Heredity (2011) 107, 205–214
© 2011 Macmillan Publishers Limited All rights reserved 0018-067X/11
www.nature.com/hdy

ORIGINAL ARTICLE

Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection

FW Albert[1,14], E Hodges[2], JD Jensen[3], F Besnier[4,10], Z Xuan[2,11], M Rooks[2], A Bhattacharjee[5], L Brizuela[5], JM Good[1,12], RE Green[1,13], HA Burbano[1], IZ Plyusnina[6], L Trut[6], L Andersson[7], T Schöneberg[8], Ö Carlborg[4], GJ Hannon[2] and S Pääbo[1,9]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA; [3]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA; [4]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden; [5]Agilent Technologies Inc., Life Sciences Group, Santa Clara, CA, USA; [6]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; [7]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden; [8]Institute of Biochemistry, University of Leipzig, Leipzig, Germany and [9]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden

The identification of the causative genetic variants in quantitative trait loci (QTL) influencing phenotypic traits is challenging, especially in crosses between outbred strains. We have previously identified several QTL influencing tameness and aggression in a cross between two lines of wild-derived, outbred rats (*Rattus norvegicus*) selected for their behavior towards humans. Here, we use targeted sequence capture and massively parallel sequencing of all genes in the strongest QTL in the founder animals of the cross. We identify many novel sequence variants, several of which are potentially functionally relevant. The QTL contains several regions where either the tame or the aggressive founders contain no sequence variation, and two regions where alternative haplotypes are fixed between the founders. A re-analysis of the QTL signal showed that the causative site is likely to be fixed among the tame founder animals, but that several causative alleles may segregate among the aggressive founder animals. Using a formal test for the detection of positive selection, we find 10 putative positively selected regions, some of which are close to genes known to influence behavior. Together, these results show that the QTL is probably not caused by a single selected site, but may instead represent the joint effects of several sites that were targets of polygenic selection.

Heredity (2011) **107,** 205–214; doi:10.1038/hdy.2011.4; published online 9 February 2011

Keywords: high-throughput sequencing; sequence capture; positive selection; behavior; QTL mapping

Introduction

Most quantitative traits are influenced by several genes, the environment and by interactions among genes, and among genes and the environment (Mackay *et al*., 2009). To find the corresponding genes, experimental crosses have been performed in a wide range of species. These studies identify quantitative trait loci (QTL) that span large regions of the genome, each of which explains a fraction of the phenotypic variance in the trait of interest. Subsequent gene identification is usually difficult, even in experimental crosses between inbred lines of organisms where the individuals have identical genomes (Flint *et al*., 2005). Gene identification can be even more difficult in crosses between outbred lines of organisms that contain genetic variation.

We have earlier mapped several QTL influencing tameness and aggression in a cross between two lines of rats that differ drastically in their response to humans (Albert *et al*., 2009). The 'tame' rats calmly tolerate human presence and handling, whereas the 'aggressive' rats ferociously attack or flee from an approaching human hand (Naumenko *et al*., 1989; Plyusnina and Oskina, 1997; Albert *et al*., 2008). Several features distinguish this model system from other commonly used experimental rodent models. First, the two lines are derived from one wild population, caught in the Novosibirsk, Russia area in 1972. This ancestry makes it unclear whether sequence

Correspondence: Dr FW Albert, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany.
E-mail: falbert@eva.mpg.de
[10]Current address: Department for Population Genetics and Ecology, Havforskningsinstituttet, 5817 Bergen, Norway.
[11]Current address: Center for Systems Biology, and Department of Molecular and Cell Biology, University of Texas at Dallas, Richardson, TX 75080, USA.
[12]Current address: Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA.
[13]Current address: Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA.
[14]Current address: Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.

Received 6 September 2010; revised 14 November 2010; accepted 17 December 2010; published online 9 February 2011

differences among laboratory strains of rats also segregate in these animals. Second, the rats are outbred, that is, they have been maintained without mating close relatives. Hence, they contain genetic variation within each of the lines, and share alleles between the lines. Consequently, any genetic marker has to be tested experimentally to ensure that alleles segregate in an informative manner (ideally, markers will be fixed for different alleles in the two lines). Further, the causative alleles underlying tameness and aggression may themselves segregate within one or both of the two lines of rats.

Third, the difference in tameness and aggression is the result of strong artificial positive selection. In each of 64 generations, ~30% of animals were allowed to breed based on their behavioral response to humans. Behavioral change was initially rapid in both lines, leading to strong differences within 10–12 generations. The response to selection then slowed, and although selection for tameness and aggression is continued to the present, no change in mean behavior has occurred in either line in the last 10–20 generations. Positive selection affects patterns of polymorphism in several ways. The signature of selection created by the fixation of a new mutation is characterized by a deficit of variation around the target of selection, local excesses of rare alleles and of high-frequency-derived alleles, and linkage disequilibrium flanking (but not across) the site of fixation (Maynard Smith and Haigh, 1974; Tajima, 1989; Fay and Wu, 2000; Kim and Stephan, 2002; Stephan *et al*., 2006). However, given the rapid phenotypic response to the selection regime in the rats, selection will likely have acted on variation that already existed in the wild population. Patterns of genetic variation expected after selection on such ‘standing’ variation can differ dramatically from those produced after selection on new mutations. Levels of variation may not be reduced much, and there may be a loss of rare alleles and an excess of intermediate frequency alleles owing to the hitchhiking of multiple genetic backgrounds. There will also be strong linkage disequilibrium across the target of selection (Hermisson and Pennings, 2005; Przeworski *et al*., 2005; Pennings and Hermisson, 2006). In addition, if there are many sites influencing a trait, selection may result in polygenic adaptation, where many alleles undergo only small shifts in frequency, rather than going to fixation (Burke *et al*., 2010; Pritchard and Di Rienzo, 2010).

If regions showing evidence of positive selection are present in the QTL regions, they would be excellent candidates for containing the causative alleles influencing tameness and aggression. However, the two lines of rats have been maintained at small population sizes (at most ~40 breeding females and 40 breeding males) since their capture from the wild. This is expected to have reduced genetic diversity genome-wide, as well as to have increased the variance in diversity among genomic regions (see (Thornton *et al*., 2007) for a review). Consequently, any inference of past positive selection needs to be informed by the population history of the rats.

The outbred nature of the rat lines, combined with small effect sizes of the QTL for tameness and aggression (for example, here we study the strongest QTL in our earlier study, which explained 5% of phenotypic variance (Albert *et al*., 2009)) complicate fine mapping of the QTL using established strategies, such as advanced intercrosses (Darvasi and Soller, 1995) or congenic strains (Markel *et al*., 1997). We reasoned that a potential shortcut to nominating candidate genes underlying tameness and aggression could be to scan the respective QTL regions for signals of positive selection. In addition, we wanted to catalogue putative functional exonic variants in the QTL region, a strategy which has proved successful in identifying genes underlying Mendelian disease traits in humans and other species (Drogemuller *et al*., 2010; Ng *et al*., 2010). Our targeted approach complements whole-genome shotgun sequencing of pooled DNA as recently used in domestic and wild chicken (Rubin *et al*., 2010). The approach by Rubin *et al*. (2010) holds great promise, but may not always be optimal, for example, when interest is focused on one or a few genomic regions, or when analyses demand that genotypes be called for multiple individuals.

We sequenced all exons of 1475 genes across 103 Mb in the strongest QTL for tameness and aggression from our earlier study, using targeted sequence capture followed by massively parallel sequencing. Our goals were (1) to create a catalogue of potentially causative sequence variants, (2) to identify a customized set of sequence variants that can be used as genetic markers in further crossing experiments and (3) to scan the QTL region for patterns of positive selection.

Materials and methods

Animals

Rats (*Rattus norvegicus*) were obtained from a long-running selection experiment in Novosibirsk, Russia (Naumenko *et al*., 1989; Plyusnina and Oskina, 1997). Beginning from one wild-caught population, two lines of rats had been generated by selecting for the absence of aggressive behavior towards humans (the ‘tame’ line) or for increased aggression towards humans (the ‘aggressive’ line). There was no breeding between the lines, and no addition of wild animals after initiation of the lines. Matings between close relatives were avoided. Here, we study four tame and four aggressive rats (two males and two females from each line) from the sixty-fourth generation of selection. The eight rats were at most distant relatives. They are the founders of a large pedigree used previously in a QTL mapping of tameness and aggression (Albert *et al*., 2009). Hence, any linkage signals that were obtained from that pedigree are due to alleles occurring among the eight rats studied here. The animals had been killed as part of a study approved by the regional government of Saxony (TVV Nr 29/05).

Microarray design

We designed two custom Agilent arrays to capture all exons in 1475 genes annotated in the Ensembl database within the tameness QTL on chromosome 1 (‘*Tame-1*’) identified in Albert *et al*. (2009). *Tame-1* was chosen for three reasons: (1) It was the most significant QTL in our previous study, explaining 5.1% of phenotypic variance for tameness and aggression, while the second QTL had an effect size of 2.3%; (2) It is collocated and probably identical with a highly significant QTL for adrenal gland weight (a potential endophenotype for tameness and aggression) and (3) It was least affected by epistatic interactions with other loci, indicating a relatively simple underlying molecular cause. We included 20 bp on either side of each exon, as well as 1 kb of intergenic DNA directly upstream of the annotated transcription start site of each gene. Overlapping regions were grouped into 6141 joint intervals ('target regions' or 'targets' in the remainder of this paper). Throughout the text, 'targeted bases' refers to bases that were covered by at least one array probe. By this definition, 4 090 661 bases were targeted (see Supplementary Methods and Supplementary Table S1 for further details).

DNA capture and sequencing

DNA was processed into Illumina single-end sequencing libraries (Illumina, San Diego, CA, USA) following standard procedures, and hybridized to capture arrays as described (Hodges *et al*., 2009). Eluates were amplified using PCR and sequenced on Illumina GAII instruments (Illumina), with 36 and 76 bp runs. Supplementary Table S2 summarizes the sequencing runs performed for this study. Raw data has been deposited in the Sequence Read Archive under accession ERP000389 (http://www.ebi.ac.uk/ena/data/view/ERP000389).

Sequence analyses

Bases were called using AltaCyclic (Erlich *et al*., 2008) for the 36 bp lanes and Ibis (Kircher *et al*., 2009) for the 76 bp lanes. We used BWA (Li and Durbin, 2009) to map reads to the rat reference genome (version rn4). Potential PCR duplicates were removed using a custom python script (Supplementary Methods). SAMtools (Li *et al*., 2009) was used to call consensus genotypes at all targeted bases and within 250 bp of flanking sequence, retaining bases with at least eightfold coverage, a minimum mean phred-scaled mapping quality of 20, a phred scaled consensus quality of at least 30 and at least 10 bp away from insertions/deletion (indels). A custom script was used to call indel genotypes (Supplementary Methods). Genome annotation and effects of sequence variants on transcripts were extracted from Ensembl build 57 (Hubbard *et al*., 2009). Candidate genes affecting behavior were extracted from the Rat Genome Database (Twigger *et al*., 2007) by searching for genes with Gene Ontology annotations, including the terms 'aggression', 'anxiety' and 'fear'. Vomeronasal receptor genes were annotated based on Ensembl's 'gene description' annotation.

Flexible intercross analysis (FIA)

We used FIA (Rönnegard *et al*., 2008) to assess whether the causal QTL alleles were fixed among founders of each line. Instead of assuming fixation of alternative alleles, FIA uses a variance component approach to model a correlation between the phenotypic effects of alleles coming from the founders of each line. All possible degrees of allelic differentiation, from complete fixation (high correlation between allelic affects within a line) to complete segregation (no correlation between allelic effects), can be analyzed. FIA was fitted to phenotypes and to identity-by-descent matrices between animals in the mapping pedigree generated previously (Albert *et al*., 2009). Following Rönnegard *et al*. (2008), we tested whether the null hypothesis of fixation could be rejected in favor of the alternative hypothesis of allelic segregation within the founders. If segregation was more likely, the degree of segregation was estimated as described (Rönnegard *et al*., 2008).

Patterns of sequence variation

Only sites where all eight rats had high-quality consensus genotypes were used in these analyses. Allele frequency difference was calculated in a sliding window of 20 adjacent single-nucleotide variants (SNVs) and a step size of 5 SNVs. To calculate nucleotide diversity (π), we extracted consensus sequences for each rat and used the program 'compute' based on the libseq code library (Thornton, 2003). For the sliding window analysis, π was calculated in windows of 20 kb of consensus sequence using a step size of 1 kb. Each window, therefore, spans a different genomic length, but always analyses the same number of bases. When calculating correlations between π and coverage, mean values from non-overlapping windows were used.

Demographic estimation

Demographic estimation was performed using the program δaδi (Gutenkunst *et al*., 2009). Briefly, by letting θ correspond to the parameters of a demographic model, one wishes to estimate from the observed frequency spectrum (denoted as $S[d_i, d_j,\ldots])$, and assuming no linkage between polymorphisms, each entry is an independent Poisson variable, with mean $M[d_i, d_j,\ldots]$. A likelihood function is then constructed as:

$$L(\theta|S) = \prod_{i=0\ldots P} \prod_{d_i=0\ldots n_i} \frac{e^{-M[d_i,d_j,\ldots d_p]} M[d_i, d_j, \ldots, d_p]^{S[d_i,d_j,\ldots d_p]}}{S[d_i, d_j, \ldots, d_p]!}$$

Thus, using a diffusion approach, the expected allele frequency spectrum M is calculated under a particular demographic model. The similarity between M and the observed spectrum, S, is maximized over the values of θ. Full code and documentation are available at: http://code.google.com/p/dadi/. To simulate individual demographic models, the program ms was used (Hudson, 2002).

Testing the significance of regions of reduced nucleotide diversity

The program MaCS (Chen *et al*., 2009) was used to perform coalescent simulations of the entire 103-Mb region. We simulated 1000 replicates of a model of the rats' demography, incorporating available knowledge on the history of the selection lines over the past 64 generations (Supplementary Methods), and using recombination rates estimated from the pedigree founded by the rats sequenced here (Albert *et al*., 2009). In each run of the simulations, we sampled eight chromosomes from each of the two populations, extracted the exact base positions that were analyzed in the real sequence data, and calculated nucleotide diversity per line in sliding windows. In each set of replicates, we recorded the regions where the samples from one or both of the two lines had no nucleotide diversity, and compared the real data to this simulated neutral distribution.

Test for positive selection

We used the maximized composite likelihood surface test (Nielsen *et al*., 2005), a multi-locus test for positive selection based on a modification of the CLR test (Kim and Stephan, 2002). The approach uses the background site frequency spectrum obtained from the data to test for positive selection. A grid of locations was created over the QTL region, and for each grid location the maximum composite likelihood ratio of the hypothesis of a sweep was compared with the null hypothesis of no sweep. Coalescent simulations of the estimated bottleneck model were used to determine the significance of the test statistic. We implemented an approach to test models of positive selection on both new mutations as well as on standing variation. The respective patterns of polymorphism that are expected under these two selection scenarios were generated by simulating sweeps from 'new' (initial frequency <0.05), and from 'standing' (initial frequencies drawn from a distribution of frequencies between 0.05 and 0.50) variants. The parametric approach used is described by equation (6) of Nielsen *et al*. (2005). The grid-size parameter was set at 10^4. The program for calculating this statistic is available for download at: http://fisher.berkeley.edu/cteg/software.html. To calculate *P*-values, demographic simulations incorporating selection were performed using SFSCODE (Hernandez, 2008). Note that these simulations were separate from those for the QTL region as a whole, as described in the preceding paragraph.

Results

Array capture performance

Across the individual capture experiments, 33–60% (mean = 45.4%) of mapped reads fell into the target regions (Supplementary Table S2), corresponding to an average 340-fold enrichment relative to random shotgun sequencing. In all rats, >99% of targets were hit by at least one read, and >97% of targeted bases were covered by >8 reads (Figure 1). After filtering for high consensus quality, genotype calls were available in all eight rats at 3 938 035 of targeted bases (96%). Mean ± s.d. base coverage ranged from 45 ± 21 to 81 ± 25 in target regions. The mean pairwise correlation of coverage of targeted bases was 0.78 ($P < 2e\text{-}16$), indicating systematic preferences of the capture and/ or the Illumina sequencing technology towards certain sequence features, as noted in other studies (Harismendy *et al*., 2009; Ng *et al*., 2009).

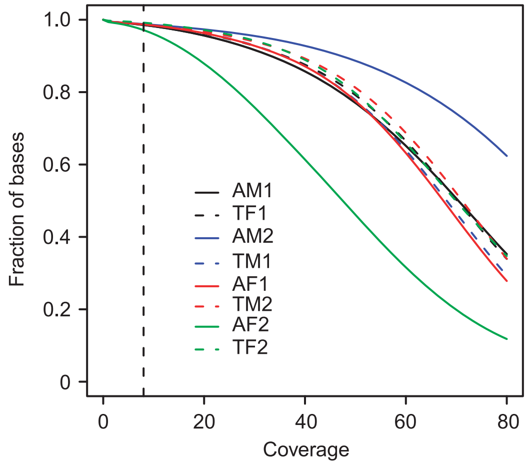

Figure 1 Inverse cumulative coverage plots. For each rat, the fraction of targeted bases that match or exceed a given depth of coverage are plotted. The dashed vertical line indicates the minimum coverage used for calling consensus sequences. Rat IDs are coded as: A, aggressive; T, tame; F, female; M, male.

Out of the 6141 targets, one was never hit in any rat. This target contains no exons annotated in Ensembl, but the sequence is similar to two coding exons in a mouse vomeronasal receptor gene (*Vmn2r65*). We found no targets that were hit only in tame or only in aggressive founders.

Single-nucleotide variants

As sequence coverage typically extended for some distance beyond the actual targets, we included regions extending 250 bp on either side of each target when calling sequence variants. In 7 858 329 analyzed base pairs, we identified 12,586 biallelic SNVs segregating among the eight founder rats (Supplementary File S1). Of these, 10 434 (83%) were novel variants not previously annotated in the dbSNP database (build 130). At a further 1861 positions (1221; 66% novel), all eight rats had the same homozygous genotype, but differed from the genome. Although these sites did have significantly lower sequence coverage than the remaining sites (Wilcoxon rank test, $P = 0.003$), they still had substantial coverage (median across rats = 33 vs 47 for other sites), making it unlikely that many of them are artifacts due to missed heterozygotes. We excluded 20 tri-allelic SNVs from further analyses.

Although introns were not *per se* targeted by our arrays, sequence coverage was typically high enough directly adjacent to exons to call intronic variants close to the exon boundaries. The majority (79%) of the SNVs were situated in intergenic regions or introns (Table 1). However, there were also 2643 SNVs within coding regions. Of these, 1168 (44%) were non-synonymous variants that alter the amino acid sequence of the protein product of their gene. This fraction of non-synonymous SNVs is slightly lower than that observed in humans (52–53%) (1000 Genomes Project Consortium, 2010; Li *et al*., 2010). Another 13 coding SNVs in 12 genes were annotated as nonsense mutations, leading to premature termination of their proteins.

Considering only sites where all rats had consensus calls, 369 SNVs were 'fixed' for alternative alleles between the founder animals of the two lines (all four tame rats were homozygous for one allele, all four

Table 1 Sequence variants in the QTL region

Sequence type	*Number of SNVs*	*Number of indels*
Intergenic	4558	1072
Intronic	4327	1008
Coding sequence		
Synonymous	1462	10
Non-synonymous	1168	19
STOP/frameshift	13	103
Noncoding transcripts	141	16
3′ UTR	536	152
5′ UTR	173	28
Splice sites[a]	9 (199)	5 (37)
Total	12 586	2450

Abbreviations: QTL, quantitative trait loci; SNVs, single-nucleotide variants; UTR, untranslated region.
The table presents the number of all variants that segregate among the founder rats, irrespective of the segregation pattern.
[a]In first and last 2 bp in intron (in first/last 3 bp in exon or 3–8 bp in intron).

aggressive rats homozygous for the other). Coverage was not significantly different for the fixed compared to non-fixed sites (medians = 64 vs 66, $P = 0.08$). Of fixed sites, 41 were synonymous, residing in 18 genes, while 43 were non-synonymous, affecting 17 genes (Supplementary File S2). The number of fixed SNVs is an underestimate, as it only considers positions where all eight rats had consensus genotype calls. When sites with missing data were included, there were 731 potentially fixed SNVs. There was no fixed stop mutation.

Insertions/Deletions

We discovered 2450 indels, mostly in introns or intergenic sequence (2080, 85%, Table 1; Supplementary File S3 contains a full list). However, 103 indels were predicted to induce frameshifts in the coding sequence of genes. These frameshift indels generally fell into five categories. First, 12 indels fell into predicted genes with unknown function, making it unclear whether these are real genes. Second, 20 indels resided in vomeronasal or olfactory receptor genes, some of them predicted (see Supplementary Discussion of this group of genes). Third, 31 indels were observed only in single rats, making them unlikely candidates for causing tameness and aggression. Fourth, 26 coding indels affected genome bases that, according to the genome annotation in the UCSC browser, lead to frameshifts of the conserved protein sequence. The indels found in the rats were either deletions of the offending genome bases or insertions next to them, in either case restoring the evolutionarily conserved reading frame. These indels were usually homozygous in all eight rats. Where they were not observed in a given rat, consensus quality was too low to have confidence in their absence. We suggest that this fourth class of ‘indels’ in fact reflect errors in the rat reference genome.

Of the remaining 14 frameshift indels, 7 occurred in both the tame and the aggressive rats, making them unlikely candidates for causing tameness. The remaining seven occurred in six genes (one each in *Sars2*, *Atp4a*, *Commd9*, *RGD1561206* and *Otud7*; two in *Mphosph10*), and may be candidates for causing tameness. We caution that some of the frameshift indels (and some of the stop mutations described above) may in fact be annotation artifacts.

Segregation analysis of QTL alleles

Many of the detected variants were not fixed between the founder animals, raising the question whether the causative alleles themselves are fixed. To test this, we used FIA (Rönnegard *et al.*, 2008) to test whether, at the position of the QTL peak, fixation or segregation of causative alleles among the founder animals is more likely. We performed this test for tameness and for the weight of the adrenal glands, a trait influenced by a major QTL coincident with the tameness QTL (Albert *et al.*, 2009). For adrenal gland weight, fixation of alternative alleles could not be rejected (likelihood ratio test with 1 d.f.: LR = 0.3, $P = 0.4$). For tameness, however, segregation of alleles among founders was more likely than fixation (LR = 3.46, $P = 0.0085$). Line-specific estimates of correlation of allele effects indicated that there was segregation only among the aggressive founders (estimated $r_{\text{aggressive}} < 0.001$), while the tame founders were likely to be fixed for one allele (estimated $r_{\text{tame}} = 1$).

Table 2 Nucleotide diversity in the QTL region

Sequence type	*Tame rats*	*Aggressive rats*
Overall	5.6	5.2
Intergenic	6.9	6.3
Intronic	5.8	5.5
Coding sequence overall	4.4	4.2
First and second codon	2.6	2.7
Third codon	7.8	7.3
3′ UTR	5.5	4.6
5′ UTR	6.3	4.8

Abbreviations: QTL, quantitative trait loci; UTR, untranslated region.
Values are nucleotide diversity (π) in units of 10^{-4} per base pair.

Patterns of genetic variation and rat population history

Nucleotide diversity (π) was calculated using the ~4.6 Mb of sequence, where data was available from all eight rats (Table 2). Overall, π was 0.00056 in the tame and 0.00052 in the aggressive rats. The level of genetic diversity in wild rats in the Novosibirsk area is not known, precluding an estimate of how much diversity was reduced in the two lines during the course of selection. However, using the likelihood based approach δaδi (Gutenkunst *et al.*, 2009), we estimated a very severe reduction in effective population size (to 0.00082 of the ancestral population size; 95% confidence interval: 0.0001–0.0017), in the very recent past (0.00008 $4N$ generations in the past; 95% confidence interval: 0.00005–0.00014). These results are strikingly consistent with the known population history. Given that the population size at which the lines were kept after capture from the wild is known (Supplementary Methods), this result implies a wild effective population size of roughly $N_e = 70\,000$.

Search for positive selection

We searched for regions that may have experienced positive selection by two approaches, both of which explicitly take into account the population history of the rats. First, we asked whether regions existed in the QTL where the founders are fixed for alternative haplotypes (Supplementary Figure S1). There were three such regions: a short stretch at ~62 Mb and two wider regions that were separated only by two neighboring SNVs segregating within the tame founders (at 118 669 118 and 118 669 144 bp). When considered together, these two regions extended for ~2.6 Mb (116 710 725–119 277 527 bp; Figure 2a). Next, we searched for regions where the tame or the aggressive founders had no genetic variation (Supplementary Figure S1). There was a significant correlation ($r = 0.38$, $P = 1.6\text{e-}9$) between π in the tame and the aggressive founders along the chromosome, probably due to the shared ancestry of the tame and the aggressive rats from the same wild population. Coverage was positively correlated with π in the aggressive founders ($r = 0.3$, $P = 2.5\text{e-}6$), but negatively correlated in the tame founders ($r = -0.2$, $P = 0.004$). The correlation in π between tame and aggressive founders was not an artifact of joint variation in coverage (correlation after correcting for coverage effects on π: $r = 0.37$, $P = 8.5\text{e-}9$). There were several regions where the founders from one but not the other line had no

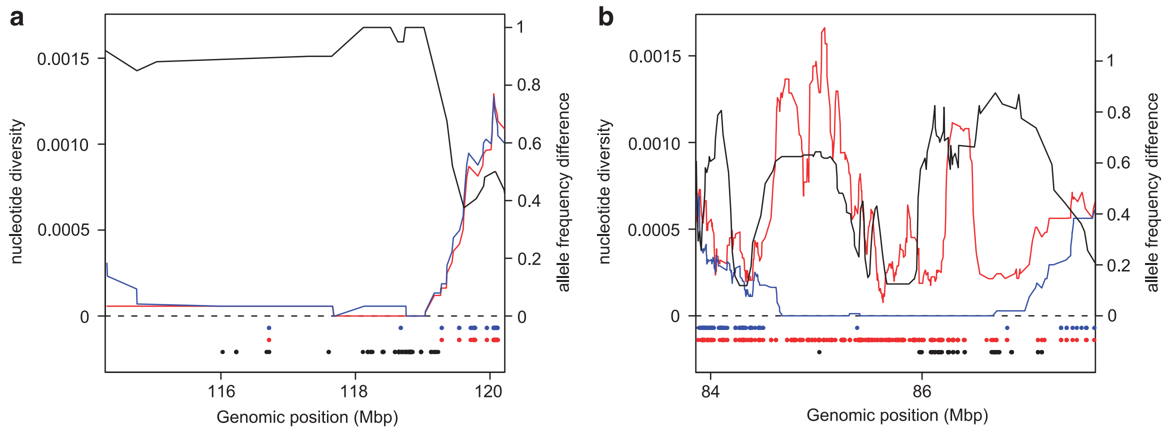

Figure 2 Examples of sequence variation in the QTL region. (**a**) A region where tame and aggressive founder rats are fixed for a continuous stretch of sequence. (**b**) A region where tame, but not aggressive founders have no sequence variation. Allele frequency difference is plotted in black, π in blue (tame rats) and red (aggressive rats). Blue (red) dots indicate SNVs segregating among tame (aggressive) rats; black dots indicate fixed SNVs.

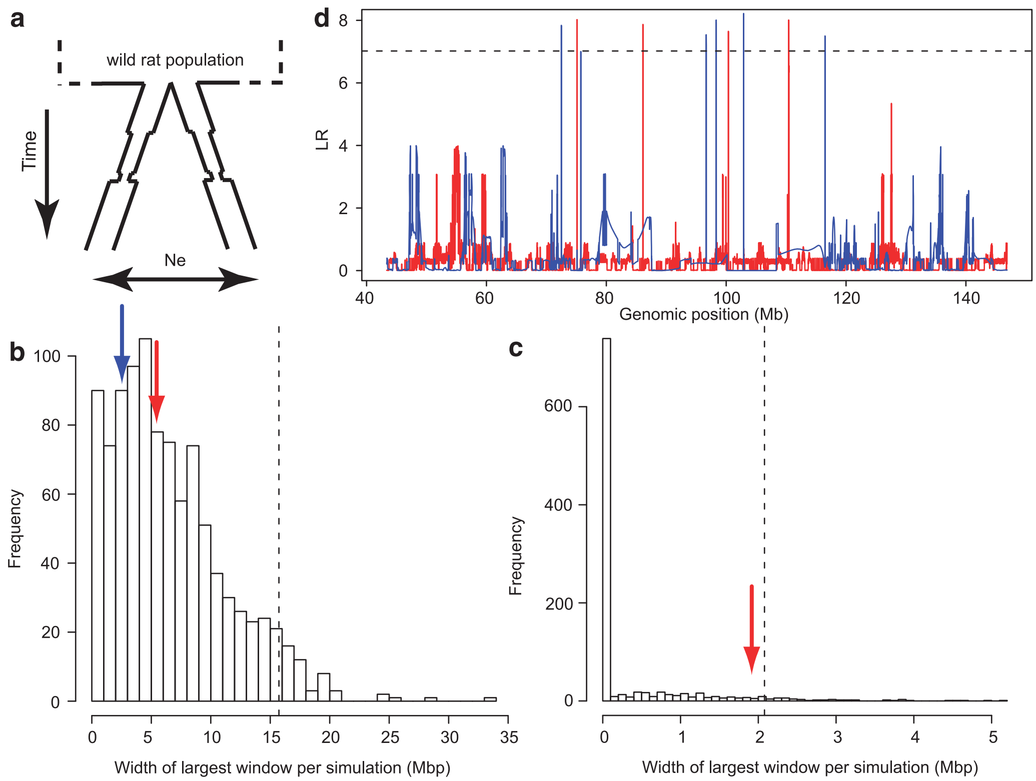

Figure 3 Scan for positive selection. (**a**) Illustration of the demographic model used in the selection scans. (**b, c**) Distributions of the width of regions with no observed nucleotide diversity ('dips') in 1000 simulations of the rat population history without any selection. The dashed lines show the 95% cutoff. (**b**) Dips in one of the two lines. Blue (red) arrow: most extreme dip in real data from tame (aggressive) rats. (**c**) Regions with fixed differences. Arrow: most extreme region in real data. (**d**) Results from the maximized composite likelihood surface-based scan for positive selection. Likelihood ratios between the respective best selection model and the background site frequency spectrum in tame (blue) and aggressive (red) rats are plotted. The dashed line is the Bonferroni-corrected significance threshold.

genetic diversity ($\pi = 0$, Figure 2b). These line-specific dips in diversity ranged in size from 74 kb to more than 5 Mb (Supplementary Table S3). The dips were not an artifact of low coverage in those regions (Supplementary Figure S2).

To gauge whether the regions of reduced diversity and of alternative fixed haplotypes could be explained by population history alone, we performed 1000 simulations of the whole QTL region without any selection (Figure 3a). Although regions with no nucleotide diversity did appear in the simulations, both the number ($P = 0.002$) and the total width of such dips ($P = 0.046$) in the aggressive founders were significantly larger than those in the simulations. In the tame founders, the

Table 3 Regions in the QTL inferred to be under positive selection

Position	*Genes of interest*[a]	*Significant population*	*Selection model*	P-*value*[b]
72.5 Mb		Tame	New	2.3×10^{-5}
75.1 Mb		Aggressive	Standing	4.5×10^{-5}
75.7 Mb		Tame	Standing	7.4×10^{-4}
86.1 Mb		Aggressive	New	3.6×10^{-5}
96.7 Mb	*Tph1*	Tame	New	8.7×10^{-5}
98.3 Mb	*Tph1*	Tame	New	7.0×10^{-5}
100.4 Mb		Aggressive	Standing	1.2×10^{-5}
102.9 Mb		Tame	Standing	9.0×10^{-5}
110.4 Mb	*Gabra5*	Aggressive	New	9.2×10^{-5}
116.5 Mb		Tame	Standing	2.3×10^{-5}

Abbreviation: QTL, quantitative trait loci.
[a]See text for details.
[b]After demographic and multiple test corrections.

number of dips was significantly larger than expected ($P < 0.001$), while the width of the dips was not unusual ($P = 0.18$). The regions of alternative fixed haplotypes were also not unusual compared with the simulations, both with respect to their number ($P = 0.07$) and total width ($P = 0.08$). When examining individual dips, we found that none exceeded significance (Figure 3b). Similarly, the regions with alternative fixed haplotypes were not unusual compared with the simulations (Figure 3c).

As a second approach to search for positive selection, we implemented a maximum likelihood approach (Nielsen *et al*., 2005). We used this framework to identify regions that are consistent with selection on new mutations and regions that are consistent with selection on standing variation (see Methods). We identified a number of significant outlier regions (Figure 3d and Table 3): under a model of selection on new mutations, five regions were significant after both multiple-test (Bonferroni correction) and demographic correction; under a model of selection on standing variation, another five regions were significant.

Candidate genes for affecting tameness and aggression

In RGD, there were four genes with annotations for aggression (Tryptophan hydroxylase 1, *Tph1*; (Haavik *et al*., 2008)), fear (γ-aminobutyric acid receptor subunit alpha 5, *Gabra5*; (Otani *et al*., 2005)) or anxiety (Protein kinase C-gamma, *Prkcc*; (Bowers *et al*., 2000) and plasminogen activator urokinase receptor; *Plaur* (Powell *et al*., 2003)). Two of these four genes are situated close to regions inferred to have experienced positive selection: *Tph1* at 97.7 Mb and *Gabra5* at 108.8 Mb. The regions around *Tph1* and *Gabra5* contained 18 and 12 SNVs, respectively, as well as two and one intronic or intergenic indels. Neither gene contained a non-synonymous variant.

Finally, the QTL contained a high number of sequence variants in vomeronasal receptor genes. For example, these genes accounted for 20 of 103 coding frameshifting indels, and for 25 of 43 fixed non-synonymous SNVs.

Discussion

In this study, we used microarray-based DNA capture and high-throughput sequencing to generate an exhaustive catalogue of exonic sequence variation in a QTL region for tameness and aggression that was previously identified in a cross between two outbred, wild-derived strains of rats. This catalogue will be useful in several ways. The sequence variants with fixed or nearly fixed alternative alleles in the founder animals will serve as genetic markers in future fine-mapping studies. Many of these variants are novel, and could, hence, not have been obtained from existing databases. Further, we identified many potentially functional sequence variants. Thus, as fine-mapping progresses, we will know which genes contain interesting variants. It is important to emphasize that, as we have only sequenced the founder animals of our earlier cross, variants that are fixed between these animals may not be fixed in the tame and aggressive selection lines as a whole. However, as the QTL signal can be caused only by variation that exists in these founders, our catalogue is essentially complete at the targeted sites, with regards to the QTL under study.

It is worth noting that 17% of the variants found in this study were previously seen in rat laboratory strains, in spite of the fact that laboratory strains are not yet fully sequenced in the genomic regions we studied here. Hence, although the Novosibirsk rats and common laboratory strains were derived many decades apart and at different geographic locations, they have sampled a shared set of polymorphisms from the wild rat population. Shared polymorphism between the tame and the aggressive rats and laboratory strains opens the possibility that the causative alleles influencing tameness and aggression might also segregate among laboratory populations. Thus, a promising approach to fine mapping could be to test for association for tameness and aggression in laboratory strains or other wild-derived populations of rats.

The causative allele(s) underlying the QTL must be present among the founder animals of the mapping pedigree. If a causative allele resides in coding exons, untranslated regions or intergenic regions upstream of genes, it is likely to have been captured by our approach. Direct identification of the causative site(s) is difficult, but we can, nevertheless, prioritize the identified variants by several means, including segregation pattern and location close to putative positively selected regions.

An intuitive expectation is that the causative locus should be fixed for alternative alleles in the tame and the aggressive founder animals. A fraction of the variants do indeed show such a segregation pattern. However, a reanalysis of the mapping data generated in our previous study (Albert *et al*., 2009), suggested that although the causative allele is likely fixed in the tame founders, there may be segregation of two or more alleles in the aggressive founders. Including sites with such a segregation pattern adds substantially to the list of variants that are candidates for underlying the QTL. In particular, this pattern is consistent with regions of the QTL where the tame founders are fixed for the same allele and where the aggressive founders segregate, but is inconsistent with regions that show the opposite pattern. Further, the fact that the causative QTL allele may not be fixed among the aggressive founders illustrates that tameness and aggression, although semantic opposites, may not necessarily be direct physiological opposites. It may be the case that the QTL studied here was caused by a site that was selected and driven to fixation in the tame, but not in the aggressive rats.

Both the tame and the aggressive rats have experienced strong positive selection, while being maintained as outbred populations. We found that the rats retain some of the genetic variation that existed in the wild, allowing a scan of the QTL region for patterns indicative of positive selection. Given that there was an immediate and strong response to selection, a model of selection acting on variation that was already segregating in the wild is intuitively appealing. We began by estimating the reduction in population size, in order to understand the expected background patterns of variation. Consistent with historical breeding records, we estimate a reduction that was both extremely strong and extremely recent.

We used this estimated demographic model to assess whether demographic forces alone, without the action of positive selection, could explain the regions with no observed nucleotide diversity. We found that despite their sometimes considerable size of up to several Mb, no individual region was significantly different from the patterns observed in simulations without any selection. Hence, although the high number of regions with no observed nucleotide diversity indicates that positive selection may have caused at least some of these regions, we cannot with confidence single out individual regions as more likely candidates than others. We note that the software used to estimate the demographic model can only fit a neutral model without selection. Consequently, if there is strong selection and a strong population size reduction in the data, the procedure will probably fit a stronger reduction to account for the effects of both processes. As the simulations are, thus, likely to be conservative with regards to selection, the fact that no individual region was significant might be a by-product of this effect.

Using more formal tests for positive selection that incorporate the site frequency spectrum, we identified five regions under a model of selection on new mutations and another five regions under a model of selection on standing variation. Note that the regions corresponding to 'new' mutations are equally consistent with selection on previously existing mutations at low frequency (Przeworski *et al*., 2005). The fact that there are several regions with evidence of positive selection in both lines of rats raises the question whether all these signals correspond to selection for tameness and aggression. For example, some of the selected sites may represent adaptations of both lines of rats to their common novel laboratory environment. If all selection signals are, indeed, due to selection for tameness and aggression, the QTL may be a composite signal caused by multiple linked sites, each probably with effects smaller than the 5.1% for the QTL as a whole. Tameness and aggression may consequently be an example of traits that can be rapidly altered by polygenic adaptation, as has recently been demonstrated in an artificial selection experiment for accelerated development in *Drosophila* (Burke *et al*., 2010). A full understanding of how positive selection has acted in these rats will require the study of additional regions of the genome.

Intriguingly, two of the ten selected regions are close to *a priori* candidate genes. *Tph1* is one of two genes encoding the protein tryptophan hydroxylase (TPH), which plays an important role in the production of the brain neurotransmitter serotonin. Serotonin has been linked to several behaviors, including aggression (Cases *et al*., 1995). Further, tame rats have lower levels of serotonin in their brains than the aggressive rats (Albert *et al*., 2008). However, *Tph1* encodes for a version of TPH that is expressed in the gut, pineal gland, spleen and thymus, rather than in the raphe nuclei, which provide serotonin to the rest of the brain (Walther and Bader, 2003). Serotonin serves several physiological functions in addition to being a neurotransmitter (Walther and Bader, 2003), raising interesting questions about how *Tph1* may potentially influence tameness and aggression. *Gabra5* encodes a subunit of the A-type receptor for γ-aminobutyric acid, which is an abundant inhibitory neurotransmitter in the brain, influencing many aspects of behavior (Kandel *et al*., 2000). Tame rats have lower levels of γ-aminobutyric acid in their brains than the aggressive rats (Albert *et al*., 2008). Given γ-aminobutyric acid's wide-reaching effects, allelic variation at a component of one of its receptors may well affect tameness and aggression. As neither Tph1 nor Gabra5 had amino acid differences between the tame and the aggressive rats, the causal sequence variant(s) may influence the level of expression of these genes in adult rats or during development. We caution that several other genes, in addition to *Tph1* and *Gabra5*, are located in the vicinity of the selection signals. Hence, it is presently not possible to tell whether selection has, in fact, targeted Tph1 and/ or Gabra5. Functional studies will be needed to clarify whether variation at these two genes, in fact, causes tameness and aggression.

Many sequence variants in the QTL region affect vomeronasal receptor genes (see also Supplementary Text 1). Copy number variation in chemosensory receptor genes among humans has been argued to be evolutionarily neutral (Nozawa *et al*., 2007), raising the possibility that the sequence variants in vomeronasal receptor genes identified here may not be functionally important. However, given the strong influence of the pheromone system on rodent behavior, including aggression (Dulac and Torello, 2003), a tantalizing alternative is that one or several of these variants may influence tameness and aggression.

In summary, we have used array-based DNA capture and high-throughput sequencing to obtain a detailed view of sequence variation in a QTL region for tameness and aggression in rats that had been artificially selected for these behaviors. The sequence content of this QTL is complex, with many individual sequence variants, considerable variation in segregation patterns along the region and several putative targets of positive selection, some of which are close to genes that are promising candidates for underlying the QTL. Functional tests of these genes, as well as further genetic fine-mapping studies, will ultimately unravel which genes cause tameness and aggression in these unique lines of animals. Finally, the fact that there are several signals of selection inside the QTL region changes our view of the genetic basis of tameness and aggression. Rather than being caused by a single causative site, this QTL may be the result of several linked causative sites. Thus, selection for tameness and aggression in these rats may have acted through polygenic selection affecting many sites, rather than through a few selective sweeps with large phenotypic effects.

Conflict of interest

AB and LB are employees of Agilent Technologies Inc. Agilent supplies arrays that can be used for targeted sequencing as described in this paper.

Acknowledgements

We are grateful to Lars Rönnegard and Elizabeth Blankenhorn for insightful discussions, and to Martin Kircher for help with sequence analysis, and to the reviewers for constructive comments on the manuscript. This work was supported by funds from the Max Planck Society, the Swedish Foundation for Strategic Research the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, a EURYI Award to ÖC, and HHMI and NIH grant #P01 CA013106-39.

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Albert FW, Carlborg Ö, Plyusnina IZ, Besnier F, Hedwig D, Lautenschläger S *et al*. (2009). Genetic architecture of tameness in a rat model of animal domestication. *Genetics* **182**: 541–554.

Albert FW, Shchepina O, Winter C, Rmpler H, Teupser D, Palme R *et al*. (2008). Phenotypic differences in behavior, physiology and neurochemistry between rats selected for tameness and for defensive aggression towards humans. *Horm Behav* **53**: 413–421.

Bowers BJ, Collins AC, Tritto T, Wehner JM (2000). Mice lacking PKC gamma exhibit decreased anxiety. *Behav Genet* **30**: 111–121.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010). Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature* **467**: 587–590. U111.

Cases O, Seif I, Grimsby J, Gaspar P, Chen K, Pournin S *et al*. (1995). Aggressive behavior and altered amounts of brain serotonin and Norephinephrine in mice lacking MAOA. *Science* **268**: 1763–1766.

Chen GK, Marjoram P, Wall JD (2009). Fast and flexible simulation of DNA sequence data. *Genome Res* **19**: 136–142.

Darvasi A, Soller M (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**: 1199–1207.

Drogemuller C, Tetens J, Sigurdsson S, Gentile A, Testoni S, Lindblad-Toh K *et al*. (2010). Identification of the bovine arachnomelia mutation by massively parallel sequencing implicates Sulfite Oxidase (SUOX) in bone development. *PLoS Genetics* **6**: e1001079.

Dulac C, Torello AT (2003). Molecular detection of pheromone signals in mammals: from genes to behaviour. *Nat Rev Neurosci* **4**: 551–562.

Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008). Alta-Cyclic: a selfoptimizing base caller for next-generation sequencing. *Nat Methods* **5**: 679–682.

Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Flint J, Valdar W, Shifman S, Mott R (2005). Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Gent* **6**: 271–286.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: 11.

Haavik J, Blau N, Thony B (2008). Mutations in human monoamine-related neurotransmitter pathway genes. *Hum Mutat* **29**: 891–902.

Harismendy O, Ng PC, Strausberg RL, Wang XY, Stockwell TB, Beeson KY *et al*. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: 13.

Hermisson J, Pennings PS (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.

Hernandez RD (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786–2787.

Hodges E, Rooks M, Xuan ZY, Bhattacharjee A, Gordon DB, Brizuela L *et al*. (2009). Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**: 960–974.

Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E *et al*. (2009). Ensembl 2009. *Nucleic Acids Res* **37**: D690–D697.

Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

Kandel ER, Schwartz JH, Jessell TM (eds). (2000). *Principles of Neural Science*. McGraw-Hill: New York.

Kim Y, Stephan W (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765–777.

Kircher M, Stenzel U, Kelso J (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**: R83.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al*. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H *et al*. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**: 969–972.

Mackay TFC, Stone EA, Ayroles JF (2009). The genetics of quantitative traits: challenges and prospects. *Nat Rev Gent* **10**: 565–577.

Markel P, Shu P, Ebeling C, Carlson GA, Nagle DL, Smutko JS *et al*. (1997). Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nat Genet* **17**: 280–284.

Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favorable gene. *Genetical Research* **23**: 23–35.

Naumenko EV, Popova NK, Nikulina EM, Dygalo NN, Shishkina GT, Borodin PM *et al*. (1989). Behavior, adrenocortical activity, and brain monoamines in norway rats selected for reduced aggressiveness towards man. *Pharmacol Biochem Behav* **33**: 85–91.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM *et al*. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**: 30–U41.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C *et al*. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276. U153.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005). Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566–1575.

Nozawa M, Kawahara Y, Nei M (2007). Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Natl Acad Sci USA* **104**: 20421–20426.

Otani K, Ujike H, Tanaka Y, Morita Y, Katsu T, Nomura A *et al*. (2005). The GABA type A receptor alpha 5 subunit gene is associated with bipolar I disorder. *Neurosci Lett* **381**: 108–113.

Pennings PS, Hermisson J (2006). Soft sweeps II-molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* **23**: 1076–1084.

Plyusnina IZ, Oskina I (1997). Behavioral and adrenocortical responses to open-field test in rats selected for reduced aggressiveness toward humans. *Physiol Behav* **61**: 381–385.

Powell EM, Campbell DB, Stanwood GD, Davis C, Noebels JL, Levitt P (2003). Genetic disruption of cortical interneuron development causes region- and GABA cell type-specific deficits, epilepsy, and behavioral dysfunction. *J Neurosci* **23**: 622–631.

Pritchard JK, Di Rienzo A (2010). Adaptation—not by sweeps alone. *Nat Rev Genet* **11**: 665–667.

Przeworski M, Coop G, Wall JD (2005). The signature of positive selection on standing genetic variation. *Evolution* **59**: 2312–2323.

Rönnegard L, Besnier F, Carlborg O (2008). An improved method for quantitative trait loci detection and identification of within-line segregation in F-2 intercross designs. *Genetics* **178**: 2315–2326.

Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT *et al*. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.

Stephan W, Song YS, Langley CH (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**: 2647–2663.

Tajima F (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**: 585–595.

Thornton K (2003). libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**: 2325–2327.

Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007). Progress and prospects in mapping recent selection in the genome. *Heredity* **98**: 340–348.

Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ (2007). The rat genome database, update 2007—easing the path from disease to data and back again. *Nucleic Acids Res* **35**: D658–D662.

Walther DJ, Bader M (2003). A unique central tryptophan hydroxylase isoform. *Biochem Pharmacol* **66**: 1673–1680.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)