A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats

Highlights

- mRatBN7.2 is a rat reference genome with improved contiguity and accuracy
- Gene annotations, from both RefSeq and Ensembl, are improved with mRatBN7.2
- Our analysis of 120 strains/substrains of rats found 20 million sequence variations
- A refined phylogenetic tree reveals the relationships between laboratory rats

Authors

Tristan V. de Jong, Yanchao Pan, Pasi Rastas, ..., Robert W. Williams, Jun Z. Li, Hao Chen

Correspondence

junzli@med.umich.edu (J.Z.L.), hchen@uthsc.edu (H.C.)

In brief

de Jong et al. evaluated the seventh assembly of the rat reference genome, mRatBN7.2, and found that it reduces base-level errors and increases contiguity, although some misassemblies remain. Gene annotations are now more complete. Analysis of whole genomes representing 120 rat strains/substrains revealed 20 million sequence variations. Phylogenetic analysis refined ancestral relationships among these strains. In addition, a new rat genetic map, along with annotated transcription start sites and alternative polyadenylation sites based on mRatBN7.2, is provided.
A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats

Tristan V. de Jong,1,27 Yanchao Pan,2,27 Pasi Rastas,3 Daniel Munro,4,5 Monika Tutaj,6,7 Huda Akil,8 Chris Benner,9 Denghui Chen,4 Apurva S. Chitre,4 William Chow,10 Vincenza Colonna,11,12 Clifton L. Dalgard,13 Wendy M. Demos,6,7 Peter A. Doris,14 Erik Garrison,12 Aron M. Geurts,15 Hakan M. Gunturkun,1 Victor Guryev,15 Thibaut Hourlier,16 Kerstin Howe,10 Jun Huang,7 Ted Kalbfleisch,17 Panjun Kim,12 Ling Li,12,18

SUMMARY

The seventh iteration of the reference genome assembly for *Rattus norvegicus*—mRatBN7.2—corrects numerous misplaced segments and reduces base-level errors by approximately 9-fold and increases contiguity by 290-fold compared with its predecessor. Gene annotations are now more complete, improving the mapping precision of genomic, transcriptomic, and proteomics datasets. We jointly analyzed 163 short-read whole-genome sequencing datasets representing 120 laboratory rat strains and substrains using mRatBN7.2. We defined ~20.0 million sequence variations, of which 18,700 are predicted to potentially impact the function of 6,677 genes. We also generated a new rat genetic map from 1,893 heterogeneous stock rats and annotated transcription start sites and alternative polyadenylation sites. The mRatBN7.2 assembly, along with the extensive analysis of genomic variations among rat strains, enhances our understanding of the rat genome, providing researchers with an expanded resource for studies involving rats.

INTRODUCTION

*Rattus norvegicus* has been used in many fields of study related to human disease.1 The earliest studies using brown rats appeared in the early 1800s.2,3 The Wistar rats, the ancestor of many laboratory strains, were bred for scientific research in 1906.4 Over 4,000 inbred, outbred, congenic, mutant, and transgenic strains have been created and are documented in the Rat Genome Database (RGD).5 Approximately 500 are available from the Rat Resource and Research Center.6 Several genetic reference populations, including the HXB/BXH7 and FXLE/LEXF8 recombinant inbred (RI) families, are also available. Both families, together with 30 diverse classical inbred strains,7 are now part of the Hybrid Rat Diversity Panel (HRDP). Several genetic reference populations, including the HXB/BXH7 and FXLE/LEXF8 recombinant inbred (RI) families, are also available. Both families, together with 30 diverse classical inbred strains,7 are now part of the Hybrid Rat Diversity Panel (HRDP), which can be used to quickly generate any of over 10,000 isogenic and replicable F1 hybrids—all of which are now essentially sequenced. The outbred N/NIH heterogeneous stock (HS) rats, derived from eight inbred strains,8 have been increasingly used for fine mapping of physiological and behavioral traits.11–15 To date, RGD has annotated nearly 2,400 rat quantitative trait loci (QTLs),16 mapped using F2 crosses, RI families, and HS rats. The *Rattus norvegicus* genome was sequenced shortly after the genomes of *Homo sapiens* and *Mus musculus*.17 The inbred Brown Norway (BN/NHsdMcwi) strain, derived by many generations of sibling matings of stock originally from a pen-bred colony of wild rats,7 was used to generate the reference. Several updates were released over the following decade.18–20 Since 2014, most rat genomic and genetic research used the incomplete and problematic Rnor_6.0 assembly.21,22 mRatBN7.2 was created in 2020 by the Darwin Tree of Life/Vertebrate Genome Project (VGP) as the new genome assembly of the BN/NHsdMcwi rat.23 The Genome Reference Consortium (GRC) (https://www.ncbi.nlm.nih.gov/grc/rat) has adopted mRatBN7.2 as the official rat reference genome.

Here, we report extensive analyses of the improvements in mRatBN7.2 compared with Rnor_6.0. To assist the rat research...
Community in transitioning to mRatBN7.2, we conducted a broad analysis of a whole-genome sequencing (WGS) dataset of 163 samples from 120 inbred rat strains and substrains. Joint variant calling led to the discovery of 19,987,273 high-quality variants from 15,804,627 sites. Additional resources created during our analysis included a rat genetic map, a comprehensive phylogenetic tree for the 120 (sub)strains, and extensive annotation of identified genes, including their transcription start sites (TSSs) and alternative polyadenylation (APA) sites. This new assembly and its associated resources create a more solid platform for research on the many dimensions of physiology, behavior, and pathobiology of rats and for more reliable and meaningful translation of findings to human populations.

RESULTS

High structural and base-level accuracy of mRatBN7.2
All sequencing data for mRatBN7.2 were generated from a male BN rat from the Medical College of Wisconsin (BN/NHsdMcwi, generation F61). The assembly is based on integrating data across multiple technologies, including long-read sequencing (PacBio CLR), 10X linked-read sequencing, BioNano DLS optical map, and Arima HiC. After automated assembly, manual curation corrected most of the apparent discrepancies among data types. While mRatBN7.2 contains an alternative pseudo-haplotype (GCA_015244455.1), we focused our analysis on the primary assembly (GCF_015227675.2).

Over the last six iterations of the rat reference, genome continuity has improved incrementally (Table S1). Contig N50, one measure of assembly quality, has been ~30 kb between m1 to m5. Rnor_6.0 was the first assembly to include some long-read data and improved contig N50 to 100.5 kb. mRatBN7.2 further improved N50 to 29.2 Mb (Figure S1). Although this measure of contiguity lags slightly behind the mouse reference genome (contig N50 = 59 Mb in GRCm39, released in 2020) and far behind the first telomere-to-telomere human genome (CHM13) (Table 1), it still marks a large improvement (~290 times higher) over Rnor_6.0 (Figure S2). In another measure, the number of contigs in mRatBN7.2 was reduced by 100-fold compared with Rnor_6.0 and is approaching the quality of GRCh38 for humans and GRCm39 for mice (Table 1).

Although aligning Rnor_6.0 with mRatBN7.2 showed a high level of structural agreement (Figures 1A and S3), we identified 36,500 discordant segments between these two assemblies that are longer than 50 bp (Figure S4). To evaluate these differences, we generated a genetic map (Table S2) using data from 378 families of 1,893 HS rats based on the recombination frequency between 150,835 markers. Comparing the order of the markers and their location on the reference confirmed that the order and orientation of genomic segments are more accurate in mRatBN7.2. For example, there is a 17.2 Mb inverted segment at proximal Chr 6 between Rnor_6.0 and mRatBN7.2 (Figure 1B). It remains inverted when the genetic distance is plotted with marker location on Rnor_6.0 (Figure 1C) but is resolved when using marker locations on mRatBN7.2 (Figure 1D). This was also true for several other regions (e.g., Figures 1E–1G for Chr 19 and Figures S5 and S6 for all autosomes). These data indicate that most of the segment-wise differences between the two assemblies are due to errors in Rnor_6.0.
We mapped linked-read WGS data for 36 samples from the HXB/BXH family of strains against both Rnor_6.0 and mRatBN7.2. These included four samples of the reference strain (BN/NHsdMcwi) and 26 samples from the two parental strains—SHR/OlaIpcv and BN-Lx/Cub—and all 30 extent HXB/BXH progeny strains. The mean read depth of the entire HXB dataset, including all parents, is 105.5x (range 23.4–184.5). From mRatBN7.2 to Rnor_6.0, the fraction of reads mapped to the reference increased by 1%–3% (Figure 2A), and regions of the genome with no coverage decreased by ~2% (Figure 2B). Genetic variants (SNPs and indels) were identified using Deepvariant\(^26\) and jointly called for the 36 samples using GLnexus.\(^27\) After quality filtering (qual ≥ 30), we identified 8,286,401 SNPs and 3,527,568 indels in Rnor_6.0. Corresponding numbers for mRatBN7.2 are 5,088,144 SNPs and 1,615,870 indels (Figures 2C and 2D). Surprisingly, variants shared by all samples—either homozygous in all samples or heterozygous in all samples—were more abundant when aligned to Rnor_6.0 (1,310,902) than to mRatBN7.2 (143,254) (Figures 2E and 2F). Because we included sequence data from 4 BN/NHsdMcwi rats, including those used for both Rnor_6.0 and mRatBN7.2, the most parsimonious explanation is that these shared variants are due to the wrong nucleotide sequence being recorded as the reference allele in the references. Therefore, mRatBN7.2 reduced base-level errors by 9.2-fold compared with Rnor_6.0.

**Improved gene model annotations in mRatBN7.2**

mRatBN7.2-based gene annotations, based on transcriptomic datasets and newly revised multi-species sequence alignments, were released in RefSeq 108 and Ensembl 105. We compared the Rnor6.0 and mRatBN7.2 annotation sets in RefSeq using BUSCO v.4.1.4,\(^28\) focusing on the glires_odb10 dataset of 13,798 genes that are expected to occur in single copy in rodents. Instead of generating a de novo annotation with BUSCO using the Augustus or MetaEuK gene predictors, we used proteins from the new annotations, picking one longest protein per gene for analysis. BUSCO reported 98.7% of the glires_odb10 genes as complete in RefSeq 108 (single copy [S] 97.2%, duplicated [D] 1.5, fragmented [F] 0.4%, missing [M] 0.9%). In comparison, analysis of Rnor_6.0 with National Center for Biotechnology Information (NCBI)’s annotation pipeline using the same code and evidence sets showed a slightly higher fraction of fragmented and missing genes and more than double the rate of duplicated genes (S 94.4%, D 3.3%, F 0.9%, M 1.4%). The Ensembl annotation was evaluated using BUSCO v.5.3.2 with the lineage dataset “glires_odb10/2021-02-19.” The “complete and single-copy BUSCO” score improved from 93.6% to 95.3%, and the overall score improved from 96.5% in Rnor_6.0 to 97.0% in mRatBN7.2. Both annotation sets demonstrate that mRatBN7.2 is a better foundation for gene annotation, with improved representation of protein-coding genes.

RefSeq (release 108) annotated 42,167 genes, each associated with a unique NCBI GeneID. These included 22,228 protein-coding genes, 7,888 long non-coding RNA (lncRNA), 1,288 small nucleolar RNA (snoRNA), 1,026 small-nuclei RNA (snRNA), and 7,972 pseudogenes. Ensembl (release 107) annotation contained 30,559 genes identified with unique “ENSRSNOG” stable IDs. These included 23,096 protein-coding genes, 2,488 IncRNA, 1,706 snoRNA, 1,512 snRNA, and 762 pseudogenes. Although the two transcriptomes have similar numbers of protein-coding genes, RefSeq annotates many more lncRNA (more than 3 times) and pseudogenes (almost 10 times), and only RefSeq annotates tRNA.

Comparing individual genes across the two annotation sets, Ensembl BioMart reported 23,074 Ensembl genes with an NCBI GeneID, and NCBI Gene reported 24,115 RefSeq genes with an Ensembl ID (Table S3). We note that 2,319 protein-coding genes were annotated with different names (Table S4) by Ensembl and RefSeq. While many of these were caused by the lack of formal gene names in one of the sources, some of them were annotated with distinct names. For example, some widely studied genes, such as Bcl2, Cd4, Atrx, etc., were not annotated with GenelD in Ensembl BioMart. We found that 18,722 gene symbols were annotated in both RefSeq and Ensembl. Among the 22,247 gene symbols found only in RefSeq, 20,185 were genes without formal names. Further, RefSeq contained a total of 100,958 transcripts, with an average of 2.9 (range: 1–51) transcripts per gene. Ensembl had 54,991 transcripts, with an average of 1.8 (range: 1–10) transcripts per gene (Figure S7).
mRatBN7.2 improved the mapping of WGS data

We compared mapping results of long-read datasets. Unmapped read numbers declined from 298,422 in Rnor_6.0 to 260,947 in mRatBN7.2, a 12.6% reduction using a PacBio CLR dataset from an SHR rat with a total of 4,508,208 reads. Mapping of Nanopore data from one outbred HS rat (∼55× coverage) showed much lower secondary alignment—from ∼10 million in Rnor_6.0 to 4 million in mRatBN7.2. Similar to the linked-read
data, structural variants (SVs) detected in the Nanopore dataset were reduced (Figure S8) using mRatBN7.2.

We also evaluated the effect of reference genome on the identification of SVs. We identified 19,538 unique SVs against Rnor_6.0 in the HXB dataset. In contrast, only 5,458 SVs were found using mRatBN7.2 (see STAR Methods, key resources table, https://doi.org/10.5281/zenodo.10398554). We illustrated these findings by focusing on a small panel of eight samples (Figure S9).

mRatBN7.2 improved the analysis of transcriptomic and proteomic data

We analyzed an RNA sequencing (RNA-seq) dataset of 352 HS rat brains (see https://RatGTEx.org) to compare the effect of the reference genome. The fraction of reads aligned to the reference increased from 97.4% in Rnor_6.0 to 98.3% in mRatBN7.2. The average percentage of reads aligned concordantly only once to the reference genome increased from 89.3% to 94.6%. The average percentage of reads aligned to the Ensembl transcriptome increased from 67.7% to 74.8%. Likewise, we examined the alignment of RNA-seq data from ribosomal RNA-depleted total RNA and short RNA in the HRDP. For the total RNA (>200 bp) samples, alignment to the genome increased from 92.4% to 94.0%, while the percentage of reads aligned concordantly only once to the reference increased from 76.1% to 79.2%. For the short RNA (<200 bp; targeting transcripts 20–50 bp long), genome alignment increased from 95.0% to 96.2% and unique alignment increased from 33.2% to 35.9%. In an snRNA-seq dataset (Figures S10A–S10C), the percentage of reads that mapped confidently to the reference increased from 87.4% on Rnor_6.0 to 91.4% on mRatBN7.2. In contrast, reads mapped with high quality to intergenic regions were reduced from 24.5% on Rnor_6.0 to 10.3% on mRatBN7.2.

We analyzed datasets containing information about transcript start and polyadenylation in a capped short RNA-seq dataset. The rate of unique alignment of TSSs to the reference genome increased by 5% in mRatBN7.2 (Figures S10D and S10E). In this dataset, we identified 42,420 TSSs when using Rnor_6.0 and 44,985 sites when using mRatBN7.2 (see STAR Methods, key resources table, https://doi.org/10.5281/zenodo.10398387). We analyzed 83 whole-transcriptome termini site sequencing datasets using total RNA derived from rat brains and found that 76.97% were mapped against Rnor_6.0, while 80.49% were mapped against mRatBN7.2 (Table S5). We identified 167,136

Figure 2. mRatBN7.2 improved mapping statistics of whole-genome sequencing data

Summary statistics from mapping 36 HXB/BXH WGS samples against Rnor_6.0 and mRatBN7.2 were compared. Using mRatBN7.2 increased the percentage of reads mapped (A), reduced regions on the reference genome with zero coverage (B), total number of SNPs (C), and indels (D). The presence of a large number of SNPs (E) and indels (F) that are shared by all samples (arrows), including BN/NHsdMcwi, indicates that they are base-level errors in the reference genome.
and quantified 8,002 unique proteins on Rnor_6.0 compared to a tide-identification false discovery rate (FDR) of 1%, we identified we analyzed a set of data from 29 HXB/BXH strains. At a pep-

tions, this demonstrates a clear improvement in the accuracy of eQTLs when using mRatBN7.2. Given the much lower difference-eQTL SNP-gene pairs. In contrast, only 0.01% (51 of 404,302) of the Rnor_6.0 trans-eQTL SNP-gene pairs became trans-eQTLs when using mRatBN7.2. When we analyzed a WGS dataset of 163 inbred rats containing 12 BN samples and 151 non-BN samples, which represented two RI panels and more than 30 inbred strains. After read mapping to mRatBN7.2 and variant calling, we used the anomalies in the genotype data to further assess the quality of mRatBN7.2. This analysis was performed at two levels to detect issues at both the segment level and the individual nucleotide level.

We examined the impact of the upgraded reference assembly on the yield and relative numbers of expression QTL (eQTL) using a large RNA-seq dataset for nucleus accumbens.30 We identified seven genes associated with one or more strong trans-eQTLs using Rnor_6.0 that converted to cis-eQTLs using mRatBN7.2 (Figure 3). This constitutes 5.2% (3,261 of 63,148) of the Rnor_6.0 trans-eQTL SNP-gene pairs. In contrast, only 0.01% (51 of 404,302) of the Rnor_6.0 cis-eQTL SNP-gene pairs became trans-eQTLs when using mRatBN7.2. When we analyzed a set of data from 29 HXB/BXH strains. At a peptide-identification false discovery rate (FDR) of 1%, we identified and quantified 8,002 unique proteins on Rnor_6.0 compared with 8,406 unique proteins on mRatBN7.2 (5% increase). For protein local expression quantitative trait locus (i.e., cis-pQTL), 536 were identified using Rnor_6.0 and 541 were identified using mRatBN7.2 at FDR < 5%. Distances between pQTL peaks and the corresponding gene start site tended to be shorter on mRatBN7.2 than on Rnor_6.0 (Figure 3C). Similar to eQTLs, four proteins with trans-pQTL in Rnor_6.0 were converted to cis-pQTL using mRatBN7.2. For example, RPA1 protein (Chr10:60,148,794–60,199,949 bp in mRatBN7.2) mapped as a significant trans-pQTL (p = 1.71 × 10^{-13}) on Chr 4 in Rnor_6.0 but as a significant cis-pQTL using mRatBN7.2 (p = 4.12 × 10^{-3}) (Figure 3E). In addition, the expression of RPA1 protein displayed a high correlation between mRatBN7.2 and in Rnor_6.0 (r = 0.79; p = 3.7 × 10^{-7}) (Figure 3D). Lastly, the annotations for RPA1 were different between the two references (Figure 3F).

WGS mapping data suggesting potential errors remaining in mRatBN7.2

We analyzed a WGS dataset of 163 inbred rats containing 12 BN samples and 151 non-BN samples, which represented two RI panels and more than 30 inbred strains. After read mapping to mRatBN7.2 and variant calling, we used the anomalies in the genotype data to further assess the quality of mRatBN7.2. This analysis was performed at two levels to detect issues at both the segment level and the individual nucleotide level.

First, we analyzed segment-wise distribution patterns of heterozygous genotypes and no-calls. Since inbred lines are expected with 8,406 unique proteins on mRatBN7.2 (5% increase). For protein local expression quantitative trait locus (i.e., cis-pQTL), 536 were identified using Rnor_6.0 and 541 were identified using mRatBN7.2 at FDR < 5%. Distances between pQTL peaks and the corresponding gene start site tended to be shorter on mRatBN7.2 than on Rnor_6.0 (Figure 3C). Similar to eQTLs, four proteins with trans-pQTL in Rnor_6.0 were converted to cis-pQTL using mRatBN7.2. For example, RPA1 protein (Chr10:60,148,794–60,199,949 bp in mRatBN7.2) mapped as a significant trans-pQTL (p = 1.71 × 10^{-13}) on Chr 4 in Rnor_6.0 but as a significant cis-pQTL using mRatBN7.2 (p = 4.12 × 10^{-3}) (Figure 3E). In addition, the expression of RPA1 protein displayed a high correlation between mRatBN7.2 and in Rnor_6.0 (r = 0.79; p = 3.7 × 10^{-7}) (Figure 3D). Lastly, the annotations for RPA1 were different between the two references (Figure 3F).

WGS mapping data suggesting potential errors remaining in mRatBN7.2

We analyzed a WGS dataset of 163 inbred rats containing 12 BN samples and 151 non-BN samples, which represented two RI panels and more than 30 inbred strains. After read mapping to mRatBN7.2 and variant calling, we used the anomalies in the genotype data to further assess the quality of mRatBN7.2. This analysis was performed at two levels to detect issues at both the segment level and the individual nucleotide level.

First, we analyzed segment-wise distribution patterns of heterozygous genotypes and no-calls. Since inbred lines are expected with 8,406 unique proteins on mRatBN7.2 (5% increase). For protein local expression quantitative trait locus (i.e., cis-pQTL), 536 were identified using Rnor_6.0 and 541 were identified using mRatBN7.2 at FDR < 5%. Distances between pQTL peaks and the corresponding gene start site tended to be shorter on mRatBN7.2 than on Rnor_6.0 (Figure 3C). Similar to eQTLs, four proteins with trans-pQTL in Rnor_6.0 were converted to cis-pQTL using mRatBN7.2. For example, RPA1 protein (Chr10:60,148,794–60,199,949 bp in mRatBN7.2) mapped as a significant trans-pQTL (p = 1.71 × 10^{-13}) on Chr 4 in Rnor_6.0 but as a significant cis-pQTL using mRatBN7.2 (p = 4.12 × 10^{-3}) (Figure 3E). In addition, the expression of RPA1 protein displayed a high correlation between mRatBN7.2 and in Rnor_6.0 (r = 0.79; p = 3.7 × 10^{-7}) (Figure 3D). Lastly, the annotations for RPA1 were different between the two references (Figure 3F).
Furthermore, regions with high heterozygosity were genome (Table S6). This rate is much reduced from that of all, we identified 673 such ‘‘flagged regions,’’ with an average density of heterozygous or multiple-allelic sites. We focused on as many reads will map to this region and will produce a high ‘‘folded’’ into a single segment in the reference assembly, twice doubled of read depth for the heterozygous variants strongly sug-

ted, and heterozygous variants were often clustered in a region of these variants: homozygous SNPs were more evenly distrib-

tuted at 1 kb intervals. Of the 2.78 M simulated variants, 92.1% were successfully lifted. Thus, ~8% of the Rnor_6.0 genome does not have a unique match in BN7.2. These ‘‘unliftable’’ sites do not distribute evenly along Rnor_6.0 (Figure S13); rather, they tend to aggregate in regions with no credible match, i.e., lost in BN7.2 (Figure S14).

The rate of ‘‘Liftover loss’’ varies by the type of genomic feature, and as such will be study dependent. We used WGS data for one sample to call variants on Rnor_6.0, then lifted them to mRatBN7.2 and compared the results against the variants called directly on mRatBN7.2. Variants called on mRatBN7.2 had a higher proportion (87.99%) of passing the quality filter than those called on Rnor_6.0 (83.64%), and 97.93% of the variants called on Rnor_6.0 were liftable to mRatBN7.2 (Figure 4A). Among the lifted variant sites, 94.48% had a match in the variant set called directly. However, 11.91% of the variants called on mRatBN7.2 were absent in the lifted variant set (Figure 4B). Thus, complete remapping of the data to the new reference is preferable despite its time and resource costs.

A comprehensive survey of the genomic landscape of Rattus norvegicus based on mRatBN7.2

To facilitate a smooth transition to mRatBN7.2, we conducted a joint analysis of WGS data. We collected WGS for a panel of 163 rats (hereafter referred to as RatCollection, Table S8). RatCollection includes all 30 HXB/BXH strains, 25 FXLE/LEXF strains, and 33 other inbred strains. In total, we covered 88 strains (120 samples at the strain level)—approximately 80% of the HRDP.

The mean depth of coverage of these samples was 60.4 ± 39.3. Additional sample statistics are provided in Table S9. After removing variants that were potential errors in mRatBN7.2 (above) and filtering for site quality (qual ≥ 30), we identified 19,987,273 variants (12,661,110 SNPs, 3,805,780 insertions, and 3,514,345 deletions) across 15,804,627 variant sites (Table 2). Across the genome, 89.4% of the sites were bi-allelic, and the mean variant density was 5.96 ± 2.20/kb (mean ± SD). The highest variant density of 30.5/kb was found on Chr 4 at 98 Mb (Figure S15). Most (97.9% ± 1.4%) of the variants were
Table 2. Genetic variants in laboratory populations

<table>
<thead>
<tr>
<th>Variants sites</th>
<th>Variants</th>
<th>SNPs</th>
<th>Indels</th>
<th>Mixed</th>
<th>Predicted impact</th>
<th>Alt/Alt homozygous % (mean ± SD)</th>
<th>Genotype qual (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RatCollection</td>
<td>15,804,627</td>
<td>19,987,273</td>
<td>12,661,110</td>
<td>7,313,702</td>
<td>12,461</td>
<td>18,646</td>
<td>262,685</td>
</tr>
<tr>
<td>HS progenitors</td>
<td>12,418,243</td>
<td>16,438,302</td>
<td>9,947,112</td>
<td>6,479,485</td>
<td>11,705</td>
<td>6,980</td>
<td>205,316</td>
</tr>
<tr>
<td>FXLE/LEXFb</td>
<td>9,183,562</td>
<td>13,070,345</td>
<td>7,036,182</td>
<td>6,023,391</td>
<td>9,440</td>
<td>10,121</td>
<td>143,514</td>
</tr>
<tr>
<td>HXB/BXH</td>
<td>7,520,223</td>
<td>11,256,227</td>
<td>5,705,592</td>
<td>5,541,195</td>
<td>9,440</td>
<td>13,988</td>
<td>143,514</td>
</tr>
<tr>
<td>SS/WR</td>
<td>7,171,447</td>
<td>10,522,763</td>
<td>5,444,220</td>
<td>4,999,398</td>
<td>9,145</td>
<td>8,606</td>
<td>111,658</td>
</tr>
<tr>
<td>LL/LN/LH</td>
<td>6,923,575</td>
<td>10,722,255</td>
<td>5,433,556</td>
<td>4,670,291</td>
<td>8,908</td>
<td>4,142</td>
<td>111,040</td>
</tr>
</tbody>
</table>

The RatCollection includes 163 rats (88 strains and 32 substrains, with some biological replicates). The HRDP contains the HXB/BXH and FXLE/LEXF panels as well as 30 or so classic inbreds. Our analysis includes ~80% of the HRDP. The variants were jointly called using Deepvariant and GLNexus. Variant impact was annotated using Snpeff.

Homozygous at the sample level, confirming the inbred nature of most strains, with a few exceptions (Figure S16).

To analyze the phylogenetic relationships of these 120 strains/substrains, we created an identity-by-state (IBS) matrix using 11,585,238 high-quality bi-allelic SNPs (Table S10). Distance-based phylogenetic trees of all strains and substrains are shown in Figure 5A. The mean IBS for classic inbred strains was 72.30% ± 2.35%, where BN has the smallest IBS (64.76%).

This phylogenetic tree included all major populations of laboratory rats, such as the eight inbred progenitor strains of the outbred HS rats (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N, and WN/N). The number of sites with a non-reference allele in each of these strains is shown in Figure 5B; WKY/N contributed the largest number of non-reference alleles (Figure 5C). The distribution of the variant sites from all eight strains across the chromosomes is shown in Figure 5D. Collectively, these 8 strains accounted for 78.6% of all non-reference alleles in the RatCollection. Conversely, 141,556 of these variant sites were not found in any other strains. Although none of these founder strains are alive today, based on IBS, we identified 6 unique to SR and 920,234 variants unique to SS. These strain-specific variants were found throughout the genome (Figure S18). A similar pattern was found for the Lyon hypertensive (LH), hypertensive (LL), and normotensive (LN) rats (Table 2) selected from outbred Sprague-Dawley rats for studying blood pressure regulation. Only 281,972, 289,112, and 262,574 variants were unique to LH, LL, and LN, respectively. In agreement with a prior report, these variants were clustered in a handful of genomic hotspots (Figure S19).

Impact of rat variants on genetic studies of human diseases

We used SnpEff (v.5.0e) to predict the impact of the 19,987,273 variants in the RatCollection based on RefSeq annotation. Among these, 18,646 variants near coding genes were predicted to have a high impact (i.e., causing protein truncation, loss of function, or triggering nonsense-mediated decay, etc.) on 6,667 genes, including 3,930 protein-coding genes (Table S12). Among the predicted high-impact variants, annotation by RGD disease ontology identified 2,601 variants affecting 2,079 genes that were associated with 3,261 distinct disease terms. Cancer/tumor (878), psychiatric (612), intellectual disability (505), epilepsy (319), and cardiovascular (304) comprised the top 5 disease terms, with many genes associated with more than one disease term. A mosaic representation of the number of high-impact variants per disease term for each strain is shown in (Figure S20). The top disease categories for the two RI panels are comparable, including cancer, psychiatric disorders, and cardiovascular disease. The disease ontology annotations for...
Figure 5. Phylogenetic relationship and genetic diversity of 120 strain/substrains of laboratory rats

(A) The phylogenetic tree was constructed using 11.6 million biallelic SNPs from 163 samples. Strains/substrains with duplicated samples were condensed. Strains highlighted with bold fonts are parental strains for RI panels. Green, HXB/BXH RI panel; blue, FXLE/LEXF RI panel; orange, progenitors of the HS outbred population.

(B) The HS progenitors contain 16,438,302 variants (i.e., 82.2% of the variants in our collection of 120 strain/substrains) based on analysis using mRatBN7.2. Among these, 10,895 are shared by all eight progenitor strains. The number of variants that are unique to each specific founder is noted.

(C) The total number of variants per strain, with the total number unique to each strain marked.

(D) The number of variants shared across strains per chromosome.
variants in the HXB/BXH, FXLE/LEXF RI panels, and SS/SR, as well as LL/LN/LH selective bred strains, are summarized in Figures S21 and S22 and Tables S13 and S14.

We further annotated genes using the human genome-wide association studies (GWASs) catalog. Among the rat genes with high-impact variants, 2,034 have human orthologs with genome-wide significant hits associated with 1,393 mapped traits (Table S15). The most frequent variant type among these rat genes was frameshift (1,557 genes), followed by splice donor variant (136 genes) and gain of stop codon (116 genes). Although rats and humans do not share the same variants, strain with these variants can potentially be a useful model for related human diseases at the gene level.

**DISCUSSION**

We systematically evaluated mRatBN7.2 and confirmed that it improved assembly quality over its predecessor, Rnor_6.0. As a result, mRatBN7.2 improved the analysis of omics datasets and now enables powerful and efficient genome-to-phenome analysis of 20 million variants segregating in the laboratory rat.

The improvements in mRatBN7.2 are based on new assembly methods and long-read data. However, the PacBio CLR reads used in mRatBN7.2 have lower base accuracy than Illumina short reads. Although Illumina data were used to polish the assembly, polishing methods do not correct all base-level errors. Our joint analysis of 163 WGS datasets identified 129,186 sites with likely nucleotide errors in mRatBN7.2. Another source of potential error is the assembly method itself. mRatBN7.2 was assembled with VGP pipeline v.1.6, which assembles the long reads into a diploid genome containing a primary and an alternative assembly. This pipeline is well suited to assemble diploid genomes. When applied to BN/NHsdMcwi, a fully inbred laboratory rat, the pipeline classified some duplications with small variants as two haploids, resulting in a collapsed repeat.

A telomere-to-telomere assembly is needed to fully resolve these issues. Liftover enables direct translation of genomic coordinates between different references, thereby reducing the cost of transitioning to a new reference. We found that 92.05% of simulated variants and 97.93% of variants from a WKY/N sample identified on Rnor_6.0 were lifted successfully to mRatBN7.2. This difference is attributable to the large number of simulated variants located in regions of low complexity. While promising, we found that reanalysis of the original sequencing data using mRatBN7.2 discovered an additional 507,700 variant sites, located primarily in regions not present in Rnor_6.0. Thus, remapping to mRatBN7.2 is preferred over using Liftover even though the latter is more convenient.

To facilitate the transition for studies of specific rat models, we mapped WGS data of 88 strains (120 substrains and 163 samples) to mRatBN7.2. Joint analysis identified 20.0 million variants from 15.8 million sites. This analysis expanded many prior analyses of rat genomes. For example, Baud et al. analyzed 8 strains against Rnor3.4 and found 7.9 million variants. Atanur et al. analyzed 27 rat strains against Rnor3.4 and reported 13.1 million variants, and Hermsen et al. analyzed 40 strains against Rnor5.0 and reported 12.2 million variants. Most recently, Ramdas et al. analyzed 8 strains and reported 16.4 million variants. In addition to using the latest reference genome and an expanded number of strains, our pipeline depends on Deepvariant and GLNexus, which have been shown to improve call set quality, especially on indels. Thus, our data provide the most comprehensive analysis of genetic variants in the laboratory rat population to date and, of comparative interest, is twice the variant count of the highly used mouse BXD family.

This collection of rats represented a wide variety of rats that are used in genetics studies today, such as the full HXB/BXH panel and 27 strains/substrains of the FXLE/LEXF RI panel. Together with the inbred strains, they cover about 80% of the rats in the HRDP. The new variant set from our analysis provide a much-needed boost in mapping precision.

The outbred N/NIH HS rats are another widely used rat genetic mapping population. Our analysis is in agreement with Ramdas et al. in identifying 16 million variants in the progenitors. Updated genetic data will be useful in HS genetic mapping studies, such as for the imputation of variants. Although live colonies of the original NIH HS progenitor strains are no longer available, we identified six living substrains that are genetically almost identical to the original progenitors (99.5% IBS), while the closest match to the other two strains has IBS of approximately 70% (Table S11). These living substrains will be useful in many ways, such as studying the effect of variants identified in genetic mapping studies. Our analysis also included groups of inbred rats segregated by less than 2 million variants. For example, the LH/LN/LL family, the SS/SR family (Table S14), as well as several pairs of near isogenic lines with distinctive phenotypes. These rats can be exploited to identify causal variants using reduced complexity crosses.

Updated genotyping data for these populations will benefit all ongoing studies that use these rats.

A trove of genetic associations has been established using laboratory rats. For example, studies using RI panels have identified loci that affect cardiovascular disease, hypertension, diabetes, immunity, tumorigenesis, and tissue-specific gene expression profiles. Using the outbred HS rats, Baud et al. reported 355 QTLs associated with phenotypes representing six diseases. Other studies using the HS population have revealed genetic control of behavior, obesity, and metabolic phenotypes. Many of these data are available from the RGD. While all these data were analyzed using previous versions of the reference genome, reanalyzing them using updated genomic data could lead to novel discoveries.

By generating F1 hybrids from these inbred lines with sequenced genomes, novel phenotypes can be mapped onto completely defined genomes: the 82 sequenced HRDP strains can produce any of 6,642 isogenic but entirely replicable F1 hybrids with completely defined genomes. Studies of these “sequenced” F1 hybrids avoid the homozygosity of the parental HRDP strains and enable a new phase of genome-phenome mapping and prediction.

While several genetic mapping studies using the HRDP are currently underway, the large number of variants in the HRDP (cf., the hybrid mouse diversity panel contains about 4 million SNPs and WGS-based genotype data will further encourage the use of this resource.

We identified 18,646 variations likely to have a high impact on 6,667 genes and many more variants predicted to have regulatory
effects on gene expression (Table S13). Some of these variants are supported by the literature. For example, the SS rats develop renal lesions with hypertension and have high impact mutations in Capn1\(^6\) and Procr\(^7\), both associated with kidney injuries, as well as in Klk1c12, associated with hypertension.\(^6\) These annotations identified many genes with variants that either result in a gain of STOP codon or cause a frameshift, which could lead to the loss of function. Strains harboring these variants can be explored to investigate the function of these genes.\(^6\)

These functional annotations are highly relevant when the human ortholog of these genes is associated with certain traits in human GWASs (Table S15). For example, CDHR3 is associated with smoking cessation.\(^6\) A gain of STOP mutation in the Cdh3r gene was found in only one parent of both RI panels (SHR/OlaIpcv and LE/Stm). Using these RI panels to study the reinstatement of nicotine self-administration, a model for smoking cessation, will likely provide insights into the role of CDHR3 in this behavior. Furthermore, a near-complete catalog of strain-specific alleles and functional prediction provide a stable source of potential candidate variants for interpreting genetic mapping results.

Our phylogenetic analysis agrees with those published previously\(^6\),\(^6\) and confirmed that BN/NHsdMcwi is an outgroup to all other common laboratory rats (Figure 5A). This is consistent with its derivation from a pen-bred colony of wild-caught rats\(^3\) and previous studies that included wild rats.\(^6\) Thus, mapping sequence data from other strains to the BN reference yields a greater number of variants (but at slightly reduced mapping quality) than using a hypothetical reference that is genetically closer to the commonly used laboratory strains. This so-called reference bias has been observed in genomic\(^6\) and transcriptomic data analyses.\(^6\) It should be noted that no individual strain is a perfect representation of a population. Instead, the nascent field of pangenomics,\(^6\) where the genomes of all strains can be directly compared with each other, provides a promising future in which all variants can be compared between individuals directly without the use of a single reference genome. While this pangenomic approach can be applied to short-read data,\(^6\) it will be especially powerful when individual genomes are all assembled from long-read sequence data, some of which are already available.\(^1\) This approach will enable a complete catalog of all genomic variants, including SVs and repeats, that differ between individuals.

Additional rat genomic resources we generated include a new rat genetic map with 150,835 markers. We envision that this map will have multiple applications. For example, this map can be integrated into the de novo assembly process, as demonstrated by the new assembly of the stickleback genome.\(^7\),\(^8\) Unlike human and mouse genomes,\(^4\) functional elements in the rat remain poorly annotated. Our analysis produced a list of TSS locations for genes expressed in the prefrontal cortex or nucleus accumbens and APA sites of genes expressed in the brain. These new datasets will further the study of gene regulatory mechanisms in rats.

Research using the Rattus norvegicus has made important contributions to the understanding of human physiology and diseases. An updated reference genome provides not only a valuable resource for future studies but also opportunities for analyzing existing data to gain new insights. The rich literature on laboratory rats, combined with the complex genomic landscape revealed in our survey, demonstrates that the rat is an excellent model organism for the next chapter of biomedical research.

**Limitations of the study**
While our comparative analyses reported many improvements in mRatBN7.2 over Rnor_6.0, we also found that mRatBN7.2 contains a few hundred segments with potential disassembly, as well as base-level errors. These segments affect gene annotation and the accuracy and completeness of Liftover. While we are confident in the informatic analysis of these segments, we did not empirically validate these findings. Future iterations of the rat reference genome need to reach telomere-to-telomere completion and eventually be expanded to a multi-strain pan-genome reference.

**STAR+METHODS**
Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
- **METHOD DETAILS**
  - Calculating genome assembly statistics
  - Analysis of WGS data
  - Sample quality control
  - Identification of genomic regions with potential mis-assembly in mRatBN7.2
  - Evaluating Liftover
  - Identify live strains that are close to HS progenitors
  - Constructing a genetic map using genetic data from a large HS cohort
  - Phylogenetic tree
  - RNA-seq data
  - eQTL relocation analysis
  - Capped small (cs)RNA-seq data
  - Single nucleid (sn) RNA-seq data
  - Transcriptome termini site sequencing
  - Brain proteome data
  - Identifying potentially mislabeled samples
  - Ensembl annotation
  - RefSeq annotation

**SUPPLEMENTAL INFORMATION**
Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100527.

**ACKNOWLEDGMENTS**
This work is supported by the Academy of Finland (grant no. 343656) to P.R.; NIH NHLBI R01HL064541 and P01HL149620 and Office of the Director.
R24OD024617 to M.T.; NIH NIDA U01DA043098 to H.A.; NIH NIDA U01DA051972 to C.B.; NIH NHLBI R01HL064541 and Office of the Director R24OD024617 to W.M.D.; NIH NHLBI P01HL149620 and Office of the Director R24OD024617 to A.M.G.; NIH Office of the Director grant R24OD024617 to M.R.D.; NIH NHLBI R01DA051972 to R.W.W.; NIH NHLBI R01DA047368 and P03DA044223 to T.V.d.J.; NIH NHLBI R01DA043098 to J.R.S.; NIH NHLBI P50DA037844 to L.C.S.W. (HS rats); NIH NIAAA R24AA013162 to B.T.; NIH NHLBI U01DA050239 and U01DA051972 to F.T.; NIH NHLBI P30DA044223 to L.C.S.; the National Center for Biotechnology Information (NCBI) National Library of Medicine (NLM), National Institutes of Health to T.V.d.J.; NIH NHLBI R01HL064541 and P01HL149620, NIHGR R24OD0410859, and Office of the Director R24OD024617 to A.E.K.; NIH Office of the Director-grant R24OD024617 to M.R.D. (HRPD); NIH NHLBI U01DA047368 and P03DA044223 to R.W.W.; NIH U01DA043098 to J.Z.L.; and NIH U01DA047638, P05DA063844, and R01DA40817 to H.C. The majority of the computation for this work was performed on the University of California for Scientific Applications and Advanced Computing (ISAC) computational resources.

This work is dedicated to the memory of Dr. Mary Shimoyama.

AUTHOR CONTRIBUTIONS

DECLARATION OF INTERESTS
The authors declare no competing interests.

Received: October 2, 2023
Revised: December 26, 2023
Accepted: February 29, 2024
Published: March 26, 2024

REFERENCES


66. Mashimo, T., Voigt, B., Tsurumi, T., Naoki, K., Nakashima, S., Yasumasaki, K., Li, K., Kuramoto, T., and Serikawa, T. (2006). A set of highly informative rat SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3; Fly 6, 80–92.


## KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposited data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural variants in HXB RI panel (VCF)</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10398554">https://doi.org/10.5281/zenodo.10398554</a></td>
</tr>
<tr>
<td>Join calling of 163 WGS rat samples (VCF)</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10398344">https://doi.org/10.5281/zenodo.10398344</a></td>
</tr>
<tr>
<td>Transcription start site for 2 brain regions</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10398387">https://doi.org/10.5281/zenodo.10398387</a></td>
</tr>
<tr>
<td>Alternative polyadenylation sites</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10398476">https://doi.org/10.5281/zenodo.10398476</a></td>
</tr>
<tr>
<td>Potential misassembled regions</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10428255">https://doi.org/10.5281/zenodo.10428255</a></td>
</tr>
<tr>
<td>Chromosomal dot plots between Rnor_6.0 and mRatBN7.2</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10515796">https://doi.org/10.5281/zenodo.10515796</a></td>
</tr>
<tr>
<td>The order of genetic markers and the distances from a rat genetic map compared to their locations in Rnor_6.0</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10552387">https://doi.org/10.5281/zenodo.10552387</a></td>
</tr>
<tr>
<td>The order of genetic markers and the distances from a rat genetic map compared to their locations in mRatBN7.2</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10552453">https://doi.org/10.5281/zenodo.10552453</a></td>
</tr>
<tr>
<td>Code and data for phylogenetic analysis nucleus accumbens core eQTL dataset</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10520242">https://doi.org/10.5281/zenodo.10520242</a></td>
</tr>
<tr>
<td>csRNA-seq data</td>
<td>Munro et al.</td>
<td>NAcc, 75 samples</td>
</tr>
<tr>
<td>Annotation of the mRatBN7.2 assembly</td>
<td>O’Leary et al.</td>
<td>Release 108</td>
</tr>
<tr>
<td>Code and computational notebook</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10552887">https://doi.org/10.5281/zenodo.10552887</a></td>
</tr>
<tr>
<td>Experimental models: Organisms/strains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample metadata for RatCollection</td>
<td>This paper</td>
<td><a href="https://doi.org/10.5281/zenodo.10552790">https://doi.org/10.5281/zenodo.10552790</a></td>
</tr>
<tr>
<td>Software and algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QUAST</td>
<td>Mikheenko et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>paftools</td>
<td>Kalikar et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>BWA mem</td>
<td>Li and Durbin</td>
<td>N/A</td>
</tr>
<tr>
<td>GATK</td>
<td>Poplin et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>LongRanger</td>
<td>Li</td>
<td>N/A</td>
</tr>
<tr>
<td>Minimap2</td>
<td>Li</td>
<td>N/A</td>
</tr>
<tr>
<td>Deepvariant</td>
<td>Poplin et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>GLNexus</td>
<td>Yun et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>SURVIVOR</td>
<td>Jeffares et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>SnpEff</td>
<td>Cingolani et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>RepeatMasker</td>
<td>Tarailo-Graovac and Chen</td>
<td>N/A</td>
</tr>
<tr>
<td>GeneCup</td>
<td>Gunturkun et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>UCSC Liftover tool</td>
<td>Hinrichs et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>PLINK</td>
<td>Purcell et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>Lep-MAP3</td>
<td>Rastas</td>
<td>N/A</td>
</tr>
<tr>
<td>bcftools</td>
<td>Danecek et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>MEGA</td>
<td>Tamura et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>gttree</td>
<td>Yu et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>STAR aligner</td>
<td>Dobin et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>HOMER’s findPeaks tool</td>
<td>Duttke et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>QTLtools</td>
<td>Delaneau et al.</td>
<td>N/A</td>
</tr>
<tr>
<td>UniProt</td>
<td>UniProt Consortium</td>
<td>N/A</td>
</tr>
</tbody>
</table>
RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Hao Chen (hchen@uthsc.edu).

Materials availability
The Hybrid Rat Diversity Panel (HRDP) consists of 96 inbred strains, and includes 30 classic inbred strains and both of the large RI families discussed in the prior sections (BXH/HXB and LEXF/FXLE). The panel is being cryo-resuscitated and cryopreserved at the Medical College of Wisconsin for use by the scientific community.

Data and code availability
- The RatCollection contains 163 WGS samples from 88 strains and 32 substrains. It includes new data from 128 rats (key resources table) and 36 datasets downloaded from NIH SRA. The detailed sample metadata are provided in the Table S8. A total of 82 members of the HRDP have been sequenced using Illumina short-read technology, including all 30 extant HXB strains, 23 of 27 FXLE/LEXF strains, and 25 of 30 classic inbred strains. RatCollection includes all 30 strains of the HXB/BXH family, 27 strains in the FXLE/LEXF family, and 33 other inbred strains. In total, we covered 88 strains and 32 substrains. It contains approximately 80% of the HRDP. WGS data generated for this work have been uploaded to NIH SRA (see Table S8 for SRA IDs). Other resources are available from Zenodo (See key resources table). The code for the custom R, Python and Bash scripts for data analysis is available from a github repository: https://github.com/hanyoupan/rat-manuscript (https://doi.org/10.5281/zenodo.10552887)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All rats used for whole genome sequencing are naive adult males. Their strain and substrain identification as well as RRIDs are provided in the Table S8.

METHOD DETAILS

Calculating genome assembly statistics
We obtained all assemblies from UCSC Goldenpath, with the exception of CHM13_T2T_v1.1, which was downloaded from the T2T Consortium GitHub page. We used QUAST25 to calculate common assembly metrics, such as contig and scaffold N50, using a consistent standard across all assemblies. We defined each entry in the fasta file as a scaffold, breaking them into contigs based on continuous Ns of 10 or more. No scaffolds below a certain length were excluded from the analysis.23 Our scripts and intermediate results can be found in the supplementary. Nx plots were generated using a custom Python script and the fasta files as inputs. Structural differences between Rnor_6.0 and mRatBN7.2 were evaluated using paftools.77

Analysis of WGS data
We used different methods for mapping data, based on the sequencing technologies. For illumina short read data, fastq files were mapped to the reference genome (either Rnor_6.0 or mRatBN7.2) using BWA mem.16 GATK79 was then used to mark PCR duplicates in the bam files. For 10x chromium linked reads, fastq files were mapped against the reference genomes using LongRanger (version 2.2.2). Long read data were mapped using Minimap2.80 Deepvariant26 (ver 1.0.0) was then used to call variants for each sample. Joint calling of variants for all the samples was conducted using GLNexus.27 Large SVs detected by LongRanger were merged using SURVIVOR81 per reference genome. SVs detected in less than two samples were removed. Variants with the QUAL score less than 30 were removed. The impact of variants were predicted using SnpEff,82 using RefSeq annotations. Overlap between SNPs and repetitive regions were identified with RepeatMasker.83 Disease ontology was retrieved from the Rat Genome Database.16 Disease associations were related to nearest genes as predicted by SnpEff. Subsequent analyses were conducted using custom scripts in R or bash. Circular plots were generated with the Circos plots package in R. Functional consequences of variants on genes were searched in PubMed via GeneCup.84

Sample quality control
To ensure the quality of data from 168 rats, we conducted a thorough examination of the missing call rate and read depth per sample. We determined that a per sample missing rate of 4% and average read depth of 10 were appropriate based on the data distribution. We identified and removed 4 samples with high missing rates (SHR/NcrlPrin_BT.ILM, WKY/NHsd_TA.ILM, GK/Ox_TA.ILM, FHL/EurMcwi_TA.ILM) and 2 samples with low read depth (WKY/NHsd_TA.ILM, BBDP/Wor_TA.ILM), resulting in a total of 5 samples removed.
Identification of genomic regions with potential mis-assembly in mRatBN7.2

We used WGS-derived genotype data for 163 samples to detect genomic regions with unusually high densities of heterozygous genotypes (i.e., the “high-het” regions). Since the samples are from inbred animals, high-het regions could arise from tandemly repeated segments in the BN genome “folded” into a single region in mRatBN7.2. While read-depth data represent an independent source of information, here we focused on the distribution patterns of heterogeneous genotypes in the 12 BN animals. Indeed, along the genome, the per-site heterozygote counts, ranging 0 to 12 across the 12 BN samples, tend to be high in certain discrete regions, which also tend to have high counts of NA (the “no-calls”). We used the het+NA counts as the scanning statistics for segment-wise switches between a high-state and a low-state, akin to using array-CGH intensity data or sequencing read depths to detect DNA copy-number variants (CNVs). Specifically, we added 2 to the per-site het+NA counts, to mimic the situation with 2 DNA copies in baseline regions and up to 14 copies in the high-regions. The logged values, ranging from log2(2) to log2(14), are “segmented” by using the segment command in R package “DNAncopy,” with parameters min.width = 5, alpha = 0.001. The results tend to be over-segmented, and they are merged by a custom R script with the following merging rules. Starting from the first three segments, we merge the first and the second segments if one of the four conditions are met: (1) both have mean values above 1.5, which is an empirically derived threshold in our log2(x+2) scale, (2) both have mean values below 1.5, (3) the three segments are high-low-high but the middle segment is less than 5 kb, hence a “positive flicker,” or (4) the three segments are low-high-low with the middle segment shorter than 5 kb: a “negative flicker.” If the first two segments are merged, the next “triplet” to be evaluated are the merged 1–2, the previous #3 segment, and the newly called-up $4 segment. If no merging is executed, the scan moves by one segment to evaluate the next “triplet”: #2–4. This operation was repeated until we reached the end of the chromosome. Altering the merging parameters will yield a somewhat different set of “flagged” regions. The current merged list contains 673 high-segments over the 20 autosomes, covering about 1.4% of mRatBN7.2 and having an average length of 52,199 nt. Future iterations will incorporate distribution patterns of read depth, multi-allelic sites and, if possible, linked read information. Similar scans can be performed for the 151 non-BN samples. Our preliminary data show that most of the regions flagged by the 12 BN samples also show high Het+NA counts in non-BN samples. In addition, some other heterozygous genotypes in non-BN samples fall outside the flagged regions, either individually or in new segments not flagged in BN samples. Many of them likely represent single-site or segment-wise differences between these strains and the reference strain: BN, and they may vary in a strain-specific fashion.

Evaluating Liftover

The Liftover tool is evaluated using both simulated and real data. The simulated data is an evenly 1000-base pair-spaced bed file covering Rnor_6.0 (2,782,023 sites within the bed file). The simulated dataset is used to study what portion of the genome is liftable, and the distribution of the unliftable and liftable sites. The real dataset is used to evaluate the accuracy and utility of the Liftover process, we compared the variants obtained through the Liftover process to those directly called from the data. We mapped WGS data from a WKY/N rat to both Rnor_6.0 and mRatBN7.2 and called variants following each respective genome. We then used the UCSC Lift tool85 and corresponding chain files to lift these variants to the other reference genome. The resulting lifted variant sets are then compared to the directly called variant sets.

Identify live strains that are close to HS progenitors

Currently, there are two colonies of the HS rats: one located at Wake Forest University (RRID: RGD_13673907), which was moved from the Medical College of Wisconsin (RRID: RGD_2314009); the University of California San Diego houses the other colony (RRID: RGD_15526810). While the original progenitor population is no longer alive, we obtained DNA samples that were preserved in 1984, from the Medical College of Wisconsin (RRID: RGD_2314009); the University of California San Diego houses the other colony (RID: RGD_15526810). We used WGS-derived genotype data for 163 samples to detect genomic regions with unusually high densities of heterozygous genotypes (i.e., the “high-het” regions). Since the samples are from inbred animals, high-het regions could arise from tandemly repeated segments in the BN genome “folded” into a single region in mRatBN7.2. While read-depth data represent an independent source of information, here we focused on the distribution patterns of heterogeneous genotypes in the 12 BN animals. Indeed, along the genome, the per-site heterozygote counts, ranging 0 to 12 across the 12 BN samples, tend to be high in certain discrete regions, which also tend to have high counts of NA (the “no-calls”). We used the het+NA counts as the scanning statistics for segment-wise switches between a high-state and a low-state, akin to using array-CGH intensity data or sequencing read depths to detect DNA copy-number variants (CNVs). Specifically, we added 2 to the per-site het+NA counts, to mimic the situation with 2 DNA copies in baseline regions and up to 14 copies in the high-regions. The logged values, ranging from log2(2) to log2(14), are “segmented” by using the segment command in R package “DNAncopy,” with parameters min.width = 5, alpha = 0.001. The results tend to be over-segmented, and they are merged by a custom R script with the following merging rules. Starting from the first three segments, we merge the first and the second segments if one of the four conditions are met: (1) both have mean values above 1.5, which is an empirically derived threshold in our log2(x+2) scale, (2) both have mean values below 1.5, (3) the three segments are high-low-high but the middle segment is less than 5 kb, hence a “positive flicker,” or (4) the three segments are low-high-low with the middle segment shorter than 5 kb: a “negative flicker.” If the first two segments are merged, the next “triplet” to be evaluated are the merged 1–2, the previous #3 segment, and the newly called-up $4 segment. If no merging is executed, the scan moves by one segment to evaluate the next “triplet”: #2–4. This operation was repeated until we reached the end of the chromosome. Altering the merging parameters will yield a somewhat different set of “flagged” regions. The current merged list contains 673 high-segments over the 20 autosomes, covering about 1.4% of mRatBN7.2 and having an average length of 52,199 nt. Future iterations will incorporate distribution patterns of read depth, multi-allelic sites and, if possible, linked read information. Similar scans can be performed for the 151 non-BN samples. Our preliminary data show that most of the regions flagged by the 12 BN samples also show high Het+NA counts in non-BN samples. In addition, some other heterozygous genotypes in non-BN samples fall outside the flagged regions, either individually or in new segments not flagged in BN samples. Many of them likely represent single-site or segment-wise differences between these strains and the reference strain: BN, and they may vary in a strain-specific fashion.

Constructing a genetic map using genetic data from a large HS cohort

The collection of genotypes from 1893 HS rats and 753 parents (378 families) was described previously.15 Briefly, genotypes were determined using genotyping-by-sequencing.93 This produced approximately 3.5 million SNP with an estimated error rate <1%. Variants for X- and Y chromosomes were not called. The genotype data used for this study can be accessed from the C-GORD database at https://doi.org/10.48810/P44W2 or through https://www.genenetwork.org. The genotype data were further cleaned to remove nonmonomorphic SNPs. Genotypes with high Mendelian inheritance error rates (>2% error across the cohort) were identified by PLINK85 and removed. We further used an unbiased selection procedure, which yielded a final list of 150,835 binned markers distributed across the genome. Mean distance between markers is 18.5 kb across the genome. We used Lep-MAP3 (LM3)87 to construct the genetic map. The following LM3 functions were used: 1) the ParentCall2 function was used to call parental genotypes by taking into account the genotype information of grandparents, parents, and offspring; 2) the Filtering2 function was used to remove those markers with segregation distortion or those with missing data using the default setting of LM3; and 3) the OrderMarkers2 function...
was used to compute cM distances (i.e., recombination rates) between all adjacent markers per chromosome. The resulting map had consistent marker order that supported the mRatBN7.2.

**Phylogenetic tree**

We used bcftools to filter for high-quality, bi-allelic SNP sites from the VCF files for the 20 autosomes. We then employed PLINK to calculate a pairwise identity-by-state (IBS) matrix using the 11.5 M variants. We imported the resulting matrix into R and converted it to a meg format as described in the MEGA manual. We then used MEGA to construct a distance-based UPGMA tree using the megfile as input. We used MEGA’s “flip subtree” function to adjust the position of a few internal nodes for improved visualization using the MEGA GUI. The modified tree was exported as a nwk file. We then imported this nwk file into R and used the ggtree package to plot the phylogenetic tree. There is a pdf file document the detailed steps of phylogenetics analysis available in key resources table “Code and data for phylogenetic analysis”.

**RNA-seq data**

RNA-seq data was downloaded from RatGTeX. Brains were extracted from 88 HS rats. Rats were housed under standard laboratory conditions. Rat brains were extracted and cryosectioned into 60 μm sections, which were mounted onto RNase-free glass slides. Slides were stored in −80°C until dissection and before RNA-extraction post dissection. AllPrep DNA/RNA mini kit (Qiagen) was used to extract RNA. RNA-seq was performed on mRNA from each brain region using Illumina HiSeq 4000 to obtain 100 bp single-end reads for 435 samples. RNA-Seq reads were aligned to the Rnor_6.0 and mRatBN7.2 genomes from Ensembl using STAR v2.7.8a.

**eQTL relocation analysis**

We obtained the nucleus accumbens core (NAcc, 75 samples) eQTL dataset from RatGTeX (https://ratgtx.org), which was mapped using Rnor_6.0. We considered associations with p < 1 × 10^-8 between any observed SNP and any gene. We labeled those for which the SNP was within 1Mb of the gene’s transcription start site as cis-eQTLs, and those with TSS distance greater than 5Mb, or with SNP and gene on different chromosomes, as trans-eQTLs. SNP-gene pairs with TSS distance 1–5Mb were not counted in either group. We estimated the set of cross-chromosome genome segment translocations between Rnor_6.0 and mRatBN7.2 using Minimap2 with the “asm5” setting. Examples of the relocations were visualized using the NCBI Comparative Genome Viewer (https://ncbi.nlm.nih.gov/genome/cgv/).

**Capped small (cs)RNA-seq data**

csRNA-seq data used for the alignment metrics were previously published. Briefly, small RNAs of ~15–60 nt were size selected by denaturing gel electrophoresis starting from total RNA extracted from 14 rat brain tissue dissections. For csRNA libraries, cap selection was followed by decapping, adapter ligation, and sequencing. For input libraries, 10% of small RNA input was used for decapping, adapter ligation, and sequencing. After library quality check by gel electrophoresis, the samples were sequenced using the Illumina NextSeq 500 platform using 75 cycles single end. Sequencing reads were aligned to the Rnor_6.0 and rat mRatBN7.2 genome assembly using STAR v2.5.3a aligner with default parameters. Transcriptional start regions were defined using HOMER’s findPeaks tool.

**Single nuclei (sn) RNA-seq data**

snRNA-seq data used for the alignment metrics were obtained from rat amygdala using the Droplet-based Chromium Single-Cell 3’ solution (10x Genomics, v3 chemistry), as previously described. Briefly, nuclei were isolated from frozen brain tissues and purified by flow cytometry. Sorted nuclei were counted and 12,000 were loaded onto a Chromium Controller (10x Genomics). Libraries were generated using the Chromium Single-Cell 3’ Library Construction Kit v3 (10x Genomics, 1000075) with the Chromium Single-Cell B Chip Kit (10x Genomics, 1000153) and the Chromium i7 Multiplex Kit for sample indexing (10x Genomics, 120262) according to manufacturer specifications. Final library concentration was assessed by Qubit dsDNA HS Assay Kit (Thermo-Fischer Scientific) and post library QC was performed using Tapestation High Sensitivity D1000 (Agilent) to ensure that fragment sizes were distributed as expected. Final libraries were sequenced using the NovaSeq6000 (Illumina). Sequencing reads were aligned to the Rnor_6.0 and rat mRatBN7.2 genome assembly using Cell Ranger 3.1.0.

**Transcriptome termini site sequencing**

A total of 83 WTTS-seq (whole transcriptome termini site sequencing) libraries were constructed individually using total RNA samples derived from several brain tissues of rats. Library sequencing produced a total of 312,092,803 raw reads, but the mapped reads were 240,225,046 (76.97%) on Rnor6.0, while 251,188,567 (80.49%) on mRatBN7.2, respectively. Using 25 reads per clustered site as a cutoff, we identified 173,124 APA sites mapped to the new reference genome, while 167,136 APA sites were assigned to the old reference genome. For Rnor6.0, only 127,460 (76.26%) APA sites were assigned to the genome regions with 18,177 annotated genes. In contrast, 141,399 (81.67%) APA sites were mapped to the 20,102 annotated genes on mRatBN7.2. In brief, our results provide evidence that mRatBN7.2 has improved qualities of both genome assembly and gene annotation in rats.
Brain proteome data
Deep proteome data were generated using whole brain tissue from both parents and 29 members of the HXB family, one male and one female per strain. Proteins in these samples were identified and quantified using the tandem-mass-tag (TMT) labeling strategy coupled with two-dimensional liquid chromatography-tandem mass spectrometry (LC/MS/MS). We used the QTLtools program for protein expression quantitative trait locus (pQTL) mapping. cis-pQTL are defined when the transcriptional start sites for the tested protein are located within ±1 Mb of each other.

Identifying potentially mislabeled samples
Phylogenetic analysis identified that the metadata of 15 samples contradicted their genetic relationships. These contradictions could happen at many steps during breeding, samples collection, sequencing, or data analysis and the true source is difficult to pinpoint. An advantage of having multiple biological samples for each strain is that these inconsistencies can be identified. One of the 15 samples is mislabeled and the correct label can be inferred (sample name denoted with ***); two samples are mislabeled and the correct labels cannot be inferred (sample names denoted with **); 12 samples are potentially mislabeled (sample names denoted with *). Details of these samples are described below.

(1) There is enough evidence to indicate that this sample is mislabeled, and we can conclusively infer what the true label is. One sample falls into this category: F344/Strm_HCJL.CRM. Despite being named F344, this sample has less than 70% IBS with the other 14 F344 samples, yet has about 99% IBS with 2 LE samples from different institutes (Table S10). We think this sample is mislabeled as F344, and the true label should be LE. Therefore, we changed the sample name accordingly and appended *** to the end of the sample name to denote such a change has been made.

(2) There is enough evidence to indicate that this sample is mislabeled, but we cannot conclusively infer what the true label is. Two samples fall into this category: LE/Strm_HCJL.CRM has about 83% IBS with the other 3 LE samples from different institutes (Table S10). The identity of this sample is likely one of the LEXF or FXLE recombinant inbred, but there isn’t another sample that has high IBS with it. We think this sample is mislabeled, but the true label is unknown. WKY/Gia_TA.ILM is a sample we downloaded from SRA. This WKY strain (WKY/Gia) clusters closer to SHR strains than other WKY substrains. The same pattern was also observed in one prior study, and was thought to be caused by the incomplete inbreeding before sample distribution. To further investigate this, we performed regional similarity analysis and found that the pattern observed is consistent with that of a congenic strain created by using SHR as the recipient and WKY as the donor. A literature search confirmed that such strains were indeed once created at the same institute from where WKY/Gia was derived.

(3) There is evidence to suggest that this sample could potentially be mislabeled, but evidence is not conclusive. A total of 12 samples fall in this category. The majority of the samples of the same strain but sequenced by different institutes have IBS over 99%; however, we observed a few instances of unexpected low IBS between samples of the same strain but with unexpected high IBS between samples of a different strain. These could be caused by the mis-labeling at either of the institutes. Although we think these samples are at the risk of being mislabeled, it is also possible that individual differences with these strains/substrains could be a cause of the unexpected IBS values. For example, both 7.7% of the variants of the BXH2 samples are heterozygous, while the rate of heterozygosity in the LEXF/FXLE in general is higher than the rest of the inbred strains. These pairs include (Table S10):

- BXH2_MZ.ILM & BXH2_HCRW.CRM: unexpected low IBS at 92%
- LEXF4_MZ.ILM & LEXF5_HCJL.CRM: unexpected high IBS at 99%
- LEXF3_MZ.ILM & LEXF4_HCJL.CRM: unexpected high IBS at 99%
- LEXF1A_MZ.ILM & LEXF1_HCJL.CRM: unexpected high IBS at 99%
- LEXF1A_MZ.ILM & LEXF2A_HCJL.CRM: unexpected high IBS at 99%
- LEXF2B_MZ.ILM & LEXF1A_HCJL.CRM: unexpected high IBS at 98%
- LEXF4_MZ.ILM & LEXF4_HCJL.CRM: unexpected low IBS at 83%
- LEXF1A_MZ.ILM & LEXF1A_HCJL.CRM: unexpected low IBS at 84%
- LEXF1C_MZ.ILM & LEXF1C_HCJL.CRM: unexpected low IBS at 95%

Ensembl annotation
Annotation of the assembly was created via the Ensembl gene annotation system. A set of potential transcripts was generated using multiple techniques: primarily through alignment of transcriptomic datasets, cDNA sequences, curated evidence, and also through gap filling with protein-to-genome alignments of a subset of mammalian proteins with experimental evidence from UniProt. Additionally, a whole genome alignment was generated between the genome and the GRCm39 mouse reference genome using LastZ and the resulting alignment was used to map the coding regions of mouse genes from the GENCODE reference set.

The short-read RNA-seq data was retrieved from two publicly available projects; PRJEB6938, representing a wide range of different tissue samples such as liver or kidney, and PRJEB1924 which is aimed at understanding olfactory receptor genes. A subset of samples from a long-read sequencing project PRJNA517125 (SRR8487230, SRR8487231) were selected to provide high quality full length cDNAs.
From the 104 Ensembl release annotation on the rat assembly Rnor_6.0, we retrieved the sequences of manually annotated transcripts from the HAVANA manual annotation team. These were primarily clinically relevant transcripts, and represented high confidence cDNA sequences. Using the *Rattus norvegicus* taxonomy id 10116, cDNA sequences were downloaded from ENA and sequences with the accession prefix ‘NM’ from RefSeq.

The UniProt mammalian proteins had experimental evidence for existence at the protein or transcript level (protein existence level 1 and 2).

At each locus, low quality transcript models were removed, and the data were collapsed and consolidated into a final gene model plus its associated non-redundant transcript set. When collapsing the data, priority was given to models derived from transcriptomic data, cDNA sequences and manually annotated sequences. For each putative transcript, the coverage of the longest open reading frame was assessed in relation to known vertebrate proteins, to help differentiate between true isoforms and fragments. In loci where the transcriptomic data were fragmented or missing, homology data was used to gap fill if a more complete cross-species alignment was available, with preference given to longer transcripts that had strong intron support from the short-read data.

Gene models were classified, based on the alignment quality of their supporting evidence, into three main types: protein-coding, pseudogene, and long non-coding RNA. Models with hits to known proteins, and few structural abnormalities (i.e., they had canonical splice sites, introns passing a minimum size threshold, low level of repeat coverage) were classified as protein-coding. Models with hits to known protein, but having multiple issues in their underlying structure, were classified as pseudogenes. Single-exon models with a corresponding multi-exon copy elsewhere in the genome were classified as processed pseudogenes.

If a model failed to meet the criteria of any of the previously described categories, did not overlap a protein-coding gene, and had been constructed from transcriptomic data then it was considered as a potential IncRNA. Potential IncRNAs were filtered to remove transcripts that did not have at least two valid splice sites or cover 1000bp (to remove transcriptional noise).

A separate pipeline was run to annotate small non-coding genes. miRNAs were annotated via a BLAST against miRbase, with the results passed into RNAfold. Poor quality and repeat-ridden alignments were discarded. Other types of small non-coding genes were annotated by scanning Rfam against the genome and passing the results into Infernal.

The annotation for the rat assembly was made available as part of Ensembl release 105.

**RefSeq annotation**

Annotation of the mRatBN7.2 assembly was generated for NCBI’s RefSeq dataset using NCBI’s Eukaryotic Genome Annotation Pipeline. The annotation, referred to as NCBI *Rattus norvegicus* Annotation Release 108, includes gene models from curated and computational sources for protein-coding and non-coding genes and pseudogenes, and is available from NCBI’s genome FTP site and web resources.

Most protein-coding genes and some non-coding genes are represented by at least one known RefSeq transcript, labeled by the method “BestRefSeq” and assigned a transcript accession starting with NM_ or NR_, and corresponding RefSeq proteins designated with NP_. These are predominantly based on rat mRNAs subject to manual and automated curation by the RefSeq team for over 20 years, including automated quality analyses and comparisons to the Rnor_6.0 and mRatBN7.2 assemblies to refine the annotations. Nearly 80% of the protein-coding genes in AR108 include at least one NM_RefSeq transcript, of which 33% have been fully reviewed by RefSeq curators.

Additional gene, transcript, and protein models were predicted using NCBI’s Gnomon algorithm using alignments of transcripts, proteins, and RNA-seq data as evidence. The evidence datasets used for Release 108 are described at [https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rattus_norvegicus/108/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rattus_norvegicus/108/), and included alignments of available rat mRNAs and ESTs. 10.7 billion RNA-seq reads from 303 SRA runs from a wide range of samples, 1 million Oxford Nanopore or PacBio transcript reads from 5 SRA runs, and known RefSeq proteins from human, mouse, and rat. BestRefSeq and Gnomon models were combined to generate the final annotation, compared to the previous Release 106 annotation of Rnor_6.0 to retain GeneID, transcript, and protein accesses for equivalent annotations, and compared to the RefSeq annotation of human GRCh38 to identify orthologous genes. Gene nomenclature was based on data from RGD, curated names, and human orthologs.
Supplemental information

A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats

Figure S1 (related to Figure 1). The Contig Nx plot for mRatBN7.2 and 6 prior rat reference genomes. The red dashed line indicates the contig N50 values. Small improvements in contig continuity are observed for the previous updates, with the most significant improvement (~290X) coming from mRatBN7.2
**Figure S2 (related to Figure 1). The Contig Nx plot for the Rat, Mouse, and Human reference genomes.** The red dashed line indicates the contig N50 values. The Nx curve for Rnor6.0 is very low and partially overlapping with the X axis. The CHM13, being the first truly gapless human genome, is in its own tier in terms of assembly continuity. The GRCh38 released in 2014 and the GRCm39 released in 2020 have similar continuity. Although mRatBN7.2 is still lagging behind in continuity compared to the current human and mouse reference genomes, it represents a very significant improvement over the current rat reference genome Rnor_6.0 (Orange vs. Cyan line).
Figure S3 (related to Figure 1). Chromosomal dot plots between Rnor_6.0 and mRatBN7.2. mRatBN7.2 is on the x-axis and Rnor_6.0 is on the y-axis. mRatNor1 is the initial name of the assembly released by the Darwin Tree of Life/Vertebrate Genome Project. chr1-chr4 are shown here. To see the complete set, see the Key Resources Table “Chromosomal dot plots between Rnor_6.0 and mRatBN7.2” https://doi.org/10.5281/zenodo.10515796
Figure S4 (related to Figure 1). SNP and structural variants between Rnor_6.0 and mRatBN7.2. **A)** The substitution frequency per chromosome for Rnor_6.0 and mRatBN7.2. **B)** The size of indels between Rnor 6.0 and mRatBN7.2. **C)** The size of SVs between Rnor6.0 and mRatBN7.2.
Figure S5 (related to Figure 1). The order of genetic markers and the distances from a rat genetic map compared to their locations in Rnor_6.0. Note the large inversion at proximal Chr 6. chr1-chr6 are shown here. To see the complete set, see the Key Resources Table “The order of genetic markers and the distances from a rat genetic map compared to their locations in Rnor_6.0” https://doi.org/10.5281/zenodo.10520087
Figure S6 (related to Figure 1). The order of genetic markers and the distances from a rat genetic map compared to their locations in mRatBN7.2 chr1-chr6 are shown here. To see the complete set, see the Key Resources Table “The order of genetic markers and the distances from a rat genetic map compared to their locations in mRatBN7.2” https://doi.org/10.5281/zenodo.10520119
Figure S7 (related to Figure 1). The number of transcripts annotated by Ensembl and RefSeq. RefSeq annotated a greater number of genes with multiple transcripts than Ensembl. The average number of transcripts per gene was 2.9 for RefSeq and 1.8 for Ensembl.
Figure S8 (related to Figure 2). Mapping results of nanopore data of one HS rat with 55x coverage. Using mRatBN7.2 as the reference decreased the number of structural variants compared to using Rnor_6.0 as the reference. In the context of our other findings, this is likely due to the improved quality of the reference genome.
Figure S9 (related to Figure 2). Locations of large deletions from multiple linked-read samples. A) Rnor_6.0 B) mRatBN7.2. From outer circle to inner circle are the following strains: BN, SHR/Olalpcv, BXH10, BXH8, BXH2, HXB17, HXB2, and HXB21.
Figure S10 (related to Figure 2). Mapping metrics of single nuclei RNA-seq and small capped RNA-seq data generated using rat brain tissues. Percentage of snRNA-seq reads mapped to A) the genome, B) the transcriptome, and C) an intergenic region of the genome for mRatBN7.2 and Rnor_6.0. D) Overall mapping rates of csRNA-seq data to two reference genomes. E) Unique alignment of csRNA-seq data to two reference genomes.
Figure S11 (related to Figure 6). Distribution of the total number of variants shared by 1 or more samples across all 163 samples for A) Homozygous alternative SNPs/Indels. The lack of variants shared by approximately 140-155 samples and the distinct peak after 155 samples indicate variants shared by more than 156 samples are likely caused by errors in the reference genome. B) Heterozygous SNPs/Indels. Similar pattern as A can be observed.
Figure S12 (related to Figure 6). SNPs and indels indicate remaining errors in mRatBN7.2. Base-level errors are indicated by homozygous variants that are shared by the majority (i.e. more than 153 out of 163) of samples, including all seven BN/NHsdMcwi rats, one of them is part of the data used to assemble mRatBN7.2. Variants that are heterozygous for the majority of the samples are clustered in a few regions and have significantly higher read-depth. This suggests that they originated from collapsed repeats in mRatBN7.2.
Figure S13 (related to Figure 4). Simulated liftover analysis: unliftable and lifted site distribution. An evenly 1000 bp spaced bed file covering Rnor_6.0 is generated and then lifted to mRatBN7.2. Out of the 2,782,023 sites, 92.04% are liftable, 7.96% are not liftable. A) The distribution of the unliftable sites are plotted on Rnor_6.0. B) The distribution of the lifted sites are plotted on mRatBN7.2. The data are subsampled by 50 for better visualization.
Figure S14 (related to Figure 4). Simulated liftover analysis and comparative genome view. Top red track is the unliftable sites distributed on the Rnor_6.0. Bottom green track is the lifted sites distributed on the mRatBN7.2. The middle is the NCBI comparative genome viewer between chr1 of Rnor_6.0 and mRatBN7.2. The red arrow highlights a region that’s not liftable in Rnor_6.0. According to the comparative genome view, the region has no corresponding region in mRatBN7.2 (not just chr1). The green arrow highlights a region in mRatBN7.2 that has no lifted sites from Rnor_6.0. According to the comparative genome view, this region has matching region in Rnor_6.0. A) Chromosome 1. B) Chromosome 2.
Figure S15 (related to Figure 5). Density of variants across the genome in a collection of 163 rat WGS samples. The mean variant density was 5.96 ± 2.20/Kb (mean ± SD). The highest variant density of 30.5/Kb was found on Chr 4 at 98 Mb.
Figure S16 (related to Figure 5). Homozygosity of inbred strains. Variants in most sample were homozygous, confirming the inbred nature of most strains. A few exceptions were noted. For example, 15.6% of the variants from FXLE24 were heterozygous. Additionally, BXH2, which we sequenced two samples, has ~7.7% heterozygous variants.
Figure S17 (related to Figure 5). The majority of the variants in HXB/BXH RI panel are inherited from the parental strains. De Novo Mutations on BN.Lx (Red) Mutations originating from BN (Yellow), Heterozygous (Green), Mutations originating from SHR (Blue), SNPs found on both BN and SHR (Grey), De Novo SNPs from neither parental strains (Pink).
Figure S18 (related to Figure 5). Distribution of variants different between SS and SR rats. Total number of variants unique to either SS or SR rats across the genome. The SS strain contains more unique variants than SR.
Figure S19 (related to Figure 5). Distribution of variants unique to LL/LH/LN rats. Strain-specific variants tend to cluster by strain.
Figure S20 (related to Figure 5). Overview of disease ontology annotation of variants in RatCollection. Heat maps shows relative number of SNPs in proximity of genes associated per disease category within the reference, classic inbreds, HS progenitors and recombinant inbred panels.
Figure S21 (related to Figure 5). Disease ontology of genetic variants found in the HXB/BXH RI panel. Numbers within the grid show the absolute number of SNPs with high impact on genes within the disease annotation. Numbers following the disease name shows the total number of annotated genes with at least 1 high impact variant within the panel.
Figure S22 (related to Figure 5). Disease ontology annotation of variants in the LEXF/FXLE RI panel. Numbers within the grid show the absolute number of SNPs with high impact on genes within the disease annotation. Numbers following the disease name shows the total number of annotated genes with at least 1 high impact variant within the panel.
Supplemental references in Excel files


PMID: 23146691.


[72] W. D’Souza, S. Pradhan, and D. Saranath. Multiple single nucleotide polymorphism analysis and association of specific genotypes in fhit, samd4a, and ankrd17 in indian


[81] B. Chen, W. Wei, L. Ma, B. Yang, RM. Gill, MS. Chua, AJ. Butte, and S. So. Computational discovery of niclosamide ethanolamine, a repurposed drug candidate that reduces


the dihydropyrimidine dehydrogenase gene interpreted by analysis of the three-
PMC1222557, PMID: 11988088.

[136] AB. Van Kuilenburg, R. Meinsma, L. Zoetekouw, and AH. Van Gennip. Increased risk of
grade iv neutropenia after administration of 5-fluorouracil due to a dihydropyrimidine
dehydrogenase deficiency: high prevalence of the ivs14+1g>a mutation. Int J Cancer,

[137] AB. Van Kuilenburg, R. Meinsma, L. Zoetekouw, and AH. Van Gennip. High prevalence of
the ivs14 + 1g>a mutation in the dihydropyrimidine dehydrogenase gene of patients with
PMID: 12360106.

chromatography analysis of the dpyd gene in patients with lethal 5-fluorouracil toxicity.

Santos, MR. Teixeira, and S. Castedo. Mutations in exon 14 of dihydropyrimidine
dehydrogenase and 5-fluorouracil toxicity in portuguese colorectal cancer patients.

5-fluorouracil/irinotecan induced lethal toxicity as a result of a combined
PMCID: PMC1770668, PMID: 15858133.

[141] NA. Al-Sanna’a, AB. Van Kuilenburg, TM. Atrak, MA. Abdul-Jabbar, and AH. Van Gennip.
Dihydropyrimidine dehydrogenase deficiency presenting at birth. J Inherit Metab Dis,

[142] HH. Ezzeldin, AM. Lee, LK. Mattison, and RB. Diasio. Methylation of the dpyd promoter:
an alternative mechanism for dihydropyrimidine dehydrogenase deficiency in cancer

Pharmacogenetics of capecitabine in advanced breast cancer patients. Clin Cancer Res,

Gamelin. 5-fluorouracil-related severe toxicity: a comparison of different methods for the
pretherapeutic detection of dihydropyrimidine dehydrogenase deficiency. Cancer Lett,

Clinical relevance of different dihydropyrimidine dehydrogenase gene single nucleotide
PMID: 17121937.


[182] TK. Froehlich, U. Amstutz, S. Aebi, M. Joerger, and CR. Largiadèr. Clinical importance of risk variants in the dihydropyrimidine dehydrogenase gene for the prediction of early-


28823277.


Roslin, M. Zaik, K. Bushby, V. Straub, M. Zatz, F. de Paula, K. Morgan, TM. Fujiwara, and K. Wrogemann. The most common mutation in fkrp causing limb girdle muscular dystrophy type 2i (lgmd2i) may have occurred only once and is present in hutterites and other populations. Hum Mutat, 25(1):38–44, 2005. PMID: 15580560.


[442] CE. Mapp, B. Beghè, A. Balboni, G. Zamorani, M. Padoan, L. Jovine, OR. Baricordi, and


[455] RS. Thomas, TM. O’Connell, L. Pluta, RD. Wolfinger, L. Yang, and TJ. Page. A


