bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Tristan V de Jong, [1†] Yanchao Pan, [2†] Pasi Rastas, [3] Daniel Munro, [4, 5] Monika Tutaj, [6, 7] Huda Akil, [8] Chris Benner, [9] Apurva S Chitre, [10] William Chow, [11] Vincenza Colonna, [12] Clifton L Dalgard, [13] Wendy M Demos, [6, 7] Peter A Doris, [14] Erik Garrison, [12] Aron Geurts, [15] Hakan M Gunturkun, [1] Victor Guryev, [16] Thibaut Hourlier, [17] Kerstin Howe, [18] Jun Huang, [1] Ted Kalbfleisch, [19] Panjun Kim, [12] Ling Li, [20] Spencer Mahaffey, [21] Fergal J Martin, [17] Pejman Mohammadi, [22] Ayse Bilge Ozel, [2] Oksana Polesskaya, [23] Michal Pravenec, [24] Pjotr Prins, [12] Jonathan Sebat, [23] Jennifer R Smith, [6, 7] Leah C Solberg Woods, [25] Boris Tabakoff, [26] Alan Tracey, [11] Marcela Uliano-Silva, [11] Flavia Villani, [12] Hongyang Wang, [27] Burt M Sharp, [12] Francesca Telese, [10] Zhihua Jiang, [27] Laura Saba, [21] Xusheng Wang, [12] Terence D Murphy, [28] Abraham A Palmer, [23] Anne E Kwitek, [6, 7] Melinda (Mindy) R Dwinell, [6] Robert W Williams, [12] Jun Z Li, [2‡] Hao Chen [1‡]

1 Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center, 2 Department of Human Genetics, University of Michigan, 3 Institute of Biotechnology, University of Helsinki, 4 Department of Psychiatry, University of California San Diego, Scripps Research, 5 Department of Integrative Structural and Computational Biology, Scripps Research, 6 Department of Physiology, Medical College of Wisconsin, 7 Rat Genome Database, Medical College of Wisconsin, 8 Michigan Neuroscience Institute, University of Michigan, 9 Department of Medicine, University of California San Diego, 10 Department of Psychiatry, University of California San Diego, 11 Tree of Life, Wellcome Sanger Institute, 12 Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, 13 Department of Anatomy, Physiology & Genetics; The American Genome Center, Uniformed Services University of the Health Sciences, 14 The Brown Foundation Institute Of Molecular Medicine, Center For Human Genetics, University of Texas Health Science Center, 15 Rat Genome Database; Department of Physiology, Medical College of Wisconsin, 16 Genome Structure and Ageing, University of Groningen, UMC Groningen, 17 European Molecular Biology Laboratory, European Bioinformatics Institute, 18 Tree of Life, Wellcome Sanger Institute, Cambridge, UK, Wellcome Sanger Institute, 19 Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, 20 Department of Biology, University of North Dakota, 21 Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, 22 Department of Integrative Structural and Computational Biology, Scripps Research Translational Institute, Scripps Research, 23 Department of Psychiatry , University of California San Diego, 24 Institute of Physiology, Czech Academy of Sciences, 25 Department of Internal Medicine, Section on Molecular Medicine, Wake Forest University School of Medicine, 26 Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, 27 Department of Animal Sciences, Washington State University, 28 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

---

* Dedicated to the memory of Dr. Mary Shimoyama.

† These authors made equal contributions to the work.

‡ Corresponding authors. E-mails: junzli@med.umich.edu hchen@uthsc.edu

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

For over a decade, a large research community has relied on a flawed reference assembly of the genome of *Rattus norvegicus* known as Rnor_6.0. The seventh assembly of the rat reference genome—mRatBN7.2, based on the inbred Brown Norway rat, corrects numerous misplaced segments, reduces base-level errors by approximately 9-fold, and increases contiguity by 290-fold, despite some remaining regions of potential misassembly. Gene annotations are now more complete, significantly improving the mapping precision of genomic, transcriptomic, and proteomics data sets. Simple *LiftOver* from Rnor_6.0 to mRatBN7.2 misses ~12% of variants. To facilitate the transition to mRatBN7.2, we performed a joint analysis of 163 whole genomes representing 120 strains/substrains. We defined 20.0 million sequence variations, of which 18.7 thousand are predicted to potentially impact the function of 6,677 genes. Phylogenetic analysis confirmed historical records and prior results and refined the ancestral relationship of these strains. Sixteen million polymorphisms segregate in the widely studied heterogeneous stock rat population, and 11–13 million variants segregate collectively in the HXB/BXH and FXLE/LEXF strain families. Some inbred strains differ by only 1–2 M variants, and closely related substrains segregate by even fewer variants. We generated a new rat genetic map based on data from 1,893 heterogeneous stock rats and annotated transcription start sites and alternative polyadenylation sites.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Rattus norvegicus was among the first mammalian species used for scientific research. The earliest studies using brown rats appeared in the early 1800s. [1,2] The Wistar rats were bred for scientific research in 1906, and is the ancestor of many laboratory rat strains. [3] Rats have been used as models in many fields of study related to human disease, due to their body sizes large enough for easy phenotyping and showing complex behavioral patterns. [4]

Over 4,000 inbred, outbred, congenic, mutant, and transgenic rat strains have been created and are documented in the Rat Genome Database (RGD). [5] Approximately 500 strains are available from the Rat Resource and Research Center. [6] Several genetic mapping populations are available, such as the HXB/BXH [7] and FXLE/LEXF [8] recombinant inbred (RI) families. Both RI families, together with 30 diverse classical inbred strains, [9] are now part of the Hybrid Rat Diversity Panel (HRDP), which has the potential to generate over 10,000 isogenic and replicable F_1 hybrids. The outbred N/NIH heterogeneous stocks (HS) rats, derived from eight inbred strains, [10] have been increasingly used for fine mapping of physiological and behavioral traits. [11–15] To date, RGD has annotated nearly 2,400 rat quantitative trait loci (QTL), mapped using F2 crosses, RI families, and HS rats.

The *Rattus norvegicus* was sequenced shortly after the genomes of *Homo sapiens* and *Mus musculus*. [16] The inbred Brown Norway (BN/SsNHsdMcwi) strain, derived from a pen-bred colony of wild rats, [3] was used to generate the reference. Several updates were released over the following decade. [17–19] Since 2014 Rnor_6.0 [20] has been the reference for all recent genomic and genetic research and has been a problematic assembly. [21] mRatBN7.2 was created in 2020 by the Darwin Tree of Life/Vertebrate Genome Project (VGP) as the new genome assembly of the BN/NHsdMcwi rat. [22] The genome reference consortium (GRC, https://www.ncbi.nih.gov/grc/rat) has adopted mRatBN7.2 as the official rat reference genome.

We report extensive analyses of the improvements in mRatBN7.2 compared to Rnor_6.0. To assist the rat research community in the transition to mRatBN7.2, we conducted a broad analysis of a whole-genome sequencing (WGS) dataset of 163 samples from 88 inbred strains and 32 substrains. Joint variant calling led to the discovery of 15,804,627 high-quality sites. Additional resources created during our analysis included a rat genetic map with 150,835 binned markers, a comprehensive phylogenetic tree for laboratory rats, and extensive annotation on transcription start sites and alternative polyadenylation sites. Together with the new reference genome,

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these resources create a springboard for future research germane to understanding many dimensions of human behavior and pathobiology.

Results

Evaluating the structural and base-level accuracy of mRatBN7.2

mRatBN7.2 [22] was generated as part of the Darwin Tree of Life/Vertebrate Genome Project. All sequencing data were generated from a male Norway rat (BN/NHsdMcwi, generation F61). The assembly is based on long-read data (PacBio CLR) and integrates data from other technologies (10X linked-reads, BioNano DLS optical map, and Arima HiC). Manual curation corrected some remaining errors. [23] In November 2020, mRatBN7.2 was accepted by the Genome Reference Consortium (GRC) as the official rat reference genome, and the Rat Genome Database (RGD) joined the GRC to oversee its curation.

Over the last six iterations of the rat reference, genome continuity has improved with each update (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 long-read PacBio data and improved contig N50 to 100.5 Kb. mRatBN7.2, primarily based on long-read PacBio data, further improved N50 to 29.2 Mb (Figure S1). Although this lags behind the mouse reference genome (contig N50=59 Mb in GRCm39, released in 2020), and is far behind the first Telomere-to-Telomere human genome (CHM13) (Table 1), it still marks a significant 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 to Rnor_6.0, and is approaching the quality of GRCh38 for humans (57.9 Mb) and GRCm39 (59.5 Mb) for mice.

Comparing Rnor_6.0 vs mRatBN7.2 showed great overall agreement (Figure 1A, Figure S3). However, we identified 36.5 K structural variants (SVs) between these two assemblies (Figure S4). To evaluate these differences, we generated a genetic map using data from 378 families of 1,893 HS rats, each genotyped at 3.5 million sites. [15]. Recombinations in these HS rats are evenly distributed over the autosomes (Table S2). Because marker ordering was independent of the reference genome, this high-resolution map (150,835 binned markers) provided us with independent information regarding the source of these structural discrepancies. For example, a significant 17.2 Mb inversion at proximal Chr 6 between Rnor_6.0 and mRatBN7.2 (Figure 1B) remains when the genetic map is compared to Rnor_6.0 (Figure 1C) but is resolved the genetic map is compared to mRatBN7.2 (Figure 1D). The same pattern was found for most other genomic regions (e.g. Figure 1E-G for Chr 19, and Figure S5

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and S6 for all autosomes). These data indicated that most of the structural differences between the two assemblies are due to errors in Rnor_6.0.

Gene annotations for mRatBN7.2

mRatBN7.2 annotations were released in RefSeq 108 and Ensembl 105 (supplemental methods). Gene annotations for both RefSeq and Ensembl were derived from multiple data sources, including transcriptomic datasets and species-specific sequences. Ensembl performed gap-filling with protein-to-genome alignments using a subset of mammalian protein sequences with experimental evidence that included a lift-over of annotation from the GENCODE GRCm39 mouse gene set. This resulted in a more comprehensive gene set compared to the previous Rnor 6.0 annotation. RefSeq 108 includes partially and fully curated models for nearly 80% of protein-coding genes in addition to computational models.

We analyzed the different annotation sets in RefSeq using BUSCO v4.1.4, [24] using the glires_odb10 dataset of 13,798 models 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 genes as complete on mRatBN7.2 in NCBI RefSeq 108 [Single-copy (S):97.2%, Duplicated (D):1.5, Fragmented (F):0.4%, Missing (M):0.9%]. In comparison, a partial run of Rnor6.0 with 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 version 5.3.2 with the lineage dataset "glires_odb10/2021-02-19". The "Complete and single-copy BUSCOs" 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. Thus, both annotation sets show that mRatBN7.2 is a better foundation for gene annotation, with excellent representation of protein-coding genes and fewer unexpected duplications.

We also compared RefSeq (release 108) and Ensembl (release 107) using their respective GTF files. The RefSeq GTF file annotated 42,167 genes, each associated with a unique NCBI GeneID. These included 22,228 protein-coding genes, 7,888 lncRNA, 1,288 snoRNA, 1,026 snRNA, and 7,972 pseudogenes. The Ensembl GTF file contained 30.1 K 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 3X) and pseudogenes (almost 10X), and only RefSeq annotates tRNA.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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). While many of these were caused by the lack of 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. Therefore, we compared the two GTF files using common gene symbols and found that 18,722 were jointly annotated. Among the 22,247 gene symbols found only in RefSeq, 20,185 were genes without formal names. We further compared transcripts annotated by each source. 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. A histogram of the number of transcripts per gene is provided in Figure S7.

mRatBN7.2 improved the mapping of short-read and long-read WGS data

To assess the overall base-level accuracy of the reference genomes, we mapped 36 linked-read WGS samples against both Rnor_6.0 and mRatBN7.2. These fully inbred samples included four samples of the reference strain, BN/NHsdMcwi, 30 members of the HXB/BXH family, and two rats from their parental strains SHR/OlaIpcv and BN-Lx/Cub. We refer to this set as the *HXB* dataset and use it for comparing various aspects of Rnor_6.0 and mRatBN7.2.

The read depth of the HXB dataset is approximately 60X. Comparing mRatBN7.2 to Rnor_6.0, reads mapped to the reference increased by 1–3% (Figure 2A), and regions of the reference genome with no coverage reduced by ~2% (Figure 2B). Genetic variants (SNPs and indels) were identified using *deepvariant* [25] and jointly called for the 36 samples using *GLnexus.* [26] After quality filtering (qual ≥ 30), 8,286,401 SNPs and 3,527,568 indels were identified in Rnor 6.0, compared to 5,088,144 SNPs and 1,615,870 indels in mRatBN7.2 (Figure 2C, 2D). The distribution of the number of strains that shared a SNP or indel (Figure 2E, 2F) was similar between the two references. However, variants shared by all 36 samples (either homozygous in all samples or heterozygous in all samples) were more abundant in Rnor_6.0 (1,310,902) than mRatBN7.2 (143,254). Because we included 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 indicated base-level errors in the references. Therefore, mRatBN7.2 reduced base-level errors by 9.2 fold compared to Rnor_6.0, but may still contain a significant number of errors. A more detailed analysis of base level errors in mRatBN7.2 is provided below.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We found 19.6 K SVs in the HXB dataset using Rnor_6.0. Among these, 3,146 were also found in more than two BN/NHsdMcwi samples. In contrast, only 5,458 SVs were found using mRatBN7.2, with 260 SVs found in more than two BN/NHsdMcwi samples. Among these SVs, 5,326 deletions were also found using Rnor_6.0, and only 1,045 deletions were reported using mRatBN7.2. We illustrated these findings by focusing on a small panel of six HXB samples as well as SHR and BN samples (Figure S8). The sample with the greatest number of deletions (HXB21) showed a reduction from 1,017 in Rnor_6.0 to 110 in mRatBN7.2; and the cumulative size of all deletions decreased from 141.8 Mbp in Rnor_6.0 to 26.4 Mbp in mRatBN7.2. These results corroborated the findings from the analysis of SNPs; that mRatBN7.2 eliminates many errors in Rnor_6.0. The remaining 5,198 SVs discovered based on mRatBN7.2 are provided as a VCF file (Table S5).

We also compared mapping results of long-read datasets. In a PacBio CLR dataset from an SHR rat with 4,508,208 reads, unmapped reads declined from 298,422 in Rnor_6.0 to 260,947 in mRatBN7.2, a 12.6% reduction. Mapping of Nanopore data from one outbred HS rat (~55X coverage) showed much lower secondary alignment—from ~10 million in Rnor_6.0 to 4 million in mRatBN7.2. Similar to the linked-read data discussed above, structural variants detected in the Nanopore dataset were reduced (Figure S9) using mRatBN7.2, demonstrating the quality of the mRatBN7.2 assembly.

mRatBN7.2 improved the analysis of rat transcriptomic and proteomic data

We analyzed an RNA-seq dataset of 352 HS rat brains (see RatGTEx.org) to compare the effect of the reference genome on the analysis of polyA-selected RNA-seq data. Reads aligned to the reference increased from 97.4% in Rnor_6.0 to 98.3% in mRatBN7.2. The average percent of reads aligned concordantly only once to the reference increased from 89.3 to 94.6%. The average percent 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 (Figure S12), 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We analyzed datasets containing information about transcript start and polyadenylation. In a capped short RNA-seq (csRNA-seq) data set, the unique alignment of transcriptional start sites to the reference genome increased 5% for mRatBN7.2 compared to Rnor_6.0 (Fig. S14B). In this dataset, we identified 42,420 transcription start sites when analyzed using Rnor_6.0, and 44,985 sites were identified on mRatBN7.2 (Table S5). We analyzed 83 whole transcriptome termini site sequencing [27] datasets using total RNA derived from rat brains. Of the total of 312 million reads, 76.97% were mapped against Rnor_6.0, while 80.49% were mapped against mRatBN7.2 (Table S7). We identified 167,136 alternative polyadenylation (APA) sites using Rnor_6.0 and 73,124 APA sites using mRatBN7.2. 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 the 20,102 annotated genes on mRatBN7.2 (Table S5).

We examined the effect of improved reference on the accuracy of expression quantitative trait locus (eQTL) mapping using an existing eQTL dataset for nucleus accumbens core. [28] We identified associations that would be labeled as trans-eQTLs using one reference but as a cis-eQTL using the other reference 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 SNP using Rnor_6.0 that converted to cis-eQTLs using mRatBN7.2 (Figure 4). 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.

We then analyzed a new set of brain proteome data from 29 strains of the HXB/BXH panel. At a protein false discovery rate (FDR) of 1%, we identified and quantified 8,002 unique proteins on Rnor_6.0, compared to 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 genome (Figure 5A). 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 the Rnor_6.0 genome, but as a

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significant cis-pQTL using mRatBN7.2 (*p-value* = 4.12×10^{-6}) (Figure 5B). In addition, the expression of RPA1 protein displayed a high correlation between mRatBN7.2 and in Rnor_6.0 (r = 0.79; p value = 3.7×10^{-7}) (Figure 5C). Lastly, the annotations for RPA1 were different between the two references (Figure 5D).

WGS mapping data suggesting potential errors in mRatBN7.2

To further assess the extent of residual errors in mRatBN7.2, we analyzed a WGS dataset containing 163 inbred rats (described in Methods) mapped against mRatBN7.2. We found 129,186 variants shared by more than 156 samples, a threshold determined from the distribution of the variants throughout the samples (Figure S10). To rule out the possibility that these variants were unique to the reference individual, we also required these non-reference alleles to be found in all seven BN/NHsdMcwi samples, including the linked-read dataset used for generating mRatBN7.2. The high mapping quality of these variants (69.1 ± 13.8) indicates that they are accurate.

These shared variants consisted of 33,550 SNPs and 95,636 indels. Among them,117,901 were homozygous, and 11,285 were heterozygous in more than 156 samples. The read depth was 32.2 ± 10.1 for homozygous and 66.3 ± 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 3). In addition, RepeatMasker masked 38.5% of shared homozygous variants, compared to 61.3% of shared heterozygous variants. LINE, LTR, and SINE were the most common repeat elements overlapping with these variants (Figure S11). These results indicated that many of the potential errors in mRatBN7.2 are caused by the collapse of repetitive sequences.

Functionally, these potential errors impact some well-studied genes, such as *Chat, Egfr, Gabrg2, and Grin2a*. The NCBI RefSeq team has referred most of these errors to the GRC for curation. Genes such as *Akap10, Cacna1c, Crhr1, Gabrg2, Grin2a, Oprm1* have been resolved by GRC curators but have not been incorporated into mRatBN7.2 because they will change the coordinates of the reference. The full list of 129,186 variants (Table S6), and a bed file indicating regions with tentative collapsed repeats (Table S5), represent approximately 0.15% of the rat genome. After removing the likely errors, we annotated the VCF files resulting from joint variant calling of 163 samples (Table S5) and used these for subsequent analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Complexities in transitioning from Rnor_6.0 to mRatBN7.2

Liftover tools can convert genomic coordinates between different versions of assemblies without having to remap raw data. We examined the proportion of variants in Rnor_6.0 that could be lifted to mRatBN7.2 by utilizing a set of variants evenly distributed at 1 kb intervals. We found that 92.1% of the 2.78 M simulated variant sites were successfully lifted. Figure S12 illustrates the distribution of the unliftable and lifted sites, which is consistent with the comparative genomic view shown in Figure S13.

We also evaluated the effectiveness of liftover on real datasets (WKY/N) by comparing a variant set lifted from Rnor_6.0 to mRatBN7.2 against a variant set obtained directly using mRatBN7.2 as the reference. A higher proportion of variants passed the quality filter when mapping to mRatBN7.2 (97.99%) than Rnor_6.0 (83.64%), and 97.93% of the variants were liftable going from Rnor_6.0 to mRatBN7.2 (Figure 8A). For the lifted variant sites, 94.48% had a match in the directly-called set. However, 507.7 K variants (11.91%) reported using mRatBN7.2 directly were absent in the lifted variant set (Figure 8B). Moreover, we noted the existence of several many-to-1 matches going from Rnor_6.0 to mRatBN7.2, which could result in conflicting genotype calls at the same location.

The significant structural improvement in mRatBN7.2 over its predecessor is likely the cause of these liftOver issues (Figure S12, S13). Therefore, despite being time and resource-intensive, complete remapping of the data to the new reference is preferable. To facilitate a smooth transition to mRatBN7.2, we conducted the largest joint analysis of WGS data for laboratory rats.

A comprehensive survey of the genetic diversity of *Rattus norvegicus*

We collected WGS for a panel of 163 rats (new data from 127 rats and 36 datasets downloaded from NIH SRA (hereafter referred to as RatCollection, Table S8). The coverage depth was 60.4 ± 39.3. Additional sample statistics are provided in Table S9. After removing the 129,186 variants that were potential errors in mRatBN7.2 (above) and filtering for quality (qual ≥ 30), 19,987,273 variants (12,661,110 SNPs, 3,805,780 insertions, and 3,514,345 deletions) across 15,804,627 sites were identified (Table 2). The mean variant density was 5.96 ± 2.20/Kb for the entire genome. The highest variant density of 30.5/Kb was found on Chr 4 at 98 Mb (Figure S15). Most (97.9 ± 1.4%) of the variants were homozygous at the sample level, confirming the inbred nature of most strains, with a few exceptions (see Figure S16).

RatCollection includes all 30 strains of the HXB/BXH family, 25 strains in the FXLE/LEXF family, and 33 other inbred strains. In total, we covered 88 strains and 32

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

substrains—approximately 80% of the HRDP. 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 6. This phylogenetic tree is in agreement with previous publications [29,30] and matches strain origins.

Our analysis included all major populations of rats actively used in genetic studies, such as the outbred HS rats generated by interbreeding eight inbred progenitor strains. [10] WGS data from these progenitors (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N, and WN/N) were included in our collection. The number of variants found in each of these strains is shown in Figure 7A; WKY/N contributed the largest number of strain-specific variant sites (Figure 7B). The distribution of the variant sites per chromosome is shown in (Figure 7C). Collectively, these accounted for 78.6% of all variant sites 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 six living proxies to the HS progenitors that were over 99.5% similar to the original strains: ACI/EurMcwi, BN/NHsdMcwi, F344/DuCrl, M520/NRrrcMcwi, MR/NRrrc, and WKY/NHsd (Table S11). Of the remaining strains, the best matches of the remaining strains were much less similar: BUF/Mna was the best approximation (73.6%) to BUF/N, and WAG/RijCrl was the closest (72.0%) to WN/N. These proxies could help HS studies estimate heritability or validate variant effects.

Our analysis also included two families of RI rats. The HXB/BXH family was generated from SHR/OlaIpcv and BN-Lx/Cub. Together with their parental strains, we identified 7,520,223 locations containing 11,256,227 variants in this population (Table 2). Approximately 24.1–53.5% of the variants from each strain were derived from SHR/OlaIpcv, and up to 1.5% of the variants were *de novo*. While the majority of the strains were highly inbred, with close to 98% of the variants being homozygous (Figure S16), one exception was BXH2, in which 7.7% of variants were heterozygous (Figure S17). The LEXF/FXLE RI family was generated from LE/Stm and F344/Stm. We discovered 8.2 M sites containing 13.1 M variants from this family (Table 2). The overall rate of homozygous variants in the FXLE/LEXF family (96.8 ± 2.4%) was lower than in other inbred rats (Figure S16). In particular, 15.6% of the variants from FXLE24 were heterozygous.

In addition to these large panels with divergent genetic differences between individual trains, the RatCollection also included groups of rats generated by selective breeding, such as the Dahl Salt Sensitive (SS) and Dahl Salt Resistant (SR) strains for hypertension. These two strains contain 7,171,447 variant sites compared to

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 for high blood pressure from outbred Sprague-Dawley rats. [31] Only 281,972, 289,112, and 262,574 variants were unique to LH, LL, and LN, respectively. In agreement with a prior report, [32] these variants were clustered in a handful of genomic hotspots (Figure S19).

Impact of variants and modeling 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 were predicted to have a high impact (i.e., have a disruptive impact on the protein, probably causing protein truncation, loss of function, or triggering nonsense-mediated decay, etc.) on 6,667 genes, including 3,930 protein-coding genes. The number of genes affected by variants with moderate or low impact variants, as well as detailed functional categories, are provided in 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 variants in the HXB/BXH, FXLE/LEXF RI panels, and SS/SR, as well as LL/LN/LH selective bred strains, are summarized in Figure S21, Figure S22, Table S13, and Table S14.

We further annotated genes with high-impact variants using the human GWAS catalog. [33] Among the genes with high-impact variants, 2,034 have human orthologs with genome-wide significant hits. 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). These human genes were significantly associated with 1,393 mapped traits (Table S15).

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We systematically evaluated mRatBN7.2, the seventh iteration of the reference genome for *Rattus Norvegicus*, and confirmed that it vastly improved assembly quality compared to its predecessor. Our extensive evaluation showed that mRatBN7.2 improved the analysis of genomic, transcriptomic, and proteomic data compared to Rnor_6.0. To facilitate the transition to mRatBN7.2, we conducted the largest joint analysis of rat WGS data to date, including 163 samples representing 88 strains and 32 substrains, and identified 19,987,273 variants from 15,804,627 sites. Our analysis generated a new rat genetic map with 150,835 binned markers, a comprehensive rat phylogenetic tree, and a collection of transcription start sites and alternative polyadenylation sites for genes expressed in the rat brain.

Previous references employed slightly different methods for calculating assembly statistics. To enable a direct comparison, we used the methods employed by the VGP team [22] to generate comparable statistical measures for all official rat references (Table 1), which indicated significant improvements in mRatBN7.2. Short-read data generated for the reference assembly is often used to evaluate and improve the base-level quality. [34] Our dataset for comparing base-level accuracy between Rnor_6.0 and mRatBN7.2 included 36 WGS samples (including the rats used for Rnor_6.0 and mRatBN7.2 and two additional male BN/NHsdMcwi rats). Using consensus results from a dataset containing multiple samples reduces the chance of falsely identifying errors in the reference. Using this dataset, we found that base-level errors were reduced by 9.2 fold in mRatBN7.2. Many methods have been developed to evaluate the structural accuracy of an assembly using sequencing data (e.g., QUAST-LG, [35] Merqury [36]). Our evaluation used independent information obtained from a rat genetic map. By comparing the physical distance between 151,380 markers and the order of markers on the reference genomes against their genetic distance calculated based on recombination frequency (Figure 1, Figure S5, S6), we confirmed that most structural differences between Rnor_6.0 and mRatBN7.2 are due to errors in Rnor_6.0. Therefore, mRatBN7.2 is an rat reference genome with improved contiguity as well as increased structural and base accuracy.

The rat genetic map with 150,835 markers provides substantial improvement to a similar map generated previously (95,769 markers). [37] By providing the centimorgan distances and the order of these markers to the community, we envision this map will have multiple applications. For example, instead of being used for assembly evaluation, this map can be integrated into the *de novo* assembly process, as demonstrated by the stickleback genome. [38,39] Additionally, Lep-MAP3 [40] imputes segregation for all markers;

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

therefore, the complex patterns of how the families are informative do not complicate QTL mapping. Hence, we expect that QTL regions will be as precisely defined as possible, eliminating the need to impute any pseudo-markers.

Both RefSeq and Ensembl reported improved BUSCO scores from Rnor_6.0 to mRatBN7.2, with a relatively small percentage increase in complete genes. RefSeq (22,228) and Ensembl (23,139) annotations of mRatBN7.2 contain a similar number of protein-coding genes. These numbers are consistent with those found in humans (20,024; RefSeq release 110) and mice (22,186; RefSeq release 109). We note that RefSeq and Ensembl had assigned different names to 2,319 protein-coding genes (Table S4). This underscores the significance of specifying the origin of gene annotation and incorporating stable identifiers in research findings. The average number of transcripts per gene (2.9 for RefSeq and 1.8 for Ensembl) was less than for mouse (3.3 RefSeq release 109) and human (4.2 RefSeq release 110). This species difference is likely an artefact of the annotation process. Additional efforts in annotating gene transcripts are needed for a better understanding of transcript regulation in rats.

We compared the results of several omics data types analyzed with mRatBN7.2 or Rnor_6.0. These include WGS (both short and long reads), RNA-seq (bulk tissue, single nuclei, targeting either 5' transcription start site or 3' polyadenylation), and proteomic data. mRatBN7.2 provided meaningful improvements in most parameters. In addition to the quantitative improvement of most QC parameters (e.g., mapping rates, depth of coverage, etc.), we also noted many qualitative differences. For example, some trans-regulatory relationships in QTL mapping based on Rnor_6.0 were corrected to cis-relationships using mRatBN7.2. These results will have different interpretations. These findings emphasize the importance of using a high-quality reference genome in omic data analysis.

The ENCODE project [41] has systematically mapped functional elements of the human and mouse genomes. [42] In contrast, such data for the rat genome remain very sparse. Our analysis produced a list of TSS locations for genes expressed in the prefrontal cortex or nucleus accumbens (Table S5); APA sites of genes expressed in the brain based on sequencing whole transcriptome termini sites (Table S5). These new data sets will further the study of gene regulatory mechanisms in rats.

The improvements in mRatBN7.2 are based on new assembly methods and long-read data. [22] However, the PacBio CLR reads used in mRatBN7.2 have lower base accuracy than Illumina short reads. [43,44] Although Illumina data were used to polish the assembly, [22] polishing methods do not correct all base-level errors. [43,45] Our joint analysis of 163 WGS datasets identified 129,186 sites that are likely errors in mRatBN7.2. Another

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

source of potential error is the assembly method itself. mRatBN7.2 was assembled with VGP pipeline v1.6., [22,46] which used Falcon and Falcon-unzip to assemble the long reads into a diploid genome containing a primary and an alternative assembly. [47] Haplotypic duplication was identified and removed with purge_dups. [48] This pipeline is well suited to assemble diploid genomes. Yet, BN/NHsdMcwi is fully inbred. The pipeline classified some regions that may contain authentic duplication as haploid variations. [49] In support of this possibility, we found that most heterozygous variants indicative of assembly errors were in regions identified by RepeatMasker (Figure S11). Although additional data can fix some of these errors, the best solution is to create a new telomere-to-telomere (T2T) assembly like the one reported for the human genome. [50] Such an assembly will require new data, including PacBio circular consensus reads (HiFi) for high accuracy (99.9%) mid-range (10-25 kb) reads and ultralong (> 100kb per read) reads from Oxford nanopore instruments to bridge the gaps between HiFi reads.

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 real 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. Although these results are promising, our analysis also revealed that reanalysis of the original sequencing data using mRatBN7.2 discovered 507.7 K novel variants not detected by Liftover. These novel variants were primarily located in regions not present in Rnor_6.0 (Figure S13). Given the substantial improvements in mRatBN7.2, remapping is likely to lead to novel discoveries.

To facilitate the adoption of mRatBN7.2 and assist the research community in transitioning to the new reference, we mapped WGS data from 163 rats to mRatBN7.2. Joint analysis identified 19,987,273 variants at 15,804,627 sites from our RatCollection, representing 88 strains and 32 substrains. This analysis is built on many prior analyses of rat genomes. For example, Baud et al. [11] analyzed 8 strains against Rnor3.4 and found 7,877,000 variants, Atanur et al. [30] analyzed 27 rat strains against Rnor3.4 and reported 13,167,457 variants, Hermsen et al. [51] analyzed 40 strains against Rnor5.0 and reported 12,249,301 variants. Most recently, Ramdas et al. [21] analyzed 8 strains and reported 16,405,184 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. [26,52] Thus, our data provide the most comprehensive analysis of genetic variants in the laboratory rat population to date.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our analysis included the full HXB/BXH panel and 27 strains/substrains of the FXLE/LEXF panel. Together with the inbred strains, they cover about 80% of the rats in the HRDP. [53] Although we do not yet have data from the full FXLE/LEXF panel, we expect the final variant count to be similar because both parental strains are included in our analysis. This large increase in variants in the RI panels will provide a much-needed boost in mapping precision: prior mapping of these RI families was based on several hundred to tens of thousands markers. [8,53,54]

By generating F_1 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 F_1 hybrids with completely defined genomes. Studies of these “sequenced” F_1 hybrids avoid the homozygosity of the parental HRDP strains and enable a new phase of genome-phenome mapping and prediction. [55] The reanalysis of existing phenotype data is yet another application of these sequencing data. Such reanalysis could lead to novel discoveries. [56] While several genetic mapping studies using the HRDP are currently underway, the large number of variants in the HRDP (c.f., the hybrid mouse diversity panel contains about 4 million SNPs [57]) and WGS-based genotype data will further encourage the use of this resource.

Outbred N/NIH HS rats are another widely used rat genetic mapping population. [13] Currently, there are two colonies of the HS rats: one located at Wake Forest University (RRID:RGD_13673907) was moved from the Medical College of Wisconsin (RRID:RGD_2314009); the University of California San Diego houses the other colony (RID:RGD_155269102). Our analysis is in agreement with Ramdas et al. [21] in identifying 16,438,30 variants in the HS progenitors. Updated genetic data will be useful in HS genetic mapping studies, such as for the imputation of variants. Since live colonies of the original NIH HS progenitor strains are no longer available, the analyses above are based on DNA samples that were preserved in 1984, when the HS colony was created. However, we identified 6 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). 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/ for strain availability and contact details).

Our analysis also included several groups of inbred rats selectively bred to express specific phenotypes. These strains, in general, are segregated by less than 2 million

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variants. For example, the LH/LN/LL family members differ by less than 300 thousand variants. [31,32] In the SS/SR family widely used to model cardiac disease, we identified genetic variants influencing genes involved in hypertension, cardiovascular disease, and kidney injury (Table S13), thus providing leads for new research directions.

Closely related inbred rat strains (a.k.a. near isogenic lines) with distinctive phenotypes can be exploited to identify causal variants using reduced complexity crosses. [58] In addition to the WMI/Eer vs WLI/Eer, [59] and LEW/NHsd vs LEW/NCrl substrains, [60] our analysis contains many other pairs that could be used for reduced complexity crosses, such as SHR/NCrl vs SHR/NHsd (differ by 269,936 SNPs) and the various F344 substrains (differ by 287,313 SNPs between the most distant pairs).

We used SnpEff to evaluate the potential functional consequences of these variants. 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 S12). While the majority of these predictions have not been verified experimentally, some are strongly supported by the literature. For example, *Klk1c12* is associated with hypertension, [61] while *Capn1* [62] and *Procr* [63] are both associated with kidney injuries. High-impact variants of these genes were found in SS rats, which develop renal lesions with hypertension. Most interestingly, 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 (LoF). Once confirmed, strains harboring these variants can be used to investigate the function of these genes, as exemplified by recent work on the LoF of FBP2. [64] Such investigations are highly relevant when the human orthologue of these genes is associated with certain traits in human GWAS (Table S15). For example, *CDHR3* is associated with smoking cessation. [65] 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.

Our phylogenetic analysis agrees with those published previously [29,30] and is consistent with the known derivation history. The inclusion of multiple individuals from the same strain facilitated the identification of mislabeled samples and discrepancies between sample metadata and the genetic relationships inferred from sequencing data (see Methods). While some conflicts can be corrected with confidence, others may be caused by a more complex sequence of events, possibly involving breeding errors (e.g. BXH2). A notable finding, also reported in a previous study, [30] is the clustering of one WKY substrain (WKY/Gla) closer to SHR strains than to WKY substrains. [30] Regional similarity analysis showed that the pattern observed is consistent with a congenic strain

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

created from the recipient SHR and the donor WKY, which occurred at the institute where WKY/Gla was derived. [66]

In agreement with previous publications, our phylogenetic analysis showed that BN/NHsdMcwi, the reference strain for *Rattus norvegicus*, is an outgroup to all other common rat strains (Figure 6). This is consistent with its derivation from a pen-bred colony of wild-caught rats. [3] Thus, mapping sequence data from other strains to the BN reference yields lower mapping quality but a greater number of variants, than using a hypothetical reference that is genetically closer to the commonly used strains. This so-called reference bias has been observed in genomic [67] and transcriptomic data analyses. [68] While alternative strains can be selected, it should be noted that no individual strain is a perfect representation of a population. Instead, the nascent field of pangenomics, [69] where the genome of all strains can be directly compared to each other, provides a promising future where all variants can be compared between individuals directly without the use of a single reference genome. This pangenomic approach will be especially powerful when individual genomes are all assembled from long-read sequence data, some of which are already available. [70] This will enable a complete catalog of all genomic variants, including SVs and repeats, that differ between individuals.

Research using *Rattus norvegicus* has made significant contributions to the understanding of human physiology and diseases. An updated reference genome provides a much-needed road map for using laboratory rats to understand the genetics of human disease. The rich literature on physiology, behavior, and disease models in laboratory rats combined with the complex genomic landscape revealed in our survey demonstrates that the rat is a superb model organism for the next chapter of biomedical research.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Data availability. The RatCollection contains 163 WGS samples from 88 strains and 32 substrains. It includes new data from 127 rats and 36 datasets downloaded from NIH SRA. The detailed sample metadata are provided in Table S8. 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. 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 are been uploaded to NIH SRA (see Table S8 for SRA IDs). Additional resources are listed in Table S5.

Code availability. The code for the custom R, Python and Bash scripts for data analysis is available upon request.

Lead contact. Hao Chen (hchen@uthsc.edu). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

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 to calculate common assembly metrics, such as contig and scaffold N50 (Mikheenko et al., 2018), 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. [22] 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. [71]

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. [72] GATK [73] 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. Long read

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data were mapped using Minimap2, [74] Deepvariant [25] (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. [26] Large SVs detected by LongRanger were merged using SURVIVOR [75] 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, [76] using RefSeq annotations. Overlap between SNPs and repetitive regions were identified with RepeatMasker. [77] Disease ontology was retrieved from the Rat Genome Database. [78] 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. [79]

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

Evaluating Liftover. To evaluate the accuracy and utility of the Liftover process, we compared the variants obtained through the Liftover process to those directly called from the data. We mapped WGS data from a WKY/N rat to both Rnor_6.0 and mRatBN7.2 and called variants against each respective genome. We then used the UCSC Liftover tool [80] 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. 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/DuCrl, M520/NRrrcMcwi, MR/NRrrc WKY/NHsd. The best matches of the remaining strains were less similar: BUF/Mna (73.6%) for BUF/N and WAG/RijCrl (72.0%) for WN/N. Better alternatives for these two strains may be identified in the future as more inbred rat strains are sequenced.

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

described previously. [15] Briefly, genotypes were determined using genotyping-by-sequencing. [81] 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 doi: 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 [82] and removed. We further used an unbiased selection procedure, which yielded a final list of 150,835 binned markers uniformly distributed across the genome. We used Lep-MAP3 (LM3) [40] 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 [83] to filter for high-quality, bi-allelic SNP sites from the VCF files for the 20 autosomes. We then employed PLINK [82] 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 [84] to construct a distance-based UPGMA tree using the meg file as input. We also manually adjusted 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 [85] package to plot the phylogenetic tree.

RNA-seq data. RNA-seq data was downloaded from RatGTEx. [28] Brains were extracted from 88 HS rats. Rats were housed under standard laboratory conditions. Rat brains were extracted and cryosectioned into 60 µm sections, which were mounted onto RNase-free glass slides. Slides were stored in -80°C until dissection and before RNA-extraction post dissection. AllPrep DNA/RNA mini kit (Qiagen) was used to extract RNA. RNA-seq was performed on mRNA from each brain region using Illumina HiSeq 4000 to obtain 100 bp single-end reads for 435 samples. RNA-Seq reads were aligned to the Rnor_6.0 and mRatBN7.2 genomes from Ensembl using STAR v2.7.8a. [86]

eQTL relocation analysis. We obtained the nucleus accumbens core (NAcc, 75 samples) eQTL dataset from RatGTEx [28] (https://ratgtex.org), which was mapped using

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rnor_6.0. We considered associations with p<1e-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 [74] 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. [87] Briefly, small RNAs of ~15–60 nt were size selected by denaturing gel electrophoresis starting from total RNA extracted from 14 rat brain tissue dissections. For csRNA libraries, cap selection was followed by decapping, adapter ligation, and sequencing. For input libraries, 10% of small RNA input was used for decapping, adapter ligation, and sequencing. After library quality check by gel electrophoresis, the samples were sequenced using the Illumina NextSeq 500 platform using 75 cycles single end. Sequencing reads were aligned to the Rnor_6.0 and rat mRatBN7.2 genome assembly using STAR v2.5.3a [86] aligner with default parameters. Transcriptional start regions were defined using HOMER's findPeaks tool. [88]

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 [89]. 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

Mapping of alternative polyadenylation sites to the mRatBN7.2 reference genome. A total of 83 WTTS-seq (whole transcriptome termini site sequencing [27]) libraries were constructed individually using total RNA samples derived from several brain tissues of

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 rat.

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 [90] for protein expression quantitative trait locus (pQTL) mapping. cis-pQTL are defined when the transcriptional start sites for the tested protein are located within ± 1 Mb of each other.

Supplementary Methods

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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, [30] 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. [66]. 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 strain/substrain 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. [91] A set of potential transcripts was generated using multiple

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

techniques: primarily through alignment of transcriptomic data sets, 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. [92] 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. [93]

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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [94] of miRbase [95] against the genome, before passing the results in to RNAfold. [96] Poor quality and repeat-ridden alignments were discarded. Other types of small non-coding genes were annotated by scanning Rfam [97] against the genome and passing the results into Infernal. [98]

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 [93] using NCBI's Eukaryotic Genome Annotation Pipeline [99]. 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/, and

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

The work of PR is supported by the Academy of Finland (grant number 343656). The work of MT is supported by NIH NHLBI R01HL064541, P01HL149620, and Office of the Director R24OD024617. The work of HA is supported by NIH NIDA U01DA043098. The work of CB is supported by NIH NIDA U01DA051972. The work of WMD is supported by NIH NHLBI R01HL064541 and Office of the Director R24OD024617. The work of AG is supported by NIH NHLBI R01HL064541 and Office of the Director R24OD024617. The work of TH is supported by Wellcome Trust [WT222155/Z/20/Z]. The work of TK is supported by NIH R01HG011252. The work of FJM is supported by Wellcome Trust [WT222155/Z/20/Z]. The work of MP is supported by 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. The work of JS is supported by NIH NIDA U01DA051234. The work of JRS is supported by NIH NHLBI R01HL064541, NHGRI U24HG010859, and Office of the Director R24OD024617. The work of LCS and HS rats is supported by NIH NIDA P50DA037844. The work of FT is supported by NIH NIDA U01DA050239 and U01DA051972. The work of LS is supported by NIH NIDA P30DA044223. The work of TDM is supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. The work of AEK is supported by NIH NHLBI R01HL064541, P01HL149620, NHGRI U24HG010859, and Office of the Director R24OD024617. The work of MRD and HRDP is supported by NIH Office of the Director grant R24OD024617. The work of RWW is supported by NIH NIDA U01DA047638 and P30DA044223. The work of JZL is supported by NIH NIDA U01DA043098. The work of HC is supported by NIH NIDA U01DA047638, P50DA037844, and R01DA048017.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Richter, C.P. (1954). The effects of domestication and selection on the behavior of the Norway rat. J. Natl. Cancer Inst. *15*, 727–738.

2. Hulme-Beaman, A., Orton, D., and Cucchi, T. (2021). The origins of the domesticate brown rat (Rattus norvegicus) and its pathways to domestication. Anim Front *11*, 78–86.

3. Modlinska, K., and Pisula, W. (2020). The Norway rat, from an obnoxious pest to a laboratory pet. Elife *9*. 10.7554/eLife.50651.

4. 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. (2013). Rats are the smart choice: Rationale for a renewed focus on rats in behavioral genetics. Neuropharmacology *76 Pt B*, 250–258.

5. Smith, J.R., Hayman, G.T., Wang, S.-J., Laulederkind, 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.

6. RRRC (2021). Rat Resource & Research Center - Rat Models. https://www.rrrc.us/.

7. Pravenec, M., Klír, 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.

8. 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.

9. Tabakoff, B., Smith, H., Vanderlinden, L.A., Hoffman, P.L., and Saba, L.M. (2019). Rat Genomics book. Methods Mol. Biol. *2018*, 213–231.

10. Hansen, C., and Spuhler, K. (1984). Development of the National Institutes of Health genetically heterogeneous rat stock. Alcohol. Clin. Exp. Res. *8*, 477–479.

11. Baud, A., Hermsen, R., Guryev, V., Stridh, P., Graham, D., McBride, M.W., Foroud, T., Calderari, S., Diez, M., Ockinger, J., et al. (2013). Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat. Genet. *45*, 767–775.

12. Woods, L.C.S., and Mott, R. (2017). Heterogeneous Stock Populations for Analysis of Complex Traits. Methods Mol. Biol. *1488*, 31–44.

13. Solberg Woods, L.C., and Palmer, A.A. (2019). Using Heterogeneous Stocks for Fine-Mapping Genetically Complex Traits. Methods Mol. Biol. *2018*, 233–247.

14. Chitre, A.S., Polesskaya, 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Glucose. Obesity *28*, 1964–1973.

15. Gunturkun, M.H., Wang, T., Chitre, A.S., Garcia Martinez, A., Holl, K., St Pierre, C., Bimschleger, H., Gao, J., Cheng, R., Polesskaya, 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.

16. 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.

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

18. 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.

19. 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 Genomics *14*, 257.

20. Tutaj, M., Smith, J.R., and Bolton, E.R. (2019). Rat Genome Assemblies, Annotation, and Variant Repository. In Rat Genomics, G. T. Hayman, J. R. Smith, M. R. Dwinell, and M. Shimoyama, eds. (Springer New York), pp. 43–70.

21. 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.

22. 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.

23. 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*. 10.1093/gigascience/giaa153.

24. 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.

25. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 10.1038/nbt.4235.

26. Yun, T., Li, H., Chang, P.-C., Lin, M.F., Carroll, A., and McLean, C.Y. (2020). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Cold Spring Harbor

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Laboratory, 2020.02.10.942086. 10.1101/2020.02.10.942086.

27. 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.

28. Munro, D., Wang, T., Chitre, A.S., Polesskaya, 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. 10.1093/nar/gkac912.

29. Canzian, F. (1997). Phylogenetics of the laboratory rat Rattus norvegicus. Genome Res. *7*, 262–267.

30. 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.

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 Genomics *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 Genomics *15*, 197.

33. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, 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. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One *9*, e112963.

35. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics *34*, i142–i150.

36. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245.

37. Littrell, J., Tsaih, S.-W., Baud, A., Rastas, P., Solberg-Woods, L., and Flister, M.J. (2018). A High-Resolution Genetic Map for the Laboratory Rat. G3 *8*, 2241–2248.

38. Rastas, P. (2020). Lep-Anchor: automated construction of linkage map anchored haploid genomes. Bioinformatics *36*, 2359–2364.

39. 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*,

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2166–2176.

40. Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics *33*, 3726–3732.

41. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

42. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889.

43. 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.

44. Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. Nat. Biotechnol. *37*, 124–126.

45. 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*. 10.1093/bib/bbab170.

46. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.

47. 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.

48. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898.

49. 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.

50. 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.

51. 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 Genomics *16*, 357.

52. Supernat, A., Vidarsson, O.V., Steen, V.M., and Stokowy, T. (2018). Comparison of three

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variant callers for human whole genome sequencing. Sci. Rep. *8*, 17851.

53. 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.

54. 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.

55. 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.

56. 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. bioRxiv, 2022.03.11.483993. 10.1101/2022.03.11.483993.

57. Lusis, 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.

58. 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.

59. de Jong, T.V., Kim, P., Guryev, V., Mulligan, M.K., Williams, R.W., Redei, E.E., and Chen, H. (2021). Whole genome sequencing of nearly isogenic WMI and WLI inbred rats identifies genes potentially involved in depression and stress reactivity. Sci. Rep. *11*, 14774.

60. Gabriel, D.B.K., Liley, A.E., Franks, H., Tutaj, M., Dwinell, M.R., de Jong, T., Williams, R.W., Mulligan, M.K., Chen, H., and Simon, N.W. (2022). Divergent risky decision-making and impulsivity behaviors in Lewis rat substrains with low genetic difference. bioRxiv, 2022.08.01.501451. 10.1101/2022.08.01.501451.

61. 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.

62. 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.

63. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

receptor levels to attenuate septic kidney injury. Shock *40*, 136–143.

64. 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.

65. 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., et al. (2018). The genetics of smoking in individuals with chronic obstructive pulmonary disease. Respir. Res. *19*, 59.

66. 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.

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. 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.

71. Kalikar, S., Jain, C., Vasimuddin, and Misra, S. (2022). Accelerating minimap2 for long-read sequencing applications on modern CPUs. Nature Computational Science *2*, 78–83.

72. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

73. 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. bioRxiv, 201178. 10.1101/201178.

74. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

75. 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.

76. 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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

w1118; iso-2; iso-3. Fly *6*, 80–92.

77. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, Unit 4.10.

78. 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.

79. 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 Genes|Genomes|Genetics *12*, jkac059.

80. 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.

81. 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 *10*, 2195–2205.

82. 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., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

83. 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*. 10.1093/gigascience/giab008.

84. Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol. Biol. Evol. *38*, 3022–3027.

85. 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.

86. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

87. 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., et al. (2022). Glucocorticoid Receptor-Regulated Enhancers Play a Central Role in the Gene Regulatory Networks Underlying Drug Addiction. Front. Neurosci. *16*, 858427.

88. 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.

89. 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., et al. (2022). Cocaine addiction-like behaviors are

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated with long-term changes in gene regulation, energy metabolism, and GABAergic inhibition within the amygdala. bioRxiv, 2022.09.08.506493. 10.1101/2022.09.08.506493.

90. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat. Commun. *8*, 15452.

91. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database *2016*. 10.1093/database/baw093.

92. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

93. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, 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.

94. 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.

95. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, D155–D162.

96. 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.

97. 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.

98. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

99. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1. Global statistics for the Rat, Mouse, and Human reference genomes. Rnor_6.0 is the current rat reference genome, mRatBN7.2 is the new rat reference genome, GRCm39 is the current mouse reference genome, GRCh38 is the current human reference genome, CHM 13 is the first truly gapless human genome from a haploid cell line recently published by the Telomere-to-Telomere Consortium. Genomes are downloaded from UCSC Goldenpath. Summary statistics are calculated based on the fasta files of each release using QUAST.

	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
Number 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
Number of contigs	75695	757	347	1431	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 number of chromosomes and plasmids	23	23	22	25	24

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Genetic variants in laboratory populations. 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. * Variants that are combinations of insertions, deletions or SNPs. † Including parental strains.

	Variants sites	Variants	Variant Type			Predicted Impact			Homozygous %(mean±SD)	Genotype Qual (mean±SD)
			SNPs	Indels	Mixed*	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†	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†	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

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. This preprint license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

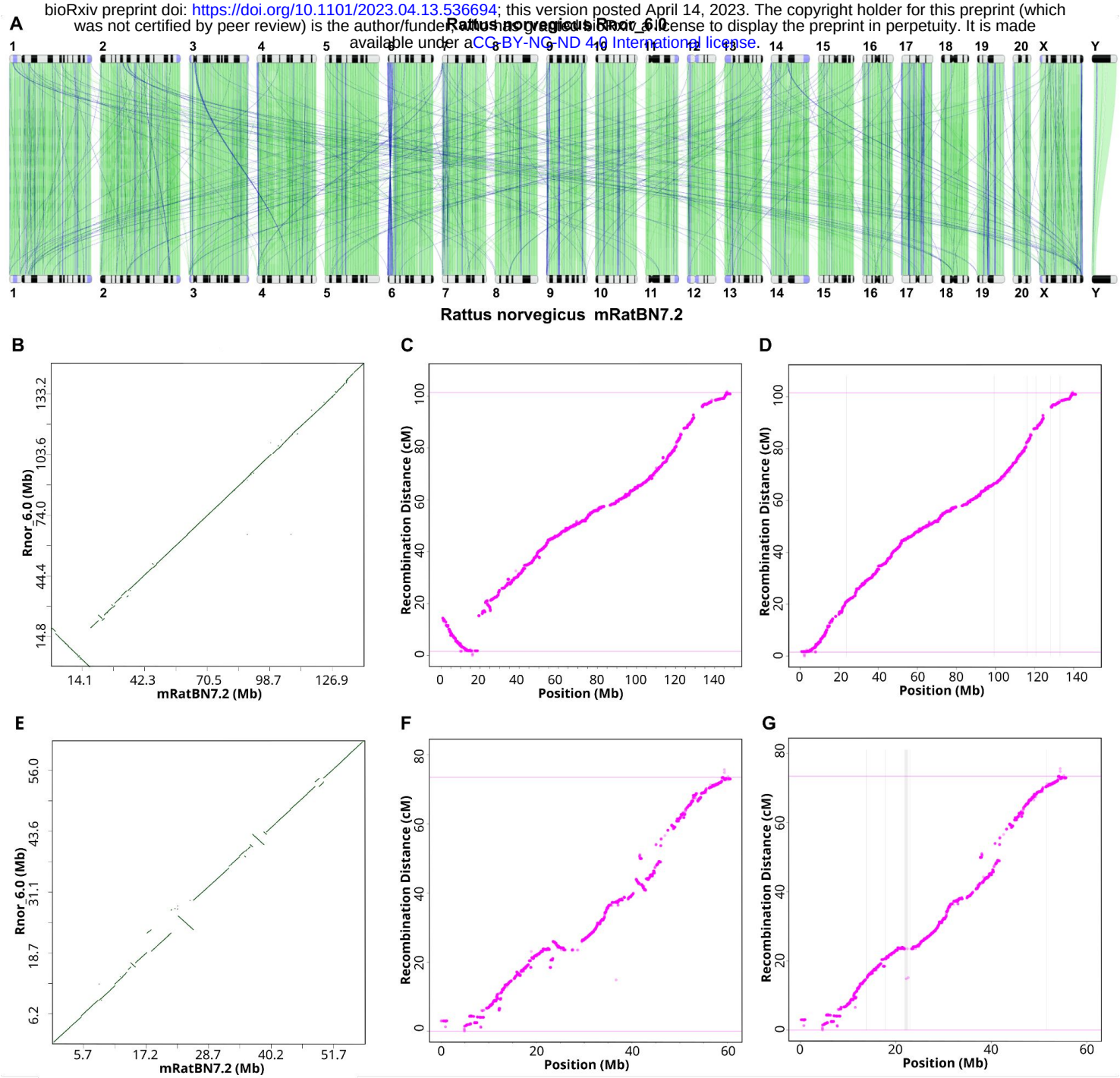

Figure 1. Rat genetic map confirms that mRatBN7.2 corrected structural errors in Rnor6.0. A) Genome-wide comparison between Rnor_6.0 and mRatBN7.2. Numbers indicate chromosomes. Green lines indicate sequences in the forward alignment. Blue lines indicate reverse alignment. Note the inversion at proximal Chr 6 and many translocations between chromosomes. Image generated using the NCBI Comparative Genome Viewer. **B)** Dot plot between Rnor_6.0 and mRatBN7.2 showing large inversion on proximal Chr 6. **C)** Recombination distances and the order of markers on Chr 6 from a rat genetic map independently confirm the inversion at proximal Chr 6 is an assembly error in Rnor_6.0. **D)** Recombination distances and the order of markers on Chr 6 from a new rat genetic map are in agreement with mRatBN7.2. **E)** Dot plot between Rnor_6.0 and mRatBN7.2 showing two large inversions in the middle of Chr 19. **F)** Recombination distances and the order of markers on Chr 19 from a rat genetic map independently confirm the inversions in Chr 19 are assembly errors in Rnor_6.0. **G)** Recombination distances and the order of markers on Chr 19 from a rat genetic map are in agreement with mRatBN7.2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

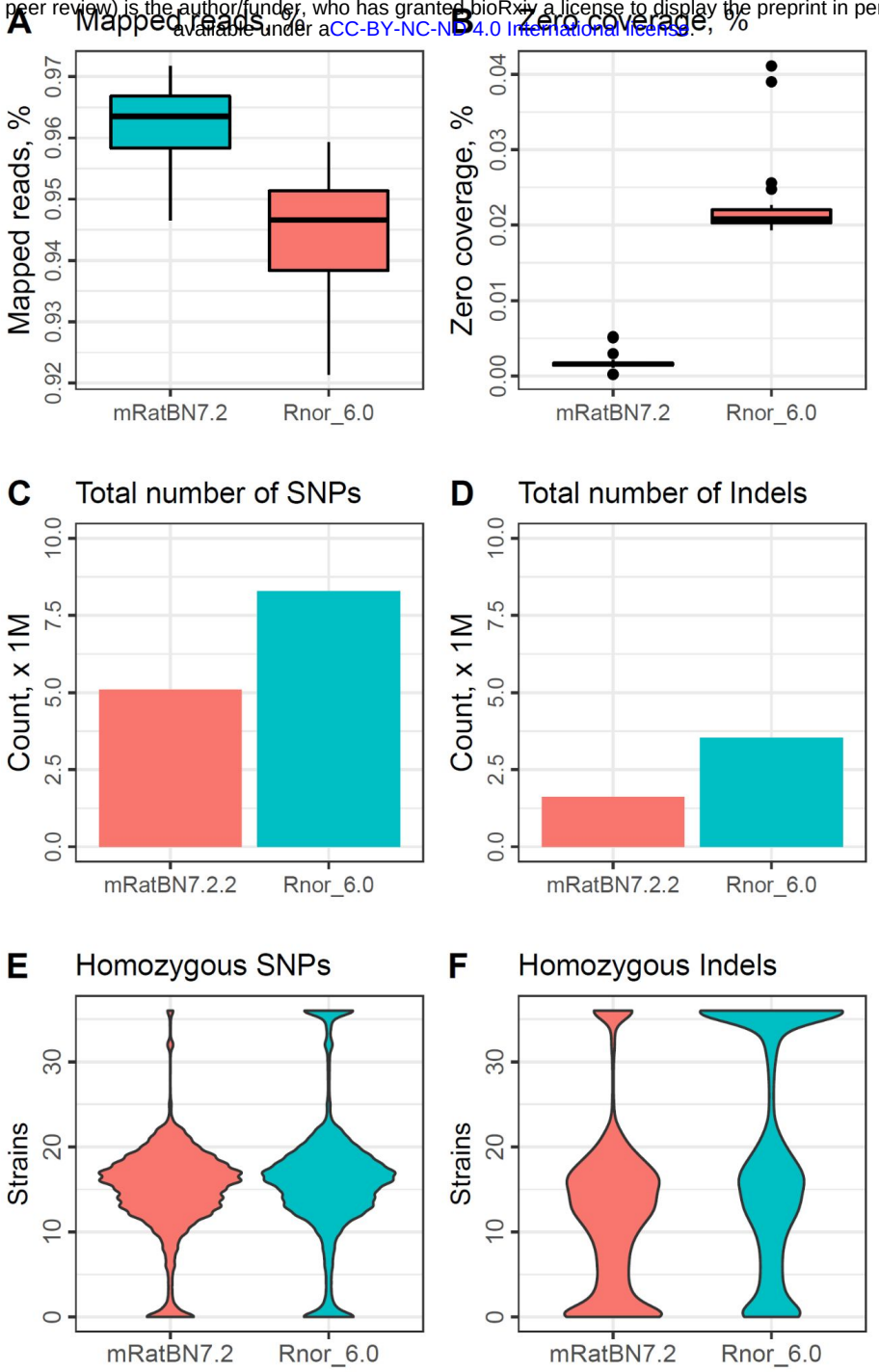

Figure 2. Summary statistics on mapping 36 HXB/BXH WGS samples against Rnor_6.0 or mRatBN7.2. A) Box plot of percentage of reads mapped. **B)** Box plot of percentage of regions on the reference genome with zero coverage. **C)** Total number of SNPs discovered by calling variants jointly over 36 samples. **D)** Total number of indels discovered by calling variants jointly over 36 samples. **E)** Distribution of homozygous SNPs. **F)** Distribution of homozygous indels.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

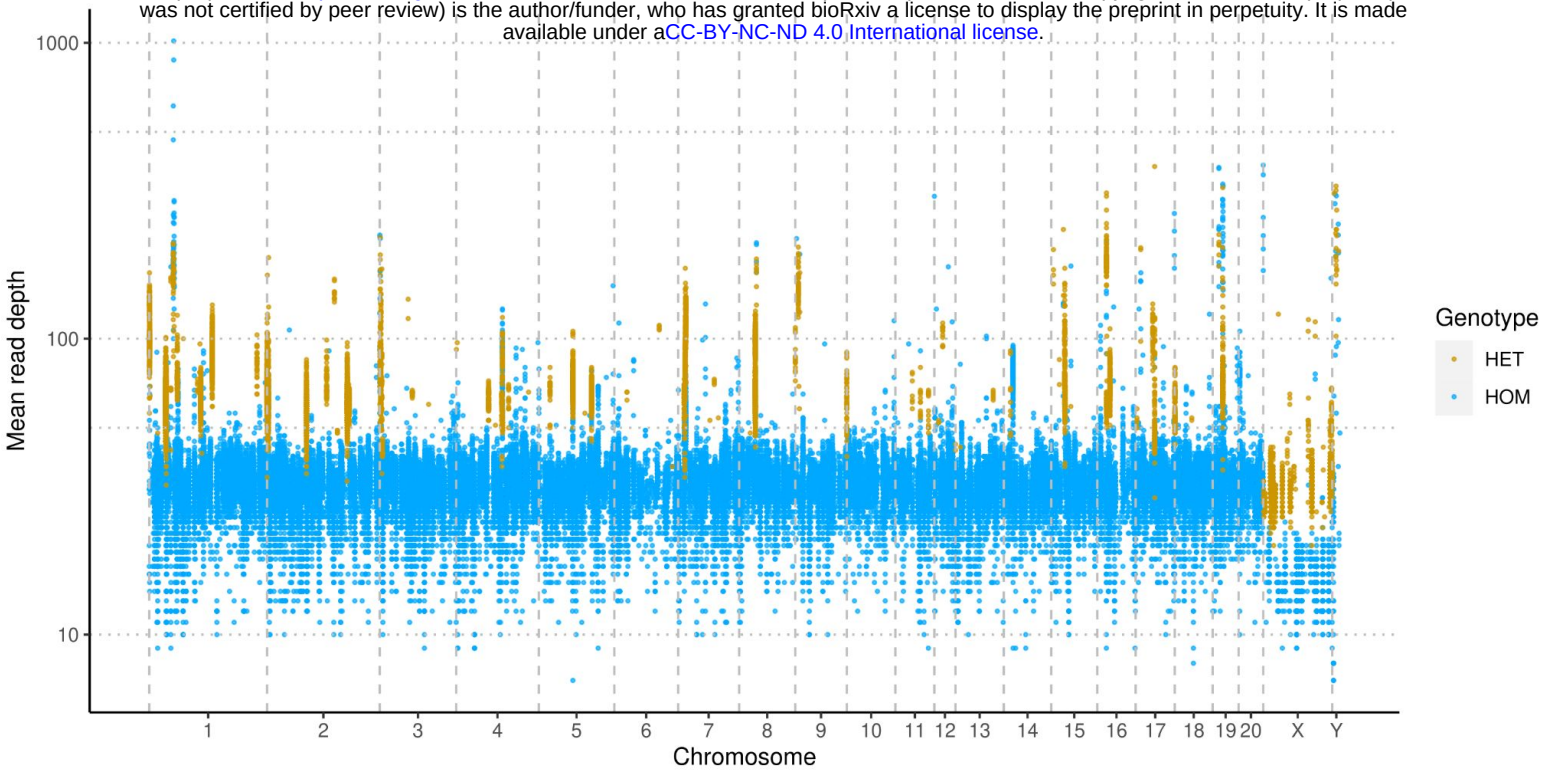

Figure 3. SNPs and indels indicate remaining errors in mRatBN7.2. Base-level errors are indicated by homozygous variants that are shared by over 153 of the 165 samples, including all seven BN/NHsdMcwi rats. Variants that are heterozygous for the majority of the samples are clustered in a few regions and have significantly higher read-depth. This suggests that they originated from collapsed repeats in mRatBN7.2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

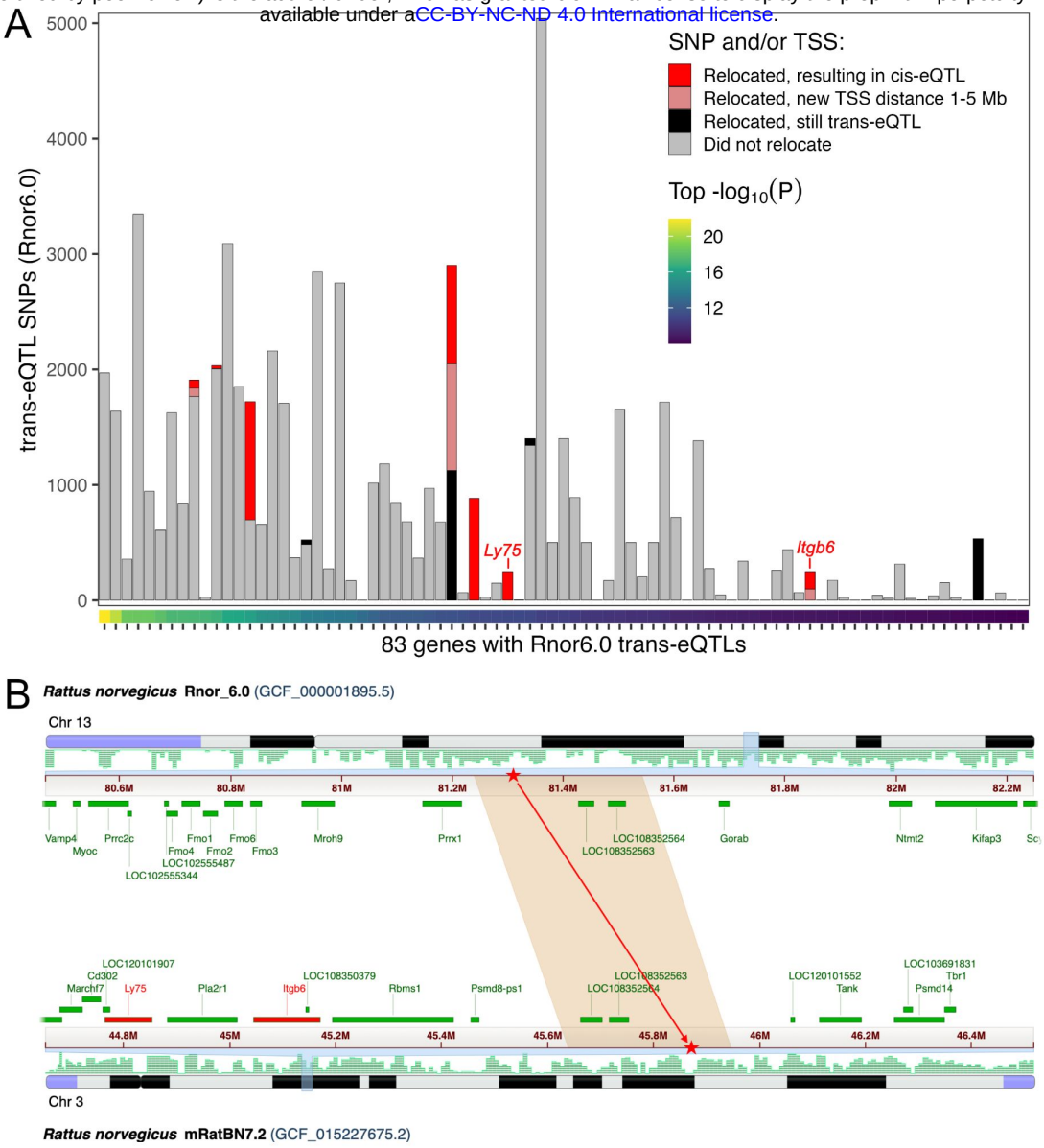

Figure 4. Effect of switching reference genome on eQTL analysis. **A)** Each column represents a gene for which at least one trans-eQTL was found at P<1e-8 using Rnor_6.0 as the reference. The color of bars indicate the number of trans-eQTL SNP-gene pairs in which the SNP and/or gene transcription start 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 1-5 Mb). **B)** One relocated trans-eQTL SNP from A) shown as an example. 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 e-QTL from trans-eQTL to cis-eQTL for both *Ly75* and *Itgb6* genes (red bars).

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

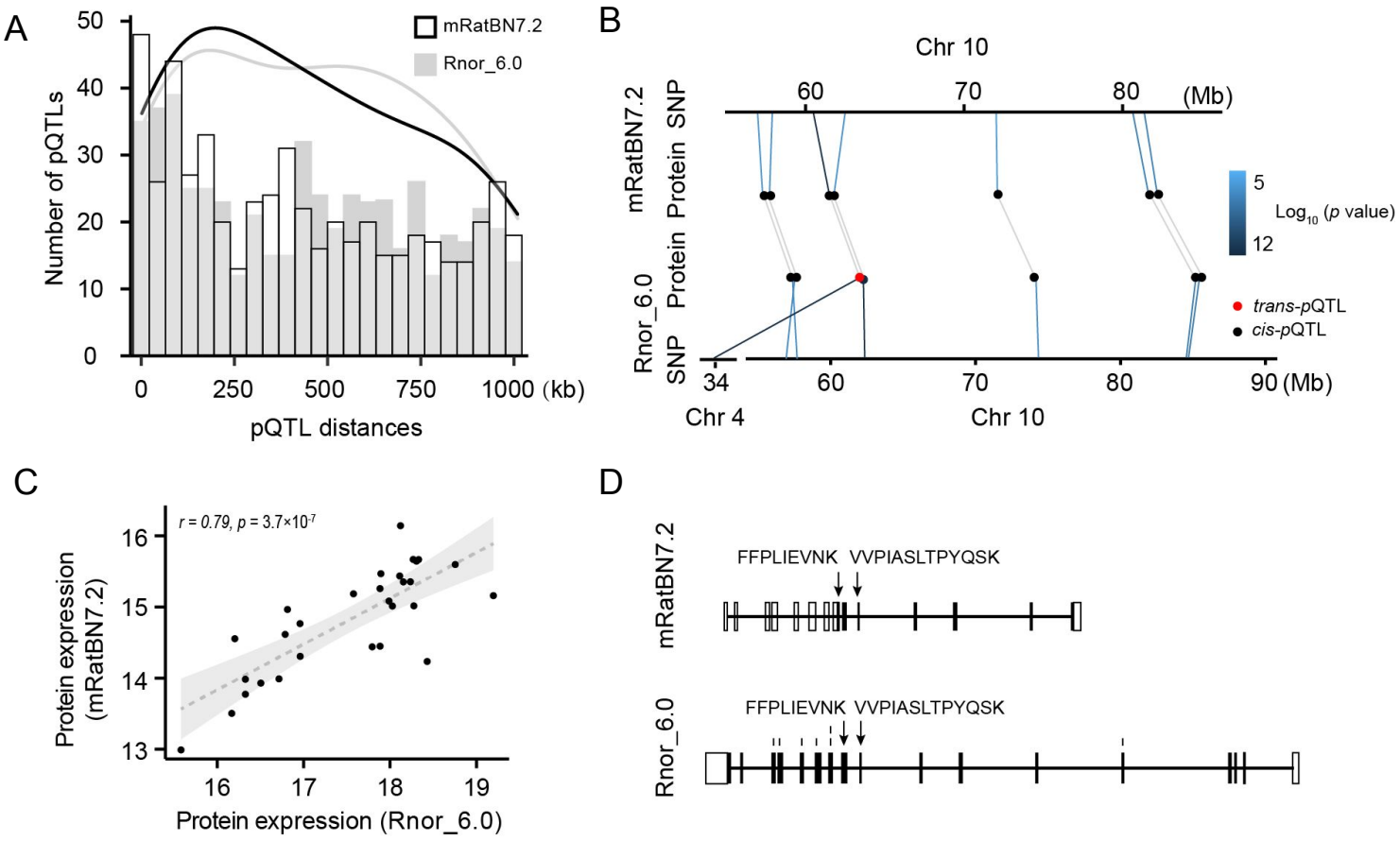

Figure 5. Comparison of the analysis of rat proteomics data between Rnor_6.0 and mRatBN7.2. **A)** Histogram showing the distance between cis-pQTLs and transcription starting site (TSS) of the corresponding proteins. The plot shows that the distances of pQTLs in mRatBN7.2 tend to be closer than those in Rnor_6.0. **B)** An example of trans-pQTL in Rnor_6.0 was detected as a cis-pQTL in mRatBN7.2. **C)** Correlation of expression of the protein (the example in B) in Rnor_6.0 and mRatBN7.2. **D)** Different annotations of the exemplar gene in Rnor_6.0 and mRatBN7.2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

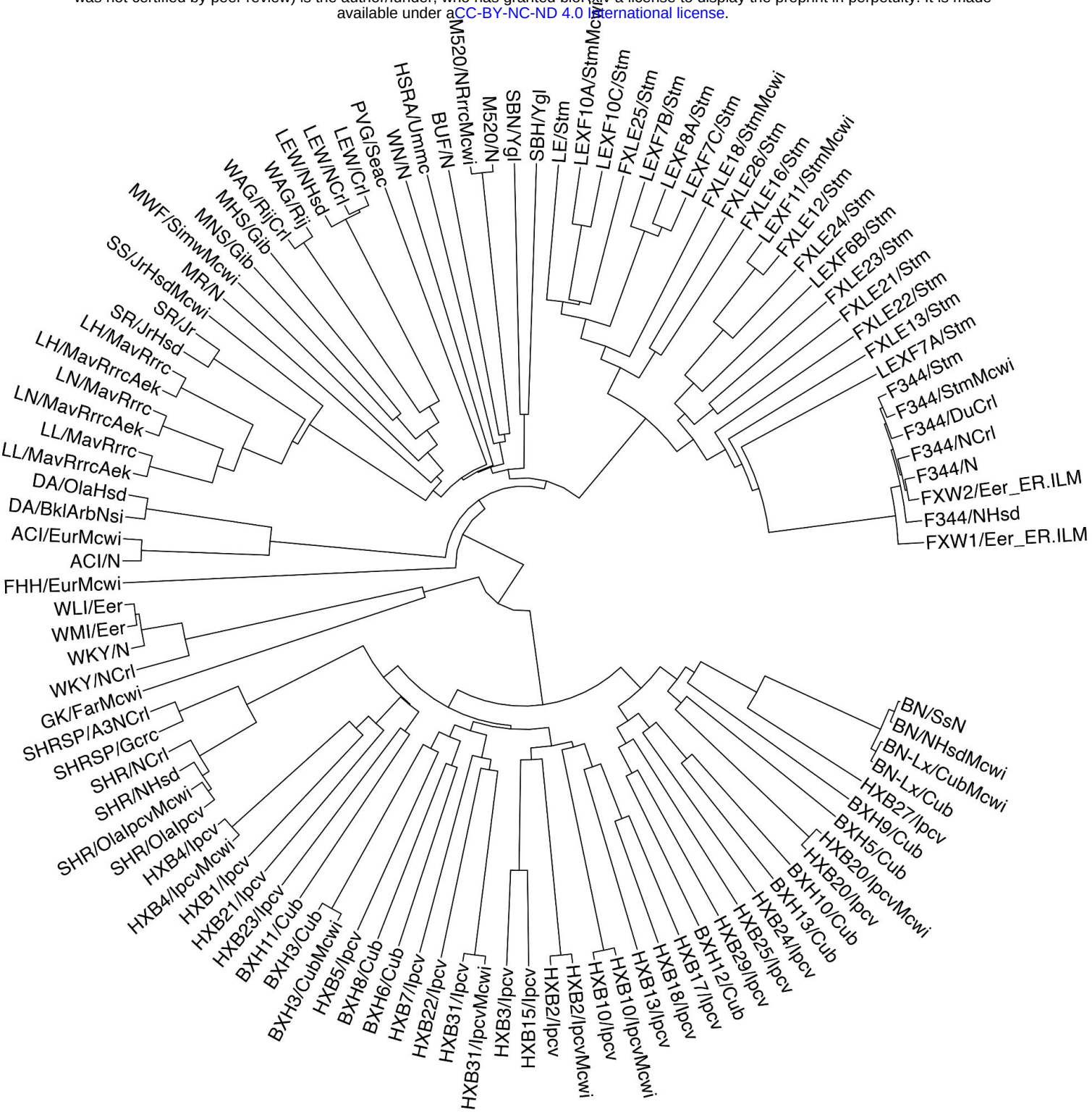

Figure 6. Distance-based phylogenetic tree 120 strain/substrains. The phylogenetic tree was constructed using ~11.6 million SNPs. Samples from the same strain/substrains are condensed. Unweighted pair group method with arithmetic mean (UPGMA) was used for tree construction.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

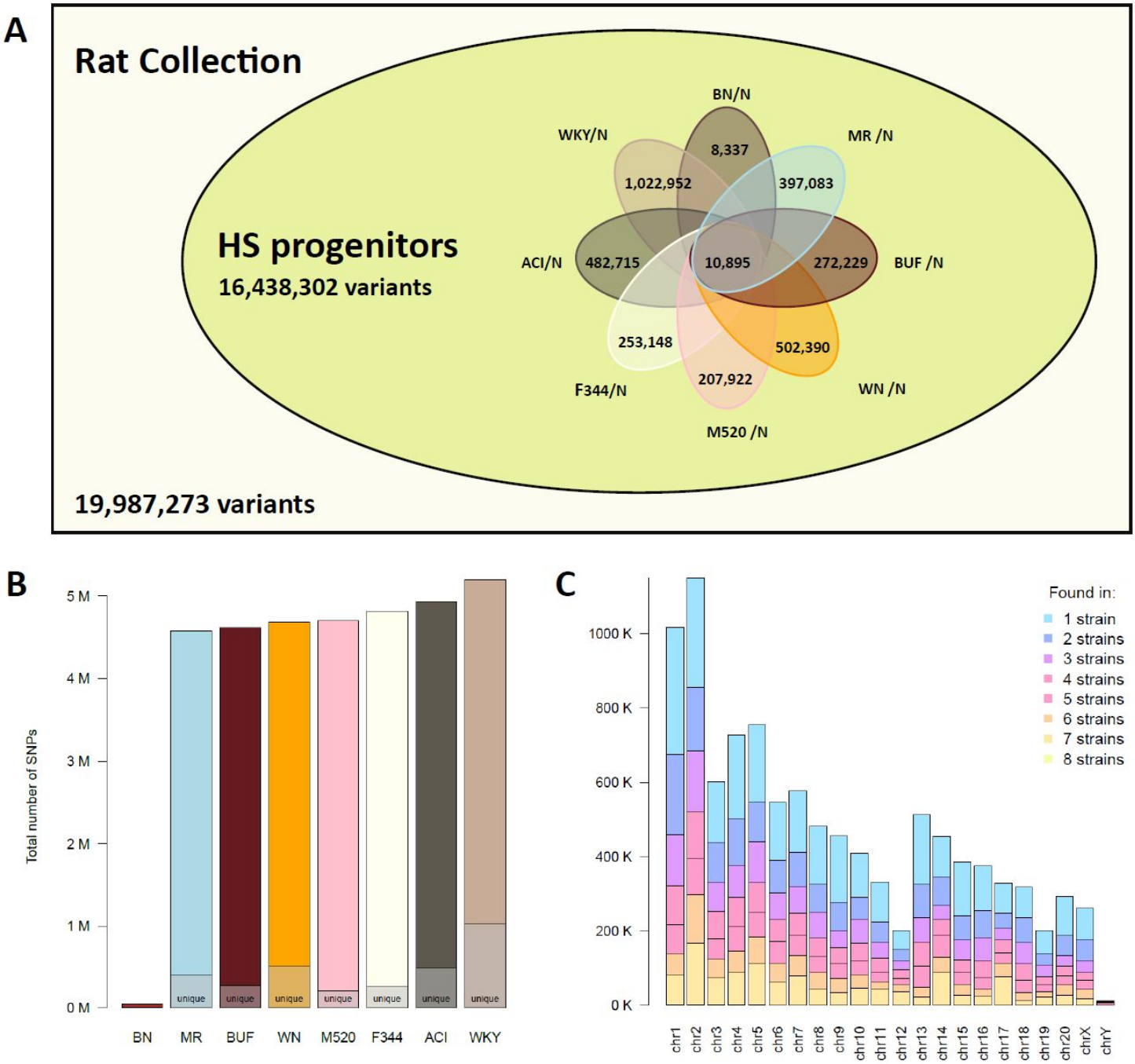

Figure 7. **Genetic diversity among progenitors of heterogeneous stock rats**. **A)** The HS progenitors contain 16,438,302 variants (i.e., 82.2% of the variants in our RatCollection) based on analysis using mRatBN7.2. Among these, 10,895 are shared by all eight progenitor strains. The number of variants that are unique to each specific founder is noted. **B)** The total number of variants per strain, with the total number unique to each strain marked. **C)** The number of variants shared across N strains, shown per chromosome.

bioRxiv preprint doi: https://doi.org/10.1101/2023.04.13.536694; this version posted April 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

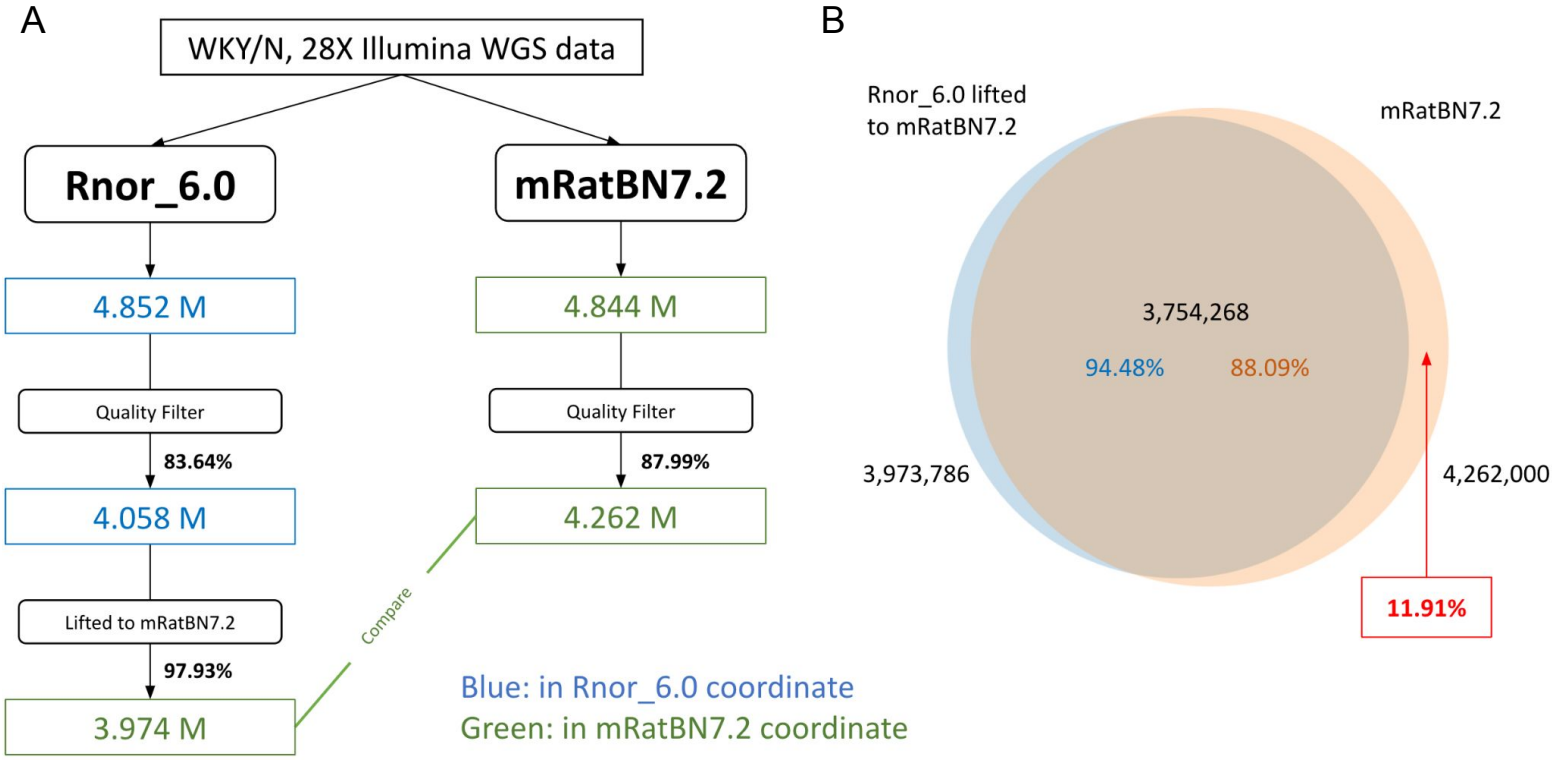

Figure 8. Using WGS data to assess the quality of the Liftover. We compared the lifted variant set to the directly-called variant set for overlapping. **A)** Overview of the pipeline used to assess the quality of the Liftover. A higher portion of variants passed the quality filter for mRatBN7.2. Among them, 97.93% of the variants are liftable from Rnor_6.0 to mRatBN7.2. **B)** The overlap between lifted variants and the remapping variants. Approximately 11.9% of the variants that can be found with a complete remapping will be missed if a liftover is done instead.