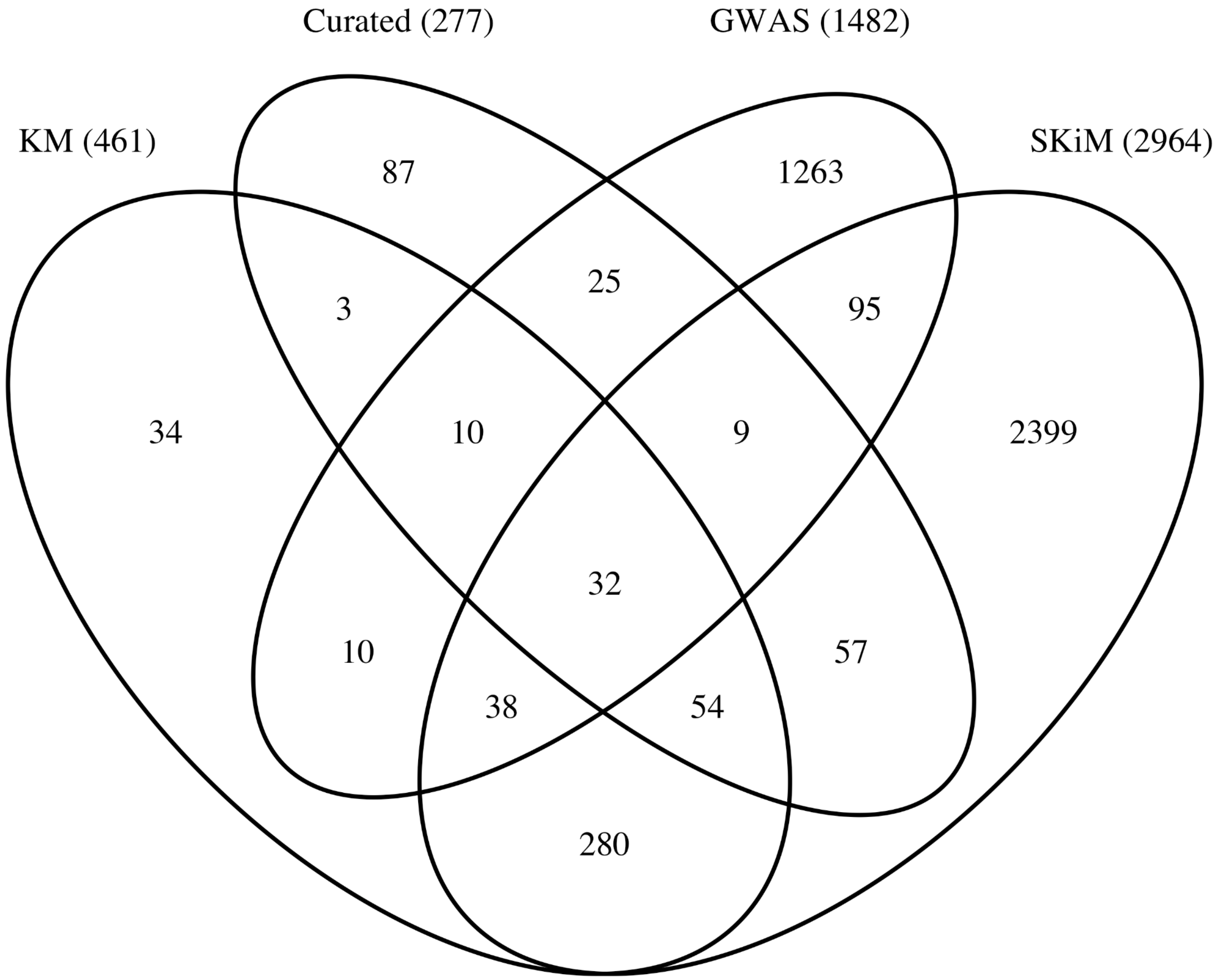
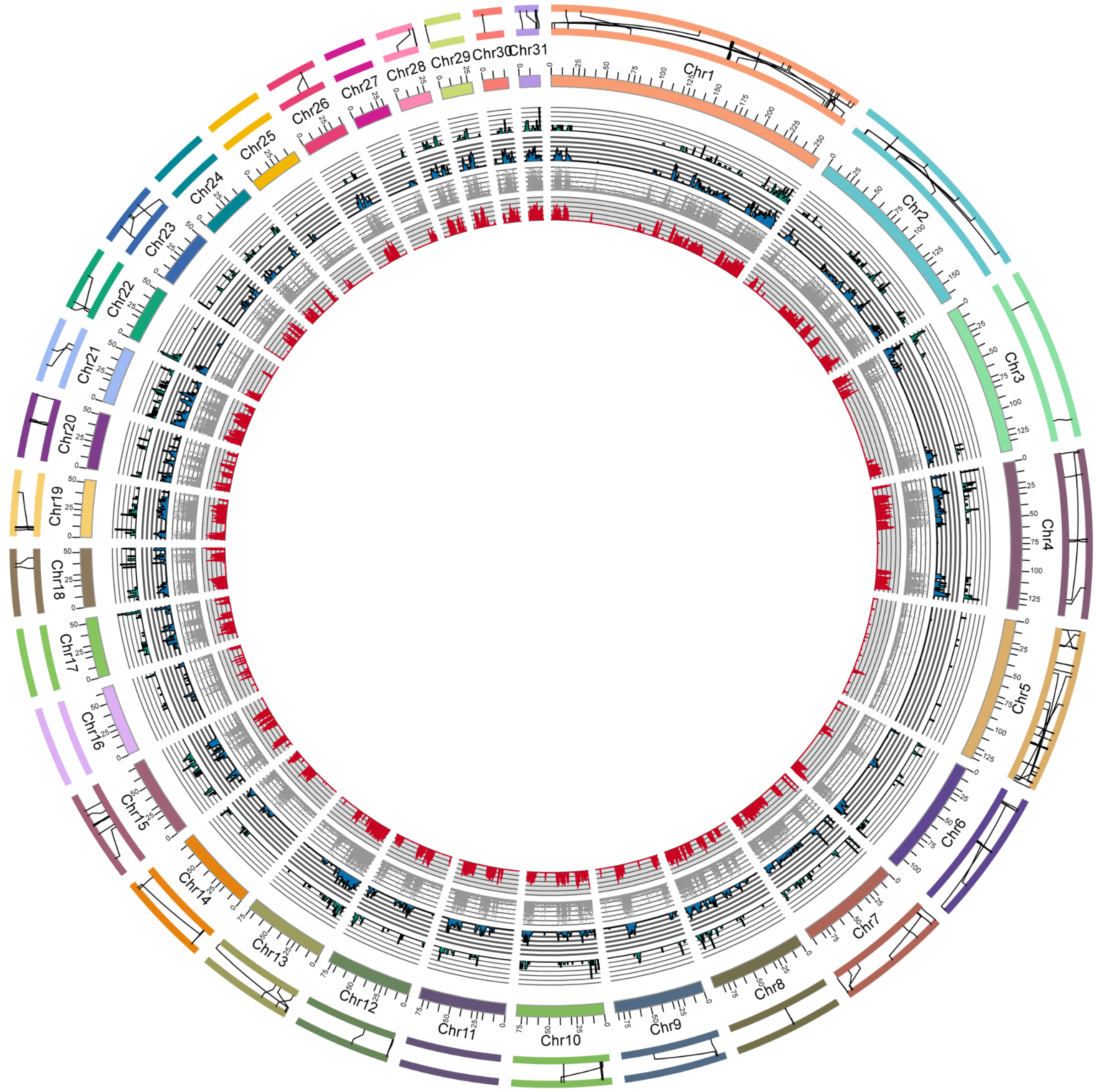
Genes linked to type 2 diabetes



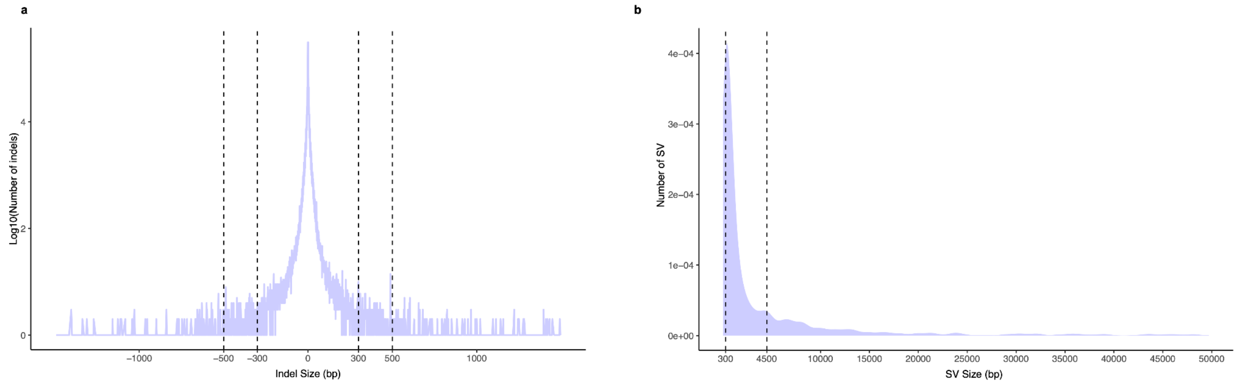
Supplementary Figure 1. Venn diagram of gene lists linked to type 2 diabetes by different types of evidence. KM = KinderMiner, a text-mining tool. SKiM - Serial KinderMiner, a literature-based discovery system.

Heterozygosity spectrum of the Nile rat



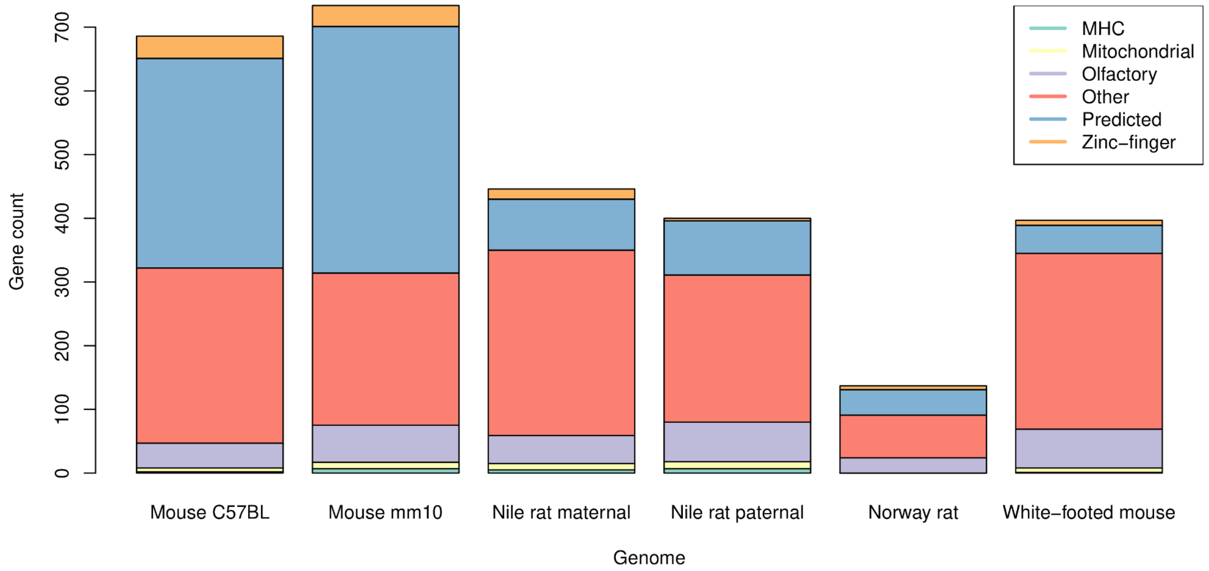
Supplementary Figure 2. Heterozygosity inferred by comparing the paternal and maternal scaffolded contigs, shown on the paternal scaffolds. From inner ring, the plots show density of heterozygous SNV, indels up to 50 bases, and finally insertion and deletion structural variants (≥ 50 bases). The outer rim shows translocations.

Length distributions of structural variants



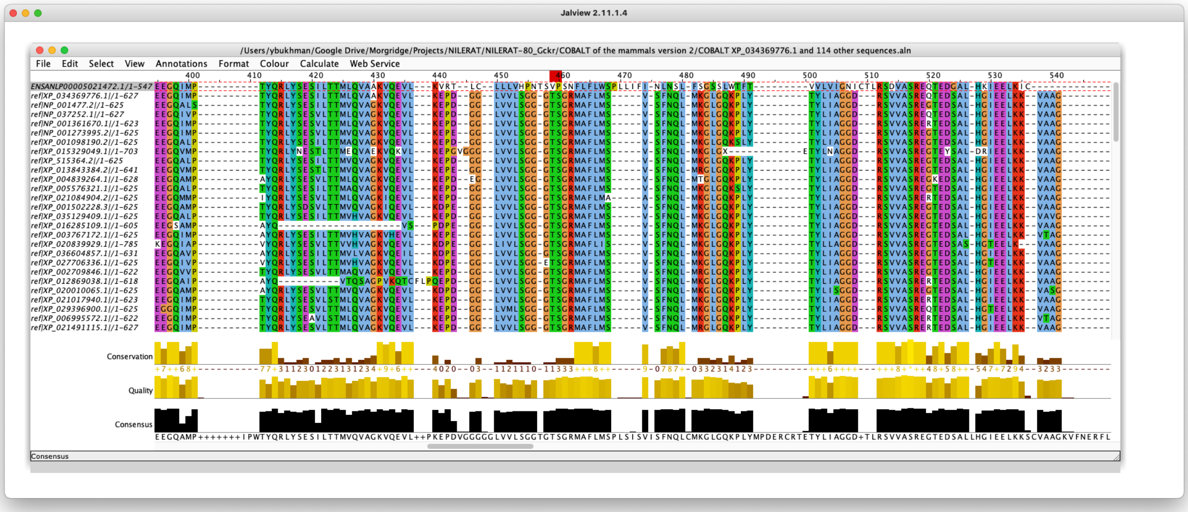
Supplementary Figure 3. Length distributions of structural variants. A. Indels. **B**. Other structural variants.

Functional classification of duplicated genes.



Supplementary Figure 4. Functional classification of duplicated genes. Genes are classified by name according to large categories of gene function, immune (MHC), Mitochondrial, Olfactory, Predicted, and Zinc-finger. The remaining category (Other) represents multi-copy genes not part of large gene families.

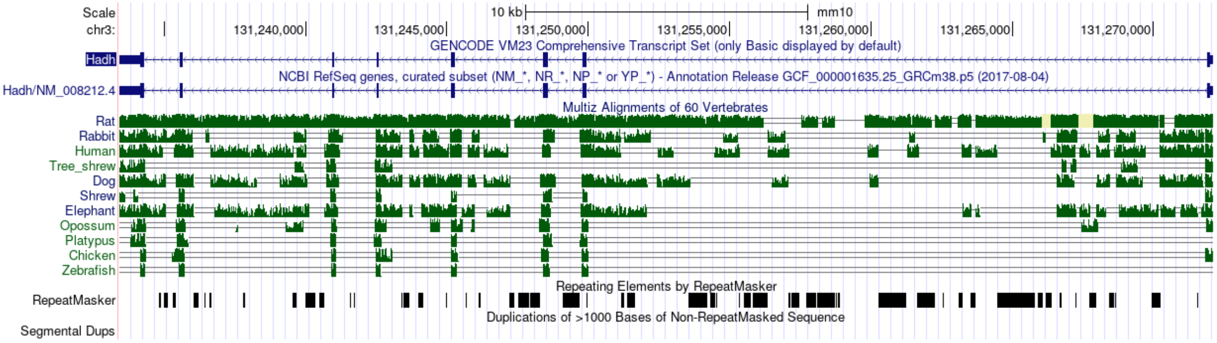
Gckr

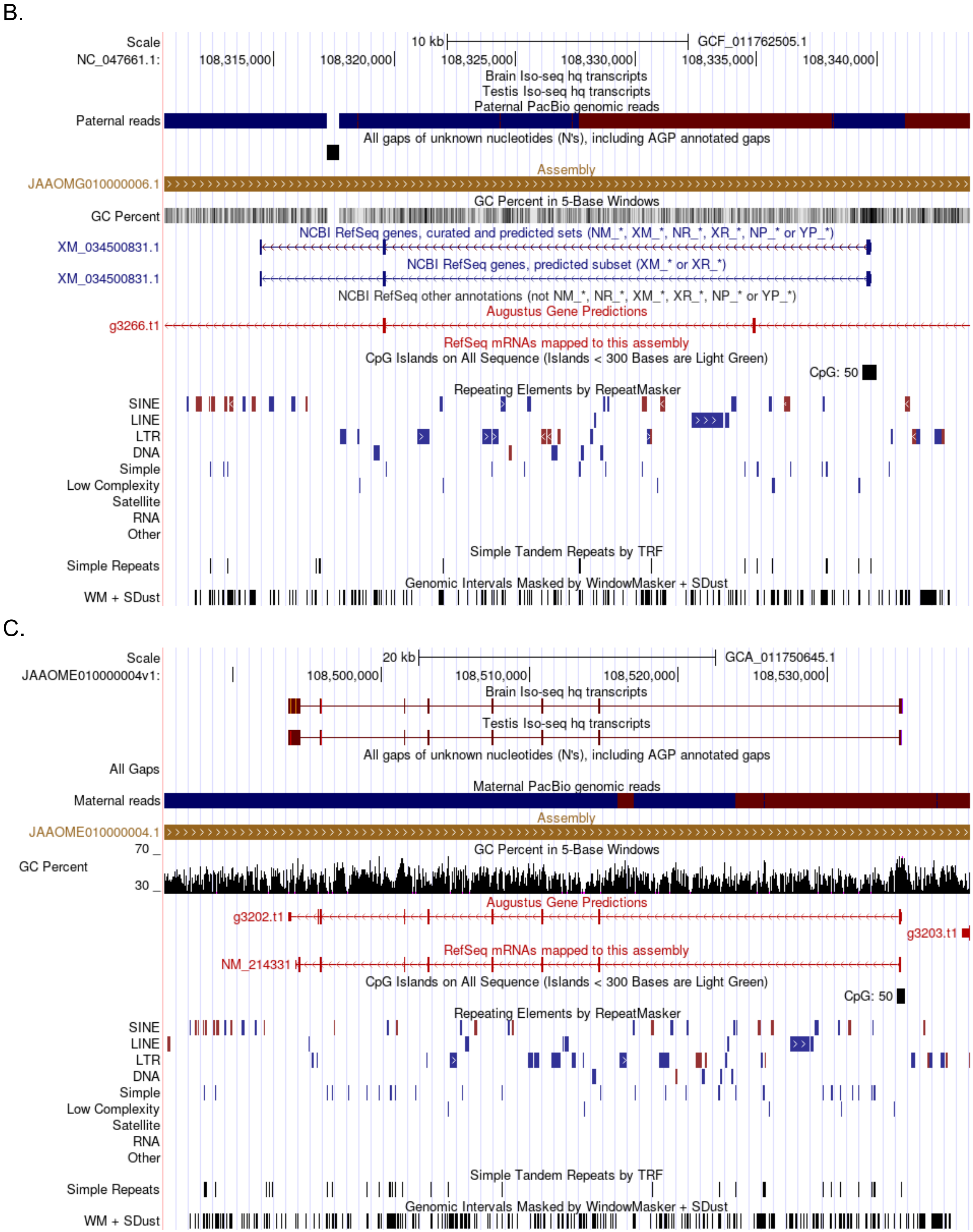

Jalview 2.11.1.4
/Users/ybukhman/Google Drive/Morgridge/Projects/NILERAT/NILERAT-80_Gckr/COBALT of the mammals version 2/COBALT XP_034369776.1 and 114 other sequences.aln
File
Edit
Select
View
Annotations
Format
Colour
Calculate
Web Service
400
410
420
430
440
450
460
470
480
490
500
510
520
530
540
ENSANLP00005021472.1/1-547
ref|XP_034369776.1|/1-627
ref|NP_001477.2|/1-625
ref|NP_037252.1|/1-627
ref|NP_001361670.1|/1-623
ref|NP_001273995.2|/1-625
ref|XP_001098190.2|/1-625
ref|XP_015329049.1|/1-703
ref|XP_515364.2|/1-625
ref|XP_013843384.2|/1-641
ref|XP_004839264.1|/1-628
ref|XP_005576321.1|/1-625
ref|XP_021084904.2|/1-625
ref|XP_001502228.3|/1-625
ref|XP_035129409.1|/1-625
ref|XP_016285109.1|/1-605
ref|XP_003767172.1|/1-625
ref|XP_020839929.1|/1-785
ref|XP_036604857.1|/1-631
ref|XP_027706336.1|/1-625
ref|XP_002709846.1|/1-622
ref|XP_012869038.1|/1-618
ref|XP_020010065.1|/1-625
ref|XP_021017940.1|/1-623
ref|XP_029396900.1|/1-625
ref|XP_006995572.1|/1-622
ref|XP_021491115.1|/1-627
Conservation
+7++68+---------77+31123012231312 34+9+6++---4020--03--1121110-11333+++8++----9-0787+-0332314123---------+++6++++---+++8+*++48+58++-547+7294--3233-------
Quality
Consensus
EEGQAMP+++++++IPWTYQRLYSESILTTMVQVAGKVQEVL++PKEPDVGGGGGLVVLSGGTGTSGRMAFLMSPLSTSVISFNQLCMKGLGQKPLYMPDERCRTETYLIAGGD+TLRSVVASREGTEDSALLHGIEELKKSCVAAGKVFNERFL
Consensus

Supplementary Figure 5. A portion of a multiple alignment of the two Nile rat Gckr proteins to 113 mammalian orthologs displayed in JalView. The second copy of Nile rat Gckr is in the top row, canonical Nile rat Gckr in the second row, followed by human, rat, mouse, and other mammals. Positions 459 and 460 of the alignment correspond to T109 and S110 of the human protein. These constitute a part of the F1P binding site, with T109 involved in polar contacts with hydroxyl substituents of fructopyranose and S110 forming a hydrogen bond with a terminal phosphate oxygen (Pautsch et al. 2013). The predicted sequence of the second putative Nile rat Gckr retrieved from Ensembl has different residues in these positions and their immediate vicinity.

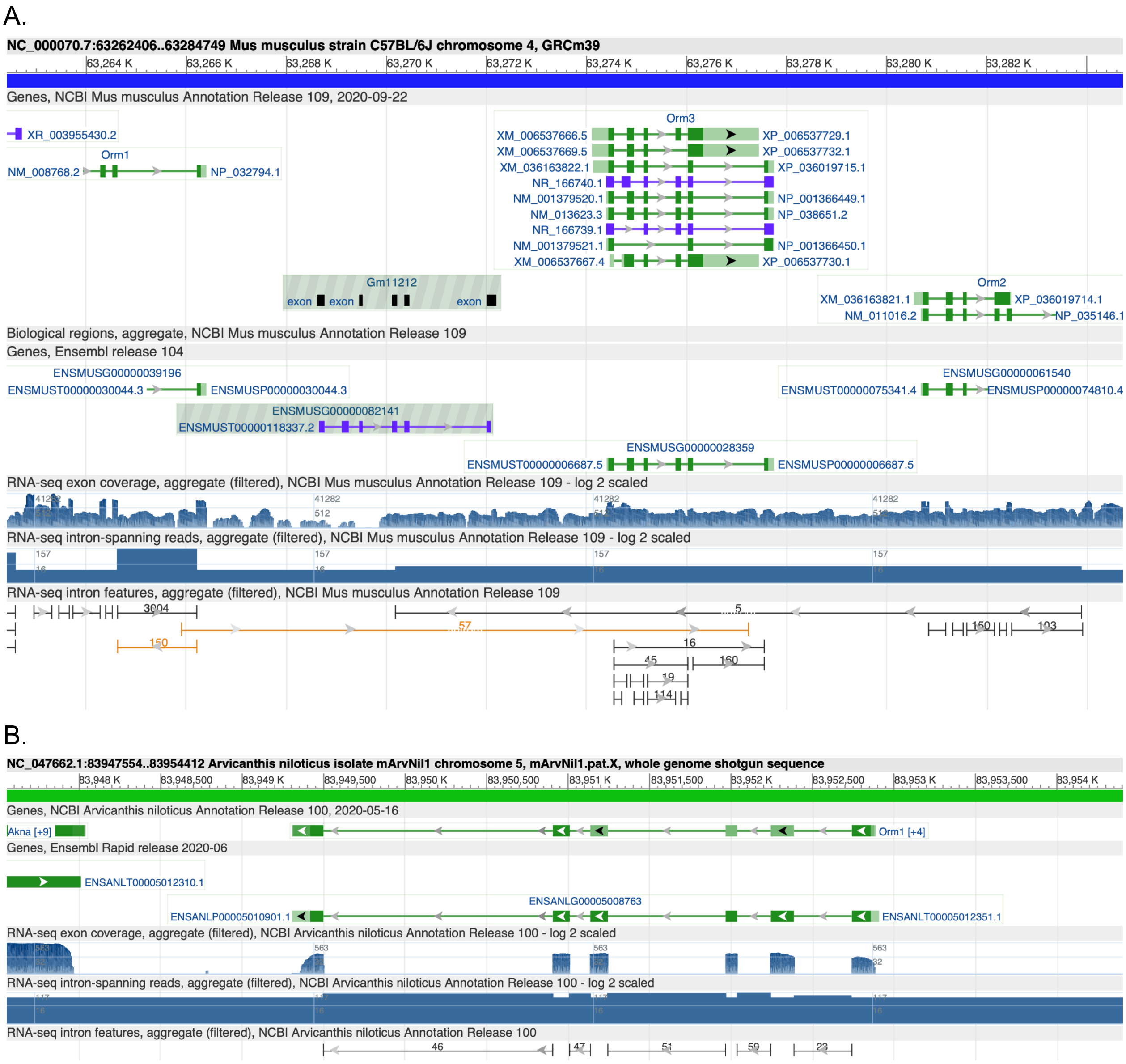
Mouse genes that appear to be absent in Nile rat

A.

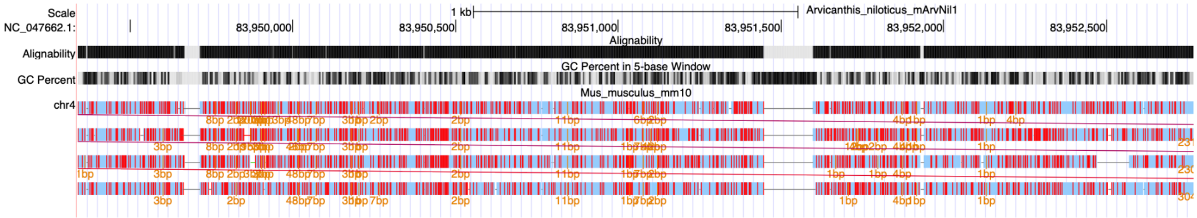

Scale
chr3:
10 kb
mm10
131,240,000|
131,245,000|
131,250,000|
131,255,000|
131,260,000|
131,265,000|
131,270,000|
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
Hadh
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRCm38.p5 (2017-08-04)
Hadh/NM_008212.4
Multiz Alignments of 60 Vertebrates
Rat
Rabbit
Human
Tree_shrew
Dog
Shrew
Elephant
Opossum
Platypus
Chicken
Zebrafish
Repeating Elements by RepeatMasker
RepeatMasker
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups


B.
Scale
10 kb
GCF_011762505.1
NC_047661.1:
108,315,000
108,320,000
108,325,000
108,330,000
108,335,000
108,340,000
Brain Iso-seq hq transcripts
Testis Iso-seq hq transcripts
Paternal PacBio genomic reads
Paternal reads
All gaps of unknown nucleotides (N's), including AGP annotated gaps
Assembly
JAAOMG010000006.1
GC Percent in 5-Base Windows
GC Percent
NCBI RefSeq genes, curated and predicted sets (NM_*, XM_*, NR_*, XR_*, NP_* or YP_*)
XM_034500831.1
NCBI RefSeq genes, predicted subset (XM_* or XR_*)
XM_034500831.1
NCBI RefSeq other annotations (not NM_*, NR_*, XM_*, XR_*, NP_* or YP_*)
Augustus Gene Predictions
g3266.t1
RefSeq mRNAs mapped to this assembly
CpG Islands on All Sequence (Islands < 300 Bases are Light Green)
CpG: 50
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Simple Tandem Repeats by TRF
Simple Repeats
Genomic Intervals Masked by WindowMasker + SDust
WM + SDust
C.
Scale
20 kb
GCA_011750645.1
JAAOME010000004v1:
108,500,000
108,510,000
108,520,000
108,530,000
Brain Iso-seq hq transcripts
Testis Iso-seq hq transcripts
All gaps of unknown nucleotides (N's), including AGP annotated gaps
All Gaps
Maternal PacBio genomic reads
Maternal reads
Assembly
JAAOME010000004.1
70 _
GC Percent in 5-Base Windows
GC Percent
30 _
Augustus Gene Predictions
g3202.t1
g3203.t1
RefSeq mRNAs mapped to this assembly
NM_214331
CpG Islands on All Sequence (Islands < 300 Bases are Light Green)
CpG: 50
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Simple Tandem Repeats by TRF
Simple Repeats
Genomic Intervals Masked by WindowMasker + SDust
WM + SDust

Supplementary Figure 6. Hadh. TOGA failed to find a functional Hadh gene in the primary assembly but did find it in the alternate assembly. **A**. Hadh gene in the mouse has 8 exons. **B**. A putative Hadh gene predicted by RefSeq in Nile rat primary haplotype assembly has only 3 exons. It has been disrupted by a gap in the assembly, which can be seen within the second intron. This gene is not supported by any Iso-seq transcripts. **C**. A full length Iso-seq transcript containing 8 exons and expressed in both brain and testis maps to the alternate haplotype assembly. This indicates haplotype variation and emphasizes the importance of analyzing both haplotypes.


A.
NC_000070.7:63262406..63284749 Mus musculus strain C57BL/6J chromosome 4, GRCm39
63,264 K
63,266 K
63,268 K
63,270 K
63,272 K
63,274 K
63,276 K
63,278 K
63,280 K
63,282 K
Genes, NCBI Mus musculus Annotation Release 109, 2020-09-22
XR_003955430.2
Orm1
NM_008768.2
NP_032794.1
Orm3
XM_006537666.5
XP_006537729.1
XM_006537669.5
XP_006537732.1
XM_036163822.1
XP_036019715.1
NR_166740.1
NM_001379520.1
NP_001366449.1
NM_013623.3
NP_038651.2
NR_166739.1
NM_001379521.1
NP_001366450.1
XM_006537667.4
XP_006537730.1
Gm11212
exon
exon
exon
Orm2
XM_036163821.1
XP_036019714.1
NM_011016.2
NP_035146.1
Biological regions, aggregate, NCBI Mus musculus Annotation Release 109
Genes, Ensembl release 104
ENSMUSG00000039196
ENSMUST00000030044.3
ENSMUSP00000030044.3
ENSMUSG00000061540
ENSMUST00000075341.4
ENSMUSP00000074810.4
ENSMUSG00000082141
ENSMUST00000118337.2
ENSMUSG00000028359
ENSMUST00000006687.5
ENSMUSP00000006687.5
RNA-seq exon coverage, aggregate (filtered), NCBI Mus musculus Annotation Release 109 - log 2 scaled
41282
512
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Mus musculus Annotation Release 109 - log 2 scaled
157
16
RNA-seq intron features, aggregate (filtered), NCBI Mus musculus Annotation Release 109
3004
5
57
150
150
103
16
45
160
19
114
B.
NC_047662.1:83947554..83954412 Arvicanthis niloticus isolate mArvNil1 chromosome 5, mArvNil1.pat.X, whole genome shotgun sequence
83,948 K
83,948,500
83,949 K
83,949,500
83,950 K
83,950,500
83,951 K
83,951,500
83,952 K
83,952,500
83,953 K
83,953,500
83,954 K
Genes, NCBI Arvicanthis niloticus Annotation Release 100, 2020-05-16
Akna [+9]
Orm1 [+4]
Genes, Ensembl Rapid release 2020-06
ENSANLT00005012310.1
ENSANLG00005008763
ENSANLP00005010901.1
ENSANLT00005012351.1
RNA-seq exon coverage, aggregate (filtered), NCBI Arvicanthis niloticus Annotation Release 100 - log 2 scaled
563
32
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Arvicanthis niloticus Annotation Release 100 - log 2 scaled
117
16
RNA-seq intron features, aggregate (filtered), NCBI Arvicanthis niloticus Annotation Release 100
46
47
51
59
23

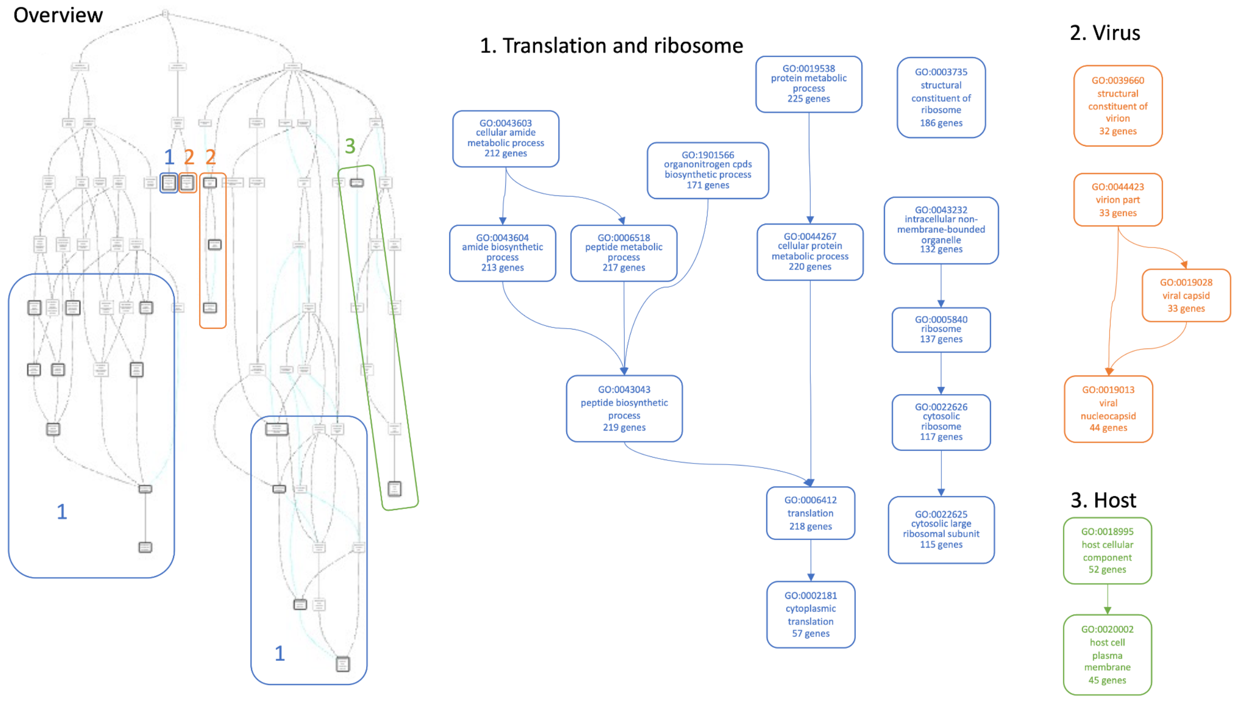
C.



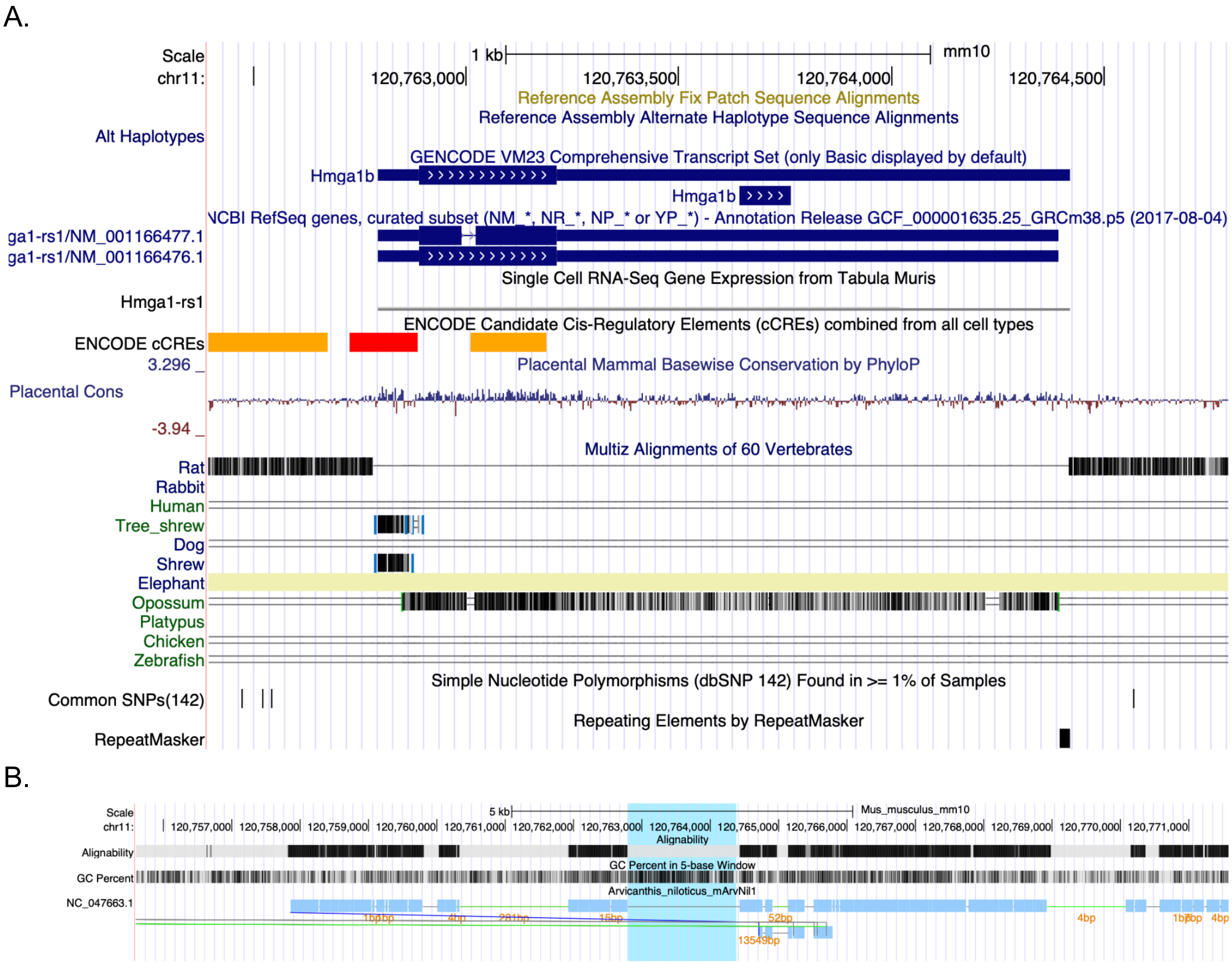
Supplementary Figure 7. Orm2. **A**. *Orm* gene cluster in house mouse visualized in NCBI Genome Data Viewer. The cluster contains 3 protein coding genes and 1 pseudogene. **B**. Nile rat has only one *Orm* gene. **C**. A whole genome alignment of mouse genome vs. Nile rat shows the presence of a 4-fold duplication of the *Orm* region in the mouse genome.

GO enrichment analysis of 1601 Nile rat-specific genes

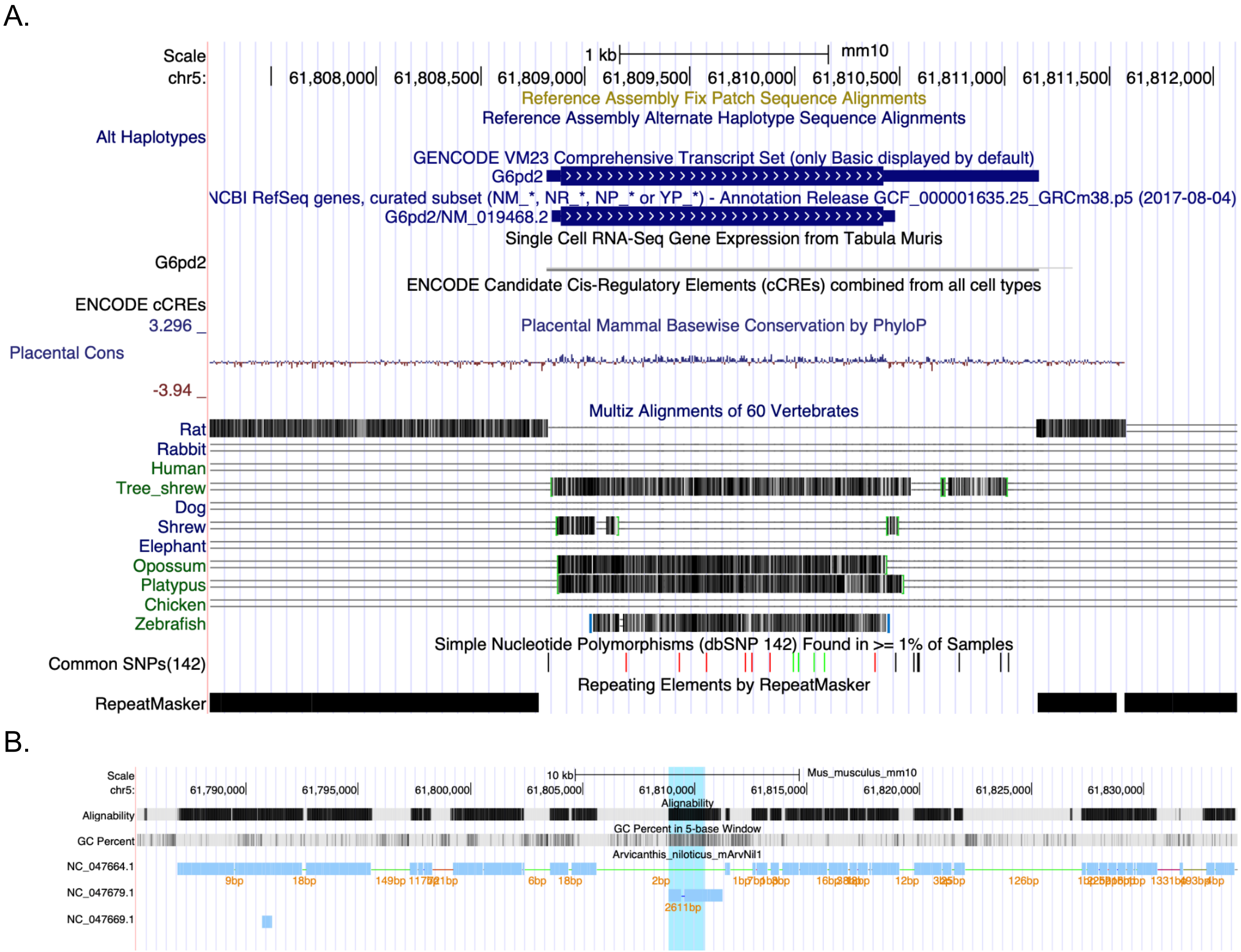
A gene set enrichment analysis of the 1601 genes that do not overlap TOGA projections from house mouse on the NaviGO server (Wei et al. 2017) revealed that several of the most highly enriched GO terms were linked to protein biosynthesis (supplementary data: https://osf.io/fh62m/). Of the top 20 statistically significant terms, the most specific protein biosynthesis-related term was GO:0022625, cytosolic large ribosomal subunit (**supplementary figure NRnotMouseTop20GOTerms**). Examination of several genes linked to this GO term revealed that these are retrogenes derived from ribosomal proteins such as L19, L21, and L23. All of these exist in multiple copies in the Nile rat genome. Multiple copies of retrogenes derived from ribosomal proteins have been known to exist in many other species of mammals (Dharia et al. 2014)


Overview
1
2
2
3
1
1
1. Translation and ribosome
GO:0019538 protein metabolic process 225 genes
GO:0003735 structural constituent of ribosome 186 genes
GO:0043603 cellular amide metabolic process 212 genes
GO:1901566 organonitrogen cpds biosynthetic process 171 genes
GO:0043604 amide biosynthetic process 213 genes
GO:0006518 peptide metabolic process 217 genes
GO:0044267 cellular protein metabolic process 220 genes
GO:0043232 intracellular non-membrane-bounded organelle 132 genes
GO:0005840 ribosome 137 genes
GO:0043043 peptide biosynthetic process 219 genes
GO:0022626 cytosolic ribosome 117 genes
GO:0006412 translation 218 genes
GO:0022625 cytosolic large ribosomal subunit 115 genes
GO:0002181 cytoplasmic translation 57 genes
2. Virus
GO:0039660 structural constituent of virion 32 genes
GO:0044423 virion part 33 genes
GO:0019028 viral capsid 33 genes
GO:0019013 viral nucleocapsid 44 genes
3. Host
GO:0018995 host cellular component 52 genes
GO:0020002 host cell plasma membrane 45 genes

Supplementary Figure 8. Top 20 GO terms overrepresented in Nile rat genes that do not overlap TOGA projections from the house mouse. Overview of the 20 terms in the context of the GO hierarchy is shown on the left. The terms are grouped into 3 broad categories: 1. Translation and ribosome, 2. Virus, and 3. Host. Hierarchical relationships between terms in each category are shown on the right. The number of genes annotated with each term is shown in the corresponding box.


A.
Scale
chr11:
1 kb
mm10
120,763,000
120,763,500
120,764,000
120,764,500
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
Hmga1b
Hmga1b
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRCm38.p5 (2017-08-04)
ga1-rs1/NM_001166477.1
ga1-rs1/NM_001166476.1
Single Cell RNA-Seq Gene Expression from Tabula Muris
Hmga1-rs1
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
3.296 _
Placental Mammal Basewise Conservation by PhyloP
Placental Cons
-3.94 _
Multiz Alignments of 60 Vertebrates
Rat
Rabbit
Human
Tree_shrew
Dog
Shrew
Elephant
Opossum
Platypus
Chicken
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker
B.
Scale
chr11:
5 kb
Mus_musculus_mm10
120,757,000
120,758,000
120,759,000
120,760,000
120,761,000
120,762,000
120,763,000
120,764,000
120,765,000
120,766,000
120,767,000
120,768,000
120,769,000
120,770,000
120,771,000
Alignability
Alignability
GC Percent in 5-base Window
GC Percent
Arvicanthis_niloticus_mArvNil1
NC_047663.1
1bp
4bp
281bp
15bp
52bp
13549bp
4bp
1bp
4bp

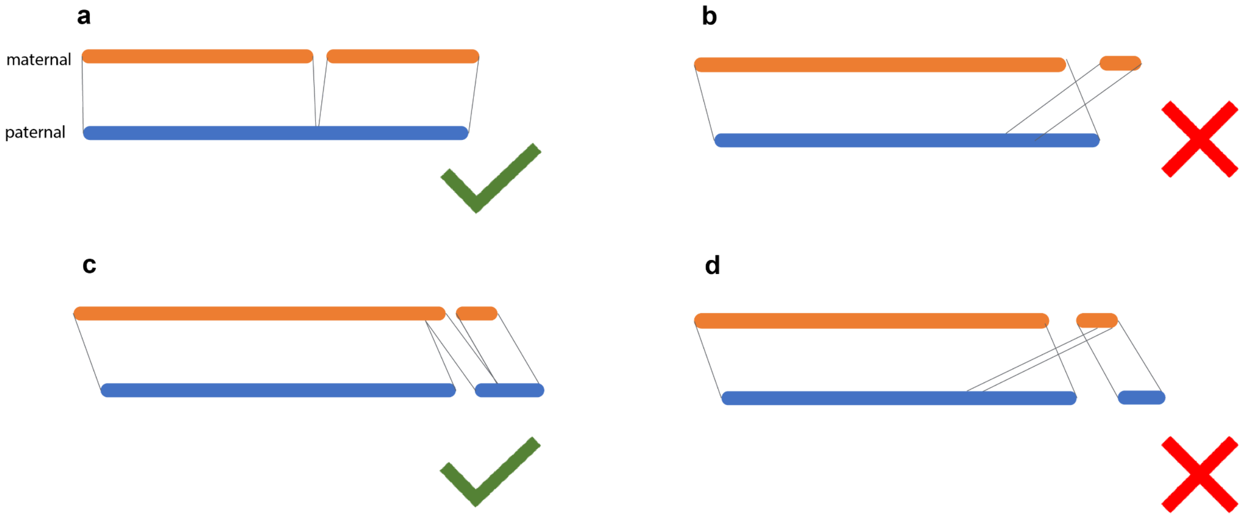
Supplementary Figure 9. Hmga1b. A. Mouse *Hmga1b* gene visualized in the UCSC genome browser. *Hmga1b* is a likely retrogene, as it lacks introns. Additionally, this gene is not present in the Norway rat. **B**. *Hmga1b* gene insertion in mouse relative to Nile rat. A whole genome alignment of Nile rat genome vs. mouse in the vicinity of the mouse *Hmga1b* locus (highlighted). The mouse *Hmga1b* locus does not align to the Nile rat genome, while the surrounding region aligns to chromosome 6.


A.
Scale
chr5:
1 kb
mm10
61,808,000
61,808,500
61,809,000
61,809,500
61,810,000
61,810,500
61,811,000
61,811,500
61,812,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
G6pd2
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRCm38.p5 (2017-08-04)
G6pd2/NM_019468.2
Single Cell RNA-Seq Gene Expression from Tabula Muris
G6pd2
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
3.296
Placental Mammal Basewise Conservation by PhyloP
Placental Cons
-3.94
Multiz Alignments of 60 Vertebrates
Rat
Rabbit
Human
Tree_shrew
Dog
Shrew
Elephant
Opossum
Platypus
Chicken
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
RepeatMasker
B.
Scale
chr5:
10 kb
Mus_musculus_mm10
61,790,000
61,795,000
61,800,000
61,805,000
61,810,000
61,815,000
61,820,000
61,825,000
61,830,000
Alignability
Alignability
GC Percent in 5-base Window
GC Percent
Arvicanthis_niloticus_mArvNil1
NC_047664.1
NC_047679.1
NC_047669.1
9bp
18bp
149bp
6bp
18bp
2bp
12bp
126bp
2611bp

Supplementary Figure 10. G6pd2. A. Mouse *G6pd2* gene visualized in the UCSC genome browser. *G6pd2* is a likely retrogene, as it lacks introns. Additionally, this gene is not present in the Norway rat. **B**. *G6pd2* gene insertion in mouse relative to Nile rat. A whole genome alignment of Nile rat genome vs. mouse in the vicinity of the mouse *G6pd2* locus (highlighted). The mouse *G6pd2* locus aligns to the *G6pd* locus on Nile rat chromosome X, while the surrounding region aligns to chromosome 7. Mouse *G6pd* is likewise located on chromosome X and is the likely parent of the *G6pd2* retrogene.

Manual modifications of haplotype assemblies for heterozygous variation detection

Having several scaffolds not anchored to chromosomes can result in overestimating overall heterozygosity and the number of structural variations. To minimize this effect, we manually modified the haplotype assemblies using a Mummer (Marçais et al. 2018) alignment. Only unanchored scaffolds longer than 500 kb, with alignment similarity over 98% and aligned blocks longer than 10 kb were used in this analysis. We applied the following three rules: 1) when maternal and paternal assemblies had a different orientation, the maternal assembly was reversed; 2) when small scaffolds of maternal (or paternal) assembly uniquely mapped to existing chromosomes of paternal (or maternal) and had no overlap with other mapped scaffolds, these small scaffolds were linked to existing ones with the insertion of 1000 bp gaps; 3) in more complicated cases, especially for those scaffolds that had overlaps with others, the overlapped regions were carefully trimmed as shown in (Supplemental Figure 9). This resulted in an improved version of haplotype assemblies, described by an AGP ("AGP Specification v2.1" n.d.) file available from https://osf.io/v4ypz/.



Supplementary Figure 11. Schematic diagram of trimming alignment. A and c can potentially link unplaced scaffolds, but b and d are not likely to be the right placements.

Comparison of the mouse segmental duplications computed by our workflow to prior work

To gauge how our SD annotations compare with existing ones, we compared our mm10 annotations using the sedef software with those from the UCSC genome browser generated using the software WGAC (Bailey et al. 2004). The unfiltered sedef annotations cover 149 Mb of sequence, or 5.4% of the genome, compared to 215 Mb of sequence (7.9%) by WGAC, with the majority of the difference accounted for by the Y-chromosome annotations where 69.5 Mb of additional SD are annotated by WGAC. On other chromosomes, there are roughly 20 Mb of duplicated sequences annotated by each method individually indicating that our approach mostly replicates the existing annotations of SD sequence on autosomal chromosomes.

The comparison of SD annotations for the long-read assembly of the C57BL strain of house mouse and mm10 illustrates the differences of SD resolution between the whole-genome shotgun (C57BL) assembly, and manually curated mm10 assembly. A total of 204 Mb of sequence is annotated as SD in the C57BL assembly, though the distribution of percent identity of duplications is shifted lower in C57BL (Figure 1a). Furthermore, the average length of SD annotated in C57BL is lower than mm10 (5.6 kb versus 8.4 kb), indicating that long-read sequencing assemblies may contain shorter repetitive DNA that is not removed by masking approaches and annotated as SD. When measuring gene duplication using gene multi-mapping and read depth, a similar number of duplicated genes are observed in the mm10 and C57BL (Figure 1b), indicating that the difference in total bases annotated as duplication may be accounted for by repeat masking or alignment artifacts.

Positively selected single amino-acid substitutions in *Xiap*

Xiap 122 G>A substitution appears to be unique among rodents. However, this residue is not well-conserved across mammalian genomes. Xiap 135T is conserved in 149 mammals, but Nile rat has 135 T>P, and four other mammals have N, I, or S. Notably, S, T, and N are all amino acids with polar uncharged side chains, but P is a nonpolar cyclic amino acid. Human variants have not been reported in residue 135, but nearby disease variants exist in positions 130, 145, and 152 (UniProt). Finally, Xiap 190Y is conserved in all mammals except African woodland thicket rat, Jamaican fruit bat, and Nile rat with Y>F substitution. The African woodland thicket rat is closely related to the Nile rat (Steppan and Schenk 2017). Although no human variant has been reported at position 190, there are nearby disease variants at positions 188 and 189 (Karczewski et al. 2020).

Bibliography

“AGP Specification v2.1.” n.d. Accessed November 21, 2021. https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/.

Bailey, Jeffrey A., Deanna M. Church, Mario Ventura, Mariano Rocchi, and Evan E. Eichler. 2004. “Analysis of Segmental Duplications and Genome Assembly in the Mouse.” *Genome Research* 14 (5): 789–801.

Dharia, Asav P., Ajay Obla, Matthew D. Gajdosik, Amanda Simon, and Craig E. Nelson. 2014. “Tempo and Mode of Gene Duplication in Mammalian Ribosomal Protein Evolution.” *PloS One* 9 (11): e111721.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. 2018. “MUMmer4: A Fast and Versatile Genome Alignment System.” *PLoS Computational Biology* 14 (1): e1005944.

Pautsch, Alexander, Nadja Stadler, Adelheid Löhle, Wolfgang Rist, Adina Berg, Lucia Glocker, Herbert Nar, et al. 2013. “Crystal Structure of Glucokinase Regulatory Protein.” *Biochemistry* 52 (20): 3523–31.

Steppan, Scott J., and John J. Schenk. 2017. “Muroid Rodent Phylogenetics: 900-Species Tree Reveals Increasing Diversification Rates.” *PloS One* 12 (8): e0183070.

Wei, Qing, Ishita K. Khan, Ziyun Ding, Satwica Yerneni, and Daisuke Kihara. 2017. “NaviGO: Interactive Tool for Visualization and Functional Similarity and Coherence Analysis with

Gene Ontology.” *BMC Bioinformatics* 18 (1): 177.