



OPEN

DATA DESCRIPTOR

# Genome assembly and annotation of the naked mole rat *Heterocephalus glaber* reared in Japan

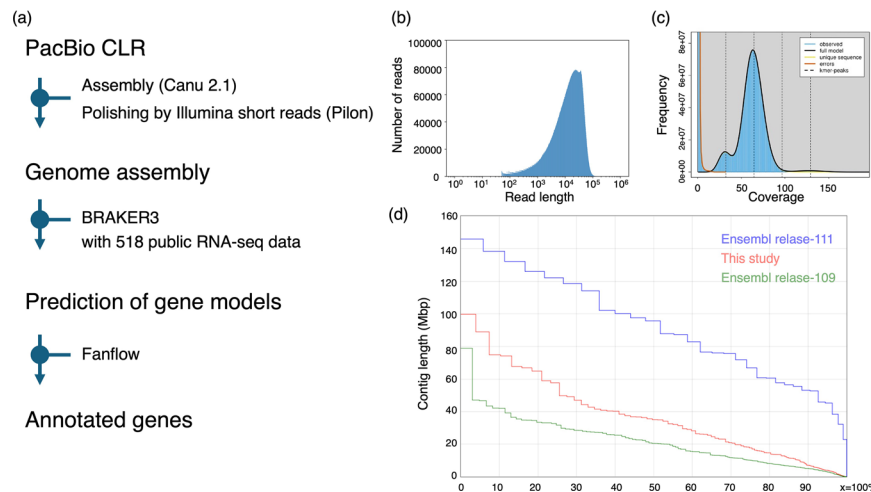
Kouhei Toga<sup>1,2</sup>, Kaori Oka<sup>3,4</sup>, Hiroyuki Tanaka<sup>5</sup>, Takehiko Itoh<sup>5</sup>, Atsushi Toyoda<sup>6,7</sup>, Hidemasa Bono<sup>1,2</sup>✉ & Kyoko Miura<sup>3,4</sup>✉

The naked mole rat (NMR, *Heterocephalus glaber*) is a eusocial rodent that is native to northeastern Africa. NMRs exhibit extraordinary traits such as longevity, resistance to age-related decline, and remarkable hypoxia tolerance. Although the reference genome of this species has been determined because of its unique characteristics, the significance or role of intraspecific genomic variations remains unknown. In this study, we used PacBio long-read sequencing to generate a genome assembly of NMR reared in Japan. The assembled genome is 2.56 Gb. Benchmarking Universal Single-Copy Orthologs (BUSCO) revealed high completeness (95.2%). BRAKER3 estimated 26,714 protein-coding genes, and we successfully added functional annotations for 26,232 protein-coding genes using the functional annotation workflow. We identified 417 gene models that were previously undetectable in the reference genome of this species. We also identified structural and amino acid sequence variations between our assembly and the reference genome, suggesting the presence of intraspecific genomic variations. This new genomic resource could help uncover the molecular mechanisms underlying the behavioral and physiological traits of NMR.

## Background & Summary

The naked mole rat (NMR) has received significant attention in the field of biology because of its unique behavior and physiology. NMR is the longest-lived rodent, exhibiting extraordinary longevity and resistance to age-related diseases, including cancer, with a maximum lifespan of 40 years despite its small size<sup>1</sup>. This remarkably long lifespan deviates from the typical positive correlation observed between body mass and lifespan in mammals<sup>2</sup>. The NMR is a eusocial rodent that lives cooperatively in subterranean colonies consisting of a single breeding female, one to three breeding males, and non-reproductive workers<sup>3</sup>. Adapted to group-living in semi-enclosed subterranean environments, they exhibit significant tolerance to both hypoxia and anoxia<sup>4</sup>. Comparative genomic analysis between NMR and other mammals revealed that genomic mutations in NMR contribute to longevity, adaptation to the subterranean environment, and resistance to carcinogenesis<sup>5–10</sup>. Additionally, amino acid substitutions are enriched at disease-causing sites in human<sup>6</sup>. Therefore, the NMR genome is an important resource for identifying genetic factors related to its unique ecology, physiology, and resistance to age-related diseases.

<sup>1</sup>Laboratory of BioDX, PtBio Co-Creation Research Center, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan. <sup>2</sup>Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan. <sup>3</sup>Department of Aging and Longevity Research, Faculty of Life Sciences, Kumamoto University, 2-2-1 Honjo, Chuo-ku, Kumamoto city, Kumamoto, 860-0811, Japan. <sup>4</sup>Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, 812-8582, Japan. <sup>5</sup>Department of Life Science and Technology, Institute of Science Tokyo, Tokyo, 152-8550, Japan. <sup>6</sup>Comparative Genomics Laboratory, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan. <sup>7</sup>Advanced Genomics Center, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan. ✉e-mail: [bonohu@hiroshima-u.ac.jp](mailto:bonohu@hiroshima-u.ac.jp); [miura.kyoko.314@m.kyushu-u.ac.jp](mailto:miura.kyoko.314@m.kyushu-u.ac.jp)



**Fig. 1** Strategies for genome assembly and annotation in this study (a). The main tools adopted in assembly and annotation are described. Histogram of read lengths after log transformation (b). GenomeScope2 results for genome size, heterozygosity rate, and repeat sequences (c). Heterozygous and homozygous peaks are shown by the first and second peaks, respectively. Nx plots for our assembly and Ensembl assemblies (release-111 and –109) (d). Different colored lines indicate different assemblies.

The concept of a reference genome is based on a single representative genome sequence for each species and is broadly applied to many organisms, as seen in the Earth BioGenome Project<sup>11</sup>. However, this concept does not account for intraspecific variation. Therefore, it remains unclear how genomic structures are preserved within the same species. Supporting these limitations, intraspecific genomic variations and chromosomal rearrangements have been observed in mammals, larvaceans, and insects<sup>12–14</sup>. Notably, maximal genetic distances within NMR species are approaching interspecific values observed among other Bathyergid species, based on calculations for the cytochrome-b gene between the Ethiopian and southern Kenyan populations<sup>15</sup>. This suggests the accumulation of genetic differences within the NMR species. Additionally, the rivers in Kenya may cause reproductive isolation, leading to local genetic differences<sup>16</sup>. However, genome resources are currently insufficient to identify genomic differences in NMR species. To address this gap, we sequenced the whole genome of a single NMR individual reared in Japan using Illumina short reads and PacBio long reads. Gene prediction and functional annotation were performed using BRAKER3<sup>17</sup> and the functional annotation workflow (Fanflow)<sup>18</sup> pipeline. We verified the presence of variations at both the genome and gene levels between NMR species, demonstrating that our genome assembly provides a valuable resource for advancing the understanding of genomic functions related to NMR physiology and behavior.

## Methods

**Sample collection, sequencing, and genome assembly.** NMRs were maintained at  $30 \pm 0.5^\circ\text{C}$  and  $55 \pm 5\%$  humidity with 12 h light and 12 h dark cycles at Kumamoto University. All animal experiments were performed in accordance with the Guide for the Care and Use of Laboratory Animals (United States National Institutes of Health). The ethics committee of Kumamoto University approved all the procedures (approval No. A2020-042, A2022-079 and A2024-063). The naked mole-rats (*Heterocephalus glaber*) used in this study originate from colonies previously maintained in the laboratory of Dr. Kazuo Okanoya (formerly at RIKEN, now at Teikyo University). These animals are derived from two lineages that were exported from northeastern Africa before 1992 and subsequently bred and maintained in research institutions and zoos in South Africa and the United States. Although the exact geographic origin of the wild-caught founders is unknown, they are believed to have come from the species' native range, which spans Ethiopia, Kenya, Somalia, and Djibouti. Since 2011, the colony has been continuously maintained in our laboratory. Genomic DNA was extracted from the quadriceps muscle tissues from a male NMR using a QIAGEN Genomic-tip (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. Our strategy of genome assembly and the annotation was described (Fig. 1a). The quality and quantity of the extracted genomic DNA were assessed using a Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Pippin Pulse system (Agilent Technologies). DNA was sheared into 30–100 kb fragments using a g-tube device (Covaris Inc., MA, USA). A continuous long-read (CLR) single-molecule real-time (SMRT) bell library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Bioscience, Menlo, CA, USA) according to the manufacturer's instructions. The CLR libraries were size-selected using the BluePippin system (Saga Science, MA, USA) with a lower cutoff of 30 kb. The library was run on the PacBio Sequel II platform using two SMRT Cell 8Ms with Binding Kit 2.0, and Sequencing Kit 2.0 with 20-h movies. A total of 15.2 M subreads were generated from two SMRT cells, yielding 282.41 Gb raw data (109.9-fold whole-genome coverage). The average read length of CLR was 18.5 kb (Fig. 1b). Additionally, genomic DNA was fragmented to an average size of 500–600 bp using a focused ultrasonicator M220 (Covaris Inc., MA., USA). Paired-end libraries with 450–550 bp insert sizes were constructed using the TruSeq DNA PCR-Free Library Prep Kit (Illumina, CA, USA) and size-selected on an agarose gel using a Zymoclean Large Fragment DNA Recovery

|                                             | parameter                     | this study         | Ensembl release-111 | Ensembl release-109 (HetGla_female_1.0) |
|---------------------------------------------|-------------------------------|--------------------|---------------------|-----------------------------------------|
| Assembly statistics                         | Length                        | 2,561,275,506      | 2,500,150,894       | 2,618,204,639                           |
|                                             | N50                           | 35,055,060         | 95,752,777          | 20,532,749                              |
|                                             | L50                           | 25                 | 11                  | 42                                      |
|                                             | No. of contigs or scaffold    | 383                | 32                  | 4,229                                   |
|                                             | Largest contig                | 99,770,558         | 145,805,425         | 78,885,850                              |
|                                             | GC %                          | 40.01              | 39.97               | 35.55                                   |
| BUSCO results (OrthoDB v10 Glires database) | Complete [single, duplicated] | 95.2% [90.6%,4.6%] | 92.6% [88.2%,4.4%]  | 94.7% [90.3%,4.4%]                      |
|                                             | Fragmented                    | 0.8%               | 1.2%                | 1.0%                                    |
|                                             | Missing                       | 4.0%               | 6.2%                | 4.3%                                    |

**Table 1.** Comparison of assembly statistics with reference genomes in Ensembl.

|                                             | Parameter                          | This study          | Ensembl release-111 | Ensembl release-109 |
|---------------------------------------------|------------------------------------|---------------------|---------------------|---------------------|
|                                             | The number of protein-coding genes | 26,714              | 20,742              | 23,216              |
|                                             | The number of mRNA                 | 43,234              | 41,841              | 72,114              |
| BUSCO results (OrthoDB v10 Glires database) | Complete [single, duplicated]      | 92.2% [49.2%,43.0%] | 90.1% [38.6%,51.5%] | 93.8% [68.6%,25.2%] |
|                                             | Fragmented                         | 0.8%                | 1.9%                | 1.2%                |
|                                             | Missing                            | 7.0%                | 8.0%                | 5.0%                |

**Table 2.** Comparison of gene prediction statistics between our assembly and the reference genome in Ensembl.

Kit (Zymo Research, CA., USA). The final libraries were sequenced following the 2 × 250 bp paired-end protocol for NovaSeq. 6000 systems (Illumina, San Diego, CA, USA), yielding 429.39 Gb of raw data. This data size corresponds to 167.13 times the genome coverage. GenomeScope2<sup>19</sup> ( $k = 32$ ) estimated the genome size to be 2.569 Gb, with 18% consisting of repeat sequences (Fig. 1c). The rate of heterozygosity was 0.173%.

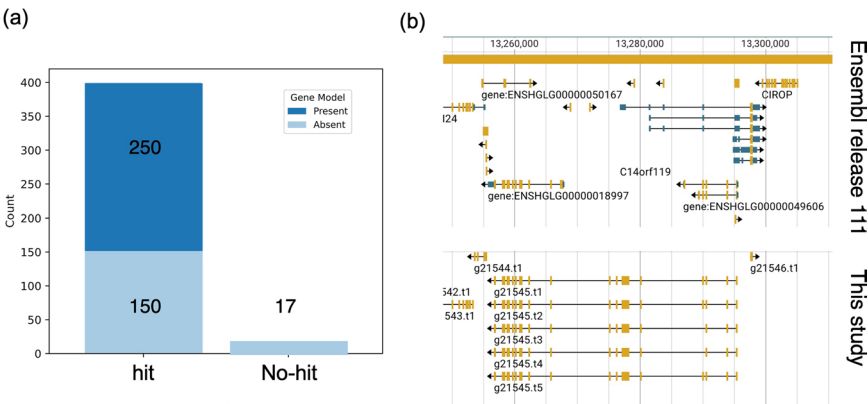
Canu 2.1<sup>20</sup> was used for *de novo* assembly of PacBio CLR reads with parameters ‘-genomesize 3.0 g -coverage 200 -minlength 1000 -raw -pacbio’. Three rounds of polishing were performed using Pilon<sup>21</sup> with default settings and with the aforementioned Illumina short reads. Purge\_dups<sup>13</sup> was used with default parameters on the polished assembly to purge haplotigs, resulting in a genome size of 2.561 Gb, which was comparable to that estimated by GenomeScope2<sup>19</sup> (Table 1). Merquy v1.3 ( $k = 21$ )<sup>22</sup> was used for the validation of our genome assembly. Completeness and quality value was 98.2% and 50.6, respectively. We compared the assembly statistics of the current and conventional assemblies in Ensembl release-111<sup>23</sup> and release-109<sup>24</sup> (Table 1 and Fig. 1d). Contiguity of our genome assembly was inferior to that of the current Ensembl genome (release-111), but Benchmarking Universal Single-Copy Orthologs (BUSCO)<sup>25</sup> showed our assembly had the highest completeness (95.2%).

**Gene prediction and functional annotation.** Since repetitive sequences can lead to overestimation of gene predictions, repetitive sequences were filtered using RepeatModeler v2.0.3 and RepeatMasker v4.0.6 in Extensive *de novo* TE Annotator v2.0.1<sup>26</sup>, and were soft-masked. We identified 978 Mb of repetitive sequences, accounting for 38.19% of the genome assembly<sup>27</sup>. Gene prediction was performed using BRAKER3<sup>17</sup>. Public RNA sequencing (RNA-Seq) data<sup>28</sup> and Vertebrate protein dataset ([https://bioinf.uni-greifswald.de/bioinf/partitioned\\_odb11/](https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/), accessed on October 10, 2023) were used as extrinsic evidence for BRAKER3<sup>17</sup>. Metadata of public RNA-Seq data was collected using SRA Run Selector in National Center for Biotechnology Information (NCBI), and manually filtered to retain only samples with well-defined metadata. Sequence read archive (SRA) retrieval of public RNA-Seq and conversion to FASTQ files was performed using SRA toolkit v3.0.5 (<https://github.com/ncbi/sra-tools>, accessed on June 30, 2023). Trimming and quality control of the FASTQ files were performed using Trim Galore! v0.6.10<sup>29</sup> with ‘-fastqc,-trim1’ option. Mapping RNA-Seq reads to the genome was performed using HISAT2 v2.2.1<sup>30</sup> with ‘-q,-dta’ option, and then obtained SAM files were converted to BAM files using SAMtools v1.12<sup>31</sup>. The obtained BAM files were supplied to the BRAKER3 workflow. Gene set coverage was assessed using BUSCO v5.2.2<sup>25</sup> with ‘-l glires\_odb10’. BRAKER3 predicted 26,714 protein-coding genes in the assembly (Table 2). BUSCO completeness values were comparable to those of the current and conventional genomes in Ensembl.

Fanflow<sup>18</sup> was used as the reference protein set in Ensembl release-111. This workflow used GGSEARCH v36.3.8 g in the FASTA package (<https://fasta.bioch.virginia.edu/>, accessed on July 19, 2023) to compare amino acid sequences against datasets from humans, mice, guinea pigs, female NMR, male NMR, and UniProtKB/Swiss-Prot. Domain searches were conducted using HMMSCAN in HMMER v3.3.2, with Pfam v35.0, which served as the domain database<sup>32</sup>. Of the 26,714 predicted genes, 98% (26,232/26,714) showed matches to at least one protein in the reference protein sets or the Pfam database, leaving only a small number of unclassified proteins (Table 3).

| Annotation category                | Reference                 | The number of gene annotated |
|------------------------------------|---------------------------|------------------------------|
| 1. Gene homolog (GGSEARCH)         | Human                     | 18,730                       |
|                                    | Mouse                     | 18,251                       |
|                                    | Guinea pig                | 17,080                       |
|                                    | NMR (female)              | 17,897                       |
|                                    | NMR (male)                | 17,983                       |
|                                    | UniProtKB/Swiss-Prot      | 19,149                       |
|                                    | At least one of the above | 26,120                       |
| 2. Genes annotated only by HMMSCAN | Pfam domain database      | 112                          |
| 1 + 2                              | —                         | 26,232                       |
| Hypothetical protein               | No hit                    | 482                          |
| Total                              |                           | 26,714                       |

**Table 3.** Summary of the functional annotation results.

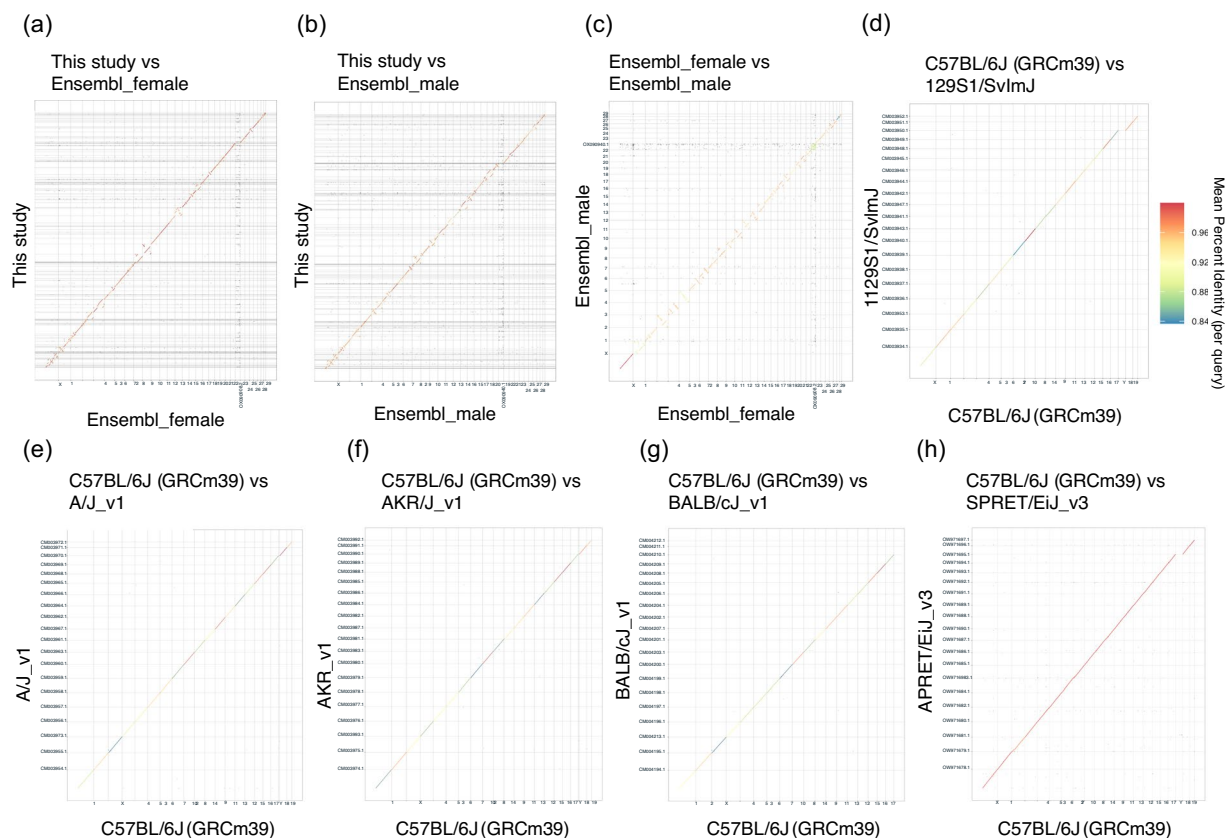


**Fig. 2** Searches for the gene model differences between our assembly and the Ensembl assembly (release-111). The number of regions mapped by miniprot is shown, with columns colored differently to indicate the presence of overlaps with the Ensembl gene model (a). Gene structures of g21545 and their corresponding regions in the Ensembl genome in Ensembl release-111 are shown (b). Yellow and blue boxes indicate the exons and untranslated regions (UTRs), respectively.

**Comparison of gene models between our assembly and the Ensembl assembly.** During Fanflow execution, 417 transcripts with counterparts in the protein sets of all Ensembl reference species were found, except for those in NMR (female and male). This implied that these transcript models were not present in the Ensembl NMR genome. To confirm whether these 417 transcripts were missing from the NMR genome in Ensembl, we ran miniprot<sup>33</sup> to map the translated proteins predicted from our assembly onto the female NMR genome in Ensembl. Subsequently, bedtools intersect (v 2.31.1)<sup>34</sup> with ‘-s -loj’ options was used to identify overlapping regions between the mapped proteins and GFF3 file of the female NMR genome in Ensembl release-111 (accessed on May 30, 2024). Amino acid sequences were aligned using MAFFT<sup>35</sup> with ‘-auto’ option, and then the alignments were visually examined to evaluate the differences of gene structures between our assembly and the female NMR genome in Ensembl.

Of the 417 transcripts, 400 were mapped to the female NMR genome in Ensembl, suggesting that these 400 transcripts are encoded in Ensembl assembly. The 400 transcripts were further categorized into two groups: one consisting of those present in the gene model of the female NMR genome in Ensembl (250 transcripts), and the other consisting of those absent from the gene models (150 transcripts) (Fig. 2a)<sup>36</sup>. Among the 250 transcripts, 39 overlapped across multiple gene models in the female NMR genome of Ensembl (e.g., the g21545 gene, as shown in Fig. 2b). The exon of g21545 overlapped with those of ENSHGLG00000049606 and ENSHGLG00000018997. Meanwhile, remaining 17 transcripts were not mapped by miniprot (Fig. 2a)<sup>36</sup>, suggesting that these transcripts might be absent from the Ensembl assembly. These results show that previously undetected gene models in the Ensembl genome were included in our assembly.

**Comparison of genomic differences between our assembly and the Ensembl assembly.** Dot plots were generated to visualize the genomic differences between our assembly and the Ensembl assembly<sup>23,37</sup>. To evaluate genomic differences within the NMR, dot plots were generated among mouse strains (C57BL/6J vs. 129S1/SvImJ\_v1, C57BL/6J vs. A/J\_v1, C57BL/6J vs. AKR/J\_v1, and C57BL/6J vs. BALB\_cJ\_v1) and between mouse species (C57BL/6J vs. *Mus spretus* (SPRET/EiJ)). The same process was applied to combinations of mouse strains to evaluate intraspecific differences in NMR. C57BL/6J in Ensembl release-111<sup>38</sup>, GCA\_001624185.1 (129S1/SvImJ\_v1)<sup>39</sup>, GCA\_001624215.1 (A/J\_v1)<sup>40</sup>, GCA\_001624295.1 (AKR/J\_v1)<sup>41</sup>, GCA\_001632525.1



**Fig. 3** Differences in genome structures within or between species. Dot plots indicate the comparisons of genome structures between our assembly and Ensembl assembly release-111 (female and male) (a–c), between C57BL/6J (GRCm39) and each mouse strain (d–g), and between C57BL/6J (GRCm39) and SPRET/EiJ (*Mus spretus*) (h). The color of plots indicates the mean percent identity (per query).

(BALB/cJ\_v1)<sup>42</sup>, and GCA\_921997135.2 (SPRET\_EiJ\_v3)<sup>43</sup> were used in this analysis. Minimap2 v2.28<sup>44</sup> was used to align genome sequences with the ‘-cx asm5’ option. Subsequently, dot plots were visualized using dot-Plotly with ‘-s -t -m 500 -q 500000 -l’ options (<https://github.com/tpoorten/dotPlotly>, accessed on July 17, 2023). Several inversions were observed in the NMR genome, but most plots showed a nearly linear pattern (Fig. 3a–c). In contrast, the dot plot between mouse strains and mouse species showed an almost completely linear pattern (Fig. 3d–g). The presence of gaps at chromosomes 18 and 19 in Ensembl was manually checked to confirm false positive inversions. All inversions examined occurred near gaps regions<sup>45,46</sup>. These results suggest that intraspecific chromosomal variations may occur due to scaffold assembly errors.

In addition, structural variants (SV) between our assembly and the Ensembl assembly were searched. Minimap2 v2.28<sup>44</sup> with ‘asm5–cs -r2k’ option was used to map our assembly to female NMR assembly in Ensembl. Svmm-asm v1.0.3<sup>47</sup>, using the default option, was used to identify SVs greater than 40 bp. As a result of Svmm-asm, 2,263 deletions and 2,668 insertions were identified between our assembly and the Ensembl assembly (Table 4)<sup>48</sup>. To evaluate these statistics, we examined the SVs between the mouse strains (Table 4). The number of SVs (deletions or insertions) exceeded 100,000 in all mouse pairs examined in this study. The number of SVs (deletions or insertions) was lower in NMR than in the mouse strains, indicating that structural variations within NMR may be smaller than those within mouse strains. Intraspecific variations generated by Minigraph-Cactus<sup>49</sup> were also available<sup>50</sup>. This program revealed 1,090,118 variations including SNPs and SVs between our assembly<sup>51</sup> and Ensembl assembly<sup>23</sup>. In addition, the SVs were searched using PacBio raw reads by Sniffles<sup>52</sup>. Detected deletions and insertions were increased compared with assembly-based methods of SV search (Table 5). Obtained VCF files are available in European Variation Archive<sup>53,54</sup>.

**Search for sequence differences of genes between our assembly and the Ensembl assembly.** To compare the gene sequences between our genome assembly and the female NMR assembly in Ensembl, the gene annotations of our assembly were transferred to the female NMR assembly in Ensembl using Liftoff v1.5.2<sup>55</sup>. Transferred gene annotations were filtered using the parameter “valid\_ORF = True” and “match\_ref\_proteins = FALSE” to identify transcripts with intraspecific variations of translated amino acid sequence between our assembly and female NMR assembly in Ensembl. The same processes were applied to combinations of mouse strains (C57BL/6J vs. 129S1/SvImJ\_v1, C57BL/6J vs. A/J\_v1, C57BL/6J vs. AKR/J\_v1, and C57BL/6J vs. BALB/cJ\_v1) to evaluate intraspecific differences in NMR. C57BL/6J in Ensembl release-111<sup>38</sup>, GCA\_001624185.1 (129S1/SvImJ\_v1)<sup>39</sup>, GCA\_001624215.1 (A/J\_v1)<sup>40</sup>, GCA\_001624295.1 (AKR/J\_v1)<sup>41</sup>, and GCA\_001632525.1

| SV                       | This study vs Ensembl | C57BL/6J vs 129S1/SvImJ_v1 | C57BL/6J vs A/J_v1 | C57BL/6J vs AKR/J_v1 | C57BL/6J vs BALB/c_v1 |
|--------------------------|-----------------------|----------------------------|--------------------|----------------------|-----------------------|
| Inversion                | 40                    | 50                         | 64                 | 70                   | 64                    |
| Deletion                 | 2,263                 | 104,041                    | 113,225            | 119,318              | 112,688               |
| Insertion                | 2,668                 | 18,614                     | 17,804             | 18,583               | 16,961                |
| Tandem duplication       | 0                     | 0                          | 0                  | 0                    | 0                     |
| Interspersed duplication | 0                     | 0                          | 0                  | 0                    | 0                     |
| Breakend                 | 0                     | 0                          | 0                  | 0                    | 0                     |

**Table 4.** Structural variants within NMR and within mice.

| SV          | Our assembly vs our raw reads | Ensembl assembly vs our raw reads |
|-------------|-------------------------------|-----------------------------------|
| Inversion   | 47                            | 219                               |
| Deletion    | 2023                          | 5323                              |
| Insertion   | 41,013                        | 42263                             |
| Duplication | 27                            | 95                                |
| Breakend    | 589                           | 1019                              |

**Table 5.** Structural variants within NMR for SV search using raw reads.

(BALB/c\_v1)<sup>42</sup> were used in this analysis. The transcripts with intraspecific variations in NMR were filtered using the Liftoff parameter “Sequence\_ID,” which indicates the sequence identity. The transcript ID of NMR and each mouse strain were mapped to those of the C57BL/6J protein ID using the annotation results from Fanflow<sup>56</sup> and Ensembl Biomart, respectively. Intervene<sup>57</sup> (accessed October 18, 2024) was used to identify the intersections of the gene sets. Enrichment analysis was performed using the Metascape software with default settings (<https://metascape.org/>, accessed on October 25, 2024)<sup>58</sup>.

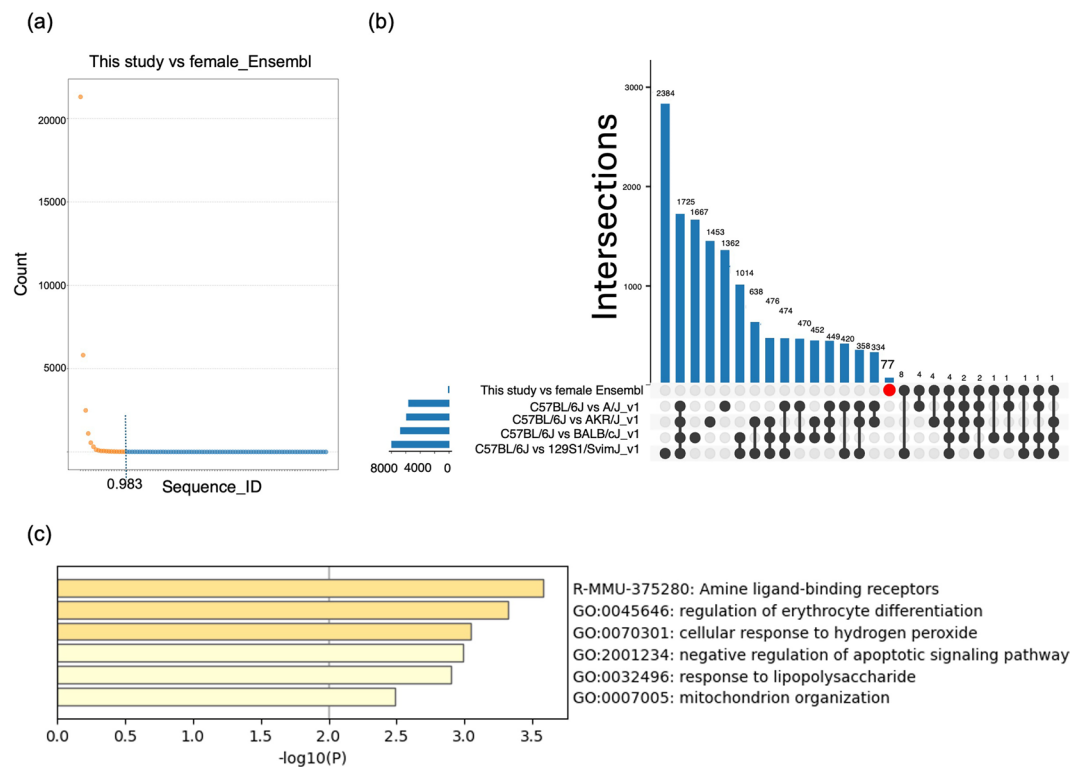
The annotation of 32,263 transcripts from our assembly was successfully transferred to the female NMR assembly in Ensembl, and sorted by indicators of sequence identity, “Sequence\_ID” (Fig. 4a). In the NMR comparison, 177 transcripts had a “Sequence\_ID” below 0.983, and the count per “Sequence\_ID” below this threshold was consistently under 10, suggesting that intraspecific variation in these transcripts was pronounced in the gene sets analyzed in this study. When comparing intraspecific transcripts between NMR and mouse strains, 77 genes exhibited NMR-specific intraspecific variations (Fig. 4b)<sup>59</sup>. Enrichment analysis of the 77 transcripts revealed significant enrichment of genes involved in amine ligand-binding receptors, including 5-hydroxytryptamine (serotonin) receptor 2B (*Htr2b*), G protein-coupled receptor 143 (*Gpr143*, a receptor for L-3,4-dihydroxyphenylalanine [L-DOPA]), erythropoietin (*Epo*), and thyroid hormone receptor beta (*Thrb*) (Fig. 4c)<sup>60</sup>. Intraspecific variations in *Htr2b* and *Gpr143* suggest that responses to serotonin and L-DOPA may differ even within NMR populations, potentially influencing social behavior. Notably, NMR fibroblasts exhibit unique serotonin management, in which serotonin accumulates and induces cell death through serotonin metabolism in response to cellular senescence<sup>61</sup>. Although SVs were also detected in these gene loci, they were located in intron regions<sup>62</sup>. Additionally, intraspecific variations in *Epo* and *Thrb* may contribute to adaptation to hypoxic underground environments through changes in erythropoiesis and metabolic regulation. Examining the functions of these genes could help deepen our understanding of the molecular mechanisms underlying the unique traits of NMRs, such as eusociality, hypoxia tolerance, and delayed aging. Overall, genes exhibiting intraspecific variations may provide valuable insights into the behavioral and physiological adaptations of NMR.

Data Records

All raw sequence data for genome assembly (Illumina short reads and PacBio CLR) have been deposited in the DNA Data Bank of Japan under accession numbers DRR401650<sup>63</sup>, DRR401651<sup>63</sup>, DRR401652<sup>63</sup> and DRR401653<sup>63</sup>. Genome assembly and WGS projects were deposited under the accession numbers GCA\_053883595.1<sup>51</sup> and BAAHMU010000000-BAAHMU010000382, respectively. Full information on functional annotations added using Fanflow is available at *figshare*<sup>56</sup>. Structural variants within NMRs have been deposited in EVA<sup>48,50,53,54</sup>. Gene models estimated by BRAKER3 for our assembly have been deposited in *figshare*<sup>64</sup>.

Technical Validation

The performance of the assemblers was evaluated by comparing four assemblers. The four assemblers (wtdbg2<sup>65</sup>, Flye v2.8<sup>66</sup>, Canu v2.0<sup>20</sup>, and Canu v2.1<sup>20</sup>) were used to compare the assembly quality with respect to contiguity. The assembly, constructed by Canu v2.1, had the longest average and maximum lengths in contigs, and showed the minimum L50 value (Table 5). Pilon was used to polish the genome assembly constructed using Canu v2.1, and Purge\_dups<sup>67</sup> was used to remove haplotigs. This process improved genome statistics, particularly regarding contiguity (Table 6).



**Fig. 4** Mutations for amino acid sequences between our assembly and the female Ensembl assembly. **(a)** Total numbers of transcripts per Sequence\_ID. The dot color changes based on a Sequence\_ID threshold of 0.983. **(b)** Intersections of transcripts with intraspecific variations in NMR and mouse strains. Red circle indicates transcript sets with intraspecific differences only within NMR. **(c)** Enriched terms in NMR-intraspecific 77 transcripts indicated by red circle in **(b)**.

| Assembler                  | wtdbg2 | Flye v2.8 | Canu v2.0 | Canu v2.1 |
|----------------------------|--------|-----------|-----------|-----------|
| Genome size (Gb)           | 2.519  | 2.555     | 2.741     | 2.683     |
| No. of contigs or scaffold | 4196   | 1450      | 1855      | 1459      |
| Average length (Mb)        | 0.6    | 1.76      | 1.46      | 1.84      |
| Max length (Mb)            | 73.97  | 71.69     | 93.24     | 99.77     |
| N50 (Mb)                   | 18.64  | 24.74     | 36.24     | 34.94     |
| L50                        | 43     | 34        | 27        | 26        |
| Gap                        | 0%     | 0.00%     | 0%        | 0%        |

**Table 6.** Comparison of genome assemblers.

Data availability

All raw sequencing data generated in this study, including Illumina short reads and PacBio CLR long reads, have been deposited in the DNA Data Bank of Japan (DDBJ) Sequence Read Archive under accession numbers DRR401650–DRR401653 (<https://ddbj.nig.ac.jp/search/entry/sra-study/DRP012667>)<sup>63</sup>. The genome assembly has been deposited under accession GCA\_053883595.1, and the whole-genome shotgun (WGS) project under BAAHMU010000000–BAAHMU010000382. Functional annotation results produced using Fanflow (TSV format) and gene models predicted by BRAKER3 (GTF format and FASTA format) are available on figshare (<https://doi.org/10.6084/m9.figshare.28171166> and <https://doi.org/10.6084/m9.figshare.28180430>). The Fanflow annotation file includes the following columns: braker\_id (BRAKER3 transcript ID), human\_pid / human\_gene / human\_description (human best-hit protein, gene symbol, and description), mouse\_pid / mouse\_gene / mouse\_description (mouse best-hit protein, gene symbol, and description), guinea\_pig\_pid / guinea\_pig\_gene / guinea\_pig\_description (guinea pig best-hit protein, gene symbol, and description), Hgla\_female\_pid / Hgla\_female\_gene / description (Naked mole-rat female best-hit protein, gene symbol, and description), Hgla\_male\_pid / Hgla\_male\_gene / description (Naked mole-rat male best-hit protein, gene symbol, and description), uniprot\_id / uniprot\_description (UniProtKB best-hit protein and description), pfam\_id / pfam\_description (Pfam domain IDs and domain annotations).

Structural variants identified in this study (VCF format) have been deposited in the European Variation Archive (EVA) under project accession numbers ERZ28787178 (<https://identifiers.org/ena.embl:ERZ28787178>)<sup>53</sup>,

ERZ28787179 (<https://identifiers.org/ena.embl:ERZ28787179>)<sup>54</sup>, ERZ28787181 (<https://identifiers.org/ena.embl:ERZ28787181>)<sup>48</sup>, and ERZ28787182 (<https://identifiers.org/ena.embl:ERZ28787182>)<sup>50</sup>. All datasets are publicly accessible at the repositories listed above.

## Code availability

All programs and pipelines were executed following their official manuals or help pages. Version and parameter information used in our analysis are provided in the Methods section. No custom scripts were used.

Received: 20 May 2025; Accepted: 27 February 2026;

Published online: 18 March 2026

## References

1. Oka, K., Yamakawa, M., Kawamura, Y., Kutsukake, N. & Miura, K. The Naked Mole-Rat as a Model for Healthy Aging. *Annu. Rev. Anim. Biosci.* **11**, 207–226 (2023).
2. Buffenstein, R. The Naked Mole-Rat: A New Long-Living Model for Human Aging Research. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **60**, 1369–1377 (2005).
3. Jarvis, J. U. M. Eusociality in a Mammal: Cooperative Breeding in Naked Mole-Rat Colonies. *Science* **212**, 571–573 (1981).
4. Park, T. J. *et al.* Fructose-driven glycolysis supports anoxia resistance in the naked mole-rat. *Science* **356**, 307–311 (2017).
5. Fang, X. *et al.* Adaptations to a Subterranean Environment and Longevity Revealed by the Analysis of Mole Rat Genomes. *Cell Reports* **8**, 1354–1364 (2014).
6. Zhou, X. *et al.* Beaver and Naked Mole Rat Genomes Reveal Common Paths to Longevity. *Cell Reports* **32**, 107949 (2020).
7. Oka, K. *et al.* Resistance to chemical carcinogenesis induction via a dampened inflammatory response in naked mole-rats. *Commun Biol* **5**, 287 (2022).
8. Tian, X. *et al.* High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* **499**, 346–349 (2013).
9. Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223–227 (2011).
10. Sokolowski, D. J. *et al.* An updated reference genome sequence and annotation reveals gene losses and gains underlying naked mole-rat biology. Preprint at, <https://doi.org/10.1101/2024.11.26.625329> (2024).
11. Lewin, H. A. *et al.* The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. USA* **119**, e2115635118 (2022).
12. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu Rev Ecol Evol Syst* **39**, 21–42 (2008).
13. Plessy, C. *et al.* Extreme genome scrambling in marine planktonic *Oikopleura dioica* cryptic species. *Genome Res.*, <https://doi.org/10.1101/gr.278295.123> (2024).
14. Dobigny, G., Britton-Davidian, J. & Robinson, T. J. Chromosomal polymorphism in mammals: an evolutionary perspective: Chromosomal polymorphism in mammals. *Biol Rev* **92**, 1–21 (2017).
15. Faulkes, C. G. *et al.* Micro- and macrogeographical genetic structure of colonies of naked mole-rats *Heterocephalus glaber*. *Molecular Ecology* **6**, 615–628 (1997).
16. Ingram, C. M., Troendle, N. J., Gill, C. A., Braude, S. & Honeycutt, R. L. Challenging the inbreeding hypothesis in a eusocial mammal: population genetics of the naked mole-rat, *Heterocephalus glaber*. *Molecular Ecology* **24**, 4848–4865 (2015).
17. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
18. Bono, H., Sakamoto, T., Kasukawa, T. & Tabunoki, H. Systematic Functional Annotation Workflow for Insects. *Insects* **13**, 586 (2022).
19. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).
20. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
21. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **9**, e112963 (2014).
22. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
23. European Nucleotide Archive. [https://identifiers.org/insdc.gca:GCA\\_944319715.1](https://identifiers.org/insdc.gca:GCA_944319715.1) (2022).
24. European Nucleotide Archive [https://identifiers.org/insdc.gca:GCA\\_000230445.1](https://identifiers.org/insdc.gca:GCA_000230445.1) (2011).
25. Manni, M., Berkeley, M. R., Seppely, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols* **1**, e323 (2021).
26. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
27. Toga, K. Repetitive sequences statistic of NMR, mice and human. *figshare* <https://doi.org/10.6084/m9.figshare.28170533.v1> (2025).
28. Toga, K. List of public RNA-Seq data of naked mole rat. *figshare* <https://doi.org/10.6084/m9.figshare.28171121.v1> (2025).
29. Krueger, F. *et al.* FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. *Zenodo* <https://doi.org/10.5281/ZENODO.5127898> (2023).
30. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
33. Li, H. Protein-to-genome alignment with minimap2. *Bioinformatics* **39**, btad014 (2023).
34. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
35. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**, 1160–1166 (2019).
36. Toga, K. Differences of gene model between our assembly and Ensembl assembly. *figshare* <https://doi.org/10.6084/m9.figshare.28171271.v1> (2025).
37. European Nucleotide Archive. [https://identifiers.org/insdc.gca:GCA\\_944319725.1](https://identifiers.org/insdc.gca:GCA_944319725.1) (2022).
38. European Nucleotide Archive. [https://identifiers.org/insdc.gca:GCA\\_000001635.9](https://identifiers.org/insdc.gca:GCA_000001635.9).
39. NCBI GenBank. [https://identifiers.org/ncbi/insdc.gca:GCA\\_001624185.1](https://identifiers.org/ncbi/insdc.gca:GCA_001624185.1).
40. NCBI GenBank. [https://identifiers.org/ncbi/insdc.gca:GCA\\_001624215.1](https://identifiers.org/ncbi/insdc.gca:GCA_001624215.1).
41. NCBI GenBank. [https://identifiers.org/ncbi/insdc.gca:GCA\\_001624295.1](https://identifiers.org/ncbi/insdc.gca:GCA_001624295.1).
42. NCBI GenBank. [https://identifiers.org/ncbi/insdc.gca:GCA\\_001632525.1](https://identifiers.org/ncbi/insdc.gca:GCA_001632525.1).
43. NCBI GenBank. [https://identifiers.org/ncbi/insdc.gca:GCA\\_921997135.2](https://identifiers.org/ncbi/insdc.gca:GCA_921997135.2).

44. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
45. Toga, K. Gap regions at chromosome 19. *figshare* <https://doi.org/10.6084/m9.figshare.29670398.v1> (2025).
46. Toga, K. Gap regions at chromosome 18. *figshare* <https://doi.org/10.6084/m9.figshare.29670218.v1> (2025).
47. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
48. European Variation Archive <https://identifiers.org/ena.embl:ERZ28787181> (2026).
49. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* **42**, 663–673 (2024).
50. European Variation Archive <https://identifiers.org/ena.embl:ERZ28787182> (2026).
51. DDBJ [https://identifiers.org/ncbi/insdc.gca:GCA\\_053883595.1](https://identifiers.org/ncbi/insdc.gca:GCA_053883595.1) (2025).
52. Smolka, M. *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* **42**, 1571–1580 (2024).
53. European Variation Archive <https://identifiers.org/ena.embl:ERZ28787178> (2026).
54. European Variation Archive <https://identifiers.org/ena.embl:ERZ28787179> (2026).
55. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
56. Toga, K. Full information on functional annotation added using Fanflow. *figshare* <https://doi.org/10.6084/m9.figshare.28171166.v1> (2025).
57. Khan, A. & Mathelier, A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics* **18**, 287 (2017).
58. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
59. Toga, K. Transcripts with intraspecific variations in NMR. *figshare* <https://doi.org/10.6084/m9.figshare.28218143.v1> (2025).
60. Toga, K. Results of enrichment analysis for 77 transcripts with intraspecific variants in NMR. *figshare* <https://doi.org/10.6084/m9.figshare.28180361.v1> (2025).
61. Kawamura, Y. *et al.* Cellular senescence induction leads to progressive cell death via the INK4a-RB pathway in naked mole-rats. *The EMBO Journal* **42**, e111133 (2023).
62. Toga, K. SV detection at Gpr143 and Htr2b gene loci. *figshare* <https://doi.org/10.6084/m9.figshare.29835212.v2> (2025).
63. DDBJ <https://identifiers.org/ncbi/insdc.sra:DRP012667> (2025).
64. Toga, K. Gene models generated from BRAKER3. *figshare* <https://doi.org/10.6084/m9.figshare.28180430.v1> (2025).
65. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155–158 (2020).
66. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
67. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP16H06279 (PAGS), supported in part by the JSPS KAKENHI Grants (JP18H02365) to K.M., JSPS KAKENHI Grants (JP24H00542) and JST COI-NEXT (JPMJPF2010) to K.M. and H.B. and JST FOREST Program (JPMJFR216C) to K.M. We would like to thank all laboratory members at Hiroshima University and Kumamoto University for their valuable comments. Computations were performed on the computers at the Hiroshima University Genome Editing Innovation Center. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

## Author contributions

K.M. and H.B. coordinated and designed this study. K.O. conducted the sampling of *Heterocephalus glaber* and H.T., T.I., and A.T. performed sequencing and de novo assembly. K.T. and H.B. performed bioinformatics analyses and generated figures and tables. K.T. wrote the first draft of the manuscript. All authors revised, edited, and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to H.B. or K.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026