

ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	387912
ToLID	mCroZim1
Species	<i>Crocidura zimmermanni</i>
Class	Mammalia
Order	Eulipotyphla

Genome Traits	Expected	Observed
Haploid size (bp)	2,224,927,356	2,571,728,606
Haploid Number	17 (source: direct)	20
Ploidy	2 (source: ancestor)	2
Sample Sex	XX	XX

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q62

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

- . Observed Haploid Number is different from Expected

Curator notes

- . Interventions/Gb: 2
- . Contamination notes: ""
- . Other observations: "The assembly of *Crocidura zimmermanni* (mCroZim1.1) is based on 37X PacBio data and 189X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity Genomics Europe project (BGE, <https://biodiversitygenomics.eu/>). The assembly process included the following steps: initial PacBio assembly generation with Hifiiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 8 contigs were identified as contaminants (bacterial), totaling 690.5 Kb (with the largest being 117.9 Kb) ; TAPAs has classified two scaffolds (scaffold_102 & scaffold_138) as bacteria, but they are host rDNA, therefore they are not removed. Additionally, 195 regions totaling 76.2 Mb (with the largest being 20.8 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, no regions were tagged as allelic duplications nor contaminants ; 24 sequences were tagged Unloc, their orientation is uncertain ; the coverage of mCroZim1_5 seems lower than the others and its alignment with *Suncus etruscus* (CM043992.1 chromosome_X) suggests that mCroZim1_5 is the X chromosome. the

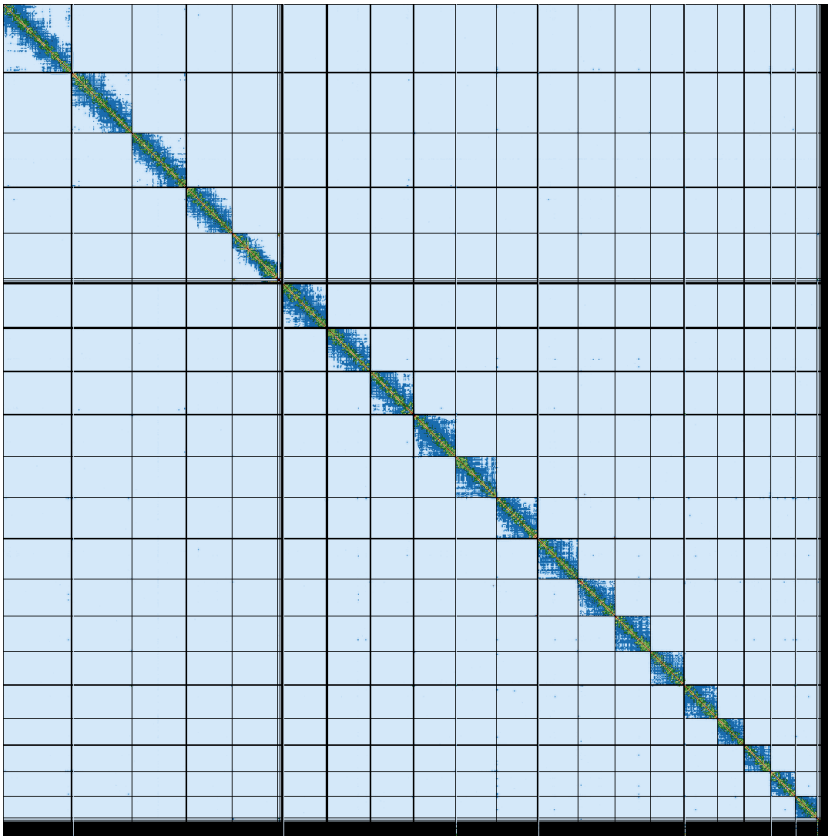
difference in expected vs observed haploid number is noteworthy, but there is a great variability of karyotypes within this genus and also between individuals of the same species ; also, *Crocidura zimmermanni* is endemic (Crete) and shows a threatened status (Endangered), which could also contribute. Despite counting with karyotype information, that information is rather limited for this species ; there is no evidence of false-joins with the contact map. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	2,571,745,479	2,571,728,606
GC %	39.85	39.85
Gaps/Gbp	62.6	62.6
Total gap bp	16,100	16,400
Scaffolds	151	146
Scaffold N50	127,229,103	127,229,103
Scaffold L50	9	9
Scaffold L90	18	18
Contigs	312	307
Contig N50	28,351,139	28,351,139
Contig L50	25	25
Contig L90	97	97
QV	62.0833	62.0867
Kmer compl.	92.9631	92.9629
BUSCO sing.	90.6%	90.6%
BUSCO dupl.	2.2%	2.2%
BUSCO frag.	0.9%	0.9%
BUSCO miss.	6.3%	6.3%

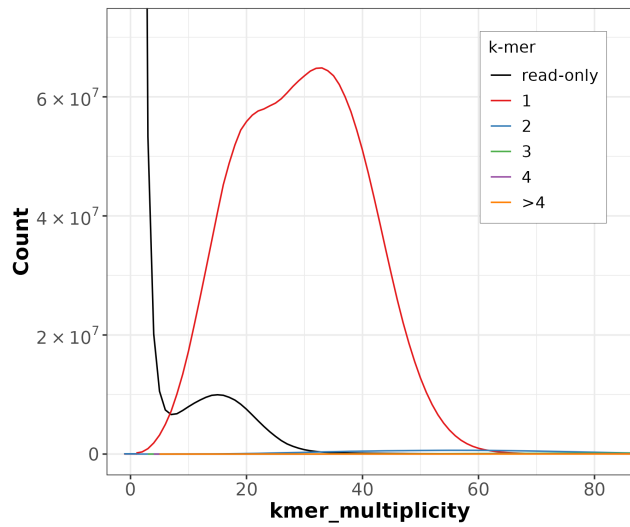
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: laurasiatheria_odb10 (genomes:52, BUSCOs:12234)

HiC contact map of curated assembly

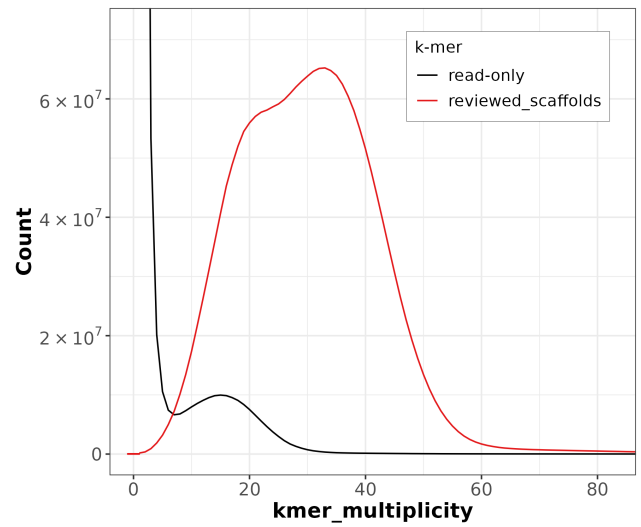


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

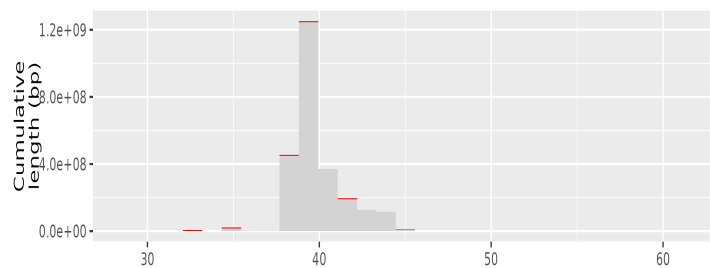


Distribution of k-mer counts per copy numbers found in asm

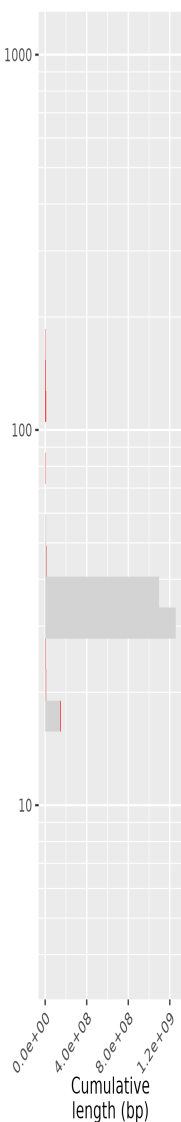
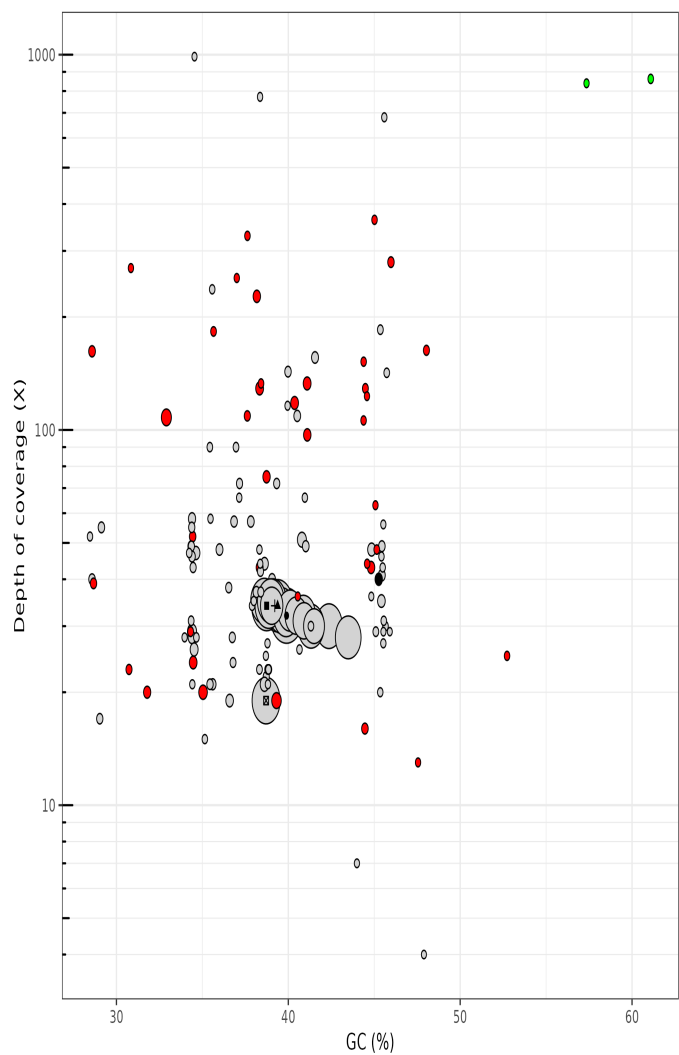


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



superkingdom

- Bacteria
- Eukaryota
- N/A
- Viruses

Longest sequences (bp)

- mCroZim1_1 - 210976412 (N/A)
- ▲ mCroZim1_2 - 185699113 (N/A)
- mCroZim1_3 - 166773580 (N/A)
- + mCroZim1_4 - 141266018 (N/A)
- ▣ mCroZim1_5 - 140510615 (N/A)

Length (bp)

- 5.0e+07
- 1.0e+08
- 1.5e+08
- 2.0e+08

collapsed. Bubble plot circles are scaled by sequence length, positioned by coverage and GC proportion, and coloured by taxonomy. Histograms show total assembly length distribution on each axis.

Data profile

Data	PACBIO Hifi	Arima
Coverage	37	189

Assembly pipeline

- **Hifiasm**
 - |_ *ver*: 0.19.5-r593
 - |_ *key param*: NA
- **purge_dups**
 - |_ *ver*: 1.2.5
 - |_ *key param*: NA
- **YaHS**
 - |_ *ver*: 1.2
 - |_ *key param*: NA

Curation pipeline

- **PretextMap**
 - |_ *ver*: 0.1.9
 - |_ *key param*: NA
- **PretextView**
 - |_ *ver*: 0.2.5
 - |_ *key param*: NA

Submitter: Lola Demirdjian

Affiliation: Genoscope

Date and time: 2025-03-11 17:37:22 CET