Novel roles of DNA variation in maintenance of AT rich genome and conserved function in rats

Amy Zinski
Washington State University
Weimin Wang
Washington State University
Jennifer Michal
Washington State University
Leah Solberg Woods
Wake Forest University School of Medicine
Ryan McLaughlin
Washington State University
Zhihua Jiang (jiangz@wsu.edu)
Washington State University https://orcid.org/0000-0003-1986-088X

Article

Keywords: Genome sequencing, DNA variants, SNPs, INDELs, genome stability, functional conservation.

Posted Date: March 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-343453/v1

Novel roles of DNA variation in maintenance of AT rich genome and conserved function in rats

Amy L. Zinski[1,*], Weimin Wang[1,*], Jennifer J. Michal[1], Leah C. Solberg Woods[2], Ryan J. McLaughlin[3] and Zhihua Jiang[1,**]

[1]Department of Animal Sciences, Washington State University, Pullman, WA, USA

[2]Department of Internal Medicine, Section on Molecular Medicine, Wake Forest University School of Medicine, Winston Salem, NC, USA

[3]Department of Integrative Physiology and Neuroscience, Washington State University, Pullman, WA, USA

***Contributed equally to the work.**

****Corresponding author:** Dr. Zhihua Jiang (ORCID ID: 0000-0003-1986-088X), Professor of Comparative Genomics. Phone: 509-335 8761; Email: jiangz@wsu.edu

Data accession number at NCBI: PRJNA642750 (SRX8786728 and SRX8786729).
https://www.ncbi.nlm.nih.gov/sra/PRJNA642750

Author contributions
Amy Zinski: Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization.
Weimin Wang: Software, Formal analysis, Investigation, Data Curation, Writing - Review & Editing.
Jennifer Michal: Validation, Investigation, Writing - Review & Editing, Project administration.
Leah Solberg Woods: Resources, Writing - Review & Editing.
Ryan McLaughlin: Resources, Writing - Review & Editing.
Zhihua Jiang: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Visualization, Funding acquisition

Abstract

Whether or not DNA variation changes genome-wide nucleotide compositions remains largely unknown. By examining 4,604,291 DNA variants between two rat strains, we observed that sequencing depth is strongly correlated with genome content as 21.41, 38.36, 44.26 and 6512.70 average reads per locus were collected for Y, X, autosomes and mitochondrial (MT) genomes; respectively ($P<0.0001$). The mutation rates corresponding to these four genome subsets were 0.055, 0.401, 1.733 and 4.475 variants per kb ($P<0.0001$), confirming the links between recombination frequencies and DNA variability. Although SNPs (single nucleotide polymorphisms) tend to reduce AT content, more CG deletions than CG insertions (INDELs) implies the GC content would not increase. Therefore, the SNP-INDEL interplay may play a key role in maintenance of the AT-rich genomes in rat during evolution. Formation of CpG sites appear to be hindered because genome-wide G INDELs (1.38%) with C as the 5'-nucleotide and CG INDELs (1.19%) are rare. However, the relatively high C—>G/G—>C rate in 5'UTRs (untranslated regions) and G/C INDELs in the 5'UTR and/or exonic regions highlight their importance for execution of gene function. Our study provides evidence that DNA variation does not jeopardize genome stability and functional conservation during evolution.

Key words: Genome sequencing, DNA variants, SNPs, INDELs, genome stability, functional conservation.

Highlights:

Our present study supports our hypothesis that the NIH-HS rats are genetically more variable than any other single rat strain because it was synthesized using eight inbred progenitor strains.

Our present study provides initial evidence that SNPs and INDELs may serve as antagonistic forces to keep genome-wide base composition unchanged during evolution.

Our present study indicates that the creation of *de novo* CpG sites in genomes is rare so that significant phenome shifts may not occur naturally during evolution.

1. Introduction

The National Institute of Health heterogeneous stock or NIH-HS rats were established in the early 1980s to create an outbred strain with a large number of segregating alleles that can be used in experiments to generate a normal distribution of responses to treatments or selectively bred with specific lines to generate divergent phenotypes [1]. Shortly after their creation, these rats were used in an alcohol drinking behavior study where both high and low consumers of alcohol were identified [2]. An additional study used selective breeding of the NIH-HS rats to produce lines with high and low alcohol sensitivities using a within-family approach under a thirteen-generation selection [3]. Other long-term selection experiments conducted using the NIH-HS rats included high or low cocaine preference [4], large or small serotonin-1A sensitivity [5], fast or slow running capacity [6] and high or low body temperature [7]. This unique outbred rat strain has also been broadly explored to map quantitative trait loci (QTLs) for complex phenotypes. To date, reported QTLs and/or phenotyping assessment include behavioral, bone, cardiovascular, depression, diabetic, exercise, glucose tolerance, hematology, immunology, metabolic, neuroinflammation, obese and renal traits, for example [8 – 17].

Genetically, the NIH-HS rats were synthesized using eight founder strains, including ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N, and WN/N [1]. In 2013, Baud and colleagues conducted whole genome sequencing of each of these progenitors using the SOLiD 4 and SOLiD 5500 sequencers with a depth of at least 22x base coverage per strain [17]. As the reference genome of the rat was derived from the BN/NHsdMcwi strain, only 71,038 SNPs (single nucleotide polymorphisms), zero INDELs (insertions and deletions) and 14,839 SVs (structure variants) were detected in BN/SsN strain. In the remaining seven strains, however, there were 2,664,124 to 3,088,953 SNPs, 151,099 to 305,705 INDELs and 18,306 to 48,553 SVs. The same datasets were re-analyzed two years later using a different data analysis pipeline, which detected 59,402 SNPs, 660,918 INDELs and 14,126 SVs in the BN/SsN strain, while 2,848,992 to 3,368,008 SNPs, 1,302,710 to 1,573,573 INDELs and 3,950 to 21,832 SVs were identified in the seven other strains [18]. These results indicate that the current bioinformatics pipelines are designed to call SNPs with high accuracy rather than calling the latter two categories of variants. In addition, we found that INDELs and SVs were often not well defined in these studies [19].

Unfortunately, genome wide DNA variants have never been thoroughly uncovered in the NIH-HS rats. A recent study reported an adapted genotyping by sequencing (GBS) approach that was used to discover and genotype genome variants in this outbred strain [9,20]. No doubt, GBS is powerful, but this genome sampling sequencing approach may not reveal a comprehensive picture of DNA variants [21]. Here we present the first report on whole genome sequencing of NIH-HS rats in comparison to Wistar (WI) rats. Our study revealed at least 4,604,291 high quality DNA variants residing in both nuclear and mitochondrial genomes of rats. We observed some striking features in genome variation between two strains. For example, only 21% DNA variants are commonly polymorphic and only 7% loci have fixed alleles between strains, while the remaining loci are within-strain polymorphic markers. Interestingly, 97% of DNA variants reside in the intergenic and intronic regions. In particular, the NIH-HS rats have 3.8-fold more within-strain DNA variants than WI rats. Understanding of genome-wide variant categories, dynamics, patterns and functional constraints will provide a solid foundation of essential knowledge that will enhance the NIH-HS strain as a primary model organism choice for biomedical research.

2. Materials and Methods

2.1. Animals, DNA extraction and genome sequencing. DNA was extracted from tail snips collected from 128 NIH-HS rats (the NMcwi:HS rats colony, half males and half females) and 48 WI rats (half males and half females). The use of tissues was approved by the Institutional Animal Care and Use Committee of Washington State University. In brief, approximately 0.6 cm of tissue was excised from each rat tail, minced, and lysed with buffer and proteinase K as described in the DNeasy Blood & Tissue kit (Qiagen, Germantown, MD). Pure genomic DNA was isolated from the lysate with the spin column included in the kit. Gel electrophoresis was

performed to check fragment size of the extracted DNA samples. Degraded samples were removed from further analysis. Each sample was adjusted to a concentration of 100 ng/μl, which was then used to form two DNA pools. The NIH-HS pool contained 122 rats, while the WI pool involved a total of 40 rats. Both DNA pools were submitted to Novogene (Sacramento, CA) and sequenced with collection of short reads using the Illumina sequencing platform.

2.2. Quality control, read mapping and variant calling. We used the Fastp software (0.19.4) tool to preprocess raw sequence reads (https://github.com/OpenGene/fastp) by removing the index and barcoded sequences and by discarding the unpaired reads. All clean short reads were then mapped to the newest *Rattus norvegicus* reference genome (Rnor_6.0) using Bowtie2 [22]. The "bowtie2-built" command line was used to create a Bowtie index from the rat reference genome and alignments were prepared using an "end-to-end" strategy. The "-no-mixed" parameter was used to discard the unpaired and multiple alignments for pair-end reads. Alignment files (SAM files) were merged and converted to BAM files prior to sorting and indexing using SAMTOOLS [23]. Potential PCR duplicates were filtered using the "MarkDuplicates" command line of Picard software (http://broadinstitute.github.io/picard/). Gap realignment, base recalibration and variant calling were performed using the Genome Analysis Toolkit (GATK) [24]. Because pooled DNA samples were used to detect genome-wide variants, we developed our own processing scripts to sort alleles and count their associated numbers of reads per allele within each polymorphic site [25].

2.3. Data analysis and validation. Our data outputs were in Excel format, with one file per chromosome or mitochondrial genome. Chromosome position (coordinate), reference allele, mutant allele, genotype, number of reads per allele and total number of reads for both alleles in each of the two strains were recorded for each DNA variant. Data quality was improved by removing the DNA variants that had 1) less than 30 total reads for both strains; 2) no genotype clearly scored in one of the strains and 3) mono-polymorphic markers observed in both NIH-HS and WI rats that were different from the reference genome. All DNA variants with 15 reads or more per site were retained for the Y chromosome. The DNA variants that passed quality control were then classified into SNPs or INDELs. Bioinformatics processing assigned these variants into 11 genomic regions and multiple Excel functions were used to sort the data for analysis. Variants were graphed using karyoploteR, an R package, to determine and plot the density SNPs and INDELs among the two rat strains using a 1-megabase (Mb) non-overlapping window (default window size) [26]. Genome-wide Fst (F statistics) values were also calculated to detect the divergent genome regions (DGRs, $P<0.05$) and divergent variant loci (DVLs, $P<0.05$) between the two rat strains [25,27,28]. Ten identified SNPs were subsequently validated in individual samples through Sanger sequencing.

3. Results

3.1. Genome copies and DNA variation rates. Mammalian somatic cells contain both nuclear and mitochondrial (MT) genomes. The former can be further split into autosomes and sex chromosomes (X and Y). As usual, all autosomes are diploid in both males and females. However, the X chromosome is diploid in females, but both sex chromosomes are haploid in males. Cells usually have many copies of MT genomes, differing by source of tissues/organs. Therefore, DNA contents vary in cells, depending on these genome parts. In the present study, genome sequencing of two pooled DNA samples, one from NIH-HS rats and another from WI outbred rats, identified 6,936,048 total DNA variants before quality control. After a three-step quality control process, 4,604,291 total genetic variants were retained for further analysis. After examining read coverage, we noticed that genome sequencing could truly measure the copies of autosomes, sex chromosomes and MT genomes.

As shown in Table S1, the genome coverage for all autosomes after filtering was 44.26 reads per locus, ranging from 43.35 (chromosome 12) to 45.27 (chromosome 16). Overall, autosomes had highly consistent read coverages. In contrast, there was an average 21.41 reads per locus in the Y chromosome, which is about half of that of the autosomes, and 38.36 average reads per locus in the X chromosome, which is higher than that of Y

chromosome, but slightly lower than that of autosomes. Interestingly, the sequencing coverage of the MT genome was 6,512.70 reads per locus (Table S1), suggesting that rat tail cells may contain approximately 300 copies of MT genomes. These results clearly indicate that whole genome sequencing of pooled DNA samples results in high quality data because the sequencing depth per locus was well correlated with the genome contents.

Though highly consistent, some variation in the read coverage per locus across autosomes may be due to fluctuation in chromosome-wide GC-contents. We noticed that the standard deviation (SD) of reads per locus along each autosome was weakly correlated to its GC content (Figure 1A). This correlation was significant at a 90% confidence ($p < 0.05$) level with an R^2 value of only 0.194, showing that there is a significant weak positive correlation between GC content and the variance in number of reads per locus across the autosomes (Figure 1A). This reflects the slight dependence of Illumina sequencing to read and record DNA sequences on the GC content of the DNA. The GC-rich regions of a genome are harder to amplify, resulting in less copies of GC-rich regions for sequencing than AT-rich sections.

We found that sex chromosomes are less variable than both autosomes and MT genomes. As shown in Table S1, the DNA variants are not equally distributed throughout the genome. The average autosomal variant density ranges from 1.34 (Chromosome 17) to 2.22 genetic variants per kilobase (kb) (Chromosome 13). Conversely, the sex chromosomes are much less variable, with 0.401 variants per kb in the X chromosome and only 0.055 DNA variants per kb in the Y chromosome (Table S1). Moreover, the frequency of mutations among autosomes is significantly greater than the mutation frequency of the sex chromosomes ($p<0.001$). On the other hand, most of the MT genome contains coding regions. As illustrated in Figure 1B, intronic and intergenic regions of the rat genome contain 36.22% and 60.64% of the total genetic variants, respectively. Thus, the absence of introns and intergenic regions would result in a loss of genetic variation potential in the MT genome. Conversely, the total variants per kb of the MT genome (4.47) is significantly higher than the nuclear genome ($p<0.001$).

In addition, the MT genome contains fewer INDELs than the nuclear genome. Of the identified genetic variants, 3,631,269 are SNPs and 973,022 are INDELs (Table S1). Overall, SNPs are 3.73 times more common than INDELs and account for 78.87% of the total variants, while the latter account for 21.13%. As shown in Table 1, the NIH-HS genome harbors 2,860,884 SNPs and 745,262 INDELs, while the WI genome contains 1,285,136 SNPs and 384,384 INDELs. The SNP to INDEL ratios (SNP:INDEL) are variable, ranging from 3.14 in chromosome Y to 4.1 in chromosome 13 (Table S1). Conversely, the MT genome is almost devoid of INDELs with a SNP:INDEL of 17.25 (Table S1).

3.2. Nucleotide types and DNA variation balances/imbalances. Among a total of 3,631,269 SNPs, 99.9% (3,628,900) are bi-allelic DNA variants (Table S1). Chemically, the four nucleotides are either purines (A and G) or pyrimidines (C and T). Bi-allelic SNPs within each group are called transitions, while those between groups are transversions. When mutation directions are considered, there are 12 types of bi-allelic SNPs (Figure 2 and Table S2). Due to DNA double-strand and base-pairing rules, frequencies are well balanced in each of these six pairs, but dramatically imbalanced across pairs: G—>A and C—>T (each with 19%), A—>G and T—>C (each with 15%), C—>A and G—>T (each with 5%), T—>G and A—>C (each with 4%), T—>A and A—>T (each with 4%) and G—>C and C—>G (each with 3%). Overall, transitions and transversions accounted for 68% and 32% of the mutations, respectively. On the other hand, the genome-wide mutation tendency is also imbalanced with mutations ending with nucleotides A/T (A or T, 56%) being higher than those ending with C/G (44%), or 1.27:1 ratio. A similar ratio (1.26:1) was also observed between G/C to A/T (48%) and A/T to G/C (38%) (Figure 2).

Of the 973,022 INDELs, 442,388 and 155,801 were those with 1 bp and 2 bp, respectively (Table S2). The INDEL events largely depend on the 5'-nucleotide and the nucleotide itself (Figure 3). For one A or one G

INDELs, the most common 5'-nucleotide is T, while C is the least frequent 5' nucleotide (Figure 3A). In contrast, the most and the least common 5'-nucleotide pair is A and G for one T or one C INDELs. Consequently, the INDEL frequencies are extremely variable from one G INDELs (1.38% or 6,103/442,388) with C as the 5'-nucleotide to one T INDELs (16.09% or 71,199/442,388) with A as the 5'-nucleotide (Figure 3A). Overall, 73% of INDELs were one A and one T insertions and deletions, while 27% were one C and one G INDELs. Furthermore, frequencies of one G and one C deletions were higher than their insertion frequencies by 1 or 2%, while the one T insertion frequency was 1% higher than its deletion frequency (Figure 3A).

For 2-bp INDELs, the combination frequencies ranged from CG (0.17% or 268/155081) with T as the 5'-nucleotide to TT (5.68% = (8,842/155,801)*100%) with C as the 5'-nucleotide (Table S2). Regardless of the 5'-nucleotide type, the CG combination appeared the least (1.19%; 1,848/155,801), while the most common combination was TT (13.90%; (21,659/155,081) (Figure 3B). With two-nucleotide combinations, A/T INDELs were the most frequent, occurring 66% of the time compared 34% for C/G INDELs (Figure 3B). In addition, deletion events were more widespread than insertion events for all four nucleotides.

3.3. Genomic regions and DNA variation preferences. Because only a few DNA variants reside in the splicing and overlapping regions of the genome, our current analysis focused on seven major genomics regions including the intergenic, upstream, 5'UTR (untranslated region), exonic, 3'UTR, downstream and intronic regions plus whole genome parameters. Based on the bi-allelic SNPs (Table S2), we observed that variant types differed by genomic region (Figure 4A). Mutations in the exonic regions were 45.15% C—>T/G—>A, which is 8.88% higher than in the intergenic regions (36.27%). Our analysis showed that the C—>T/G—>A frequencies across different genomic regions were negatively correlated with those of C—>A/G—>T ($r = 0.94$, $P<0.05$), A—>T/T—>A ($r = 0.98$, $P<0.05$) and A—>C/T—>G ($r = 0.98$, $P<0.05$). After arranging SNP types in the following order: C—>T/G—>A, T—>C/A—>G, C—>A/G—>T, A—>T/T—>A, A—>C/T—>G and C—>G/G—>C (Figure 4A), we classified genomic regions into three clusters: 1) intergenic, upstream, intronic and whole genome; 2) 5'UTR and exonic regions and 3) 3'UTR and downstream regions. In cluster 1, we observed a continuous decline in incidence of SNP types along the order. The frequencies of SNP types in cluster 2 declined in the first four types and increased in the last two types. Numbers of bi-allelic SNP types in the defined order in cluster 3 were more variable, decreasing in the first four types, followed by an upturn and then a downturn for the last two types of mutations (Figure 4A).

Generally speaking, the coding region takes up only ~3% of the genome with intergenic and intronic regions making up the rest. As shown in Figure 1B, therefore, DNA variants are abundant in intergenic and intronic regions (~97%), but are rare in other genomic regions (~3%). More specifically, 61.50%, 35.73%, 0.69%, 0.74%, 0.3%5, and 0.06% of the 1-bp INDELs occurred in intergenic, intronic, upstream, downstream, 5'UTR, exonic and 3'UTR regions, respectively (Table S2). However, the 1-bp insertion/deletion patterns showed that the 5'UTR and exonic regions are relatively different from other genomic regions (Figure 4B). That is, the C/G INDELs in both 5'UTR and exonic regions were more frequent (37% - 42%) than those of other regions, ranging from 26% to 30% (Figure 4B). For 2-bp INDELs, the exonic region pattern was also quite different from other regions because C/G INDEL events were more frequent than A/T INDEL events (52% vs. 48%, respectively) (Figure 4C). In other regions, C/G INDEL frequencies ranged from 32% in the 3'UTRs to 38% in the 5'UTRs.

3.4. Rat strains and DNA variation abundancy. We classified these 4,604,291 total genetic variants into four groups: 1) 975,636 (21%) common markers that are polymorphic in both WI and NIH-HS lines; 2) 693,884 (15%) WI exclusive markers that are monomorphic in NIH-HS; 3) 2,630,510 (57%) NIH-HS exclusive markers that are not polymorphic in WI; and 4) 304,261(7%) between-strain polymorphisms because the two stains each are fixed with a monomorphic allele (Table 1). As such, there are 3.79-fold more variants in the NIH-HS genome than in the WI genome based on the strain-exclusive markers. When total variants (common and exclusive combined) are compared, on the other hand, NIH-HS rats have 2.16-fold more variants than the WI

rats (3,606,146 vs.1,669,520 total variants). Nevertheless, these quantities between the two strains are significantly different ($p<2.2e-16$). In addition, the SNP:INDEL ratio also varies, ranging from 3.29 with the common markers to 4.08 with NIH-HS exclusive variants ($P<0.00001$). The same ratio in total DNA polymorphic markers also differs significantly between two strains (3.34 in WI and 3.84 in NIH-HS) ($P<0.000001$) (Table 1).

These four types of strain-related DNA variations are distributed differently among autosomes, sex chromosomes and MT genomes (Figure 5A). For the 20 autosomes, the percentages of NIH-HS exclusive variants increased from 44.9% on Chr12 to 69.7% on Chr20, while those of common polymorphic markers decreased from 30.5% on Chr12 to 13.1% on Chr20. A negative relationship was also observed between percentages of WI exclusive (decreased from 23.4% on Chr12 to 7.0% on Chr17) and between-strain polymorphic markers (increased from 1.2% on Chr12 to 13.3% on Chr17). DNA variants on the X chromosome included 66.1% NIH-HS exclusive, 13.1% both-strain common markers, 8.2% WI exclusive and 12.6% between-strain polymorphic markers, indicating that distribution falls into the ranges along with the autosomes. In contrast, the Y chromosome had the lowest percentage of shared DNA variants (4.4%), but the highest proportion of fixed markers (17.0%) between NIH-HS and WI rat strains in comparison to autosomes and the X chromosome. The distribution among these four types of strain-related DNA variations on MT was dramatically different with 97.3% NIH-HS exclusive, 0.0% common, 1.4% WI exclusive polymorphisms and 1.4% between-strain makers.

Because the WI genome contains fewer DNA variants than the NIH-HS genome, it is not surprising that the homozygosity distance (HD) is much longer in the former strain than in the latter strain (Table S1). Briefly, the HD average ranges from 1,013 bp (Chr10) to 3,310 bp (Chr17) in WI rats, but only from 553 bp (Chr13) to 938 bp (Chr17) in NIH-HS rats among 20 autosomes. The HD values were 11,738 bp (WI) vs. 3,150 bp (NIH-HS) for ChrX and 98,332 bp (WI) vs. 26,034 bp (NIH-HS) for ChrY. The overall distributions of genome variation are illustrated in Figure 5B. In addition, Fst analysis revealed a total of 1,890 divergent genome regions (DGRs, $P<0.05$) for autosomes and X chromosome between two strains (Table S3 and Figure 5B).

3.5. Genes and DNA variation gains/losses. There were 14,584 genes with polymorphic markers that had at least one DNA variant segregating in both strains, while 673 and 4,073 genes contained DNA variants that were exclusive in WI and NIH-HS rats, respectively (Table S1). In addition, an Fst test also collected a total of 275,565 divergent variant loci (DVLs) between the two strains ($P<0.05$), including 117,060 DVLs in 7,788 genes (single genes or overlapped genes) (Table S4). When these datasets were combined, 96 WI-exclusive and 745 NIH-HS exclusive genes with DVLs were identified and used for pathway enrichment analysis (Figure 6). The 96 DVL WI-exclusive genes play important roles in drug metabolism and gland/skeletal muscle development while the 745 DVL genes found exclusively in NIH-HS rats are mainly involved in protein/amino acids/lipid metabolism, DNA/RNA processing events, cell division and regulation of epigenetic processes.

Here we use the solute carrier family 6 member 3 (*Slc6a3*) gene, also known as dopamine transporter (*Dat1*), as an example to illustrate the dramatic difference in polymorphic status between WI and NIH-HS rats. Based on information from NCBI, we found that the annotation of the rat gene is incomplete. In human and mouse, the full-length cDNA sequence of the gene is 3,942 bp (NM_001044.5) and 3,456 bp (NM_010020.3), respectively, but only 1,985 bp (NM_012694.2) in rat. After exploring the GenBank databases, we realized that there are two key entries (M80570 and FQ140726) related to the rat gene. Merging both sequence entries produced a 3,492 bp cDNA sequence in rat. In addition, we also found that the gene is completely sequenced in the species – the genomic DNA sequence is 45,393 bp and spans a total of 15 exons.

The whole genome sequencing in the present study detected a tentative list of 95 DNA variants, including 77 SNPs and 18 INDELs in the rat *Dat1* gene (Table S6). Among them, 94 are polymorphic in WI rats, but monomorphic in NIH-HS rats. Only one exonic SNP is a bi-allelic locus in NIH-HS rats but fixed with a mono-

allele in WI rats. Table S6 is an example of a data output file, which includes information on chromosome (scaffold), mutation position, allele nucleotides (reference genome and mutant sequence), SNP or INDEL, genomic region and genotype, total reads and allele frequency in each rat strain. Among 77 SNPs, seven are potentially located in CpG sites (highlighted in green, Table S6). In addition, there are DNA variant hotspots in the rat genome, too. An example is shown in Figure 7A to demonstrate a hotspot of 10 SNPs located in the rat *Dat1* gene, spanning a 412-bp region from 32,356,153 bp to 32,356,564 bp (see Table S6). Using Sanger sequencing, we confirmed these ten SNPs exist in the WI rat genome (Figure 7B). These results clearly indicate that genome sequencing of pooled DNA discovers genome-wide variations with high accuracy.

4. Discussion

4.1. The NIH-HS genome is full of DNA markers that are segregated in the outbred rat strain. Although the NIH-HS was created more than 35 years ago [1], our present study is the first report of whole genome sequencing of the outbred rats themselves using a DNA pooling strategy. Among the eight progenitors used to synthesize the outbred line, three strains (MR/N, WN/N, and WKY/N) were derived from WI stock [8]. Interestingly, only 21% (975,636/4,604,291) of the polymorphic markers are common to both WI and NIH-HS rats. As such, we postulate that the 2,630,510 (57%) DNA markers exclusive to the NIH-HS genome may originate from the five non-WI progenitor strains – ACI/N, BN/SsN, BUF/N, F344/N and M520/N. Due to the evolutionarily close relationship, only 7% (304,261/4,604,291) of genome variants are fixed with different monomorphic alleles between WI and NIH-HS rats. Nevertheless, the large number of DNA variants decreases the average homozygosity distances in NIH-HS rats (553 bp to 938 bp for autosomes, 3,150 bp for X and 26,034 bp for Y) compared to WI rats (1,013 bp to 3,310 bp for autosomes, 11,738 bp for X and 98,332 bp for Y) (Table S1). In addition, we observed 1,890 DGRs ($P<0.05$) and 275,565 DVLs ($P<0.05$) between the two strains (Tables S3 and S4). Therefore, the genetic polymorphisms harbored by these eight progenitors have been well integrated in the NIH-HS line and have been well maintained over many generations of outbreeding. No doubt, the NIH-HS rat is a key resource for fine-mapping of QTLs underlying complex phenotypes and can be used to efficiently model human health and diseases [16].

4.2. Are SNPs and INDELs antagonistic forces to maintain an AT-rich genome? In the present study, we simply classified DNA variants into two categories: SNPs and INDELs. The former type remains over three times more common than the latter type in rat, a prominent trend identified in plants, fish, parasites, bacteria, and mammals including humans [29 – 39] with some exceptions [40,41]. We grouped SNPs into 12 types. As G is paired with C at a position, there are three complementary pairs of point mutations in a one-way direction: G—>A and C—>T, G—>T and C—>A or G—>C and C—>G (Figure 2). The A and T pair at a position also produces three complementary pairs of point mutations in a one-way direction: A—>G and T—>C, A—>C and T—>G or A—>T and T—>A (Figure 2). Based on the frequencies we observed in the present study, the single nucleotide mutation processes resulted in 56% AT and 44% CG. Currently we know that the average GC-content of the *Rattus norvegicus* genome is 41.61% (Table S1). As such, SNPs alone would contribute to accumulation of more CG content in the genome. Fortunately, we observed that C/G INDEL events are imbalanced, with deletion being more common than insertion (Figure 3). This means that INDELs could accumulate less CG content in the genome. Therefore, SNPs and INDELs may be antagonistic to each other so that an AT-rich genome in rats can be well maintained during evolution. Evidence shows that the AT-rich genes are often expressed in a tissue-specific manner [42].

4.3. Sex chromosomes, autosomes and MT differ significantly in mutational processes. Our results revealed that sex chromosomes had the lowest mutation rates (0.055 DNA variants per kb in the Y chromosome and 0.401 variants per kb in the X chromosome), while the MT genomes were extremely rich in genetic variation (4.47 DNA variants per Kb). The mutation rate in autosomes varied from 1.34 to 2.22 genetic variants per Kb. Conversely, the runs of homozygosity (ROH) are the shortest in the MT genome with 223 bp (for NIH-HS line), but 3,150 bp to 11,738 bp in chromosome X and 26,034 bp to 98,332 bp in chromosome Y, respectively (Table

S1). As such, there is a negative correlation between mutation rate and ROH. In mammals, the Y chromosome is clonal and the X chromosome has one-third of its chance in the male germline. Therefore, recombination events are relatively rare in the sex chromosomes with the exception of the small pseudo-autosomal region. Therefore, low recombination results in low mutation rate and long stretches of ROH [43]. On the other hand, recombination should occur frequently in MT because a mammalian cell can have up to 8,000 copies of MT genomes [44]. Furthermore, breeding or mating systems can dramatically influence population dynamics of DNA mutants, particularly for the Y chromosome and MT genomes. For example, we found that 17% of DNA variants in the Y chromosome were fixed each with a monomorphic allele between WI and NIH-HS rats and over 97% of SNPs in MT were NIH-HS exclusive (Figure 5A).

4.4. Does a genome avoid accumulation of additional *de novo* methylation sites during evolution? Previously, we found that the 5'- and 3'-flanking nucleotide combinations significantly affected the frequencies of transition SNPs [45]. For example, when SNPs are flanked by 5'T and 3'A, the transition frequency was 54%. However, when SNPs are located between 5'C and 3'G, the rate increases to 83% ($P<0.0001$). Obviously, CpG structures can be formed in the latter cases but cannot be created in the former cases. When a cytosine is methylated at the CpG site, it undergoes deamination and produces thymine [46]. This mechanism may minimize additional CpG accumulation due to mutation in genomes. In the present study, we noticed that one G INDEL events with C flanking at the 5'-side and CG INDELs following 5'T are relatively rare, occurring 1.38% and 0.17% of the time, respectively (Table S2). In particular, deletions are often more common than insertions for these INDEL events. Furthermore, the higher SNP:INDEL ratio in NIH-HS rats than in WI rats may indicate that genome may not favor integration of INDELs during the crossbreeding process (Table 1).

4.5. Why are eight progenitor sequencing datasets not used for comparison in the present study? In addition to work done by the Rat Genome Sequencing and Mapping Consortium [18] and others [19], these eight progenitor lines were also recently sequenced by Ramdas and coworkers [47]. We decided not to integrate those datasets in the present study, because these teams used only a few rats per strain in whole genome sequencing. The limited information from these types of studies makes it difficult to determine which DNA variants are segregating within each progenitor strain, resulting in a higher false negative rate (17.2% to 65%) than the false positive rate (2.7% to 16.7%), depending on the types of DNA variants [18]. Furthermore, whole genome sequencing of a few animals per strain detects only the homozygous loci, in general, which may account for more than 95% of total variants [48,49]. In the present study, we used a total of 122 NIH-HS rats and 40 WI rats so that we were able to detect the segregating makers in both strains. Use of a relatively large number of animals per progenitor strain warrants future investigation so we may understand how the NIH-HS genome structures were derived from each of progenitor strain.

In conclusion, our present study revealed that 1) the ratio between SNPs and INDELs is almost 80:20 in rats; 2) the sex chromosomes are less variable in comparison to autosomes and MT genomes; 3) 97% of DNA variants reside in the intergenic and intronic regions and 4) there are many DNA variant hotspots in a genome. In addition, characterization of both SNPs and INDELs in rats provides novel insights into mechanisms on how genome composition and function are maintained during evolution. Understanding genome variation will help us maximize use of the information to benefit the current and future generations.

ACKNOWLEDGEMENTS
This work was supported by the National Institute of Food and Agriculture, United States Department of Agriculture under Award Numbers 2016-67015-24470/2018-67015-27500 (sub-contract)/ 2020-67015-31733 and by funds provided for medical and biological research by the State of Washington Initiative Measure No. 171 and the Washington State University Agricultural Experiment Station (Hatch funds 1014918) received from the National Institutes for Food and Agriculture, United States Department of Agriculture to ZJ.

CONFLICT OF INTEREST STATEMENT

None declared.

Literature Cited

[1] C. Hansen, K. Spuhler. Development of the National Institutes of Health genetically heterogeneous rat stock. Alcohol Clin. Exp. Res. 8 (1984) 477-479, https://doi.org/10.1111/j.1530-0277.1984.tb05706.x

[2] J.M. Murphy, W.J. McBride, L. Lumeng, T.K. Li. Alcohol preference and regional brain monoamine contents of N/Nih heterogeneous stock rats. Alcohol Drug Res. 7 (1987) 33-39.

[3] L.J. Draski, K.P. Spuhler, V.G. Erwin, R.C. Baker, R.A. Deitrich. Selective breeding of rats differing in sensitivity to the effects of acute ethanol administration. Alcohol Clin. Exp. Res. 16 (1992) 48-54, https://doi:10.1111/j.1530-0277.1992.tb00634.x.

[4] M.D. Schechter. Rats bred for differences in preference to cocaine: Other behavioral measurements. Pharmacol. Biochem. Behav. 43 (1992) 1015-1021, https://doi:10.1016/0091-3057(92)90475-u.

[5] D.H. Overstreet, A.H. Rezvani, O. Pucilowski, L. Gause, D.S. Janowsky. Rapid selection for serotonin-1A sensitivity in rats. Psychiatr. Genet. 4 (1994) 57-62, https://doi:10.1097/00041444-199421000-00008.

[6] L.G. Koch, S.L. Britton. Artificial selection for intrinsic aerobic endurance running capacity in rats. Physiol. Genomics. 5 (2001) 45-52, https://doi:10.1152/physiolgenomics.2001.5.1.45.

[7] C.J. Gordon, A.H. Rezvani. Genetic selection of rats with high and low body temperatures. J Therm. Biol. 26 (2001) 223-229, https://doi:10.1016/s0306-4565(00)00046-2.

[8] M. Johannesson, R. López -Aumatell, P. Stridh, et al. A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. Genome Res. 19 (2009) 150-158, https://doi:10.1101/gr.081497.108.

[9] A.S. Chitre, O. Polesskaya, K. Holl, et al. Genome-Wide Association Study in 3,173 Outbred Rats Identifies Multiple Loci for Body Weight, Adiposity, and Fasting Glucose. Obesity (Silver Spring). 28 (2020) 1964-1973, hppps://doi: 10.1002/oby.22927.

[10] L.C. Solberg Woods, K. Holl, M. Tschannen, W. Valdar. Fine-mapping a locus for glucose tolerance using heterogeneous stock rats. Physiol. Genomics. 41 (2010) 102-108, https://doi:10.1152/physiolgenomics.00178.2009.

[11] L.C. Solberg Woods, C. Stelloh, K.R. Regner, T. Schwabe, J. Eisenhauer, M.R. Garrett. Heterogeneous stock rats: a new model to study the genetics of renal phenotypes. Am. J. Physiol. Renal Physiol. 298 (2010) F1484-F1491. https://doi:10.1152/ajprenal.00002.2010.

[12] L.C. Solberg Woods, K.L. Holl, D. Oreper, Y. Xie, S.W. Tsaih, W. Valdar. Fine-mapping diabetes-related traits, including insulin resistance, in heterogeneous stock rats. Physiol. Genomics. 44 (2012) 1013-1026, https://doi.org/10.1152/physiolgenomics.00040.2012.

[13] I. Alam, D.L. Koller, Q. Sun, et al. Heterogeneous stock rat: a unique animal model for mapping genes influencing bone fragility. Bone. 48 (2011) 1169-1177, https://doi:10.1016/j.bone.2011.02.009.

[14] A. Sánchez-González, A. Esnal, C. Río-Álamos, et al. Association between prepulse inhibition of the startle response and latent inhibition of two-way avoidance acquisition: A study with heterogeneous NIH-HS rats. Physiol. Behav. 155 (2016) 195-201, https://doi.org/10.1016/j.physbeh.2015.12.011.

[15] G.R. Keele, J.W. Prokop, H. He, et al. Genetic fine-mapping and identification of candidate genes and variants for adiposity traits in outbred rats. Obesity (Silver Spring). 26 (2018) 213-222, https://doi:10.1002/oby.22075.

[16] L.C. Solberg Woods, A.A. Palmer. Using heterogeneous stocks for fine-mapping genetically complex traits. Methods Mol. Biol. 2018 (2019) 233-247, https://doi:10.1007/978-1-4939-9581-3_11.

[17] Rat Genome Sequencing and Mapping Consortium, Baud A, Hermsen R, et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat. Genet. 45 (2013) 767-775, https://doi:10.1038/ng.2644.

[18] R. Hermsen, J. de Ligt, W. Spee, et al. Genomic landscape of rat strain and substrain variation. BMC Genomics. 16 (2015) 357, https://doi:10.1186/s12864-015-1594-1.

[19] A.L. Zinski, S. Carrion, J.J. Michal, M.A. Gartstein, R,M. Quock, J.F. Davis, Z. Jiang. Genome-to-phenome research in rats: progresses and perspectives. Int J Biol Sci.(2020) in press.

[20] A.F. Gileta, J. Gao, A.S. Chitre, et al. Adapting genotyping-by-sequencing and variant calling for heterogeneous stock rats [published online ahead of print, 2020 May 12]. G3 (Bethesda). 10 (2020) 2195-2205, https://doi:10.1534/g3.120.401325.

[21] Z. Jiang, H. Wang, J.J. Michal, X. Zhou, B. Liu, L.C. Woods, R.A. Fuchs. Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. Int J Biol Sci. 12(2016):100-8. https://doi: 10.7150/ijbs.13498.

[22] B. Langmead, S.L. Salzberg. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9 (2012) 357-359, https://doi:10.1038/nmeth.1923.

[23] H. Li, B. Handsaker, A. Wysoker, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 16 (2009) 2078-2079, https://doi:10.1093/bioinformatics/btp352.

[24] A. McKenna, M. Hanna, E. Banks, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (2010) 1297-1303. https://doi:10.1101/gr.107524.110.

[25] W. Wang, X. Zhang, X. Zhou, et al. Deep genome resequencing reveals artificial and natural selection for visual deterioration, plateau adaptability and high prolificacy in Chinese domestic sheep. Front. Genet. 10 (2019) 300, https://doi.org/10.3389/fgene.2019.00300.

[26] B. Gel, E. Serra. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*. 33 (2017) 3088-3090, https://doi:10.1093/bioinformatics/btx346.

[27] C.C. Cockerham, B.S. Weir. Estimation of gene flow from F-statistics. Evolution. 47 (1993) 855-863, https://doi:10.1111/j.1558-5646.1993.tb01239.x.

[28] E. Axelsson, A. Ratnakumar, M.L. Arendt, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 495 (2013) 360-364, https://doi:10.1038/nature11837.

[29] Y.J. Shen, H. Jiang, J.P. Jin, et al. Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 13 (2004) 1198-1205, https://doi:10.1104/pp.103.038463.

[30] V. Guryev, M.J. Koudijs, E. Berezikov, et al. Genetic variation in the zebrafish. Genome Res. 16 (2006) 491-497, https://doi:10.1101/gr.4791006.

[31] R.E. Mills, C.T. Luttig, C.E. Larkins, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 16 (2006) 1182-1190, https://doi:10.1101/gr.4565806.

[32] M. Brandström, T.H. Ellegren. The genomic landscape of short insertion and deletion polymorphisms in the chicken (*Gallus gallus*) genome: a high frequency of deletions in tandem duplicates. Genetics. 176 (2007) 1691-1701, https://doi:10.1534/genetics.107.070805.

[33] T.M. Keane, L. Goodstadt, P. Danecek, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 477 (2011) 289-294, https://doi:10.1038/nature10413.

[34] B. Zhan, J. Fadista, B. Thomsen, J. Hedegaard, F. Panitz, C. Bendixen. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC Genomics. 12 (2011) 557, https://doi:10.1186/1471-2164-12-557.

[35] 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 491 (2012) 56-65, https://doi:10.1038/nature11632.

[36] J. Evans, R.F. McCormick, D. Morishige, et al. Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. PLoS One. 8 (2013) e79192, https://doi:10.1371/journal.pone.0079192.

[37] J. Qi, X. Liu, D. Shen, et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat. Genet. 45 (2013) 1510-1515, https://doi:10.1038/ng.2801

[38] F. Coll, M. Preston, J.A. Guerra-Assunção, et al. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb). 94 (2014) 346-354, https://doi:10.1016/j.tube.2014.02.005

[39] Y. Yan, G. Yi, C. Sun, L. Qu, N. Yang. Genome-wide characterization of insertion and deletion variation in chicken using next generation sequencing. PLoS One. 9 (2014) e104652, https://doi:10.1371/journal.pone.0104652.

[40] D.C. Jeffares, A. Pain, A. Berry, et al. Genome variation and evolution of the malaria parasite *Plasmodium falciparum* [published correction appears in Nat. Genet. 39 (2007) 567] [published correction appears in Nat Genet. 39 (2007) 422]. Nat Genet. 39 (2007) 120-125, https://doi:10.1038/ng1931.

[41] G. Ramakrishna, P. Kaur, D. Nigam, et al. Genome-wide identification and characterization of InDels and SNPs in *Glycine max* and *Glycine soja* for contrasting seed permeability traits. BMC Plant Biol. 18 (2018) 14, https://doi:10.1186/s12870-018-1341-2.

[42] A.E. Vinogradov, O V. Anatskaya. DNA helix: the importance of being AT-rich. Mamm Genome. 28 (2017) 455-464, https://doi: 10.1007/s00335-017-9713-8.

[43] F.C. Ceballos, P.K. Joshi, D.W. Clark, M. Ramsay, J.F. Wilson. Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. 19 (2018) 220-234, https://doi:10.1038/nrg.2017.109.

[44] F.J. Miller, F.L. Rosenfeldt, C. Zhang, A.W. Linnane, P. Nagley. Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. Nucleic Acids Res. 31 (2003) e61, https://doi:10.1093/nar/gng060.

[45] Z. Jiang, X.L. Wu, M. Zhang, J.J. Michal, R.W. Wright Jr. The complementary neighborhood patterns and methylation-to-mutation likelihood structures of 15,110 single-nucleotide polymorphisms in the bovine genome. Genetics. 180 (2008) 639-647, https://doi:10.1534/genetics.108.090860.

[46] C.S. Nabel, S.A. Manning, R.M. Kohli. The curious chemical biology of cytosine: deamination, methylation, and oxidation as modulators of genomic potential. ACS Chem. Biol. 7 (2012) 20-30, https://doi:10.1021/cb2002895.

[47] S. Ramdas, A.B. Ozel, M.K. Treutelaar, et al. Extended regions of suspected mis-assembly in the rat reference genome. Sci. Data. 6 (2019) 39, https:// doi:10.1038/s41597-019-0041-6.

[48] S.S. Atanur, A.G. Diaz, K. Maratou, et al. Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. Cell. 154 (2013) 691-703, https://doi:10.1016/j.cell.2013.06.040.

[49] X. Guo, M. Brenner, X. Zhang, et al. Whole-genome sequences of DA and F344 rats with different susceptibilities to arthritis, autoimmunity, inflammation and cancer. Genetics. 194 (2013) 1017-1028, https://doi:10.1534/genetics.113.153049.

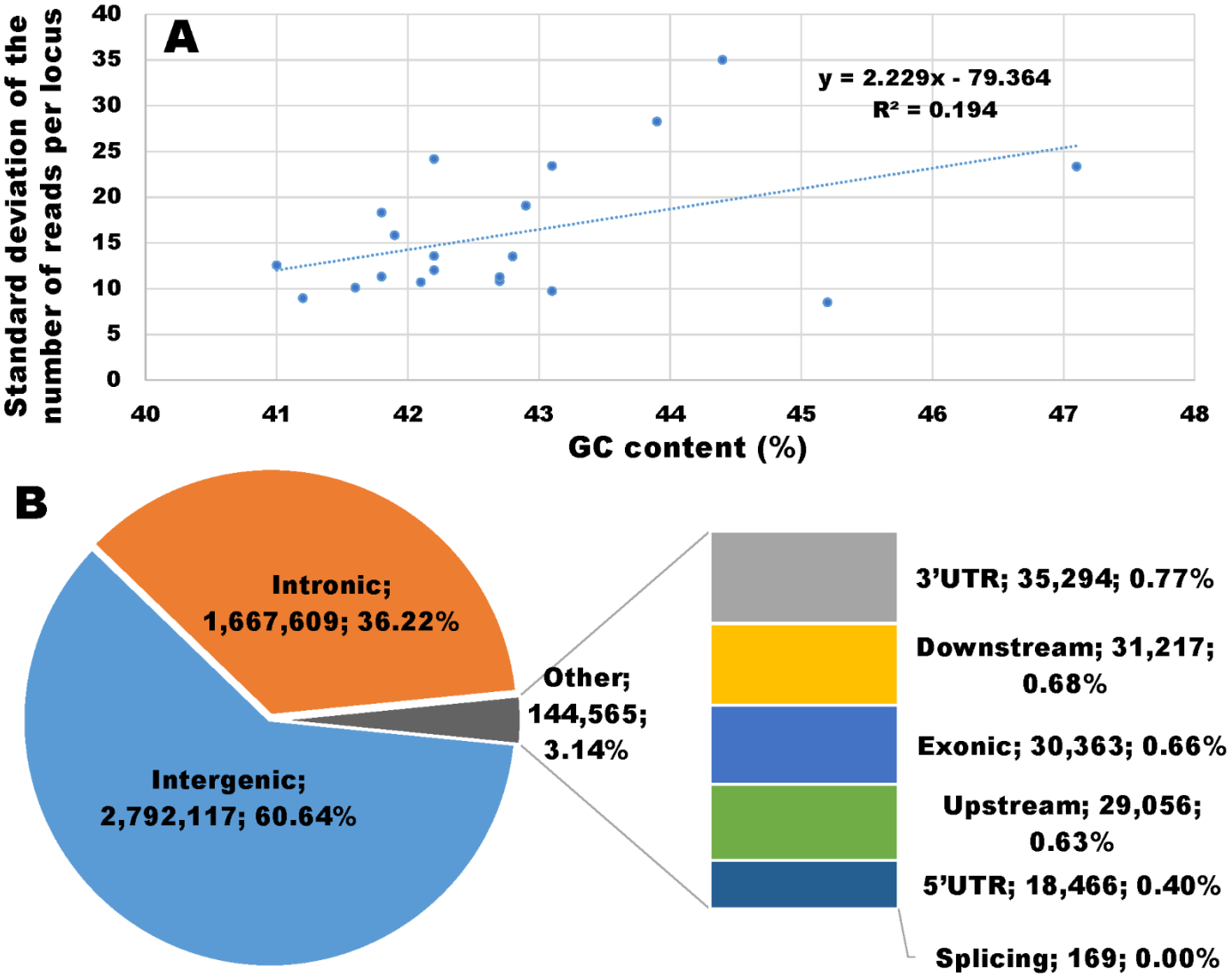

Figure 1. Characterization of DNA variants in rats. (A) Standard deviation of reads per locus is affected by GC-content. Each point represents one chromosome. The line of best fit is represented by the dashed line. (B) Distribution of DNA variants by genomic regions.

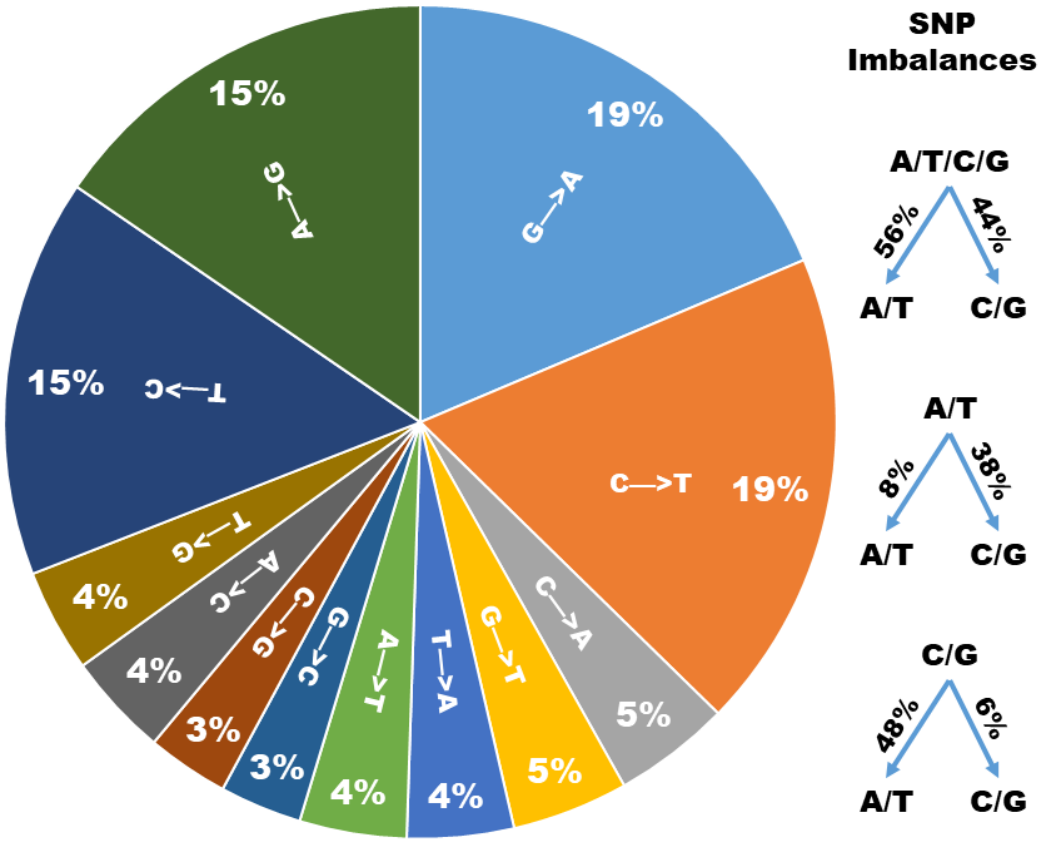

Figure 2. Characterization of bi-allelic SNPs in rats. When a one-way model is considered, there are a total of 12 bi-allelic SNP types, which form six balanced pairs with equal frequencies. However, imbalances appear across the pairs. Genome-wide, A/T as the endpoints of mutation is favored. It is highly likely that A/T are mutated to C/G, while C/G to A/T.

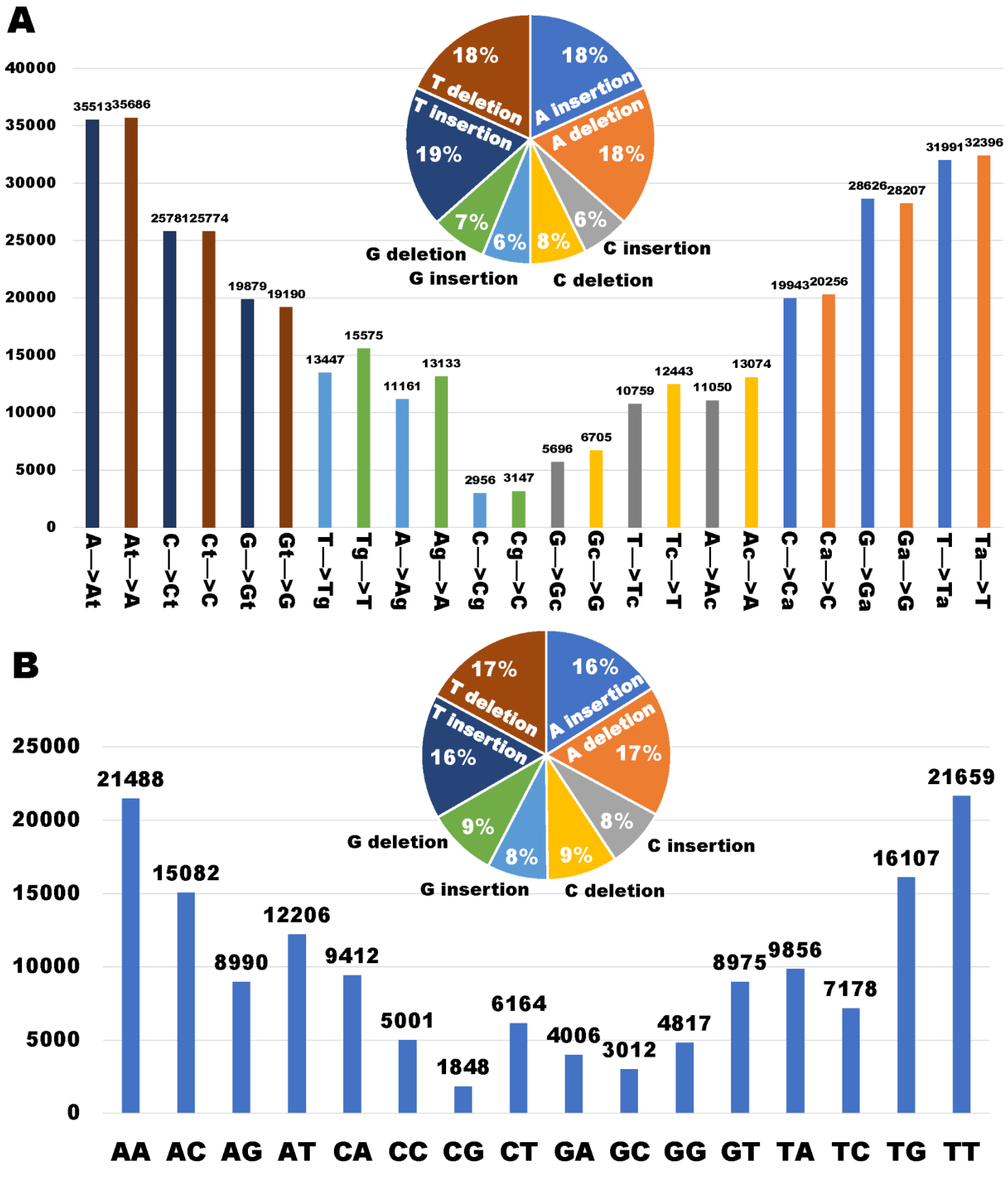

Figure 3. Characterization of INDELs in rats. (A) 1-bp INDELs in rats with a total of 12 INDEL pairs. (B) 2-bp INDELs in rats with a total of 16 nucleotide combinations.

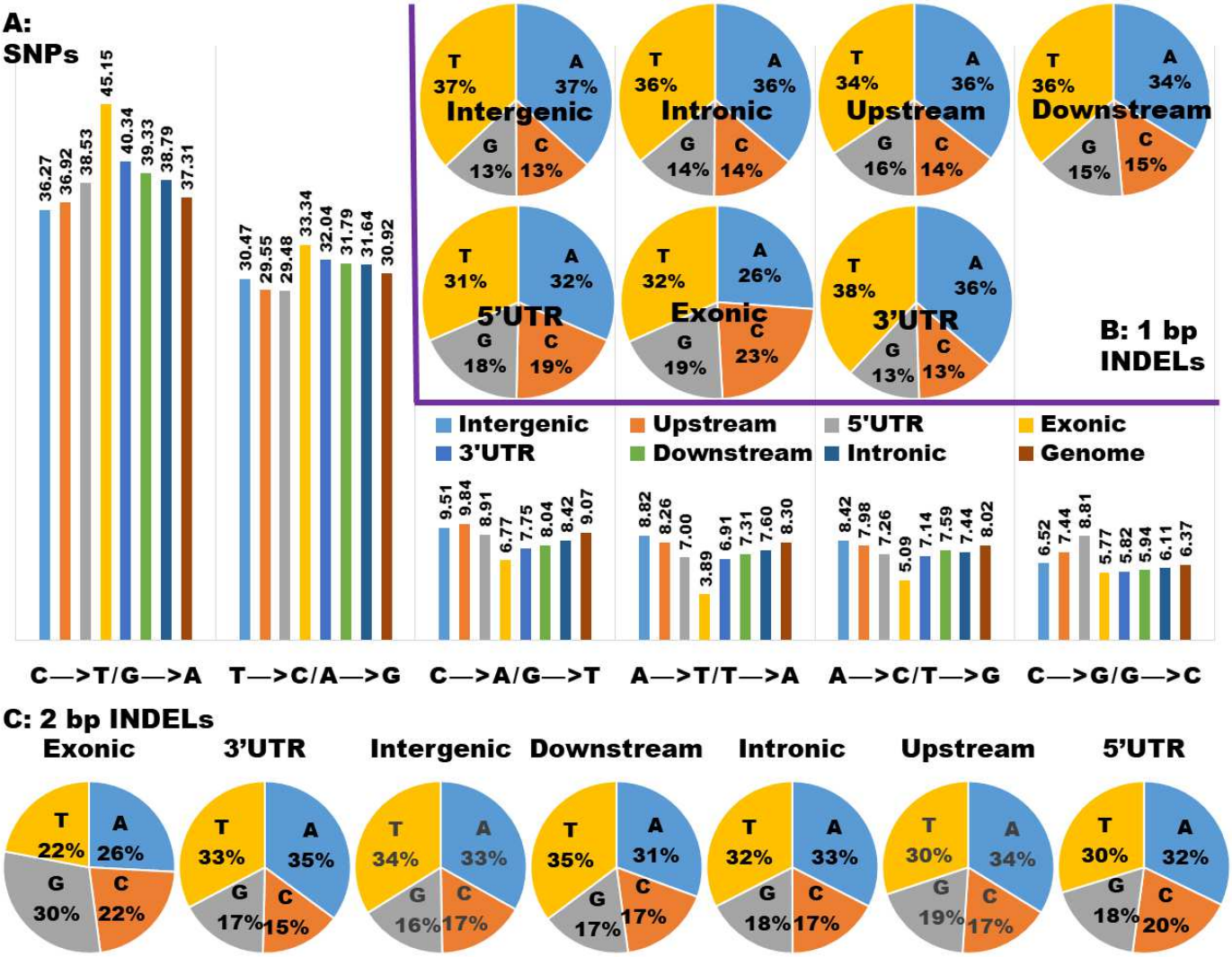

Figure 4. Genomic regions and characteristics of DNA variation in rats. Effects of genomics regions on frequencies (%) of (A) bi-allelic SNPs, (B) 1-bp INDELs and (C) 2-bp INDELs (C).

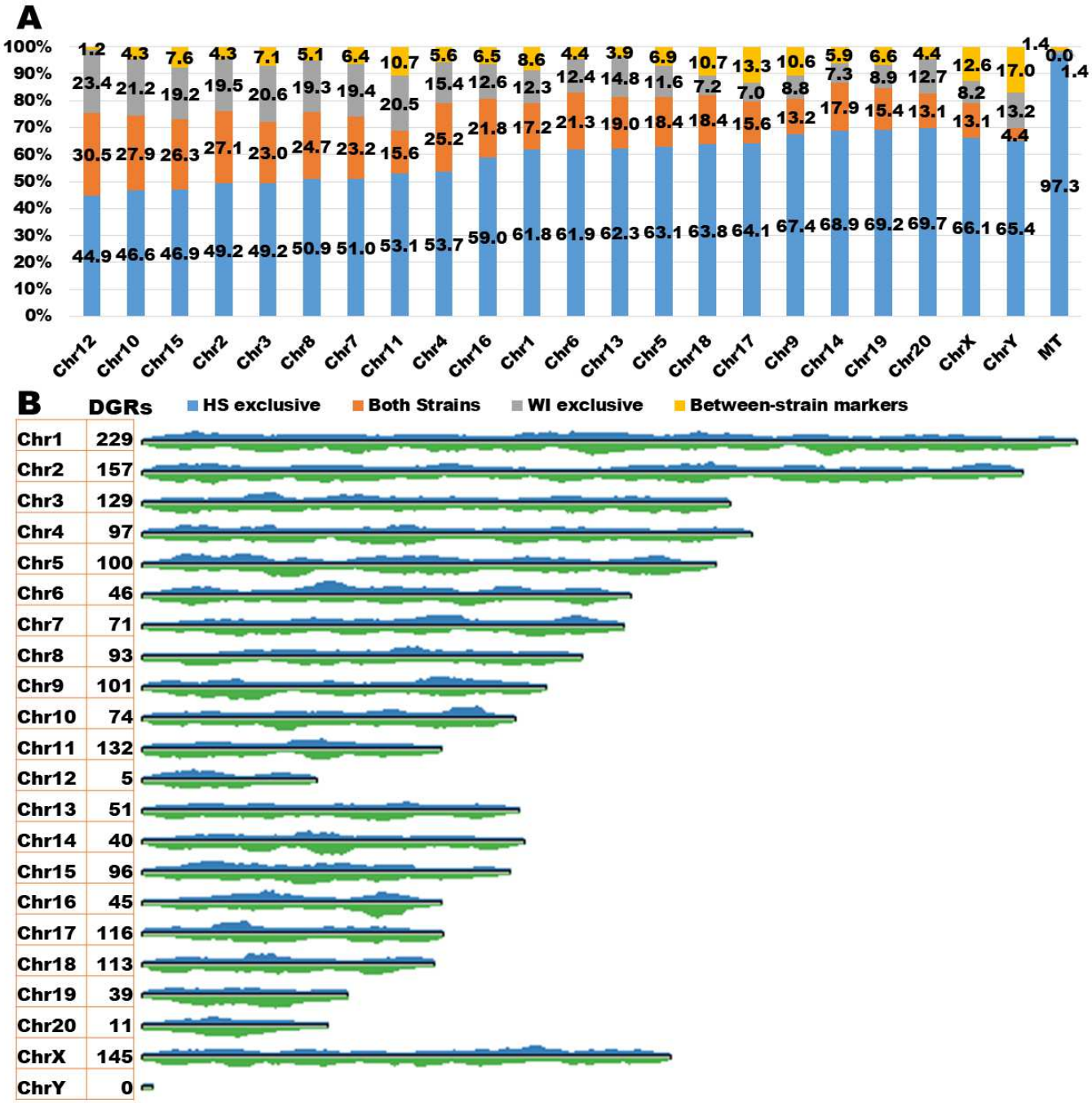

Figure 5. Characterization of genome variation between WI and NIH-HS rats. (A) Strain related DNA variants in autosomes, sex chromosomes and/or MT genomes. (B) Homozygosity distances along each chromosome. DGRs: divergent genome regions. Chr: chromosome.

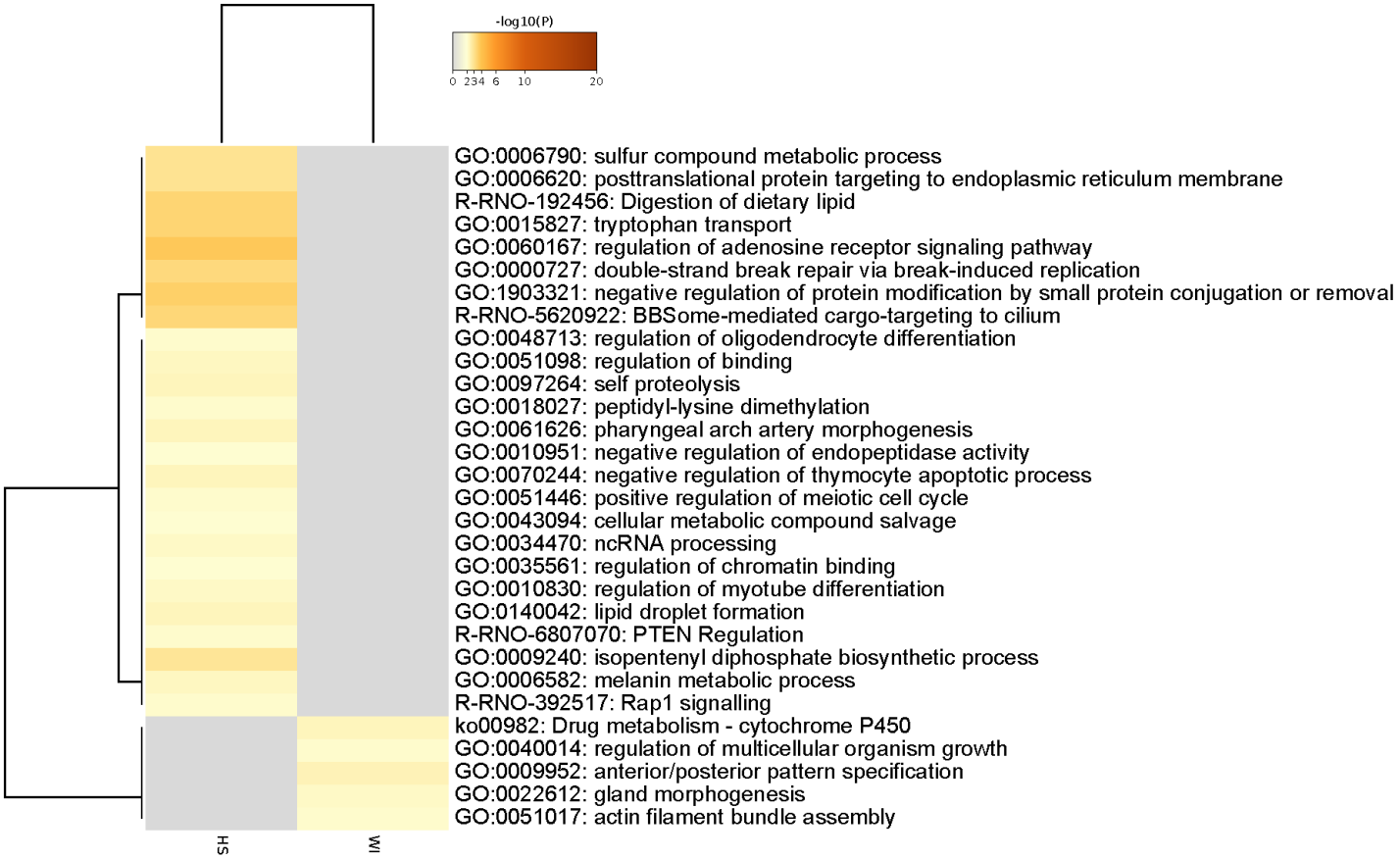

Figure 6. Gain or loss of mutations in genes and their enriched pathways in WI and NIH-HS rats. Pathways were enriched using 745 genes with polymorphisms in NIH-HS (but with monomorphism in WI) and using 96 genes with polymorphisms in WI (but with monomorphisms in NIH-HS rats).

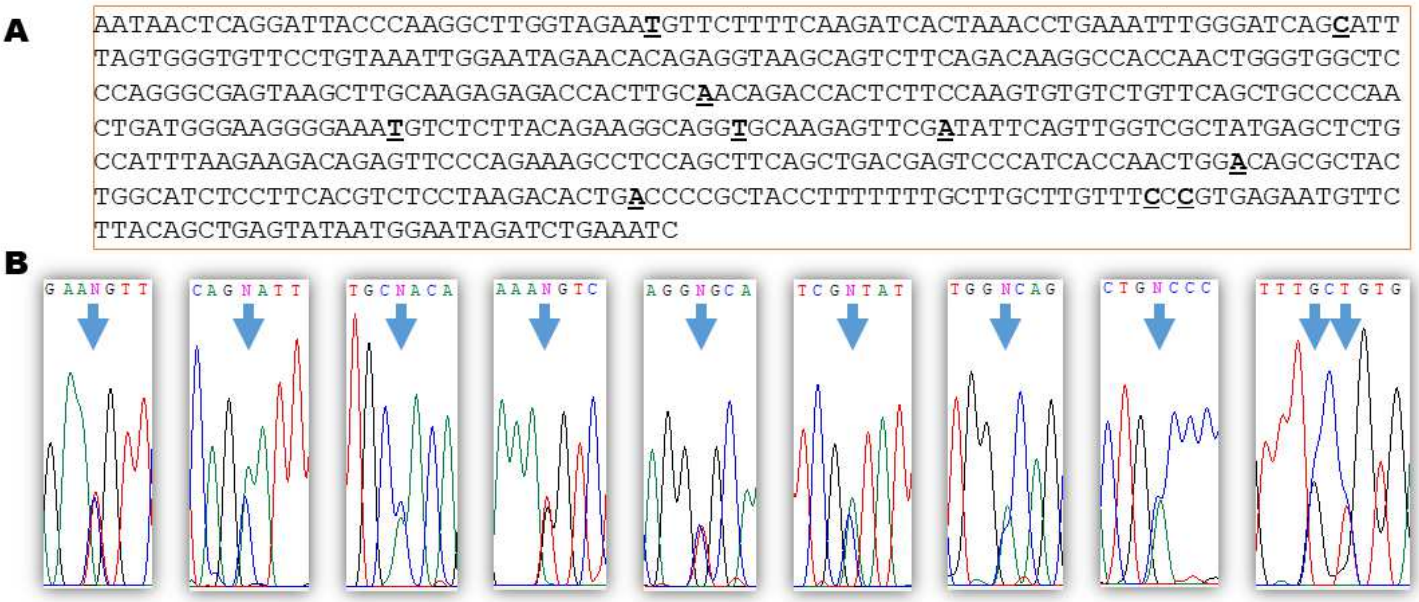

Figure 7. Validation of SNPs in the rat *Dat1* gene. (A) A hotspot of 10 SNPs detected in a region of the *Dat1* gene. (B) Sanger sequencing of individual DNA samples confirmed all 10 SNPs.

Table 1. Basics of genetic variants in WI and NIH-HS rats

Group	INDELs	SNPs	Total	SNPs:INDELs
Commonly polymorphic	227,172	748,464	975,636	3.29
NIH-HS exclusive	518,090	2,112,420	2,630,510	4.08
Between-strain	70,548	233,713	304,261	3.31
WI exclusive	157,212	536,672	693,884	3.41
Total	973,022	3,631,269	4,604,291	3.73
NIH-HS total variants	745,262	2,860,884	3,606,146	3.84
WI total variants	384,384	1,285,136	1,669,520	3.34

Figures

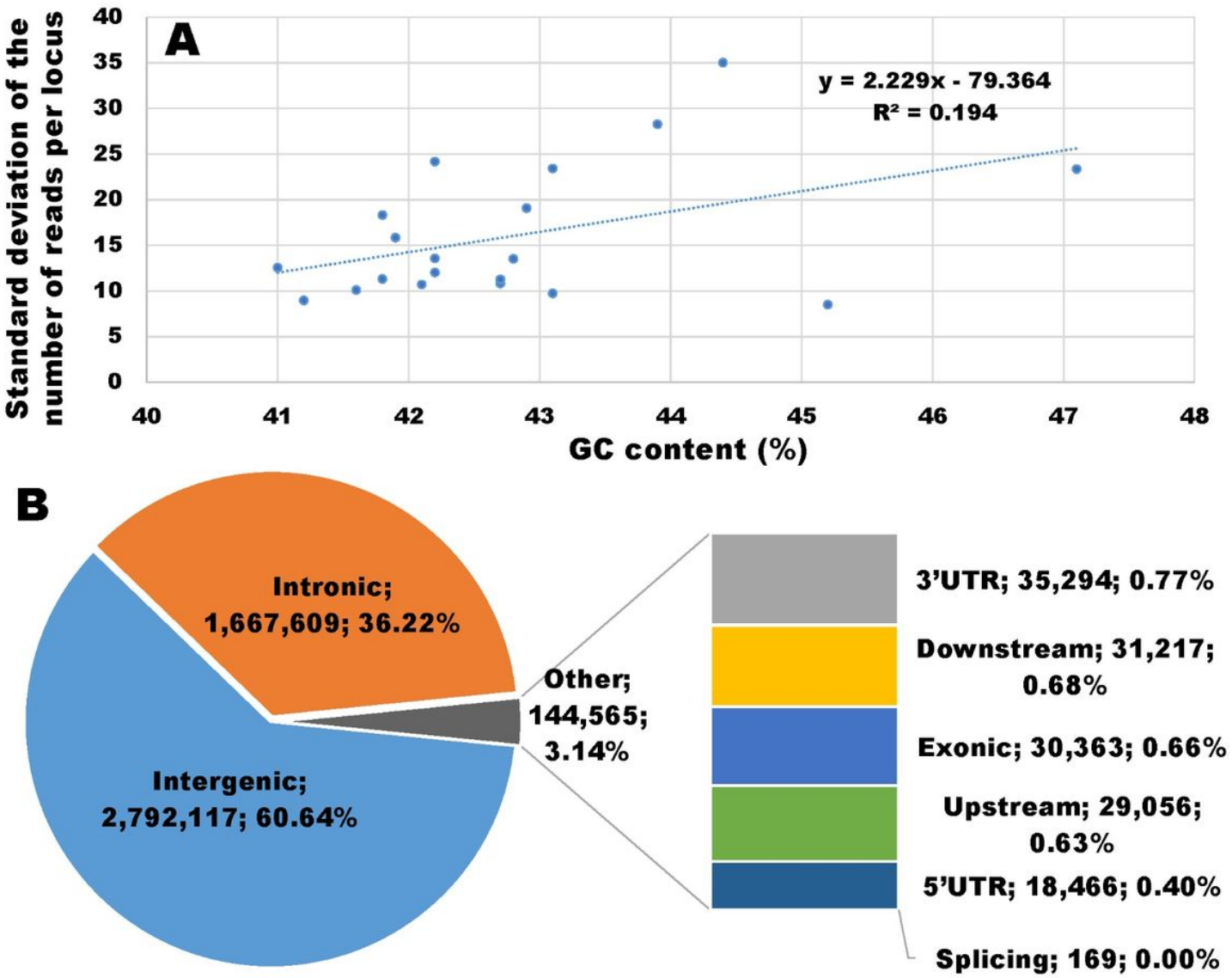

Figure 1

Characterization of DNA variants in rats. (A) Standard deviation of reads per locus is affected by GC-content. Each point represents one chromosome. The line of best fit is represented by the dashed line. (B) Distribution of DNA variants by genomic regions.

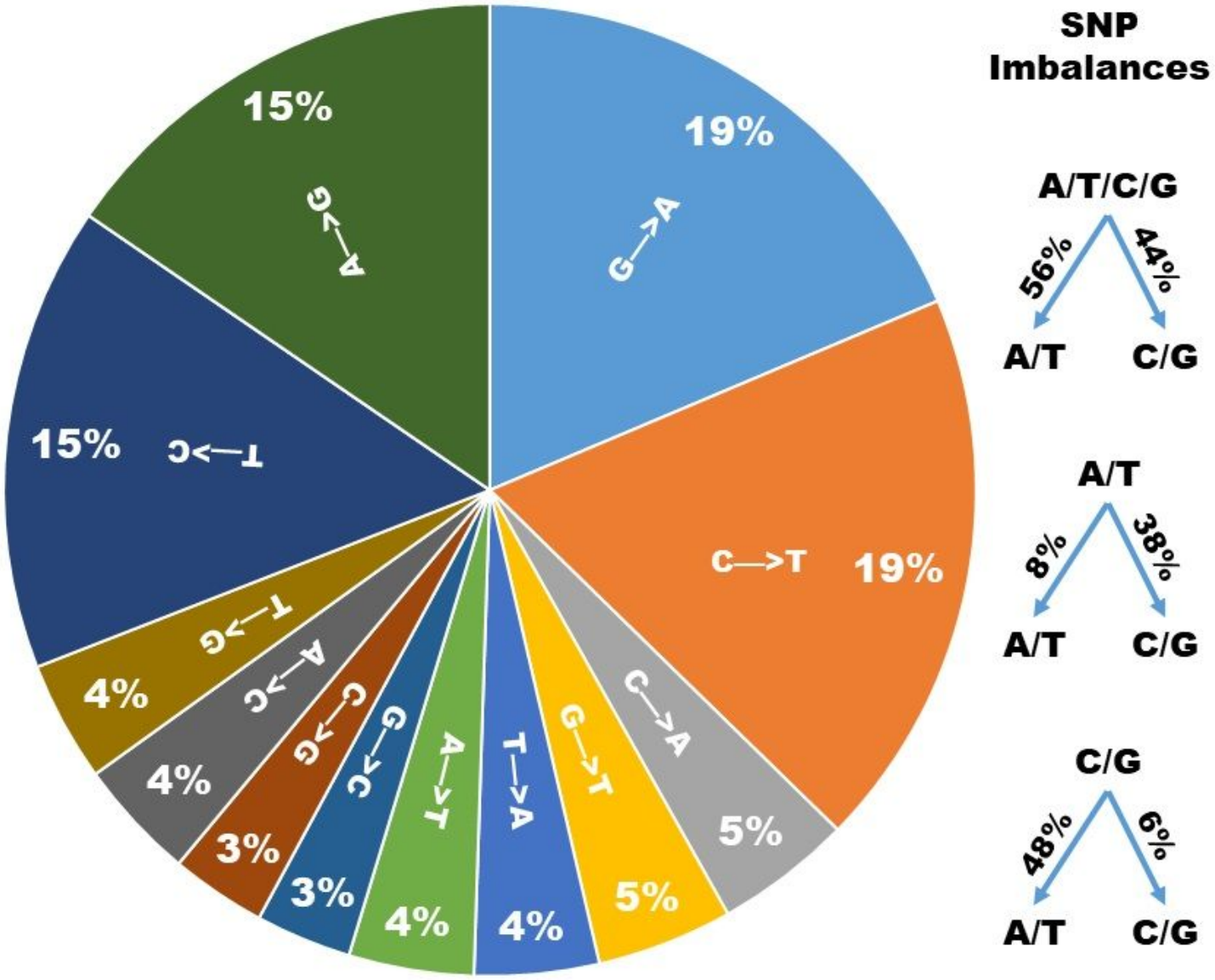

Figure 2

Characterization of bi-allelic SNPs in rats. When a one-way model is considered, there are a total of 12 bi-allelic SNP types, which form six balanced pairs with equal frequencies. However, imbalances appear across the pairs. Genome-wide, A/T as the endpoints of mutation is favored. It is highly likely that A/T are mutated to C/G, while C/G to A/T.

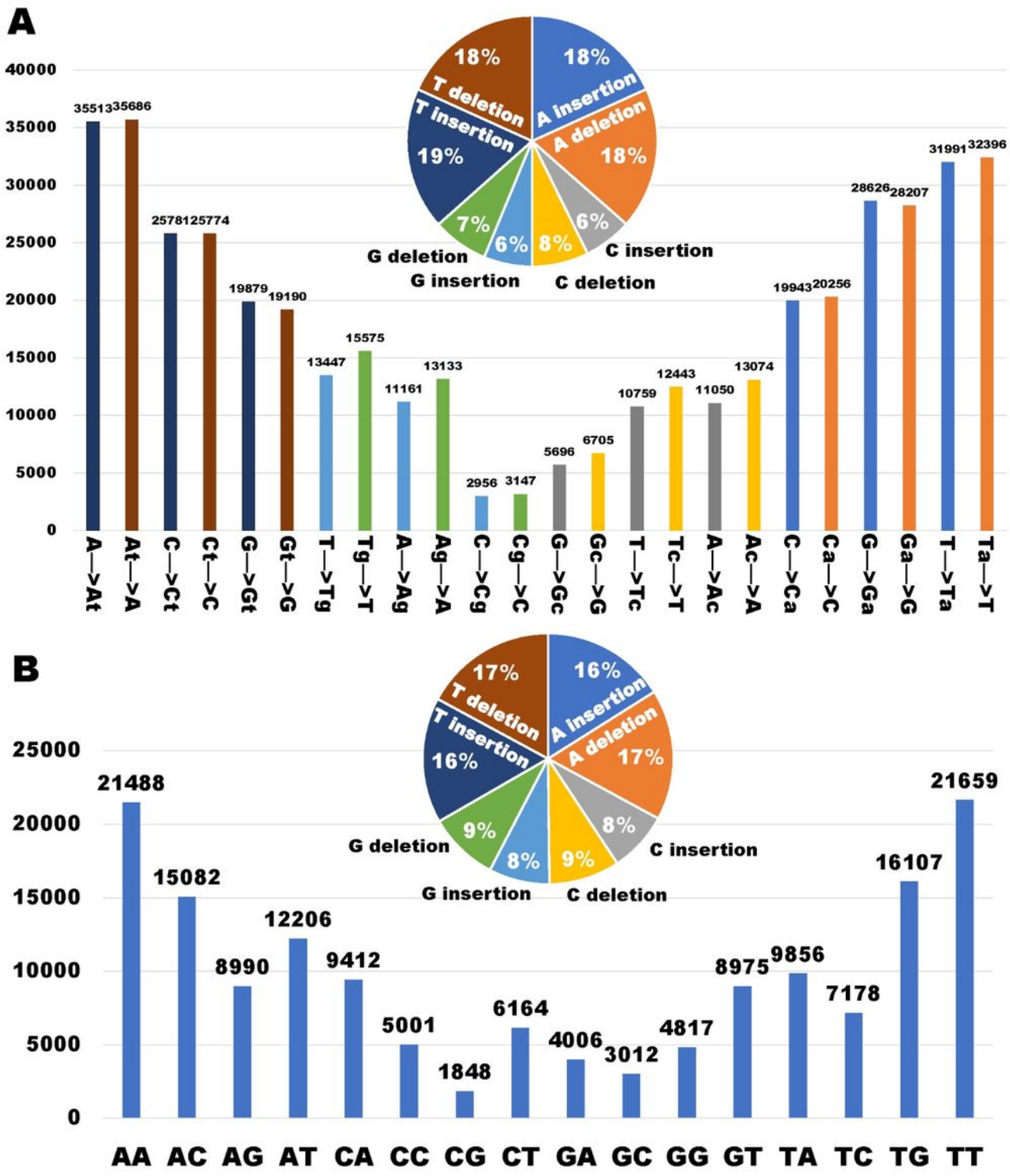

Figure 3

Characterization of INDELs in rats. (A) 1-bp INDELs in rats with a total of 12 INDEL pairs. (B) 2-bp INDELs in rats with a total of 16 nucleotide combinations.

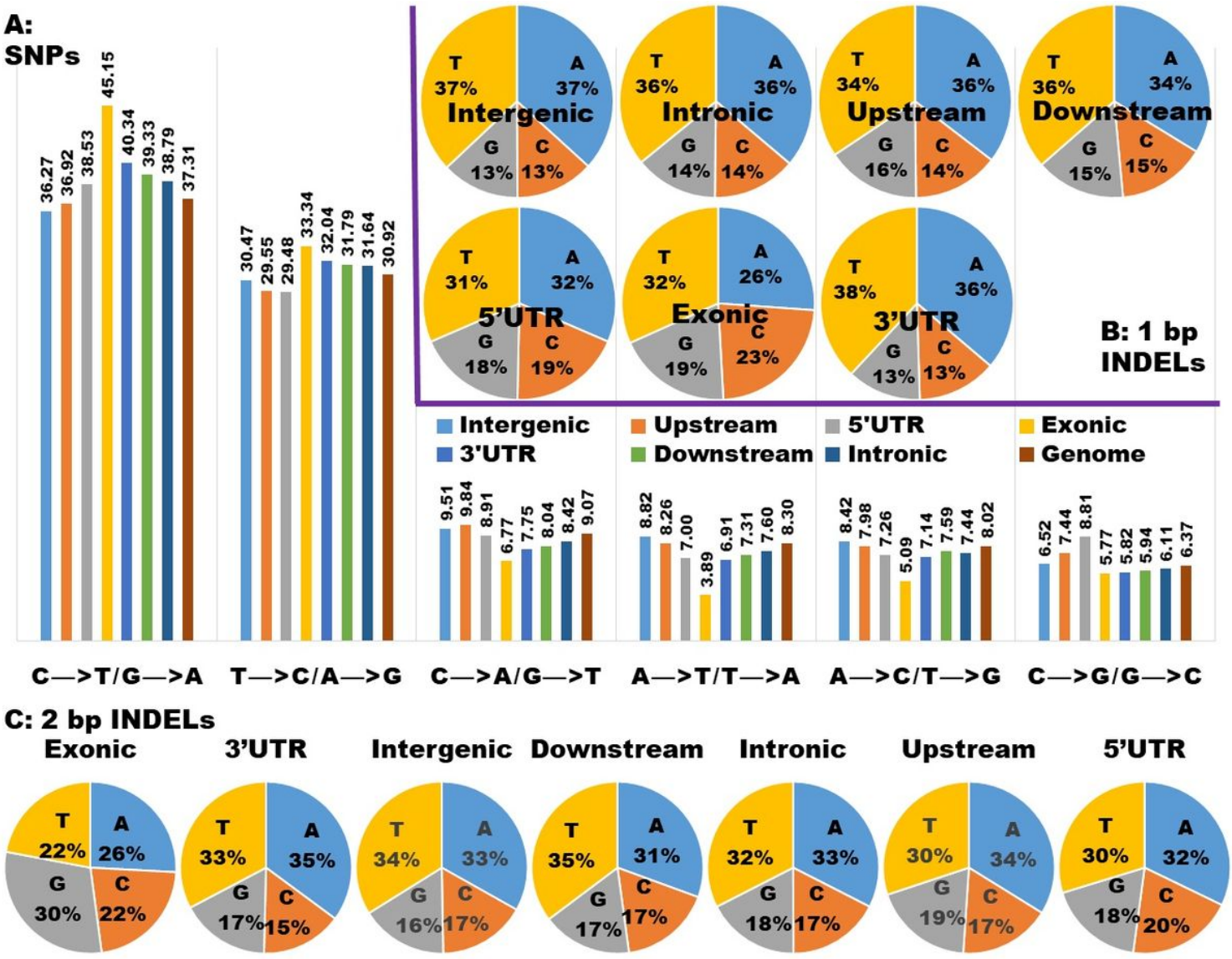

Figure 4

Genomic regions and characteristics of DNA variation in rats. Effects of genomics regions on frequencies (%) of (A) bi-allelic SNPs, (B) 1-bp INDELs and (C) 2-bp INDELs (C).

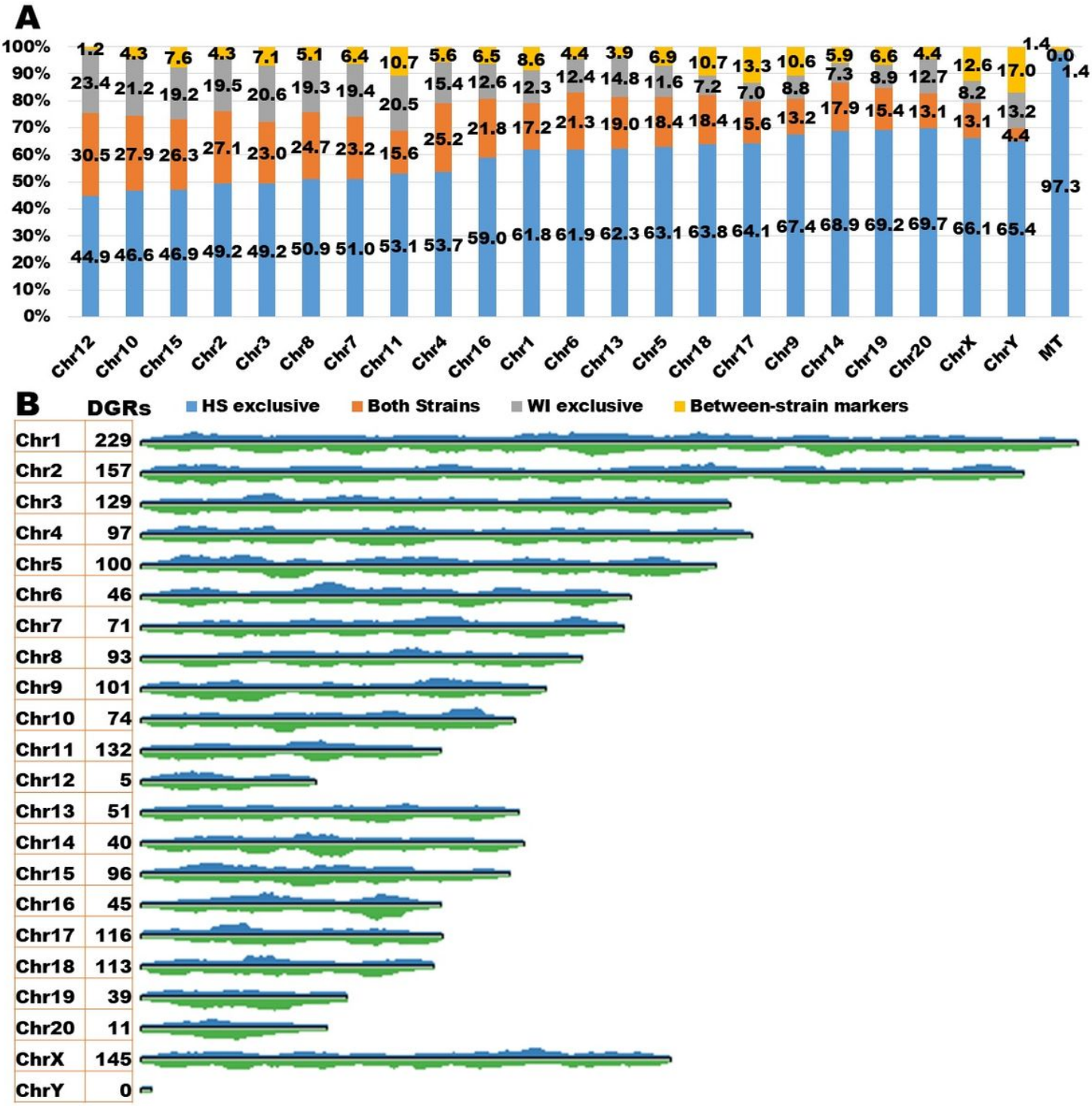

Figure 5

Characterization of genome variation between WI and NIH-HS rats. (A) Strain related DNA variants in autosomes, sex chromosomes and/or MT genomes. (B) Homozygosity distances along each chromosome. DGRs: divergent genome regions. Chr: chromosome.

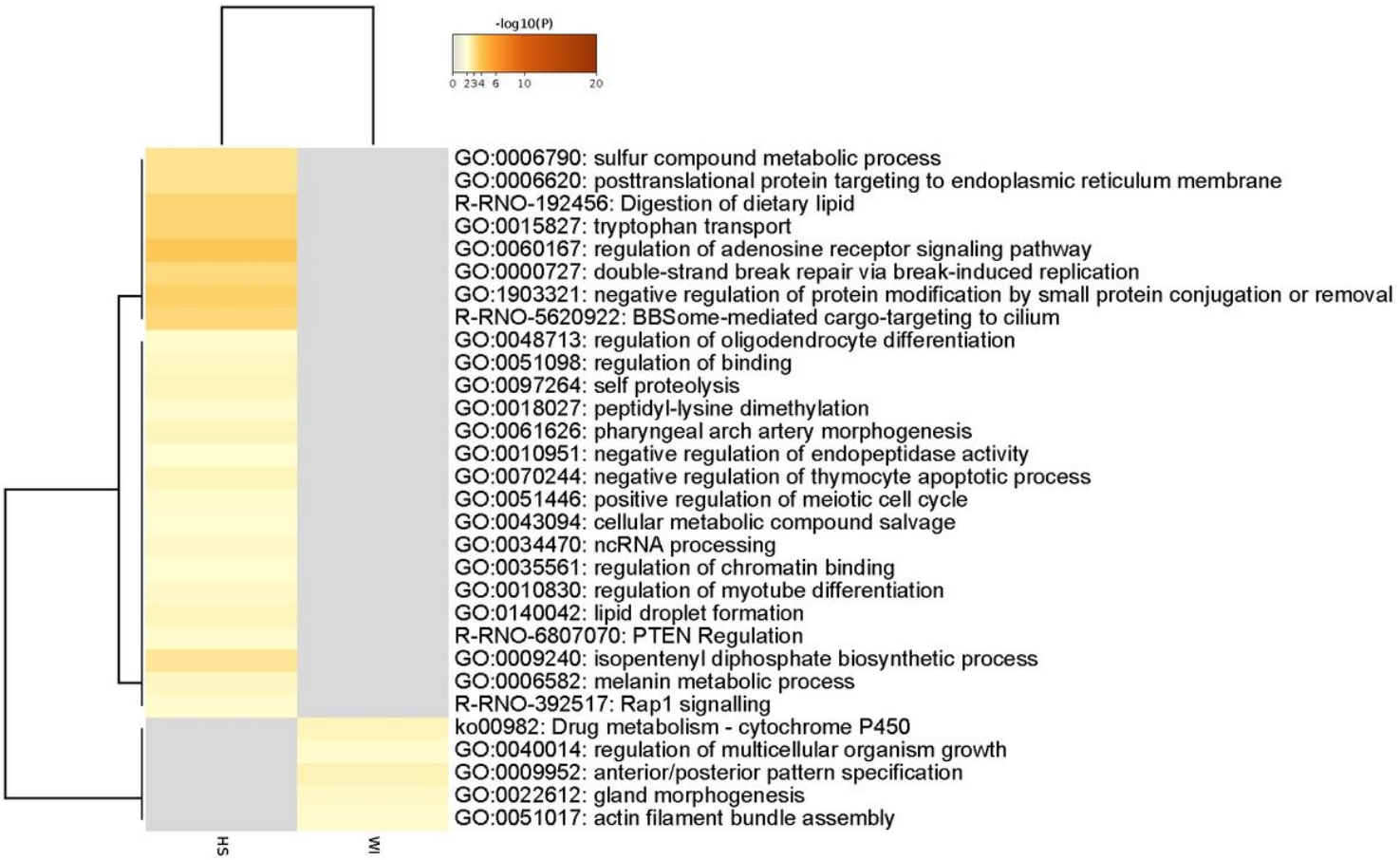

Figure 6

Gain or loss of mutations in genes and their enriched pathways in WI and NIH-HS rats. Pathways were enriched using 745 genes with polymorphisms in NIH-HS (but with monomorphism in WI) and using 96 genes with polymorphisms in WI (but with monomorphisms in NIH-HS rats).

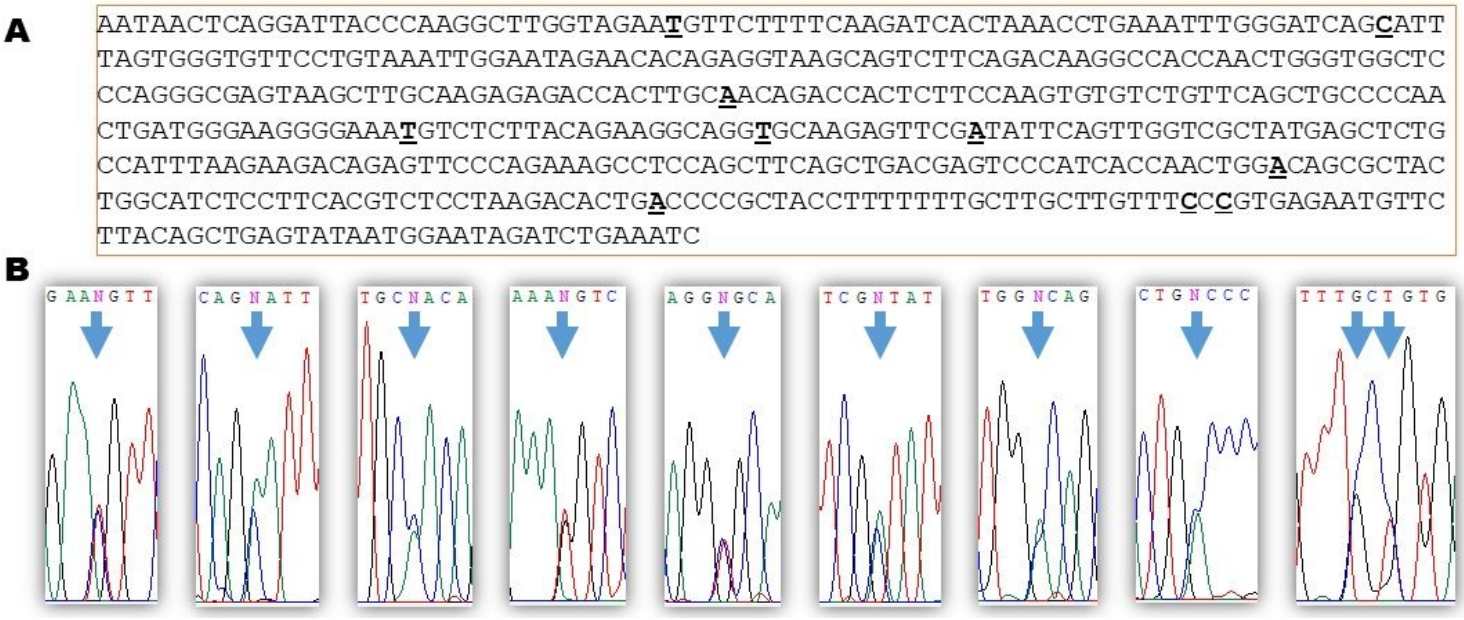

Figure 7

Validation of SNPs in the rat Dat1 gene. (A) A hotspot of 10 SNPs detected in a region of the Dat1 gene. (B) Sanger sequencing of individual DNA samples confirmed all 10 SNPs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1RatGenomeVariantBasics.xlsx
- TableS2RatGenomeVariantImbalances.xlsx
- TableS3DivergentGenomeRegionsbetweenHSandWIFst.xlsx
- TableS4DivergentVariantLocibetweenHSandWIFst.xlsx
- TableS5RatSlc6a3geneannotation.docx
- TableS6GeneticVariantsinRatDat1gene.xlsx
- nrreportingsummary.pdf