

# UC San Diego

## UC San Diego Previously Published Works

### Title

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

### Permalink

<https://escholarship.org/uc/item/9p74s81d>

### Journal

Cell Genomics, 4(4)

### ISSN

2666-979X

### Authors

de Jong, Tristan V

Pan, Yanchao

Rastas, Pasi

et al.

### Publication Date

2024-04-01

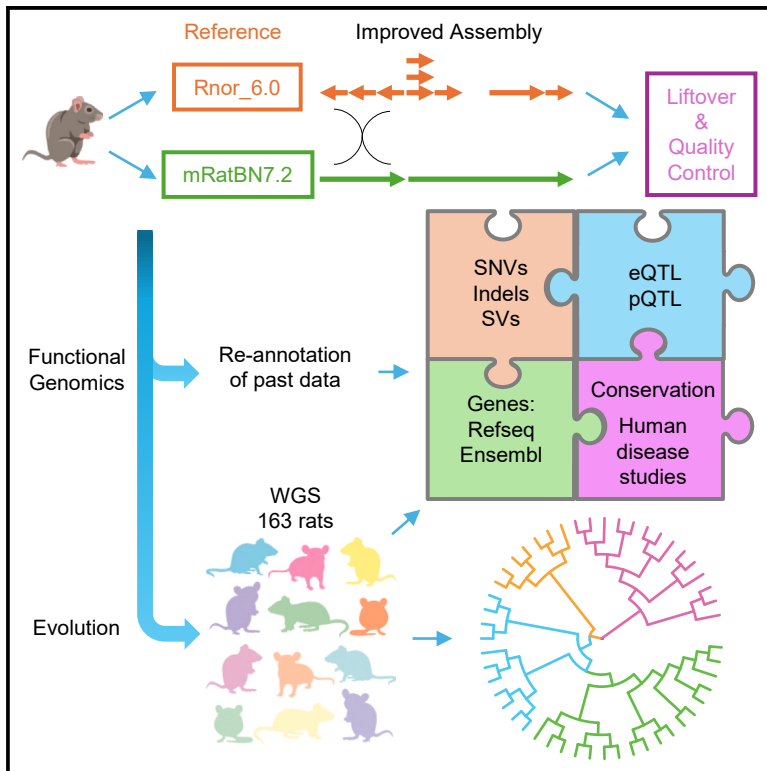
### DOI

10.1016/j.xgen.2024.100527

Peer reviewed

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

### Graphical abstract



### 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.

### 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



## Resource

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

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

(Author list continued on next page)

<sup>1</sup>Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center, Memphis, TN, USA

<sup>2</sup>Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

<sup>3</sup>Institute of Biotechnology, University of Helsinki, Helsinki, Finland

<sup>4</sup>Department of Psychiatry, University of California San Diego, San Diego, CA, USA

<sup>5</sup>Department of Integrative Structural and Computational Biology, Scripps Research, San Diego, CA, USA

<sup>6</sup>Department of Physiology, Medical College of Wisconsin, Milwaukee, WI, USA

<sup>7</sup>Rat Genome Database, Medical College of Wisconsin, Milwaukee, WI, USA

<sup>8</sup>Michigan Neuroscience Institute, University of Michigan, Ann Arbor, MI, USA

<sup>9</sup>Department of Medicine, University of California San Diego, San Diego, CA, USA

<sup>10</sup>Tree of Life, Wellcome Sanger Institute, Cambridge, UK

(Affiliations continued on next page)

## 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.<sup>1</sup> The earliest studies using brown rats appeared in the early 1800s.<sup>2,3</sup> The Wistar rats, the ancestor of many laboratory strains, were bred for scientific research in 1906.<sup>4</sup> Over 4,000 inbred, outbred, congenic, mutant, and transgenic strains have been created and are documented in the Rat Genome Database (RGD).<sup>5</sup> Approximately 500 are available from the Rat Resource and Research Center.<sup>6</sup> Several genetic reference populations, including the HXB/BXH<sup>7</sup> and FXLE/LEXF<sup>8</sup> recombinant inbred (RI) families, are also available. Both families, together with 30 diverse classical inbred strains,<sup>9</sup> 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,<sup>10</sup> have been increasingly used for fine map-

ping of physiological and behavioral traits.<sup>11–15</sup> To date, RGD has annotated nearly 2,400 rat quantitative trait loci (QTLs),<sup>16</sup> 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*.<sup>17</sup> The inbred Brown Norway (BN/NHsdMcow) strain, derived by many generations of sibling matings of stock originally from a pen-bred colony of wild rats,<sup>4</sup> was used to generate the reference. Several updates were released over the following decade.<sup>18–20</sup> Since 2014, most rat genomic and genetic research used the incomplete and problematic Rnor\_6.0 assembly.<sup>21,22</sup> mRatBN7.2 was created in 2020 by the Darwin Tree of Life/Vertebrate Genome Project (VGP) as the new genome assembly of the BN/NHsdMcow rat.<sup>23</sup> The Genome Reference Consortium (GRC) (<https://www.ncbi.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



Spencer Mahaffey,<sup>19</sup> Fergal J. Martin,<sup>16</sup> Pejman Mohammadi,<sup>20,21</sup> Ayse Bilge Ozel,<sup>2</sup> Oksana Poleskaya,<sup>4</sup> Michal Pravenec,<sup>22</sup> Pjotr Prins,<sup>12</sup> Jonathan Sebat,<sup>4</sup> Jennifer R. Smith,<sup>6,7</sup> Leah C. Solberg Woods,<sup>23</sup> Boris Tabakoff,<sup>19</sup> Alan Tracey,<sup>10</sup> Marcela Uliano-Silva,<sup>10</sup> Flavia Villani,<sup>12</sup> Hongyang Wang,<sup>24</sup> Burt M. Sharp,<sup>12</sup> Francesca Telese,<sup>4</sup> Zhihua Jiang,<sup>24</sup> Laura Saba,<sup>19</sup> Xusheng Wang,<sup>12,18</sup> Terence D. Murphy,<sup>25</sup> Abraham A. Palmer,<sup>4,26</sup> Anne E. Kwitek,<sup>6,7</sup> Melinda R. Dwinell,<sup>6,7</sup> Robert W. Williams,<sup>12</sup> Jun Z. Li,<sup>2,\*</sup> and Hao Chen<sup>1,28,\*</sup>

<sup>11</sup>Institute of Genetics and Biophysics, National Research Council, Naples, Italy

<sup>12</sup>Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

<sup>13</sup>Department of Anatomy, Physiology & Genetics, The American Genome Center, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

<sup>14</sup>The Brown Foundation Institute of Molecular Medicine, Center for Human Genetics, University of Texas Health Science Center, Houston, TX, USA

<sup>15</sup>Genome Structure and Ageing, University of Groningen, UMC, Groningen, the Netherlands

<sup>16</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus in Hinxton, Cambridgeshire, UK

<sup>17</sup>Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Louisville, KY, USA

<sup>18</sup>Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN, USA

<sup>19</sup>Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

<sup>20</sup>Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA, USA

<sup>21</sup>Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA

<sup>22</sup>Institute of Physiology, Czech Academy of Sciences, Prague, Czechia

<sup>23</sup>Department of Internal Medicine, Section on Molecular Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA

<sup>24</sup>Department of Animal Sciences, Washington State University, Pullman, WA, USA

<sup>25</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

<sup>26</sup>Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA

<sup>27</sup>These authors contributed equally

<sup>28</sup>Lead contact

\*Correspondence: [junzli@med.umich.edu](mailto:junzli@med.umich.edu) (J.Z.L.), [hchen@uthsc.edu](mailto:hchen@uthsc.edu) (H.C.)

<https://doi.org/10.1016/j.xgen.2024.100527>

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.<sup>24</sup> While mRatBN7.2 contains an alternative pseudohaplotype (GCA\_015244455.1), we focused our analysis on the primary assembly (GCF\_015227675.2).

Over the last six iterations of the rat reference, genome contiguity has improved incrementally (Table S1). Contig N50, one measure of assembly quality, has been ~30 kb between rn1 to

rn5. 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.<sup>15</sup> Comparing the order of the markers and their location on the reference confirmed that the order and orientation of genomic segments are much 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.

**Table 1. Global statistics for the rat, mouse, and human reference genomes**

	Rnor_6.0	mRatBN7.2	GRCm39	GRCh38	CHM13
Year published	2014	2021	2020	2014	2021
Total sequence length	2,870,182,909	2,647,915,728	2,728,222,451	3,209,286,105	3,054,832,041
Total ungapped length	2,729,984,219	2,626,580,772	2,654,621,837	3,049,316,098	3,054,832,041
No. of scaffolds	953	176	61	455	24
Scaffold N50	145,729,302	135,012,528	130,530,862	145,138,636	154,259,566
Scaffold L50	8	8	9	9	8
No. of contigs	75,695	757	347	1,431	24
Contig N50	100,511	29,198,295	59,462,871	56,413,054	154,259,566
Contig L50	7,346	27	15	19	8
Total no. of chromosomes and plasmids	23	23	22	25	24

Genomes are downloaded from UCSC Goldenpath. Summary statistics are calculated based on the fasta files of each release using QUAST.<sup>25</sup>

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), two 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 parentals, is 105.5× (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*<sup>26</sup> and jointly called for the 36 samples using *GLnexus*.<sup>27</sup> 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 36 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

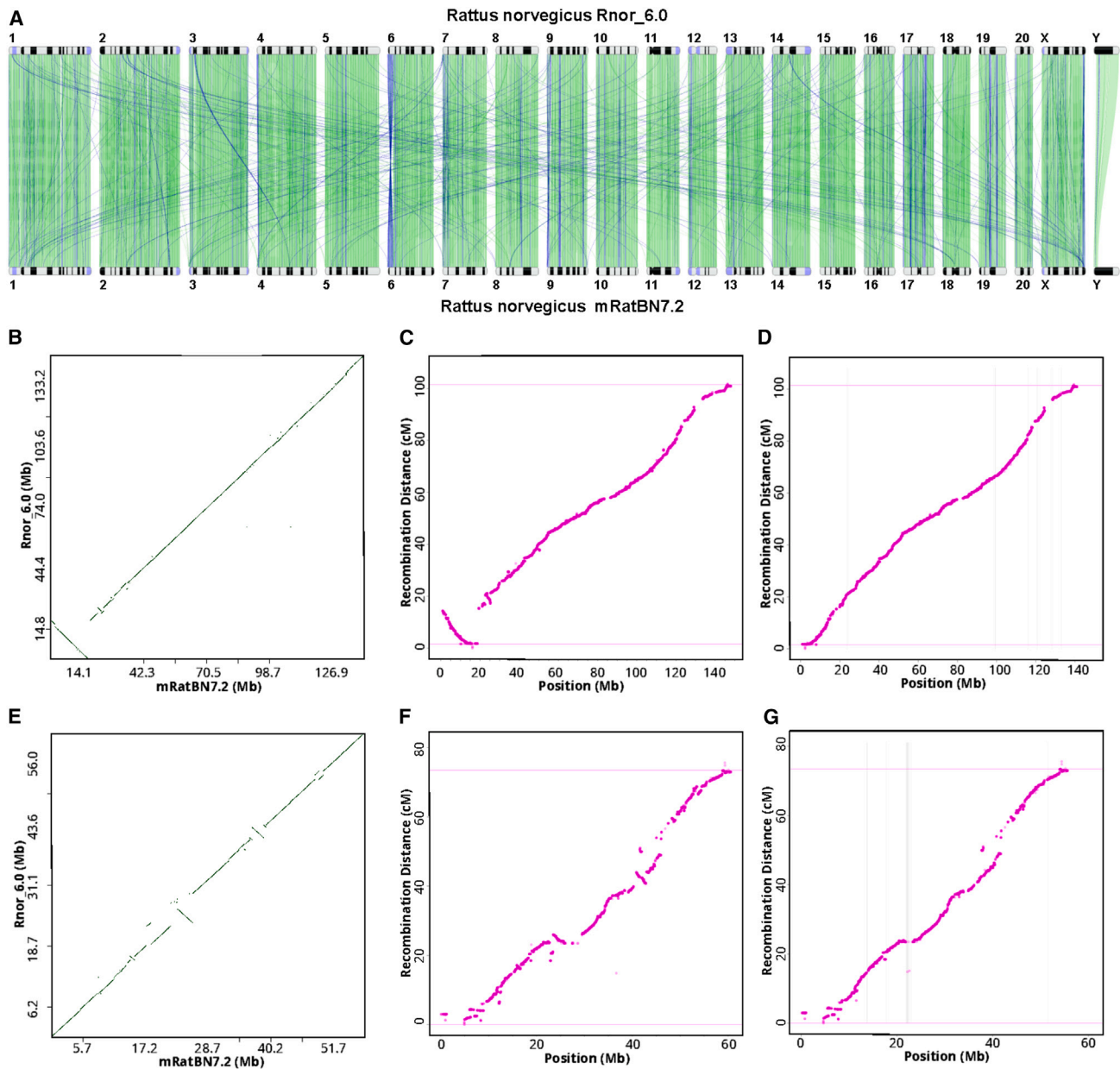
mRatBN7.2-based gene annotations, based on transcriptomic datasets and newly revised multi-species sequence alignments, were released in RefSeq 108 and Ensembl 105. We compared the Rnor6.0 and mRatBN7.2 annotation sets in RefSeq using BUSCO v.4.1.4,<sup>28</sup> focusing on the glires\_odb10 dataset 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 “ENSRNOG” stable IDs. These included 23,096 protein-coding genes, 2,488 lncRNA, 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×) and pseudogenes (almost 10×), 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*, *Adrb2*, etc., were not annotated with GeneID 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).





**Figure 1. mRatBN7.2 corrects structural errors in Rnor\_6.0**

(A) Genome-wide comparison between Rnor\_6.0 and mRatBN7.2 showed many structural differences, such as a large inversion at proximal Chr 6 and many translocations between chromosomes. Image generated using the NCBI Comparative Genome Viewer. Numbers indicate chromosomes. Green lines indicate sequences in the forward alignment. Blue lines indicate reverse alignment.

(B) The large inversion on proximal Chr 6 is shown in a dot plot between Rnor\_6.0 and mRatBN7.2.

(C) A rat genetic map generated using 150,835 binned markers from 1,893 heterogeneous stock rats showed an inversion at proximal Chr 6 between genetic distance and physical distance based on Rnor\_6.0, indicating that the inversion is caused by assembly errors in Rnor\_6.0.

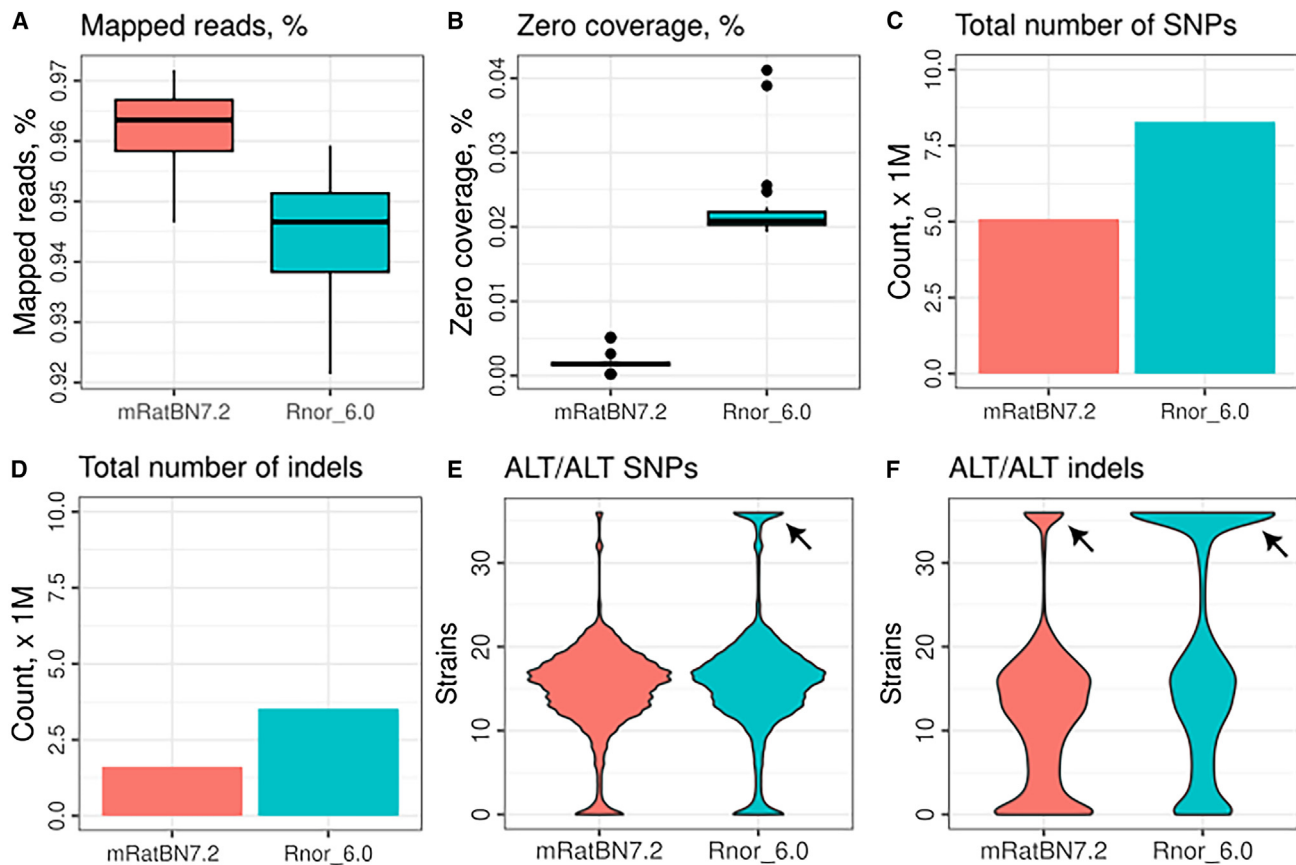
(D) Marker order and genetic distance from the genetic map on Chr 6 are in agreement with physical distance based on mRatBN7.2, indicating that the mis-assembly is fixed.

(E–G) Genetic map confirms that many assembly errors on Chr 19 in Rnor\_6.0 are fixed in mRatBN7.2.

### mRatBN7.2 improved the mapping of WGS data

We compared mapping results of long-read datasets. Un-mapped 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



**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.

data, structural variants (SVs) detected in the Nanopore dataset were reduced (Figure S8) using mRatBN7.2.

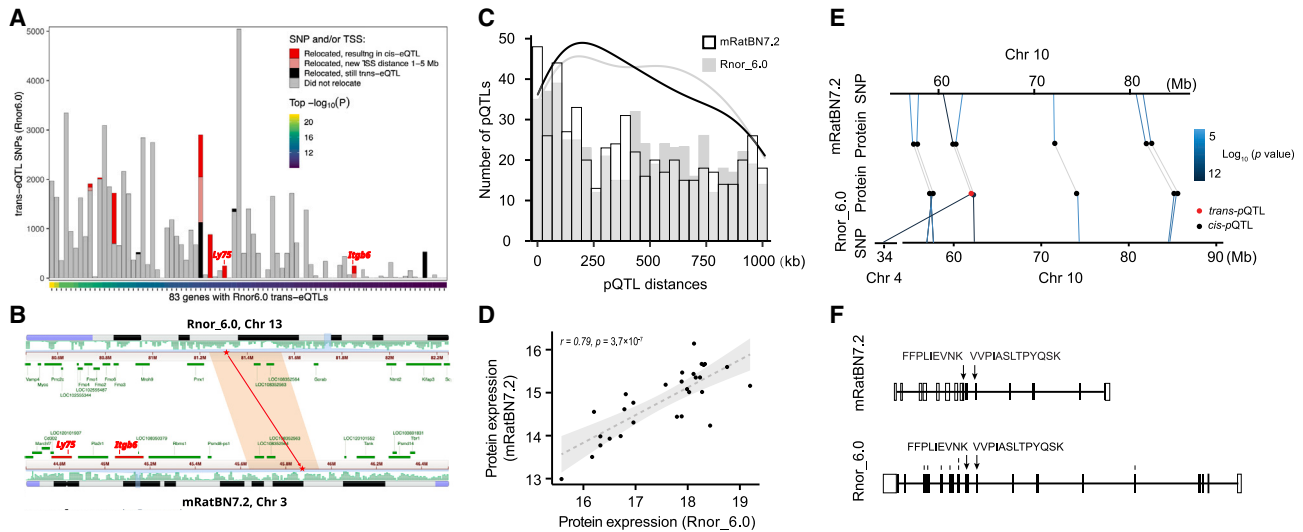
We also evaluated the effect of reference genome on the identification of SVs. We identified 19,538 unique SVs against Rnor\_6.0 in the HXB dataset. In contrast, only 5,458 SVs were found using mRatBN7.2 (see STAR Methods, key resources table, <https://doi.org/10.5281/zenodo.10398554>). We illustrated 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<sup>29</sup> 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 3. mRatBN7.2 improves eQTL and proteomic analysis**

Genome misassembly is associated with increased rates of calling spurious *trans*-eQTLs.

(A) Each column represents a gene for which at least one *trans*-eQTL was found at  $p < 1 \times 10^{-8}$  using Rnor\_6.0. The color of bars indicate the number of *trans*-eQTL SNP-gene pairs in which the SNP and/or gene transcription start site (TSS) relocated to a different chromosome in mRatBN7.2 and whether the relocation would result in a reclassification to *cis*-eQTL (TSS distance < 1 Mb) or ambiguous (TSS distance is between 1–5 Mb).

(B) Genomic location of one relocated *trans*-eQTL SNP from (A). The SNP is in a segment of Chr 13 in Rnor\_6.0 that was relocated to Chr 3 in mRatBN7.2 (red stars), reclassifying the eQTL from *trans*-eQTL to *cis*-eQTL for both *Ly75* and *Itgb6* genes (red bars).

(C) Histogram showing the distance between *cis*-pQTLs and TSS of the corresponding proteins. The distances of pQTLs in mRatBN7.2 tend to be closer than those in Rnor\_6.0.

(D) An example of *trans*-pQTL in Rnor\_6.0 was detected as a *cis*-pQTL in mRatBN7.2.

(E) Correlation of expression of the protein (the example in B) in Rnor\_6.0 and mRatBN7.2.

(F) Different annotations of the exemplar gene in Rnor\_6.0 and mRatBN7.2.

APA sites using Rnor\_6.0 and 73,124 APA sites using mRatBN7.2 (see STAR Methods, key resources table, <https://doi.org/10.5281/zenodo.10398476>). For Rnor\_6.0, only 76.26% APA sites were assigned to the genomic regions with 18,177 annotated genes (Table S5). In contrast, 81.67% APA sites were mapped to 20,102 annotated genes on mRatBN7.2 (Table S5).

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.<sup>30</sup> We identified associations that would be labeled as *trans*-eQTLs using one reference but *cis*-eQTL using the other, due to relocation of the SNP and/or the TSS. The expectation is that assembly errors will give rise to spurious *trans*-eQTLs. We found seven genes associated with one or more strong *trans*-eQTL SNPs 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. Given the much lower probability of a distant SNP-gene pair remapping to be in close proximity than vice versa under a null model of random relocations, this demonstrates a clear improvement in the accuracy and interpretation of eQTLs when using mRatBN7.2.

To evaluate the effect of reference genome on proteome data, 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

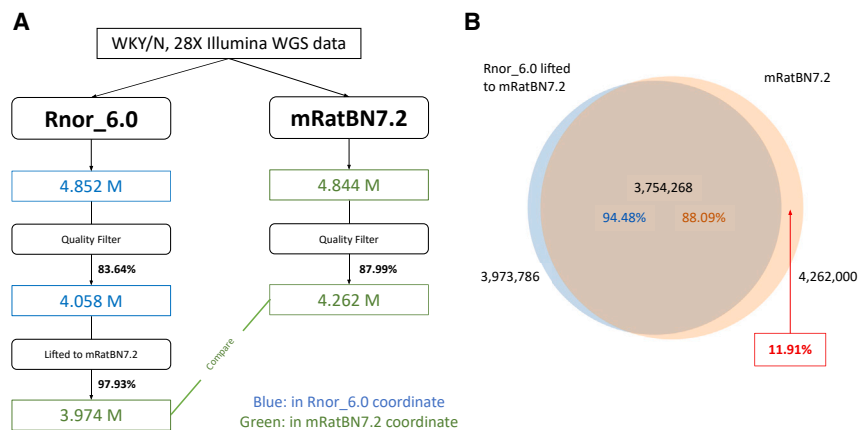
with 8,406 unique proteins on mRatBN7.2 (5% increase). For protein local expression quantitative trait locus (i.e., *cis*-pQTL), 536 were identified using Rnor\_6.0 and 541 were identified using mRatBN7.2 at FDR < 5%. Distances between pQTL peaks and the corresponding gene start site tended to be shorter 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 ( $p = 1.71 \times 10^{-12}$ ) on Chr 4 in Rnor\_6.0 but as a significant *cis*-pQTL using mRatBN7.2 ( $p = 4.12 \times 10^{-6}$ ) (Figure 3E). In addition, the expression of RPA1 protein displayed a high correlation between mRatBN7.2 and in Rnor\_6.0 ( $r = 0.79$ ;  $p = 3.7 \times 10^{-7}$ ) (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





**Figure 4. Using WGS data to assess the quality of the Liftover from Rnor\_6.0 to mRatBN7.2** (A) Overview of the workflow using a real WGS sample from a WKY rat. A higher portion of variants passed the quality filter for mRatBN7.2. Among them, 97.93% of the variants were liftable from Rnor\_6.0 to mRatBN7.2. (B) The overlap between variants lifted from Rnor\_6.0 and variants obtained by direct mapping sequence data to mRatBN7.2. Approximately 11.9% of the variants that were found from direct mapping were missing from the Liftover.

to show a negligible number of heterozygous sites, genomic regions with an unusually high density of heterozygous genotypes may indicate a segment-wise assembly error. For example, if two segments are tandem repeats of each other and have been “folded” into a single segment in the reference assembly, twice as many reads will map to this region and will produce a high density of heterozygous or multiple-allelic sites. We focused on the 12 BN samples, because BN is the basis of the reference. In all, we identified 673 such “flagged regions,” with an average length of 52,199 bp, which collectively covered ~1.4% of the genome (Table S6). This rate is much reduced from that of Rnor\_6.0.<sup>22</sup> Furthermore, regions with high heterozygosity were observed among the 151 non-BN samples outside the list of 673 flagged regions. These regions are likely to reflect shared structural variants in the non-BN samples.

We further examined per-site anomalies by identifying sites with Alt/Alt genotypes in most of the 163 strains, including the 12 BN strains. These 129,186 shared variant sites consisted of 33,550 SNPs and 95,636 indels (Table S7). Among them, 117,901 were homozygous alternative genotypes and 11,285 were heterozygous in more than 156 samples (Figure S11). The read depth was  $32.2 \pm 10.1$  for homozygous and  $66.3 \pm 26.2$  for heterozygous variants. Because all samples were inbred, the doubling of read depth for the heterozygous variants strongly suggests that they mapped to regions of the reference genome with collapsed repetitive sequences. This is supported by the location of these variants: homozygous SNPs were more evenly distributed, and heterozygous variants were often clustered in a region (Figure S12). These results indicated that many of the potential errors in mRatBN7.2 are caused by the collapse of repetitive sequences. Functionally, these regions of potential misassembly impact some well-studied genes, such as *Chat*, *Egfr*, *Gabrg2*, and *Grin2a*. After removing the likely errors, we annotated the VCF files resulting from the joint variant calling of 163 samples (see STAR Methods, key resources table, <https://doi.org/10.5281/zenodo.10398344>) and used these for subsequent analysis.

### Complexities in transitioning from Rnor\_6.0 to mRatBN7.2

Liftover tools can convert genomic coordinates between different versions of the reference assembly. However, Liftover can only

return results for regions with one-to-one matches between the two references. We examined Liftover from Rnor\_6.0 to mRatBN7.2 by testing a mock set of variant sites evenly distributed 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/LEFX strains, and 33 other inbred strains. In total, we covered 88 strains (120 samples at the substrain level)—approximately 80% of the HRDP.

The mean depth of coverage of these samples was  $60.4 \pm 39.3$ . Additional sample statistics are provided in Table S9. After removing variants that were potential errors in mRatBN7.2 (above) and filtering for site quality ( $qual \geq 30$ ), 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  $5.96 \pm 2.20/\text{kb}$  (mean  $\pm$  SD). The highest variant density of 30.5/kb was found on Chr 4 at 98 Mb (Figure S15). Most ( $97.9\% \pm 1.4\%$ ) of the variants were

**Table 2. Genetic variants in laboratory populations**

	Variants sites	Variant type				Predicted impact			Alt/Alt homozygous % (mean ± SD)	Genotype qual (mean ± SD)
		Variants	SNPs	Indels	Mixed <sup>a</sup>	High	Low	Moderate		
RatCollection	15,804,627	19,987,273	12,661,110	7,313,702	12,461	18,646	262,685	125,523	97.9 ± 1.4	70.1 ± 21.2
HS progenitors	12,418,243	16,438,302	9,947,112	6,479,485	11,705	6,980	205,316	94,659	98.5 ± 0.3	71.4 ± 22.0
FXLE/LEXF <sup>b</sup>	9,183,562	13,070,345	7,036,182	6,023,391	10,772	13,988	143,623	66,186	96.8 ± 2.4	72.3 ± 21.0
HXB/BXH <sup>b</sup>	7,520,223	11,256,227	5,705,592	5,541,195	9,440	10,121	115,081	53,819	97.9 ± 1.2	72.2 ± 22.8
SS/SR	7,171,447	10,522,763	5,544,220	4,969,398	9,145	8,606	111,658	51,149	98.0 ± 0.9	72.3 ± 21.8
LL/LN/LH	6,923,575	10,112,755	5,433,556	4,670,291	8,908	4,142	111,040	52,364	98.5 ± 0.3	73.2 ± 21.4

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.

<sup>a</sup>Variants that are combinations of insertions, deletions, or SNPs.

<sup>b</sup>Including parental strains.

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  $72.30\% \pm 2.35\%$ , 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).<sup>10</sup> 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 living proxies of the HS progenitors that were over 99.5% similar to the original progenitor strains: ACI/EurMcwi, BN/NHsdMcwi, F344/DuCrI, M520/NRrcMcwi, MR/NRrc, and WKY/NHsd (Table S11). The best matches of the remaining strains were much less similar: BUF/Mna was the best approximation (73.6%) to BUF/N, and WAG/RijCrI was the closest (72.0%) to WN/N.

Our phylogenetic tree also included two families of RI rats. The HXB/BXH family was generated from SHR/Olalpcv and BN-Lx/Cub. Together with their parental strains, we identified 7,520,223 variant sites in this population (Table 2). Approximately 24.1%–53.5% of the alleles at these sites of each RI strain were derived from SHR/Olalpcv. While the majority of the strains were highly inbred, with close to 98% of the variants being homozygous (Figure S17), one exception was BXH2, in which 7.7% of variants were heterozygous (Figure S16), likely due to a recent breeding error. The LEXF/FXLE RI family was generated from LE/Stm and F344/Stm. We discovered 9,183,562 variant sites from the parental strains and 25 strains of this family (Table 2). We expect the final variant count of the LEXF/FXLE panel

to be similar because both parental strains are included in our analysis. The overall rate of homozygous variants in the FXLE/LEXF family ( $96.8\% \pm 2.4\%$ ) was lower than in other inbred rats (Figure S16). In particular, 15.6% of the sites from the FXLE24 samples are heterozygous, indicating likely breeding errors.

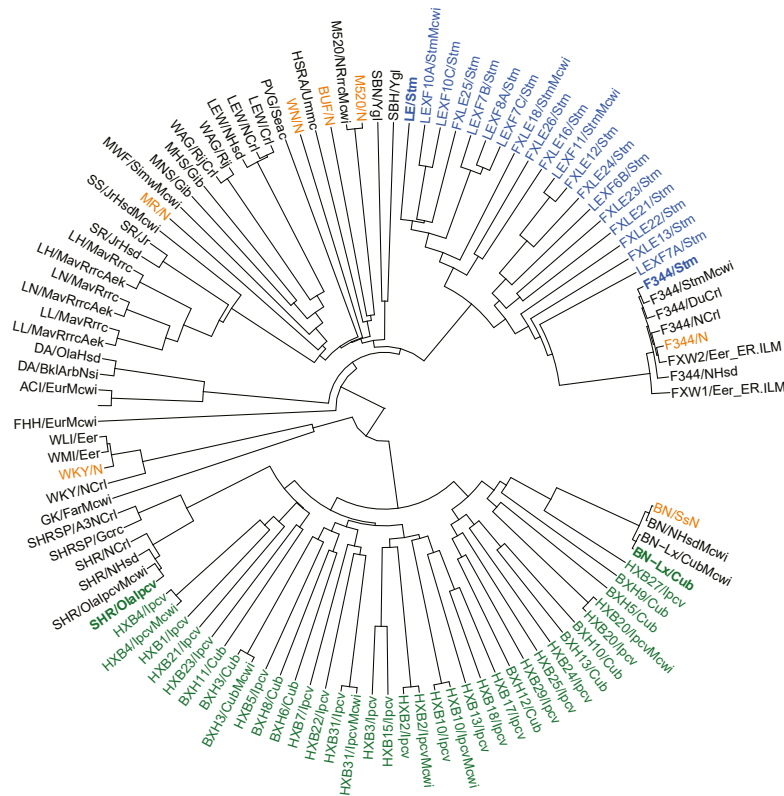
In addition, our analysis also included sets of two or three strains generated by selective breeding for certain traits, such as the Dahl salt-sensitive (SS) and Dahl salt-resistant (SR) strains for studying hypertension. These two strains contain 7,171,447 variant sites compared with mRatBN7.2 (Table 2), with 1,024,283 variants 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), hypotensive (LL), and normotensive (LN) rats (Table 2) selected from outbred Sprague-Dawley rats for studying blood pressure regulation.<sup>31</sup> Only 281,972, 289,112, and 262,574 variants were unique to LH, LL, and LN, respectively. In agreement with a prior report,<sup>32</sup> 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

A



**Figure 5. Phylogenetic relationship and genetic diversity of 120 strain/substrains of laboratory rats**

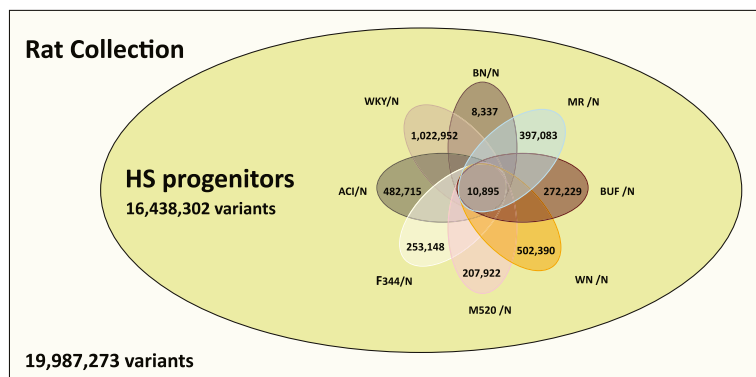
(A) The phylogenetic tree was constructed using 11.6 million biallelic SNPs from 163 samples. Strains/substrains with duplicated samples were condensed. Strains highlighted with bold fonts are parental strains for RI panels. Green, HXB/BXH RI panel; blue, FXLE/LEXF RI panel; orange, progenitors of the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

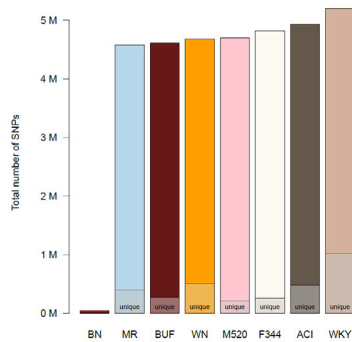
(C) The total number of variants per strain, with the total number unique to each strain marked.

(D) The number of variants shared across strains per chromosome.

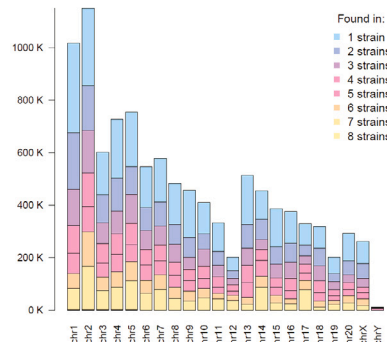
B



C



D



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.<sup>33</sup> 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.<sup>23</sup> However, the PacBio CLR reads used in mRatBN7.2 have lower base accuracy than Illumina short reads.<sup>34,35</sup> Although Illumina data were used to polish the assembly,<sup>23</sup> polishing methods do not correct all base-level errors.<sup>34,36</sup> 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,<sup>23,37</sup> which assembles the long reads into a diploid genome containing a primary and an alternative assembly.<sup>38</sup> This pipeline is well suited to assemble diploid genomes. When applied to BN/NHsdMcwi, a fully inbred rat, the pipeline classified some duplications with small variants as two haploids, resulting in a collapsed repeat.<sup>39</sup> A telomere-to-telomere assembly<sup>40</sup> 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.<sup>11</sup> analyzed 8 strains against Rnor3.4 and found 7.9 million variants, Atanur et al.<sup>41</sup> analyzed 27 rat strains against Rnor3.4 and reported 13.1 million variants, and Hermsen et al.<sup>42</sup> analyzed 40 strains against Rnor5.0 and reported 12.2 million variants. Most recently, Ramdas et al.<sup>22</sup> 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.<sup>27,43</sup> 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.<sup>44,45</sup>

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.<sup>46</sup> The new variant set from our analysis provide a much-needed boost in mapping precision.<sup>8,46,47</sup>

The outbred N/NIH HS rats are another widely used rat genetic mapping population.<sup>13</sup> Our analysis is in agreement with Ramdas et al.<sup>22</sup> 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,<sup>31,32</sup> 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.<sup>48</sup> 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,<sup>49</sup> hypertension,<sup>50,51</sup> diabetes,<sup>52</sup> immunity,<sup>52</sup> tumorigenesis,<sup>53</sup> and tissue-specific gene expression profiles.<sup>54</sup> Using the outbred HS rats, Baud et al.<sup>11</sup> reported 355 QTLs associated with phenotypes representing six diseases. Other studies using the HS population have revealed genetic control of behavior,<sup>15,55</sup> obesity,<sup>56</sup> and metabolic phenotypes.<sup>57</sup> Many of these data are available from the RGD.<sup>16</sup> While all these data were analyzed using previous versions of the reference genome, reanalyzing them using updated genomic data could lead to novel discoveries.<sup>58</sup>

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.<sup>44</sup> 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<sup>59</sup>) 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*<sup>60</sup> and *Procr*,<sup>61</sup> both associated with kidney injuries, as well as in *Klk1c12*, associated with hypertension.<sup>62</sup> 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.<sup>63</sup>

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.<sup>64</sup> A gain of STOP mutation in the *Cdhr3* 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<sup>41,65</sup> 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<sup>4</sup> and previous studies that included wild rats.<sup>66</sup> 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<sup>67</sup> and transcriptomic data analyses.<sup>68</sup> It should be noted that no individual strain is a perfect representation of a population. Instead, the nascent field of pangenomics,<sup>69</sup> 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,<sup>70</sup> it will be especially powerful when individual genomes are all assembled from long-read sequence data, some of which are already available.<sup>71</sup> 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.<sup>72,73</sup> Unlike human and mouse genomes,<sup>74</sup> 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 Lifter. 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 pangenome 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 Lifter
  - 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 nuclei (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.; Wellcome Trust WT222155/Z/20/Z to T.H.; NIH R01HG011252 to T.K.; Wellcome Trust WT222155/Z/20/Z to F.J.M.; NIH R01GM140287 to P.M.; a program from the National Institute for Research of Metabolic and Cardiovascular Diseases (Program EXCELES, ID project no. LX22NPO5104) funded by the European Union – Next Generation EU to M.P.; National Institute of Food and Agriculture, United States Department of Agriculture (2016-67015-24470/2020-67015-31733/2022-51300-38058/2023-67015-39566/2023-67015-40080) to Z.J.; NIH NIDA U01DA051234 to J.S.; NIH NHLBI R01HL064541, NHGRI U24HG010859, and Office of the Director R24OD024617 to J.R.S.; NIH NIDA P50DA037844 to L.C.S.W. (HS rats); NIH NIAAA R24AA013162 to B.T.; NIH NIDA U01DA050239 and U01DA051972 to F.T.; NIH NIDA P30DA044223 to L.S.; the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health to T.D.M.; NIH NHLBI R01HL064541 and P01HL149620, NHGRI U24HG010859, and Office of the Director R24OD024617 to A.E.K.; NIH Office of the Director grant R24OD024617 to M.R.D. (HRDP); NIH NIDA U01DA047638 and P30DA044223 to R.W.W.; NIH NIDA U01DA043098 to J.Z.L.; and NIH NIDA U01DA047638, P50DA037844, and R01DA048017 to H.C. The majority of the computation for this work was performed on the University of Tennessee Infrastructure for Scientific Applications and Advanced Computing (ISAAC) computational resources.

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

#### AUTHOR CONTRIBUTIONS

Conceptualization, A.A.P., R.W.W., J.Z.L., and H.C.; formal analysis, T.V.d.J., Y.P., P.R., D.M., M.T., X.W., T.D.M., J.Z.L., and H.C.; visualization, T.V.d.J., Y.P., P.R., D.M., F.T., Z.J., L.S., X.W., and J.Z.L.; investigation, H.A., C.B., D.C., A.S.C., W.C., V.C., W.M.D., P.A.D., E.G., A.M.G., H.M.G., V.G., T.H., K.H., J.H., T.K., P.K., L.L., S.M., F.J.M., P.M., A.B.O., O.P., P.P., J.S., J.R.S., L.C.S.W., B.T., A.T., M.U.-S., F.V., H.W., F.T., Z.J., and L.S.; resources, C.L.D., M.P., A.E.K., and M.R.D.; data curation, W.M.D., A.M.G., J.R.S., A.E.K., and M.R.D.; writing – original draft, T.V.d.J., Y.P., P.R., D.M., T.H., L.S., X.W., T.D.M., J.Z.L., and H.C.; writing – review & editing, A.M.G., B.M.S., L.S., X.W., T.D.M., A.A.P., A.E.K., M.R.D., R.W.W., J.Z.L., and H.C.; supervision, J.Z.L. and H.C.

#### 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

- Parker, C.C., Chen, H., Flagel, S.B., Geurts, A.M., Richards, J.B., Robinson, T.E., Solberg Woods, L.C., and Palmer, A.A. (2014). Rats are the smart choice: Rationale for a renewed focus on rats in behavioral genetics. *Neuropharmacology* 76, 250–258.
- Richter, C.P. (1954). The effects of domestication and selection on the behavior of the Norway rat. *J. Natl. Cancer Inst.* 15, 727–738.
- Hulme-Beaman, A., Orton, D., and Cucchi, T. (2021). The origins of the domesticated brown rat (*Rattus norvegicus*) and its pathways to domestication. *Anim. Front.* 11, 78–86.
- Modlinska, K., and Pisula, W. (2020). The Norway rat, from an obnoxious pest to a laboratory pet. *Elife* 9, e50651. <https://doi.org/10.7554/eLife.50651>.
- Smith, J.R., Hayman, G.T., Wang, S.-J., Laudekerkind, S.J.F., Hoffman, M.J., Kaldunski, M.L., Tutaj, M., Thota, J., Nalabolu, H.S., Ellanki, S.L.R., et al. (2020). The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res.* 48, D731–D742.
- RRRC (2021). Rat Resource & Research Center - Rat Models. <https://www.rrrc.us/>.
- Pravenec, M., Klir, P., Kren, V., Zicha, J., and Kunes, J. (1989). An analysis of spontaneous hypertension in spontaneously hypertensive rats by means of new recombinant inbred strains. *J. Hypertens.* 7, 217–221.
- Voigt, B., Kuramoto, T., Mashimo, T., Tsurumi, T., Sasaki, Y., Hokao, R., and Serikawa, T. (2008). Evaluation of LEXF/FXLE rat recombinant inbred strains for genetic dissection of complex traits. *Physiol. Genomics* 32, 335–342.
- Tabakoff, B., Smith, H., Vanderlinden, L.A., Hoffman, P.L., and Saba, L.M. (2019). Networking in Biology: The Hybrid Rat Diversity Panel. In *Rat Genomics*, G. Hayman, J. Smith, M. Dwinell, and M. Shimoyama, eds. (Humana), pp. 213–231.
- Hansen, C., and Spuhler, K. (1984). Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcohol Clin. Exp. Res.* 8, 477–479.
- Rat Genome Sequencing and Mapping Consortium (2013). Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nat. Genet.* 45, 767–775.
- Woods, L.C.S., and Mott, R. (2017). Heterogeneous Stock Populations for Analysis of Complex Traits. In *Systems Genetics*, K. Schughart and R. Williams, eds. (Humana Press), pp. 31–44.
- Solberg Woods, L.C., and Palmer, A.A. (2019). Using Heterogeneous Stocks for Fine-Mapping Genetically Complex Traits. In *Rat Genomics*, G. Hayman, J. Smith, M. Dwinell, and M. Shimoyama, eds. (Humana), pp. 233–247.
- Chitre, A.S., Poleskaya, O., Holl, K., Gao, J., Cheng, R., Bimschleger, H., Garcia Martinez, A., George, T., Gileta, A.F., Han, W., et al. (2020). Genome-Wide Association Study in 3,173 Outbred Rats Identifies Multiple Loci for Body Weight, Adiposity, and Fasting Glucose. *Obesity* 28, 1964–1973.
- Gunturkun, M.H., Wang, T., Chitre, A.S., Garcia Martinez, A., Holl, K., St Pierre, C., Bimschleger, H., Gao, J., Cheng, R., Poleskaya, O., et al. (2022). Genome-Wide Association Study on Three Behaviors Tested in an Open Field in Heterogeneous Stock Rats Identifies Multiple Loci Implicated in Psychiatric Disorders. *Front. Psychiatry* 13, 790566.
- Kaldunski, M.L., Smith, J.R., Hayman, G.T., Brodie, K., De Pons, J.L., Demos, W.M., Gibson, A.C., Hill, M.L., Hoffman, M.J., Lamers, L., et al. (2022). The Rat Genome Database (RGD) facilitates genomic and phenotypic data integration across multiple species for biomedical research. *Mamm. Genome* 33, 66–80.
- Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493–521.
- Worley, K.C., Weinstock, G.M., and Gibbs, R.A. (2008). Rats in the genomic era. *Physiol. Genomics* 32, 273–282.
- Twigger, S.N., Pruitt, K.D., Fernández-Suárez, X.M., Karolchik, D., Worley, K.C., Maglott, D.R., Brown, G., Weinstock, G., Gibbs, R.A., Kent, J., et al. (2008). What everybody should know about the rat genome and its online resources. *Nat. Genet.* 40, 523–527.
- van Heesch, S., Kloosterman, W.P., Lansu, N., Ruzius, F.-P., Levandowsky, E., Lee, C.C., Zhou, S., Goldstein, S., Schwartz, D.C., Harkins, T.T., et al. (2013). Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genom.* 14, 257.
- Tutaj, M., Smith, J.R., and Bolton, E.R. (2019). Rat Genome Assemblies, Annotation, and Variant Repository. In *Rat Genomics*, G. Hayman, J. Smith, M. Dwinell, and M. Shimoyama, eds. (Humana), pp. 43–70.
- Ramdas, S., Ozel, A.B., Treutelaar, M.K., Holl, K., Mandel, M., Woods, L.C.S., and Li, J.Z. (2019). Extended regions of suspected mis-assembly in the rat reference genome. *Sci. Data* 6, 39.

23. Howe, K., Dwinell, M., Shimoyama, M., Corton, C., Betteridge, E., Dove, A., Quail, M.A., Smith, M., Saba, L., Williams, R.W., et al. (2021). The genome sequence of the Norway rat, *Rattus norvegicus* Berkenhout 1769. *Wellcome Open Res.* 6, 118.
24. Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., and Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. *GigaScience* 10, g1aa153. <https://doi.org/10.1093/gigascience/g1aa153>.
25. Mikheenko, A., Pribelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34, i142–i150.
26. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Djimco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. <https://doi.org/10.1038/nbt.4235>.
27. Yun, T., Li, H., Chang, P.-C., Lin, M.F., Carroll, A., and McLean, C.Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 36, 5582–5589. <https://doi.org/10.1093/bioinformatics/btaa1081>.
28. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 38, 4647–4654.
29. Zhou, X., Li, R., Michal, J.J., Wu, X.-L., Liu, Z., Zhao, H., Xia, Y., Du, W., Wildung, M.R., Pouchnik, D.J., et al. (2016). Accurate Profiling of Gene Expression and Alternative Polyadenylation with Whole Transcriptome Termini Site Sequencing (WTTs-Seq). *Genetics* 203, 683–697.
30. Munro, D., Wang, T., Chitre, A.S., Poleskaya, O., Ehsan, N., Gao, J., Gusev, A., Woods, L.C.S., Saba, L.M., Chen, H., et al. (2022). The regulatory landscape of multiple brain regions in outbred heterogeneous stock rats. *Nucleic Acids Res.* 50, 10882–10895. <https://doi.org/10.1093/nar/gkac912>.
31. Martín-Gálvez, D., Dunoyer de Segonzac, D., Ma, M.C.J., Kwitek, A.E., Thybert, D., and Flicek, P. (2017). Genome variation and conserved regulation identify genomic regions responsible for strain specific phenotypes in rat. *BMC Genom.* 18, 986.
32. Ma, M.C.J., Atanur, S.S., Aitman, T.J., and Kwitek, A.E. (2014). Genomic structure of nucleotide diversity among Lyon rat models of metabolic syndrome. *BMC Genom.* 15, 197.
33. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Solis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012.
34. Koren, S., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. (2019). Reply to “Errors in long-read assemblies can critically affect protein prediction.”. *Nat. Biotechnol.* 37, 127–128.
35. Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 37, 124–126.
36. Sacristán-Horcajada, E., González-de la Fuente, S., Peiró-Pastor, R., Carrasco-Ramiro, F., Amils, R., Requena, J.M., Berenguer, J., and Aguado, B. (2021). ARAMIS: From systematic errors of NGS long reads to accurate assemblies. *Brief. Bioinform.* 22, bbab170. <https://doi.org/10.1093/bib/bbab170>.
37. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Functamman, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746.
38. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O’Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054.
39. de Jong, T.V., Chen, H., Brashear, W.A., Kochan, K.J., Hillhouse, A.E., Zhu, Y., Dhande, I.S., Hudson, E.A., Sumlut, M.H., Smith, M.L., et al. (2022). mRatBN7.2: familiar and unfamiliar features of a new rat genome reference assembly. *Physiol. Genomics* 54, 251–260.
40. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53.
41. Atanur, S.S., Diaz, A.G., Maratou, K., Sarkis, A., Rotival, M., Game, L., Tschannen, M.R., Kaisaki, P.J., Otto, G.W., Ma, M.C.J., et al. (2013). Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell* 154, 691–703.
42. Hermsen, R., de Ligt, J., Spee, W., Blokzijl, F., Schäfer, S., Adami, E., Boymans, S., Flink, S., van Boxtel, R., van der Weide, R.H., et al. (2015). Genomic landscape of rat strain and substrain variation. *BMC Genom.* 16, 357.
43. Supernat, A., Vidarsson, O.V., Steen, V.M., and Stokowy, T. (2018). Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.* 8, 17851.
44. Ashbrook, D.G., Arends, D., Prins, P., Mulligan, M.K., Roy, S., Williams, E.G., Lutz, C.M., Valenzuela, A., Bohl, C.J., Ingels, J.F., et al. (2021). A platform for experimental precision medicine: The extended BXD mouse family. *Cell Syst.* 12, 235–247.e9.
45. Ashbrook, D.G., Sasani, T., Maksimov, M., Gunturkun, M.H., Ma, N., Villani, F., Ren, Y., Rothschild, D., Chen, H., Lu, L., et al. (2022). Private and sub-family specific mutations of founder haplotypes in the BXD family reveal phenotypic consequences relevant to health and disease. Preprint at bioRxiv. <https://doi.org/10.1101/2022.04.21.489063>.
46. Pattee, J., Vanderlinden, L.A., Mahaffey, S., Hoffman, P., Tabakoff, B., and Saba, L.M. (2022). Evaluation and characterization of expression quantitative trait analysis methods in the Hybrid Rat Diversity Panel. *Front. Genet.* 13, 947423.
47. Senko, A.N., Overall, R.W., Silhavy, J., Mlejnek, P., Malínská, H., Hüttl, M., Marková, I., Fabel, K.S., Lu, L., Stuchlik, A., et al. (2022). Systems genetics in the rat HXB/BXH family identifies Tti2 as a pleiotropic quantitative trait gene for adult hippocampal neurogenesis and serum glucose. *PLoS Genet.* 18, e1009638.
48. Bryant, C.D., Smith, D.J., Kantak, K.M., Nowak, T.S., Jr., Williams, R.W., Damaj, M.I., Redei, E.E., Chen, H., and Mulligan, M.K. (2020). Facilitating Complex Trait Analysis via Reduced Complexity Crosses. *Trends Genet.* 36, 549–562.
49. Witte, F., Ruiz-Orera, J., Mattioli, C.C., Blachut, S., Adami, E., Schulz, J.F., Schneider-Lunitz, V., Hummel, O., Patone, G., Mücke, M.B., et al. (2021). A trans locus causes a ribosomopathy in hypertrophic hearts that affects mRNA translation in a protein length-dependent fashion. *Genome Biol.* 22, 191.
50. Pravenec, M., Kožich, V., Krijt, J., Sokolová, J., Zidek, V., Landa, V., Mlejnek, P., Šilhavý, J., Šimáková, M., Škop, V., et al. (2016). Genetic Variation in Renal Expression of Folate Receptor 1 (Folr1) Gene Predisposes Spontaneously Hypertensive Rats to Metabolic Syndrome. *Hypertension* 67, 335–341.
51. Pravenec, M., Churchill, P.C., Churchill, M.C., Viklicky, O., Kazdova, L., Aitman, T.J., Petretto, E., Hubner, N., Wallace, C.A., Zimdahl, H., et al. (2008). Identification of renal Cd36 as a determinant of blood pressure and risk for hypertension. *Nat. Genet.* 40, 952–954.
52. Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S.R., Bauerfeind, A., et al. (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* 467, 460–464.
53. Lu, L.M., Shisa, H., Tanuma, J., and Hiai, H. (1999). PropylNitrosourea-induced T-lymphomas in LEXF RI strains of rats: genetic analysis. *Br. J. Cancer* 80, 855–861.
54. Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005). Integrated

transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* 37, 243–253.

55. Holl, K., He, H., Wedemeyer, M., Clopton, L., Wert, S., Meckes, J.K., Cheng, R., Kastner, A., Palmer, A.A., Redei, E.E., and Solberg Woods, L.C. (2018). Heterogeneous stock rats: a model to study the genetics of despair-like behavior in adolescence. *Genes Brain Behav.* 17, 139–148.
56. Keele, G.R., Prokop, J.W., He, H., Holl, K., Littrell, J., Deal, A., Francic, S., Cui, L., Gatti, D.M., Broman, K.W., et al. (2018). Genetic Fine-Mapping and Identification of Candidate Genes and Variants for Adiposity Traits in Outbred Rats. *Obesity* 26, 213–222.
57. Solberg Woods, L.C., Holl, K.L., Oprea, D., Xie, Y., Tsaih, S.-W., and Valdar, W. (2012). Fine-mapping diabetes-related traits, including insulin resistance, in heterogeneous stock rats. *Physiol. Genomics* 44, 1013–1026.
58. Lemen, P.M., Hatoum, A.S., Dickson, P.E., Mittleman, G., Agrawal, A., Reiner, B.C., Berrettini, W., Ashbrook, D., Gunturkun, H., Mulligan, M.K., et al. (2022). Opiate responses are controlled by interactions of *Oprm1* and *Fgf12* loci in the murine BXD family: Correspondence to human GWAS finding. Preprint at bioRxiv. <https://doi.org/10.1101/2022.03.11.483993>.
59. Lusi, A.J., Seldin, M.M., Allayee, H., Bennett, B.J., Civelek, M., Davis, R.C., Eskin, E., Farber, C.R., Hui, S., Mehrabian, M., et al. (2016). The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits. *J. Lipid Res.* 57, 925–942.
60. Ulusoy, S., Ozkan, G., Alkanat, M., Mungan, S., Yuluğ, E., and Orem, A. (2013). Perspective on rhabdomyolysis-induced acute kidney injury and new treatment options. *Am. J. Nephrol.* 38, 368–378.
61. Kang, K., Nan, C., Fei, D., Meng, X., Liu, W., Zhang, W., Jiang, L., Zhao, M., Pan, S., and Zhao, M. (2013). Heme oxygenase 1 modulates thrombomodulin and endothelial protein C receptor levels to attenuate septic kidney injury. *Shock* 40, 136–143.
62. Yuan, G., Deng, J., Wang, T., Zhao, C., Xu, X., Wang, P., Voltz, J.W., Edin, M.L., Xiao, X., Chao, L., et al. (2007). Tissue kallikrein reverses insulin resistance and attenuates nephropathy in diabetic rats by activation of phosphatidylinositol 3-kinase/protein kinase B and adenosine 5'-monophosphate-activated protein kinase signaling pathways. *Endocrinology* 148, 2016–2026.
63. Osipova, E., Barsacchi, R., Brown, T., Sadanandan, K., Gaede, A.H., Monte, A., Jarrells, J., Moebius, C., Pippel, M., Altshuler, D.L., et al. (2023). Loss of a gluconeogenic muscle enzyme contributed to adaptive metabolic traits in hummingbirds. *Science* 379, 185–190.
64. Obeidat, M. 'en, Zhou, G., Li, X., Hansel, N.N., Rafaels, N., Mathias, R., Ruczinski, I., Beaty, T.H., Barnes, K.C., Paré, P.D., and Sin, D.D. (2018). The genetics of smoking in individuals with chronic obstructive pulmonary disease. *Respir. Res.* 19, 59.
65. Canzian, F. (1997). Phylogenetics of the laboratory rat *Rattus norvegicus*. *Genome Res.* 7, 262–267.
66. Mashimo, T., Voigt, B., Tsurumi, T., Naoi, K., Nakanishi, S., Yamasaki, K.-I., Kuramoto, T., and Serikawa, T. (2006). A set of highly informative rat simple sequence length polymorphism (SSLP) markers and genetically defined rat strains. *BMC Genet.* 7, 19.
67. Chen, N.-C., Solomon, B., Mun, T., Iyer, S., and Langmead, B. (2021). Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* 22, 8.
68. Munger, S.C., Raghupathy, N., Choi, K., Simons, A.K., Gatti, D.M., Hinerfeld, D.A., Svenson, K.L., Keller, M.P., Attie, A.D., Hibbs, M.A., et al. (2014). RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics* 198, 59–73.
69. Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., et al. (2020). Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* 21, 139–162.
70. Villani, F., Guarracino, A., Ward, R.R., Green, T., Emms, M., Pravenec, M., Prins, P., Garrison, E., Williams, R.W., Chen, H., and Colonna, V. (2024). Pangenome reconstruction in rats enhances genotype-phenotype mapping and novel variant discovery. Preprint at bioRxiv. <https://doi.org/10.1101/2024.01.10.575041>.
71. Kalbfleisch, T.S., Hussien AbouEl Ela, N.A., Li, K., Brashear, W.A., Kochan, K.J., Hillhouse, A.E., Zhu, Y., Dhande, I.S., Kline, E.J., Hudson, E.A., et al. (2023). The Assembled Genome of the Stroke-Prone Spontaneously Hypertensive Rat. *Hypertension* 80, 138–146.
72. Rastas, P. (2020). Lep-Anchor: automated construction of linkage map anchored haploid genomes. *Bioinformatics* 36, 2359–2364.
73. Kivikoski, M., Rastas, P., Löytynoja, A., and Merilä, J. (2021). Automated improvement of stickleback reference genome assemblies with Lep-Anchor software. *Mol. Ecol. Resour.* 21, 2166–2176.
74. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
75. Duttke, S.H., Montilla-Perez, P., Chang, M.W., Li, H., Chen, H., Carrette, L.L.G., de Guglielmo, G., George, O., Palmer, A.A., Benner, C., and Telese, F. (2022). Glucocorticoid Receptor-Regulated Enhancers Play a Central Role in the Gene Regulatory Networks Underlying Drug Addiction. *Front. Neurosci.* 16, 858427.
76. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciuffo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745.
77. Kailkar, S., Jain, C., Vasimuddin, M., and Misra, S. (2022). Accelerating minimap2 for long-read sequencing applications on modern CPUs. *Nat. Comput. Sci.* 2, 78–83.
78. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595.
79. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. <https://doi.org/10.1101/201178>.
80. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100.
81. Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8, 14061.
82. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92.
83. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 25, 4.10.1–4.10.14.
84. Gunturkun, M.H., Flashner, E., Wang, T., Mulligan, M.K., Williams, R.W., Prins, P., and Chen, H. (2022). GeneCup: mining PubMed and GWAS catalog for gene-keyword relationships. *G3 (Bethesda)*. 12, jkac059.
85. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34, D590–D598.
86. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association

- and population-based linkage analyses. *Am. J. Hum. Genet.* *81*, 559–575.
87. Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* *33*, 3726–3732.
  88. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* *10*, giab008. <https://doi.org/10.1093/gigascience/giab008>.
  89. Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* *38*, 3022–3027.
  90. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* *8*, 28–36.
  91. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15–21.
  92. Duttke, S.H., Chang, M.W., Heinz, S., and Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* *29*, 1836–1846.
  93. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermizakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* *8*, 15452.
  94. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* *47*, D506–D515.
  95. Gileta, A.F., Gao, J., Chitre, A.S., Bimschleger, H.V., St Pierre, C.L., Gopalakrishnan, S., and Palmer, A.A. (2020). Adapting Genotyping-by-Sequencing and Variant Calling for Heterogeneous Stock Rats. *G3 (Bethesda)* *10*, 2195–2205.
  96. Zhou, J.L., de Guglielmo, G., Ho, A.J., Kallupi, M., Li, H.-R., Chitre, A.S., Carrette, L.L.G., George, O., Palmer, A.A., McVicker, G., and Telese, F. (2022). Cocaine addiction-like behaviors are associated with long-term changes in gene regulation, energy metabolism, and GABAergic inhibition within the amygdala. Preprint at bioRxiv. <https://doi.org/10.1101/2022.09.08.506493>.
  97. Jeffs, B., Negrin, C.D., Graham, D., Clark, J.S., Anderson, N.H., Gauguier, D., and Dominiczak, A.F. (2000). Applicability of a “speed” congenic strategy to dissect blood pressure quantitative trait loci on rat chromosome 2. *Hypertension* *35*, 179–187.
  98. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Girón, C., Hourlier, T., et al. (2016). The Ensembl Gene Annotation System. *Database* *2016*, baw093. <https://doi.org/10.1093/database/baw093>.
  99. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* *215*, 403–410.
  100. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* *47*, D155–D162.
  101. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA Websuite. *Nucleic Acids Res.* *36*, W70–W74.
  102. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* *46*, D335–D342.
  103. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* *29*, 2933–2935.
  104. McGarvey, K.M., Goldfarb, T., Cox, E., Farrell, C.M., Gupta, T., Joardar, V.S., Kodali, V.K., Murphy, M.R., O’Leary, N.A., Pujar, S., et al. (2015). Mouse genome annotation by the RefSeq project. *Mamm. Genome* *26*, 379–390.

## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Deposited data</b>		
Structural variants in HXB RI panel (VCF)	This paper	<a href="https://doi.org/10.5281/zenodo.10398554">https://doi.org/10.5281/zenodo.10398554</a>
Join calling of 163 WGS rat samples (VCF)	This paper	<a href="https://doi.org/10.5281/zenodo.10398344">https://doi.org/10.5281/zenodo.10398344</a>
Transcription start site for 2 brain regions	This paper	<a href="https://doi.org/10.5281/zenodo.10398387">https://doi.org/10.5281/zenodo.10398387</a>
Alternative polyadenylation sites	This paper	<a href="https://doi.org/10.5281/zenodo.10398476">https://doi.org/10.5281/zenodo.10398476</a>
Potential misassembled regions	This paper	<a href="https://doi.org/10.5281/zenodo.10428255">https://doi.org/10.5281/zenodo.10428255</a>
Chromosomal dot plots between Rnor_6.0 and mRatBN7.2	This paper	<a href="https://doi.org/10.5281/zenodo.10515796">https://doi.org/10.5281/zenodo.10515796</a>
The order of genetic markers and the distances from a rat genetic map compared to their locations in Rnor_6.0	This paper	<a href="https://doi.org/10.5281/zenodo.10552387">https://doi.org/10.5281/zenodo.10552387</a>
The order of genetic markers and the distances from a rat genetic map compared to their locations in mRatBN7.2	This paper	<a href="https://doi.org/10.5281/zenodo.10552453">https://doi.org/10.5281/zenodo.10552453</a>
Code and data for phylogenetic analysis	This paper	<a href="https://doi.org/10.5281/zenodo.10520242">https://doi.org/10.5281/zenodo.10520242</a>
nucleus accumbens core eQTL dataset	Munro et al. <sup>30</sup>	N/Acc, 75 samples
csRNA-seq data	Duttke et al. <sup>75</sup>	N/A
Annotation of the mRatBN7.2 assembly	O'Leary et al. <sup>76</sup>	Release 108
Code and computational notebook	This paper	<a href="https://doi.org/10.5281/zenodo.10552887">https://doi.org/10.5281/zenodo.10552887</a>
<b>Experimental models: Organisms/strains</b>		
Sample metadata for RatCollection	This paper	<a href="https://doi.org/10.5281/zenodo.10552790">https://doi.org/10.5281/zenodo.10552790</a>
<b>Software and algorithms</b>		
QUAST	Mikheenko et al. <sup>25</sup>	N/A
paftools	Kalibar et al. <sup>77</sup>	N/A
BWA mem	Li and Durbin <sup>78</sup>	N/A
GATK	Poplin et al. <sup>79</sup>	N/A
LongRanger		N/A
Minimap2	Li <sup>80</sup>	N/A
Deepvariant	Poplin et al. <sup>26</sup>	N/A
GLNexus	Yun et al. <sup>27</sup>	N/A
SURVIVOR	Jeffares et al. <sup>81</sup>	N/A
SnEff	Cingolani et al. <sup>82</sup>	N/A
RepeatMasker	Tarailo-Graovac and Chen <sup>83</sup>	N/A
GeneCup	Gunturkun et al. <sup>84</sup>	N/A
UCSC Liftover tool	Hinrichs et al. <sup>85</sup>	N/A
PLINK	Purcell et al. <sup>86</sup>	N/A
Lep-MAP3	Rastas <sup>87</sup>	N/A
bcftools	Danecek et al. <sup>88</sup>	N/A
MEGA	Tamura et al. <sup>89</sup>	N/A
ggtree	Yu et al. <sup>90</sup>	N/A
STAR aligner	Dobin et al. <sup>91</sup>	N/A
HOMER's findPeaks tool	Duttke et al. <sup>92</sup>	N/A
QTLtools	Delaneau et al. <sup>93</sup>	N/A
UniProt	UniProt Consortium <sup>94</sup>	N/A



## 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](mailto: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 QUAST<sup>25</sup> 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.<sup>23</sup> 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*.<sup>77</sup>

### 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.<sup>78</sup> GATK<sup>79</sup> 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,<sup>80</sup> Deepvariant<sup>26</sup> (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.<sup>27</sup> Large SVs detected by LongRanger were merged using SURVIVOR<sup>31</sup> 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,<sup>82</sup> using RefSeq annotations. Overlap between SNPs and repetitive regions were identified with RepeatMasker.<sup>83</sup> Disease ontology was retrieved from the Rat Genome Database.<sup>16</sup> 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.<sup>84</sup>

### 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/NCrPrin\_BT.ILM, WKY/NHsd\_TA.ILM, GK/Ox\_TA.ILM, FHL/EurMcowi\_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 arrayCGH 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  $\log_2(2)$  to  $\log_2(14)$ , are “segmented” by using the `segment` command in R package “DNAcopy,” 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  $\log_2(x+2)$  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 Lifter

The Lifter 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 Lifter process, we compared the variants obtained through the Lifter 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 against each respective genome. We then used the UCSC Lifter tool<sup>85</sup> 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\_155269102). While the original progenitor population is no longer alive, we obtained DNA samples that were preserved in 1984, when the HS colony was created. Using the 163-by-163 IBS matrix previously generated (see method: tree generation). We identified six closely-related substrains of the progenitors that were over 99.5% similar to the original strains based on identity by state (IBS): ACI/EurMcwi, BN/NHsdMcwi, F344/DuCrI, M520/NRrcMcwi, MR/NRrc WKY/NHsd. The best matches of the remaining strains were less similar: BUF/Mna (73.6%) for BUF/N and WAG/RijCrI (72.0%) for WN/N. Better alternatives for these two strains may be identified in the future as more inbred rat strains are sequenced. At the time of this report, all 8 of these inbred strains are available from Hybrid Rat Diversity Program at the Medical College of Wisconsin (see [https://rgd.mcw.edu/wg/hrdp\\_panel/](https://rgd.mcw.edu/wg/hrdp_panel/) for strain availability and contact details).

### 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.<sup>15</sup> Briefly, genotypes were determined using genotyping-by-sequencing.<sup>95</sup> 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 monomorphic SNPs. Genotypes with high Mendelian inheritance error rates (>2% error across the cohort) were identified by PLINK<sup>36</sup> 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)<sup>87</sup> 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<sup>88</sup> to filter for high-quality, bi-allelic SNP sites from the VCF files for the 20 autosomes. We then employed PLINK<sup>86</sup> 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<sup>89</sup> to construct a distance-based UPGMA tree using the meg file 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<sup>90</sup> 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.<sup>30</sup> Brains were extracted from 88 HS rats. Rats were housed under standard laboratory conditions. Rat brains were extracted and cryosectioned into 60  $\mu\text{m}$  sections, which were mounted onto RNase-free glass slides. Slides were stored in  $-80^{\circ}\text{C}$  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.<sup>91</sup>

### eQTL relocation analysis

We obtained the nucleus accumbens core (NAcc, 75 samples) eQTL dataset from RatGTEx<sup>30</sup> (<https://ratgtex.org>), which was mapped using Rnor\_6.0. We considered associations with  $p < 1 \times 10^{-8}$  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<sup>80</sup> 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.<sup>75</sup> Briefly, small RNAs of  $\sim 15$ –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<sup>91</sup> aligner with default parameters. Transcriptional start regions were defined using HOMER's findPeaks tool.<sup>92</sup>

### 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.<sup>96</sup> 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<sup>29</sup>) 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/LC-MS/MS). We used the QTLtools program<sup>93</sup> for protein expression quantitative trait locus (pQTL) mapping. *cis*-pQTL are defined when the transcriptional start sites for the tested protein are located within  $\pm 1$  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/Stm\_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/Stm\_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/Gla\_TA.ILM is a sample we downloaded from SRA. This WKY substrain (WKY/Gla) clusters closer to SHR strains than other WKY substrains. The same pattern was also observed in one prior study,<sup>41</sup> 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/Gla was derived.<sup>97</sup> We think this sample is mislabeled, but we don't know what the correct label should be.
- (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 substrain 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\_MD.ILM & BXH2\_HCRW.CRM: unexpected low IBS at 92%
  - LEXF4\_MD.ILM & LEXF5\_HCJL.CRM: unexpected high IBS at 99%
  - LEXF3\_MD.ILM & LEXF4\_HCJL.CRM: unexpected high IBS at 99%
  - LEXF1A\_MD.ILM & LEXF1C\_HCJL.CRM: unexpected high IBS at 99%
  - LEXF1C\_MD.ILM & LEXF2A\_HCJL.CRM: unexpected high IBS at 99%
  - LEXF2B\_MD.ILM & LEXF1A\_HCJL.CRM: unexpected high IBS at 98%
  - LEXF4\_MD.ILM & LEXF4\_HCJL.CRM: unexpected low IBS at 83%
  - LEXF1A\_MD.ILM & LEXF1A\_HCJL.CRM: unexpected low IBS at 84%
  - LEXF1C\_MD.ILM & LEXF1C\_HCJL.CRM: unexpected low IBS at 95%

### Ensembl annotation

Annotation of the assembly was created via the Ensembl gene annotation system.<sup>98</sup> 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.<sup>94</sup> 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.<sup>76</sup>

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 lncRNA. Potential lncRNAs 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<sup>99</sup> of miRbase<sup>100</sup> against the genome, before passing the results into RNAfold.<sup>101</sup> Poor quality and repeat-ridden alignments were discarded. Other types of small non-coding genes were annotated by scanning Rfam<sup>102</sup> against the genome and passing the results into Infernal.<sup>103</sup>

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<sup>76</sup> using NCBI's Eukaryotic Genome Annotation Pipeline.<sup>104</sup> 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\_ accessions. 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 accessions 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.