Whole-genome sequences of DA and F344 rats with different susceptibilities to arthritis, autoimmunity, inflammation and cancer

Guo, Xiaosen; Brenner, Max; Zhang, Xuemei; Laragione, Teresina; Tai, Shuaishuai; Li, Yanhong; Bu, Junjie; Yin, Ye; Shah, Anish A.; Kwan, Kevin; Li, Yingrui; Wang, Jun; Gulko, Percio S.

Published in:
Genetics

DOI:
10.1534/genetics.113.153049

Publication date:
2013

Document version
Publisher's PDF, also known as Version of record

Whole-Genome Sequences of DA and F344 Rats with Different Susceptibilities to Arthritis, Autoimmunity, Inflammation and Cancer

Xiaosen Guo,*,[1] Max Brenner,†,‡,[1] Xuemei Zhang,* Teresina Laragione,†,‡ Shuaishuai Tai,* Yanhong Li,* Junjie Bu,*,§ Ye Yin,* Anish A. Shah,† Kevin Kwan,† Yingrui Li,* Wang Jun,*,,††,[2] and Pércio S. Gulko†,‡,‡‡,[2]**

*BGI-Shenzhen, Shenzhen 518083, China, †Laboratory of Experimental Rheumatology, Feinstein Institute for Medical Research, Manhasset, New York 11030, ‡Hofstra North Shore–Long Island Jewish School of Medicine, Manhasset, New York 11030, §Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China, **Department of Biology and ††The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-1165 Copenhagen, Denmark, and ‡‡The Elmezzi Graduate School of Molecular Medicine, Manhasset, New York 11030

ABSTRACT DA (D-blood group of Palm and Agouti, also known as Dark Agouti) and F344 (Fischer) are two inbred rat strains with differences in several phenotypes, including susceptibility to autoimmune disease models and inflammatory responses. While these strains have been extensively studied, little information is available about the DA and F344 genomes, as only the Brown Norway (BN) and spontaneously hypertensive rat strains have been sequenced to date. Here we report the sequencing of the DA and F344 genomes using next-generation Illumina paired-end read technology and the first *de novo* assembly of a rat genome. DA and F344 were sequenced with an average depth of 32-fold, covered 98.9% of the BN reference genome, and included 97.97% of known rat ESTs. New sequences could be assigned to 59 million positions with previously unknown data in the BN reference genome. Differences between DA, F344, and BN included 19 million positions in novel scaffolds, 4.09 million single nucleotide polymorphisms (SNPs) (including 1.37 million new SNPs), 458,224 short insertions and deletions, and 58,174 structural variants. Genetic differences between DA, F344, and BN, including high-impact SNPs and short insertions and deletions affecting >2500 genes, are likely to account for most of the phenotypic variation between these strains. The new DA and F344 genome sequencing data should facilitate gene discovery efforts in rat models of human disease.

THE laboratory rat (*Rattus norvegicus*) has been a model organism for the study of human biology and diseases for nearly 200 years (Jacob 1999). Rats differing in susceptibility to disease models and other traits have been extensively studied to better understand human physiology, pharmacology, toxicology, nutrition, behavior, immunology, and diseases such as diabetes, autoimmunity, arthritis, and cancer. These traits have a strong genetic component, making rat models of human disease highly useful for the identification and validation of causative genes and pathways, as well as for testing new therapeutic approaches.

The sequencing of the Brown Norway (BN/SsNHsdMcwi) rat genome was a milestone for the identification, positional cloning, and study of disease model and trait regulatory genes. The BN rat genome was first drafted using a strategy that combined bacterial artificial chromosome (BAC) end sequencing, whole-genome shot gun sequencing, and BAC fingerprinting mapping (Gibbs *et al.* 2004). The BN rat genome was later expanded and reassembled, leading to the draft assembly RGSC v3.4 (Worley *et al.* 2008). The BN strain was chosen because it has been commonly used in many different fields and studies and was also a founder strain for panels of consomic and recombinant inbred rat strains (Worthey *et al.* 2010).

doi: 10.1534/genetics.113.153049
Manuscript received December 14, 2012; accepted for publication May 14, 2013
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153049/-/DC1.
SNPs and indels have been submitted to NCBI and will be available in dbSNP.
[1]These authors contributed equally to this article.
[2]Corresponding authors: Feinstein Institute for Medical Research, 350 Community Dr., Room 1240, Manhasset, NY 11030. E-mail: pgulko@nshs.edu; Beijing Genome Institute, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. E-mail: wangj@genomics.org.cn

The DA (D-blood group of Palm and Agouti, also known as Dark Agouti) and the F344 (Fischer) strains have been extensively studied due to their phenotypic differences in complex traits as diverse as nociception and behavior (Brodkin *et al.* 1999; Terner *et al.* 2006), resistance to infections and parasites (Ishih 1994; Suzuki *et al.* 2006; Zhang *et al.* 2011), severity of autoimmune and inflammatory diseases such as arthritis (Dahlman *et al.* 1998; Sun *et al.* 1999; Wilder *et al.* 1999), oxygen-induced retinopathy (van Wijngaarden *et al.* 2007), muscular strength (Biesiadecki *et al.* 1998), bone mineral density (Turner *et al.* 2001), taste preference (Tordoff *et al.* 2008), cellular phenotypes (Brenner *et al.* 2007; Laragione *et al.* 2007, 2008; Zhang *et al.* 2011), metabolic traits (van Den Brandt *et al.* 2000), and costicosterone levels (Potenza *et al.* 2004). These and other complex traits have been mapped in linkage studies, and the Rat Genome Database (http://rgd.mcw.edu) presently curates 257 quantitative trait loci (QTL) in crosses involving DA and 362 QTL in crosses involving F344 rats, including congenics. Yet detailed genomic information for DA and F344 is lacking and would be instrumental for the identification of the genes accounting for each QTL and for the understanding of the genetic regulation of several complex traits.

Next-generation whole-genome sequencing (NGS) technology enables ultrahigh depth and high-resolution sequencing projects at a cost significantly lower than the traditional dideoxynucleotide-based capillary method. NGS has been successfully used to resequence the human (Bentley *et al.* 2008; Wang *et al.* 2008; Wheeler *et al.* 2008; Ahn *et al.* 2009; Kim *et al.* 2009; G. Li *et al.* 2009; Fujimoto *et al.* 2010; Tong *et al.* 2010), mouse (Keane *et al.* 2011; Yalcin *et al.* 2011), and the spontaneously hypertensive rat (SHR) (Atanur *et al.* 2010) genomes. Here we report the high-depth sequencing of the DA and F344 strains using NGS to generate the first two *de novo* assemblies of the rat genome and the identification of >2 million new variants likely to account for many of the phenotypic differences between DA, F344, and BN.

Materials and Methods

Rats and DNA

DA (DA/BklArbNsi) rats were originally purchased from Bantin and Kingman, transferred to the Arthritis and Rheumatism Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, and maintained since 2002 at the Laboratory of Experimental Rheumatology at the Feinstein Institute for Medical Research (formerly North Shore-Long Island Jewish Research Institute) via brother–sister mating. F344 (F344/NHsd) rats were purchased from Harlan Laboratories. Genomic DNA was extracted from the liver of one male DA and one male F344 rat using the phenol–chloroform–isoamyl alcohol method (Strauss 2001). The quantity of DNA was determined using a NanoDrop spectrophotometer (Thermo Scientific), and the integrity was evaluated using electrophoresis.

Construction and sequencing of DNA libraries

Illumina pair-end index libraries were constructed according to the manufacturer's protocol. Briefly, ~3 μg of DNA was randomly fragmented by nebulization with compressed nitrogen gas. Overhangs (5′ or 3′) of double-stranded DNA fragments were converted to blunt ends using T4 DNA polymerase and Klenow polymerase. An "A" base was added to the end of double-stranded DNA fragments using exo- Klenow polymerase, followed by ligation to adaptors with a "T" base overhang. After electrophoresis, DNA fragments of 500 bp on average were gel-purified. To minimize bias in library preparation, two DNA libraries were built for each sample. The adaptor-modified DNA fragments were loaded on an Illumina Cluster Station and underwent 10 cycles of bridge amplification PCR to generate sequencing template clusters on flow cells. Samples were processed on the HiSequation 2000 platform (Illumina) according to the manufacturer's instructions for template hybridization, isothermal amplification, linearization, blocking and denaturing, and hybridization of the sequencing primers. Base-calling was done using Illumina's pipeline, HiSeq Control (HCS) + OLB + GAPipeline-1.6 (Illumina), and the sequences of each lane were generated as 90-bp reads. Data were processed and analyzed according to a pipeline summarized in Supporting Information, Figure S9 and described in detail below.

Reference genome

The rat (*R. norvegicus*) reference genome (RGSC v3.4) was downloaded from the University of California at Santa Cruz database (http://genome.ucsc.edu/) along with data on gene annotation, ESTs, gaps, repeats, and position of the centromeres. Single nucleotide polymorphisms (SNPs) were downloaded from dbSNP build 136.

Read filtering and mapping

The raw data were refined using two filtering steps: (1) Contaminant filtering: adapter sequences may be introduced into raw reads during the library construction process. Therefore reads containing sequences similar to the adapter (mismatch ≤3) were considered contaminated and discarded, as were reads <30 bp in length. (2) Quality value filtering: to obtain high-quality data, reads with 40% or more low-confidence bases (quality value = 2) were discarded. All cleaned reads were mapped onto the BN rat reference genome using SOAP2.21 (Li *et al.* 2009b), allowing a maximum of five mismatches for each read. The alignment parameters were the following: -a –b –D –o -2 –u –m –x (-g) –l 32 –s 30 –v 3. Duplicated reads caused by PCR were removed using an in-house C++ script.

Detection of SNPs

To identify SNPs against the reference genome, the genotype probability of each site in DA and F344 was calculated using SOAPsnp (Li *et al.* 2009a), which is based on the Bayesian statistical model. A consensus sequence (CNS) was generated to contain the genotype with the highest

probability for each position. SNPs between the reference sequence and the CNS were considered high-quality SNPs when they fulfilled all of the following criteria: (1) quality value >20 (indicating an inferred base call accuracy >99%); (2) estimated copy number of flanking sequences <2; (3) minimum distance between adjacent SNPs of 5 bp; (4) at least six uniquely mapped reads supporting homozygous SNPs or three for each allele of heterozygous SNPs; and (5) a maximum depth of each site of 75 (depth value was limited to twice the mean depth to avoid incorrect SNP calls supported by reads in repeats). DA and F344 genomic DNA were extracted from male rats, so we considered all SNP sites in chromosome X to be hemizygous and required them to be covered by only two reads.

Detection of short insertions/deletions

The clean reads were realigned to the BN genome with SOAP2 set to tolerate gaps of up to 10 bp. Then we clustered mapped read pairs containing gaps in only one end to detect insertions/deletions (indels) of up to 5 bp. Candidate indels overlapping SNP sites were filtered out. The remaining candidate events were considered high-quality indels when supported by 15–55 reads.

Experimental validation of SNPs and indels

Primers were designed to cover 1045 variants (SNPs and indels) on the chromosome 4 locus *Cia3d* (Brenner *et al.* 2011) and on the chromosome 10 loci *Cia5a* and *Cia5d* (Brenner *et al.* 2005). PCR products were generated using AmpliTaq Gold (Life Technologies) and 10 ng of genomic DNA. Excess primers and dNTPs were removed from the PCR reaction by treatment with Exosap-IT (USB) according to the manufacturer's instructions. Samples were then diluted to 20–40 ng/μl and sequenced at Genewiz, Inc. (South Plainfield, NJ) using BigDye Terminator v.3.1 on a 3730xl capillary analyzer (Life Technologies). Base calls were manually determined using LaserGene v.8 (Dnastar, Madison, WI).

Detection of structural variation and copy-number variation candidates

We identified structural variation using the paired-end method (Wang *et al.* 2008). The accuracy of this method depends on the distribution of the insert size of the DNA library. A Perl script was written to compile the mean and the standard deviation of the insert sizes used for the paired-end mapping. Paired-end reads that could both be aligned but did not meet the insert size and/or orientation inferred from the reference genome were classified as abnormal paired-end reads. Regions supported by at least three abnormal paired-end reads and differing from the inferred insert size by at least 3 standard deviations were considered to contain structural variation. Abnormal paired-end reads were analyzed by clustering, and structural variants were categorized as insertions, tandem or dispersed duplications, deletions, and combinations of inversions and deletions.

Segmental duplication or deletion events are also evident as regions of increased or decreased copy number (Yoon *et al.* 2009). To locate copy-number variation (CNV) candidates using the alignment results, we first obtained the depth of each base along the reference genome using SOAPcoverage (http://soap.genomics.org.cn/). We then used CNVDetector (Chen *et al.* 2008), a program developed by BGI, to calculate the mean depth of 100-bp sliding windows along each chromosome and to select the candidate regions of CNV based on the difference of depth between each consecutive window and the overall mean. Events with a high absolute difference in depth (*i.e.*, outside the 0.75- to 1.25-fold range) and >10 kb were considered an effective CNV candidate. Some candidate regions had to be subdivided because of gaps (N-region) in the BN genome.

Simulation

To evaluate our optimal sequencing depth and the accuracy of our methods, we simulated short reads of different lengths using the BN genome. We also simulated mismatch sequencing errors using a sampling of the quality scores from the DA and F344 sequencing data, as well as SNPs, indels, and structural variants (separately and with occurrence rates of 1×10^{-3}, 1×10^{-4}, and 1×10^{-5}, respectively). The length of indels ranged from 1 to 10 bp and the length of structural variants ranged from 100 bp to 100 kb. The simulated reads were then realigned back to the whole BN genome. We used the rate of misplacement to calculate the sensitivity and specificity to detect SNPs, indels, and structural variation and how their detection rates were affected by coverage and quality scores.

GapCloser Tool

The GapCloser tool (http://soap.genomics.org.cn/) (Li *et al.* 2010) adopts a greedy algorithm to fill gaps. It extends contig ends iteratively by using reads overlapping with the contig end. Contig-end extension terminates when (1) the extended sequence overlaps with the other contig end at the other side of a gap, (2) an extended sequence with no overlap with the contig end at the other side of gap is 1000 bp longer than the size of the gap in the reference genome, or (3) no reads can be found to make a new round of extension. If extension of one strand fails to close a given gap, GapCloser will perform another extension on the complementary strand.

Construction of the DA and F344 genome drafts

The DA and F344 genomes were assembled using the reference-aided assembly method (RAM), a novel strategy for genome assembly based on resequencing data. RAM contains three main steps: (1) construction of semifinished genome, (2) independent *de novo* assembly to generate contigs and scaffolds, and (3) generation of the genome draft by anchoring scaffolds onto the semifinished genome.

Cleaned sequencing reads were aligned onto the BN reference genome using SOAPaligner (Li *et al.* 2009b) to construct DA and F344 CNS equal in length to the BN genome but

tolerating SNPs at a rate of 10^{-3}. Then gaps in each chromosome's CNS were closed with each line's own clean sequencing reads using GapCloser (Li *et al.* 2010). Each gap-closed CNS constituted a coordinated, semifinished genome.

To obtain the *de novo* genome assembly of each line, SOAPdenovo (Li *et al.* 2010) was used to reassemble clean reads and to generate contigs and scaffolds for DA and F344. Gaps between scaffolds were closed using GapCloser.

The final step to obtain the genome drafts of DA and F344 was anchoring the scaffolds onto the semifinished genome. To avoid scaffold contamination, only qualified scaffolds—>200 bp and containing <50% of Ns—were selected for anchoring. Tag sequences with a length of 100 bp and containing no Ns were extracted from each end of the qualified scaffolds, with additional tags extracted for >5000 bp. Tag sequences were mapped to the semifinished genomes using BLAST (Altschul *et al.* 1990), and the aligned tag sequences were filtered according to the following criteria: (a) *e*-value $<1 \times 10^{-40}$, (b) identity value >95, (c) alignment length >95, and (d) number of mismatches fewer than five. We used these qualified tag sequences to anchor the high-confidence scaffolds onto the semifinished genome and thus obtain the genome assembly for each strain.

To evaluate the accuracy of the DA and F344 genome drafts, we retrieved all 194,363 ESTs available in the rat genome and aligned them to the assembled drafts using BLAST, set to cover at least 95% of each EST. We also estimated the single-base error for each genome draft by comparing their sequence to the corresponding positions containing homozygous SNPs at the same strain's semifinished genome.

Data access

All reads have been deposited in the European Bioinformatics Institute (EBI)/NCBI Short Read Archive (accession no. SRA046343). All DA and F344 data have been released for public use and can be freely accessed at NCBI's Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra or at http://dx.doi.org/10.5524/100042). The data set includes all reads, semifinished genome sequences, genome drafts, annotation of variants including SNPs, short indels (1–5 bp), structural variations, and the bioinformatics tools used.

Results

Sequencing

Genomic DNA was extracted from the liver of male DA (DA/BklArbNsi, The Feinstein Institute for Medical Research) and F344 (F344/NHsd, Harlan Laboratories) rats using the phenol–chloroform–isoamyl alcohol method (Strauss 2001). Massively parallel whole-genome sequencing was performed using the Illumina HiSeq2000 sequencing platform. To minimize systematic bias in library preparation, for each genome we prepared two paired-end DNA libraries with a read length of 90 bp and insert sizes of 475–504 bp (Table S1). A total of 1.02 billion reads from the DA and 1.07 billion reads from the F344 genomes were generated, corresponding to 92.03 and 96.80 Gb of sequencing data, respectively. The proportion of high-quality data (*Q*-score ≥20) obtained for DA was 96.67% and for F344 was 97.63%.

The current assembly of the BN reference genome has an effective size of 2.57 Gb. Using the Short Oligonucleotide Alignment Program (SOAP) (Li *et al.* 2008), 83.87 Gb of DA and 88.18 Gb of F344 sequence—91.3% of each strain's reads—aligned with the BN genome. These reads covered 98.9% of the BN reference genome with at least one read and 98.0% with a sequencing depth of three or more reads and resulted in genome-wide average sequencing depths of 32.68-fold for DA and 34.36-fold for F344 reads (Table S2). The sequencing depth did not vary significantly between autosomes, indicating euploidy (Figure S1). Sequencing depth followed a Poisson distribution, and regions of lower depth correlated with extremes of GC content (Figure S2).

Regions of sequence ambiguity and breaks between contigs in the BN genome form 876,652 gaps that limit alignment with DA and F344 reads. These gaps contain 267.83 million positions of undetermined sequence (Ns). Using the GapCloser tool (Li *et al.* 2010) to bridge gaps with aligned reads, we were able to assign sequences to 59.31 million positions in DA and to 59.70 million positions in F344 and to effectively close 359,392 gaps in DA and 361,412 in F344 (Figure 1, Table S2).

SNPs

We used SOAPsnp (Li *et al.* 2010) to identify the SNP sets for each inbred line based on the alignment results of all sequencing data with the BN genome sequence. Unreliable sites were excluded from the analysis by filtering the SNPs for quality, copy number, distance between SNPs, number of supporting reads for each allele, and total depth. After filtering, we identified 2,964,158 high-quality nuclear DNA SNPs in DA and 2,973,513 in F344, compared with the BN genome (Figure 2A). We also identified 156 mitochondrial SNPs in DA and 163 in F344. Mitochondrial SNPs had a frequency of 9.8×10^{-3}.

We detected a total of 5,632,694 homozygous SNPs: 2,816,017 in the DA set and 2,816,677 in the F344 set. A total of 2,059,492 homozygous SNPs were polymorphic between DA and F344, and 1,786,600 homozygous SNPs were identical (Table 1, Figure 2A). The frequency of homozygous SNPs was 1.1×10^{-3} for both strains. More than 1.37 million homozygous SNPs were new and not represented in the dbSNP database (build 136). These novel SNPs included 502,994 SNPs with alleles unique to DA (F344 and BN carried the same allele), 496,368 SNPs with alleles unique to F344 (DA and BN carried the same allele), and 370,879 SNPs with alleles unique to BN (DA and F344 carried the same allele). A percentage of 39.38 of DA and F344 homozygous SNPs mapped to repeat regions, in agreement with the 40% interspersed repetitive DNA described in the rat genome (Gibbs *et al.* 2004). To estimate the accuracy of our SNP set, we sequenced three gene regions of

Figure 1 Genetic variation in the DA and F344 genomes. Distribution and frequency of (A) SNPs, (B) short insertion/deletions (InDel), (C) structural variants (SV), (D) copy-number variant (CNV) candidates, and (E) filled gaps along the rat genome (numbers outside the circle represent each chromosome), using the BN genome as reference, are shown. The F344 genome is in dark blue, and DA is in light blue.

chromosomes 4 and 10 using the Sanger method and confirmed 99.68% of 933 homozygous SNPs (Table S3).

A percentage of 5.14 of the SNPs were detected in the heterozygous state, with a genomic distribution rate of 5.9×10^{-5}. Heterozygous SNPs were predominantly detected in regions with high alignment rates (median sequencing depth; homozygous SNPs = 28-fold for DA and 27-fold for F344, heterozygous SNPs = 40-fold for both strains; Figure S3) and mapped to repeat regions at a rate significantly higher than homozygous SNPs (60.02% *vs.* 39.38%, respectively; $P < 0.001$, chi-square test), suggesting that improperly aligned reads might account for some of heterozygous SNP calls. To confirm heterozygous SNPs in these highly inbred strains, 45 heterozygous SNPs were sequenced with Sanger methodology, and 6 (13.33%) were in fact homozygous SNPs, while 39 (86.87%) were false SNP calls (Table S3). Therefore, heterozygous SNPs were not included in subsequent analyses.

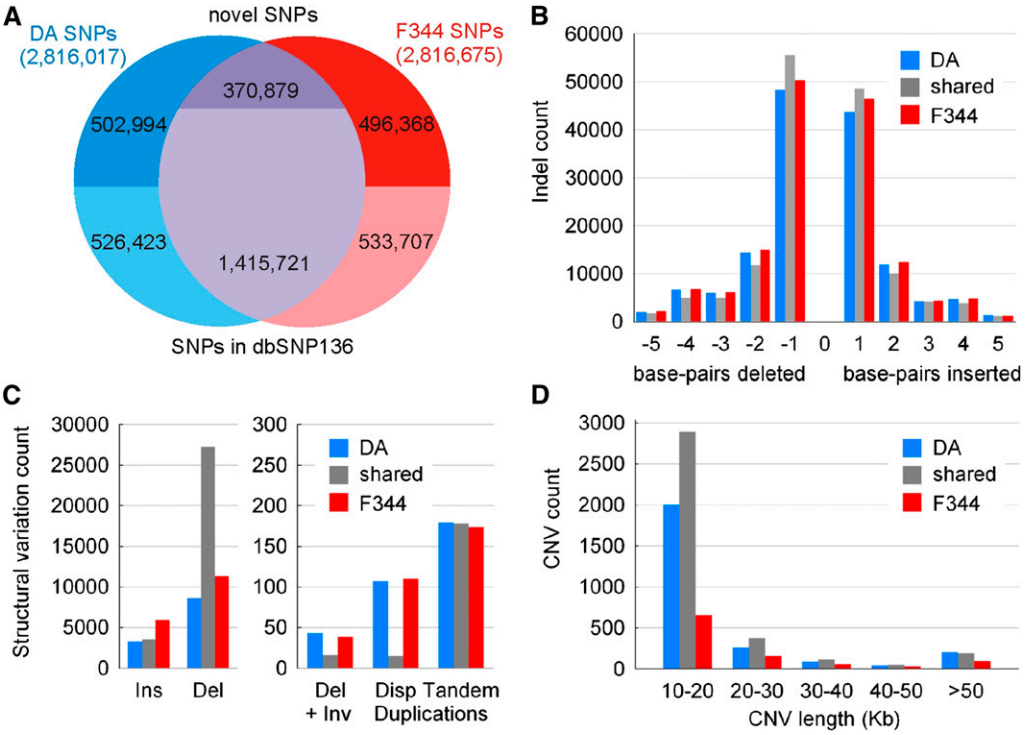

Figure 2 Variation between DA, F344, and BN. (A) DA (blue, light blue) and F344 (red, pink) each had 1.03 million unique homozygous SNPs and shared alleles for 1,786,600 homozygous SNPs (purple, light purple). Forty-eight percent of the strain-specific SNPs and 20% of the shared SNPs were novel and were not present in dbSNP v.136, including 502,994 SNPs with alleles unique to DA (blue), 496,368 SNPs with alleles unique to F344 (red), and 370,879 SNPs for which DA and F344 had the same alleles (purple). (B) DA and F344 had the same alleles for 146,502 homozygous indels; 143,058 were unique to DA, and 149,621 were unique to F344. Most homozygous indels were 1 bp long, and distribution of short insertions and deletions according to size was similar in DA and F344. (C) DA and F344 had the same alleles covering 80% of 30,978 structural variants; 15,151 were unique to DA, and 17,575 were unique to F344. The most frequent structural variants were deletions and insertions, followed by tandem and dispersed duplications. (D) There were 2594 CNV candidates unique to DA, 3611 unique to F344, and 994 identical between DA and F344. Most CNV candidates were in the 10- to 20-kb range (blue: DA; red: F344; purple or gray: shared).

Many of the homozygous SNPs detected had the potential to impact gene function, including 422 SNPs predicted to cause loss or gain of start codons, 231 SNPs impacting splicing sites, and 140 SNPs causing loss or gain of stop codons. Additionally, 15,477 SNPs were nonsynonymous, mapping to 3174 Refseq genes and 4724 Ensembl genes in the DA genome and to 3074 Refseq genes and 4632 Ensembl genes in the F344 genome (Table S4 and Table S5). A total of 4.3 million SNPs (88.4% of all homozygous SNPs) were intergenic, intronic, or synonymous.

Indels

We detected indels using SOAP2 (Li *et al.* 2009b). Indels were defined as alignment gaps of up to 5 bp, supported by three or more nonredundant pairs of reads and present in at least one-third of reads for autosomic indels or in all of the reads for X-chromosome indels. In total, we identified 299,532 indels in DA and 305,705 in F344 (Figure 2B, Table 1). Of the indels, 96.8% were homozygous and had a genomic distribution rate of 1.1×10^{-4}, and 3.2% were heterozygous and had a genomic distribution rate of 3.8×10^{-6}. Insertions or deletions of 1 bp accounted for 67.76% of the indels (Figure S4). Sanger sequencing confirmed 100% of 77 homozygous indels tested (Table S6).

While DA and F344 shared 146,502 homozygous indels, 292,679 were polymorphic between these two strains with 143,058 indels unique to DA and 149,621 unique to F344 (Figure 2B). Most indels were intronic or intergenic (Figure S5), but 605 homozygous indels were predicted to cause codon insertions/deletions or frameshift in coding genes.

Table 1 SNPs and indels in the DA and F344 consensus assemblies

	SNPs				Short indels	
Sample	Homozygous	Heterozygous	Known[a,b]	Novel[b]	Homozygous	Heterozygous
DA[c]	1,029,417	91,848	526,423	502,994	143,058	9,461
F344[d]	1,030,075	100,545	533,707	496,368	149,621	9,071
Shared[e]	1,786,600	56,293	1,415,721	370,879	146,502	511
Total	3,846,092	248,686	2,475,851	1,370,241	439,181	19,043

[a] Homozygous SNPs mapping to SNP positions in dbSNP 136.
[b] Include only homozygous SNPs.
[c] Allele is unique to DA (F344 allele = BN allele)
[d] Allele is unique to F344 (DA allele = BN allele)
[e] DA and F344 have the same allele, which is different from BN.

Table 2 Genetic variation annotation of DA and F344

	SNPs[a]			Indels			Structural variant	
	DA	F344	Shared	DA	F344	Shared	DA	F344
Intergenic region	658,646	660,019	1,150,467	85,753	91,326	97,035	—	—
Intron	485,087	489,738	837,376	70,958	73,231	75,716	17516	19607
Downstream (up to 5 kb)	71,596	71,646	123,030	10,565	10,620	10,722	—	—
Upstream (up to 5 kb)	70,831	71,107	119,219	10,109	10,413	9,853	—	—
3′ UTR	4,094	3,766	6,955	685	723	811	291	305
5′ UTR	752	647	1,061	48	37	47	226	239
Start gained in 5′ UTR	128	122	157	—	—	—	—	—
Coding sequences and splice sites							2572	2829
Synonymous coding	7,146	6,981	12,104	—	—	—	—	—
Nonsynonymous coding	4,230	4,060	7,187	—	—	—	—	—
Frameshift	—	—	—	156	135	217	—	—
Start lost	1	6	8	—	—	—	—	—
Stop gained	35	35	59	—	—	—	—	—
Stop lost	4	1	6	—	—	—	—	—
Splice-site acceptor	30	31	59	35	21	58	—	—
Splice-site donor	26	27	58	32	16	62	—	—
Synonymous stop	3	5	10	—	—	—	—	—
Nonsynonymous start	1	0	2	—	—	—	—	—
Codon deletion	—	—	—	19	10	14	—	—
Codon change + codon deletion	—	—	—	10	2	6	—	—
Codon insertion	—	—	—	12	4	5	—	—
Codon change plus codon insertion	—	—	—	7	4	4	—	—
Within noncoding gene								
Nonsynonymous coding	434	428	866	—	—	—	—	—
Synonymous coding	232	228	395	—	—	—	—	—
Stop gained	24	14	27	—	—	—	—	—
Stop lost	6	15	18	—	—	—	—	—
Synonymous stop	3	0	6	—	—	—	—	—
Start lost	3	1	1	—	—	—	—	—
Synonymous start	1	0	0	—	—	—	—	—
Codon change + codon deletion	—	—	—	3	0	2	—	—
Codon deletion	—	—	—	2	1	2	—	—
Codon insertion	—	—	—	1	0	2	—	—
Codon change + codon insertion	—	—	—	0	1	0	—	—

Calculated using SNPEff v.1.9.5 (Cingolani *et al.* 2012) and Ensembl's *R. norvegicus* build 3.4.64.
[a] Homozygous SNPs.

Of these, 204 transcript-affecting indels were unique to DA, 155 were unique to F344, and 246 were found in both DA and F344 (Table 2).

The frequencies of homozygous SNPs along the DA and F344 genomes varied from 0 to 3 $\pm$ 10^{-3} and strongly correlated with that of homozygous indels (Figure 1, Figure S6), suggesting a progressive increase in variation density from shared haplotypes.

Structural variation

We used paired-end alignment to identify structural variation. Regions containing structural variants were detected when read pairs aligned to the reference genome abnormally—differing in orientation and/or inferred insert size with the support of at least three read pairs. We identified a total of 58,174 structural variants: 12,151 unique to DA, 17,575 unique to F344, and 30,978 present in both DA and F344 (Figure 2C, Figure S7, and Figure S8). Deletions and insertions >5 bp were the most frequently detected class of structural variants, followed by tandem duplication, dispersed duplication, and combined insertion–deletion. Structural variants overlapping coding sequences have a high potential to disrupt the function of those genes. In total, 2572 structural variants in the DA and 2829 in the F344 genomes overlapped coding sequences of Ensembl genes (Table 2). And 1398 structural variants in the DA and 1502 in the F344 genomes overlapped coding sequences of RefSeq genes (Table S7).

Based on the mean depth of 100-bp sliding windows along each chromosome, we detected 7199 candidate regions of copy-number variation: 2594 unique to DA, 3691 unique to F344, and 994 in both DA and F344 (Figure 1D, Table S8). Seventy-seven percent of copy-number variant candidates were in the 10- to 20-kb range (Figure 2D).

Sensitivity and specificity

To evaluate the accuracy of read mapping, we generated a variation of the BN genome containing SNPs, indel, and

structural variants with frequencies similar to those observed in the DA and F344 sequencing data. We also simulated short reads of different lengths containing mismatch sequencing errors and quality scores similar to those in the DA and F344 sequences. We then aligned the simulated reads back to the BN reference assembly to quantify the precision of alignment for the detection of variants.

For an average 35-fold coverage with simulated reads, sensitivity for SNP detection was inversely proportional to read-quality threshold and varied from slightly over 96% for reads with $Q = 22$, to 96.6% and for reads with $Q = 15$. Specificity for SNP detection was more dependent on sequencing coverage, and it increased from 99.78% with a depth of 1-fold to 99.82% with a depth of 5-fold to 99.94% with a depth of 10-fold (Figure 3A). Sensitivity and specificity for indel detection were similar to those of the simulated SNPs (data not shown).

Specificity for the detection of structural variants increased sharply with the number of supporting reads from 47% (1–2 reads) to 91% (3 reads) and continued increasing at a lower rate to plateau at 99.68% with seven or more reads (Figure 3B). Sensitivity for the detection of structural variants was inversely correlated with the number of supporting reads, sharply declining from 62.1% (1 read) to 49.92% (3 reads) and then to 47.34% (10 reads).

Construction of the DA and F344 genome drafts

To generate the DA and F344 genome drafts, we created a new strategy for *de novo* genome assembly using NGS data: the reference-aided assembly method (Figure 4). Briefly, semifinished genomes were generated for each strain by aligning their reads to the BN genome using SOAPaligner (Li *et al.* 2009a) to form a consensus sequence, followed by assembly of reads to bridge gaps in the BN genome using GapCloser (Li *et al.* 2010). In parallel, contigs and scaffolds were independently assembled for each strain using SOAPdenovo (Li *et al.* 2010), followed by closure of gaps between scaffolds using GapCloser. Finally, sequences from both ends of each scaffold were mapped onto each coordinated semifinished genome using BLAST to anchor the scaffolds and obtain the DA and F344 genome drafts.

The DA and F344 genome drafts include 2,616,053,766 and 2,615,410,193 effective bases and are 1.94% and 1.91% larger than the BN genome, respectively. The DA and F344 genome drafts also contain 49.76 and 49.11 million novel base pairs bridging 391,057 and 401,069 gaps of the BN genome. Of the novel base pairs, 20.47 million (41.13%) and 19.35 (39.41%) million base pairs are in novel scaffolds. And 2.55% and 2.42% more reads could be mapped to each coordinated draft compared with the consensus sequences (Table 3).

We evaluated the quality of DA and F344 genome drafts using two methods. First, we retrieved all 194,363 ESTs available in the rat genome and aligned them to the assembled drafts using BLAST to cover at least 95% of each EST. Of the ESTs, 97.97% aligned to each *de novo* assembly, and 836 (0.43%) and 1088 (0.56%) ESTs aligned exclusively to novel scaffolds in DA and F344, respectively (Table S9). Second, we estimated the single-base error rates for *de novo* assemblies by comparing the draft genome sequences to corresponding positions containing homozygous SNPs in the semifinished genome of each strain. The estimated single-base error for these two newly assembled drafts was 3.06×10^{-5} for DA and 2.99×10^{-5} for F344 (Table S10).

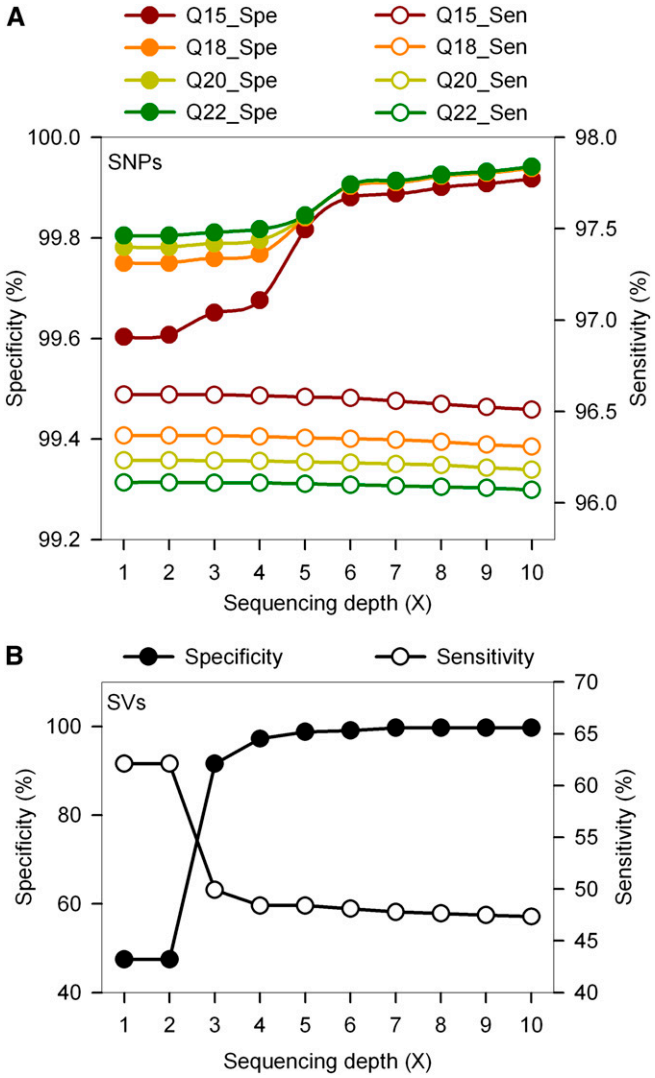

Figure 3 Accuracy of variant detection. We produced a copy of the BN rat genome with a read coverage of 35-fold and aligned the simulated reads back onto the RGSC3.4 genome scaffold to measure the rate of misplacement. Simulated reads contained simulated mismatch sequencing errors, SNPs, indels, and structural variants to the RGSC3.4 reference at rates similar to those detected in the DA and F344 genomes. (A) SNP detection sensitivity (open circles) was inversely proportional to the read quality threshold. The SNP detection specificity (solid circles) was more dependent on the number of supporting reads. (B) The detection sensitivity for structural variants (open circles) was inversely proportional to the number of supporting reads. The detection specificity for structural variants (solid circles) increased with the number of supporting reads and remained >99% with six or more reads.

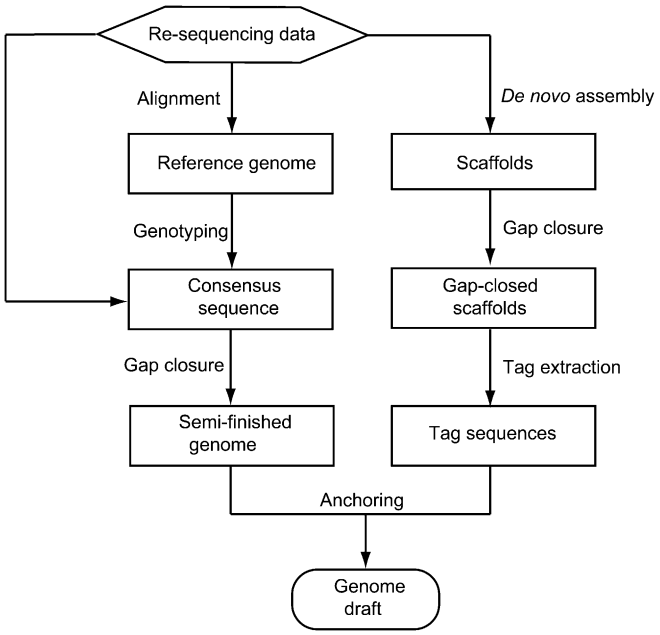

Figure 4 Construction of the DA and F344 genome drafts using the Reference-Aided Assembly Method. The strategy to construct the DA and F344 genome drafts from NGS data consisted of (1) generating a coordinated, semifinished genome, (2) producing a *de novo* assembly, and (3) anchoring the *de novo* assembly onto the semifinished genome. Each semifinished genome was created by alignment of reads onto the BN reference genome, inference of a consensus sequence, and closure of gaps (left arm). In parallel, reads were assembled independently into scaffolds, followed by closure of gaps and extraction of tag sequences (right arm). Tag sequences were then mapped onto the semifinished genome using BLAST, anchoring the affiliated scaffolds to finalize each genome draft (bottom of diagram).

Discussion

DA and the F344 rats have unique dichotomous phenotypes that have been used to better understand development, human physiology, and disease. DA rats are highly susceptible to autoimmunity, including models of rheumatoid arthritis, multiple sclerosis, and uveitis (Dahlman *et al.* 1998; Sun *et al.* 1999; Wilder *et al.* 1999). DA rats are also susceptible to bladder and tongue carcinomas (Kitano *et al.* 1992), have reduced variation in circadian corticosteroid production (Brodkin *et al.* 1999), and are more easily addicted to morphine (Brodkin *et al.* 1999). F344 rats, on the other hand, are typically resistant to the above conditions, but are susceptible to chemically induced hepatocarcinoma and lymphoma (Lu *et al.* 1999; De Miglio *et al.* 2006) and have decreased bone mineral density (Turner *et al.* 2001). Genetic variation between these strains accounts for most of such strain-specific phenotypes. Therefore, sequencing the DA and F344 genomes constitutes a major step toward identifying the genetic causes and pathogenic processes underlying these traits and models of human diseases such as rheumatoid arthritis, multiple sclerosis, and cancer (Table S11) and should facilitate the development of novel disease treatments and biomarkers, as well as new pathways to be tested for disease prevention.

The sequencing of the DA and F344 genomes identified a large number of variants between each of these two strains and BN. The 5.6 million SNPs identified in the DA and F344 genomes increased the total number of known SNPs between these two strains and BN by 150-fold from 19,326 (Saar *et al.* 2008) to 2.2 million SNPs between DA and F344 and 2.9 million SNPs between BN and each of the other two strains. Furthermore, 1.37 million SNPs and 0.44 million indels were novel. The addition of these novel variants significantly expands known variation in the rat genome.

A large number of variants were predicted to significantly disrupt gene structure. High-impact variants included deletion of coding sequences, loss of start codons, premature stops, frameshifts, codon insertions/deletions, nonsynonymous SNPs, and changes at splicing sites. In addition to these effects on gene structure, other variants can potentially alter gene expression. Upstream and 5′-UTR variants can disrupt epigenetic regulation and transcription factor-binding sites, 3′-UTR variants can modify messenger RNA stability (Boffa *et al.* 2008), intronic SNPs can influence expression breadth (Park *et al.* 2012), and synonymous SNPs can affect translation efficiency (Plotkin and Kudla 2011). The DA and F344 genome sequencing provides a detailed framework for future studies aimed at characterizing how these variants alter gene function.

At 32- and 34-fold redundancy, the DA and F344 genomes were assembled at a sequencing depth almost five times that of the BN genome (Gibbs *et al.* 2004) and three times that of the SHR genome (Atanur *et al.* 2010). The DA and F344 genome assemblies also used stringent quality criteria to define variants. The combination of high-quality and high-sequencing depth resulted in increased accuracy to detect SNPs and indels, as was confirmed with Sanger sequencing and *in silico* simulations. The frequency of SNPs was 10-fold higher than that of indels, revealing a SNP/indel ratio similar to that of other resequencing projects (Ahn *et al.* 2009; Atanur *et al.* 2010). SNPs in mitochondrial DNA were 8.9-fold

Table 3 Construction of the DA and F344 genome drafts

	Genome size (bp)			Novel scaffolds (bp)		Reads mapped	
	Total	Effective[a]	Gaps[b]	Total	Effective[a]	DA (%)	F344 (%)
BN	2,834,127,293	2,566,294,765	876,652	—	—	91.49	91.84
DA	2,798,712,224	2,616,053,766	485,595	20,558,331	20,465,987	94.03	—
F344	2,793,938,348	2,615,410,193	475,583	19,441,505	19,355,322	—	94.27

[a] Genome length without Ns.
[b] Number of gaps in each genome draft.

more frequent than in nuclear DNA in agreement with its 9–25 times higher mutation rate (Lynch *et al.* 2006).

A small percentage of the SNPs (5%) and indels (3%) were detected as heterozygous, and Sanger-based resequencing showed that a fraction were in fact homozygous SNPs, while the majority were false calls. Misalignment of reads mapping to repeats or highly homologous segmental duplications and sequencing errors may have contributed to false detection of heterozygous SNPs in the DA and F344 genomes. Eventual residual heterozygosity cannot be entirely excluded; it might result from selection against recessive alleles that are embryonically lethal or are associated with infertility or unproductive breeding behavior (Bailey 1977; Saar *et al.* 2008).

The importance of copy-number and copy-neutral structural variants in the genome has only recently begun to be understood (Korbel *et al.* 2007). Structural variation accounts for an even higher proportion of the genetic diversity between individuals than SNPs (Li *et al.* 2011) and has been associated with disease in both rats and humans (Aitman *et al.* 2006). Copy number variants can also correlate with levels of gene expression in rats (Guryev *et al.* 2008; Charchar *et al.* 2010) and have been estimated to account for 20% of expression differences in humans (Stranger *et al.* 2007). We identified variants in the DA and F344 genomes that caused duplications, deletions, or potential disruptions of the structure of >2500 genes. This frequency of potentially gene-disrupting structural/copy-number variants has also been seen in other interstrain comparisons such as that described between DBA and B6 mice (Quinlan *et al.* 2010). Insert size of libraries can be a limiting factor for the identification of insertion events in NGS (Pang *et al.* 2010). And in fact, using the simulated reads we estimated that our method of detecting structural variants had a sensitivity of 45–50%. Therefore, DA and F344 structural variants are most likely underrepresented.

DA and F344 rats shared alleles for 60% of the SNPs, 50% of the indels, and 70% of the structural variants, an indication of the phylogenetic proximity between these two strains. The high levels of allele sharing between DA and F344 are in agreement with a previous observation that BN was the most divergent of 167 commonly used laboratory inbred strains, including DA and F344 (Saar *et al.* 2008).

We devised and employed a new strategy to generate the first *de novo* assembly of a rat genome using NGS technology data. As a result, the DA and F344 genome drafts are more extensive and more complete than the BN genome and should facilitate the study of discrepancies with genetic maps (Saar *et al.* 2008) and areas of sequencing collapse (Guryev *et al.* 2008). The new DA and F344 genome drafts contain 49 million base pairs of novel sequence each, nearly half the number of gaps present in the BN genome, and ~1000 ESTs uniquely mapped to novel scaffolds of each strain.

The BN and SHR are the only rat nuclear genomes drafted to date. As additional rat genomes become available, investigators will be able to construct detailed haplotype maps, a key resource for both targeted and genome-wide studies in the rat. Sequencing additional genomes will help resolve regions of poor coverage in the BN and other rat genomes, as well as alignment and sequencing errors and undetected duplications.

Over 615 inbred rat strains and substrains are presently registered at the Rat Genome Database. These strains are an important resource for gene identification and studies of gene function and are currently being used by several laboratories worldwide. The SNPs, indels, and structural variants reported here compose a large collection of new informative markers that can be used to increase the precision of genetic mapping and genotype-guided breeding, as well as for studies in advanced intercross lines and for genome-wide association studies using heterogeneous stocks. Indeed, with an average density of one SNP per 0.86 kb, SNPs identified in this study will facilitate mapping at a resolution 100-fold higher than with previously available SNPs (Saar *et al.* 2008).

Acknowledgments

Funded by the National Institutes of Health grants R01-AR46213, R01-AR052439 (NIAMS) and R01-AI54348 (NIAID) to Dr. P. Gulko, and by The Shenzhen Municipal Government of China (grants JC201005260191A and CXB201108250096A) and from the National Gene Bank Project of China to Dr. Wang Jun.

Literature Cited

Ahn, S. M., T. H. Kim, S. Lee, D. Kim, H. Ghang *et al.*, 2009 The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19: 1622–1629.

Aitman, T. J., R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson *et al.*, 2006 Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439: 851–855.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Atanur, S. S., I. Birol, V. Guryev, M. Hirst, O. Hummel *et al.*, 2010 The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. Genome Res. 20: 791–803.

Bailey, D. W., 1977 Genetic drift: the problem and its possible solution by frozen-embryo storage. Ciba Found. Symp. 52: 291–303.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.

Biesiadecki, B. J., P. H. Brand, P. J. Metting, L. G. Koch, and S. L. Britton, 1998 Phenotypic variation in strength among eleven inbred strains of rats. Proc. Soc. Exp. Biol. Med. 219: 126–131.

Boffa, M. B., D. Maret, J. D. Hamill, N. Bastajian, P. Crainich *et al.*, 2008 Effect of single nucleotide polymorphisms on expression of the gene encoding thrombin-activatable fibrinolysis inhibitor: a functional analysis. Blood 111: 183–189.

Brenner, M., H. C. Meng, N. C. Yarlett, B. Joe, M. M. Griffiths *et al.*, 2005 The non-MHC quantitative trait locus Cia5 contains three major arthritis genes that differentially regulate disease severity, pannus formation, and joint damage in collagen- and pristane-induced arthritis. J. Immunol. 174: 7894–7903.

Brenner, M., T. Laragione, N. C. Yarlett, and P. S. Gulko, 2007 Genetic regulation of T regulatory, CD4, and CD8 cell numbers by the arthritis severity loci Cia5a, Cia5d, and the MHC/Cia1 in the rat. Mol. Med. 13: 277–287.

Brenner, M., T. Laragione, A. Shah, A. Mello, E. F. Remmers *et al.*, 2011 Identification of two new arthritis severity loci that regulate levels of autoantibodies, IL-1beta and joint damage. Arthritis Rheum. 64: 1369–1378.

Brodkin, E. S., T. A. Kosten, C. N. Haile, G. R. Heninger, W. A. Carlezon, Jr. *et al.*, 1999 Dark Agouti and Fischer 344 rats: differential behavioral responses to morphine and biochemical differences in the ventral tegmental area. Neuroscience 88: 1307–1315.

Charchar, F. J., M. Kaiser, A. J. Bingham, N. Fotinatos, F. Ahmady *et al.*, 2010 Whole genome survey of copy number variation in the spontaneously hypertensive rat: relationship to quantitative trait loci, gene expression, and blood pressure. Hypertension 55: 1231–1238.

Chen, P. A., H. F. Liu, and K. M. Chao, 2008 CNVDetector: locating copy number variations using array CGH data. Bioinformatics 24: 2773–2775.

Cingolani, P., V. M. Patel, M. Coon, T. Nguyen, S. J. Land *et al.*, 2012 Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet. 3: 35.

Dahlman, I., J. C. Lorentzen, K. L. de Graaf, A. Stefferl, C. Linington *et al.*, 1998 Quantitative trait loci disposing for both experimental arthritis and encephalomyelitis in the DA rat: impact on severity of myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis and antibody isotype pattern. Eur. J. Immunol. 28: 2188–2196.

De Miglio, M. R., P. Virdis, D. F. Calvisi, M. Frau, M. R. Muroni *et al.*, 2006 Mapping a sex hormone-sensitive gene determining female resistance to liver carcinogenesis in a congenic F344. BN-Hcs4 rat. Cancer Res. 66: 10384–10390.

Fujimoto, A., H. Nakagawa, N. Hosono, K. Nakano, T. Abe *et al.*, 2010 Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat. Genet. 42: 931–936.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren *et al.*, 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521.

Guryev, V., K. Saar, T. Adamovic, M. Verheul, S. A. van Heesch *et al.*, 2008 Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. 40: 538–545.

Ishih, A., 1994 Worm burden and mucosal mast cell response in DA and F344/N rat strains infected with Hymenolepis diminuta. Int. J. Parasitol. 24: 295–298.

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

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong *et al.*, 2011 Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477: 289–294.

Kim, J. I., Y. S. Ju, H. Park, S. Kim, S. Lee *et al.*, 2009 A highly annotated whole-genome sequence of a Korean individual. Nature 460: 1011–1015.

Kitano, M., H. Hatano, and H. Shisa, 1992 Strain difference of susceptibility to 4-nitroquinoline 1-oxide-induced tongue carcinoma in rats. Jpn. J. Cancer Res. 83: 843–850.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert *et al.*, 2007 Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420–426.

Laragione, T., N. C. Yarlett, M. Brenner, A. Mello, B. Sherry *et al.*, 2007 The arthritis severity quantitative trait loci Cia4 and Cia6 regulate neutrophil migration into inflammatory sites and levels of TNF-alpha and nitric oxide. J. Immunol. 178: 2344–2351.

Laragione, T., M. Brenner, A. Mello, M. Symons, and P. S. Gulko, 2008 The arthritis severity locus Cia5d is a novel genetic regulator of the invasive properties of synovial fibroblasts. Arthritis Rheum. 58: 2296–2306.

Li, G., L. Ma, C. Song, Z. Yang, X. Wang *et al.*, 2009 The YH database: the first Asian diploid genome database. Nucleic Acids Res. 37: D1025–D1028.

Li, R., Y. Li, K. Kristiansen, and J. Wang, 2008 SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713–714.

Li, R., Y. Li, X. Fang, H. Yang, J. Wang *et al.*, 2009a SNP detection for massively parallel whole-genome resequencing. Genome Res. 19: 1124–1132.

Li, R., C. Yu, Y. Li, T. W. Lam, S. M. Yiu *et al.*, 2009b SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.

Li, R., H. Zhu, J. Ruan, W. Qian, X. Fang *et al.*, 2010 De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20: 265–272.

Li, Y., H. Zheng, R. Luo, H. Wu, H. Zhu *et al.*, 2011 Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat. Biotechnol. 29: 723–730.

Lu, L. M., H. Shisa, J. Tanuma, and H. Hiai, 1999 Propylnitrosourea-induced T-lymphomas in LEXF RI strains of rats: genetic analysis. Br. J. Cancer 80: 855–861.

Lynch, M., B. Koskella, and S. Schaack, 2006 Mutation pressure and the evolution of organelle genomic architecture. Science 311: 1727–1730.

Pang, A. W., J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq *et al.*, 2010 Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11: R52.

Park, J., K. Xu, T. Park, and S. V. Yi, 2012 What are the determinants of gene expression levels and breadths in the human genome? Hum. Mol. Genet. 21: 46–56.

Plotkin, J. B., and G. Kudla, 2011 Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12: 32–42.

Potenza, M. N., E. S. Brodkin, B. Joe, X. Luo, E. F. Remmers *et al.*, 2004 Genomic regions controlling corticosterone levels in rats. Biol. Psychiatry 55: 634–641.

Quinlan, A. R., R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang *et al.*, 2010 Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 20: 623–635.

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.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley *et al.*, 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.

Strauss, W. M., 2001 Preparation of genomic DNA from mammalian tissue. Curr. Protoc. Mol. Biol. Chapter 2: Unit 2.2. DOI: 10.1002/0471142727.mb0202s42.

Sun, S. H., P. B. Silver, R. R. Caspi, Y. Du, C. C. Chan *et al.*, 1999 Identification of genomic regions controlling experimental autoimmune uveoretinitis in rats. Int. Immunol. 11: 529–534.

Suzuki, T., A. Ishih, H. Kino, F. W. Muregi, S. Takabayashi *et al.*, 2006 Chromosomal mapping of host resistance loci to Trichinella spiralis nematode infection in rats. Immunogenetics 58: 26–30.

Terner, J. M., A. C. Barrett, L. M. Lomas, S. S. Negus, and M. J. Picker, 2006 Influence of low doses of naltrexone on morphine antinociception and morphine tolerance in male and female rats of four strains. Pain 122: 90–101.

Tong, P., J. G. Prendergast, A. J. Lohan, S. M. Farrington, S. Cronin *et al.*, 2010 Sequencing and analysis of an Irish human genome. Genome Biol. 11: R91.

Tordoff, M. G., L. K. Alarcon, and M. P. Lawler, 2008 Preferences of 14 rat strains for 17 taste compounds. Physiol. Behav. 95: 308–332.

Turner, C. H., R. K. Roeder, A. Wieczorek, T. Foroud, G. Liu *et al.*, 2001 Variability in skeletal mass, structure, and biomechanical properties among inbred strains of rats. J. Bone Miner. Res. 16: 1532–1539.

van Den Brandt, J., P. Kovacs, and I. Kloting, 2000 Metabolic variability among disease-resistant inbred rat strains and in comparison with wild rats (Rattus norvegicus). Clin. Exp. Pharmacol. Physiol. 27: 793–795.

van Wijngaarden, P., H. M. Brereton, D. J. Coster, and K. A. Williams, 2007 Genetic influences on susceptibility to oxygen-induced retinopathy. Invest. Ophthalmol. Vis. Sci. 48: 1761–1766.

Wang, J., W. Wang, R. Li, Y. Li, G. Tian *et al.*, 2008 The diploid genome sequence of an Asian individual. Nature 456: 60–65.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen *et al.*, 2008 The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876.

Wilder, R. L., E. F. Remmers, Y. Kawahito, P. S. Gulko, G. W. Cannon *et al.*, 1999 Genetic factors regulating experimental arthritis in mice and rats. Curr. Dir. Autoimmun. 1: 121–165.

Worley, K. C., G. M. Weinstock, and R. A. Gibbs, 2008 Rats in the genomic era. Physiol. Genomics 32: 273–282.

Worthey, E. A., A. J. Stoddard, and H. J. Jacob, 2010 Sequencing of the rat genome and databases. Methods Mol. Biol. 597: 33–53.

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane *et al.*, 2011 Sequence-based characterization of structural variation in the mouse genome. Nature 477: 326–329.

Yoon, S., Z. Xuan, V. Makarov, K. Ye, and J. Sebat, 2009 Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19: 1586–1592.

Zhang, Y., X. Lin, K. Koga, K. Takahashi, H. M. Linge *et al.*, 2011 Strain differences in alveolar neutrophil infiltration and macrophage phenotypes in an acute lung inflammation model. Mol. Med. 17: 780–789.

Communicating editor: T. R. Magnuson

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153049/-/DC1

Whole-Genome Sequences of DA and F344 Rats with Different Susceptibilities to Arthritis, Autoimmunity, Inflammation and Cancer

Xiaosen Guo, Max Brenner, Xuemei Zhang, Teresina Laragione, Shuaishuai Tai, Yanhong Li, Junjie Bu, Ye Yin, Anish A. Shah, Kevin Kwan, Yingrui Li, Wang Jun, and Pércio S. Gulko

DOI: 10.1534/genetics.113.153049

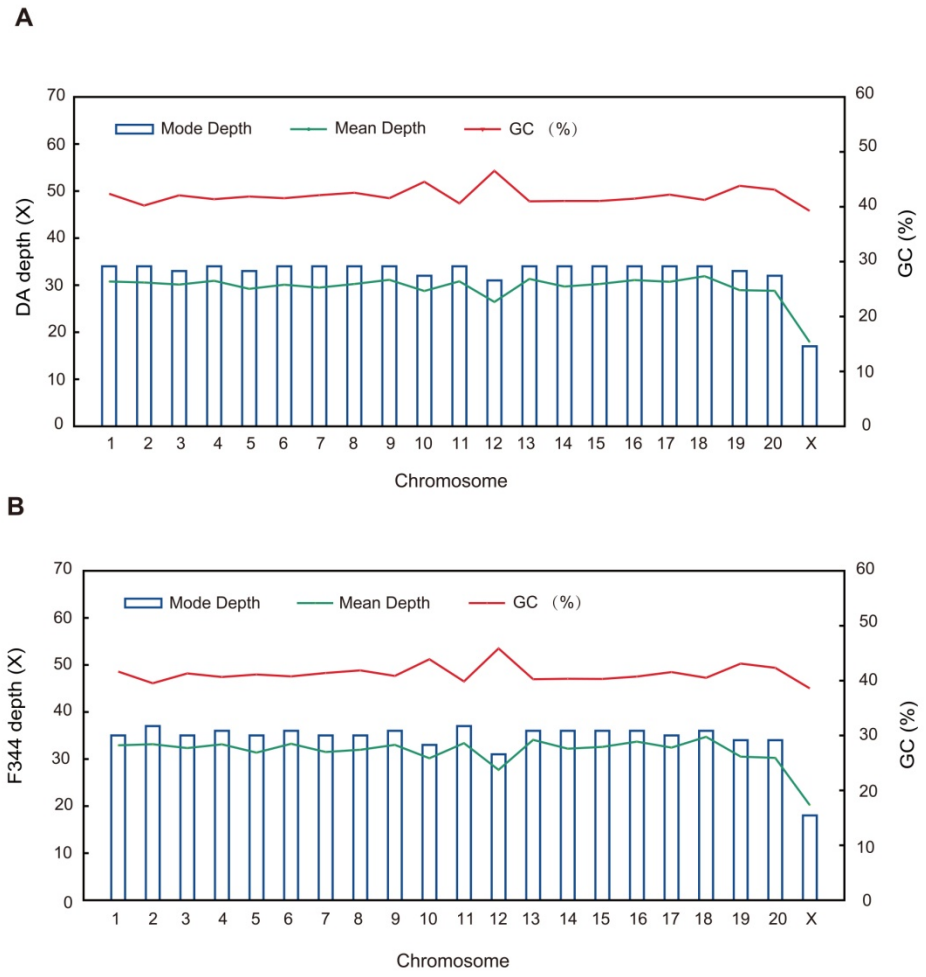

Figure S1 Distribution of sequencing depth per chromosome. The sequencing depth of each chromosome among DA **(A)** and F344 **(B)** autosomes did not vary significantly, indicating euploidy. Sequencing depth of the X-chromosome is approximately half that of autosomes due to haploidy. The slight decrease in coverage on chromosomes 10 and 12 was associated with the higher overall GC content of those chromosomes. M: mitochondrial chromosome, Un: position unknown

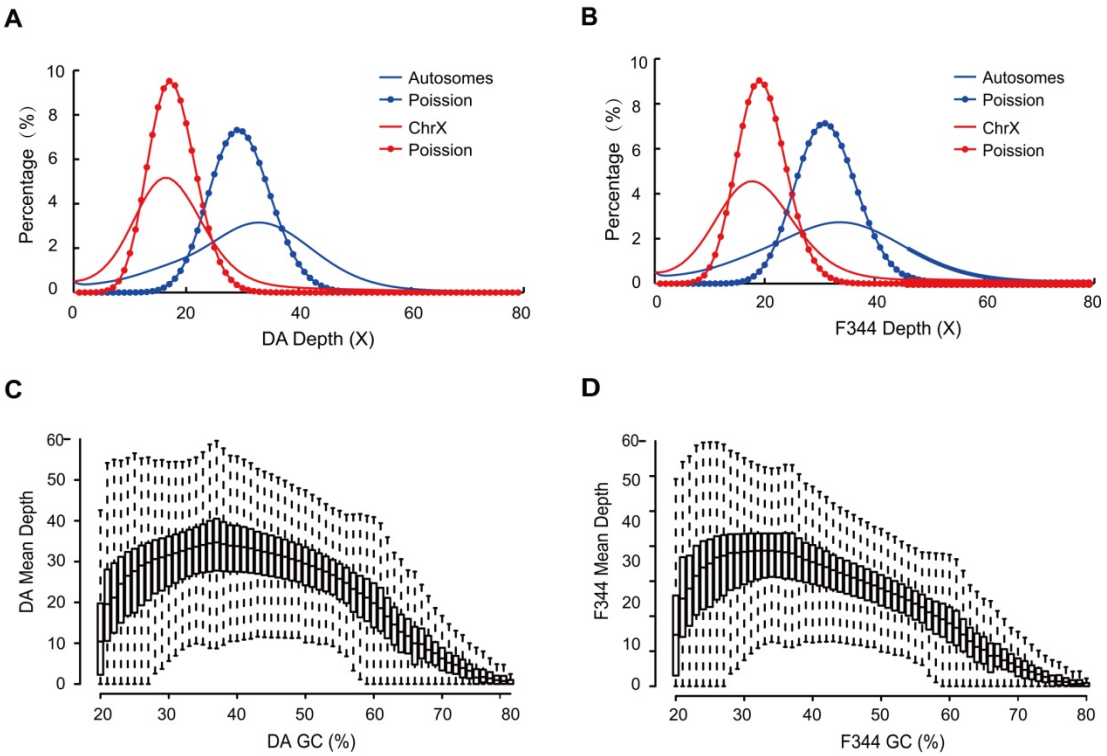

Figure S2 Distribution of DA and F344 sequencing depth in autosomes and X-chromosome. The observed sequencing depth of **(A)** DA and **(B)** F344 X-chromosome (red) and autosomes (blue) followed a Poisson-like distribution. The median depth for autosomes was 31-fold for DA and 32-fold for F344. The median depth for the X-chromosome was 16-fold for DA and 18-fold for F344. The lower sequencing depth for the X-chromosome is due to hemizygozity because both genomes are of male rats. **(C)** We calculated the GC content and average sequencing depth of 500-bp non-overlapping sliding windows along the assembled sequence. DA and **(D)** F344 sequencing depth was negatively correlated with extremes of GC content. The box-plot was created using the R package.

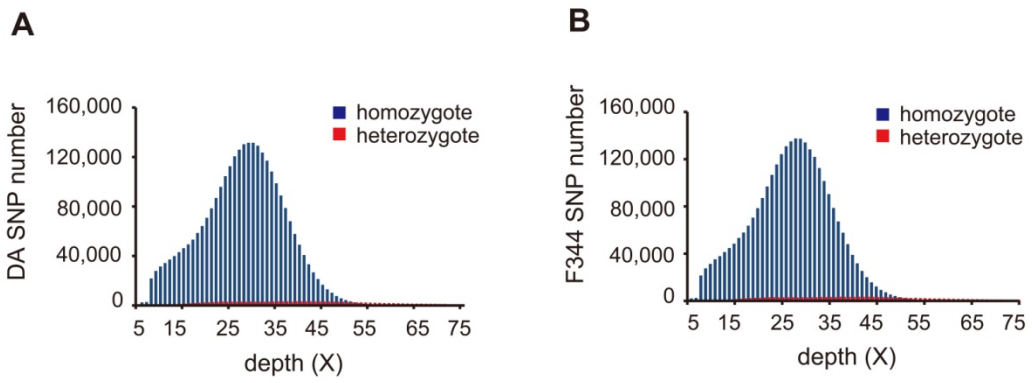

Figure S3 Distribution of SNPs according to sequencing depth. (A) The sequencing depth of homozygous SNPs (blue) was normally distributed with a median depth of 28 for DA and **(B)** 27 for F344. Heterozygous SNPs (red) were detected in regions with deeper coverage, with a median sequencing depth of 40 for both strains.

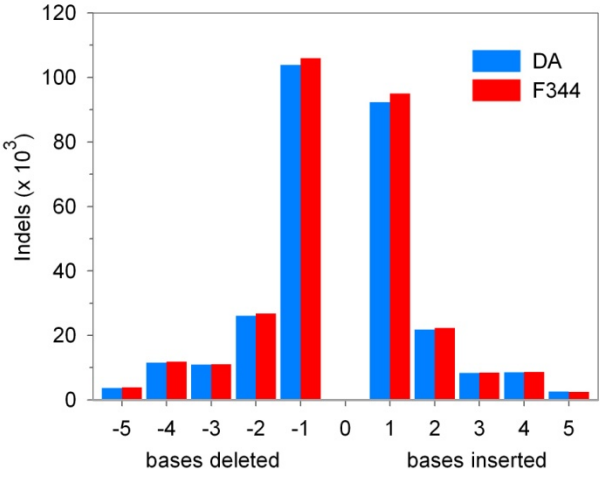

Figure S4 Distribution of homozygous indels relative to size. The majority of homozygous indels were 1-bp deletions and 1-bp insertions. Indel size distribution was similar in DA (blue) and F344 (red).

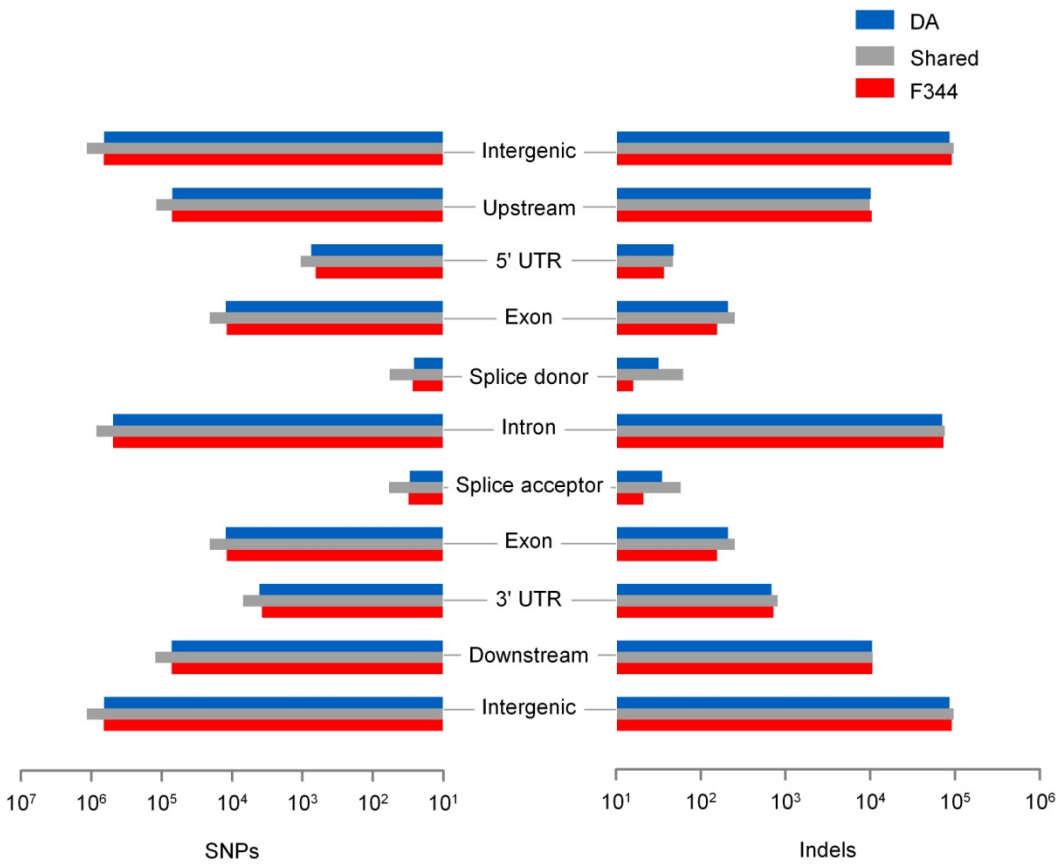

Figure S5 Distribution of homozygous SNPs and indels relative to gene structure. Most SNPs and indels were intergenic, intronic, and downstream or upstream of genes. There were less SNPs and indels in exons, splicing sites and the 5'- and 3'-untranslated regions (UTR). The distribution of strain-specific and shared SNPs and indels were similar.

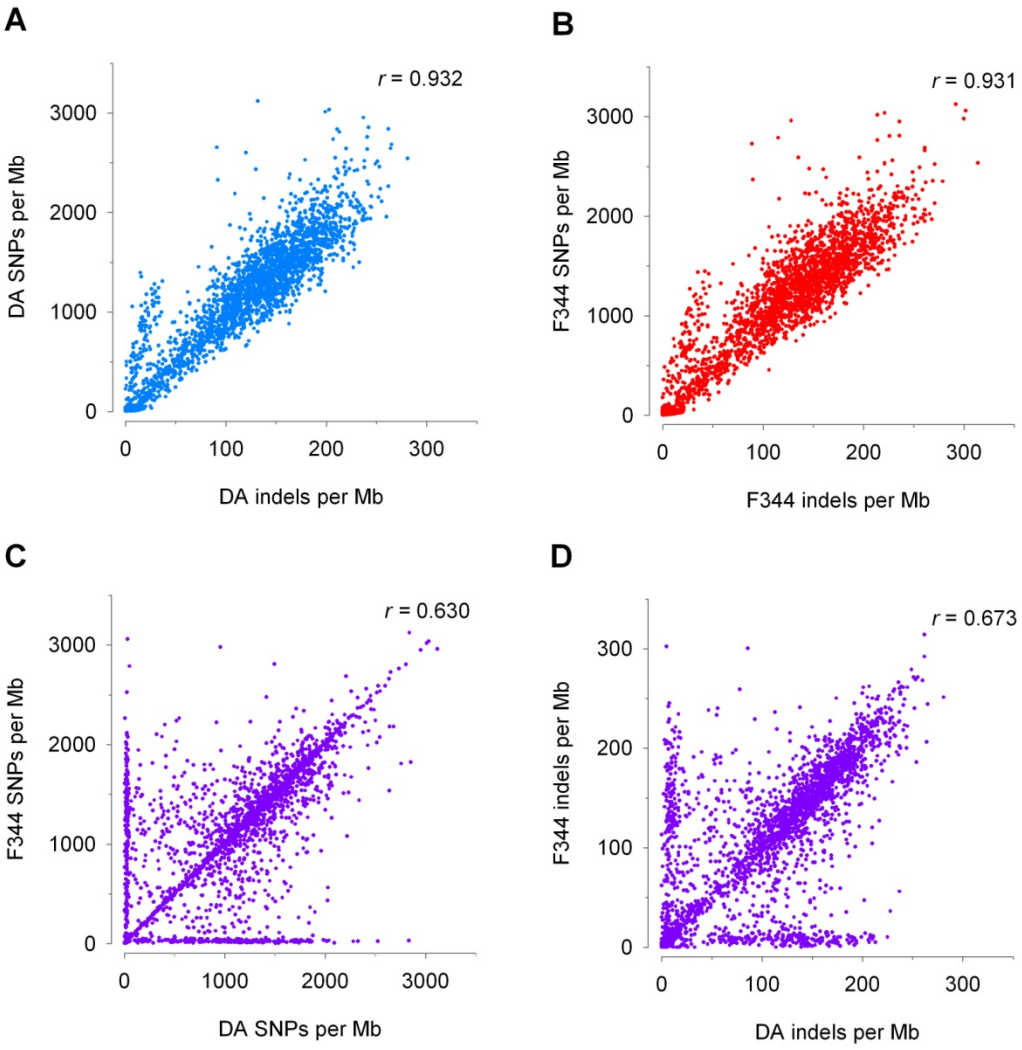

Figure S6 Correlation between the genomic distribution of homozygous SNPs and indels in DA and F344. There was a near perfect correlation between the regional density of homozygous SNPs and homozygous indels in 1-Mb non-overlapping windows for both **(A)** DA (r=0.932, P≈0) and **(B)** F344 events (r=0.931, P≈0). There was also a strong correlation between the regional density of **(C)** DA and F344 homozygous SNPs (Pearson correlation coefficient [r]=0.630, P=4.38x10^{-302}) and of **(D)** DA and F344 homozygous indels (r=0.673, P≈0). In (C) and (D), events plotted at low density on the X-axis indicate shared haplotypes between BN and F344, and on the Y-axis indicate shared haplotypes between BN and DA.

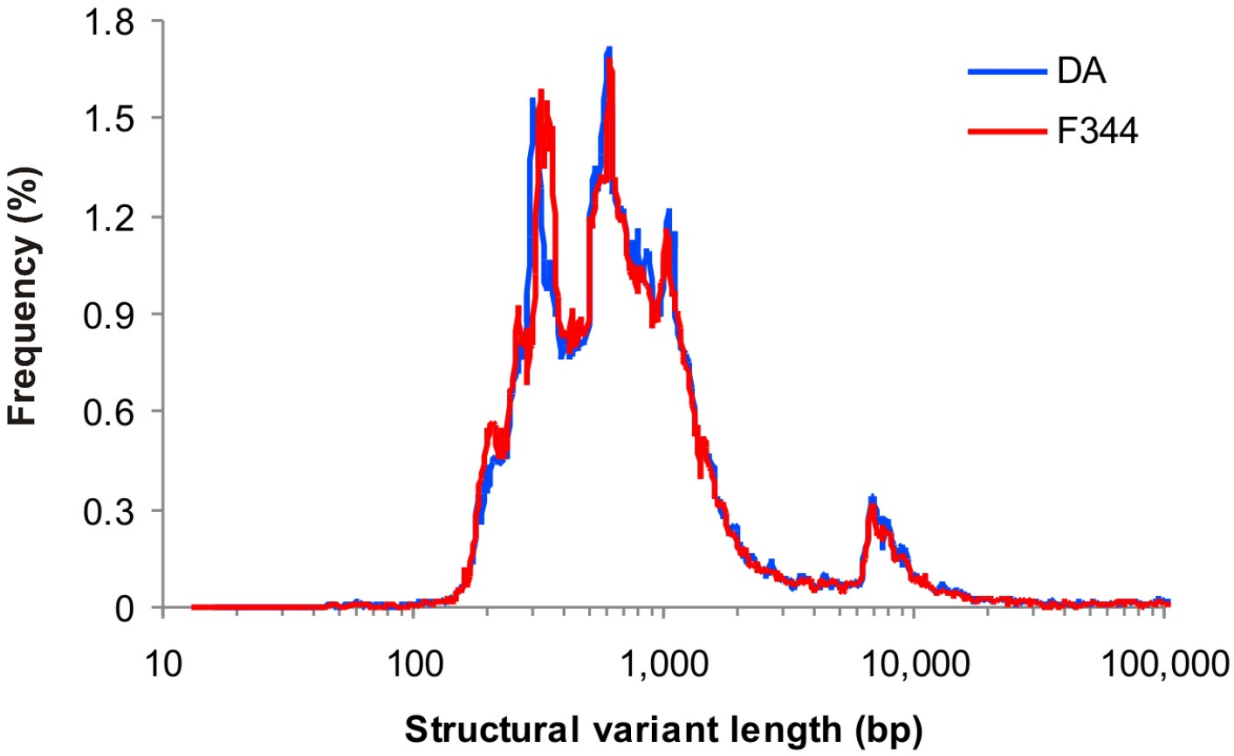

Figure S7 Distribution of structural variants according to size. Structural variants were similarly distributed in the DA and F344 genomes. Most structural variants were in the 200-2000bp range. There were four distinct frequency peaks at 302-324, 602, 1023-1047, and 6760-6918 bp.

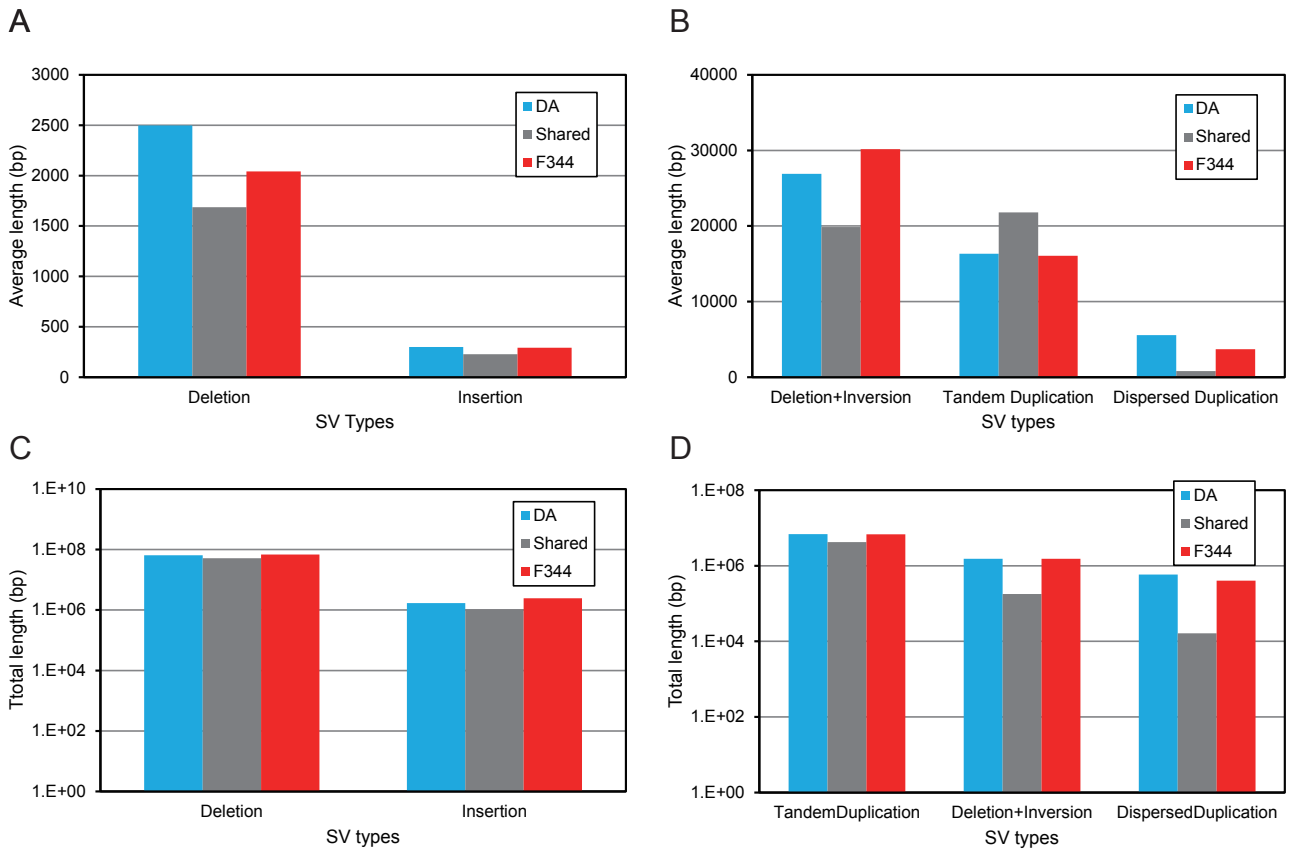

Figure S8 Types of structural variant in the DA and F344 genomes. (A) Insertions had lower average length than other structural variants, which is expected since detection of insertions in next-generation sequencing is constrained by the insert-size. **(B)** The structural variants with the longest average length were combined deletions and inversions, followed by tandem duplications and dispersed duplications. **(C)** Most of the sequence affected by structural variants was in deletions, insertions, and **(D)** in tandem duplications. Blue: variants specific to DA, red: variants specific to F344, grey variants found in both DA and F344.

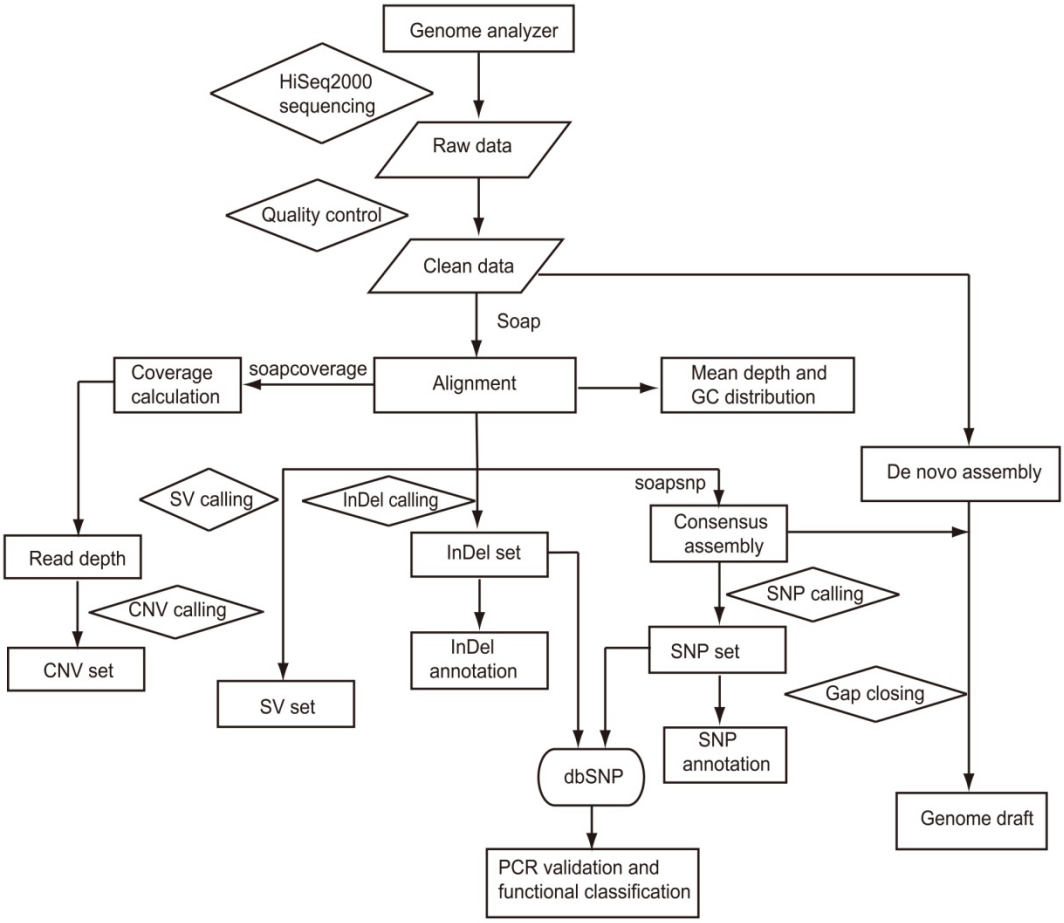

Figure S9 Construction and analyses of the DA and F344 semi-finished genomes. Genomic DNA from DA and F344 male rats was sequenced using Genome Analyzer. After quality filtering, over one billion high-quality reads for each strain were aligned to the BN rat reference genome (RGSC3.4) using SOAP2.21. 91% of the reads were mapped, covering 98.9% of the BN genome. SOAP software package was used to generate consensus sequences (CNS) and to call single-nucleotide polymorphisms (SNP), short insertion-deletions (Indel), and structural variations (SV), which were further quality filtered, functionally annotated, compared to variations in dbSNP, and validated using Sanger sequencing. Gap-containing regions of the reference genome were covered using the GapCloser tool. Assembly of consensus sequences and contig end extension generated an additional 59-Mb of sequence for either strain, corresponding to 22% of the unknown sequences in the BN genome. Distribution of copy number variation (CNV) candidates and coverage of repetitive elements were also analyzed.

¢l·ōťŜ {м [λōℕl·ℕλŜă l·yŘ ℕŜl·Řă dzăŜŘ ű2 l·ăăŜΥōťŜ űKŜ 5! l·yŘ Coпп ∃Ŝy2ΥŜă

Sample	Library	Insert Size	Read Length	Cleaned Data		High quality data *	
				Reads ($x10^9$)	Bases (Gb)	Bases (Gb)	Rate (%)
	RATvubRAADIAAPEI-2	475	90	0.745	66.96	64.88	96.89
DA	RATvubRAADIBAPEI-3	499	90	0.279	25.07	24.09	96.09
	Total			1.024	92.03	88.97	96.67
	RATvubRABDIAAPEI-1	487	90	0.667	59.90	58.59	97.81
F344	RATvubRABDIBAPEI-4	504	90	0.410	36.90	35.92	97.34
	Total			1.077	96.80	94.51	97.63

* *Q*-value ≥ 20

[illegible]

Sample	Alignment to BN genome				Coverage of BN genome			Gaps closed	
	Reads ($x10^6$)	Rate (%)	Bases (Gb)	Rate (%)	At least one read (%)	Three or more reads (%)	Average depth (fold)	Gaps	Base pairs
DA	936.01	91.38	83.87	91.13	98.90	98.05	32.68	359,392	59,317,377
F344	983.64	91.32	88.18	91.10	98.90	98.03	34.36	361,412	59,700,347

¢l·ōtŜ {o /2ӯ͡Ɏl·ű2ӯ 2Ŧ 5! l·ӯŘ Conn {btă ǿűK {l·ӯƎŜ ăŜljdzŜӯŎӯƎ

Sample	Homozygous SNPs			Heterozygous SNPs		
	Tested	Confirmed	Rate	Tested	Confirmed *	Rate
DA	587	584	99.49%	25	5	20.00%
F344	346	346	100.00%	20	1	5.00%
Total	933	930	99.68%	45	6	13.33%

* Confirmed as homozygous SNPs; no SNPs were confirmed in the heterozygous state

¢l·õfŜ {n 5 ! l·yŘ Conn {b t ă ø↕ůK↕y˝Ô2Ř↕yƎ ăŜljʤŜyŐŜă ó / 5{ú 2Ŧ wŜŧăŜlj l·yŘ 9yăŜY˝õf ƎŜyŜă◊

SNPs	Database	DA SNPs in CDS		F344 in CDS	
		Synonymous	Non-synonymous	Synonymous	Non-synonymous
All SNPs	Refseq	9,679	6,328	9,597	6,143
	Ensembl	15,174	9,066	14,914	8,998
Homozygous	Refseq	8,901	5,307	8,850	5,163
	Ensembl	14,195	7,723	13,987	7,670

¢l·ōfŝ {p wŝŦăŝlj l·yŘ 9yãŝY ōf ƎŝySă Ô2yűl·ʎyʎyƎ 5 ! l·yŘ Conn {b tă¢

SNPs	Database	Genes containing DA SNPs		Genes containing F344 SNPs	
		Total	Non-synonymous	Total	Non-synonymous
All SNPs	Refseq	6,405	3,506	6,353	3,420
	Ensembl	9,660	5,209	9,532	5,112
Homozygous	Refseq	6,083	3,174	6,007	3,074
	Ensembl	9,185	4,724	9,034	4,632

¢l·ōƒŜ {c {l·yƎŜǊ Ŏƨyƭ ǊY l·ŭʎƨyʹ ƨƭ 5! l·yŘ Conn ƘƨY ƨÎèƎƨdză ʎyŘŜƭăφ

Strain	Tested	Confirmed
DA	28	28
F344	49	49
Total	77	77

¢l·ōſŜ {т /ƨπſƨŎl·ſʎÎl·ǘʎƨỿ ƨŦ 5 ! l·yŘ Conn ǎǘNJdzŎǘdzNJl·ſ Øl·NJʎl·yǘǎ ǿʎǘK ŎƨŘʎyƎ ǎŜljdzŜyŎŜǎ ƨŦ wŜŦ{Ŝlj ƎŜyŜǎ ғ

Strain	RefSeq Gene Region			
	3’ UTR	5’ UTR	CDS	intron
DA	269	165	1398	11314
F344	292	177	1502	12622

* UTR: untranslated region; CDS: coding sequence plus start and stop codons

¢l·ōŧŜ {y bǳҮōŜǋ l·yŘ ăʎÎŜ 2ŧ ǋŜ∃ʎ2yă Ŏ2yǘl·ʎyʎy∃ Ŏ2Ǉè yǳҮōŜǋ Ǿl·ǋʎl·ǘʎ2y ʎy ǘKŜ 5 ! l·yŘ Cопп ∃Ŝy2ҮŜă

Length (Kb)	DA [1]	F344 [2]	Shared [3]
10-20	2,004	2,891	652
20-30	259	373	159
30-40	90	112	57
40-50	41	47	31
50-60	68	78	26
60-70	24	18	14
70-80	17	8	3
80-90	11	10	6
90-100	11	10	6
>100	69	64	40
Total	2,594	3,611	994

[1] Allele is unique to DA (F344 allele = BN allele)
[2] Allele is unique to F344 (DA allele = BN allele)
[3] DA and F344 have the same allele, different from BN

¢I·ōfŜ {ф ! fλƎyϒŜyű ƻŦ 9{¢ă ƻyűƻ űKŜ 5 ! I·yŘ Conn ƎŜyƻϒŜ ŘǊI·Ŧűăф[a]

	Total mapped (%)	95% Covered in Genome (%)		95% Covered in Chromosomes (%)		95% Covered in Novel scaffolds (%)	
		Total	Unique	Total	Unique	Total	Unique
Rn3.4_EST vs. DA	97.97	97.53	78.82	97.53	78.88	0.64	0.43
Rn3.4_EST vs. F344	97.97	97.47	78.92	97.45	79.14	0.76	0.56

[a] 194,363 ESTs of BN were downloaded from UCSC; columns show the percentage of ESTs that are covered ≥ 95% of the total length using BLAST

Table S10 Estimation of single base error rates in the DA and F344 genome drafts.

Samples	Effective genome size	Incorrect assembly sites	Single base error rate (10^{-5})
DA	2,616,053,766	79,970	3.05
F344	2,615,410,193	78,233	2.99

Table S11 Selected 23 genes implicated in specific diseases or phenotypes with non-synonimous SNPs or insertion-deletions in between DA and F344 genomes*.

Disease/Phenotype	Gene Symbol	Gene name	Reference PMID	Ensembl Gene_ID	RefSeq ID	Chr	Position	BN	DA	F344	Type	Exon_ID	Effect
Rheumatoid arthritis	**RT1-Db1**	RT1 class II, locus Db1	23381558	ENSRNOG00000033215	294270	20	4672641	T	C	T	SNP	exon_20_4672633_4672743	NON_SYNONYMOUS_CODING
							4673534	C	T	C	SNP	exon_20_4673348_4673629	NON_SYNONYMOUS_CODING
Rheumatoid arthritis	**Prdm1**	PR domain containing 1, with ZNF domain	19898481	ENSRNOG00000000323	309871	20	48442690	C	G	C	SNP	exon_20_48442007_48443112	NON_SYNONYMOUS_CODING
Multiple sclerosis	**Tagap**	T-cell activation RhoGTPase activating protein	22190364	ENSRNOG00000018915	308097	1	41385244	T	T	C	SNP	exon_1_41384339_41385585	NON_SYNONYMOUS_CODING
							41385526	C	C	T	SNP	exon_1_41384339_41385585	NON_SYNONYMOUS_CODING
Multiple sclerosis	**Evi5**	ecotropic viral integration site 5	19525955	ENSRNOG00000002039	100360066	14	2571837	G	A	G	SNP	exon_14_2571645_2571956	NON_SYNONYMOUS_CODING
Multiple sclerosis	**Cyp24a1**	cytochrome P450, family 24, subfamily a, polypeptide 1	22725956	ENSRNOG00000013062	25279	3	161550223	C	C	A	SNP	exon_3_161550179_161550275	NON_SYNONYMOUS_CODING
							161553779	C	C	G	SNP	exon_3_161553544_161554146	NON_SYNONYMOUS_CODING
Multiple sclerosis	**Vcam1**	vascular cell adhesion molecule 1	22725956	ENSRNOG00000014333	25361	2	212294171	C	T	C	SNP	exon_2_212294006_212294326	NON_SYNONYMOUS_CODING
Multiple sclerosis	**Cd6**	Cd6 molecule	22725956	ENSRNOG00000020884	25752	1	213292169	G	G	C	SNP	exon_1_213292139_213292240	NON_SYNONYMOUS_CODING
Psoriasis	**Rnf114**	ring finger protein 114	18364390	ENSRNOG00000009525	362277	3	158658588	*	-C	*	DEL	exon_3_158658548_158658590	FRAME_SHIFT: ENSRNOT00000012698
Chronic granulomatous disease, X-linked	**Cybb**	cytochrome b-245, beta polypeptide	2556453	ENSRNOG00000003622	1536	X	25520395	C	T	C	SNP	exon_X_25520380_25520400	STOP_GAINED
Familial cold autoinflammatory syndrome 2	**Nlrp12**	NLR family, pyrin domain containing 12	18230725	ENSRNOG00000014812	292541	1	64252963	*	-T	*	DEL	exon_1_64251445_64253125	FRAME_SHIFT: ENSRNOT00000043650
TLRs (1, 2, and 4) chaperone, macrophage activation	**Hsp90b1**	heat shock protein 90, beta, member 1	17275357	ENSRNOG00000026963	362862	7	23333834	*	-TTC	*	DEL	exon_7_23333819_23333944	CODON_DELETION
TGFb signaling regulator	**Mtmr4**	myotubularin related protein 4	20061380	ENSRNOG00000007496	287607	10	75893083	*	-G	*	DEL	exon_10_75893027_75893084	FRAME_SHIFT: ENSRNOT00000009943
Oxygen-induced retinopathy	**Cyr61**	cysteine-rich, angiogenic inducer, 61		ENSRNOG00000014350	83476	2	243825827	A	G	A	SNP	exon_2_243825795_243826139	NON_SYNONYMOUS_CODING
							243825889	G	G	T	SNP	exon_2_243825795_243826139	NON_SYNONYMOUS_CODING
							243826673	C	A	C	SNP	exon_2_243826481_243826694	NON_SYNONYMOUS_CODING
Oxygen-induced retinopathy	**Mmp2**	matrix metallopeptidase 2	23048035	ENSRNOG00000016695	81686	19	15267575	C	T	C	SNP	exon_19_15267566_15267665	NON_SYNONYMOUS_CODING
Morphine antinociception	**Oprm1**	opioid receptor, mu 1	21212276	ENSRNOG00000018191	25601	1	37567576	G	A	G	SNP	exon_1_37567434_37567786	NON_SYNONYMOUS_CODING
							37570778	C	C	A	SNP	exon_1_37570718_37570822	NON_SYNONYMOUS_CODING
							37777671	G	G	A	SNP	exon_1_37777620_37777720	NON_SYNONYMOUS_CODING
Bone mineral density	**Alox15**	arachidonate 15-lipoxygenase	14716014	ENSRNOG00000019183	81639	10	57186448	*	-G	*	DEL	exon_10_57186404_57186448	FRAME_SHIFT: ENSRNOT00000026038
Bone mineral density	**Dkk1**	dickkopf-related protein 1 precursor	22504420	ENSRNOG00000011692	293897	1	234396245	A	A	T	SNP	exon_1_234396048_234396426	NON_SYNONYMOUS_CODING
Cancer-prone Rothmund-Thomson Syndrome	**Recql4**	RecQ protein-like 4	18716613	ENSRNOG00000032446	300057	7	114753590	*	-TCC	*	DEL	exon_7_114753538_114753697	CODON_CHANGE_PLUS_CODON_DELETION
Cancer, ovarian	**Rsf1**	remodeling and spacing factor 1	16172393	ENSRNOG00000024194	308839	1	154912001	*	-ACA	*	DEL	exon_1_154910304_154912078	CODON_DELETION
Cancer, prostate	**Lmtk2**	lemur tyrosine kinase 2	18264097	ENSRNOG00000025155	304286	12	10769797	*	*	-T	DEL	exon_12_10769760_10769807	FRAME_SHIFT: ENSRNOT00000060813
Cancer, uterine endometrium	**Chd4**	chromodomain helicase DNA binding protein 4	23104009	ENSRNOG00000018309	117535	4	161250411	*	-C	*	DEL	exon_4_161250411_161250425	FRAME_SHIFT: ENSRNOT00000055970
							64252963	*	-T	*	DEL	exon_1_64251445_64253125	FRAME_SHIFT: ENSRNOT00000066836
Cancer, inhibitor of metastasis and p53 regulator	**Hexim1**	hexamethylene bis-acetamide inducible 1 ,	22948151 22964639	ENSRNOG00000003203	498008	10	92291030	*	+AGA	*	INS	exon_10_92290289_92292399	CODON_INSERTION
Cancer, neuroblastoma, Th2 cytokine responses	**Il31ra**	interleukin 31 receptor A	21436895, 17353366	ENSRNOG00000042080	688622	2	43901028	*	*	+AAA	INS	exon_2_43901008_43901063	CODON_INSERTION

* All variants detected in the homozygous state.