ERGA Assembly Report

v24.02.09_beta

Tags: ERGA-BGE

ToLID	**ilGraIsab1.1**
Species	***Graellsia isabellae***
Class	Insecta
Order	Lepidoptera

Genome Traits	Expected	Observed
Haploid size (bp)	553,262,926	560,905,695
Haploid Number	30 (source: ['ancestor'])	31
Ploidy	1 (source: ['ancestor'])	2
Sample Sex	Z0	Z0

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q49

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

. Observed Haploid Number is different from Expected
. Observed Ploidy is different from Expected

Curator notes

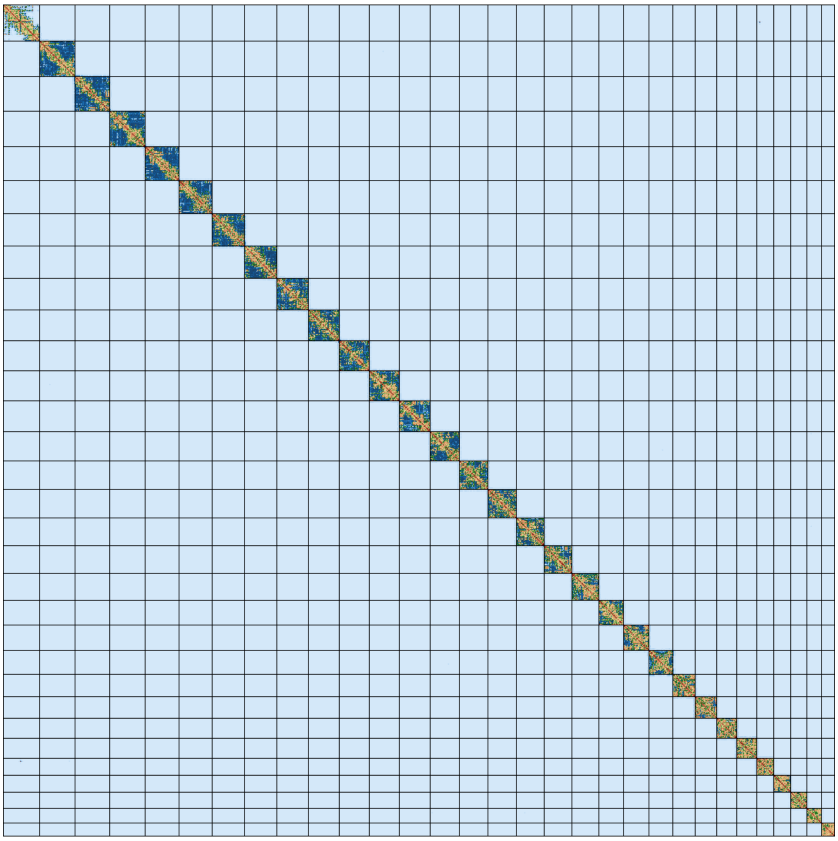
. Interventions/Gb: 0
. Contamination notes: "No contaminants detected. All sequences were assigned to the order Lepidoptera using blobtoolkit. The complete mitochondrial genome was assembled into a single circular contig of 15,247 base pairs with excellent base accuracy using the FOAM pipeline (https://github.com/cnag-aat/FOAM)."
. Other observations: "Genome Assembly obtained with CLAWS v2.1 pipeline (https://workflowhub.eu/workflows/567). Input assembly (nextdenovo.hypo2.purged) for HiC scaffolding was already highly contiguous (N50=18.8Mb), YaHS made 1 break and 6 joins. Manual curation was not required. We tagged the Z SUPER based on its haploid coverage while the autosomal SUPER were numbered by descending scaffold size. SCAFFOLD_32: is 58Kb long and remains unplaced, it shows inespecific but weak contacts with several SUPER. 18.70% of its length is occupied by repeats, of which 12.79% are interspersed repeats. It has low mappability and this explains why diagonal contact are only visible when we allow multimappings (mq=0). The blastn hits of scaffold_32 correspond to autosomes of other Lepidotera genomes reassuring our decission of including it in the assembly. WARNING: The shared contact map (.pretext) of the curated assembly shows all scaffolds sorted by length. Thus most of the unloc are towards the end and separated from their corresponding SUPER. This will be fixed in future versions of CLAWS."

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	560,905,695	560,905,695
GC %	35.57	35.57
Gaps/Gbp	10.7	10.7
Total gap bp	1,200	1,200
Scaffolds	32	32
Scaffold N50	20,417,927	20,417,927
Scaffold L50	13	13
Scaffold L90	26	26
Contigs	38	38
Contig N50	18,864,205	18,864,205
Contig L50	14	14
Contig L90	28	28
QV	49.1145	49.1145
Kmer compl.	92.3242	92.3242
BUSCO sing.	98.9%	98.9%
BUSCO dupl.	0.1%	0.1%
BUSCO frag.	0.2%	0.2%
BUSCO miss.	0.8%	0.8%

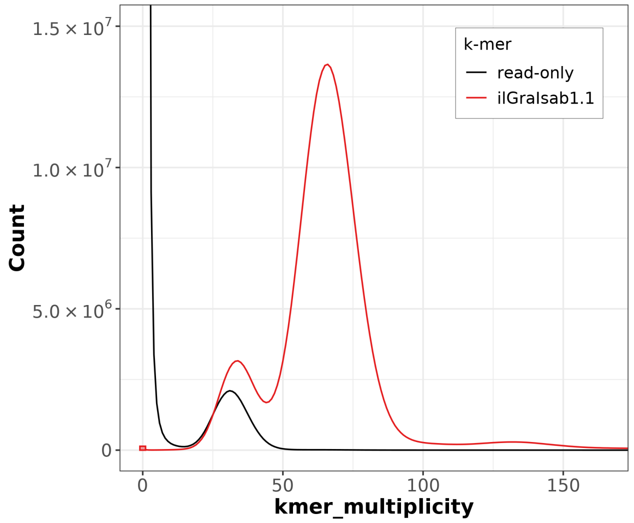
BUSCO 5.4.0 Lineage: insecta_odb10 (genomes:75, BUSCOs:1367)

HiC contact map of curated assembly

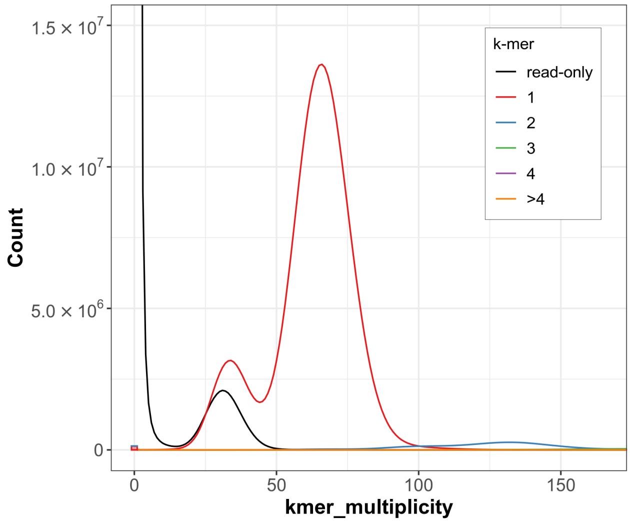


collapsed [LINK]

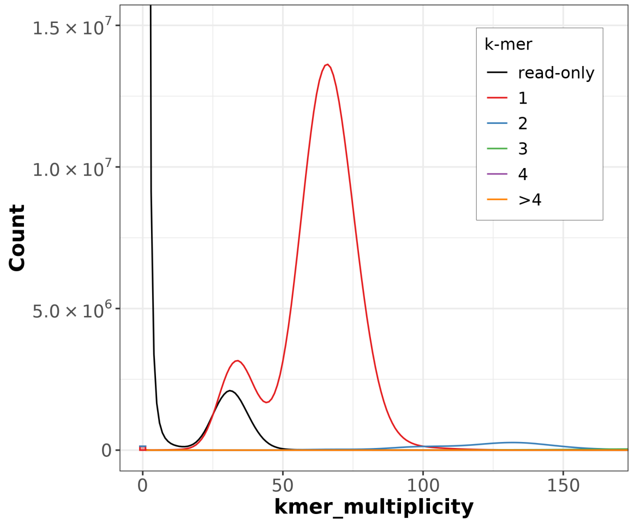
K-mer spectra of curated assembly



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

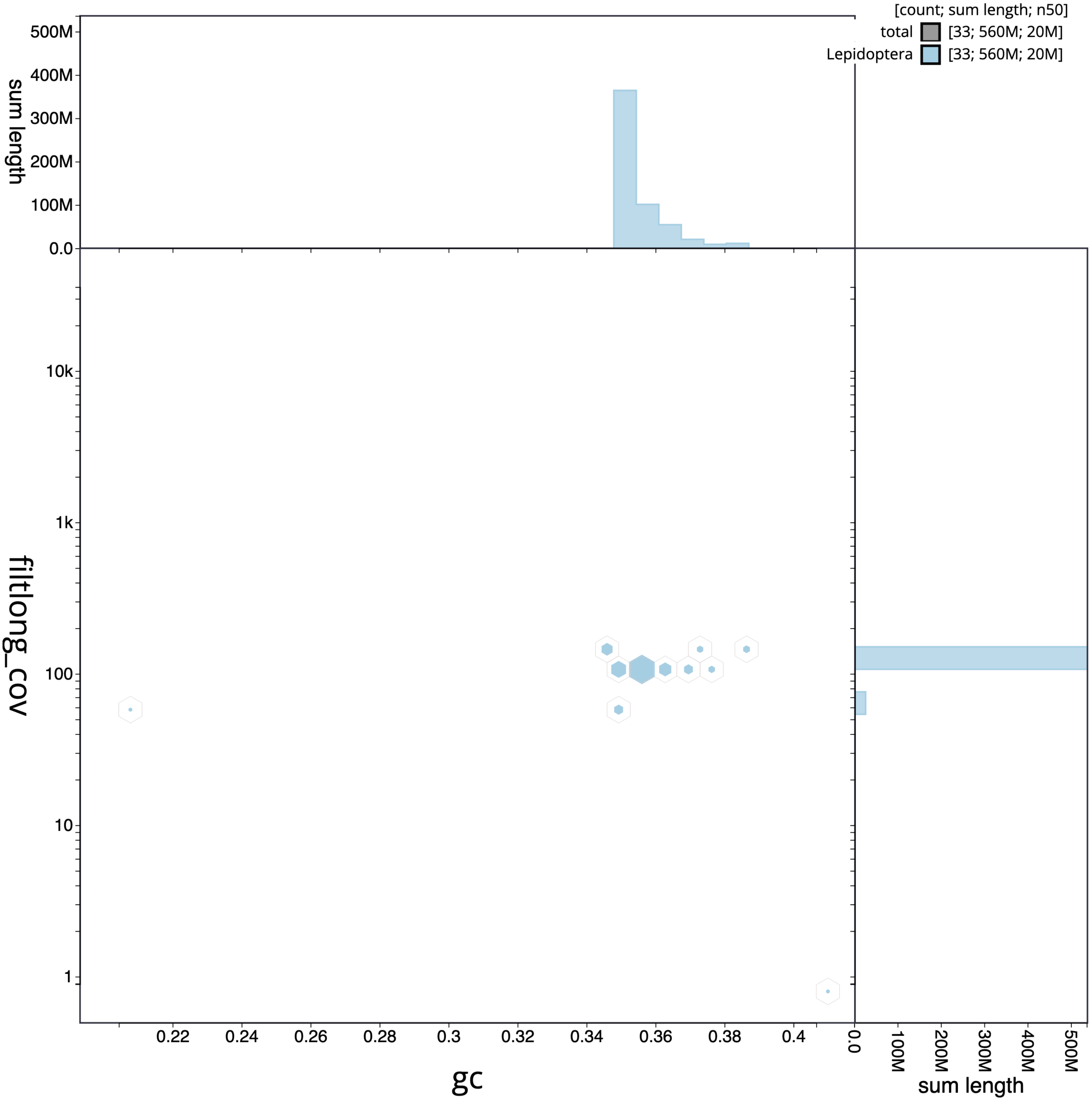


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	ONT	Illumina	OmniC
Coverage	265x	78x	63x

Assembly pipeline

```
- Trim_Galore
    |_ ver: 0.6.7
    |_ key param: "--gzip -q 20"
    |_ key param: "--paired"
    |_ key param: "--retain_unpaired"
    |_ key param: "--max_n 0"
- Filtlong
    |_ ver: 0.2.1
    |_ key param: "--target_bases 70000000000"
- nextdenovo
    |_ ver: 2.5.0
    |_ key param: NA
- hypo
    |_ ver: 1.0.3
    |_ key param: NA
- purge_dups
    |_ ver: 1.2.6
    |_ key param: NA
- YaHS
    |_ ver: 1.2a
    |_ key param: NA
```

Curation pipeline

```
- PretextView
    |_ ver: 0.2.5
    |_ key param: NA
```

Submitter: Fernando Cruz
Affiliation: CNAG Barcelona

Date and time: 2024-03-21 16:01:14 CET