parameters on the switch error rates produced by Shapeit2 using the
192 X-chromosomes of the pseudo-females. We filtered SNPs based on combinations of
193 variant and genotype quality scores (QUAL and GQ, respectively) and on an
194 individual's sequencing depth (DP) (Table S1). For the individual-specific statistics
195 (DP and GQ), if a single individual failed a particular filter, then that SNP was not
196 included in further analyses. By comparing the known X-chromosome haplotypes
197 and those inferred by Shapeit2, we calculated switch error rates as the ratio of
198 incorrectly resolved heterozygous SNPs to the total number of heterozygous SNPs
199 for each pseudo-female individual. We used these results to choose filter parameters
200 to apply to the autosomal data that generated a low switch error rate in Shapeit2,

201 while maintaining a high number of heterozygous SNPs. We obtained 20 phased
202 haplotypes for each of the 19 mouse autosomes. With these, we estimated the
203 recombination rate landscape for *M. m. castaneus*.

204

205 Estimating recombination maps and validation of the approach

206

207 LDhelmet (v1.7; Chan *et al.* 2012) generates a sex-averaged map of
208 recombination rates from a sample of haplotypes that are assumed to be drawn from
209 a randomly mating population. Briefly, LDhelmet examines patterns of LD in a
210 sample of phased chromosomal regions and uses a composite likelihood approach
211 to infer recombination rates that are best supported between adjacent SNPs.

212 LDhelmet appears to perform well for species with large effective population size
213 (N_e) and has been shown to be robust to the effects of selective sweeps, which may
214 be prevalent and reduce diversity in and around functional elements of the *M. m.*
215 *castaneus* genome (Halligan *et al.* 2013). However, the analyses conducted by Chan
216 *et al.* (2012), in which the software was tested, were performed with a larger number
217 of haplotypes than we have in our sample. To assess whether our smaller sample
218 size gives reliable recombination maps, we validated and parameterized LDhelmet
219 using simulated datasets.

220

221 A key parameter in LDhelmet is the block penalty, which determines the
222 extent by which likelihood is penalized by spatial variation in the recombination rate,
223 such that a high block penalty results in a smoother recombination map. We
224 performed simulations to determine the block penalty that leads to the most accurate
225 estimates of the recombination rate in chromosomes that have levels of diversity and

226 base content similar to *M. m. castaneus*. Chromosomes with constant values of ρ
227 ($4N_e r$) ranging from 2×10^{-6} to 2×10^1 were simulated in SLiM v1.8 (Messer 2013).
228 For each value of ρ , 0.5Mbp of neutrally evolving sequence was simulated for
229 populations of $N = 1,000$ diploid individuals. Mutation rates in the simulations were
230 set using the compound parameter $\theta = 4N_e \mu$, where μ is the per-base, per-
231 generation mutation rate. The mutation and recombination rates of the simulations
232 were scaled to $\theta/4N$ and $\rho/4N$, respectively. θ was set to 0.01 for all simulations, as
233 this is close to the genome-wide average for our data, based on pairwise differences.
234 Simulations were run for 10,000 generations to achieve equilibrium levels of
235 polymorphism, at which time 10 diploid individuals were sampled from the
236 population. Each simulation was repeated 20 times, resulting in 10Mbp of sequence
237 for each value of ρ . The SLiM output files were converted to sequence data, suitable
238 for analysis by LDhelmet, using a custom Python script that incorporated the
239 mutation rate matrix estimated for non-CpG prone sites in *M. m. castaneus* (see
240 below). We inferred recombination rates from the simulated data in windows of 4,400
241 SNPs with a 200 SNP overlap between windows, following (Chan *et al.* 2012). We
242 analyzed the simulated data using LDhelmet with block penalties of 10, 25, 50 and
243 100. The default parameters of LDhelmet are tuned to analyze *Drosophila*
244 *melanogaster* data (Chan *et al.* 2012). Since the *D. melanogaster* population studied
245 by Chan *et al.* (2012) has comparable levels of genetic diversity to *M. m. castaneus*
246 we used the defaults for all other parameters, other than the block penalty and
247 estimate of θ .

248

249 Errors in phase inference, discussed above, may bias our estimates of the
250 recombination rate, since they appear to break apart patterns of LD. We assessed

251 the impact of these errors on recombination rate inference by incorporating them into
252 the simulated data at a rate estimated from the pseudo-female individuals. For each
253 of the 10 individuals drawn from the simulated populations, switch errors were
254 randomly introduced at heterozygous positions at the rate estimated using the
255 chosen SNP filter set (*see Results*). We then inferred the recombination rates, as
256 above, for the simulated population using these error-prone data. We assessed the
257 effect of switch errors on recombination rate inference by comparing estimates
258 based on the simulated data both with and without switch errors. It is worth noting
259 that there is the potential for switch errors to undo crossing-over events, reducing
260 inferred recombination rates, if they affect heterozygous SNPs that are breakpoints
261 of recombinant regions.

262

263 Recombination rate estimation for *M. m. castaneus*

264

265 We used LDhelmet (Chan *et al.* 2012), to estimate recombination rates for each
266 of the *M. m. castaneus* autosomes. It is well established that autosomal
267 recombination rates differ between the sexes in *M. musculus* (Cox *et al.* 2009; Liu *et*
268 *al.* 2014). A drawback of LD-based approaches is that they give sex-averaged
269 recombination rates.

270

271 We used both *M. famulus* and *R. norvegicus* as outgroups to assign ancestral
272 alleles to polymorphic sites. LDhelmet incorporates both the mutation matrix and a
273 prior probability on the ancestral allele at each variable position as parameters in the
274 model. We obtained these parameters as follows. For non-CpG prone polymorphic
275 sites, if the outgroups shared the same allele, we assigned that allele as ancestral

276 and these sites were then used to populate the mutation matrix, following Chan *et al.*
277 (2012). This approach ignores the possibility of both back mutation and homoplasy.
278 To account for this uncertainty, LDhelmet incorporates a prior probability on the
279 ancestral base. Following Singhal *et al.* (2015), at resolvable sites (i.e. when both
280 outgroups agreed), the ancestral base was given a prior probability of 0.91, with 0.03
281 assigned to each of the three remaining bases. This was done to provide high
282 confidence in the ancestral allele, but to also include the possibility of ancestral allele
283 misinference. At unresolved sites (i.e., if the outgroup alleles did not agree or there
284 were alignment gaps in either outgroup), we used the stationary distribution of allele
285 frequencies from the mutation rate matrix as the prior (Table S2).

286

287 We analyzed a total of 43,366,235 SNPs in LDhelmet to construct
288 recombination maps for each of the *M. m. castaneus* autosomes. Following Chan *et*
289 *al.* (2012), windows of 4,400 SNPs, overlapping by 200 SNPs on either side, were
290 analysed. We ran LDhelmet with a block penalty of 100 for a total of 1,000,000
291 iterations, discarding the first 100,000 as burn-in. The block penalty value was
292 chosen to obtain a conservatively estimated recombination map, on the basis of the
293 simulation analysis. We analyzed all sites that passed the filters chosen using the
294 pseudo-female phasing regardless of CpG status; note that excluding CpG-prone
295 sites removes ~50% of the available data and thus would substantially reduce the
296 power to infer recombination rates. We assumed $\theta = 0.01$, the approximate genome-
297 wide level of neutral diversity in *M. m. castaneus*, and included ancestral allele priors
298 and the mutation rate matrix for non-CpG sites as parameters in the model.
299 Following the analyses, we removed overlapping SNPs and concatenated SNP
300 windows to obtain recombination maps for whole chromosomes.

301

302 It is worthwhile noting that our map was constructed with genotype calls made
303 using the mm9 version of the mouse reference genome. This version was released
304 in 2007 and there have been subsequent versions released since then. However,
305 previously published genetic maps for *M. musculus* were constructed using mm9, so
306 we used that reference to make comparisons (see below).

307

308 Comparison to previously published maps

309

310 The recombination rate map inferred for *M. m. castaneus* was compared with
311 two previously published genetic maps for *M. musculus*. The first map was
312 generated by analyzing the inheritance patterns of markers in crosses between
313 inbred lines (Cox *et al.* 2009)(downloaded from
314 <http://cgd.jax.org/mousemapconverter/>). Hereafter, this map shall be referred to as
315 the Cox map. The second map was generated by Brunschwig *et al.* (2012), by
316 analyzing SNPs in a large number of inbred mouse lines using LDhat (Auton and
317 Mcvean 2007), the software upon which LDhelmet is based (available at
318 <http://www.genetics.org/content/early/2012/05/04/genetics.112.141036>). Hereafter,
319 this map shall be referred to as the Brunschwig map. Both maps were generated
320 using classical strains of laboratory mice, which are predominantly of *M. m.*
321 *domesticus* origin (Yang *et al.* 2011). Both the Brunschwig and Cox maps were
322 constructed using far fewer markers than the present study, ~250,000 and ~10,000
323 SNPs, respectively.

324

325 Recombination rates in the Brunshwig map and our *castaneus* map were
326 inferred in terms of the population recombination rate ($\rho = 4N_e r$), units that are not
327 directly convertible to centimorgans (cM), but were converted to cM/Mb for
328 comparison purposes using frequency weighted means, as follows. Both LDhat and
329 LDhelmet give estimates of ρ (per Kbp and bp, respectively) between pairs of
330 adjacent SNPs. To account for differences in the physical distance between adjacent
331 SNPs when calculating cumulative ρ , we used the number of bases between a pair
332 of SNPs to weight that pair's contribution to the sum. By setting the total map
333 distance for each chromosome to be equal to those found by Cox *et al.* (2009), we
334 scaled the cumulative ρ at each analyzed SNP position to cM values.

335

336 At the level of whole chromosomes, we compared mean recombination rates
337 from the *castaneus* map with several previously published maps. The frequency-
338 weighted mean recombination rates (in terms of ρ) for each of the autosomes from
339 the *castaneus* and Brunshwig maps were compared with the cM/Mb values
340 obtained by Cox *et al.* (2009) as well as independent estimates of the per
341 chromosome recombination rates from Jensen-Seaman *et al.* (2004). Pearson
342 correlations were calculated for each comparison. Population structure in the inbred
343 line data analyzed by Brunshwig *et al.* (2012) may have elevated LD, thus
344 downwardly biasing estimates of ρ . To investigate this, we divided the frequency-
345 weighted mean recombination rates per chromosome from the *castaneus* and
346 Brunshwig maps by the rates given in Cox *et al.* (2009) to obtain estimates of
347 effective population size.

348

349 At a finer scale, we compared variation in recombination rates across the
350 autosomes in the different maps using windows. We calculated Pearson correlations
351 between the frequency weighted-mean recombination rates (in cM/Mb) in non-
352 overlapping windows for the *castaneus*, Cox and Brunschwig maps. The window
353 size considered may affect the correlation between maps, so we calculate Pearson
354 correlations in windows of 1Mbp to 20Mbp in size. For visual comparison of the
355 *castaneus* and Cox maps, we plotted recombination rates in sliding windows of
356 10Mbp, offset by 1Mb.

357

358 *Examining the correlation between nucleotide diversity and recombination rate*

359

360 There is evidence that natural selection is pervasive in the protein-coding
361 genes and conserved non-coding elements in the murid genome (Halligan *et al.*
362 2010; Halligan *et al.* 2011; Halligan *et al.* 2013). Directional selection acting on
363 selected sites within exons may reduce diversity at linked neutral sites through the
364 processes of background selection and/or selective sweeps. These processes have
365 the largest effect in regions of low recombination, and can therefore generate
366 positive correlations between diversity and the recombination rate, as has been
367 observed in multiple species (Cutter and Payseur 2013). We used our *castaneus*
368 map to examine the relationship between nucleotide diversity and recombination
369 rates as follows. We obtained the coordinates of the canonical spliceforms of protein
370 coding genes, orthologous between mouse and rat from Ensembl Biomart (Ensembl
371 Database 67; <http://www.ensembl.org/info/website/archives/index.html>). We
372 calculated the frequency-weighted mean recombination rate and the GC content for
373 each gene. Using the approximate *castaneus* reference, described above, and the

374 outgroup alignment, we obtained the locations of 4-fold degenerate synonymous
375 sites. If a site was annotated as 4-fold in all three species considered, it was used for
376 further analysis. We removed poor quality alignments between mouse and rat,
377 exhibiting a spurious excess of diverged sites, where >80% of sites were missing.
378 We also excluded five genes that were diverged at all non-CpG prone 4-fold sites, as
379 it is likely that these also represent incorrect alignments. After filtering, there were a
380 total of 18,171 protein-coding genes for analysis.

381

382 We examined the correlation between local recombination rates in protein
383 coding genes with nucleotide diversity and divergence. Variation in the mutation rate
384 across the genome may influence genome-wide analyses of nucleotide
385 polymorphism, so we also examined the correlation between the ratio of nucleotide
386 diversity and divergence from *R. norvegicus* at neutral sites and the rate of
387 recombination. We used non-parametric Kendall rank correlations for all
388 comparisons.

389

390 All analyses were conducted using custom Python scripts, except correlation
391 analyses which were conducted using R (R Core Team 2016).

392

393 **Results**

394

395 *Phasing SNPs and estimating the switch error rate*

396

397 In order to infer recombination rates from our sample of individuals, we
398 required phased SNPs. Taking advantage of the high sequencing depth of the

399 sample, we phased SNPs using Shapelt2, an approach that makes use of both LD
400 and sequencing reads to resolve haplotypes. We phased each of the mouse
401 autosomes, giving a total of 43,366,235 SNPs for estimation of recombination rates.

402

403 By constructing pseudo-female individuals, we quantified the switch error rate
404 incurred when inferring phase from our data. After filtering of variants, Shapelt2
405 achieved low switch error rates for all parameter combinations tested (Table S1). We
406 chose a set of filters (GQ > 15, QUAL > 30) that resulted in a mean switch error rates
407 across the three pseudo-females of 0.46% (Table S1). More stringent filtering
408 resulted in slightly lower mean switch error rates, but also resulted in the removal of
409 many more variants from the dataset (Table S1), thus reducing power to resolve
410 recombination rates in downstream analyses.

411

412 Simulations to validate LDhelmet for the population sample of *M. m. castaneus*

413

414 We assessed the performance of LDhelmet when applied to our dataset by
415 simulation. In the absence of switch errors, LDhelmet accurately infers the average
416 recombination rate down to values of $\rho/\text{bp} = 2 \times 10^{-4}$ (Figure 1). Below this value,
417 LDhelmet overestimated the scaled recombination rate for the simulated populations
418 (Figure 1). With switch errors incorporated into simulated data, LDhelmet accurately
419 estimated ρ/bp in the range 2×10^{-3} to 2×10^2 . When the true ρ/bp was $< 2 \times 10^{-3}$,
420 however, LDhelmet overestimated the mean recombination rate for 0.5Mbp regions
421 (Figure 1). This behavior was consistent for all block penalties tested (Figure S1).
422 Given that the simulations incorporated the mutation rate matrix (Table S2) and
423 mutation rate ($\theta = 4N_e\mu$) estimated for *M. m. castaneus* we concluded that LDhelmet

424 is applicable to the dataset of 10 *M. m. castaneus* individuals sequenced by Halligan
425 *et al.* (2013).

426

427 Recombination rates across the *M. m. castaneus* autosomes

428

429 A recombination rate map for each *M. m. castaneus* autosome was
430 constructed using LDhelmet. We analyzed a total of 43,366,235 phased SNPs
431 across the 19 mouse autosomes. The frequency weighted mean value of p/bp for all
432 autosomes was 0.009. This value is greater than the lower detection limit suggested
433 by both the simulations with and without switch errors (Figure 1).

434

435 We assessed variation in whole-chromosome recombination rates between
436 our LD-based *castaneus* map and direct estimates of recombination rates published
437 in earlier studies. Comparing the mean recombination rates for whole chromosomes
438 provides us with a baseline comparison for which we have an *a priori* expectation:
439 we expect that chromosome 19, the shortest in physical length, should have the
440 highest mean recombination rate, since at least one crossing over event is required
441 per meiosis per chromosome in mice. This has been demonstrated in previous
442 studies of recombination in *M. musculus* (Jensen-Seaman *et al.* 2004; Cox *et al.*
443 2009). Indeed, we find that the frequency-weighted mean recombination rate for
444 chromosome 19 is the highest among the autosomes (Table 1). We also found that
445 the frequency-weighted mean recombination rates for each of the autosomes were
446 highly correlated with the direct estimates given in Jensen-Seaman *et al.* (2004)
447 (Pearson correlation = 0.66, $p = 0.002$) and Cox *et al.* (2009) (Pearson correlation =
448 0.88, $p < 0.0001$), suggesting that our analysis captures real variation in

449 recombination rates at the scale of whole chromosomes in the *M. m. castaneus*
450 genome.

451

452 Comparison of the *M. m. castaneus* map to maps constructed using inbred lines

453

454 We compared the intra-chromosomal variation in recombination rates
455 between our *castaneus* map and previously published maps. Figure 2 shows the
456 variation in recombination rates across the largest and smallest autosomes in the
457 mouse genome, chromosomes 1 and 19, respectively. It is clear that the *castaneus*
458 and Cox maps are very similar (see also Figure S2 showing a comparison of all
459 autosomes). Correlation coefficients between the maps are >0.8 for window sizes of
460 8Mbp and above (Figure 3). Although the overall correlation between the *castaneus*
461 and Cox maps is high (Figure 3), there were several regions of the genome that
462 substantially differ, for example in the centre of chromosome 9 (Figure S2). The Cox
463 and *castaneus* maps are more similar to one another than either are to the
464 Brunschwig map (Figure 3), presumably because the Brunschwig map was
465 constructed with a sample of 60 inbred mouse strains. Population structure in the
466 lines or the subspecies from which they were derived would elevate LD, resulting in
467 downwardly-biased chromosome-wide values of ρ . This is also reflected in the N_e
468 values estimated from the frequency-weighted average recombination rates for each
469 chromosome. The estimates of N_e are substantially different between the *castaneus*
470 and Brunschwig maps, i.e. the *castaneus* estimates are consistently $\sim 500x$ higher
471 (Table 1). The estimates of N_e from the *castaneus* map are in broad agreement with
472 the estimates of N_e based on polymorphism data (Geraldes *et al.* 2008).

473

474 Correlations between recombination rate and properties of protein coding genes in

475 *M. m. castaneus*

476

477 By examining the correlation between genetic diversity and recombination
478 rate, we determined whether our map captures variation in N_e across the genome.
479 We found that recombination rates at protein coding genes are significantly and
480 positively correlated with levels of neutral genetic diversity (Table 2), at all sites
481 regardless of base context and at non-CpG-prone sites only (Table 2). Divergence
482 from the rat at 4-fold sites was also significantly and positively correlated with
483 recombination rate when analyzing all sites. However, for non-CpG-prone sites we
484 found a small negative correlation (Table 2). There was also a significant and
485 positive relationship between recombination rate and a gene's GC content ($r =$
486 0.125 , $p < 2.2 \times 10^{-16}$). The correlation between recombination rate and neutral
487 diversity divided by divergence from the rat was both positive and significant,
488 regardless of base context (Table 2; Figure S3). This indicates that natural selection
489 may have a role in reducing diversity via hitchhiking and/or background selection.

490

491 **Discussion**

492

493 By constructing fine-scale maps of the recombination rate for the autosomes
494 of *M. m. castaneus*, we have shown that there is a high degree of similarity between
495 the recombination landscape for wild-caught mice and their laboratory counterparts.
496 Our map captures variation in the recombination rate, similar to that observed in a
497 more traditional linkage map, at the level of both whole chromosomes and genomic
498 windows of varying size.

499

500 Recombination landscapes inferred using coalescent approaches, as in this
501 study, reflect ancestral variation in recombination rates. We show that this ancestral
502 variation is highly correlated with contemporaneous recombination rates in inbred
503 mice of a different subspecies, suggesting that the broad-scale variation in
504 recombination rate has not evolved dramatically since the subspecies diverged
505 around 500,000 years ago (Gerald *et al.* 2008). At a finer scale, however,
506 Smagulova *et al.* (2016) showed that there is considerable variation in the locations
507 of recombination hotspots between the *M. musculus* subspecies. Their findings,
508 taken together with ours, parallel results in hominids and the great-apes, which
509 suggest that, although the locations of recombination hotspots are strongly diverged
510 between species, broad-scale patterns of recombination rate are relatively
511 conserved (Leseque *et al.* 2014; Stevison *et al.* 2015). However, there do seem to
512 be multiple regions of the genome that distinguish *M. m. castaneus* and *M. m.*
513 *domesticus*. For example, we observe peaks in recombination rate for *M. m.*
514 *castaneus* on chromosomes 4, 5, 14 and 15 that are not present in the Cox map
515 (Figure S2). These results are seemingly consistent with those of Dumont *et al.*
516 (2011), who found that there are significant differences in genetic length between *M.*
517 *m. castaneus* and *M. m. musculus* (when crossed to *M. m. domesticus*) in multiple
518 regions of the genome (though a large proportion of the differences they detected
519 were on the X-chromosome, which was not analyzed in our study). Performing a
520 comparative analysis of recombination rates in the different subspecies of house
521 mice, as well as sister species, using LD-based methods would help elucidate the
522 time-scale of recombination rate evolution in wild mice.

523

524 The *castaneus* map constructed in this study appears to be more similar to
525 the Cox map than the Brunshwig map (Figure 3). There are number of potential
526 reasons for this. Firstly, we used a much larger number of markers to resolve
527 recombination rates than Brunshwig *et al.* (2012), giving us more power to capture
528 variation in the recombination rate. Secondly, it seems probable that population
529 structure within and between the inbred and wild-derived lines studied by
530 Brunshwig *et al.* (2012) could have resulted in biased estimates of the
531 recombination rate. By dividing the mean estimated ρ /bp values (inferred using
532 LDhelmet) for each chromosome by the corresponding recombination rate estimated
533 from crosses (Cox *et al.* 2009), we showed that N_e estimates from the Brunshwig
534 map are much lower than estimates based on our map (Table 1). This is consistent
535 with the presence of elevated LD between the SNPs in the inbred lines analyzed by
536 Brunshwig *et al.* (2012). It should be noted, however, that the estimates of N_e will
537 be biased, as $\theta = 4N_e\mu$ is a parameter in both LDhat and LDhelmet. In spite of this
538 potential bias, the differences in N_e estimated from the Brunshwig and *castaneus*
539 maps shown in Table 1 are striking, given that the ancestral effective population
540 sizes of *M. m. domesticus* and *M. m. castaneus* are expected to be ~150,000 and
541 ~350,000, respectively (Gerald *et al.* 2008). The Brunshwig map does, however,
542 capture true variation in recombination rates, because their map is also highly
543 correlated with the Cox map (Pearson correlation >0.6) for all genomic windows
544 wider than 8Mbp (Figure 3). Indeed, Brunshwig *et al.* (2012) showed by simulation
545 that hotspots are detectable by analysis of inbred lines and validated their inferred
546 hotspots against the locations of those observed in crosses among classical strains
547 of *M. m. domesticus* (Smagulova *et al.* 2011). This suggests, that while estimates of
548 the recombination rate in the Brunshwig *et al.* (2012) map may have been

549 downwardly biased by population structure, variation in the rate and locations of
550 hotspots were still accurately detected in their study.

551

552 We obtained an estimate of the switch error rate, taking advantage of the
553 hemizygous sex chromosomes of males present in our sample. This allowed us to
554 assess the extent by which switch errors affected our ability to infer recombination
555 rates in *M. m. castaneus*. It should be noted, however, that our inferred switch error
556 rate may not fully represent that of the autosomes. This is because multiple factors
557 influence the ability to phase variants using Shapelt2 (i.e. LD, SNP density, sample
558 size, depth of coverage and read length) and some of these factors differ between
559 the X-chromosome and the autosomes. Firstly, as the sex-averaged recombination
560 rate for the X-chromosome is expected to be 3/4 that of the autosomes, it likely has
561 elevated LD, and thus there will be higher power to infer phase. In contrast, the level
562 of X-linked nucleotide diversity in *M. m. castaneus* is approximately one half that of
563 the autosomes (Kousathanas *et al.* 2014), and thus there would be a higher
564 probability of phase informative reads on the autosomes. While it is difficult to assess
565 whether the switch error rates we estimated from the X-chromosome analysis will be
566 the same as on the autosomes, the analysis allowed us to explore the effects of
567 different SNP filters on the error rate.

568

569 By simulating the effect of switch errors on estimates of the recombination
570 rate, we inferred the range over which ρ /bp is accurately estimated in our data.
571 Switch errors appear identical to legitimate crossing-over events and, if they are
572 randomly distributed along chromosomes, a specific rate of error will resemble a
573 constant rate of crossing over. The rate of switch error will then determine a

574 detection threshold below which recombination cannot be accurately inferred. We
575 introduced switch errors at random into the simulation data and estimates of ρ /bp
576 obtained from these datasets reflect this detection threshold; below $2 \times 10^{-3} \rho$ /bp, we
577 found that LDhelmet consistently overestimates the recombination rate in the
578 presence of switch errors (Figure 1; Figure S1). This highlights a possible source of
579 bias affecting LD-based recombination mapping studies using inferred haplotypes. In
580 a recent study, Singhal *et al.* (2015) showed that the power to detect recombination
581 hotspots is reduced when the recombination rate in the regions surrounding a
582 hotspot is low. Though we did not attempt to locate recombination hotspots in this
583 study, our findings and those of Singhal *et al.* (2015) both suggest that error in phase
584 inference needs to be carefully considered before attempting to estimate
585 recombination rates and/or recombination hotspots using LD-based approaches.

586

587 Consistent with studies in a variety of organisms, we found a positive
588 correlation between genetic diversity at putatively neutral sites and the rate of
589 recombination. Both unscaled nucleotide diversity and diversity divided by
590 divergence between mouse and rat, a proxy for the mutation rate, are positively
591 correlated with recombination (Table 2). Cai *et al.* (2009) found evidence suggesting
592 that recombination may be mutagenic, though insufficient to account for the
593 correlations they observed between recombination and diversity. The Kendall
594 correlation between π/d_{rat} and recombination rate of 0.20 for all 4-fold sites, a value
595 that is similar in magnitude to the corresponding value of 0.09 reported by Cai *et al.*
596 (2009) in humans. The correlations we report may be downwardly biased, however,
597 because switch errors may result in inflated recombination rates inferred for regions
598 of the genome where the true recombination rate is low (see above). Genes that

599 have recombination rates lower than the detection limit set by the switch error rate
600 may be reported as having inflated ρ/bp (Figure 1; Figure S1), and this would have
601 the effect of reducing correlation statistics. It is difficult to assess the extent of this
602 bias, however, and in any case the correlations we observed between diversity and
603 recombination suggest that our recombination map does indeed capture real
604 variation in N_e across the genome. This indicates that a recombination mediated
605 process influences levels of genetic diversity. Previously, Halligan *et al.* (2013)
606 showed that there are troughs in nucleotide diversity surrounding protein coding
607 exons in *M. m. castaneus*, characteristic of natural selection acting within exons
608 reducing diversity at linked sites. Their results and ours suggest pervasive natural
609 selection in the genome of *M. m. castaneus*.

610

611 In conclusion, we find that sex-averaged estimates of the ancestral
612 recombination landscape for *M. m. castaneus* are highly correlated with
613 contemporary estimates of the recombination rate estimated from crosses of *M. m.*
614 *domesticus* (Cox *et al.* 2009). It has been demonstrated previously that the turnover
615 of hotspots has led to rapid evolution of fine-scale rates of recombination in the *M.*
616 *musculus* subspecies complex (Smagulova *et al.* 2016). On a broad scale, however,
617 our results suggest that the recombination landscape is very strongly conserved
618 between the subspecies. In addition, our estimate of the switch-error rate implies that
619 phasing errors leads to upwardly biased estimates of the recombination rate when
620 the true recombination rate is low. This is a source of bias that should be assessed
621 in future studies. Finally, we showed that the variation in recombination rate is
622 positively correlated with genetic diversity, suggesting that natural selection reduces

623 diversity at linked sites across the *M. m. castaneus* genome, consistent with the
624 findings of Halligan et al (2013).

625

626 To further our understanding of the evolution of the rate of recombination in
627 the house mouse we need to directly compare subspecies. The comparison of our
628 results and previously published maps indicates that there is broad-scale agreement
629 in recombination rates between *M. m. castaneus* and *M. m. domesticus*. In this
630 study, we have assumed that inbred lines derived from *M. m. domesticus* reflect
631 natural variation in recombination rates in that sub-species, though this is not
632 necessarily the case. Population samples like the one studied here could be used to
633 more clearly elucidate the recombination rate maps specific to the different
634 subspecies. A broad survey of this kind would most efficiently be generated using
635 LD-based approaches.

636

637 **Acknowledgements**

638

639 We are grateful to Bettina Harr, Dan Halligan, Ben Jackson and Rory Craig for
640 discussions and helpful comments on the manuscript. Tom Booker is supported by a
641 BBSRC EASTBIO studentship. This project has received funding from the European
642 Research Council (ERC) under the European Union's Horizon 2020 research and
643 innovation programme (grant agreement No. 694212). Rob Ness was funded by the
644 BBSRC (BB/L00237X/1).

645

646

647

648 **Literature Cited**

- 649 Auton, A., and G. McVean, 2007 Recombination rate estimation in the presence of
650 hotspots. *Genome Res* 17: 1219-1227.
- 651 Baines, J. F., and B. Harr, 2007 Reduced x-linked diversity in derived populations of
652 house mice. *Genetics* 175: 1911-1921.
- 653 Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober *et al.*, 2010 Prdm9 is a major
654 determinant of meiotic recombination hotspots in humans and mice. *Science*
655 327: 836-840.
- 656 Baudat, F., Y. Imai and B. de Massy, 2013 Meiotic recombination in mammals:
657 Localization and regulation. *Nat Rev Genet* 14: 794-806.
- 658 Brick, K., F. Smagulova, P. Khil, R. D. Camerini-Otero and G. V. Petukhova, 2012
659 Genetic recombination is directed away from functional genomic elements in
660 mice. *Nature* 485: 642-645.
- 661 Brunschwig, H., L. Liat, E. Ben-David, R. W. Williams, B. Yakir *et al.*, 2012 Fine-
662 scale maps of recombination rates and hotspots in the mouse genome.
663 *Genetics* 191: 757-764.
- 664 Cai, J. J., J. M. Macpherson, G. Sella and D. A. Petrov, 2009 Pervasive hitchhiking
665 at coding and regulatory sites in humans. *PLoS Genet* 5: e1000336.
- 666 Chan, A. H., P. A. Jenkins and Y. S. Song, 2012 Genome-wide fine-scale
667 recombination rate variation in *Drosophila melanogaster*. *PLoS Genet* 8:
668 e1003090.
- 669 Cox, A., C. L. Ackert-Bicknell, B. L. Dumont, Y. Ding, J. T. Bell *et al.*, 2009 A new
670 standard genetic map for the laboratory mouse. *Genetics* 182: 1335-1344.
- 671 Cutter, A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked
672 sites: Unifying the disparity among species. *Nat Rev Genet* 14: 262-274.

- 673 Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto *et al.*, 2016 Re-
674 engineering the zinc fingers of prdm9 reverses hybrid sterility in mice. *Nature*
675 530: 171-176.
- 676 Delaneau, O., B. Howie, A. J. Cox, J. F. Zagury and J. Marchini, 2013 Haplotype
677 estimation using sequencing reads. *Am J Hum Genet* 93: 687-696.
- 678 Dumont, B. L., M. A. White, B. Steffy, T. Wiltshire and B. A. Payseur, 2011 Extensive
679 recombination rate variation in the house mouse species complex inferred
680 from genetic linkage maps. *Genome Res* 21: 114-125.
- 681 Geraldès, A., P. Basset, B. Gibson, K. L. Smith, B. Harr *et al.*, 2008 Inferring the
682 history of speciation in house mice from autosomal, x-linked, y-linked and
683 mitochondrial genes. *Mol Ecol* 17: 5349-5363.
- 684 Grey, C., P. Barthes, G. Chauveau-Le Friec, F. Langa, F. Baudat *et al.*, 2011 Mouse
685 prdm9 DNA-binding specificity determines sites of histone h3 lysine 4
686 trimethylation for initiation of meiotic recombination. *PLoS Biol* 9: e1001176.
- 687 Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eory *et al.*, 2013
688 Contributions of protein-coding and regulatory change to adaptive molecular
689 evolution in murid rodents. *PLoS Genet* 9: e1003995.
- 690 Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr and P. D. Keightley, 2010
691 Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet* 6:
692 e1000825.
- 693 Halligan, D. L., F. Oliver, J. Guthrie, K. C. Stemshorn, B. Harr *et al.*, 2011 Positive
694 and negative selection in murine ultraconserved noncoding elements. *Mol Biol*
695 *Evol* 28: 2651-2660.
- 696 Hudson, R. R., 2001 Two-locus sampling distributions and their applications.
697 *Genetics* 159: 12.

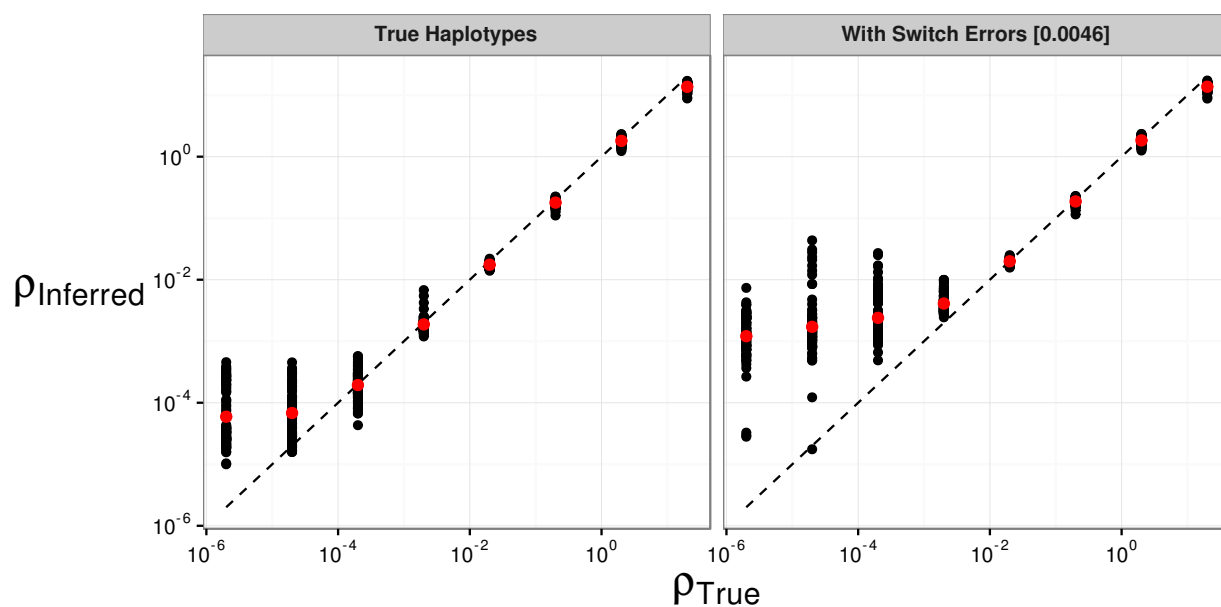
- 698 Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin *et al.*, 2004
699 Comparative recombination rates in the rat, mouse and human genomes.
700 *Genome Res* 14: 528-538.
- 701 Johnston, S. E., C. Berenos, J. Slate and J. M. Pemberton, 2016 Conserved genetic
702 architecture underlying individual recombination rate variation in a wild
703 population of soay sheep (*ovis aries*). *Genetics* 203: 583-598.
- 704 Kousathanas, A., D. L. Halligan and P. D. Keightley, 2014 Faster-x adaptive protein
705 evolution in house mice. *Genetics* 196: 1131-1143.
- 706 Lesecque, Y., S. Glemin, N. Lartillot, D. Mouchiroud and L. Duret, 2014 The red
707 queen model of recombination hotspots evolution in the light of archaic and
708 modern human genomes. *PLoS Genet* 10: e1004790.
- 709 Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with burrows-
710 wheeler transform. *Bioinformatics* 25: 1754-1760.
- 711 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence
712 alignment/map format and samtools. *Bioinformatics* 25: 2078-2079.
- 713 Liu, E. Y., A. P. Morgan, E. J. Chesler, W. Wang, G. A. Churchill *et al.*, 2014 High-
714 resolution sex-specific linkage maps of the mouse reveal polarized distribution
715 of crossovers in male germline. *Genetics* 197: 91-106.
- 716 McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for
717 detecting and estimating recombination from gene sequences. *Genetics* 160:
718 1231-1241.
- 719 McVean, G., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-
720 scale structure of recombination rate variation in the human genome. *Science*
721 304.

- 722 Messer, P. W., 2013 Slim: Simulating evolution with selection and linkage. *Genetics*
723 194: 1037-1039.
- 724 Myers, S. R., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman *et al.*, 2010 Drive
725 against hotspot motifs in primates implicates the *prdm9* gene in meiotic
726 recombination. *Science* 327.
- 727 Paigen, K., and P. Petkov, 2010 Mammalian recombination hot spots: Properties,
728 control and evolution. *Nat Rev Genet* 11: 221-233.
- 729 Paigen, K., J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov *et al.*, 2008 The
730 recombinational anatomy of a mouse chromosome. *PLoS Genet* 4:
731 e1000119.
- 732 RCoreTeam, 2016 R: A language and environment for statistical computing., pp. R
733 Foundation for Statistical Computing, Vienna Austria.
- 734 Schwartz, J. J., D. J. Roach, J. H. Thomas and J. Shendure, 2014 Primate evolution
735 of the recombination regulator *prdm9*. *Nat Commun* 5: 4370.
- 736 Singhal, S., E. Leffler, K. Sannareddy, I. Turner, O. Venn *et al.*, 2015 Stable
737 recombination hotspots in birds. *Science* 350: 6.
- 738 Smagulova, F., K. Brick, P. Yongmei, R. D. Camerini-Otero and G. V. Petukhova,
739 2016 The evolutionary turnover of recombination hotspots contributes to
740 speciation in mice. *Genes & Development* 30: 277-280.
- 741 Smagulova, F., I. V. Gregoret, K. Brick, P. Khil, R. D. Camerini-Otero *et al.*, 2011
742 Genome-wide analysis reveals novel molecular features of mouse
743 recombination hotspots. *Nature* 472: 375-378.
- 744 Smukowski, C. S., and M. A. Noor, 2011 Recombination rate variation in closely
745 related species. *Heredity (Edinb)* 107: 496-508.

- 746 Smukowski Heil, C. S., C. Ellison, M. Dubin and M. A. Noor, 2015 Recombining
747 without hotspots: A comprehensive evolutionary portrait of recombination in
748 two closely related species of drosophila. *Genome Biol Evol* 7: 2829-2842.
- 749 Stevison, L. S., K. B. Hoehn and M. A. Noor, 2011 Effects of inversions on within-
750 and between-species recombination and divergence. *Genome Biol Evol* 3:
751 830-841.
- 752 Stevison, L. S., A. E. Woerner, J. M. Kidd, J. L. Kelley, K. R. Veeramah *et al.*, 2015
753 The time scale of recombination rate evolution in great apes. *Mol Biol Evol*.
- 754 Wang, R. J., M. M. Gray, M. D. Parmenter, K. W. Broman and B. A. Payseur, 2017
755 Recombination rate variation in mice from an isolated island. *Mol Ecol* 26:
756 457-470.
- 757 Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald *et al.*, 2005
758 Comparison of fine-scale recombination rates in humans and chimpanzees.
759 *Science* 308.
- 760 Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific
761 origin and haplotype diversity in the laboratory mouse. *Nat Genet* 43: 648-
762 655.
- 763
- 764
- 765
- 766
- 767
- 768
- 769
- 770

771 **Figures and Tables**

772



773

774 **Figure 1** The effect of switch errors on the mean recombination rate inferred using LDhelmet

775 with a block penalty of 100. Each black point represents results for a window of 4,000 SNPs,

776 with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM

777 for a constant value of ρ /bp. Red points are mean values. Switch errors were randomly

778 incorporated at heterozygous SNPs with probability 0.0046. The dotted line shows the value

779 for inferred=true.

780

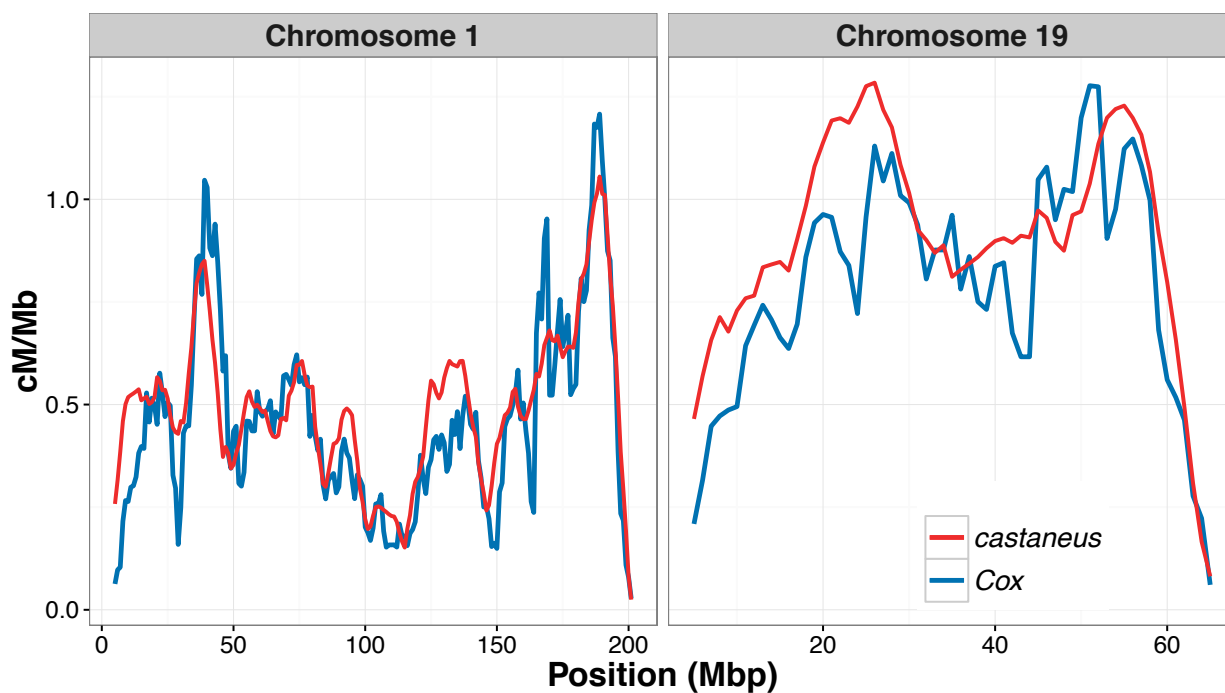
781

782

783

784

785



786

787 **Figure 2** Comparison of the sex-averaged recombination rate inferred for chromosomes 1
788 and 19 of *M. musculus castaneus* using LDhelmet in red and those estimated from the
789 pedigree-based study of Cox *et al.* (2009) in blue. Recombination rates in units of cM/Mb for
790 the *castaneus* map were obtained by setting the total genetic lengths for each chromosome
791 to the corresponding lengths from Cox *et al.* (2009).

792

793

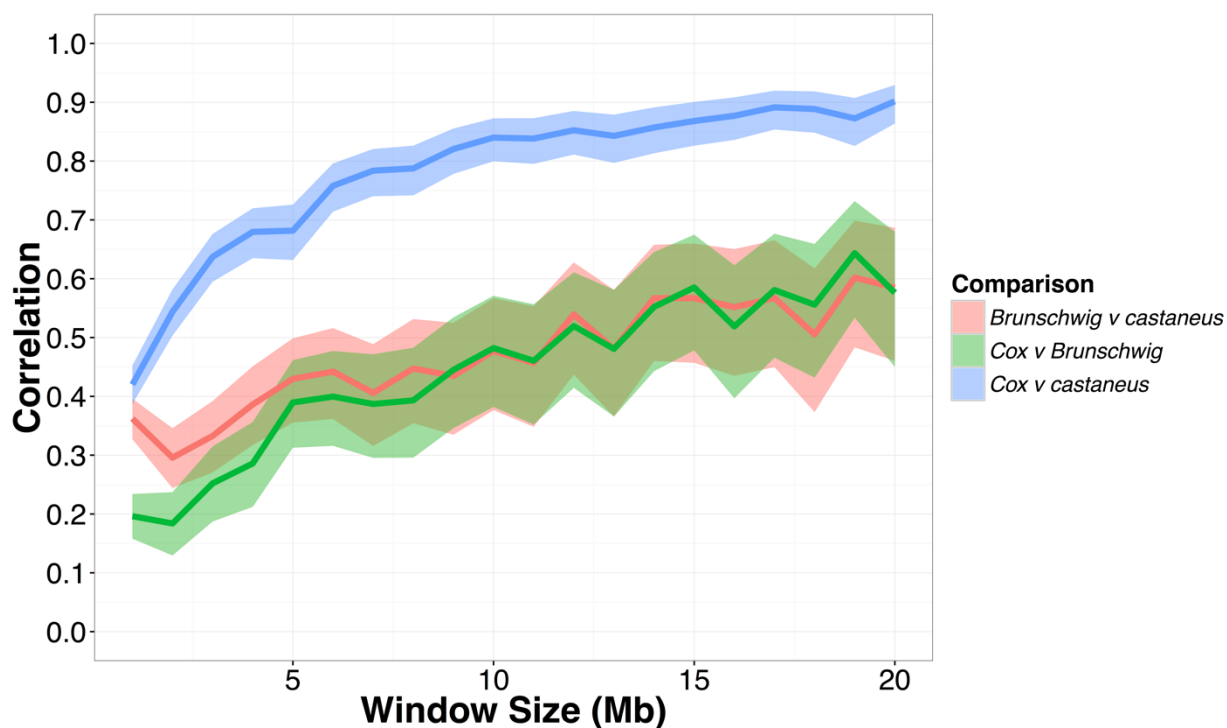
794

795

796

797

798



799

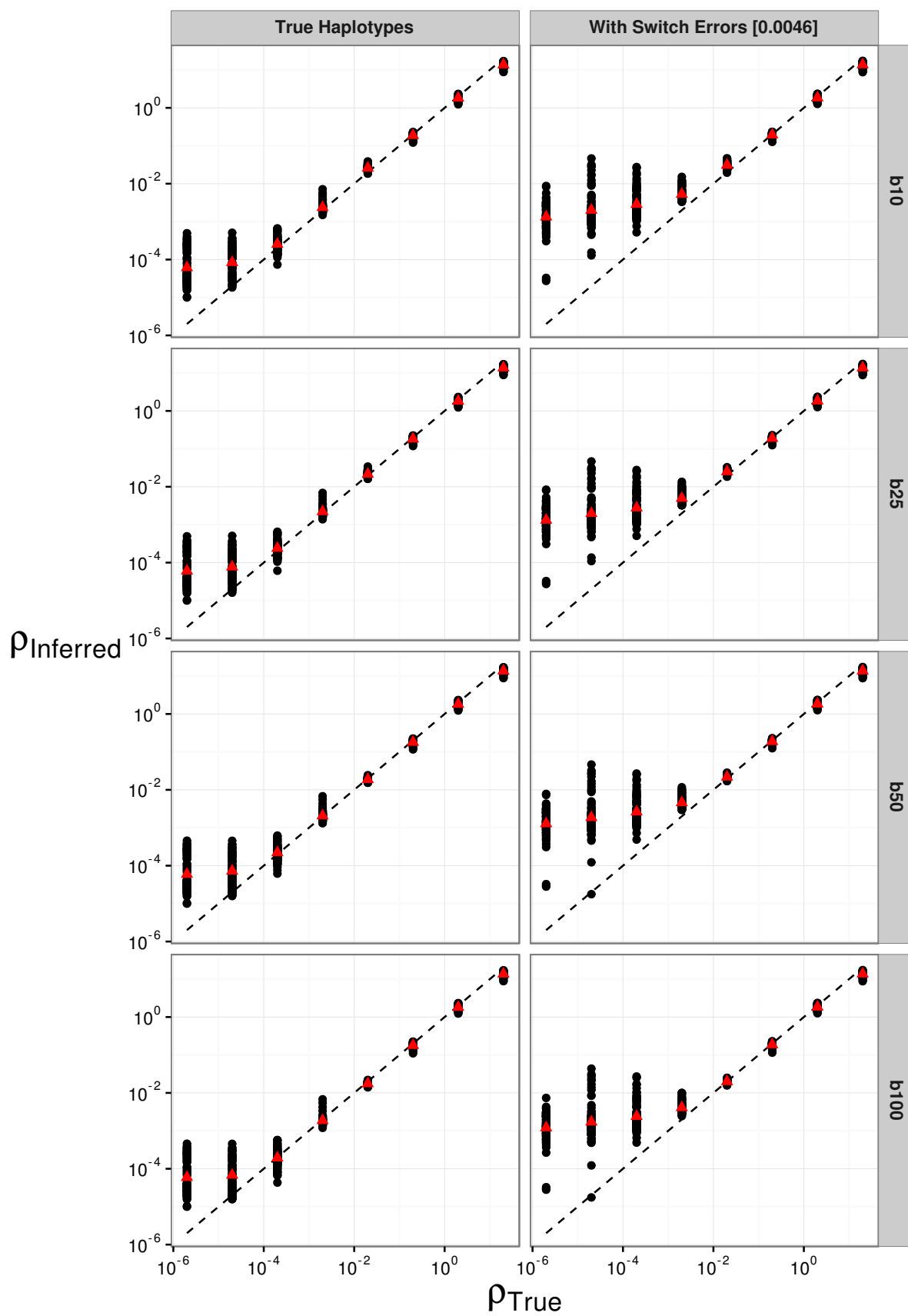
800 **Figure 3** Pearson correlation coefficients between the recombination map inferred for *M. m.*

801 *castaneus*, the Brunshwig *et al.* (2012) map and the Cox *et al.* (2009) map. Correlations

802 were calculated in non-overlapping windows of varying size across all autosomes.

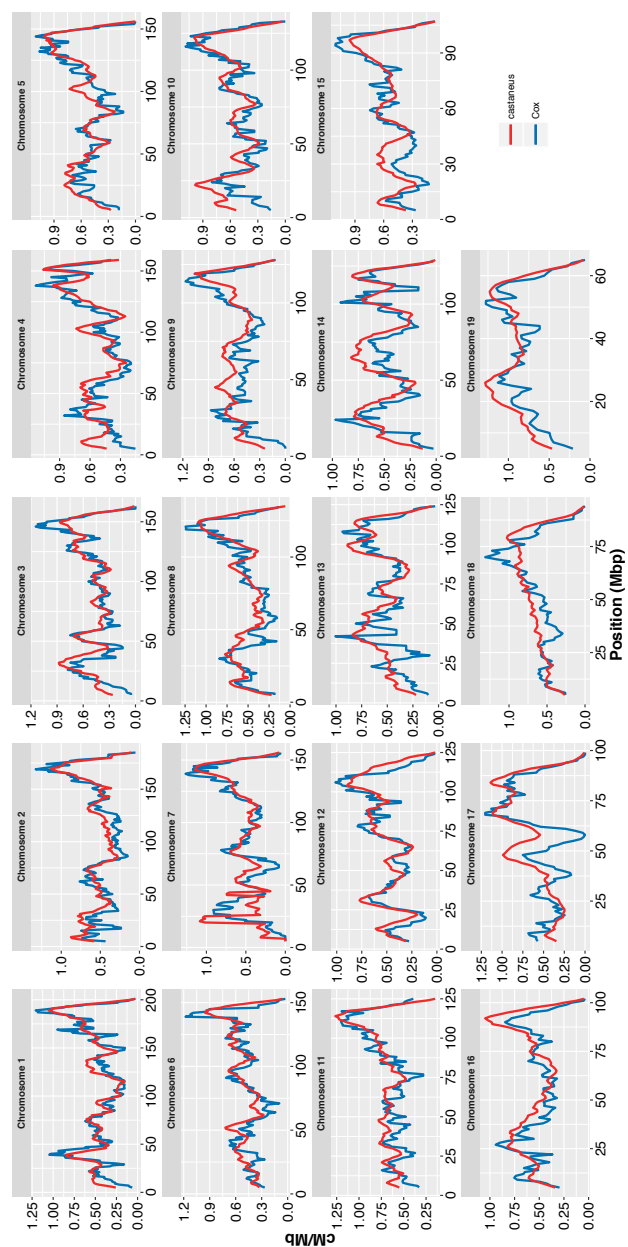
803 Confidence intervals (95%) are indicated by shading around each line.

804



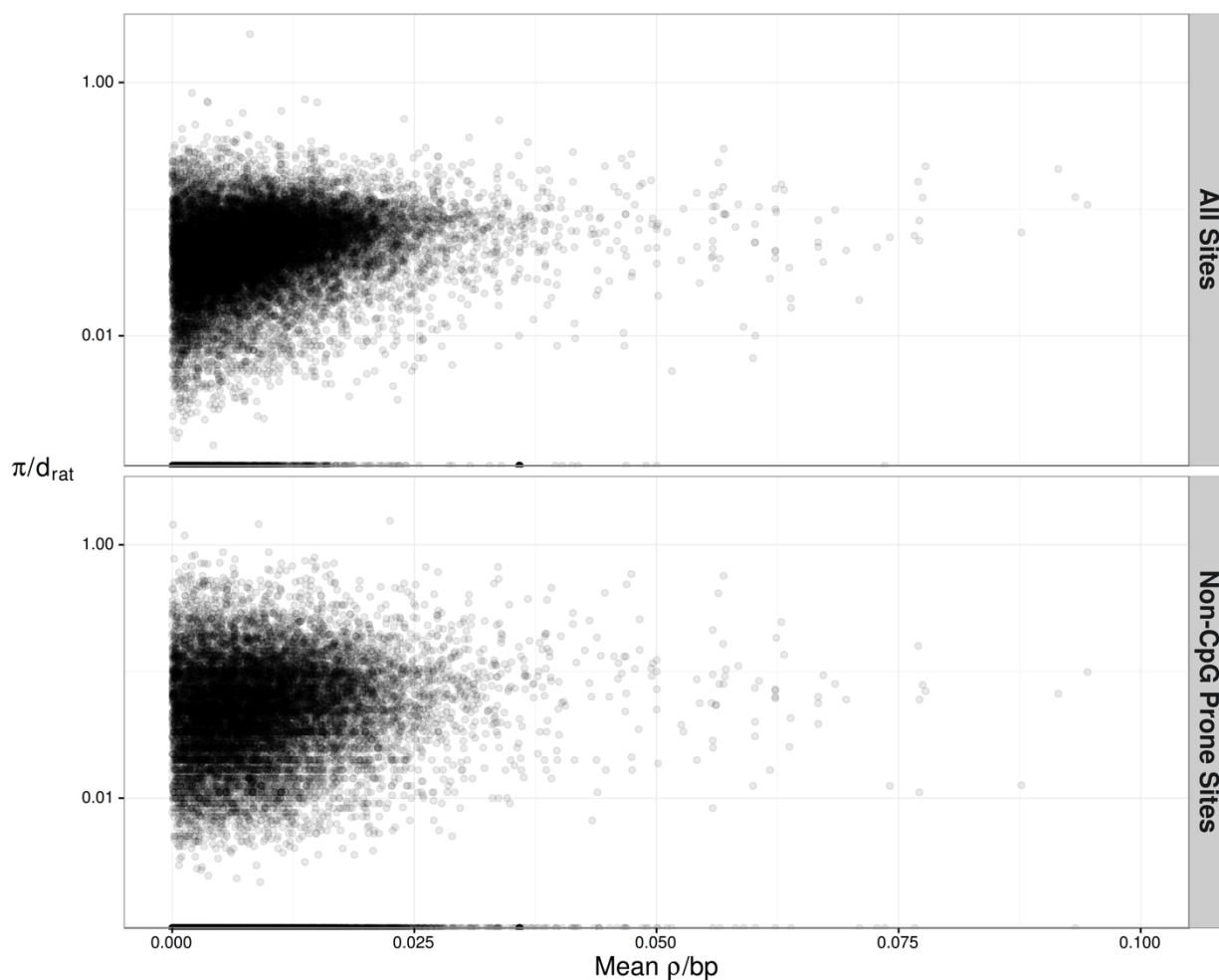
805

806 **Figure S1** The effect of switch errors and block penalty on the mean recombination rate
807 inferred using LDhelmet. Block penalties (b) of 10, 25, 50 and 100 were used, shown in the
808 vertically ordered facets from top to bottom.
809
810
811
812
813
814
815



816

817 **Figure S2** Comparison of recombination rates inferred for *M. m. castaneus* using LDhelmet
818 and recombination rates reported by Cox et al (2009). Recombination rates in units of ρ /bp
819 for the *castaneus* map were converted to cM/Mb by scaling using the genetic length of the
820 corresponding chromosome in the Cox map.



821

822

823 **Figure S3** Nucleotide diversity divided by divergence from rat (π/d_{rat}) at 4-fold degenerate
824 synonymous sites plotted against the frequency weighted mean recombination rates (ρ/bp)
825 for protein coding genes in *M. m. castaneus*. Correlation statistics are shown in the main
826 text. For the purposes of visualization, the range of ρ/bp is restricted from 0 to 0.1. There are
827 19 genes that have mean $\rho/bp > 0.1$, however.

828

829

830

831

832

833

834 **Table 1**

835 Summary of sex-averaged recombination rates estimated for the *M. m castaneus*
 836 autosomes compared with the rates from Brunschwig *et al.* (2012) and Cox *et al.* (2009).
 837 Rates for the *castaneus* and Brunschwig maps are presented in terms of $4N_e r$ /bp. Estimates
 838 of N_e were obtained by assuming the recombination rates from Cox *et al.* (2009).

839

Chromosome	Cox cM/Mb	<i>castaneus</i>		Brunschwig	
		Freq. Weighted Mean	N_e Estimate	Freq. Weighted Mean	N_e Estimate
1	0.50	0.0079	395,000	0.000015	745
2	0.57	0.0088	386,000	0.000015	653
3	0.52	0.0083	400,000	0.000014	693
4	0.56	0.0091	408,000	0.000020	889
5	0.59	0.0090	382,000	0.000015	646
6	0.53	0.0089	421,000	0.000015	728
7	0.58	0.0100	429,000	0.000019	801
8	0.58	0.0094	404,000	0.000014	610
9	0.61	0.0096	394,000	0.000018	749
10	0.61	0.0096	392,000	0.000023	928
11	0.70	0.0102	365,000	0.000019	689
12	0.53	0.0089	420,000	0.000019	897
13	0.56	0.0095	426,000	0.000014	629
14	0.53	0.0084	395,000	0.000013	632
15	0.56	0.0083	371,000	0.000024	1,080
16	0.59	0.0091	386,000	0.000017	721
17	0.65	0.0087	335,000	0.000052	2,020
18	0.66	0.0098	371,000	0.000021	785
19	0.94	0.0122	323,000	0.000026	681
Mean		0.0092		0.000020	

840 **Table 2**

841 Correlation coefficients between recombination rate and pairwise nucleotide diversity and
842 divergence from the rat at 4-fold degenerate sites for protein coding genes. Non-parametric
843 Kendall correlations were calculated for non-CpG prone sites and for all sites, regardless of
844 base context. All coefficients shown are highly significant ($p < 10^{-10}$).

845

	Non-CpG	All Sites
Nucleotide diversity (π)	0.09	0.20
Divergence to rat (d_{rat})	-0.04	0.06
Corrected diversity (π/d_{rat})	0.10	0.18

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864 **Table S1**

865 The effect of different filters on the frequency of switch errors in the haplotypes inferred
 866 based on the three pseudo-females. The values in the switch errors column are the raw
 867 numbers of switch errors and the total number of heterozygous SNPs on the X-chromosome.
 868 Variants with Quality (QUAL) <30 were excluded for all filter sets.

869

Filter Set	HWE [*]	Min DP [†]	Max DP	Min GQ [‡]	Switch Errors			Switch Error Rate
					H40	H46	H62	
1	-	-	-	-	5148 / 409486	4819 / 407422	5020 / 394778	0.0124
2	<0.0002	10	-	15	1690 / 338592	1451 / 334111	1452 / 324199	0.0046
3	<0.0002	10	100	5	2460 / 341744	2066 / 339508	2536 / 328998	0.0070
4	<0.0002	-	-	40	523 / 288471	444 / 286636	550 / 281266	0.0018

870

871 ^{*} HWE refers to the p-value for the Samtools Hardy-Weinberg equilibrium test below which
 872 variants were excluded.

873 [†] Depth of coverage per individual.

874 [‡] Per individual genotype quality scores.

875

876

877

878

879

880

881

882

883 **Table S2**

884 The normalized mutation rate matrix and stationary distribution of base frequencies

885 estimated with two out-groups, *M. famulus* and *R. norvegicus*, using the method described

886 by Chan *et al.* (2012).

	A	C	G	T
A	0.48	0.09	0.32	0.11
C	0.19	0.00	0.12	0.69
G	0.69	0.12	0.00	0.19
T	0.11	0.32	0.08	0.48
Stationary Distribution	0.34	0.16	0.16	0.34

887