Title: *De novo* assembly of the Mongolian gerbil genome and transcriptome

Authors: Shifeng Cheng[1,2*], Yuan Fu[1,2*], Yaolei Zhang[1,3], Wenfei Xian[1,2], Hongli Wang[2], Benedict Grothe[4], Xin Liu[2,3], Xun Xu[2,3], Achim Klug[5], Elizabeth A McCullagh[5#]

Institutional Affiliations: [1]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;
[2]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China;
[3]China National GeneBank, BGI-Shenzhen, Shenzhen, 518083, China;
[4]Division of Neurobiology, Ludwig-Maximilians-Universitaet Munich, Planegg-Martinsried 82152, Germany
[5]Department of Physiology and Biophysics, School of Medicine, University of Colorado Denver, Aurora, CO, 80045, USA
*co first authors
#corresponding author

Email Addresses for Authors:

Shifeng Cheng: chengshifeng@caas.cn

Yuan Fu: fuyuan@genomics.cn

Yaolei Zhang: zhangyaolei@genomics.cn

Wenfei Xian: xianwenfei@genomics.cn

Hongli Wang: wanghongli@genomics.cn

Benedict Grothe: grothe@lmu.de

Xin Lu: liuxig@genomics.cn

Xun Xu: xuxun@genomics.cn

Achim Klug: achim.klug@ucdenver.edu

Elizabeth A McCullagh: elizabeth.mccullagh@ucdenver.edu

Abstract: (250 words)

BACKGROUND: The Mongolian gerbil (*Meriones unguiculatus*) has historically been used as a model organism for the auditory and visual systems, stroke/ischemia, epilepsy and aging related research since 1935 when laboratory gerbils were separated from their wild counterparts. In this study we report genome sequencing, assembly, and annotation further supported by transcriptome data from 27 different tissues samples.

FINDINGS: The genome was assembled using Illumina HiSeq 2000 and resulted in a final genome size of 2.54 Gbp with contig and scaffold N50 values of 31.4 Kbp and 500.0 Kbp, respectively. Based on the k-mer estimated genome size of 2.48 Gbp, the assembly appears to be complete. The genome annotation was supported by transcriptome data that identified 36 019 predicted protein-coding genes across 27 tissue samples. A BUSCO search of 3023 mammalian groups resulted in 86% of curated single copy orthologs present among predicted genes, indicating a high level of completeness of the genome.

CONCLUSIONS: We report a *de novo* assembly of the Mongolian gerbil genome that was further enhanced by annotation of transcriptome data from several tissues. Sequencing of this genome increases the utility of the gerbil as a model organism, opening the availability of now widely used genetic tools.

Keywords: Gerbil genome, *Meriones unguiculatus*, transcriptome, model organism

DATA DESCRIPTION

62 **Background information on *Meriones unguiculatus***

63 The Mongolian gerbil is a small rodent that is native to Mongolia, southern Russia, and northern China.
64 Laboratory gerbils used as model organisms originated from 20 founders captured in Mongolia in 1935 [1].
65 Gerbils have been used as model organisms for sensory systems (visual and auditory) and pathologies (aging,
66 epilepsy, irritable bowel syndrome and stroke/ischemia). The gerbil's hearing range covers the human
67 audiogram while also extending into ultrasonic frequencies, making gerbils a better model than rats or mice
68 to study lower frequency human-like hearing [2]. In addition to the auditory system, the gerbil has also been
69 used as a model for the visual system because gerbils are diurnal and therefore have more cone receptors than
70 mice or rats making them a closer model to the human visual system [3]. The gerbil has also been used as a
71 model for aging due to its ease of handling, prevalence of tumors, and experimental stroke manipulability
72 [1,4]. Interestingly, the gerbil has been used as a model for stroke and ischemia due to variations in the blood
73 supply to the brain due to an anatomical region known as the "Circle of Willis" [5]. In addition, the gerbil is
74 a model for epileptic activity as a result of its natural minor and major seizure propensity when exposed to
75 novel stimuli [6,7]. Lastly, the gerbil has been used as model for inflammatory bowel disease, colitis, and
76 gastritis due to the similarity in the pathology of these diseases between humans and gerbils [8,9]. Despite its
77 usefulness as a model for all these systems and medical conditions, the utility of the gerbil as a model organism
78 has been limited due to a lack of a sequenced genome to manipulate. This is especially the case with the
79 increased use of genetic tools to manipulate model organisms.

80 Here we describe a *de novo* assembly and annotation of the Mongolian gerbil genome and transcriptome.
81 Recently, a separate group has sequenced the gerbil genome, however our work is further supported by
82 comparisons with an in-depth transcriptome analysis [10]. RNA-seq data were produced from 27 tissues that
83 were used in the genome annotation and deposited in the NCBI SRA database under the project ____. These
84 data provide a draft genome sequence to facilitate the continued use of the Mongolian gerbil as a model
85 organism and to help broaden the genetic rodent models available to researchers.

86 **Animals and Genome Sequencing**

87 All experiments complied with all applicable laws, NIH guidelines, and were approved by the University of
88 Colorado IACUC. Five young adult (postnatal day 65-71) gerbils (three males and two females) were used
89 for tissue RNA transcriptome analysis and DNA genome assembly. In addition, two old (postnatal day 1013
90 or 2.7 years) female gerbil's tissue was used for transcriptome analysis.
91 Genomic DNA was extracted from young adult animal tail and ear snips using a commercial kit (DNeasy
92 Blood and Tissue Kit, Qiagen, Venlo, Netherlands). We then used the extracted DNA to create different pair-
93 end insert libraries of 250 bp, 350 bp, 500 bp, 800 bp, 2 Kb, 4 Kb, 6 Kb, and 10 Kb. These libraries were then
94 sequenced using an Illumina HiSeq2000 Genome Analyzer (Ilumina, San Diego, CA, USA) generating a total
95 of 322.13 Gb in raw data, from which a total of 287.4 Gb of 'clean' data was obtained after removal of
96 duplicates, contaminated reads, and low-quality reads.

97 **Assembly**

98 The gerbil genome was estimated to be approximately 2.48 Gbp using a k-mer-based approach. High-quality
99 reads were then used for genome assembly using the SOAPdenovo (version 2.04) package. The final assembly
100 had a total length of 2.54 Gb and was comprised of 31,769 scaffolds assembled from 114,522 contigs. The
101 N50 sizes for contigs and scaffolds were 31.4 Kbp and 500.0 Kbp, respectively (Table 1). Given the genome
102 size estimate of 2.48 Gbp, genome coverage by the final assembly was likely complete and is consistent with
103 the previously published gerbil genome, which had a total length of 2.523 Gbp [10]. Completeness of the
104 genome assembly was confirmed by successful mapping of the RNA-seq assembly back to the genome
105 showing that 98% of the RNA-seq sequences can be mapped to the genome with >50% sequence in one
106 scaffold.

107 **Transcriptome Sequencing/Assembly/Annotation**

108 Gene expression data were produced to aid in the genome annotation process. Samples from 27 tissues were
109 collected from the seven gerbils described above (Supplementary Figure 1). The tissues were collected after
110 the animals were euthanized with isoflurane and stored on liquid nitrogen until homogenized with a pestle.
111 RNA was prepared using the RNeasy mini isolation kit (Qiagen, Venlo, Netherlands). RNA integrity was
112 analyzed using a Nanodrop Spectrophotometer (Thermo Fisher Waltham, MA, USA) followed by analysis
113 with an Agilent Technologies 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and samples
114 with an RNA integrity number (RIN) value greater than 7.0 were used to prepare libraries which were
115 sequenced using an Ilumina Hiseq2000 Genome Analyzer (Ilumina, San Diego, CA, USA). The sequenced
116 libraries were assembled with Trinity (v2.0.6 parameters: "--min_contig_length 150 --min_kmer_cov 3 --
117 min_glue 3 --bfly_opts '-V 5 --edge-thr=0.1 --stderr'") generating 131,845 sequences with a total length of
118 130,734,893 bp. Quality of the RNA assembly was assessed by filtering RNA-seq reads using SOAPnuke
119 (v1.5.2 parameters: "-l 10 -q 0.1 -p 50 -n 0.05 -t 5,5,5,5") followed by mapping of clean reads to the assembled
120 genome using HISAT2 (v2.0.4) and StringTie (v1.3.0). The initial assembled genes were then filtered using
121 CD-HIT (v4.6.1) with sequence identity threshold of 0.9 followed by a homology search (human, rat, mouse
122 proteins) and TransDecoder (v2.0.1) open reading frame (ORF) prediction. The RNA-seq assembly resulted
123 in 19,737 protein-coding genes with a total length of 29.4 Mbp, which is available in the NCBI Nucleotide,
124 Protein and Gene resources database.

125 **Genome Annotation**

126 Genomic repeat elements of the genome assembly were also identified and annotated using RepeatMasker
127 (v4.0.5 RRID:SCR_012954)[11] and RepBase library (v20.04)[12]. In addition, we constructed a *de novo*
128 repeat sequence database using LTR-FINDER (v1.0.6) [13] and RepeatModeler (v1.0.8) [13] to identify any
129 additional repeat elements using RepeatMasker. A combination of both repeat element identification
130 approaches resulted in a total length of 1016.7 Mbp of the total *M. unguiculatus* genome as repetitive,

131 accounting for 40.0% of the entire genome assembly. The repeat element landscape of *M. unguiculatus*
132 consists of long interspersed elements (LINEs)(27.5%), short interspersed elements (SINEs)(3.7%), long
133 terminal repeats (LTRs)(6.5%), and DNA transposons (0.81%) (Table 2). This is consistent with other rodent
134 species including mouse [14] and rat [15].

135 Protein-coding genes were predicted and annotated by a combination of homology searching, *ab initio*
136 prediction (using AUGUSTUS (v3.1), GENSCAN (1.0), and SNAP (v2.0)), and RNA-seq data (using TopHat
137 (v1.2 with parameters: "-p 4 --max-intron-length 50000 -m 1 –r 20 --mate-std-dev 20 --closure-search --
138 coverage-search --microexon-search") and Cufflinks (v2.2.1 http://cole-trapnell-lab.github.io/cufflinks/))
139 after repetitive sequences in the genome were masked using known repeat information detected by
140 RepeatMasker and RepeatProteinMask. Homology searching was performed using protein data from *Homo*
141 *Sapiens* (human), *Mus musculus* (mouse), *and Rattus norvegicus* (rat) from Ensembl (v80) aligned to the
142 masked genome using BLAT. Genewise (v2.2.0) was then used to improve the accuracy of alignments and to
143 predict gene models. The *de novo* gene predictions and homology-based search were then combined using
144 GLEAN. The GLEAN results were then integrated with the transcriptome dataset using an in-house program
145 (Table 3). This resulted in an identification of a total of 22,998 protein-coding genes with an average transcript
146 length of 23,846.58 bp. There were an average of 7.76 exons per gene with an average length of 197.9 bp and
147 average intron length of 3300.83 bp. The 22,998 protein-coding genes were aligned to several protein
148 databases to begin to identify their possible function. InterProScan (v5.11) was used to align the final gene
149 models to databases (ProDom, ProSiteProfiles, SMART, PANTHER, PRINTS, Pfam, PIRSF,
150 ProSitePatterns, SignalP_EUK, Phobius, IGRFAM, and TMHMM) to detect consensus motifs and domains
151 within these genes. Using the InterProScan results, we obtained the annotations of the gene products from the
152 Gene Ontology database. We then mapped these genes to proteins in SwissProt and TrEMBL (Uniprot release
153 2015.04) using blastp with an E-value <1E-5. We also aligned the final gene models to proteins in KEGG

154 (release 76) to determine the functional pathways for each gene (Table 4). This resulted in 20,760 protein-
155 coding genes that had a functional annotation, or 90.3% of the total gene set.

156 **Quality Assessment**

157 In addition to measuring standard assembly quality metrics, genome assembly and annotation quality were
158 further assessed by comparison with closely related species, gene family construction, evaluation of
159 housekeeping genes, and Benchmarking Universal Single-Copy Orthologs (BUSCO) search. The assembled
160 gerbil genome was compared with other closely related model organisms including mouse, rat, and hamster
161 (Table 5). The genomes from these species varied in size from 2.3 to 2.8 Gbp. The total number of annotated
162 proteins in gerbil (20,760) is most similar to mouse (22,598), followed by rat (23,347), and then hamster
163 (24,238). Gene family construction was performed using Treefam (http://www.treefam.org/) (Figure 1). This
164 analysis showed that single-copy orthologs in gerbil are similar to mouse and rat. To examine housekeeping
165 genes we downloaded 2169 human housekeeping genes from (http://www.tau.ac.il/~elieis/HKG/) and
166 extracted corresponding protein sequences to align to the gerbil genome using blastp (v.2.2.26). We found
167 there were 2141 genes consistent between human and gerbil housekeeping genes (this is similar to rat (2153)
168 and mouse (2146). Lastly, we employed BUSCO (v1.2) to search 3023 mammalian groups. Of these groups,
169 86% complete BUSCO groups can be detected in the final gene set. The presence of 86% complete
170 mammalian BUSCO gene groups suggests a high level of completeness of this gerbil genome assembly. A
171 BUSCO search was also performed for the gerbil transcriptome data resulting in detection of 82% complete
172 BUSCO groups in the final transcriptome dataset (Table 6). Based on the results from the quality metrics
173 described above, we are confident of the quality of the data for this assembly of the gerbil genome and
174 transcriptome.

175

In summary, we report a fully annotated Mongolian gerbil genome sequence assembly enhanced by transcriptome data from several different gerbils and tissues. The gerbil genome and transcriptome adds to the availability of alternative rodent models that may be better models for diseases than rats or mice. Additionally, the gerbil is an interesting comparative rodent model to mouse and rat since it has many traits in common, but also differs in seizure susceptibility, low-frequency hearing, cone visual processing, stroke/ischemia susceptibility, gut disorders and aging. Sequencing of the gerbil genome and transcriptome opens these areas to molecular manipulation in the gerbil and therefore better models for specific disease states.

Availability of supporting data

Genome annotation results are available at the NCBI, and supporting materials, which include transcripts and genome assembly, are available at the *Gigascience* database, *Giga*DB.

Additional files

Additional file 1: Table S1 Tissues analyzed for RNA-seq data

Abbreviations

bp: base pair

BUSCO: Benchmarking Universal Single-Copy Orthologs

CDS: coding sequence

LINEs: long interspersed elements

LTRs: long terminal repeats

Myr: million years

NCBI: National Center for Biotechnology Information

RefSeq: Reference sequence

RNA-seq: high-throughput messenger RNA sequencing

RIN: RNA integrity number

SINEs: short interspersed elements

Competing Interests: The authors declare that they have no competing interests.

Funding
Authors' contributions
Acknowledgements

References

1. Cheal ML. The gerbil: a unique model for research on aging. Exp Aging Res. 1986;12:3–21.

2. Ryan A. Hearing sensitivity of the mongolian gerbil, Merionesunguiculatis. The Journal of the Acoustical Society of America. Acoustical Society of America; 1976;59:1222–6.

3. Govardovskii VI, Röhlich P, Szél A, Khokhlova TV. Cones in the retina of the Mongolian gerbil, Meriones unguiculatus: an immunocytochemical and electrophysiological study. Vision Res. 1992;32:19–27.

4. Vincent AL, Rodrick GE, Sodeman WA. The Mongolian gerbil in aging research. Exp Aging Res. Routledge; 2007;6:249–60.

5. Small DL, Buchan AM. Animal models. Br Med Bull. Oxford University Press; 2000;56:307–17.

6. Bertorelli R, Adami M, Ongini E. The Mongolian gerbil in experimental epilepsy. Ital J Neurol Sci. 1995;16:101–6.

7. Löscher W. Genetic animal models of epilepsy as a unique resource for the evaluation of anticonvulsant drugs. A review. Methods Find Exp Clin Pharmacol. 1984;6:531–47.

8. Bleich E-M, Martin M, Bleich A, Klos A. The Mongolian gerbil as a model for inflammatory bowel disease. Int J Exp Pathol. Blackwell Publishing Ltd; 2010;91:281–7.

9. Hirayama F, Takagi S, Kusuhara H, Iwao E, Yokoyama Y, Ikeda Y. Induction of gastric ulcer and intestinal metaplasia in mongolian gerbils infected with Helicobacter pylori. J. Gastroenterol. 1996;31:755–7.

10. Zorio DAR, Monsma S, Sanes DH, Golding NL, Rubel EW, Wang Y. De novo sequencing and initial annotation of the Mongolian gerbil (Meriones unguiculatus) genome. Genomics. Elsevier Inc; 2018.

11. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2009;12:1269–4.10.14.

12. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

13. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999;27:573–80.

14. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics & Development. 1999;9:657–63.

15. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004;428:493–521.

NCBI. NCBI Annotation Release Meriones unguiculatis, 2017; https://www.ncbi.nlm.nih.gov/genome/annotation euk/, (2 October 2017, date last accessed).

262 **Table 1 Global statistics of the Mongolian gerbil genome**

Statistic	Value
Size (Gb)	2.54
Scaffold number (>2000bp)	31769
Scaffold N50 (Kb)	500.0
Contig number (>2000bp)	114522
Contig N50 (Kb)	31.4

263

264 **Table 2 Summary of mobile element types**

Type	Length (Kb)	Percentage of the genome (%)
DNA	20,498	0.81
LINE	697,185	27.5
SINE	94,229	3.7
LTR	164,504	6.5
Other	40,254	1.6
Total	1,016,671	40.0

265

266

267

268

269

270

271

272

273

274

275 **Table 3 General statistics of predicted protein-coding genes**

	Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	SNAP	76858	42227.63	742.83	5.52	134.62	9182.18
	AUGUSTUS	24675	19838.68	1133.22	5.61	201.97	4056.79
	GENESCAN	49390	24183.55	1023.1	6.25	163.54	4406.54
Homolog	*Mus musculus*	22728	26977.32	1465.18	8.02	182.61	3632.46
	Rattus norvegicus	23686	23564.96	1336.56	7.43	179.83	3455.8
	Homo sapiens	17131	31217.18	1580.27	9.11	173.55	3656.27
	GLEAN	19893	18835.39	1418.26	7.72	183.69	2691.49
	Transcriptome	36019	33752.29	1758.58	10.74	163.77	3285.43
	Final set	22998	23846.58	1535.48	7.76	197.9	3300.83

276

277 **Table 4 Functional annotation of the final gene set**

	Number	Percent (%)
Total	22,998	100
InterPro	18,570	80.7
GO	14,591	63.4
KEGG	17,572	76.4
Swissprot	20,113	87.5
TrEMBL	20,666	89.9
Annotated	20,760	90.3
Unannotated	2238	9.7

278

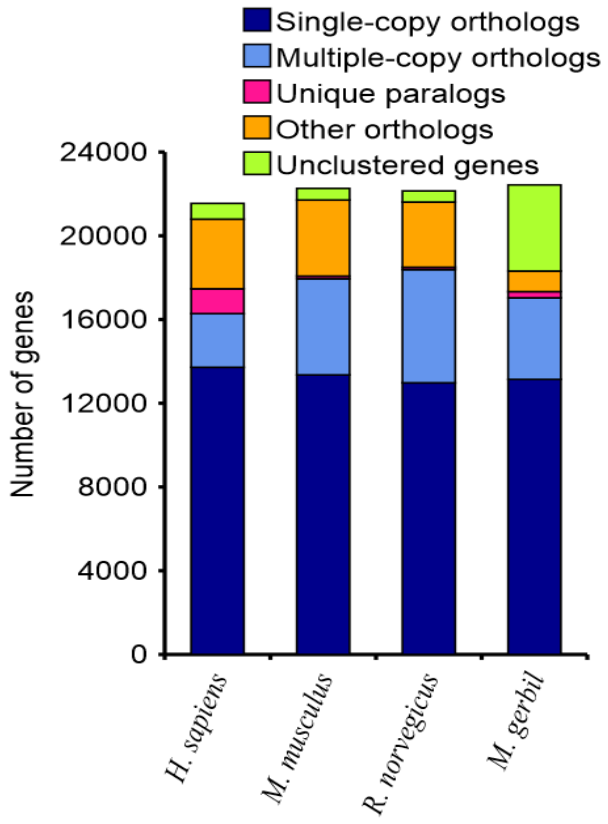
279

280

281

282 **Table 5 Genome annotation comparisons with other model organisms**

Species	Common name	Protein coding genes	Assembly Size	Divergence time to gerbils, Myr	RefSeq assembly accession	Annotation release ID
Meriones unguiculatus	Mongolian gerbil	20,760	2,537,533,819	--		
Mus musculus	mouse	22,598	2,818,974,548	22.5	GCF_000001635.26	106
Rattus norvegicus	rat	23,347	2,870,184,193	22.5	GCF_000001895.5	106
Cricetulus griseus	Chinese hamster	24,238	2,358,151,106	25	GCA_900186095.1	102



283

284 **Figure 1 Gene Family Construction**. The number of genes is similar between species compared (human,
285 mouse, rat, and gerbil.

286

287

288 **Table 6 Completeness of gerbil genome and transcriptome assembly as assessed by BUSCO**

	Genome	Transcriptome
Complete BUSCOs	2601	2508
Duplicated BUSCOs	55	46
Fragmented BUSCOs	170	293
Missing BUSCOs	252	222
Total BUSCO groups searched	3023	3023

289

290 **Supplementary Figure 1: Tissues sampled for RNA transcriptome**

Tissue	Run_accession	Sex	Age (postnatal day)	Data size (Mbp)
Lung		M	71	6733.54
Lung		F	1013	6347.26
Occipital lobe		F	1013	6231.73
Occipital lobe		F	70	5820.49
Kidney		F	1013	6412.73
Kidney		M	70	5609.90
Olfactory bulb		M	71	7467.99
Olfactory bulb		F	70	5576.19
Striatum		M	71	4596.98
Striatum		F	1013	5456.08
Striatum		M	71	6010.27
Striatum		F	71	8508.27
Cerebellum		F	1013	6021.12
Cerebellum		M	65	6724.73
Inferior colliculus		F	1013	5637.18
Inferior colliculus		M	71	6296.64

Liver	F	1013	5077.32
Liver	F	1013	6280.63
Spleen	M	71	9051.52
Spleen	F	1013	7943.03
Spleen	F	1013	6702.24
Frontal cortex	M	65	5895.65
Frontal cortex	F	1013	7202.13
Hippocampus	M	70	5189.69
Auditory brainstem	F	66	7332.74
Brainstem	M	65	5820.49
Parietal cortex	M	65	6786.95