ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	499837
ToLID	**iqSphRube1**
Species	Sphingonotus rubescens
Class	Insecta
Order	Orthoptera

Genome Traits	Expected	Observed
Haploid size (bp)	8,461,483,717	9,028,342,188
Haploid Number	11 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	XX	XX

EBP metrics summary and curation notes

Obtained EBP quality metric for hap1: 7.8.Q73
Obtained EBP quality metric for hap2: 7.8.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

Curator notes

. Interventions/Gb: 27
. Contamination notes: "FCS-GX and Blobtools detected no presence of contaminants. .Mitochondrial genome was removed from the assembly"
. Other observations: "The PacBio reads were subsampled to 40X. HiFiasm created a primary collapsed assembly of 13Gb, and HiC-phased haplotypes of 9Gb each. Therefore the haplotype-phased assemblies were used for this species. For each haplotype purge_dups, FCS-GX and yahs was run seperately. When running purge_dups with default parameter 545Mb (hap1) and 746Mb (hap2) of sequence was removed. This also resulted in the loss of some busco genes. Therefore first: contigs larger than 1Mb and second: contigs that were marked as hap but contained single Busco genes were kept. The purge_dups bed annotations was later used in the manual curation step too, in order to decide if e.g. a REPEAT tagged contigs needs to be removed. The manual curation revealed many haplotype misplacements and HapHic was used to guide to guide those corrections. Just for clarification: HapHiC was not error-free either - the correction of contig errors was sub-optimal. Therefore I decided to stick to yahs scaffolds, where I already invested some time to curate those. There are still some

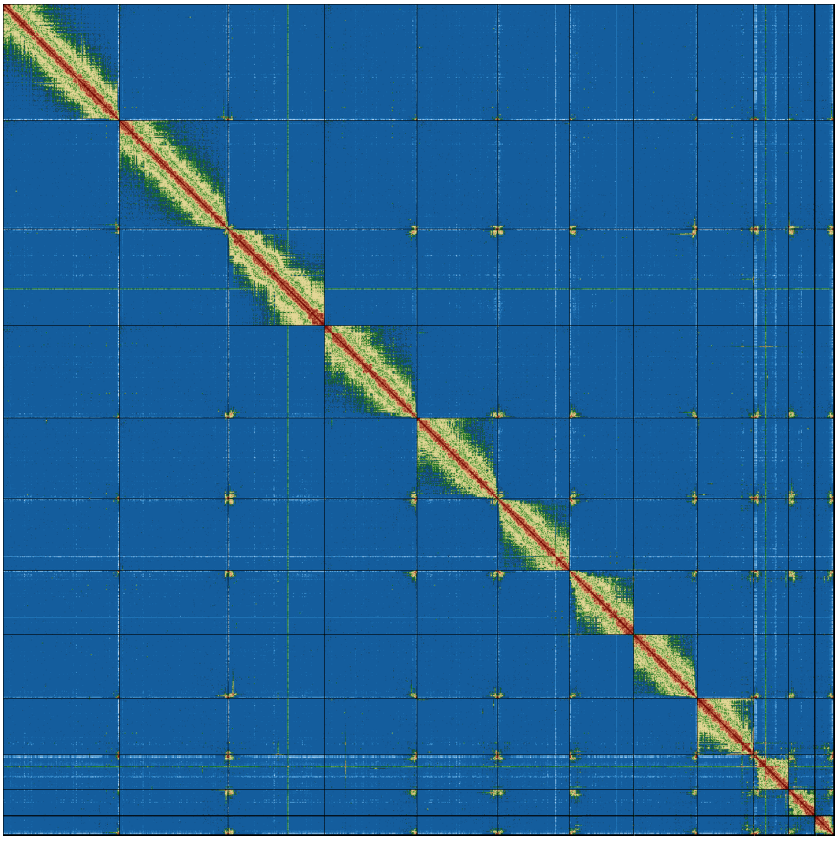
weird repetitive HiC patterns left, which look odd to me. E.g.: a large inversion in h2s7 at 638M - 668M, but when trying to remove the potential dup the HiC map looks much worse as the main diagonal has a pretty decent signal including the dup. "

Quality metrics table

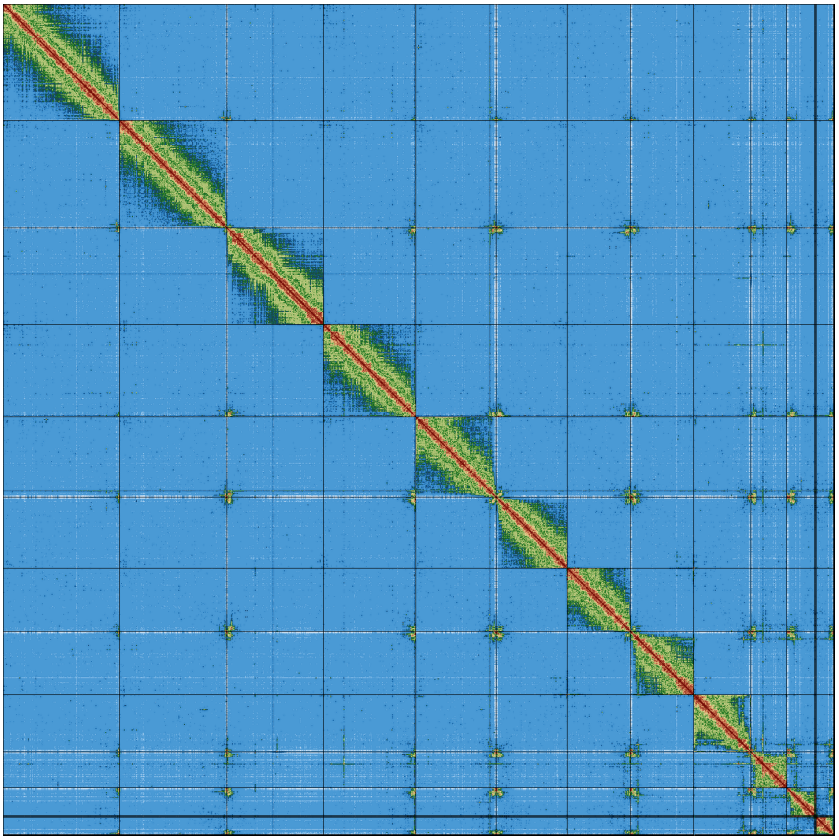
Metrics	Pre-curation hap1	Pre-curation hap2	Curated hap1	Curated hap2
Total bp	8,987,749,137	9,126,898,536	9,028,342,188	8,991,737,251
GC %	40.89	40.93	40.9	40.91
Gaps/Gbp	50.18	47.55	56.27	56.72
Total gap bp	45,100	43,400	58,700	60,100
Scaffolds	200	183	135	100
Scaffold N50	856,901,657	694,436,201	874,086,278	868,557,452
Scaffold L50	5	5	5	5
Scaffold L90	10	13	9	10
Contigs	651	617	643	610
Contig N50	37,154,065	35,069,156	36,940,391	34,890,047
Contig L50	75	81	75	81
Contig L90	266	282	269	280
QV	73.1435	72.7469	73.0762	72.827
Kmer compl.	62.0923	62.0862	62.0322	61.9684
BUSCO sing.	96.1%	96.0%	96.0%	96.2%
BUSCO dupl.	2.4%	2.5%	2.5%	2.3%
BUSCO frag.	0.9%	0.9%	0.9%	1.0%
BUSCO miss.	0.6%	0.6%	0.6%	0.6%

BUSCO: 5.8.3 (euk_genome_min, miniprot) / Lineage: insecta_odb12 (genomes:79, BUSCOs:3114)

HiC contact map of curated assembly

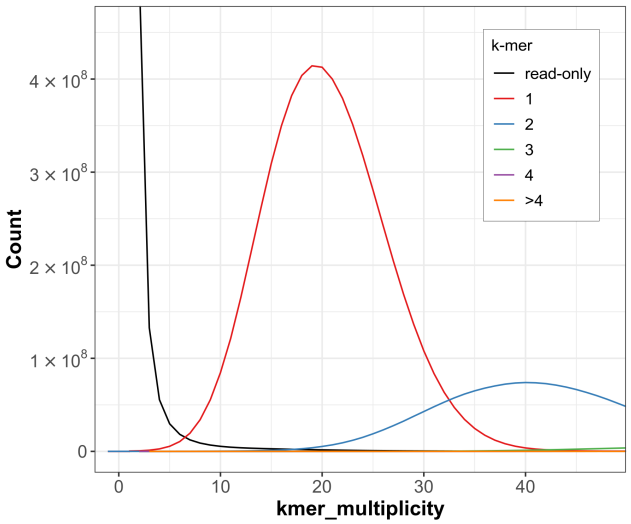


hap1 [LINK]

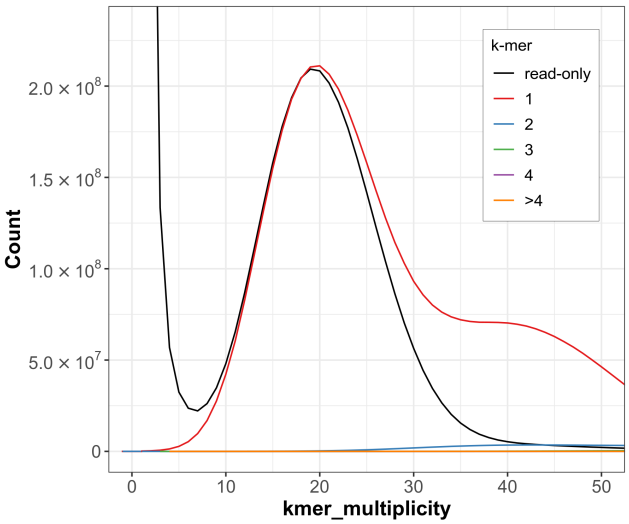


hap2 [LINK]

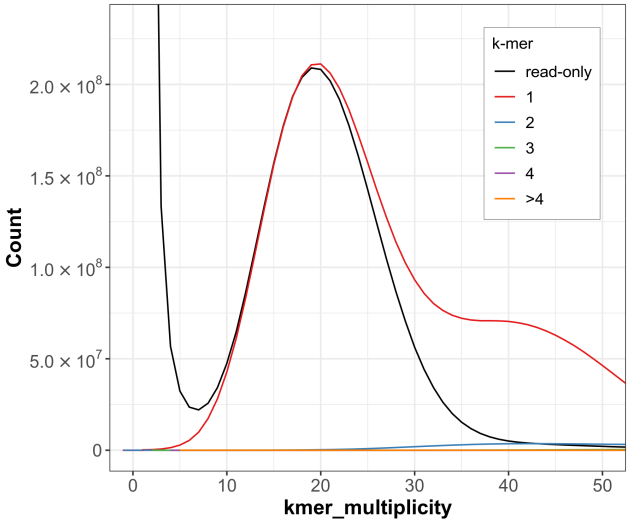
K-mer spectra of curated assembly



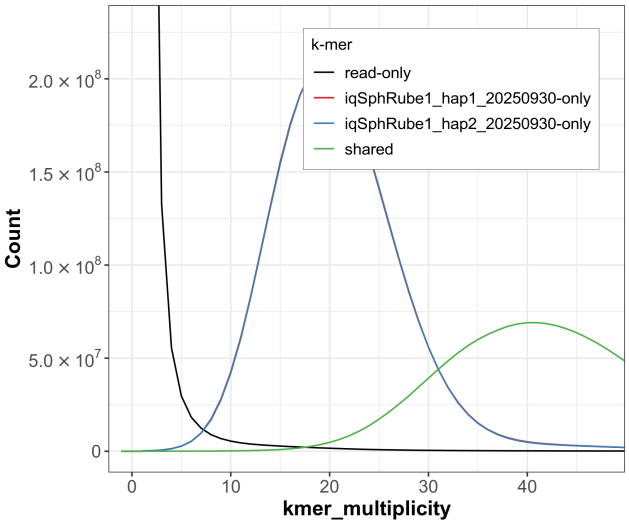
Distribution of k-mer counts per copy numbers found in asm (dipl.)



Distribution of k-mer counts per copy numbers found in **iqSphRube1_hap2_20250930** (hapl.)

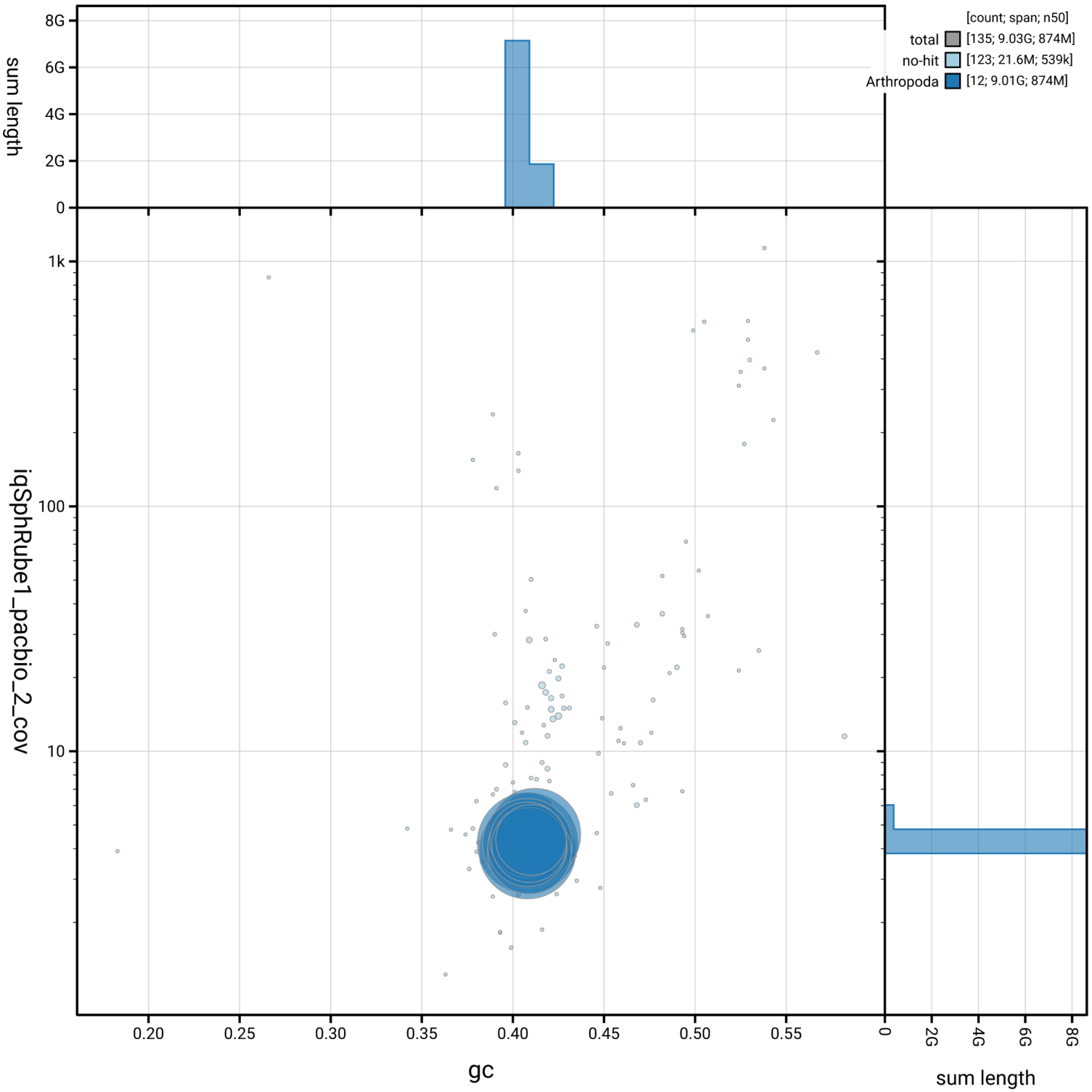


Distribution of k-mer counts per copy numbers found in **iqSphRube1_hap1_20250930** (hapl.)

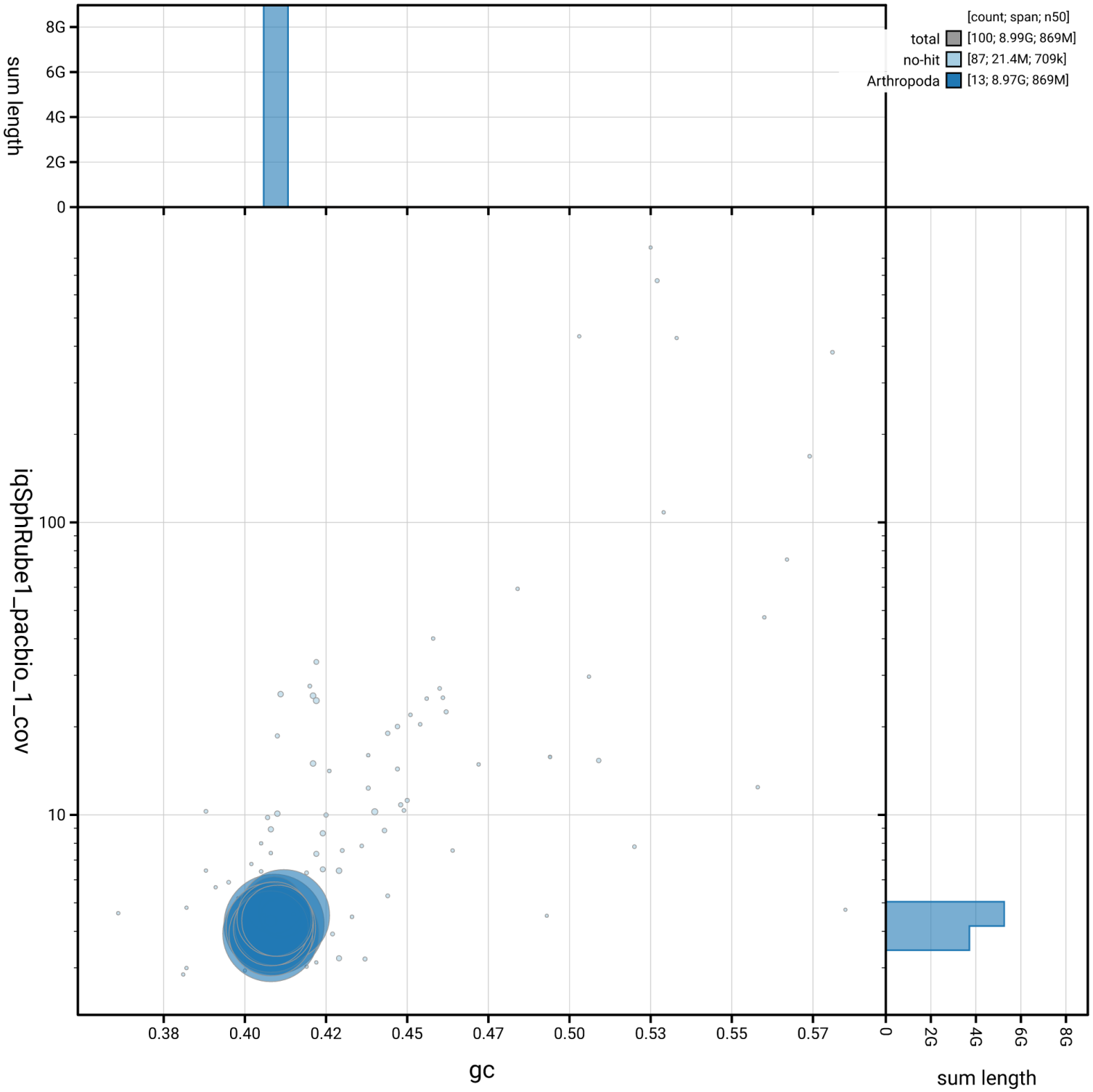


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Bionano	OmniC
Coverage	56x	NA	89x

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
- **HapHic**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Curation pipeline

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

Submitter: Martin Pippel
Affiliation: SciLifeLab

Date and time: 2025-10-01 15:54:58 CEST