ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3349997
ToLID	**icScaAbbr5**
Species	Scarites abbreviatus
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	768,563,182	910,842,350
Haploid Number	23 (source: ancestor)	27
Ploidy	2 (source: ancestor)	2
Sample Sex	unknown	unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for hap1: 7.7.Q72
Obtained EBP quality metric for hap2: 7.7.Q72

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

. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for hap1
. Kmer completeness value is less than 90 for hap2
. Assembly length loss > 3% for hap1
. Not 90% of assembly in chromosomes for hap1
. Not 90% of assembly in chromosomes for hap2

Curator notes

. Interventions/Gb: None
. Contamination notes: "FCX-GX and blobtools were used to detect contaminations but could not identify anything. Mitochondrial genome was removed from the assemblies."
. Other observations: "PacBio Hifi reads were sub-sampled to 60x read coverage. Hifiasm was run in HiC-mode and created two haplotype assemblies (hap1 - contigs: 702, yield 762Mb, N50: 17.0Mb; hap2 - contigs: 726, yield 1071Mb, N50: 10.1Mb). The phasing is quite in-balanced in terms of yield. But a prior haplotype collapsed assembly showed similar stats as the hap2 and were difficult to manually curate. (1) MitoHifi - detect mitochondrial contigs (2) Tiara - deep learning to classify eukaryotic, bacterial, organelle, archea sequences and (3) BlobtoolKit - detect contaminations. Purge_dups was used to remove haplotypic duplications and YaHS was used to scaffold the contigs. Manual curations was done in dual-curation mode (3 rounds) to fix haplotype misplacements, and misjoins and followed by one curation

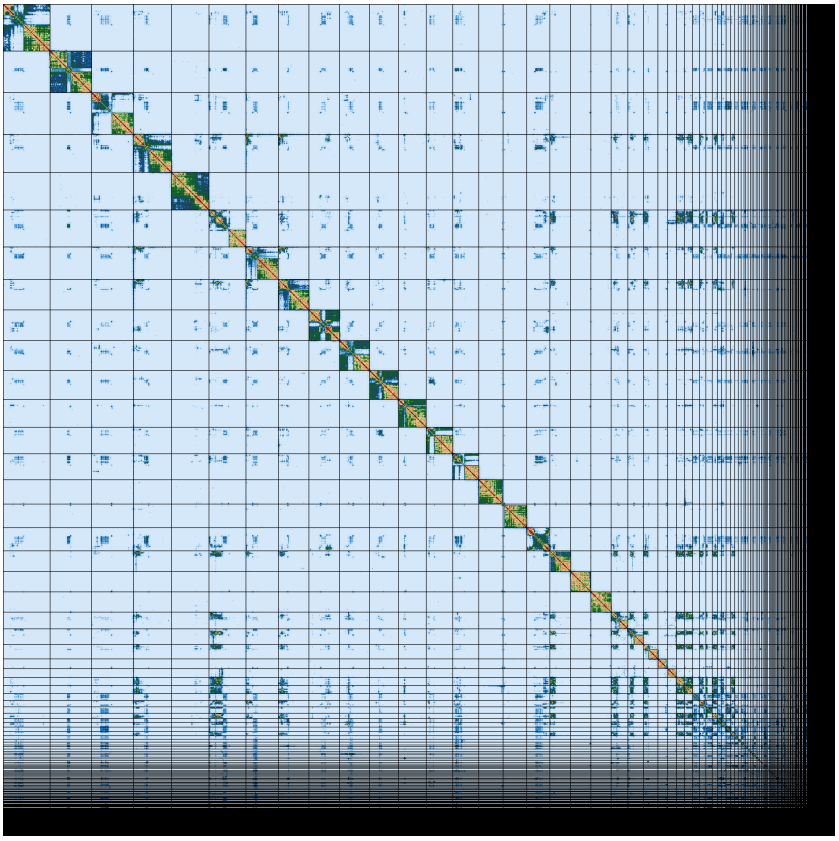
round on each haplotype. I ended up with 26 autosomes + X in hap1 and 26 chromosomes in hap2. Though I think that is probably not correct. Many chromosomes have a very repetitive block and for some of the smaller chrosomes I could not assign a repetitive block. Therfore I might have counted chromosome arms as full chromosomes. There are many repetitive contigs (also larger ones in the shrapnels) that show low PacBio read coverages. On top there are many structural variations within contigs. Long story ... this is the best that I came up with. Happy for any suggestions! I could not find any karyotypes images of this species nor any chromosome number. Within the genus the chromosome number ranges from 17-30 and the sex system from XXY, XY and XO. I could definitely find a single X chromosome that shows nice haplotype coverage but I could not find a Y chromosome. I did a synteny analysis with 26 "closely" related chromosome scale assemblies but this gave no clue about chromosome number or structure either. The mitochondrial sequence was assembled with MitoHifi from the raw reads (len=17001bp)."

Quality metrics table

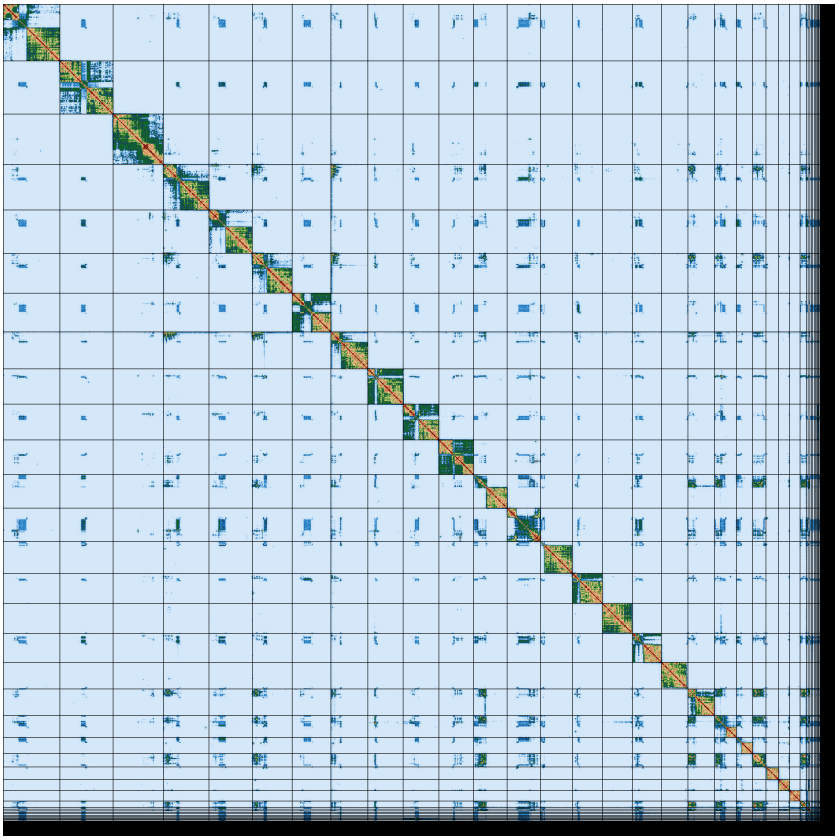
Metrics	Pre-curation hap1	Pre-curation hap2	Curated hap1	Curated hap2
Total bp	1,071,779,099	762,754,003	910,842,350	910,842,350
GC %	31.73	31.28	31.68	31.68
Gaps/Gbp	27.99	32.78	73.56	73.56
Total gap bp	3,000	2,500	10,900	10,900
Scaffolds	694	674	209	209
Scaffold N50	13,600,044	20,412,667	30,387,215	30,387,215
Scaffold L50	23	15	12	12
Scaffold L90	154	52	42	42
Contigs	724	699	276	276
Contig N50	10,145,749	17,038,473	13,411,682	13,411,682
Contig L50	29	18	22	22
Contig L90	179	75	98	98
QV	69.3952	66.9877	72.0794	72.289
Kmer compl.	88.2277	83.2452	87.7559	83.1447
BUSCO sing.	96.7%	90.6%	96.6%	90.7%
BUSCO dupl.	0.5%	0.5%	0.5%	0.5%
BUSCO frag.	1.0%	1.0%	1.0%	1.0%
BUSCO miss.	1.8%	8.0%	1.9%	7.9%

BUSCO: 5.8.3 (euk_genome_min, miniprot) / Lineage: endopterygota_odb12 (genomes:76, BUSCOs:3754)

HiC contact map of curated assembly

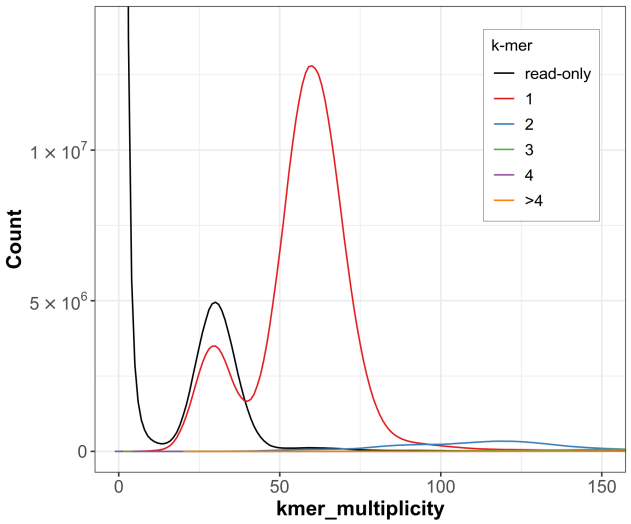


hap1 [LINK]

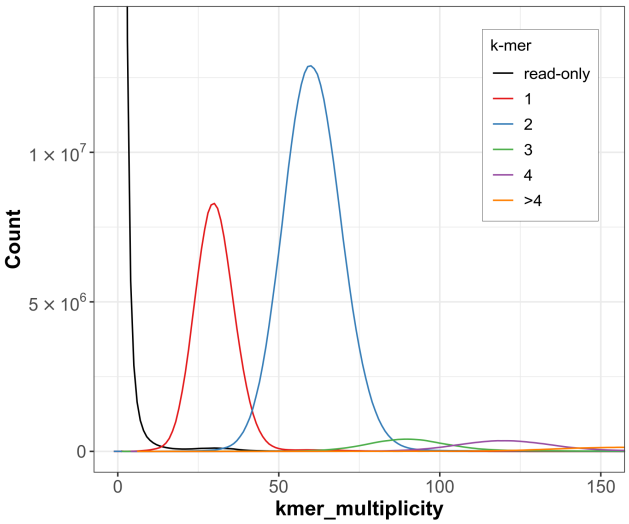


hap2 [LINK]

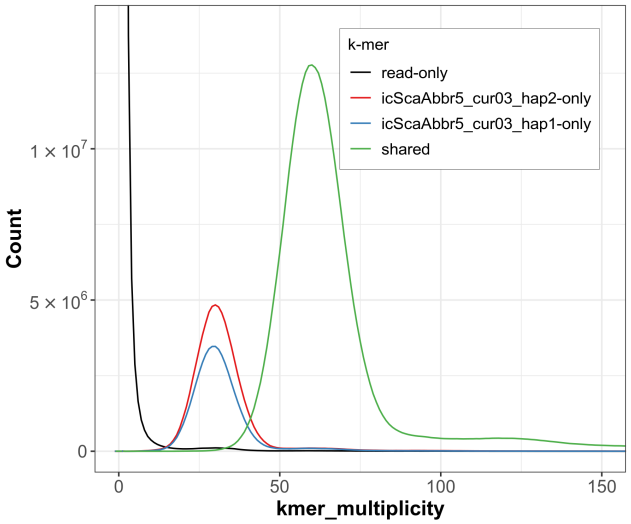
K-mer spectra of curated assembly



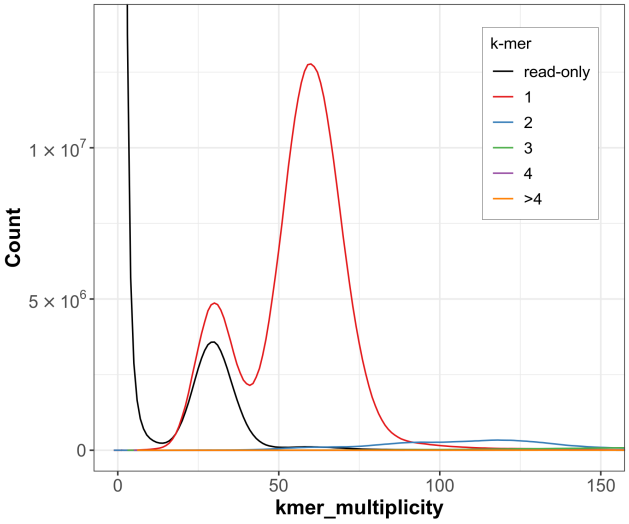
Distribution of k-mer counts per copy numbers found in **icScaAbbr5_cur03_hap1** (hapl.)



Distribution of k-mer counts per copy numbers found in asm (dipl.)

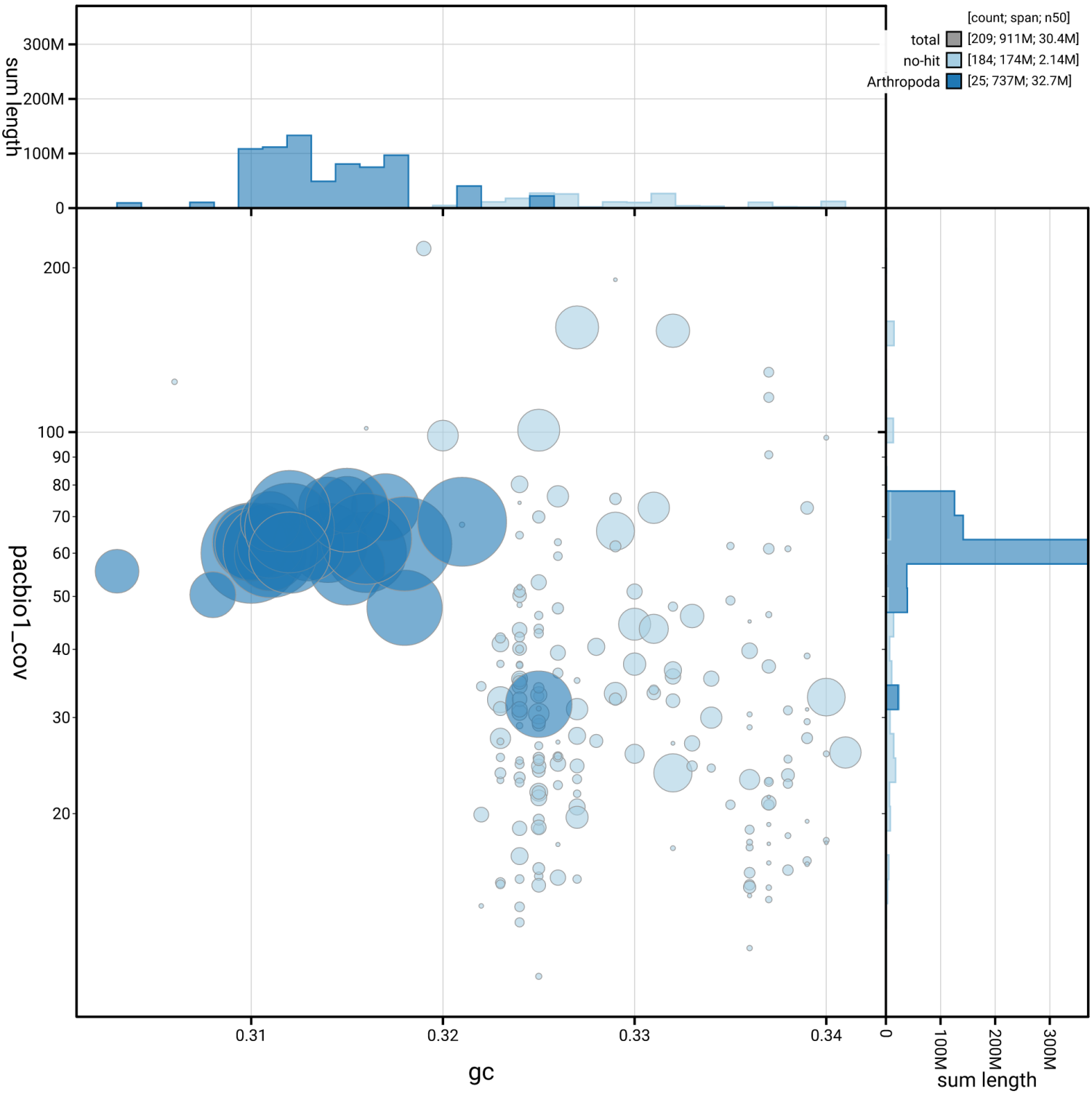


Distribution of k-mer counts coloured by their presence in reads/assemblies

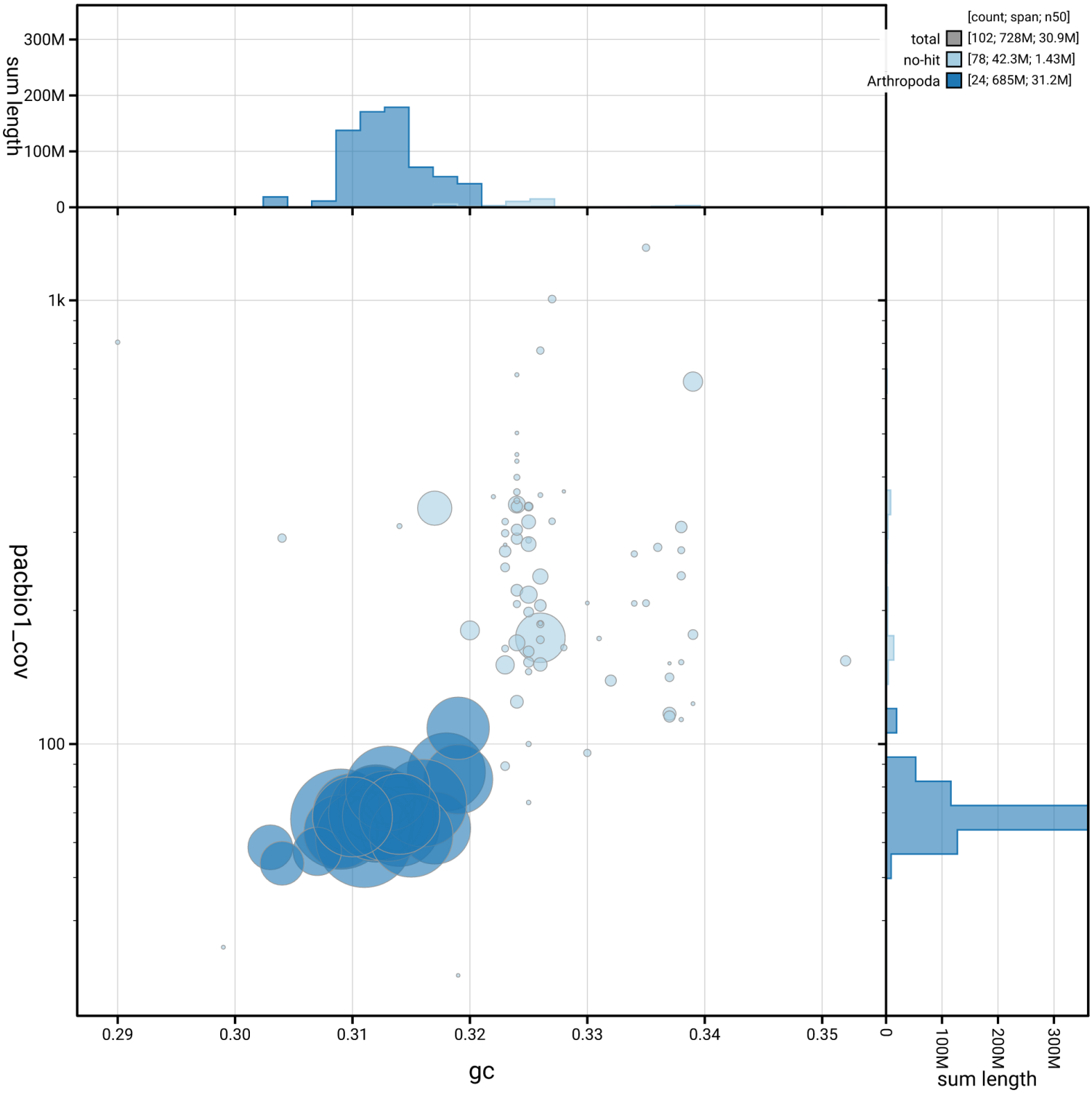


Distribution of k-mer counts per copy numbers found in **icScaAbbr5_cur03_hap2** (hapl.)

Post-curation contamination screening



hap1. 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.



hap2. 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	HiFi	Arima3
Coverage	114x	71x

Assembly pipeline

- **Hifiasm**
    |_ *ver:* 0.25.0-r726
    |_ *key param:* HiC
    |_ *key param:* l3
- **purge_dups**
    |_ *ver:* 1.2.6
    |_ *key param:* NA
- **YaHS**
    |_ *ver:* 1.2.2
    |_ *key param:* NA

Curation pipeline

- **GRIT_Rapid**
    |_ *ver:* 2.0
    |_ *key param:* NA

Submitter: Martin Pippel
Affiliation: SciLifeLab

Date and time: 2026-01-17 15:54:11 CET