Rat Reference Genome mRatBN7.2 Curation

Wendy M. Demos[1], Valerie A Schneider[2], Terence D. Murphy[2], Monika Tutaj[1], Jennifer R. Smith[1], Anne E. Kwitek[1,3]

1. Rat Genome Database, Department of Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI 53226, USA,
2. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.
3. Department of Physiology, Medical College of Wisconsin, Milwaukee, WI 53226, USA.

Abstract

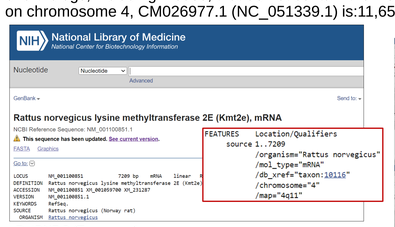
Rattus norvegicus (rat) is an important experimental model for human diseases. Previous rat genome references were highly fragmented despite periodic updates. The latest assembly, mRatBN7.2[1] addresses many deficiencies of prior assemblies but requires continued manual curation for reliability and optimization.

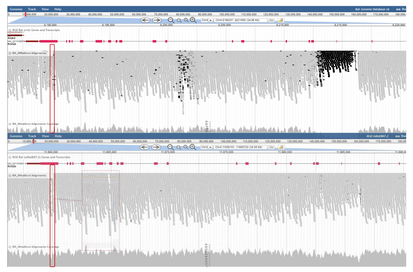
Reference genome issues are directly reported to the Genome Reference Consortium (GRC; https://www.ncbi.nlm.nih.gov/grc/report-an-issue) by RefSeq curators and the rat research community. Issues are assigned a ticket ID via the Atlassian JIRA Service Management platform and addressed by GRC curators at the Rat Genome Database (RGD https://rgd.mcw.edu/). The ticket prioritization strategy is modeled after processes established by the GRC: effects on protein function (known (first), potential (second)), then sequence differences outside the coding region, giving priority to community reported issues. Ticket resolution relies heavily on public tools such as Genome Workbench (NCBI), JBrowse Genome Browser[2] (genomics data produced and curated by RGD at the Medical College of Wisconsin), and additional curation tools available to curators on the GRC platform.

A workflow has been established to review and resolve tickets. Ticket resolution status updates are provided on the GRC webpage and are being integrated into RGD through gene pages and JBrowse Genome Browser and announced to the community through RGD social media.

RG-129 Kmt2e - Stalled

Curator Comment: JACYVU010000141.1 has a 1-nt insertion compared to NM_001100851.1, which results in frameshift in the encoded CDS, and premature termination of the protein (1569 aa vs 1856 aa). NM_001100851.1 is based on Celera genomic sequence, but its sequence is in agreement with prior assemblies, which do not contain the insertion present in the new assembly. NM_001004085.2 is also supported by the genomic sequence of orthologs, including mouse, that do not contain the insert seen in the new genome. The location of the insertion on chromosome 4, CM026977.1 (NC_051339.1) is:11,659,922, with an extra 'C'.[haddadd]



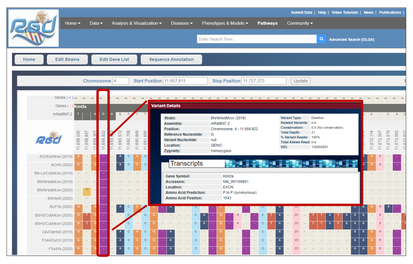


mRatBN7.2 (Top image)

GGAGGCGGAGGGGGTGGGGGGGGGCGGGGCC

RN6.0

GGAGGCGGAGGGGGTGGGGGGGGGCGGGGCCG (bottom)

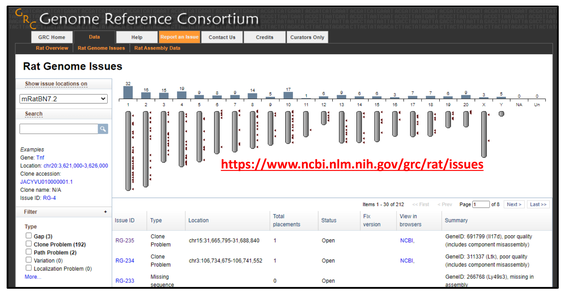


The reported issue in ticket RG-129 may be technical in nature due to the repetitiveness of the region. Variant Visualizer indicates the insertion is not present at this position across strains and the illumina short reads also imply this may be an assembly and / or transcript issue. alternative alignment available at this time to review and potentially use to patch the region, the issue is declared stalled in the JIRA system until more evidence can be provided.

Methods

Sequencing - The latest assembly mammalian Rat Brown Norway (mRatBN) version 7.2 is generated from a kidney tissue of a generation F61 male descendent of the original female generation F14 Brown Norway rat. This assembly generated by the Darwin Tree of Life Project at the Wellcome Sanger Institute provides major advancements in comparison to other clone-based reference versions.

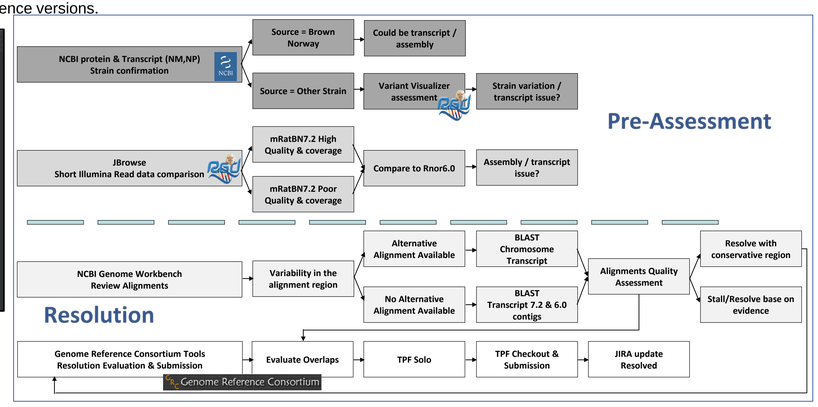
RefSeq Annotation & Issue Reporting – The genome records for mRatBN7.2 were annotated with the NCBI Eukaryotic Genome Annotation Pipeline which utilizes evidence such as RNA-seq, Transcript, known RefSeq data, orthology, and other NCBI data sources. Assembly issues are determined by RefSeq curators. A summary of the issue, gene name, transcript, and coordinates is recorded in the Atlassian JIRA ticketing system. The GRC Rat Issues Dashboard provides public access to individual views of all reported issues and resolution status.



Gene Prioritization - RGD provided the summary of issues to members of the International Rat Omics Consortium (IROC) to identify specific genes of interest in the community.

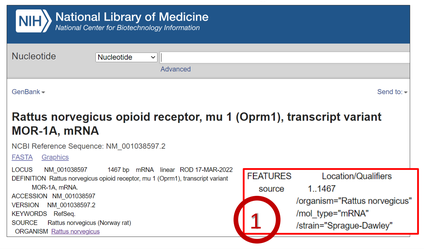
Resolution Workflow - Tickets are pre-assessed via analysis with NCBI resources including the Nucleotide and Gene Databases.

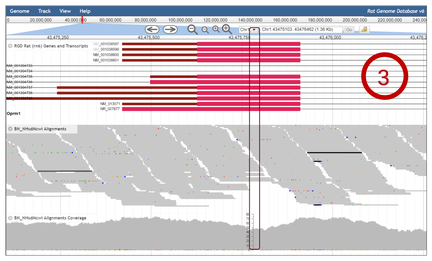
RGD tools Variant Visualizer and JBrowse utilize illumina short read data of the same animal used for reference generation. These data are useful for regional comparisons to the long-read data.

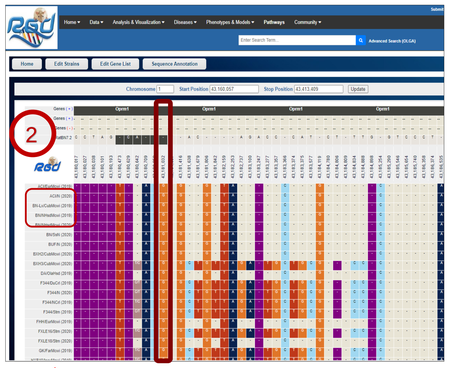


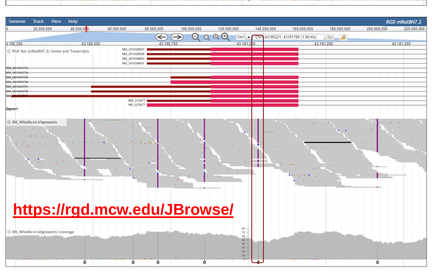
RG-8 Oprm1 - Resolved

Curator Comment: NM_001038597.2 has insertion in CDS relative to JACYVU010000016.1 in the new genome, with resulting frameshift. NM_001038597.2 is based on a transcript, and its sequence is in agreement with prior assemblies. The location of the frameshift on chromosome 1, CM026974.1 (NC_051336.1), in the new assembly is 43,181,033, with a missing 'G'. NM_001038597.2 is also supported by the genomic sequence of orthologs, including human, that do not contain this single nucleotide insertion seen in the new genome. [haddadd]

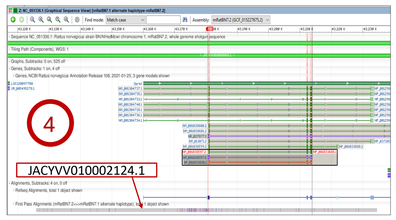




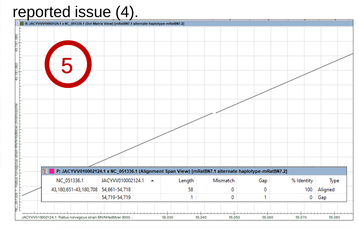




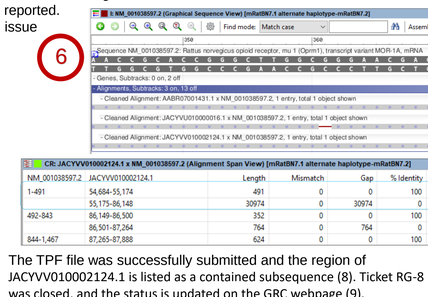
NCBI Nucleotide Database indicates the source for this transcript is the Sprague Dawley strain (1). Variant Visualizer shows Brown Norway strain agrees with other strains at this position (2). JBrowse analysis of short read data (3) illustrates the assembly for Rnor6.0 has less assembly issues in this region.

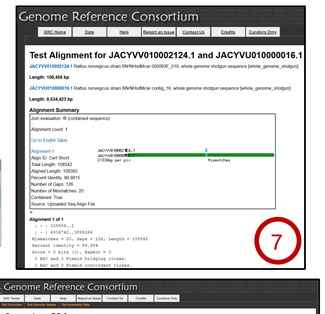


BLAST alignments to the chromosomal region (5), transcript (6) and between contigs (Rnor6.0 & mRatBN7.2, fig 7) provide evidence the alternative contig JACYVV010002124.1 can be utilized to address the reported issue

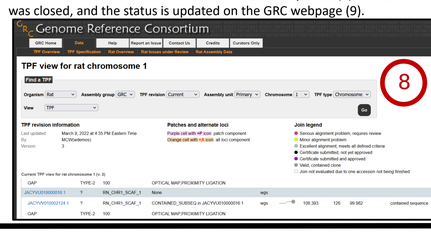


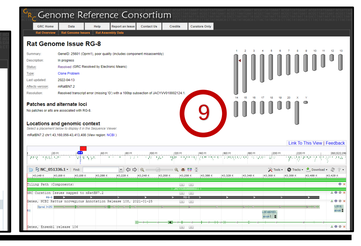
Review of the alternative alignment in Genome Workbench suggests the alternative contig JACYVV010002124.1 can be utilized to address the reported issue (4).





The TPF file was successfully submitted and the region of JACYVV010002124.1 is listed as a contained subsequence (8). Ticket RG-8 was closed, and the status is updated on the GRC webpage (9).

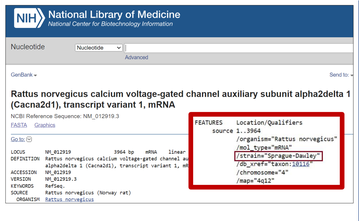


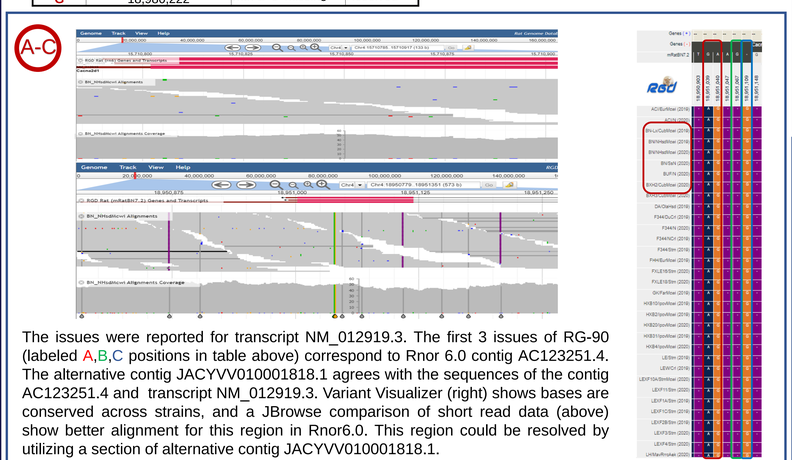


RG-90 Cacna2d1 - Partially Resolved

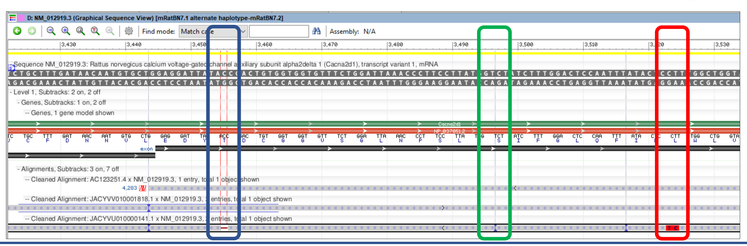
Curator Comment: JACYVU010000141.1 in the new assembly mRatBN7.2 had numerous indels and mismatches compared to NM_012919.3, which cause frameshifts in the encoded protein. NM_012919.3 is based on transcripts, and its genomic sequence is in agreement with prior assemblies. NM_012919.3 is also supported by the genomic sequence of orthologs which do not contain the indels present in the new assembly. The indels are located on chromosome 4, CM026977.1 (NC_051339.1) at these positions: 18,980,222 with a 'C' missing;18,971,771 with a 'C' missing; 18,965,942 with an 'A' missing and an extra 'G';18,963,722 with a missing 'A'; 18,951,108 with a missing 'G';18,951,068 with an extra 'C';18,951,040-18,951,039 with a missing 'CT'.[haddadd]

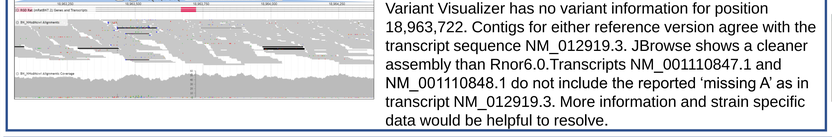
RG-90 Issue ID	Position (mRatBN7.2)	Issue	Status
A	18,951,040 - 18,951,039	with a missing 'CT	In progress
B	18,951,068	with an extra 'C';	In progress
C	18,951,108	with a missing 'G'	In progress
D	18,963,722	with a missing 'A'	Stalled
E	18,965,942	with an 'A' missing and an extra 'G'	Stalled
F	18,971,771	with a 'C' missing;	Stalled
G	18,980,222	with a 'C' missing	Resolved





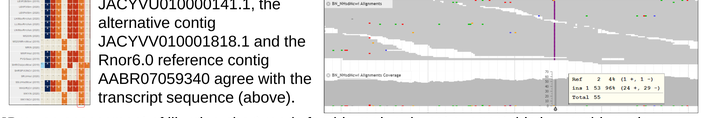
The issues were reported for transcript NM_012919.3. The first 3 issues of RG-90 (labeled A,B,C positions in table above) correspond to Rnor 6.0 contig AC123251.4. The alternative contig JACYVV010001818.1 agrees with the sequences of the contig AC123251.4 and transcript NM_012919.3. Variant Visualizer (right) shows bases are conserved across strains, and a JBrowse comparison of short read data (above) show better alignment for this region in Rnor6.0. This region could be resolved by utilizing a section of alternative contig JACYVV010001818.1.



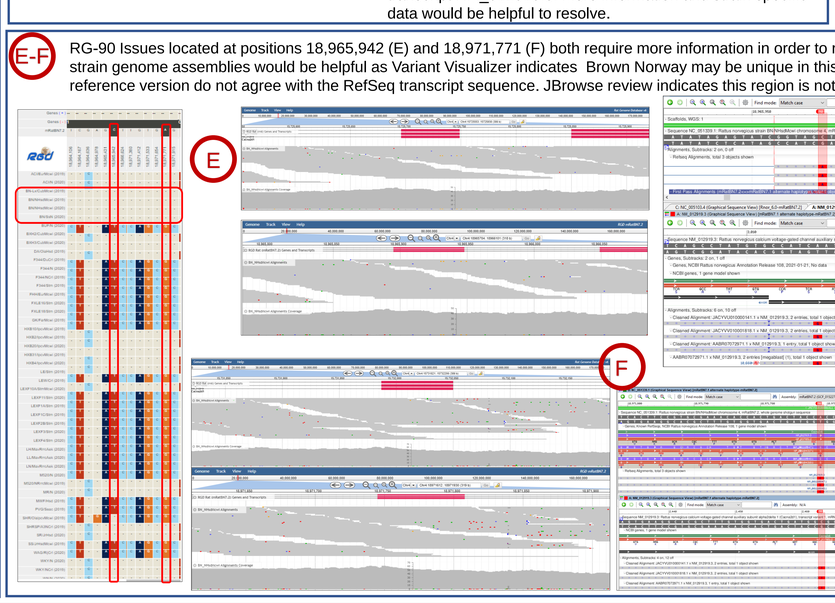


Variant Visualizer has no variant information for position 18,963,722. Contigs for either reference version agree with the transcript sequence NM_012919.3. JBrowse shows a cleaner assembly than Rnor6.0.Transcripts NM_001110847.1 and NM_001110848.1 do not include the reported 'missing A' as in transcript NM_012919.3. More information and strain specific data would be helpful to resolve.

Variant Visualizer shows transcript NM_001110847.1 which agrees with NM_012919.3 at position 18,980,222. It is displayed as an insertion to the reference. The insertion does not appear to be unique to the Brown Norway strain.

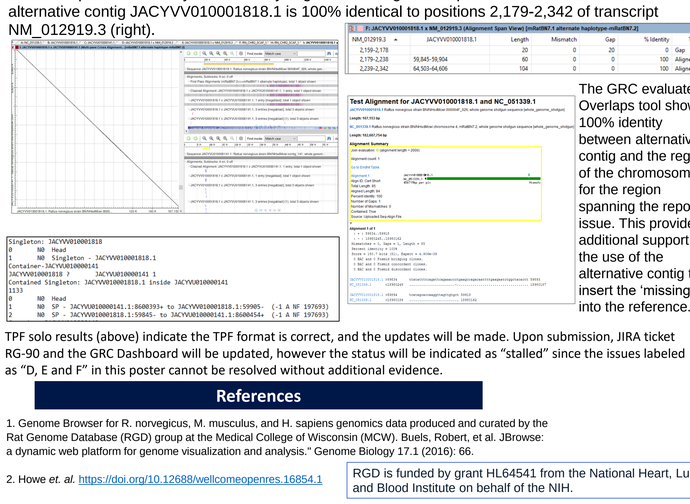


Contrary to the reference contig, JACYVU010000141.1, the alternative contig JACYVV010001818.1 and the Rnor6.0 reference contig AABR07059340 agree with the transcript sequence (above).



RG-90 Issues located at positions 18,965,942 (E) and 18,971,771 (F) both require more information in order to resolve. In both cases other strain genome assemblies would be helpful as Variant Visualizer indicates Brown Norway may be unique in this region. Contigs for either reference version do not agree with the RefSeq transcript sequence. JBrowse review indicates this region is not overly problematic.

JBrowse assessment of illumina short reads for this region do not suggest this is a problematic region to assemble. These data are indicative of a technical error. To support the use of the alternative contig for a section of this region to insert the 'C' into the reference, BLAST was used to determine the % identity between the reference and alternative contigs. The slope of the line in the below left panel show they are inversely aligned. The alignment span view of the BLAST of the alternative contig JACYVV010001818.1 is 100% identical to positions 2,179-2,342 of transcript NM_012919.3 (right).



The GRC evaluate Overlaps tool shows 100% identity between alternative contig and the region of the chromosome for the region spanning the reported issue. This provides additional support in the use of the alternative contig to insert the 'missing C' into the reference.

TPF solo results (above) indicate the TPF format is correct, and the updates will be made. Upon submission, JIRA ticket RG-90 and the GRC Dashboard will be updated, however the status will be indicated as "stalled" since the issues labeled as "D, E and F" in this poster cannot be resolved without additional evidence.

References

1. Genome Browser for R. norvegicus, M. musculus, and H. sapiens genomics data produced and curated by the Rat Genome Database (RGD) group at the Medical College of Wisconsin (MCW). Buels, Robert, et al. JBrowse: a dynamic web platform for genome visualization and analysis." Genome Biology 17.1 (2016): 66.

2. Howe *et. al.* https://doi.org/10.12688/wellcomeopenres.16854.1

RGD is funded by grant HL64541 from the National Heart, Lung, and Blood Institute on behalf of the NIH.