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.

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

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

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

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.