

1 **Title:** Recent evolution in *Rattus norvegicus* is shaped by
2 declining effective population size

3 **Running title:** Recent evolution in *R. norvegicus*

4 **Authors:** Eva E. Deinum¹, Daniel L. Halligan¹, Rob W. Ness¹, Yao-Hua
5 Zhang², Lin Cong³, Jian-Xu Zhang², and Peter D. Keightley^{1*}

6 ¹: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh,
7 United Kingdom

8 ²: State Key Laboratory of Integrated Management of Pest Insects and
9 Rodents in Agriculture, Institute of Zoology, Chinese Academy of Sciences,
10 1# Bei-Chen West Road, Beijing 100101, China

11 ³: Institute of Plant Protection, Heilongjiang Academy of Agricultural Sci-
12 ences, Harbin 150086, China

13 *: Corresponding author:

14 `peter.keightley@ed.ac.uk`

15 Institute of Evolutionary Biology

16 University of Edinburgh

17 West Mains Rd

18 Edinburgh EH9 3FL

19 UK

20

21 **Keywords:** *Rattus norvegicus*, evolutionary adaptation, comparative pop-
22 ulation genomics, effective population size, bottleneck, distribution of fitness
23 effects, PSMC, DFE- α , ILLUMINA whole genome sequencing, *Mus muscu-*
24 *lus castaneus*

Abstract

The brown rat, *Rattus norvegicus*, is both a notorious pest and a frequently used model in biomedical research. By analysing genome sequences of 12 wild-caught brown rats from their ancestral range in NE China, along with the sequence of a black rat, *R. rattus*, we investigate the selective and demographic forces shaping variation in the genome. We estimate that the recent effective population size (N_e) of this species = 1.24×10^5 , based on silent site diversity. We compare patterns of diversity in these genomes with patterns in multiple genome sequences of the house mouse (*Mus musculus castaneus*), which has a much larger N_e . This reveals an important role for variation in the strength of genetic drift in mammalian genome evolution. By a Pairwise Sequentially Markovian Coalescent (PSMC) analysis of demographic history, we infer that there has been a recent population size bottleneck in wild rats, which we date to approximately 20,000 years ago. Consistent with this, wild rat populations have experienced an increased flux of mildly deleterious mutations, which segregate at higher frequencies in protein-coding genes and conserved noncoding elements (CNEs). This leads to negative estimates of the rate of adaptive evolution (α) in proteins and CNEs, a result which we discuss in relation to the strongly positive estimates observed in wild house mice. As a consequence of the population bottleneck, wild rats also show a markedly slower decay of linkage disequilibrium with physical distance than wild house mice.

Introduction

Comparative genomics became possible between human and mouse with the publication of the mouse genome (Mouse Genome Sequencing Consortium, 2002), leading to many important new findings, including estimates

1 of the fraction of conserved nucleotide sites, corroboration of downwardly
2 revised estimates of protein-coding gene number, and the discovery of ul-
3 traconserved non-coding elements (Bejerano et al., 2004). Genome sequenc-
4 ing of single individuals naturally led to population genomics, pioneered in
5 *Drosophila simulans* (Begun et al., 2007). This allows detailed inferences
6 to be made concerning many important questions in evolutionary genetics,
7 including the demographic history of populations (Li and Durbin, 2011), the
8 nature and frequency of adaptive evolution (e.g., Hernandez et al., 2011; Sat-
9 tath et al., 2011), and the causes of the correlation between recombination
10 rate and neutral diversity (Cai et al., 2009).

11 Much population genetics has traditionally relied on comparing nucleotide
12 polymorphism in one species to divergence from another. With decreas-
13 ing sequencing costs, comparative population genomics - the comparison of
14 multiple genome sequences from different species - has now become possi-
15 ble. This allows reciprocal analysis of polymorphism *versus* divergence. For
16 example, the McDonald-Kreitman test (McDonald and Kreitman, 1991) for
17 estimating the relative extent of adaptive evolution, uses polymorphism in
18 one species (A) and the divergence between two species (A and B). From the
19 assumptions of the test, including a stable demography since the divergence
20 between the two species, the reciprocal estimate should yield the same re-
21 sult. A significant difference in the estimates could, therefore, indicate that
22 an evolutionary important demographic change has occurred since the split.
23 For such a reciprocal analysis to be biologically meaningful, the species need
24 to be closely related, but not so closely related that they share a substantial
25 fraction of nucleotide polymorphism that originated prior to the split be-
26 tween the species. This condition is comfortably met in the case of our focal
27 species, wild brown rats and mice, which are thought to have diverged at

1 least 12MYA (Benton and Donoghue, 2007), equating to at least 24 million
2 generations.

3 One of the primary determinants of the efficacy of selection across the
4 genome is the strength of genetic drift. The rate of drift is inversely propor-
5 tional to the effective population size (N_e), which represents the size of an
6 ideal population that would display the observed amount of drift. N_e pre-
7 dicts a number of fundamental properties of natural populations, including
8 the amount of genetic variation ($N_e \times$ mutation rate (μ)), the rate at which
9 linkage disequilibrium (LD) is broken down ($N_e \times$ recombination rate (r))
10 and the strength of selection ($N_e \times$ selection coefficient (s)). For example,
11 theory predicts that if the scaled selection coefficient is below one ($N_e s < 1$)
12 genetic drift will dominate over selection, rendering such mutations ‘effec-
13 tively neutral’. The proportion of such mutations is therefore predicted to
14 increase across the genome through demographic processes that reduce N_e ,
15 such as fluctuating population size or bottlenecks. The interaction between
16 genetic drift and selection can be manifest in a number of different ways,
17 including an increase in the fraction of mutations in functional regions that
18 behave as slightly deleterious or a lower rate of adaptive evolution.

19 In this paper, we compare a population genomic dataset of wild brown
20 rats with a previously published dataset from wild house mice (Baines and
21 Harr, 2007; Halligan et al., 2013) and investigate the differential effects of
22 drift on the genomic signature of selection acting in protein-coding and con-
23 served non-coding DNA in the genome. The effective population size in our
24 focal population of wild house mice (*Mus musculus castaneus*) is nearly two
25 orders of magnitude higher than recent N_e for human populations and sub-
26 stantially higher than that of inbred lab strains (Salcedo et al., 2007; Baines
27 and Harr, 2007; Phifer-Rixey et al., 2012). Halligan et al. (2013) inferred

1 that the protein-coding and conserved noncoding elements (CNE) evolved
2 more rapidly than the neutral theory expectation, suggesting that there has
3 been substantial genome-wide adaptation in proteins and CNEs of wild mice.
4 Moreover, they found reductions in neutral diversity around protein-coding
5 exons and CNEs, indicative of frequent selective sweeps and/or background
6 selection. In contrast, despite a presumably large contemporary census pop-
7 ulation size in wild brown rats (*Rattus norvegicus*), it has been estimated
8 that their N_e is five-fold smaller than wild house mice (Ness et al., 2012).
9 The difference in the N_e between mice and rats provides an opportunity to
10 investigate the effects of genetic drift in the mammalian genome and the
11 way in which selection and drift interact to shape patterns of diversity in
12 the genome.

13 Using whole genome data from 12 rats collected from their ancestral
14 range in NE China and a comparable dataset in wild house mice, we ask
15 a number of questions (1) Does reduced N_e in rats lead to reduced efficacy
16 of selection on new mutations affecting protein or CNE sequence? (2) How
17 does the effect of hitchhiking differ between mice and rats and how does this
18 compare between protein-coding exons and conserved non-coding elements?
19 (3) Does reduced N_e in rats influence the extent of LD? (4) What can
20 patterns of DNA polymorphism tell us about the recent demographic history
21 of wild brown rats? We find strong evidence for a population bottleneck that
22 has distorted the distribution of allele frequencies throughout the genome
23 and altered patterns of LD in wild rats. We also find evidence for a higher
24 frequencies of segregating deleterious mutations in wild rats, consistent with
25 a reduction in the efficacy of purifying selection. However, neutral diversity
26 reductions around protein-coding exons follow a virtually identical pattern in
27 the two species, suggesting that selection has had similar effects on diversity

1 at sites linked to exons, and that these patterns are insensitive to recent
2 changes in N_e .

3 **Results**

4 **Inference of selective forces operating in protein-coding genes** 5 **and CNEs**

6 As a first step to quantify and compare the selective forces acting on vari-
7 ation in the wild rat genome, we computed nucleotide diversity (π) within
8 brown rats, nucleotide divergence (d) from the house mouse and black rat,
9 and Tajima's D for different classes of sites. We focused on two classes
10 of conserved sequences: protein coding exons and CNEs. In exons, 0-fold
11 degenerate sites are under the strongest selection, as any nucleotide change
12 would result in an amino acid substitution. The 4-fold degenerate sites, on
13 the other hand, are typically seen as a neutral standard for these, as mu-
14 tations at these sites do not result in amino acid changes. CNEs are not
15 translated to proteins, so we consider all sites within CNEs as potentially
16 under selection for the purpose of our analysis. As a neutral standard we
17 use CNE flanking sequence at a distance of at least 500 bp with the same
18 total length as the element (following Halligan et al., 2013).

19 Both nucleotide diversity and divergence from the outgroups reflected
20 this expected ranking of selection strength (fig. 1, supplementary table S1,
21 see also figs. S1, S2, S3, S4). Diversity was lowest at 0-fold sites (0.045%),
22 followed by CNEs (0.097%), 2-fold sites (0.112%), 4-fold sites (0.147%) and
23 highest in CNE flanks (0.157%). Divergence from the mouse reference se-
24 quence (mm10) followed the same rank order (3.1%, 7.1%, 10.9%, 14.2% and
25 15.4%). In line with this, Tajima's D was similar for 0-fold sites and CNEs

1 (-0.43 and -0.41, respectively) and values for these sites were lower than
2 at 2-fold sites, 4-fold sites and CNE flanks (-0.27, -0.23 and -0.22, respec-
3 tively). Although all of these values are negative (implying a slight excess
4 of low-frequency variants), they are much closer to zero than what was pre-
5 viously found in wild mice. A recent population bottleneck would affect the
6 genome wide diversity spectrum in a way that produces less negative, or
7 even positive, Tajima's D . Additionally, π/d was higher for 0-fold sites and
8 CNEs than for the other classes (0.015 (0-fold), 0.014 (CNEs), 0.010 (other
9 classes)). This pattern is expected if there are slightly deleterious mutations
10 affecting 0-fold sites and CNEs, since they are expected to contribute more
11 to π than to d .

12 We then estimated distributions of fitness effects (DFE) of new mutations
13 using the program DFE- α (Keightley and Eyre-Walker, 2007) and compared
14 the results to previous estimates from wild house mice (Halligan et al., 2013)
15 (fig. 2). DFE- α compares the folded site frequency spectra (SFSs) of two
16 classes of sites, one neutrally evolving and one under selection, to assess
17 the DFE. It uses that mutations are purged faster from the selected sites
18 when they are more deleterious. In line with theoretical expectations for
19 a smaller N_e in the rat, we inferred that there is a substantially larger
20 proportion of mildly deleterious mutations ($N_e s < 1$), 0.29 and 0.58 in exons
21 and CNEs, respectively, than in the same classes in wild mice (0.17 and 0.44,
22 respectively). Concordantly, the proportions of highly deleterious mutations
23 ($N_e s > 10$) were lower in the rat (exons: 0.65 and CNEs: 0.29) than in the
24 mouse (0.77 and 0.37, respectively). The presence of a higher fraction of
25 slightly deleterious mutations is consistent with the increased values of π/d
26 that we found for exons and CNEs.

27 We also attempted to estimate the fraction of nucleotide differences in

1 exons and CNEs driven to fixation by positive selection (α) and the rate of
2 adaptive substitution relative to the rate of neutral substitution (ω_a). For
3 this we used an extension of the McDonald-Kreitman test incorporated into
4 DFE- α (McDonald and Kreitman, 1991; Eyre-Walker and Keightley, 2009).
5 This subtracts the number of fixed nucleotide substitutions (relative to the
6 mouse) in the selected class of sites that is expected from the fixation of
7 neutral and deleterious mutations alone from the actual number of substi-
8 tutions. The remainder is contributed to positive selection. We consistently
9 obtained negative estimates for α and ω_a for both exons (table S2). As we
10 discuss below, these negative estimates likely reflect a recent population size
11 bottleneck.

12 **Reduced diversity around exons and CNEs**

13 Selection operating within CNEs and exons is also expected to affect nu-
14 cleotide diversity in closely linked surrounding sequences as a consequence of
15 selective sweeps (Maynard Smith and Haigh, 1974) or background selection
16 (Charlesworth et al., 1993). We therefore investigated diversity statistics in
17 exonic (up to 100 kb) and CNE (up to 20 kb) flanking regions, and again
18 compared our results with those previously obtained in wild house mice
19 (Halligan et al., 2013) (fig. 3). Although wild house mice have much higher
20 diversity than what we observe in the rat, the relative reduction in diversity
21 in exon flanks was remarkably similar in both species. The reduction in di-
22 versity (π and π/d) around CNEs, on the other hand, was less pronounced
23 in rat. Moreover, in the bins directly adjacent to the CNEs, we found an
24 increase of π/d , coinciding with a less strong reduction of π than in mouse
25 at these strongly conserved sites.

26 The reduction of π and d in the CNE flanks appears to exist on two

1 length scales: a strong reduction that decays over ~ 500 bp and a second less
2 pronounced reduction that decays over tens of kb. We therefore investigated
3 to what extent this second length scale could be the result of the proximity of
4 CNEs to exons. We first computed the distribution of distances from CNEs
5 to the nearest exon (figs. S5, S6). Due to the power law-like distribution
6 of distances between neighbouring exons, bases tend to “cluster” around
7 exons; this means that more bases are located at a particular short (e.g.
8 10 bp) distance from the nearest exon than at a particular large distance
9 (e.g. 100 kb) from it (fig. 3G). Taking this into account, there remains a
10 two-fold overrepresentation of CNEs near exons (fig. S5). Yet when we
11 used the distribution of distances from CNEs to their nearest exon (fig.
12 S6) to convolute (\approx blur; see methods) the exon flanks, the resulting slope
13 “far away” from the CNEs, i.e., fitted between 5 and 20 kb away, was much
14 shallower than the long length scale in the CNE flanks (fig. S7). This implies
15 that the proximity to exons of many CNEs can only explain a small part of
16 the long length scale we observed in the decay of π in CNE flanks.

17 **LD decay in rat and mouse genomes**

18 To gain an understanding of the striking similarity of the diversity reduc-
19 tions in the exon flanks between rat and mouse and the less similar patterns
20 in CNE flanks, we investigated the decay of linkage disequilibrium (LD)
21 around focal SNPs in wild rats and mice. For this, we computed the pair-
22 wise genomic r^2 (Rogers and Huff, 2009), and averaged over all SNPs at a
23 particular distance from each focal SNP (using a bin size of 20 bp). As focal
24 SNPs, we used either all SNPs, SNPs within exons or SNPs within CNEs
25 (fig. 4AB).

26 In wild house mice, average r^2 (written $\langle r^2 \rangle$) decayed much faster than

1 in wild rats, and the peak value was lower, consistent with the larger N_e
2 in mice. To quantify the difference, we first fitted exponential functions to
3 the decay of $\langle r^2 \rangle$ with physical distance: $f(x) = (a - c) \times \exp(-x/b) + c$.
4 For our purposes, characteristic length b is the biologically most important
5 parameter, as this is the distance over which the relevant information, $\langle r^2 \rangle -$
6 c , decays by a factor $1/e$ (to $\sim 37\%$ of its original value). Maximum value a is
7 the intercept and c is the offset, which has a theoretical minimum of $1/(n-1)$
8 for a sample of n individuals (see supplementary text 2) and increased with
9 decreasing N_e (Hill, 1981).

10 We found that it was impossible, however, to obtain a good fit with this
11 kind of curve (fig. 4C). The structure of the residuals of the best fitting
12 curves (fig. S8AB) suggested that LD ($\langle r^2 \rangle$) decays first faster, then slower
13 than exponential, which is a property of a stretched exponential $g(x) =$
14 $(a - c) \times \exp(-(x/b)^d) + c$, with stretching exponent $0 < d < 1$. See discussion
15 for the biological meaning of d . We obtained good fits to the data with this
16 formula (figs. 4D, S9, table S3A). The rat:mouse ratios of $c - 1/(n - 1)$ were
17 all larger than 1, consistent with a larger N_e in mouse.

18 We fitted all curves again with a fixed stretch exponent of $d = 0.5$ to allow
19 a more direct comparison of the characteristic length parameters b (figs. 4D,
20 S9, table S3B). This effectively exploits the fact that stretched exponentials
21 are notoriously hard to fit to our kind of data. By these means, we found
22 that LD decays 6-7 times faster in mouse than in rat (rat:mouse ratios of b :
23 7.14 exons; 6.31 CNEs “noOverlap”; 5.96 CNEs “strict”; 5.76 all SNPs).

24 **Recent bottleneck in the rat population**

25 The preceding analysis provides several lines of evidence for a recent pop-
26 ulation size bottleneck in wild rats: the much lower diversity than in wild

1 mice (figs. 1B, 3), the negative estimates of α and ω_a (table S2), the values
2 of Tajima's D (fig. 1A) that are much larger than in mice and near zero for
3 some functional categories of exons.

4 To further investigate the possibility of changes in population size, we
5 used the method of Li and Durbin, PSMC, (Li and Durbin, 2011) to infer
6 population history based on the length distribution of genome stretches that
7 are identical by state. Based on the non-CpG prone sites, the 12 rat genomes
8 showed a 3-fold decline in N_e up to 10,000 – 20,000 YA (fig. 5).

9 A similar trend was found using all sites (i.e., including the CpG-prone
10 sites and adjusting mutation rates accordingly; fig. S10), using a different
11 published mutation rate for rat of $\mu = 4.2 \times 10^{-9}$ (Ness et al., 2012) (fig.
12 S11B) and by varying the parameter (md) of the proximity filter (fig. S11),
13 which is the filter that had most impact information used by PSMC, i.e.,
14 the distribution homozygous runs (HHn from MacLeod et al., 2013, see fig.
15 S12). For more details, see supplementary text 3.

16 From this we conclude that the rat genomes contain a strong and robust
17 signal for a bottleneck between 10,000 and 50,000 YA. The actual bottleneck
18 may have been sharper than the PSMC traces show – and, if it has been
19 fairly short, also more severe – as simulations have shown that PSMC has
20 a tendency to smoothen sharp transitions (Li and Durbin, 2011).

21 Discussion

22 Previous work suggests that wild brown rats have a much lower effective
23 population size than wild house mice (Ness et al., 2012). Re-estimating
24 the mutation rate based on our measured divergence from mouse at 4-fold
25 degenerate sites (14.2%), assuming a divergence time of 12 MYA (Benton
26 and Donoghue, 2007) and 2 generations per year (Halligan et al., 2013;

1 Ness et al., 2012), yields $\mu = 2.96 \times 10^{-9}(2.94 \times 10^{-9} - 2.98 \times 10^{-9})$.
2 Equating π at 4-fold degenerate sites to $4N_e\mu$, we obtain an estimate of
3 $N_e = 1.24 \times 10^5(1.20 \times 10^5 - 1.28 \times 10^5)$, marginally smaller than the
4 1.3×10^5 estimate previously published. We obtained a very similar N_e
5 estimate using divergence and diversity from our other neutral class, the
6 CNE flanking sites: $N_e = 1.22 \times 10^5 (1.21 \times 10^5 - 1.23 \times 10^5)$ (and $\mu =$
7 $3.21 \times 10^{-9} (3.20 \times 10^{-9} - 3.23 \times 10^{-9})$).

8 From our PSMC demographic inference we obtain a minimum N_e of
9 approximately 4×10^4 at 2×10^4 years ago and a 3-4 times larger ancestral
10 size (i.e., the most ancient N_e that can be detected with this method). The
11 N_e estimate based on silent site diversity lies between these values, which is
12 likely explained by the fact that it is affected by the whole recent population
13 history. An effective population size of 4×10^4 appears small for the present
14 day, given the abundance and widespread distribution of brown rats, and the
15 small amount of population structure that appears to be present in wild rat
16 brown populations (Ness et al., 2012). However, it is likely that the effective
17 size of the brown rat population has increased dramatically since the origins
18 of human agriculture, which provided them with a large and stable source
19 of food. None of the methods we have used, however, can reliably estimate
20 contemporary N_e .

21 **Negative estimates of the rate of adaptive molecular evolution**

22 Conceptually, the rate of adaptive molecular evolution should always be a
23 positive number. We now review the computational method to explain how
24 a recent population bottleneck – as suggested by the unusually high values
25 of Tajima’s D (fig. 1) and the PSMC demographic inferences (fig. 5) – could
26 cause negative estimates of α , the fraction of substitutions driven to fixation

1 by positive selection. ω_a , the rate of adaptive substitution relative to the
2 rate of neutral substitution, is defined such that it inherits the sign of α
3 (Eyre-Walker and Keightley, 2009) OR (Halligan et al., 2013).

4 To estimate α (and ω_a), DFE- α subtracts an estimate of the fraction
5 of (slightly) deleterious substitutions in the selected class (e.g., 0-fold sites)
6 from the total number of substitutions in the selected class, and assumes that
7 the remaining substitutions have been positively selected. To estimate the
8 (slightly) deleterious fraction, the inferred DFE is used along with a single
9 effective population size (weighted over the whole evolutionary time period
10 under analysis), and the selected (e.g., 0-fold) and neutral (e.g., 4-fold) di-
11 vergence from an outgroup. This implicitly assumes that the strength and
12 effectiveness of selection (i.e., the DFE) were constant over the evolutionary
13 time period under analysis (Eyre-Walker and Keightley, 2009). A recent
14 population size bottleneck is expected to cause an over-representation of
15 slightly deleterious mutations in the polymorphism data used to infer the
16 DFE, which will therefore increase the predicted divergence in the selected
17 site class. This implies that the DFE- α method, and all other similar im-
18 plementations of the McDonald-Kreitman test, will tend to under-estimate
19 the rate of adaptive evolution if there has been a recent population bottle-
20 neck. At the heart of this problem lies the fact that divergence is built up
21 continually from the moment the two lineages split (in this case 12 MYA),
22 whereas polymorphism only reflects a limited window of recent evolution-
23 ary history ($4N_e$ generations). As the size of this window is dependent on
24 N_e , pre-bottleneck information is rapidly lost from polymorphism data as
25 a consequence of a severe bottleneck. Analysing data from species with
26 shorter divergence times may mitigate the impact of long term population
27 size changes, albeit at the expense of power. Using *R. rattus* as an outgroup

1 could not mitigate the problem, as the split between both rat species is ~ 2.9
2 MYA (Robins et al., 2008), i.e., long before the bottleneck we inferred from
3 our data. Consistently, we again obtained negative estimates of α and ω_a
4 (table S2).

5 **Decay of LD with physical distance**

6 We inferred that there is a roughly 6-7 fold faster LD decay in wild mice
7 than wild rats as a function of physical distance in the genome by fitting
8 stretched exponential functions of the form $(a - c) \times \exp(-(x/b)^d) + c$. This
9 is consistent with a recent population size bottleneck, as confirmed using
10 different information such as our PSMC analysis. We found that a stretched
11 exponential including an offset was needed to obtain a satisfactory fit to the
12 data, whereas a single exponential gave a poor fit (figs. 4, S9).

13 A stretched exponential can be obtained by summing over a large number
14 of single exponentials with various exponents (parameter b in our functions),
15 (e.g., Johnston, 2006). Assuming that the recombination rate is constant
16 over the genome, population genetic theory predicts that LD ($\langle r^2 \rangle$) decay
17 follows a single exponential (Charlesworth and Charlesworth, 2010), which
18 can be offset by the theoretical minimum for the sample size ($1/(n - 1)$, see
19 supplementary text 2) + a residual offset for the expected genome wide LD
20 due to finite population size (Laurie-Ahlberg and Weir, 1979; Hill, 1981).
21 The single exponential would also fit if fluctuations in recombination rate
22 average out over sufficiently short distances.

23 Previous estimates of the recombination rate in rat and mouse (Jensen-
24 Seaman et al., 2004), however, showed evidence of fluctuations on a scale of
25 much more than 10 Mb, i.e., larger than our window for calculating $\langle r^2 \rangle$,
26 and also showed evidence for variation in whole chromosome average recom-

1 bination rate. This variation could explain the requirement for a stretched
2 exponential. As the major fluctuations in the reported data for mouse and
3 rat are on the same length scale, this also explains that we found smaller val-
4 ues of the stretch exponent d for mouse than for rat (table S3A): the smaller
5 the relevant window for LD decay, the larger the variation in the (weighted)
6 average recombination rate over the window. The roughly 6-7 fold faster
7 LD decay in mouse also explains why the relative differences among classes
8 (exons, CNEs, all data) of the fitted values are larger in mouse.

9 In principle, the fitted value of the offset c minus its theoretical minimum
10 carries information about N_e (Laurie-Ahlberg and Weir, 1979; Hill, 1981).
11 With our window length and limitations on resolution close to the focal SNP,
12 however, it is impossible to obtain a reliable estimate of the stretch exponent
13 d or the offset c independently. We, therefore, refrain from estimating N_e
14 in this way.

15 **Reductions of nucleotide diversity around protein-coding ex-** 16 **ons and CNEs**

17 A striking finding from our study is the extremely similar proportional re-
18 ductions in mean scaled neutral nucleotide diversity around protein-coding
19 exons in wild rats and in wild house mice (fig. 3). The depth, width and
20 shape of the reductions in diversity are all similar. The drops in diversity
21 are presumably caused by the hitchhiking effect of selection on variants in
22 protein-coding exons, which reduces diversity in tightly linked flanking re-
23 gions. The previous analysis of the pattern of diversity reduction in wild
24 house mice suggests that there is a substantial role for selection of advanta-
25 geous mutations, whereas a background selection model alone appears to be
26 incapable of explaining the width of the observed mean diversity reduction

1 (Halligan et al., 2013). The question then arises as to whether these similar
2 patterns around exons in the two species can be reconciled with the differ-
3 ence in the effectiveness of selection between the species (caused by lower
4 N_e in rats) and the presence of substantially greater LD in rats than in
5 mice. If diversity at linked sites is reduced by selection of newly arising ad-
6 vantageous mutations in exons of large selective effect(s) that go to fixation
7 in both species (i.e., classic selective sweeps such that $N_e s \gg 1$), and the
8 rate and strength of advantageous mutations and the rate of recombination
9 per physical distance are the same in the two species (0.555 cM/Mb in rat
10 and 0.528 cM/Mb in mouse (Jensen-Seaman et al., 2004)), then equivalent
11 patterns of diversity reduction are predicted, and these are not expected to
12 depend on N_e or LD (Maynard Smith and Haigh, 1974). A similar argument
13 can be made for the case of background selection (BGS) involving strongly
14 deleterious mutations (Nordborg et al., 1996). Alternatively, if diversity re-
15 ductions are caused by positive selection on standing variation, the pattern
16 of diversity reduction is expected to depend on the effective population size
17 during the phase in which a variant can rise to a high frequency by drift,
18 and subsequently be positively selected to fixation (Przeworski et al., 2005).
19 Specifically, a higher N_e increases the difference between the pattern of di-
20 versity change seen under a classic sweep model and a model of standing
21 variation. The similarity of the diversity reduction patterns surrounding
22 protein-coding exons we observe between mice and rats, species which differ
23 substantially in recent N_e , is therefore indirect evidence in favour of the
24 classic selective sweeps model.

25 Narrower and shallower scaled diversity reductions in the regions sur-
26 rounding CNEs are also present, but these have somewhat different patterns
27 in mice and rats (fig. 3). Specifically, diversity reductions are shallower in

1 rats and diversity returns to a genomic background level more slowly in rats
2 than mice. It was previously shown in mice that the diversity reductions can
3 be explained by a BGS model, although a role for positive selection was not
4 excluded (Halligan et al., 2013). If diversity reductions are mainly caused
5 by BGS, a weaker effect is expected in rats than mice if there are substantial
6 numbers of CNE mutations which have selective effects $< 1/N_e$ in rats and
7 $> 1/N_e$ in mice, because these would behave as nearly neutral in rats and
8 therefore have a smaller influence on linked neutral diversity than in mice.
9 This is consistent with our estimates of the DFE in the two species, which
10 suggest that there are substantially more deleterious mutations in CNEs
11 with selective effects $< 1/N_e$ in rats than mice (fig. 2).

12 **Conclusion**

13 We have conducted a whole genome polymorphism study to quantify the
14 selective forces shaping recent wild brown rat evolution and compared our
15 findings to a similar study in wild house mice. We found a larger proportion
16 of slightly deleterious mutations in rats than in mice for both protein-coding
17 exons and CNEs, in line with the theoretical expectation for a larger N_e in
18 mice. The data also shows evidence for a recent population bottleneck in
19 rats, which we dated at roughly 20,000 years ago using a PSMC analysis,
20 followed by a likely explosion of population size starting roughly at the same
21 time as the rise of agriculture in humans. The population size bottleneck
22 distorted the allele frequency distribution, leading to unusually high, but still
23 negative, Tajima's D values, and led to substantially more LD than observed
24 in wild mice. Strikingly, however, we found a very similar pattern in the
25 reduction of π/d in the tens of kbs flanking protein-coding exons, which are
26 consistent with recurrent selective sweeps on newly arising advantageous

1 mutations.

2 **Methods**

3 **Samples**

4 We obtained genomic DNA from 22 wild *R. norvegicus* trapped in a ~500-
5 km² area around the city of Harbin, Heilongjiang Province, China in 2011
6 (Ness et al., 2012) from locations a minimum of 100 m apart to avoid sam-
7 pling of closely related individuals. We selected 12 of these individuals for
8 whole-genome sequencing. DNA from one individual black rat (*R. rattus*)
9 that died of natural causes was obtained from Bristol Zoo's colony.

10 **Sequencing**

11 Genomic DNA was extracted from a small piece of kidney tissue. Standard
12 Illumina 100bp PE libraries for the HiSeq sequencer with an insert size of
13 approximately 450bp were prepared according to manufacturers recommen-
14 dations. The Illumina sequencing was performed at the Wellcome Trust
15 Sanger Institute. We obtained a modal coverage of 19x – 46x per sample
16 for *R. norvegicus* and 33x for *R. rattus* (table S4). Reads were aligned to
17 the rn5 reference (from Ensembl release 71) using BWA version 0.5.10-mt Li
18 and Durbin (2009). All lanes from the same library were then merged into
19 a single BAM file using Picard tools (<http://picard.sourceforge.net>)
20 and PCR duplicates were marked using Picard tools 'MarkDuplicates'. Fi-
21 nally, the library BAM files were merged into a single BAM containing all
22 sequencing reads for that sample.

1 SNP calling and filtering

2 We used the Genome Analysis Toolkit (GATK) UnifiedGenotyper version
3 2.8-1-g932cd3a for SNP calling (DePristo et al., 2011), using the following
4 non-default arguments: output mode: emit all confident sites; genotype
5 likelihoods model: both; stand emit conf: 10. By choosing the latter pa-
6 rameter value, we obtained information about sites called with relatively
7 low confidence, which were filtered subsequently, as described below.

8 Before SNP calling, we first performed indel realignment using GATK
9 IndelRealigner with default parameters on the BAM files containing the
10 aligned reads simultaneously on all 12 samples. There is a SNP database
11 available for *R. norvegicus* from Ensembl, but this contains only 10% of
12 the putative variant sites in our data (estimated from release 71, [ftp://ftp.
13 ensembl.org/pub/release-71/variation/vcf/rattus_norvegicus/](ftp://ftp.ensembl.org/pub/release-71/variation/vcf/rattus_norvegicus/)). Re-
14 calibrating bases using a limited SNP data set after realignment may have
15 introduced a significant bias, so we did not do this. It has been shown,
16 moreover, that the combination of local realignment and base recalibration
17 is likely to result in biased SNP calls (Guo et al., 2012). We used all putative
18 indels from a first round of SNP calling to mask all sites near putative indels
19 (for deletions: deleted bases + 1 base on either side; for insertions: insert
20 length + 1 base on either side of insertion point). We also performed a
21 second round of SNP calling using the same parameters (and GATK version
22 2.7-4-g6f46d11), but without the indel realignment step. We filtered out
23 putative SNPs that did not appear in both sets. Both these steps were done
24 because BWA (and any other alignment algorithm) is prone to introduce
25 false SNPs near indels, particularly in low complexity regions (as described
26 in the online GATK documentation). This occurs because the penalty for
27 introducing a small number of false SNPs may be lower than the penalty

1 for introducing a gap, and is most likely to occur close to sequencing read
2 ends.

3 We further filtered sites that had a GATK quality score $QUAL < 23$. This
4 threshold was chosen post hoc, based on the distribution of scores of invari-
5 ant sites. Our samples contained a very small fraction of invariant sites with
6 $QUAL < 24$, above which the density increases markedly (fig. S13). Choos-
7 ing our threshold just below this value therefore allows filtering on quality
8 without introducing substantial bias against invariant sites. We excluded all
9 sites that had an inbreeding coefficient $F < -0.8$. GATK only computes F for
10 a site if at least 10 samples are called at that site (online GATK documenta-
11 tion). Using this threshold for F , all sites that have exclusively heterozygous
12 individuals are excluded, which is a strong indication of paralogous reads
13 mapping to the same region. Following common practice, we filtered out
14 high and low coverage regions, since such regions are prone to SNP calling
15 errors. We exploited the fact that we have 12 samples by applying relatively
16 lenient bounds on a per sample basis (between 25 and 300% of the sample's
17 modal coverage) and using much stricter bounds on the average normalized
18 coverage (between 50 and 140%). The latter bounds were derived from the
19 distribution of autosome wide average coverages (fig. S14). There was more
20 than a factor of two difference between the highest and lowest modal cov-
21 erage, so we used each individual sample's modal coverage, computed from
22 the whole autosome, for normalizing coverage. Throughout our analyses,
23 we only considered sites that had at least 3/12 samples called after filtering.
24 In some analyses we applied a "**proximity filter**" that removed all variant
25 sites less than $md = 5$ bp from another variant site, regardless of site qual-
26 ity. The filter does appear to cause a large number of false negatives, so we
27 applied it only to analyses that are highly sensitive to false positives. The

1 proximity filter had a stronger impact on π and D than the precise selection
2 criteria for exons/ CNEs, but had little impact on divergence statistics (figs.
3 S1, S2, S3, S4). It never affected the rank order of different classes of sites.

4 For the *R. rattus* outgroup we used the same SNP calling pipeline, with
5 the following exceptions: 1) we applied indel realignment to the *R. rattus*
6 genome, aligned to the rn5 reference, in isolation. 2) we used a minimum
7 base quality cutoff of 13 rather than the GATK default 17, based on an
8 analysis of GC content and average base quality using Picard tools Col-
9 lectGcBiasMetrics. 3) we used a minimum QUAL cutoff of 30, which was
10 determined in the same way as for the *R. norvegicus* data, and is higher
11 because of the higher modal coverage of 33x in the *R. rattus* sample. 4) we
12 masked sites near indels using the *R. rattus* putative indels. 5) we required
13 that sites have a normalized coverage between 40% and 200% of the *R. rat-*
14 *tus* modal coverage. We have only one *R. rattus* sample, so filtering against
15 likely paralogs based on Hardy-Weinberg frequencies was not possible.

16 CpG-prone sites, defined as sites preceded by a C or followed by a G,
17 were identified based on the rn5 and mm10 reference sequences as well as
18 the *R. norvegicus* and *R. rattus* samples. We excluded sites that were CpG-
19 prone in any of these sequences from several of our analyses, because the
20 hypermutability of the CG dinucleotide strongly violates the assumption of
21 a uniform mutation rate across the genome, which is commonly made in
22 the theory underpinning many population genetic analyses. Exclusion of
23 CpG-prone is an effective way of removing this source of hypermutability
24 (Gaffney and Keightley, 2008).

25 The *R. norvegicus* sample consists of four males and eight females. For
26 consistency in filtering and statistics, we restricted all our analyses to auto-
27 somes, and excluded unplaced contigs.

1 **Exons**

2 We used the Ensembl Rnor5.0.73 annotation file to obtain the locations of
3 exons. We confined our analysis to exons that are part of complete tran-
4 scripts, i.e., starting with a start codon, terminated by a stop codon and
5 containing no premature stop codons. The annotation file contains 25,725
6 transcripts, 20,278 of which are complete: 19,530 on the autosomes, 697
7 on the X chromosome, 9 on the mitochondrial genome and 42 on unplaced
8 contigs.

9 Exonic sites were analyzed only if they were consistently 0-fold, 2-fold,
10 or 4-fold degenerate over all annotated transcripts in the rat annotation,
11 based on computational translation of all canonical and non-canonical tran-
12 scripts in *R. norvegicus* containing the site. Sites with inconsistent de-
13 generacy (e.g., a site that is 4-fold degenerate in a canonical transcript, but
14 0-fold degenerate in an overlapping non-canonical transcript) were excluded.
15 Confidence intervals were computed using n=1,000 bootstrap replicates by
16 sampling per transcript. For details see supplementary text 1.

17 **Conserved noncoding elements (CNEs)**

18 Noncoding sequences conserved across the mammalian phylogeny, CNEs,
19 were defined using phastCons on a mammal phylogeny excluding rodents
20 as described in Halligan et al. (2013). This resulted in a set of elements
21 comprising 5% of the genome.

22 In Halligan et al. (2013), the resulting set of elements was lifted over
23 from human hg18 coordinates to mouse mm9 using liftOver. Each liftOver
24 step inevitably results in the loss of a subset of the elements. Simply lifting
25 over the Halligan et al. (2013) set from mouse mm9 to rat rn5 might,
26 therefore, result in a set biased towards higher conservation. To minimize

1 this potential bias, we used three different routes for lifting over the original
2 hg18 coordinates to our rat rn5 reference: hg18 → hg19 → rn5; hg18 →
3 rn4 → rn5; and hg18 → mm9 → rn5. In case of conflicts, i.e., if different
4 liftOver routes placed the same element at different positions on rn5, we only
5 retained those that differed by no more than 25% in length from the original
6 hg18 element. If different liftOver routes resulted in partially overlapping
7 elements, we retained the one with length closest to the original (fig. S15).
8 From this set of elements, we removed those segments of the elements that
9 overlapped with annotated exons (valid and invalid). This final set of CNEs
10 we call “noOverlap”.

11 To test how sensitive our results are to the precise definition of CNEs, we
12 also created two slightly different sets of CNEs: “strict”: removing all ele-
13 ments that have any overlap with annotated exons, not just the overlapping
14 segments; and “noOverlap, <1 kb”: the noOverlap set excluding elements
15 longer than 1kb. By analyzing the normalized CNE length distributions for
16 different sets, we found that length distributions are very similar between
17 the human, mouse and rat sets, with a slight inflation of very long elements
18 in the rat (“noOverlap”) and mouse sets (fig. S15). The number of elements
19 in these inflated tails was small, but because the median CNE length is very
20 short, these elements, which are less strongly conserved than the average,
21 could have a disproportionate impact on the within-CNE statistics. From
22 the cumulative length distribution (fig. S15) it can be seen that in the human
23 set, less than 1% of CNE bases occur in elements of over 1kb, whereas in the
24 rat and mouse sets this is 5-6%. For this reason we used the “noOverlap,
25 <1kb” set as default for the within-CNE statistics. For the CNE flanks there
26 was almost no difference between both noOverlap sets, because the number
27 of long (>1 kb) elements is only a very small fraction the total number (fig.

1 S16).

2 As a neutral reference for the CNEs, we used sequence elements 500 bp
3 upstream and downstream of the CNE, each of half the length of the CNE
4 (Halligan et al., 2013). From this set we masked any segments (noOverlap)
5 or elements (strict) overlapping with exons or other CNEs (from the full
6 noOverlap set).

7 We used a bootstrapping approach to obtain confidence intervals for the
8 within CNE statistics. We subdivided the genome in 1Mb windows and
9 sampled with replacement among the non-empty windows.

10 **Estimating the impact of exons on the diversity reduction in** 11 **CNE flanks**

12 Most CNEs occur in the vicinity of exons. We hypothesized that this ex-
13 plains (part of) the reduction of diversity in CNE flanks, especially farther
14 away from the CNEs. As this effect is likely stronger for CNEs close to exons
15 than CNEs far away from them, we “blurred” the exon flanks based on the
16 distances from CNEs and compared the slopes of the CNE flanks and the
17 “blurred” exon flanks using a linear fit on the data 5-20 kb away from the
18 CNEs.

19 The idea for the “blurring”, technically called convolution, is taken from
20 image analysis and uses the same principle as simple blurring algorithms
21 in common image manipulation programs. In those, each pixel of image is
22 multiplied with the so called kernel, which determines how far information
23 is smeared out over neighbouring pixels. In our case, we used the 100 bp
24 bins with which the flanks were computed as “pixels”. Instead of a Gaussian
25 kernel, that is typical for image blurring, we used the normalized distribution
26 of distances of CNEs (using the “noOverlap” set) to their nearest exon (fig.

1 S6), up to a maximum of 200 kb, which covers >98.7% of all CNEs.

2 Before convolution, the exon flanks (not using the proximity filter, i.e.,
3 $md=0$) were padded (extended) with 200 kb of 100 bp bins containing the
4 average value of π over the regions 60-100 kb away from the exons in order
5 to accommodate the full width of the kernel.

6 **Functional annotation**

7 For functional annotation of transcripts, we used the eggNog version 4.0
8 database, downloaded from [http://eggnog.embl.de/version_4.0.beta/](http://eggnog.embl.de/version_4.0.beta/downloads.v4.html)
9 [downloads.v4.html](http://eggnog.embl.de/version_4.0.beta/downloads.v4.html) on 14/01/2014 (Powell et al., 2014).

10 We combined annotations at different taxonomic levels. In cases of
11 conflicts, we gave preference to annotations of narrower taxonomic levels
12 (e.g. rodents over mammals), with the exception of vague annotations (“R”:
13 “General function prediction only” or “S”: “Function unknown”). We used
14 “X” for transcripts that did not appear in the database. Of the 25,725 tran-
15 scripts in the annotation file, 21,242 appeared in the database, and of those
16 16,888 had a function assigned to them (i.e., not R or S). In a subsequent
17 analysis we found that the group of X labeled transcripts was a consistent
18 outlier compared to all other annotations on a variety of metrics. We there-
19 fore repeated the within-exon analysis excluding these transcripts to assure
20 that our results do not hinge on this specific subset.

21 **Genome-wide LD scans**

22 For each variant site (“focal SNP”) we computed the pairwise r^2 statistic
23 directly from the genotypic data (Rogers and Huff, 2009) using up to n_{max}
24 = 1,500 neighbouring variant sites and sites up to 40 kb away from the focal
25 SNP. We collected averages in bins of 20 bp. The value of n_{max} was chosen

1 such that it would not lead to the exclusion of sites from the analysis. This
2 value allows for an average diversity of 3.75% within the 40kb, >20 times
3 the average value we found at 4-fold synonymous sites and in 500 bp outside
4 CNEs (fig. 1), before excluding sites. To check this, we verified that the
5 number of sites in the most distant bins did not increase by further increasing
6 n_{max} in a preliminary analysis using the entire chromosome 19. Note that
7 from the highest diversity estimates reported here, the expected number
8 of SNPs within 40kb would be more than an order of magnitude smaller
9 than our n_{max} , leaving a wide margin for variation in SNP density. We
10 only considered biallelic sites for which all samples were called and passed
11 the default filters. Moreover, we applied the proximity filter with $md=5$ for
12 this analysis, because without it we observed small a depression in the $\langle r^2 \rangle$
13 curves within 200 bp of the focal SNP, whereas theory predicts a monotonic
14 decline (fig. S8C). This suggests that a fraction of the spurious SNPs that
15 are removed by the proximity filter are of a different kind than the genuine
16 SNPs, which results in a population of sites with much faster decay of $\langle r^2 \rangle$
17 than the others. Sites near indels and CpG-prone sites were not considered
18 in this analysis.

19 **Genome wide LD scans in *M. m. castaneus***

20 Genotypes were obtained from the VCF files generated by (Halligan et al.,
21 2013). To treat the data as similarly as possible to the rat data, we calcu-
22 lated modal coverage per sample from per sample coverage histograms for
23 the autosome and applied the same bounds as on the *R. norvegicus* data:
24 average normalized coverage between 50 and 140% and per sample between
25 25 and 300%. We also determined quality histograms for variant and invari-
26 ant sites, from which we obtained a minimum quality score of 15. Following

1 the original analysis, we disregarded sites with a χ^2 score for HW equilib-
2 rium ≥ 0.0002 (against paralogs) and near indels. We removed CpG prone
3 sites based on CpG prone status in *M. m. castaneus*, *M. m. famulus* and
4 *R. norvegicus* (rn4). As with the rat data, all samples had to pass all fil-
5 ters. The mouse sample has a much higher sequence diversity than the rat
6 sample, so to cut computational cost, and because a preliminary analysis
7 showed a much faster decay of $\langle r^2 \rangle$ than in rat, we only considered sites
8 up to 20 kb away from the focal site. We used $n_{max} = 7,500$, equivalent
9 to 15,000 on 40kb or 10x the n_{max} used for rat, whereas the difference in
10 diversity in regions >80 kb from exons is less than five fold. The proximity
11 filter had virtually no impact on the mouse data (fig. S8C).

12 Inference of population history

13 We used the inference method and software Pairwise Sequentially Markovian
14 Coalescent, PSMC from (Li and Durbin, 2011). For prediction of population
15 size at the most recent time scales, this method is known to be sensitive to
16 false positive heterozygote sites (MacLeod et al., 2013). We therefore re-
17 quired that at least 10/12 individuals were called, so that inbreeding coeffi-
18 cient estimates are available in the GATK output. We applied the proximity
19 filter with $md=5$ (and $md=10$ for further sensitivity analysis) and at the in-
20 dividual sample level we only considered calls with a genotype quality ≥ 20 .
21 Following (Li and Durbin, 2011), we binned the genome into 100 bp windows
22 and created a fasta-like sequence consisting of the letters “K” (at least one
23 heterozygous site in the bin), “T” (no heterozygous sites) and “N” (less than
24 10% of sites in the bin called and remaining after filtering). These sequences
25 were then directly input to PSMC using the following arguments: `-N80 -`
26 `r0.63` (for not CpG-prone only) or `1.33` (all sites) `-t15 -p “2*4+18*2+4+6”`.

1 These parameters differ from the defaults recommended for the analysis of
2 human data in the PSMC documentation <https://github.com/lh3/psmc>
3 in three ways. First, we used a much larger number of iterations, after find-
4 ing that the default ($N=25$) was insufficient for convergence (convergence
5 of all inferences checked in fig. S17). Second, we reduced the number of
6 free parameters to prevent overfitting (default: -p “4+25*2+4+6”). Third,
7 we initiated r , the ratio θ/ρ , to per base pair estimates of μ/c (default:
8 $r=5$). Parameter t (initial value for history length in units of number of
9 generations/ $2N_0$) was taken from the online documentation, claimed to give
10 good results on human data. The program failed to produce meaningful
11 results using a higher starting value of $t=20$.

12 To scale the inferred demographic histories to real time, we assumed a
13 recombination rate of 0.6 cM/Mb (Jensen-Seaman et al., 2004) and a com-
14 bination of a generation time of 0.5 years (Ness et al., 2012) and a mutation
15 rate of 2.96×10^{-9} per base pair (calculated from our data of divergence
16 from mouse on 4-fold synonymous sites) on the data excluding CpG-prone
17 sites and 5×10^{-9} or 8×10^{-9} on the full sequence data to reflect the higher
18 mutation rate of CpG-prone sites. The latter mutation rate corresponds,
19 for example, to a 2.7 or 5.3 times higher mutation rate, respectively, and
20 CpG-prone sites taking up 40% of the genome and was chosen because it
21 produced either same length genealogies (5×10^{-9}) or had the main local
22 minimum coinciding in time (8×10^{-9}). As the model assumes that the mu-
23 tation rate is constant over time and throughout the genome, we considered
24 inferences based on all sequence as less reliable than those based on the
25 non-CpG prone sites, but did produce them as a qualitative control. For
26 details on testing the robustness of the inference, see supplementary text 3.

1 **Data access**

2 FastQ and BAM files containing all reads aligned to the rn5 reference genome
3 are deposited in the European Nucleotide Archive as study ERP001276
4 <http://www.ebi.ac.uk/ena/data/view/ERP001276&display=html>.

5 **Acknowledgements**

6 We are grateful to the Wellcome Trust for funding and for grants from
7 the Strategic Priority Research Program of the Chinese Academy of Sci-
8 ences [XDB11010400 to JXZ], and the China National Science Foundation
9 [91231107 JXZ and 31301887 to YHZ]. We thank David Adams and Thomas
10 Keane for Illumina sequencing and Sue Dow and Adina Valentine (Bristol
11 Zoo) for providing us with a *R. rattus* sample.

12 **Disclosure declaration**

13 The authors declare no conflicts of interest.

1 **Figure legends**

2 **List of Figures**

3 1 Tajima's D (A), diversity (B) and divergence from mouse
4 (mm10; C) and *R. Rattus* (D) for wild *R. norvegicus* data.
5 Error bars show 95% confidence intervals based on 1000 boot-
6 strapping replicates. See supplementary material for impact
7 of proximity filter, exon/CNE selection rules (figs. S1, S2, S3
8 and S4) and restricting analysis to sites with corresponding
9 *R. rattus* bases (table S1). 34
10 2 DFE of deleterious mutations for rat exons (red), mouse exons
11 (blue), rat CNEs (magenta) and mouse CNEs (cyan). Error
12 bars indicate 95% confidence intervals from 1000 bootstrap-
13 ping replicates. Mouse data after Halligan et al. (2013). . . . 35

- 1 3 Exon (A, C, E, G) and CNE (B, D, F, H) flanks. A,B: di-
 2 vergence between rat and mouse. C,D: Diversity (π). To fa-
 3 cilitate comparison of relative changes between both species,
 4 rat values are indicated on the left axis, mouse values on the
 5 right axis. E,F: π/d . G,H: Number of sites per bin. Each site
 6 was only counted once, in relation to its nearest exon/CNE.
 7 We used a bin size of 1 kb for exon flanks (up to 100 kb)
 8 and 100 bp for CNE flanks (up to 5 kb). Rat data in cyan,
 9 mouse data in red. We assessed the impact of the proximity
 10 filter (with $md=5$) on nucleotide diversity in exon and CNE
 11 flanks. This resulted in an overall reduction of π of $\sim 17\%$ and
 12 some increase of the π/d spike directly surrounding the CNEs,
 13 without further changing the general shape of the curves (fig.
 14 S16). The magnitude of this reduction was 1-2 times as large
 15 as the proximity filter's impact within exons (0-fold: 17%,
 16 4-fold: 9%) or CNEs (11%). 36
- 17 4 LD measured by genomic $\langle r^2 \rangle$ as a function of distance from
 18 focal SNP. A,B: $\langle r^2 \rangle$ for all chromosomes combined, with all
 19 SNPs, or only those within exons or CNEs as focal SNP. A:
 20 rat, B: mouse. All values are averages over 20 bp bins. C:
 21 Average of r^2 for all chromosomes combined, all rat SNPs,
 22 with offset exponential fit $f(x) = (a - c) * \exp(-x/b) + c$. D:
 23 Same curve, with an offset stretched exponential fit: $g(x) =$
 24 $(a - c) * \exp(-(x/b)^d) + c$ 37

1 5 Inferred population history from distribution of IBS (identical
2 by state) tract length distributions Li and Durbin (2011).
3 Curves represent the inferred population history of individual
4 rat samples. Estimates are based on non-CpG prone sites only
5 and using a proximity filter of $md = 5$ bp. see figs. S10 and
6 S11 for effects of these choices. 38

1 Figures

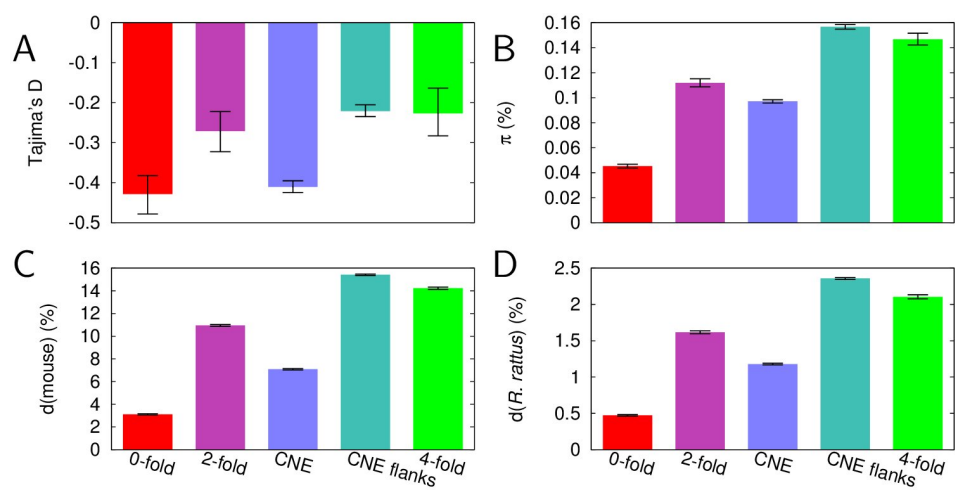


Figure 1

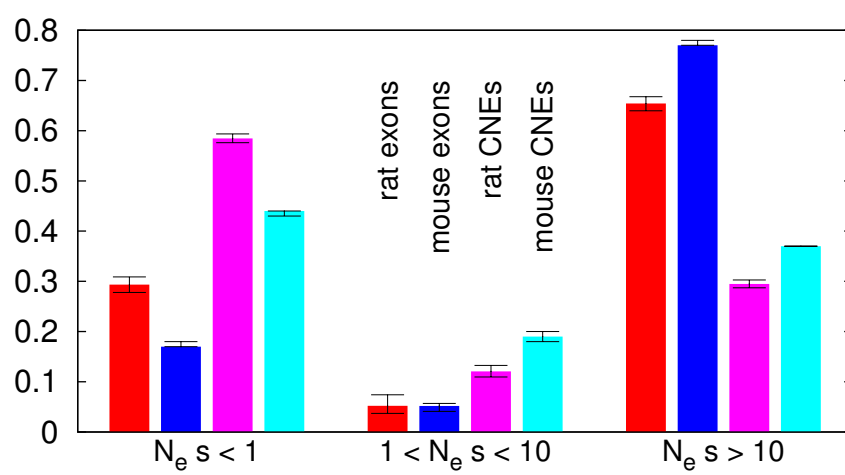


Figure 2

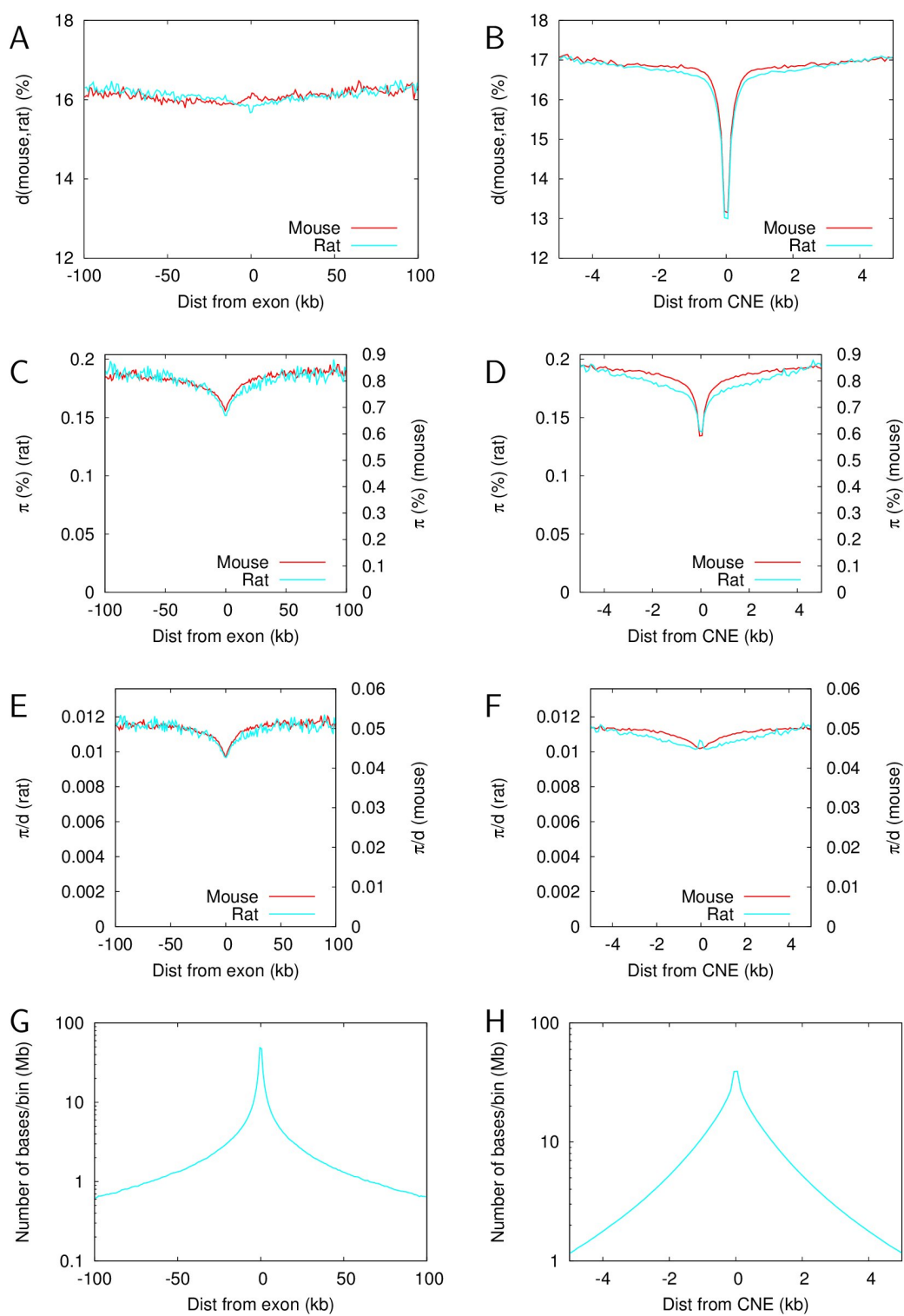


Figure 3

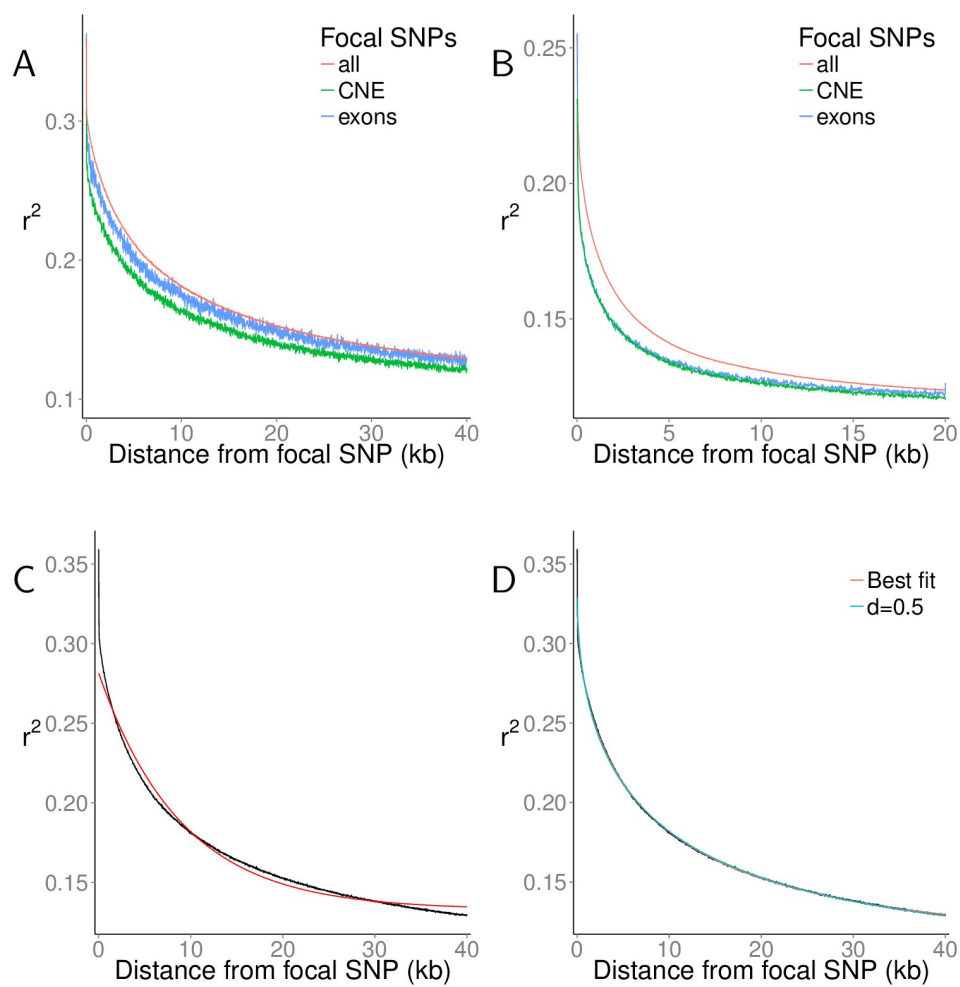


Figure 4

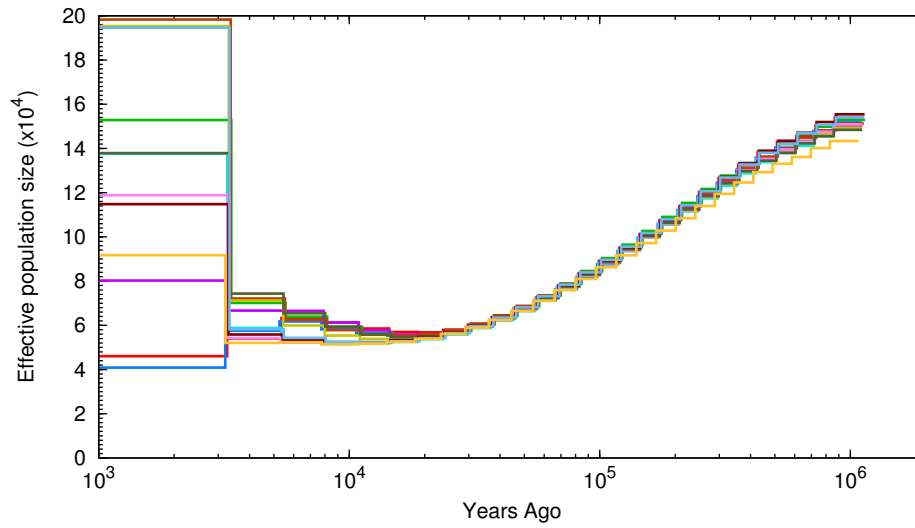


Figure 5

1 **Tables**

2 **References**

- 3 Baines JF and Harr B, 2007. Reduced x-linked diversity in derived popula-
4 tions of house mice. *Genetics*, **175**(4):1911–1921.
- 5 Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista
6 PM, Jones CD, Kern AD, Dewey CN, *et al*, 2007. Population genomics:
7 whole-genome analysis of polymorphism and divergence in drosophila sim-
8 ulans. *PLoS Biol*, **5**(11):e310.
- 9 Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, and
10 Haussler D, 2004. Ultraconserved elements in the human genome. *Science*,
11 **304**(5675):1321–1325.
- 12 Benton MJ and Donoghue PCJ, 2007. Paleontological evidence to date the
13 tree of life. *Mol Biol Evol*, **24**(1):26–53.
- 14 Cai JJ, Macpherson JM, Sella G, and Petrov DA, 2009. Pervasive hitchhik-
15 ing at coding and regulatory sites in humans. *PLoS Genet*, **5**(1):e1000336.
- 16 Charlesworth B and Charlesworth D, 2010. *Elements of evolutionary genet-*
17 *ics*. Roberts and Company Publishers.
- 18 Charlesworth B, Morgan MT, and Charlesworth D, 1993. The effect of dele-
19 terious mutations on neutral molecular variation. *Genetics*, **134**(4):1289–
20 1303.
- 21 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C,
22 Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al*, 2011. A frame-
23 work for variation discovery and genotyping using next-generation dna
24 sequencing data. *Nat Genet*, **43**(5):491–498.

- 1 Eyre-Walker A and Keightley PD, 2009. Estimating the rate of adaptive
2 molecular evolution in the presence of slightly deleterious mutations and
3 population size change. *Mol Biol Evol*, **26**(9):2097–2108.
- 4 Gaffney DJ and Keightley PD, 2008. Effect of the assignment of ancestral
5 cpg state on the estimation of nucleotide substitution rates in mammals.
6 *BMC Evol Biol*, **8**:265.
- 7 Guo Y, Li J, Li C-I, Long J, Samuels DC, and Shyr Y, 2012. The effect
8 of strand bias in illumina short-read sequencing data. *BMC Genomics*,
9 **13**:666.
- 10 Halligan DL, Kousathanas A, Ness RW, Harr B, Ery L, Keane TM, Adams
11 DJ, and Keightley PD, 2013. Contributions of protein-coding and reg-
12 ulatory change to adaptive molecular evolution in murid rodents. *PLoS*
13 *Genet*, **9**(12):e1003995.
- 14 Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G,
15 Project G, Sella G, and Przeworski M, 2011. Classic selective sweeps
16 were rare in recent human evolution. *Science*, **331**(6019):920–924.
- 17 Hill WG, 1981. Estimation of effective population size from data on linkage
18 disequilibrium. *Genetical Research*, **38**(03):209–216.
- 19 Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-
20 F, Thomas MA, Haussler D, and Jacob HJ, 2004. Comparative recom-
21 bination rates in the rat, mouse, and human genomes. *Genome Res*,
22 **14**(4):528–538.
- 23 Johnston D, 2006. Stretched exponential relaxation arising from a continu-
24 ous sum of exponential decays. *Physical Review B*, **74**(18):184430.

- 1 Keightley PD and Eyre-Walker A, 2007. Joint inference of the distribution of
2 fitness effects of deleterious mutations and population demography based
3 on nucleotide polymorphism frequencies. *Genetics*, **177**(4):2251–2261.
- 4 Laurie-Ahlberg C and Weir B, 1979. Allozymic variation and linkage dis-
5 equilibrium in some laboratory populations of *Drosophila melanogaster*.
6 *Genetics*, **92**(4):1295–1314.
- 7 Li H and Durbin R, 2009. Fast and accurate short read alignment with
8 Burrows-Wheeler transform. *Bioinformatics*, **25**(14):1754–1760.
- 9 Li H and Durbin R, 2011. Inference of human population history from
10 individual whole-genome sequences. *Nature*, **475**(7357):493–496.
- 11 MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, and Goddard ME, 2013. In-
12 ferring demography from runs of homozygosity in whole-genome sequence,
13 with correction for sequence errors. *Mol Biol Evol*, **30**(9):2209–2223.
- 14 Maynard Smith J and Haigh J, 1974. The hitch-hiking effect of a favourable
15 gene. *Genetical research*, **23**(01):23–35.
- 16 McDonald JH and Kreitman M, 1991. Adaptive protein evolution at the
17 *adh* locus in *Drosophila*. *Nature*, **351**(6328):652–654.
- 18 Mouse Genome Sequencing Consortium, 2002. Initial sequencing and com-
19 parative analysis of the mouse genome. *Nature*, **420**(6915):520–562.
- 20 Ness RW, Zhang Y-H, Cong L, Wang Y, Zhang J-X, and Keightley PD, 2012.
21 Nuclear gene variation in wild brown rats. *G3 (Bethesda)*, **2**(12):1661–
22 1664.
- 23 Nordborg M, Charlesworth B, and Charlesworth D, 1996. The effect of
24 recombination on background selection. *Genet Res*, **67**(2):159–174.

- 1 Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Pilek J, Tucker PK,
2 and Nachman MW, 2012. Adaptive evolution and effective population size
3 in wild house mice. *Mol Biol Evol*, **29**(10):2949–2955.
- 4 Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas
5 J, Gabaldon T, Rattei T, Creevey C, Kuhn M, *et al*, 2014. eggNOG v4.0:
6 nested orthology inference across 3686 organisms. *Nucleic Acids Research*,
7 **42**(D1):D231–D239.
- 8 Przeworski M, Coop G, and Wall JD, 2005. The signature of positive selec-
9 tion on standing genetic variation. *Evolution*, **59**(11):2312–2323.
- 10 Robins JH, McLenachan PA, Phillips MJ, Craig L, Ross HA, and Matisoo-
11 Smith E, 2008. Dating of divergences within the rattus genus phylogeny
12 using whole mitochondrial genomes. *Molecular Phylogenetics and Evolu-
13 tion*, **49**(2):460 – 466.
- 14 Rogers AR and Huff C, 2009. Linkage disequilibrium between loci with
15 unknown phase. *Genetics*, **182**(3):839–844.
- 16 Salcedo T, Geraldes A, and Nachman MW, 2007. Nucleotide variation in
17 wild and inbred mice. *Genetics*, **177**(4):2277–2291.
- 18 Sattath S, Elyashiv E, Kolodny O, Rinott Y, and Sella G, 2011. Pervasive
19 adaptive protein evolution apparent in diversity patterns around amino
20 acid substitutions in drosophila simulans. *PLoS Genet*, **7**(2):e1001302.
- 21 Tange O, 2011. Gnu parallel - the command-line power tool. *;login: The
22 USENIX Magazine*, **36**(1):42–47.