

## Supplementary Data

This Supplementary Data belongs to the paper “MoGAAAP: A modular Snake-make workflow for automated genome assembly and annotation with quality assessment (van Workum et al., 2025)”.

### File descriptions

- README.md: This file.
- 20241122.Lactuca\_serriola.html: The MoGAAAP report belonging to the assembly and annotation of the *Lactuca serriola* US96UC23 genome.
- 20250502.Arabidopsis.html: The MoGAAAP report belonging to the assembly and annotation of the 32 *Arabidopsis thaliana* pangenome.
- 20250508.grape.html: The MoGAAAP report belonging to the assembly and annotation of the 6 grapevine pangenome.
- 20250517.human.html: The MoGAAAP report belonging to the assembly and annotation of the human trio genome dataset (HG002, HG003, HG004).

### Input data

Here, we describe the input data that was used for each run of MoGAAAP. All data is publicly available and was downloaded from ENA (for all starting with ‘SRR’) or CNCB (for all starting with ‘CRR’). All reference genomes were downloaded from NCBI.

Example download commands:

```
# SRA
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR202/039/
  SRR20242239/SRR20242239_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR202/039/
  SRR20242239/SRR20242239_2.fastq.gz

# CRA
wget ftp://download.big.ac.cn/gsa2/CRA008584/CRR591656/
  CRR591656.fastq.gz

# Reference genomes
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/
  /000/001/735/GCF_000001735.4_TAIR10.1/GCF_000001735.4
  _TAIR10.1_genomic.fna.gz
```

### *Lactuca serriola* US96UC23

The HiFi and ONT data used for the *Lactuca serriola* US96UC23 genome assembly is generated and described in the van Workum et al. (2025) paper. We used CNS0047707 for the Illumina data.

Next to this, we used GCF\_002870075.4 (*Lactuca sativa* cv. Salinas) as the reference genome and corresponding annotation for QA.

### ***Arabidopsis thaliana* pangenome**

For the *Arabidopsis thaliana* pangenome, we used the data from the Kang et al. (2023) paper. An overview of the data as given to the MoGAAAP workflow:

Accession ID	HiFi	Hi-C	Illumina
01_col	CRR591656	SRR20242239	-
02_tibet	CRR591657	-	CRR624285
03_yilong	CRR591658	-	CRR624284
04_bor_1	CRR591659	-	-
05_cdm_0	CRR591660	-	-
08_kondara	CRR591662	-	-
12_li_of_095	CRR591665	-	-
13_got_22	CRR591666	-	-
14_st_0	CRR591667	-	-
15_kelsterbach_2	CRR591668	-	-
19_kz_9	CRR591671	-	-
20_ll_0	CRR591672	-	-
21_ms_0	CRR591673	-	-
23_sij_1	CRR591675	-	-
24_hs_0	CRR591676	-	-
25_per_1	CRR591677	-	-
26_nz_1	CRR591678	-	-
27_belmonte_4_94	CRR591679	-	-
29_sij_2	CRR591680	-	-
30_tu_sb30_3	CRR591681	-	-
31_mammo_1	CRR591682	-	-
33_sha	CRR591683	-	-
36_pra_6	CRR591684	-	-
37_pu_2_23	CRR591685	-	-
38_dra_2	CRR591686	-	-
39_ah_7	CRR591687	-	-
40_etna_2	CRR591688	-	-
41_sorbo	CRR591689	-	-
42_arb_0	CRR591690	-	-
43_elk_1	CRR591691	-	-
44_ket_10	CRR591692	-	-
45_meh_0	CRR591693	-	-

Next to this, we used the TAIR10 (GCF\_000001735.4) reference genome and ARAPORT11 annotation for QA. The ARAPORT11 annotation was downloaded from [https://v2.arabidopsis.org/download\\_files/Genes/Araport11\\_genome\\_](https://v2.arabidopsis.org/download_files/Genes/Araport11_genome_)

release/Araport11\_GFF3\_genes\_transposons.current.gff.gz and chromosome names were renamed to match the reference genome.

### Grapevine pangenome

For the grapevine pangenome, we used the data from the Liu et al. (2024) paper. An overview of the data as given to the MoGAAAP workflow:

Accession ID	HiFi	ONT	Hi-C
Manicule_Finger	SRR29686483	SRR29686485	SRR29686480
Muscat_Hamburg	SRR29686484	SRR29686479	SRR29686481
Shine_Muscat	SRR29686472	SRR29686476	SRR29686482
Baimunage	SRR30692114	SRR29686474	SRR30692115
Hongmunage	SRR30692112	SRR29686477	SRR30692113
Wolley	SRR30536351	SRR29686473	SRR30536352

Next to this, we used the GCF\_030704535.1 reference genome and corresponding annotation for QA.

### Human trio genomes

For the human trio genomes, we used the data from the Genome in a Bottle. An overview of the data as given to the MoGAAAP workflow:

Accession ID	HiFi	Hi-C
HG002	SRR26402938	SRR11347815
HG003	SRR26402937	-
HG004	SRR26402936	-

Next to this, we used the GRCh38 (GCA\_000001405.29) reference genome and the GENCODE 46 annotation for QA. The GENCODE 46 annotation was downloaded from [https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_46/gencode.v46.chr\\_patch\\_hapl\\_scaff.annotation.gff3.gz](https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_46/gencode.v46.chr_patch_hapl_scaff.annotation.gff3.gz) and chromosome names were renamed to match the reference genome.

### Running MoGAAAP

All sequencing data was registered in a samples.tsv file according to MoGAAAP specifications. The config.yaml was set up accordingly with the mentioned reference genome and annotation. Examples can be found in the config/examples directory. The workflow was run using Snakemake version 8.23.0 and MoGAAAP version v0.2.2 for the *Lactuca serriola* US96UC23 genome assembly and annotation:

```
snakemake all
```

For the other use cases, we used the most recent version of MoGAAAP at the time of writing (v1.0.2) and snakemake 8.30.0 using the following command:

```
MoGAAAP run --configfile ${CONFIG} -m 1000
```

The reason for not re-running the *Lactuca serriola* US96UC23 genome assembly is that the tools have not significantly between versions and that this is the version that has been submitted to GenBank. The other use cases were not submitted to GenBank but only run for showcasing the MoGAAAP workflow.