Whole-Genome Sequencing Reveals Genetic Variation in the Asian House Rat

Huajing Teng,*,†,‡ Yaohua Zhang,* Chengmin Shi,*,§ Fengbiao Mao,‡ Lingling Hou,‡ Hongling Guo,*,† Zhongsheng Sun,‡ and Jianxu Zhang*,[1]
*The State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, 100101 Beijing, China, †University of Chinese Academy of Sciences, 100049 Beijing, China, ‡ Beijing Institutes of Life Science, Chinese Academy of Sciences, 100101 Beijing, China, and §Beijing Institute of Genomics, Chinese Academy of Sciences, 100101 Beijing, China

ABSTRACT Whole-genome sequencing of wild-derived rat species can provide novel genomic resources, which may help decipher the genetics underlying complex phenotypes. As a notorious pest, reservoir of human pathogens, and colonizer, the Asian house rat, *Rattus tanezumi*, is successfully adapted to its habitat. However, little is known regarding genetic variation in this species. In this study, we identified over 41,000,000 single-nucleotide polymorphisms, plus insertions and deletions, through whole-genome sequencing and bioinformatics analyses. Moreover, we identified over 12,000 structural variants, including 143 chromosomal inversions. Further functional analyses revealed several fixed nonsense mutations associated with infection and immunity-related adaptations, and a number of fixed missense mutations that may be related to anticoagulant resistance. A genome-wide scan for loci under selection identified various genes related to neural activity. Our whole-genome sequencing data provide a genomic resource for future genetic studies of the Asian house rat species and have the potential to facilitate understanding of the molecular adaptations of rats to their ecological niches.

KEYWORDS
genetic landscape
next-generation sequencing
Rattus tanezumi
single-nucleotide polymorphisms
structural variations

Genetic variations of the laboratory rat have served as a valuable resource in biomedical and behavioral research for nearly 200 yr (Baker *et al.* 1979; Jacob 1999). However, the extant strains of laboratory rat originate from limited *Rattus norvegicus* founder populations (Baker *et al.* 1979; Krinke 2000; Smits *et al.* 2005; Saar *et al.* 2008). Extensive genetic variation is required to facilitate correlation of genotypes with complex behavioral or ecologically relevant traits. Wild-derived rat strains or species can provide novel genome resources to decipher the genetic mechanisms underlying such complex phenotypes (Smits *et al.* 2005).

As a diploid species of rodent in the genus *Rattus*, the Asian house rat (AHR, *R. tanezumi*; 2n = 42) (Hamta *et al.* 2006; Badenhorst *et al.* 2011) is widely distributed in Asia, and has recently invaded the South Pacific, African countries, and the United States (Wilson and Reeder 2005; Bastos *et al.* 2011; Lack *et al.* 2012; Conroy *et al.* 2013; Nakayama *et al.* 2013). As an agricultural pest (Singleton *et al.* 2007; Huang *et al.* 2013), the AHR demonstrates superior adaptive potential, compared to other rodents, with higher resistance to commonly used rodenticides (Zheng 1981; Tian *et al.* 2006; Huang *et al.* 2012). The AHR can also serve as a reservoir of a wide range of pathogens associated with human disease (Plyusnina *et al.* 2009; Jonsson *et al.* 2010; Yin *et al.* 2011; Kocher *et al.* 2015). However, although the AHR represents a tremendous threat to agriculture and human health, little is known about the genetic basis underlying its remarkable biological traits. A comprehensive genetic inventory of the AHR will not only be essential for the development of effective pest control programs, but will also advance our understanding of the diversity of rodent-borne zoonosis.

Rapid advances in next-generation sequencing (NGS) technology and reverse ecology hold great promise for deciphering the genetic basis underlying the functional variation of natural organisms (Ellison *et al.* 2011; Liu *et al.* 2014). In the present study, we used an NGS-based pooled sequencing strategy to analyze genome-wide variation, including single-nucleotide polymorphisms (SNPs) and structural variations (SVs), across the AHR genome. Our data provide a resource for future population genetic research on the AHR, and advance our understanding of the molecular adaptations of rats with respect to their ecological niches.

doi: 10.1534/g3.116.029504
Manuscript received March 22, 2016; accepted for publication April 27, 2016; published Early Online April 27, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.029504/-/DC1
[1] Corresponding author: The State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beichen West Road, Chao Yang District, Beijing 100101, China. E-mail: zhangjx@ioz.ac.cn

MATERIALS AND METHODS

Phylogenetic analysis of the AHR based on mitochondrial genomes

To estimate the phylogenetic position of the AHR, published *Rattus* genus mitochondrial genome sequences (accession numbers listed in Figure 1) were downloaded from GenBank. The sequences of 13 mitochondrial protein-coding genes were joined and aligned using MUSCLE (Edgar 2004). The best mutational model (GTR+I+G) was then selected using jModelTest v2.1.7 (Darriba *et al.* 2012). Mitochondrial genome phylogenetic analyses were conducted using MrBayes v3.22 (Ronquist *et al.* 2012), with two independent analyses running in parallel, each with four Markov Chain Monte Carlo (MCMC) chains. The final tree topology was recovered after a total of 10,000,000 generations, sampling 1 in every 10,000 generations after discarding the first 1000 as burn-in.

Samples, DNA extraction, and library preparation

In 2014, 42 individual AHRs, including 24 from Dujiangyan (31°01'N/103°40'E), Sichuan Province, China and 18 from Taiyuan (37°43'N/112°29'E), Shanxi Province, China, were trapped and identified using mitochondrial cytochrome oxidase subunit I barcode sequences. Genomic DNA samples were extracted from small pieces of tail using a TailGen DNA extraction Kit (CWBIO, Beijing, China). The quality and integrity of the extracted DNA was checked by measuring the A260/A280 ratio using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA) and by agarose gel electrophoresis. Then, equivalent amounts of DNA extracted from all individuals were pooled for library construction. A library with an insert size of approximately 300 bp was prepared and sequenced on either an Illumina HiSequation 2000 or an X Ten instrument, with 100 or 150 bp paired-end reads, respectively. After filtering out of raw sequencing reads containing adapters and reads of low quality, the remaining clean reads were aligned to the reference genome of *R. norvegicus* (RGSC5.0, Ensembl release 70), a close relative of AHR, using the Burrows-Wheeler Alignment tool (BWA v0.7.5a-r405) with default parameters (Li and Durbin 2009). Genome size was estimated on the basis of k-mer frequency distribution using KmerGenie with default parameters (Chikhi and Medvedev 2014). Duplicate reads were removed using the rmdup function in SAMtools v 0.1.19-44428cd (Li *et al.* 2009). In addition, unmapped reads were assembled *de novo* using SOAPdenovo v2.04 (Luo *et al.* 2012) and gene predictions were performed using AUGUSTUS (Hoff and Stanke 2013).

Small variant calling and genetic diversity estimation

After performing the alignment and removing duplicate reads, two different variant calling tools, SNVer v0.5.3 (Wei *et al.* 2011) and VarScan v2.3.7 (Koboldt *et al.* 2009; Koboldt *et al.* 2012), were used to detect SNPs and small insertions and deletions (InDels) in the pooled samples. For SNVer, we set the number of haploids to 84, and used the RGSC v5.0 reference assembly with the Single-Nucleotide Polymorphism database (dbSNP) build 136. Prior to running VarScan, a pileup file was created using the mpileup command in SAMtools. Then, variants were called using the VarScan subcommands mpileup2snp and mpileup2indel with the following parameters: –min-coverage 20, –min-avg-qual 20, and –p-value 1e-02. In our further analyses, we considered only those SNPs and small InDels that were obtained using both pipelines.

Two summary statistics, nucleotide diversity (π) (Nei 1987) and Watterson's theta (θ_W) (Watterson 1975), are commonly used for measuring genetic diversity within a population (Tajima 1983). We used the PoPoolation package (Kofler *et al.* 2011; Schlotterer *et al.* 2014) to estimate π and θ_W in sliding windows across the genome, with a window size of 100 kb and a step size of 20 kb, and the following parameters: –min-coverage 20, –max-coverage 500, –min-qual 20, and –pool-size 42.

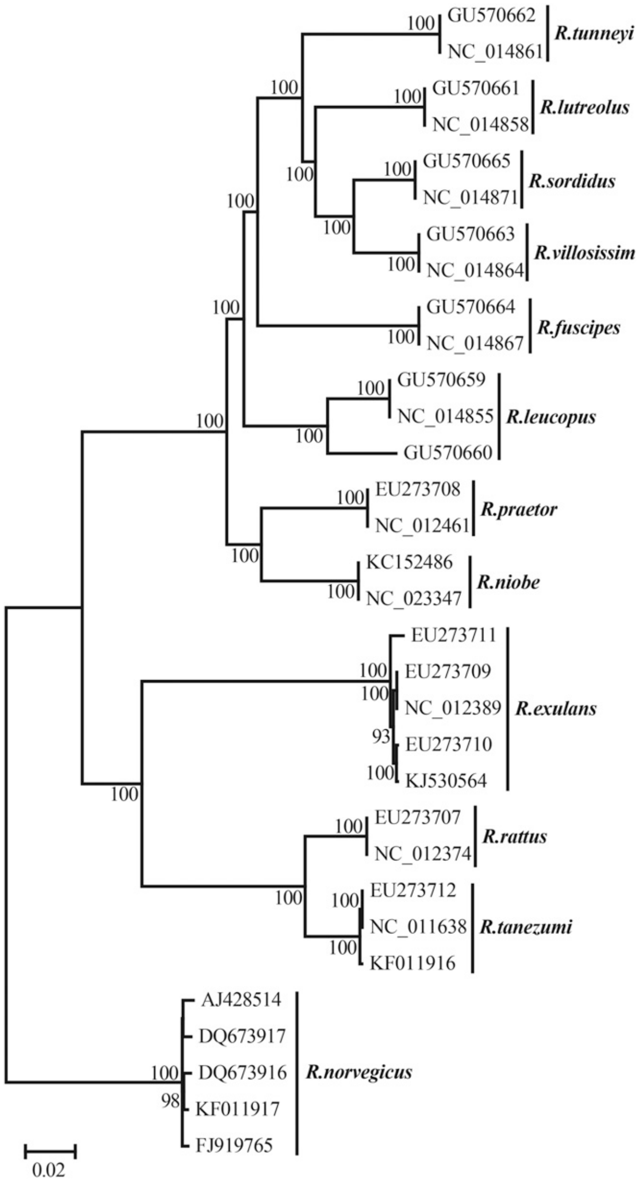
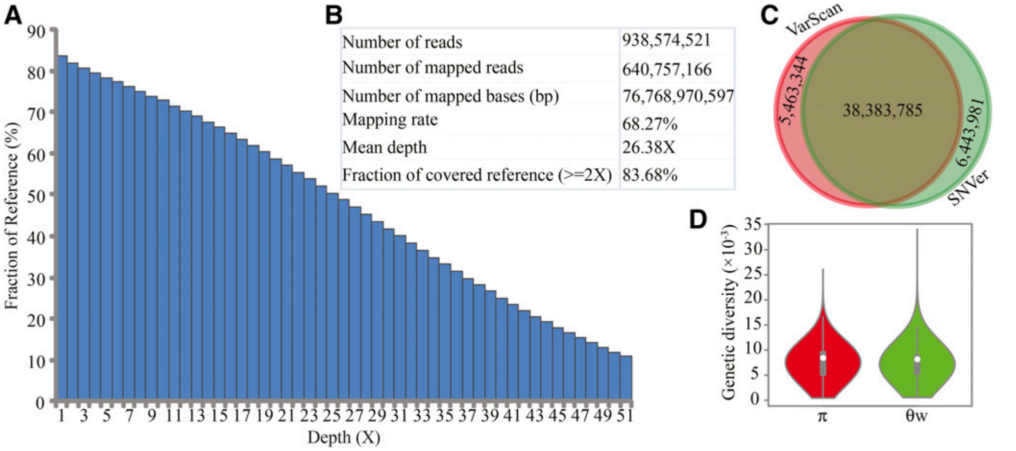


Figure 1 Phylogenetic tree of *Rattus* spp., estimated using mitochondrial genomes. Published mitochondrial genome sequences of the *Rattus* genus were downloaded from GenBank and Accession numbers are indicated.

Structural variation detection

Following the removal of duplicate reads, large InDels and inversions were detected using MetaSV v0.5 (Mohiyuddin *et al.* 2015) on the alignment to the RGSC v5.0 reference genome. The integrative structural-variant caller MetaSV merges results obtained by multiple detection methods, including BreakDancer (Chen *et al.* 2009), LUMPY (Layer *et al.* 2014), and Pindel (Ye *et al.* 2009). BreakDancer can detect a cluster of reads with abnormal length of insert size, or incorrect orientation of ends, based on paired-end read mapping. In contrast, Pindel splits the unmapped end of a one-end anchored read into a few pieces, and performs local realignment of each piece in the candidate region. Finally,



Number of reads	938,574,521
Number of mapped reads	640,757,166
Number of mapped bases (bp)	76,768,970,597
Mapping rate	68.27%
Mean depth	26.38X
Fraction of covered reference (>=2X)	83.68%

Figure 2 Genetic diversity in the AHR. Mapping rate, depth, and genome coverage of AHR sequencing data (A, B); overlap of SNPs obtained from two different callers, VarScan and SNVer (C); and distribution of summary statistics: nucleotide diversity (π) and Watterson's theta (θ_W) (D).

LUMPY incorporates both split read analysis and read-pair discordance to detect breakpoints. To evaluate the accuracy of the predicted breakpoints, 20 randomly selected deletion and inversion breakpoints were analyzed using polymerase chain reaction (PCR)-based Sanger sequencing with the primers listed in Supplemental Material, File S1. Inversions were displayed using RCircos (Zhang *et al.* 2013) and inGAP-sv (Qi and Zhao 2011).

Variant annotation and identification of neutral SNPs

SNPEff v4.0e (Cingolani *et al.* 2012) was used to annotate the identified SNPs based on the Rnor5.0.74 rat assembly. Our previously sequenced individual genome of the AHR sibling species, *R. rattus*, was also reanalyzed and annotated (Deinum *et al.* 2015). Fourfold degenerate sites have traditionally been regarded as neutral variations in mammals, due to the degenerate nature of the genetic code (Eyre-Walker and Keightley 1999; Nachman and Crowell 2000). Because of the uncertainty of the nature of predicted SNPs in segmental duplications (Fredman *et al.* 2004), SNPs residing in duplicated genes were removed from the heterozygous fourfold degenerate dataset in order to obtain reliable neutral SNPs. Furthermore, in order to exclude linkage between SNPs, the distance between neighboring neutral SNPs was set to at least 100,000 bp.

Identification of selection footprints and functional characterization of selected genes

To identify footprints of selection in the AHR genome, we estimated genome-wide allele frequencies using the Pool_hmm program (Boitard *et al.* 2009, 2012, 2013) with the following options: -c 20, –q 20, -k 1e-10, –pred, and –theta 0.0018. Pool_hmm estimates allele frequencies at each polymorphic site based on a probabilistic model, and different patterns of allele frequencies are associated with different hidden states: "Neutral", "Intermediate", and "Selection". The top 5% of polymorphic sites, based on the posterior probability of the hidden state "Selection", were selected, and neighboring loci were merged and identified as positively selected regions. Candidate sweep regions were further identified based on the fact that focal regions show reduced nucleotide diversity in the AHR population (the bottom 5% quantile of the mean genome-wide distribution). To characterize the molecular functions of the genes contained in selective sweep regions, we performed functional enrichment analyses using the clusterProfiler toolkit (Yu *et al.* 2012).

Data availability

The raw sequencing datasets were deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) under accession numbers SRX1425877 and SRX1425879. Table S1 and Table S2 contain nonsense mutations and gene ontology information, respectively, for genes with frameshifts in the AHR. BLASTP results detailing predicted proteins for the *de novo* assembled unmapped reads are listed in Table S3. Genes in selective sweep regions, and functional annotation of selective sweep regions of the AHR genome, are listed in Table S4 and Table S5, respectively. File S1 contains PCR-based Sanger sequencing of the candidate structural variation breakpoints. Fixed missense mutations in the warfarin interaction pathway and polymorphic and unlinked neutral sites of single-copy protein-coding genes are listed in File S2 and File S3, respectively. Frameshift variants and SVs of the AHR are listed in File S4 and File S5, respectively.

RESULTS

Whole-genome sequencing and mapping

More than 938 million clean reads, corresponding to 112 Gb of sequencing data, were generated. The genome size of the AHR was approximately 3.28 Gb estimated based on the k-mer frequency distribution. Approximately 68.27% of the sequencing data could be mapped to the reference genome of the closely related species, *R. norvegicus*, although the two species diverged more than 2.2 million yr ago (Figure 1) (Robins *et al.* 2008, 2010). Approximately 83.68% of the reference genome was covered by at least two reads, the effective genome-wide average sequencing depth was 26.38-fold, and the percentage of the genome with a depth of 10 or more was 73.99% (Figure 2). In an attempt to identify novel genomic sequences, we assembled unmapped reads and obtained 2979 scaffolds with a minimum length of 1 kb and a total of 3.42 Mb in length.

SNP calling

A total of 38,383,785 SNPs were identified in the AHR genome compared with the reference genome (Figure 2 and Figure 3). The means of the summary statistics, π and θ_W, were 7.133×10^{-3} and 6.887×10^{-3}, respectively (Figure 2), suggesting more sequence diversity in the AHR than was reported for *R. norvegicus* (Ness *et al.* 2012). Among the SNPs, 49.85% (19,134,953) were homozygous, and were thus regarded as fixed variants in the AHR in our study. Among the fixed SNPs, 250,606 were located in protein-coding regions, and 82,495 and 167,681 of these were nonsynonymous and synonymous variants, respectively (Figure 4). Of the 82,495 nonsynonymous variants, 239 were predicted to cause premature truncation of proteins due to the insertion of stop codons, whereas 51 were predicted to cause the

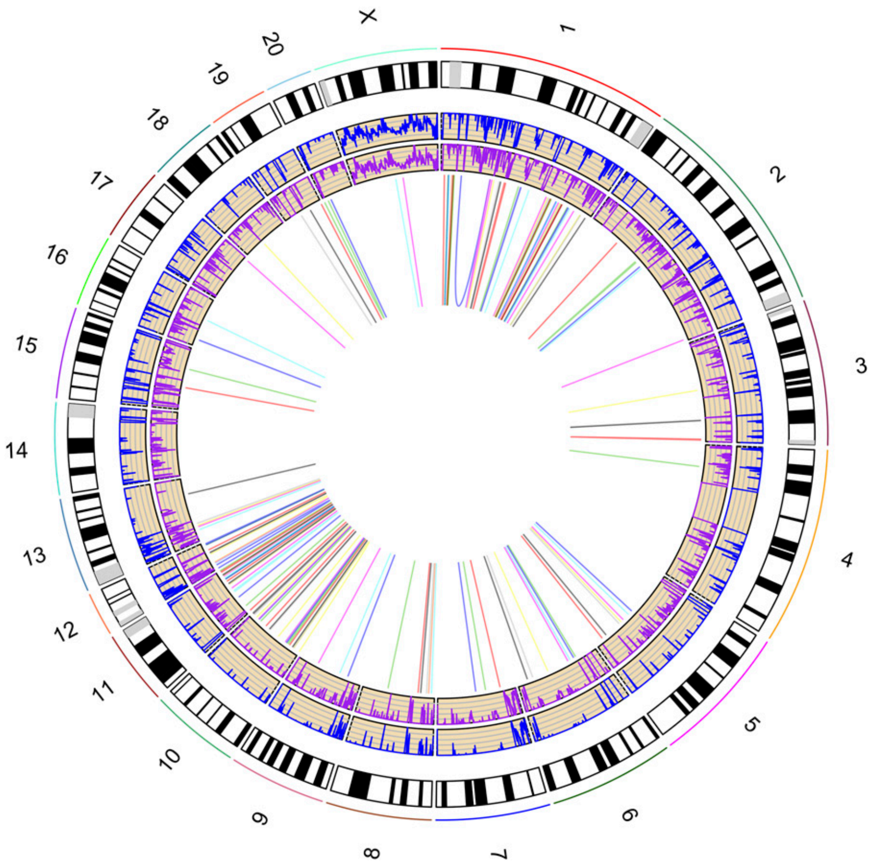


Figure 3 Distribution of SNPs, InDels, and inversions in the AHR genome relative to *R. norvegicus* chromosomes and karyotypes. The innermost circle shows inversions. The next circles show lines representing small InDels (purple) and SNP density (blue).

removal of stop codons (Table S1 and Figure 4). Of the 239 nonsense mutations, 106 were shared by the AHR sibling species, *R. rattus*, another pathogen carrier (Meerburg *et al.* 2009; Himsworth *et al.* 2015), suggesting a possible ancient origin for these mutations, dating to before the divergence of these two species. Some viral-infection-related genes, such as *Il1a* (Pruitt *et al.* 1995; Dinarello 2011) and *Srpk1* (Sciabica *et al.* 2003; Wang *et al.* 2014), exhibited gains of stop codons (Figure 4), which may correlate with the ability of the AHR to serve as a pathogen carrier. The widespread rodenticide, warfarin, can inhibit blood coagulation, and continuous intake of warfarin causes potentially fatal hemorrhages (Davis and Davies 1970; Kohn *et al.* 2000). We found various missense mutations in several genes involved in the warfarin interaction pathway (Figure 4 and File S2), including *Calu*, *Ggcx*, *Ephx1*, *Orm1*, *Pros1*, *Proz*, *Serpinc1*, and *Vkorc1* (I90L). Of note, the I90L mutation in the *Vkorc1* gene can decrease VKOR activity by 10% compared to wild type, which has led to warfarin resistance in rat populations in Argentina (Rost *et al.* 2009; Prescott 2011). These fixed missense SNPs appear to underlie anticoagulant resistance in the AHR (Long *et al.* 2010; Huang *et al.* 2012).

Heterozygous neutral SNPs are popular markers for use in assessment of genetic variation in natural populations (Gray *et al.* 2000; Suh and Vijg 2005; Holderegger *et al.* 2006). Through our pooled sequencing experiments, we identified 12,823,195 heterozygous SNPs within the genomes of AHR populations. Due to the degenerate nature of the genetic code, substitutions at fourfold degenerate sites are regarded as neutral variations (Eyre-Walker and Keightley 1999; Nachman and Crowell 2000). We excluded SNPs with high linkage disequilibrium and those residing in duplicated genes from the dataset of heterozygous fourfold degenerate SNPs, resulting in 5800 autosome neutral SNPs in single-copy protein-coding genes (Figure 5 and File S3), providing an ideal panel for future migration or dispersal pattern studies of AHR populations.

InDels and inversions

We identified 3,179,903 intraread InDels, with an average density of 1.12 InDels per kb, including 1,624,854 deletions and 1,555,049 insertions. Among all intraread InDels in coding regions, 1244 were 1 bp long, and 1919 were 3n bp long (*i.e.*, the length was a multiple of 3 bp). The proportion of 3n bp intraread InDels in coding regions was significantly higher than that in intergenic regions. In addition, we observed that the InDels within coding sequences were enriched in regions encoding the *N*- and *C*-termini of proteins (Figure 6). Notably, we identified a total of 1881 frameshift InDels, with an allele frequency of less than 0.55 in 1297 genes (File S4). Functional category analyses showed that these genes were related to the cellular response to stress (GO:0033554), sensory organ development (GO:0007423), and regulation of neurotransmitter levels (GO:0001505) (Table S2). Additionally, we found that some genes containing frameshift InDels, such as *Aldh* and *Stard7*, were not single-copy.

We detected a total of 12,423 structural variants, including 12,216 deletions, 61 duplications, and 143 inversions (File S5). Using PCR-based Sanger sequencing, we found that the accuracy of the predicted breakpoints was 85% (Figure 7 and File S1). We found an inversion region with a length of 5.46 Mb and different breakpoints (1,357,955:6,822,593; 1,358,196:6,822,593; 1,358,799: 6,821,965; 1,358,402:6,822,135) on chromosome 12. This finding has been confirmed by fluorescence *in situ* hybridization, which showed that the order of the BAC clones on the submetacentric form of chromosome 12 in *R. norvegicus* (12p12, 12p11) was inverted in the acrocentric form of the heteromorphic pairs in the

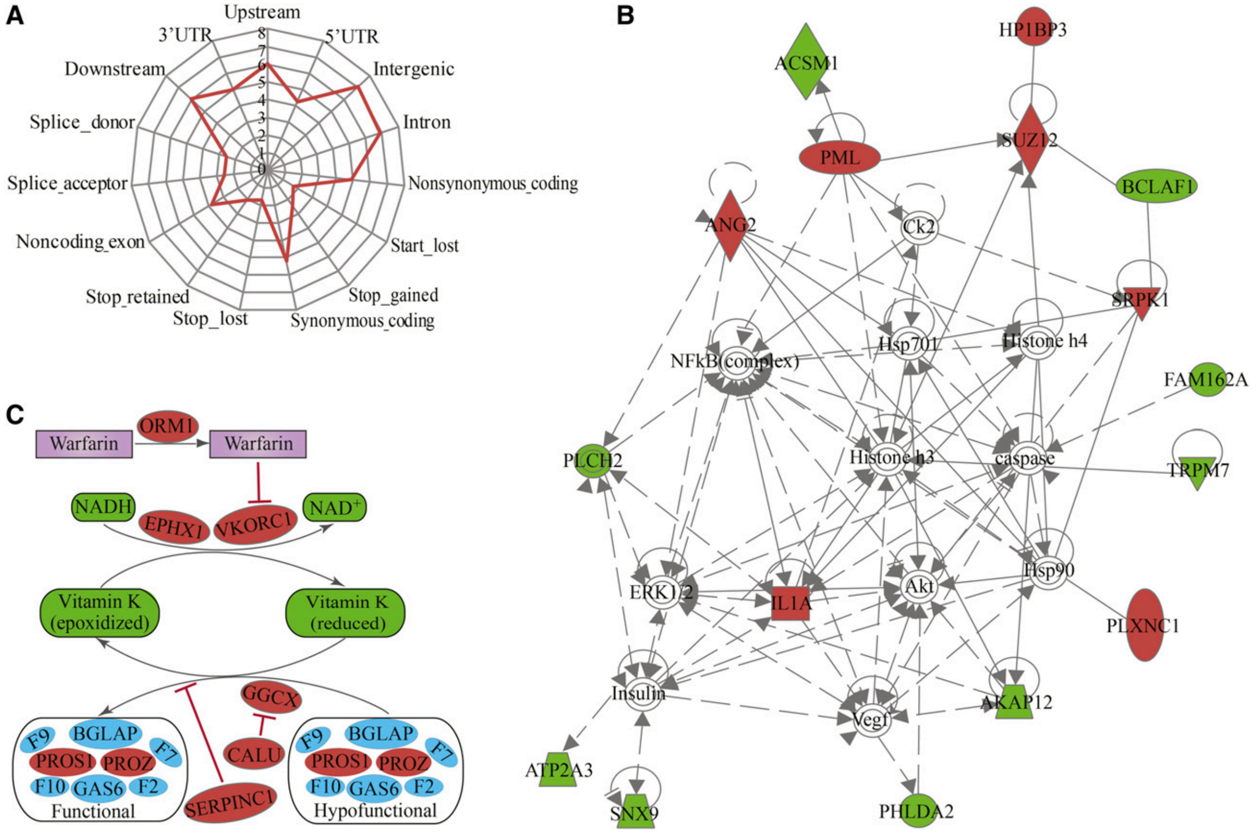


Figure 4 Functional significance of genetic variants in the AHR. (A) Annotation of homozygous AHR SNPs. (B) Interaction network analysis of proteins containing nonsense mutations using the Ingenuity Pathway Knowledge Base (IPKB). (C) Missense-mutated genes of the warfarin interaction pathway. (A) The axes in the radar chart represent annotated genomic features and Log_{10}-transformed values. (B) Color shading corresponds to identified genes containing nonsense mutations (Table S1), with red indicating viral-infection-related genes. Direct (solid lines) and indirect (dashed lines) interactions of genes are based on the IPKB database. The shape of nodes indicates the major function of the protein. (C) Depicts simplified interrelationships between vitamin K, coagulation factors (II, VII, IX, and X), and warfarin. The identified missense-mutated genes of the warfarin interactive pathway (File S2) are indicated in red.

AHR (Badenhorst *et al.* 2011). These chromosomal variations indicate the extensive chromosomal plasticity of the *Rattus* genus.

De novo assembly of unmapped reads

After assembling the unmapped reads, we obtained 2979 scaffolds longer than 1 kb, with a total sequence length of 3.42 Mb. Computational prediction of gene structure for these scaffolds revealed 20 protein-coding genes. A BLASTP search against the NCBI nonredundant animal protein database revealed that 13 of these genes were homologs of known proteins (E-value <1e-10) (Table S3). Some of these proteins were encoded by motor behavior related genes, such as *Etv1* (Lin *et al.* 1998; de Nooij *et al.* 2013), *SmtnL1* (Wooldridge *et al.* 2008; Lontay *et al.* 2015), and *Npas3* (Erbel-Sieler *et al.* 2004; Brunskill *et al.* 2005), indicating sequence divergence in these genes between the wild AHR and the reference *R. norvegicus* genomes.

Identification of selective sweeps

Following an integrative analysis of the allele frequency spectrum (Boitard *et al.* 2009, 2012, 2013) and nucleotide diversity of focal regions (Kofler *et al.* 2011) in the AHR population, we detected a total of 570 selective sweeps across 21 chromosomes (Figure 5 and Table S4). Annotation of these regions revealed 1120 genes. Moreover, gene overrepresentation analyses for these regions revealed a relationship with several neural activity related terms, including spontaneous neurotransmitter secretion (GO:0061669), spontaneous synaptic transmission (GO:0098814), positive regulation of synaptic vesicle exocytosis (GO:2000302), and regulation of synaptic vesicle transport (GO:1902803) (Table S5). Transport of synaptic vesicles and the release of neurotransmitters are essential for propagating nerve impulses between neurons. Several genes under positive selection, including *Syt1*, *Unc13b*, *Nlgn1*, and *Stx1b*, are involved in transmission of nerve impulses and synaptic transmission. Glutamate is the major neurotransmitter in the brain, regulating many kinds of behaviors and emotions, and playing an important role in cognitive ability (Petroff 2002). *Gria4* and *Grik1* are important excitatory glutamate receptors, which were also found to be under positive selection in our study. Another family of glutamate receptor genes that were found to be under selection were the inhibitory γ-aminobutyric acid receptors, such as *Gabbr2*. Hence, enrichment of the synaptic function category appears to underlie the evolution of the AHR.

DISCUSSION

Wild-derived rat species provide novel genome resources that cannot be obtained from the laboratory rat, and that can be used to explore the genetic mechanisms underlying complex behavioral or ecological phenotypes (Smits *et al.* 2005). Here, we report the spectrum of genetic

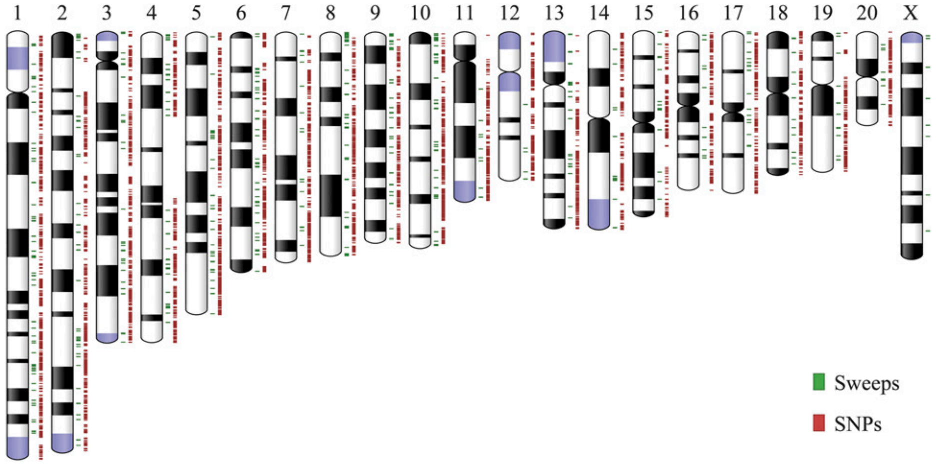


Figure 5 Distribution of selective sweep regions and SNP loci in the AHR. Distribution of AHR selective sweep regions (green) and SNP loci at fourfold degenerate sites (red), according to *R. norvegicus* chromosomes and karyotypes. A sequence gap in the *R. norvegicus* RGSC5.0 reference genome accounts for the region on the chromosome 4 lacking annotation.

variation of SNPs and SVs across the AHR genome, providing a resource for future genetic studies of the species.

As a highly adaptable pest, AHR has evolved the ability to survive in its ecological habitats. We identified eight fixed missense-mutated genes within the warfarin interaction pathway, which appear to underlie anticoagulant resistance in the AHR (Long *et al.* 2010; Huang *et al.* 2012). In addition to being a notorious pest (Singleton *et al.* 2007; Huang *et al.* 2013), the AHR is a known reservoir of a wide range of human pathogens, including hantaviruses, *Bartonella* spp., and *Leptospira interrogans* (Plyusnina *et al.* 2009; Jonsson *et al.* 2010; Yin *et al.* 2011). Loss of function of several infection-related genes in this species might contribute to its ability to survive in this adverse niche. Given the risk that the AHR constitutes to human health, due to its commensal species and pathogen-carrier characteristics, it is of major interest to investigate the migration and dispersal routes of this species. However, the sensitivity of current molecular markers limits the scope of population research studies. SNPs have been recognized as the markers of choice for population genetics, because of their genome-wide incidence and high frequency, compared to other types of polymorphisms (Collins *et al.* 1997; Landegren *et al.* 1998; Brookes 1999). Heterozygous neutral SNPs are attractive markers for assessment of genetic variation in natural populations (Gray *et al.* 2000; Suh and Vijg 2005), and they have great potential for use in investigation of processes such as gene flow, migration, or dispersal (Holderegger *et al.* 2006). Through our pooled sequencing, we obtained 5800 fourfold degenerate autosomal sites residing in single-copy protein-coding genes (Figure 5). Oligonucleotide hybridization assays or multiplex PCR coupled with NGS sequencing based on these polymorphic neutral sites will enable high-throughput genotyping of multiple individuals. Thus, we have produced an ideal SNP panel for future migration or dispersal pattern studies of AHR populations.

A great number of intraread and large InDels were observed in the AHR genome. We found that the size distribution of intraread InDels (predominantly 1 bp or 3n bp) in coding regions was significantly different from that in noncoding regions, indicating the functional importance of the coding InDels. In addition, we observed that InDels within coding sequences were enriched in regions encoding the *N*- and *C*-termini of proteins (Figure 6). InDels in these regions may be less functionally damaging than those in other coding regions because genes may have alternative translation start sites at the N-terminus or translation of the protein may be almost completed before C-terminal InDels (Ng *et al.* 2008; Pelak *et al.* 2010; Zhan *et al.* 2011). Notably, we found that some genes carrying frameshift variants, such as *Aldh* and *Stard7*, are not single-copy. Due to the functional redundancy of genes in some families, it is conceivable that functional compensation within the same family could rescue the loss of function of a gene containing a nonsense mutation (Teng *et al.* 2013).

Chromosome evolution in rodents has been driven largely by pericentric inversions (Badenhorst *et al.* 2011). An integrative structural-variant caller was used in our study to detect large InDels and inversions

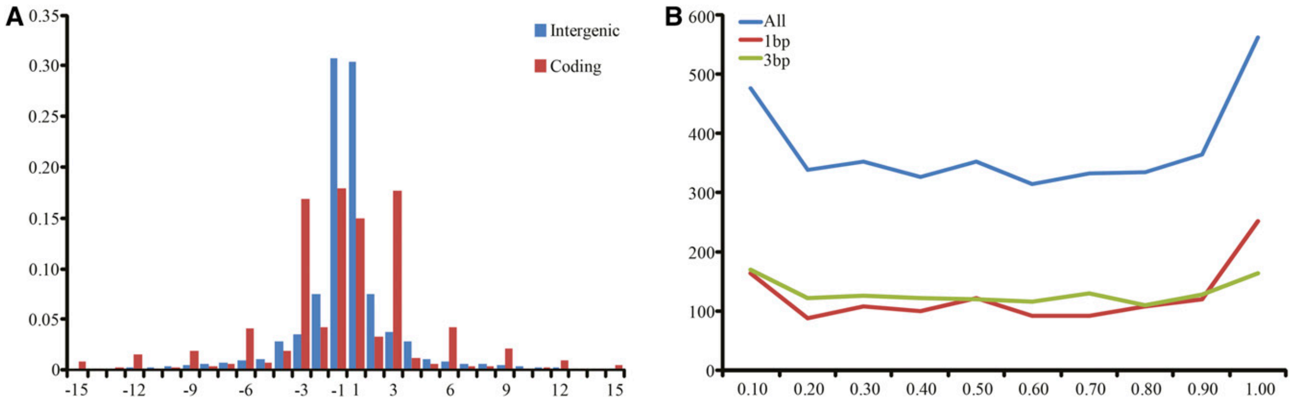


Figure 6 Distribution of InDel sizes in the genome of the AHR. (A) Negative numbers represent deletions and positive numbers represent insertions. (B) The relative locations of InDels within coding sequences are plotted as the first amino acid position of the InDel, divided by the total protein length.

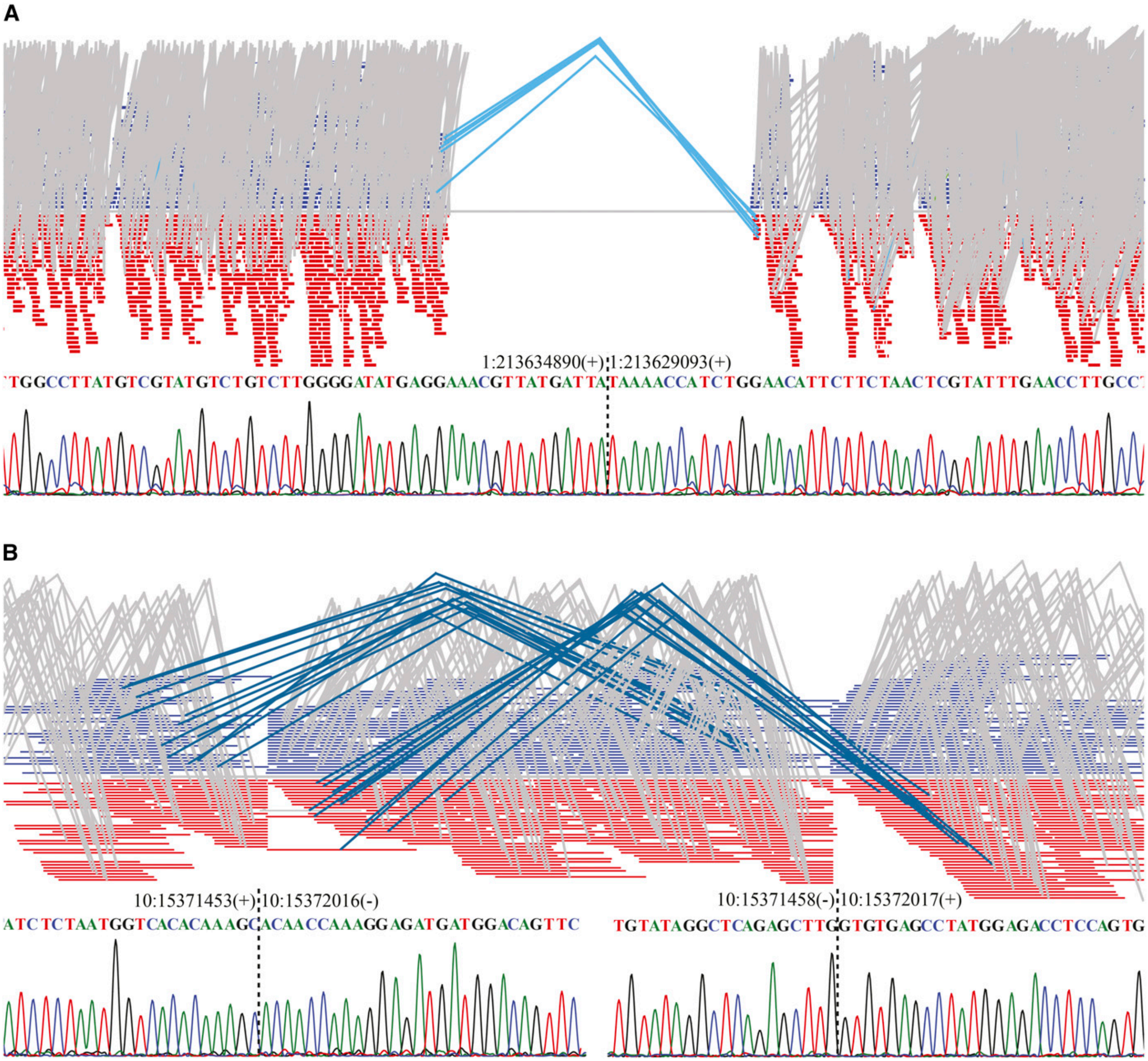


Figure 7 Structural variants in the AHR genome. Representative visualization of read alignments adjacent to (A) deletion and (B) inversion breakpoints in the AHR genome, followed by validation via PCR-based Sanger sequencing. Gray links indicate normally mapped read pairs with proper read orientation and distance. (A) Light blue links represent read pairs with proper read orientation but longer distance, indicating a deletion event in the query sequence. (B) Dark blue links indicate an inversion, causing the paired reads to demonstrate abnormal orientation, with both ends mapping to the same strand.

(Mohiyuddin *et al.* 2015). Read pairs spanning SV breakpoints produced discordant alignments with an unexpected alignment distance and/or orientation (Chen *et al.* 2009). DNA segments flanking the breakpoint aligned to disjointed locations in split read mapping (Ye *et al.* 2009; Layer *et al.* 2014). This integrative analysis can improve the accuracy of prediction through leveraging multiple orthogonal SV signals, based on both split read and read-pair discordance. The SV calling strategies perform well on mapped regions, although they may miss some SVs in low genome coverage regions. Individuals within populations may have different SV breakpoints in the same region. Due to pooled sequencing, some inversions with different breakpoints were obtained, such as the 5.46 Mb inversions in chromosome 12 (12p12, 12p11). This finding is consistent with previous research (Badenhorst *et al.* 2011). These chromosomal variations indicate the extensive chromosomal plasticity of the *Rattus* genus and provide a rich resource for comparative genomic studies of the rat.

In summary, we report genome-wide SNP and SV variations in the AHR, providing a genomic resource for future studies. Our findings have the potential to contribute to the understanding of the molecular adaptations of the rat to its ecological niches.

ACKNOWLEDGMENTS

This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (no. XDB11010400), the National Natural Science Foundation of China (nos. 91231107 and 31572277), and the State Key Laboratory of Integrated Management of Pest Insects and Rodents (ChineseIPM1415).

LITERATURE CITED

Badenhorst, D., G. Dobigny, F. Adega, R. Chaves, P. C. O'Brien *et al.*, 2011 Chromosomal evolution in Rattini (Muridae, Rodentia). Chromosome Res. 19: 709–727.

Baker, H. J., J. R. Lindsey, and S. H. Weisbroth, 1979 The Laboratory Rat. Academic Press, New York.

Bastos, A. D., D. Nair, P. J. Taylor, H. Brettschneider, F. Kirsten *et al.*, 2011 Genetic monitoring detects an overlooked cryptic species and reveals the diversity and distribution of three invasive *Rattus* congeners in South Africa. BMC Genet. 12: 26.

Boitard, S., C. Schlotterer, and A. Futschik, 2009 Detecting selective sweeps: a new approach based on hidden markov models. Genetics 181: 1567–1578.

Boitard, S., C. Schlotterer, V. Nolte, R. V. Pandey, and A. Futschik, 2012 Detecting selective sweeps from pooled next-generation sequencing samples. Mol. Biol. Evol. 29: 2177–2186.

Boitard, S., R. Kofler, P. Francoise, D. Robelin, C. Schlotterer *et al.*, 2013 Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. Mol. Ecol. Resour. 13: 337–340.

Brookes, A. J., 1999 The essence of SNPs. Gene 234: 177–186.

Brunskill, E. W., L. A. Ehrman, M. T. Williams, J. Klanke, D. Hammer *et al.*, 2005 Abnormal neurodevelopment, neurosignaling and behaviour in Npas3-deficient mice. Eur. J. Neurosci. 22: 1265–1276.

Chen, K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki *et al.*, 2009 BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6: 677–681.

Chikhi, R., and P. Medvedev, 2014 Informed and automated k-mer size selection for genome assembly. Bioinformatics 30: 31–37.

Cingolani, P., A. Platts, L. le Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92.

Collins, F. S., M. S. Guyer, and A. Charkravarti, 1997 Variations on a theme: cataloging human DNA sequence variation. Science 278: 1580–1581.

Conroy, C., K. Rowe, K. C. Rowe, P. Kamath, K. Aplin *et al.*, 2013 Cryptic genetic diversity in *Rattus* of the San Francisco Bay region, California. Biol. Invasions 15: 741–758.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada, 2012 jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9: 772.

Davis, R. J., and B. H. Davies, 1970 The biochemistry of warfarin resistance in the rat. Biochem. J. 118: 44P–45P.

Deinum, E. E., D. L. Halligan, R. W. Ness, Y. H. Zhang, L. Cong *et al.*, 2015 Recent evolution in *Rattus norvegicus* is shaped by declining effective population size. Mol. Biol. Evol. 32: 2547–2558.

de Nooij, J. C., S. Doobar, and T. M. Jessell, 2013 Etv1 inactivation reveals proprioceptor subclasses that reflect the level of NT3 expression in muscle targets. Neuron 77: 1055–1068.

Dinarello, C. A., 2011 Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. Blood 117: 3720–3732.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Ellison, C. E., C. Hall, D. Kowbel, J. Welch, R. B. Brem *et al.*, 2011 Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc. Natl. Acad. Sci. USA 108: 2831–2836.

Erbel-Sieler, C., C. Dudley, Y. Zhou, X. Wu, S. J. Estill *et al.*, 2004 Behavioral and regulatory abnormalities in mice deficient in the NPAS1 and NPAS3 transcription factors. Proc. Natl. Acad. Sci. USA 101: 13648–13653.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature 397: 344–347.

Fredman, D., S. J. White, S. Potter, E. E. Eichler, J. T. Den Dunnen *et al.*, 2004 Complex SNP-related sequence variation in segmental genome duplications. Nat. Genet. 36: 861–866.

Gray, I. C., D. A. Campbell, and N. K. Spurr, 2000 Single nucleotide polymorphisms as tools in human genetics. Hum. Mol. Genet. 9: 2403–2408.

Hamta, A., T. Adamovic, E. Samuelson, K. Helou, A. Behboudi *et al.*, 2006 Chromosome ideograms of the laboratory rat *(Rattus norvegicus)* based on high-resolution banding, and anchoring of the cytogenetic map to the DNA sequence by FISH in sample chromosomes. Cytogenet. Genome Res. 115: 158–168.

Himsworth, C. G., E. Zabek, A. Desruisseau, E. J. Parmley, R. Reid-Smith *et al.*, 2015 Prevalence and characteristics of *Escherichia coli* and *Salmonella* spp. in the feces of wild urban Norway and black rats *(Rattus norvegicus* and *Rattus rattus)* from an inner-city neighborhood of Vancouver, Canada. J. Wildl. Dis. 51: 589–600.

Hoff, K. J., and M. Stanke, 2013 WebAUGUSTUS: a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41: W123–128.

Holderegger, R., U. Kamm, and F. Gugerli, 2006 Adaptive *vs.* neutral genetic diversity: implications for landscape genetics. Landsc. Ecol. 21: 797–807.

Huang, H. W., Z. T. Zou, Z. Z. Tian, X. Y. Hu, and Y. Zhao, 2012 Test of resistance of Rattus tanezumi to warfarin in Xingyi City, Guizhou Province, China. Chin. J. Vec. Biol. Contr. 23: 554–555.

Huang, L. Q., X. G. Guo, J. R. Speakman, and W. G. Dong, 2013 Analysis of gamasid mites (Acari: Mesostigmata) associated with the Asian house rat, *Rattus tanezumi* (Rodentia: Muridae) in Yunnan Province, southwest China. Parasitol. Res. 112: 1967–1972.

Jacob, H. J., 1999 Functional genomics and rat models. Genome Res. 9: 1013–1016.

Jonsson, C. B., L. T. Figueiredo, and O. Vapalahti, 2010 A global perspective on hantavirus ecology, epidemiology, and disease. Clin. Microbiol. Rev. 23: 412–441.

Koboldt, D. C., K. Chen, T. Wylie, D. E. Larson, M. D. McLellan *et al.*, 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25: 2283–2285.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan *et al.*, 2012 VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22: 568–576.

Kocher, A., M. Desquesnes, S. Yangtara, S. Morand, and S. Jittapalapong, 2015 Is the oriental house rat *(Rattus tanezumi)* a potential reservoir for *Trypanosoma evansi* in Thailand? J. Wildl. Dis. 51: 719–723.

Kofler, R., P. Orozco-terWengel, N. De Maio, R. V. Pandey, V. Nolte *et al.*, 2011 PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6: e15925.

Kohn, M. H., H. J. Pelz, and R. K. Wayne, 2000 Natural selection mapping of the warfarin-resistance gene. Proc. Natl. Acad. Sci. USA 97: 7911–7915.

Krinke, G., 2000 The Laboratory Rat. Academic Press, San Diego, CA.

Lack, J. B., D. U. Greene, C. J. Conroy, M. J. Hamilton, J. K. Braun *et al.*, 2012 Invasion facilitates hybridization with introgression in the Rattus rattus species complex. Mol. Ecol. 21: 3545–3561.

Landegren, U., M. Nilsson, and P. Y. Kwok, 1998 Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res. 8: 769–776.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall, 2014 LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15: R84.

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

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

Lin, J. H., T. Saito, D. J. Anderson, C. Lance-Jones, T. M. Jessell *et al.*, 1998 Functionally related motor neuron pool and muscle sensory afferent subtypes defined by coordinate ETS gene expression. Cell 95: 393–407.

Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris *et al.*, 2014 Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157: 785–794.

Long, H. Y., J. J. Wang, Z. X. Huang, and Y. N. Peng, 2010 Laboratory test of the resistance of *Rattus tanezumi* to warfarin and bromadiolone. Chin. J. Vec. Biol. Contr. 21: 378–379.

Lontay, B., K. Bodoor, A. Sipos, D. H. Weitzel, D. Loiselle *et al.*, 2015 Pregnancy and Smoothelin-like Protein 1 (SMTNL1) deletion promote the switching of skeletal muscle to a glycolytic phenotype in human and mice. J. Biol. Chem. 290: 17985–17998.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang *et al.*, 2012 SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1: 18.

Meerburg, B. G., G. R. Singleton, and A. Kijlstra, 2009 Rodent-borne diseases and their risks for public health. Crit. Rev. Microbiol. 35: 221–270.

Mohiyuddin, M., J. C. Mu, J. Li, N. Bani Asadi, M. B. Gerstein *et al.*, 2015 MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics 31: 2741–2744.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

Nakayama, S. M., Y. Ikenaka, K. Hamada, K. Muzandu, K. Choongo *et al.*, 2013 Accumulation and biological effects of metals in wild rats in mining areas of Zambia. Environ. Monit. Assess. 185: 4907–4918.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Ness, R. W., Y. H. Zhang, L. Cong, Y. Wang, J. X. Zhang *et al.*, 2012 Nuclear gene variation in wild brown rats. G3 (Bethesda) 2: 1661–1664.

Ng, P. C., S. Levy, J. Huang, T. B. Stockwell, B. P. Walenz *et al.*, 2008 Genetic variation in an individual human exome. PLoS Genet. 4: e1000160.

Pelak, K., K. V. Shianna, D. Ge, J. M. Maia, M. Zhu *et al.*, 2010 The characterization of twenty sequenced human genomes. PLoS Genet. 6: e1001111.

Petroff, O. A., 2002 GABA and glutamate in the human brain. Neuroscientist 8: 562–573.

Plyusnina, A., I. N. Ibrahim, and A. Plyusnin, 2009 A newly recognized hantavirus in the Asian house rat *(Rattus tanezumi)* in Indonesia. J. Gen. Virol. 90: 205–209.

Prescott, C., 2011 Development of rodenticides and the impact of resistance on anticoagulant rodenticides. 7th International Conference on Urban Pests, Ouro Preto, Brazil.

Pruitt, J. H., E. M. Copeland, 3rd, and L. L. Moldawer, 1995 Interleukin-1 and interleukin-1 antagonism in sepsis, systemic inflammatory response syndrome, and septic shock. Shock 3: 235–251.

Qi, J., and F. Zhao, 2011 inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res. 39: W567–W575.

Robins, J. H., P. A. McLenachan, M. J. Phillips, L. Craig, H. A. Ross *et al.*, 2008 Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Mol. Phylogenet. Evol. 49: 460–466.

Robins, J. H., P. A. McLenachan, M. J. Phillips, B. J. McComish, E. Matisoo-Smith *et al.*, 2010 Evolutionary relationships and divergence times among the native rats of Australia. BMC Evol. Biol. 10: 375.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling *et al.*, 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61: 539–542.

Rost, S., H. J. Pelz, S. Menzel, A. D. MacNicoll, V. Leon *et al.*, 2009 Novel mutations in the VKORC1 gene of wild rats and mice: a response to 50 years of selection pressure by warfarin? BMC Genet. 10: 4.

Saar, K., A. Beck, M. T. Bihoreau, E. Birney, D. Brocklebank *et al.*, 2008 SNP and haplotype mapping for genetic analysis in the rat. Nat. Genet. 40: 560–566.

Schlotterer, C., R. Tobler, R. Kofler, and V. Nolte, 2014 Sequencing pools of individuals: mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15: 749–763.

Sciabica, K. S., Q. J. Dai, and R. M. Sandri-Goldin, 2003 ICP27 interacts with SRPK1 to mediate HSV splicing inhibition by altering SR protein phosphorylation. EMBO J. 22: 1608–1619.

Singleton, G. R., P. R. Brown, J. Jacob, and K. P. Aplin, 2007 Unwanted and unintended effects of culling: a case for ecologically-based rodent management. Integr. Zool. 2: 247–259.

Smits, B. M., V. Guryev, D. Zeegers, D. Wedekind, H. J. Hedrich *et al.*, 2005 Efficient single nucleotide polymorphism discovery in laboratory rat strains using wild rat-derived SNP candidates. BMC Genomics 6: 170.

Suh, Y., and J. Vijg, 2005 SNP discovery in associating genetic variation with human disease phenotypes. Mutat. Res. 573: 41–53.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Teng, H., W. Cai, K. Zeng, F. Mao, M. You *et al.*, 2013 Genome-wide identification and divergent transcriptional expression of StAR-related lipid transfer (START) genes in teleosts. Gene 519: 18–25.

Tian, J. H., J. Y. Bao, T. P. Wu, F. B. Wu, X. Huang *et al.*, 2006 Efficacy of rodenticides used in common against Rattus norvegicus and Rattus flavipectus. Chinese Journal of Hygienic Insecticides and Equipments 12: 284–286.

Wang, P., Z. Zhou, A. Hu, C. Ponte de Albuquerque, Y. Zhou *et al.*, 2014 Both decreased and increased SRPK1 levels promote cancer by interfering with PHLPP-mediated dephosphorylation of Akt. Mol. Cell 54: 378–391.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wei, Z., W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson, 2011 SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 39: e132.

Wilson, D. E., and D. M. Reeder, 2005 *Mammal Species of the World: A Taxonomic and Geographic Reference*. Johns Hopkins University Press, Baltimore.

Wooldridge, A. A., C. N. Fortner, B. Lontay, T. Akimoto, R. L. Neppl *et al.*, 2008 Deletion of the protein kinase A/protein kinase G target SMTNL1 promotes an exercise-adapted phenotype in vascular smooth muscle. J. Biol. Chem. 283: 11850–11859.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871.

Yin, J. X., A. Geater, V. Chongsuvivatwong, X. Q. Dong, C. H. Du *et al.*, 2011 Predictors for abundance of host flea and floor flea in households of villages with endemic commensal rodent plague, Yunnan Province, China. PLoS Negl. Trop. Dis. 5: e997.

Yu, G., L. G. Wang, Y. Han, and Q. Y. He, 2012 clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16: 284–287.

Zhan, B., J. Fadista, B. Thomsen, J. Hedegaard, F. Panitz *et al.*, 2011 Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC Genomics 12: 557.

Zhang, H., P. Meltzer, and S. Davis, 2013 RCircos: an R package for Circos 2D track plots. BMC Bioinformatics 14: 244.

Zheng, Z. M., 1981 Preliminary investigation on the susceptibility and tolerance of some rodents to Diphacinone and its applied significance. Acta Theriologica Sinica 1: 93–99.

Communicating editor: J. M. Comeron