Check for updates

DATA NOTE

ERGA-BGE genome of *Xanthium orientale subsp. italicum* (Moretti) Greuter. A plant of American origin, now widespread and, in some cases, invasive in Europe

[version 1; peer review: 2 approved with reservations]

Salvatore Tomasello[1], Eleonora Manzo[1], Rita Monteiro[2], Astrid Böhne[2], Thomas Marcussen[3,4], Torsten H Struck[3], Rebekah A Oomen[3-7], Genomescope Sequencing Team, Alice Moussy[8], Corinne Cruaud[8], Karine Labadie[8], Lola Demirdjian[9], Adama N'dar[9], Patrick Wincker[9], Pedro H Oliveira[9], Jean-Marc Aury[9], Vianey Paola Barrera Enriquez[10], Leanne Haggerty[10], Fergal Martin[10], Tom Brown[11,12]

[1]Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium), Georg-August-University of Göttingen, Untere Karspüle 2, Göttingen, 37073, Germany
[2]Leibniz Institute for the Analysis of Biodiversity Change - Museum Koenig Bonn, Bonn, Germany
[3]Natural History Museum, University of Oslo, Oslo, Norway
[4]Centre for Ecological & Evolutionary Synthesis, University of Oslo, Oslo, Norway
[5]Department of Biological Sciences, University of New Brunswick Saint John, Saint John, Canada
[6]Tjärnö Marine Laboratory, University of Gothenburg, Gothenburg, Sweden
[7]Centre for Coastal Research, University of Agder, Kristiansand, Norway
[8]Genoscope, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057, Evry, France
[9]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Evry, France
[10]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[11]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[12]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

v1 **First published:** 27 Nov 2025, **5**:367
https://doi.org/10.12688/openreseurope.21891.1
Latest published: 27 Nov 2025, **5**:367
https://doi.org/10.12688/openreseurope.21891.1

Abstract

The *Xanthium orientale* subsp. *italicum* (Asterales: Asteraceae: Asteroideae) reference genome offers a valuable resource for understanding the genetic basis of invasiveness and fitness differences among congeneric species. The entirety of the genome sequence was assembled into 18 contiguous chromosomal pseudomolecules. This chromosome-level assembly encompasses 2.26 Gb, composed of 46 contigs and 32 scaffolds, with contig and scaffold N50 values of 98.8 Mb and 124 Mb, respectively.

Open Peer Review

Approval Status ? ?

	1	2
version 1 27 Nov 2025	? view	? view

1. **Nunzio D'Agostino**, University of Naples Federico II, Portici, Italy
2. **Marcial Escudero**, University of Seville, Seville, Spain

Keywords
Xanthium orientale subsp. italicum, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, Asteraceae, cocklebur

Rogelio Sánchez-Villegas, Pablo de Olavide University, Seville, Spain

Inés Gómez-Ramos, University of Seville, Seville, Spain

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon Europe gateway.

This article is included in the Genome Reports from the Biodiversity Genomics Europe Project collection.

Corresponding author: Tom Brown (brown@izw-berlin.de)

Author roles: Tomasello S: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Manzo E**: Investigation, Resources, Writing – Review & Editing; **Monteiro R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Böhne A**: Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Marcussen T**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Struck TH**: Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Oomen RA**: Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Moussy A**: Investigation, Writing – Review & Editing; **Cruaud C**: Investigation, Writing – Review & Editing; **Labadie K**: Investigation, Writing – Review & Editing; **Demirdjian L**: Data Curation, Formal Analysis, Writing – Review & Editing; **N'dar A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Wincker P**: Funding Acquisition, Supervision, Writing – Review & Editing; **Oliveira PH**: Funding Acquisition, Investigation, Supervision, Writing – Review & Editing; **Aury JM**: Data Curation, Formal Analysis, Funding Acquisition, Supervision, Writing – Review & Editing; **Barrera Enriquez VP**: Data Curation, Formal Analysis, Writing – Review & Editing; **Haggerty L**: Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Martin F**: Data Curation, Formal Analysis, Funding Acquisition, Supervision, Writing – Review & Editing; **Brown T**: Visualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: Funding for the collection of the original plant material, cultivation at the Old Botanical Garden of the University of Göttingen, collection and shipment, as well as taxonomic expertise and barcode sequencing was provided by the Deutsche Forschungsgemeinschaft (DFG), project TO1400/1-1, as part of the priority program 1991 "Taxon-Omics". This work was supported by the Genoscope, the Commissariat à l'Energie Atomique et aux Énergies Alternatives (CEA), France Génomique (ANR-10-INBS-09-08) and the exploratory research programme "ATLASea: Atlas of marine genomes" and its targeted project SEQ-Sea (ANR-22-EXAT-0003-SEQ-Sea). This project received funding from Horizon Europe under the Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Tomasello S *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Tomasello S, Manzo E, Monteiro R *et al.* **ERGA-BGE genome of *Xanthium orientale subsp. italicum* (Moretti) Greuter. A plant of American origin, now widespread and, in some cases, invasive in Europe [version 1; peer review: 2 approved with reservations]** Open Research Europe 2025, **5**:367 https://doi.org/10.12688/openreseurope.21891.1

First published: 27 Nov 2025, **5**:367 https://doi.org/10.12688/openreseurope.21891.1

Introduction

The taxon *Xanthium orientale* subsp. *italicum* (Moretti) Greuter is part of a taxonomically complex genus of the sunflower family (Asteraceae) and belongs to the subtribe Ambrosiinae of the Sunflower tribe (Heliantheae), subfamily Asteroideae (Tomasello *et al.,* 2019). *Xanthium* comprises seven species (Manzo *et al.,* 2024; POWO, 2024; Tomasello, 2018) and, like other Ambrosiinae, it is quite peculiar among the Asteraceae for being wind-pollinated. This specific taxon, originally described in Italy, is nowadays widespread in Southern Europe and other areas of the world. Its original distribution range is most probably South-Western North America.

This taxon has become widespread in Europe in ruderal places as well as in coastal habitats and riverbanks. Occasionally, it is also a weed with negative economic effects on agriculture.

Developing a high-quality reference genome for *Xanthium orientale* subsp. *italicum* is essential for a better understanding of the genetic basis of invasiveness and fitness differences among congeneric species. It will also provide a fundamental tool for further taxonomic and evolutionary studies on the genus.

The generation of this reference resource was coordinated by the European Reference Genome Atlas (ERGA) initiative's Biodiversity Genomics Europe (BGE) project, supporting ERGA's aims of promoting transnational cooperation to promote advances in the application of genomics technologies to protect and restore biodiversity (Mazzoni *et al.,* 2023).

Materials & methods

ERGA's sequencing strategy includes Oxford Nanopore Technology (ONT) and/or Pacific Biosciences (PacBio) for long-read sequencing, along with Hi-C sequencing for chromosomal architecture, Illumina Paired-End (PE) for polishing (i.e. recommended for ONT-only assemblies), and RNA sequencing for transcriptomic profiling, to facilitate genome assembly and annotation.

Sample and sampling information

Salvatore Tomasello (ST) sampled one specimen of diclinous monoecious *Xanthium orientale* subsp. *italicum,* cultivated at the "Alter Botanischer Garten", Göttingen, Germany on 06 November 2023. The plant cultivated in Göttingen originated from burs collected by ST at "Quattro Finaite" (38.037267, 13.498884), Santa Flavia, Sicily, Italy. It was identified by ST based on morphology and DNA barcoding on 13 October 2020. Sampling was performed by extracting leaf samples, which were immediately frozen in liquid nitrogen by Eleonora Manzo. Samples were then stored at -80°C and shipped on dry ice within one day to the sequencing facility.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in the herbarium (GOET) of the University of Göttingen, Germany https://www.wikidata.org/wiki/Q55829124 under the accession number GOET064310, and a duplicate specimen is available at the herbarium (B) of the Botanical Garden and Botanical Museum Berlin, Germany https://ww2.bgbm.org/herbarium/default.cfm, under the accession number B 10 0240906.

Silica-gel dried leaf is available from the same individual at the DNA-Bank of the Botanical Garden and Botanical Museum Berlin (BGBM) https://www.ggbn.org/ggbn_portal/search/result?institution=BGBM%2C+Berlin under the voucher ID B GT 0027280.

An electronic voucher image of the sequenced individual is available from JACQ: http://herbarium.bgbm.org/object/B100240906.

Genetic information

The estimated genome size, based on ancestral taxa, is 3.13 Gb. This is a diploid genome with a haploid number of 18 chromosomes (2n=36) and unknown sex chromosomes. All information for this species was retrieved from Genomes on a Tree (Challis *et al.,* 2023).

DNA/RNA processing

DNA was extracted from 0.6 g of leaves using a conventional CTAB extraction followed by a commercial purification using Qiagen Genomic tips (Qiagen). A detailed protocol is available on protocols.io (https://www.protocols.io/view/hmw-dna-extraction-for-long-read-sequencing-using-bp2l694yzlqe/v1). DNA fragment size selection was performed using Short Read Eliminator (PacBio, CA, USA). Quantification was performed using a Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) and integrity was assessed in a FemtoPulse system (Agilent). DNA was stored at 4°C until usage.

RNA was extracted from leaves (50 mg) using RNeasy Powerplant Kit (Qiagen) following manufacturer instructions. Residual genomic DNA was removed with 6U of TURBO DNase (2 U/µl) (Thermo Fisher Scientific). Quantification was performed using a Qubit RNA HS Assay and integrity was assessed in a Bioanalyzer system (Agilent). RNA was stored at -80°C.

Library preparation and sequencing

Long-read DNA libraries were prepared with the SMRTbell prep kit 3.0 following manufacturers' instructions and sequenced on a Revio system (PacBio). Omni-C libraries were prepared using the Dovetail Omni-C Kit (Dovetail Genomics, Scotts Valley, CA, USA) after plant nuclei isolation. Briefly, flash-frozen leaves (1 g) were cryoground in liquid nitrogen and pure nuclei were first isolated following the "High Molecular Weight DNA Extraction from Recalcitrant Plant Species" protocol described by Workman *et al.,* 2018. Once the nuclei had been isolated, the pellets were treated as mammalian cells and Omni-C libraries were prepared according to the Mammalian Cell Protocol for Sample Preparation v1.4. The final libraries were sequenced on an Illumina NovaSeq 6000 instrument (Illumina, San Diego, CA, USA) with 2x150 bp read length. In total 91.4 Gb PacBio HiFi and 23.4 Gb HiC data were sequenced to generate the assembly.

Genome assembly methods

The genome of *Xanthium orientale* subsp. *italicum* was assembled using the Genoscope GALOP pipeline (https://workflowhub.eu/workflows/1200). Briefly, raw PacBio HiFi reads were assembled using Hifiasm v0.19.5-r593. Retained haplotigs were removed using purge_dups v1.2.5 with default parameters and

the proposed cutoffs. The purged assembly was scaffolded using YaHS v1.2 and assembled scaffolds were then curated through manual inspection using PretextView v0.2.5 to remove false joins and incorporate sequences not automatically scaffolded intotheir respective locations within the chromosomal pseudomolecules. The mitochondrial and chloroplast genomes were assembled as a single contig using Oatk v1.0 and included in the released assembly. Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (De Panis, 2024b).

Genome annotation methods

A gene set was generated using the Ensembl Gene Annotation system (Aken *et al.,* 2016), primarily by aligning publicly available short-read RNA-seq data from BioSample: SAMEA104170699, SAMD01571316-21, SAMN00717289, SAMN03445768-70, SAMN16433987, SAMN18740786-97, SAMN20514913-36, SAMN30033481-504 to the genome. Long-read transcriptomic data from BioSamples SAMN23470776 and SAMN23470777 were mapped to the genome using Minimap2 (Li, 2018) with the recommended settings for Iso-Seq and Nanopore data. Gaps in the annotation were filled via protein-to-genome alignments of a select set of clade-specific proteins from UniProt (The UniProt Consortium, 2019), which had experimental evidence at the protein or transcript level. At each locus, data were aggregated and consolidated, prioritising models derived from RNA-seq data, resulting in a final set of gene models and associated non-redundant transcript sets. To distinguish true isoforms from fragments, the likelihood of each open reading frame (ORF) was evaluated against known metazoan proteins. Low-quality transcript models, such as those showing evidence of fragmented ORFs, were removed. In cases where RNA-seq data were fragmented or absent, homology data were prioritised, favouring longer transcripts with strong intron support from short-read data. The resulting gene models were classified into two categories: protein-coding, and long non-coding. Models that did not overlap protein-coding genes, and were constructed from transcriptomic data were considered potential lncRNAs. Potential lncRNAs were further filtered to remove single-exon loci due to their unreliability. Putative miRNAs were predicted by performing a BLAST search of miRBase (Kozomara *et al.,* 2019) against the genome, followed by RNAfold analysis (Gruber *et al.,* 2008). Other small non-coding loci were identified by scanning the genome with Rfam (Kalvari *et al.,* 2018) and passing the results through Infernal (Nawrocki & Eddy, 2013). Summary analysis of the released annotation was performed using the ERGA-BGE Genome Report ANNOT Galaxy workflow (De Panis, 2024a), incorporating tools such as AGAT v1.2, BUSCO v5.5 and OMArk v0.3.

Results

Genome assembly

The genome assembly has a total length of 2,257,535,428 bp in 32 scaffolds including the mitochondrial and chloroplast genomes (Figure 1 & Figure 2), with a GC content of 37.03%. The assembly has a contig N50 of 98,804,340 bp and L50 of

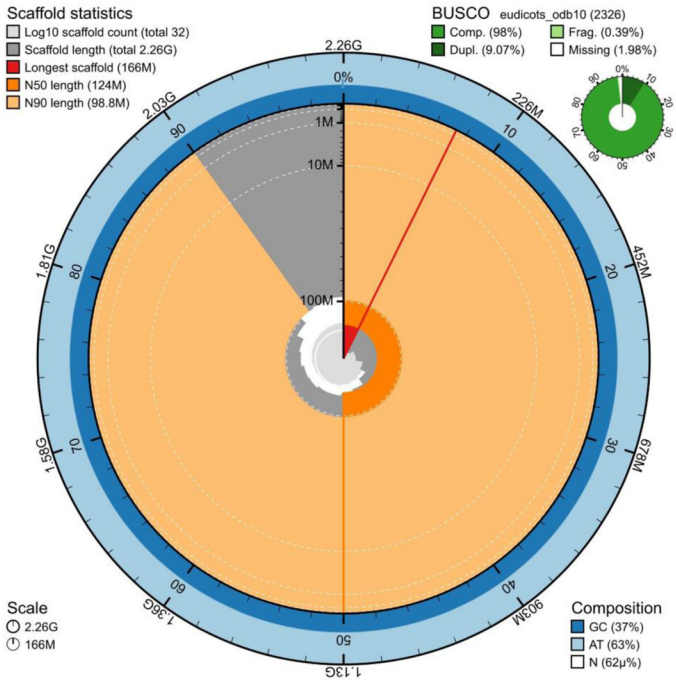

Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 2,257,535,428 bp assembly including the mitochondrial genome. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (164,997,723 bp, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (124,155,882 and 98,804,340 bp), respectively. The pale grey spiral shows the cumulative sequence count on a log-scale, with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes found in the assembled genome from the Eudicots database (odb10) is shown in the top right.

9 and a scaffold N50 of 124,155,882 bp and L50 of 8. The assembly has a total of 14 gaps, totaling 1.4 kb in cumulative size. The single-copy gene content analysis using the Eudicots database with BUSCO (Manni *et al.*, 2021) resulted in 98.1% completeness (89.0% single and 9.1% duplicated). 99.6% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 69.8 as calculated by Merqury (Rhie *et al.*, 2020).

Genome annotation

The genome annotation consists of 41,760 protein-coding genes with an associated 63,957 transcripts, in addition to 22,843 non-coding RNA genes of various types (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the eudicot_odb10 database with BUSCO, run in protein mode, resulted in 94.7% completeness. Using the OMAmer Asterids database for OMArk resulted in 96.29% completeness and 83.88% consistency (Table 2).

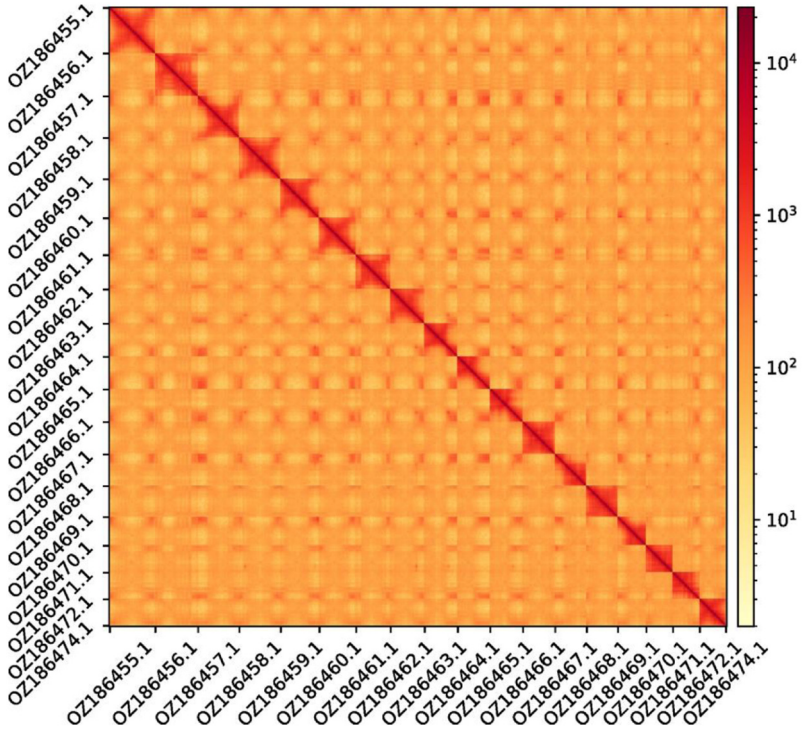

Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 100 kb bin size for plotting.

Table 1. Statistics from assembled gene models.

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
Protein-coding	41,760	63,957	3,483	2,956	5.5
lncRNA	17,930	23,660	2,193	119	2.5
snRNA	175	175	142	175	1.0
snoRNA	823	823	106	823	1.0
rRNA	769	769	454	769	1.0
tRNA	813	813	75	813	1.0
Other non-coding	2,333	2,333	84	2,333	1.0

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (eudicots_odb) and OMArk (Asterids).

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	2,202 (94.7%)	2,009 (86.4%)	193 (8.3%)	35 (1.5%)	89 (3.8%)
OMArk	10,783 (96.29%)	6,502 (58.06%)	4,281 (38.23%)	-	415 (3.71%)
	Consistent	**Inconsistent**	**Contaminants**	**Unknown**	
OMArk	35,029 (83.88%)	899 (2.15%)	0 (0.05)	5,832 (13.97%)	

Data availability

X. orientale subsp. *italicum* and the related genomic study were assigned to Tree of Life ID (ToLID) 'daXanOrie1' and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB77262. The sample information is available at the following BioSample accessions: SAMEA115083666 and SAMEA116115597. The genome assembly is accessible from GenBank under accession number GCA_964271845.1.

Sequencing data produced as part of this project are available from SRA at the following accessions: ERX13095982, ERX13095983, and ERX13158964. Documentation related to the genome assembly and curation can be found in the ERGA Assembly Report (EAR) document available at https://github.com/ERGA-consortium/EARs/tree/main/Assembly_Reports/Xanthium_orientale/daXanOrie1. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/1534734.

Author contributions

ST collected the species, ST identified the species, ST and EM sampled and preserved biological material, ST provided metadata, TM, RM, RO, THS and AsB provided support in sampling, shipping of biological material, metadata collection, and management, the GST extracted DNA, prepared libraries, and performed sequencing under the supervision of AM, CC, KL, PHO and PW; LD, AN and JMA performed genome assembly and curation, LH and FM performed genome annotation and TB generated the analysis and report. All authors contributed to the writing, review, and editing of this genome note and read and approved the final version. This work is part of the species assigned to Genoscope, which was instrumental in the wet lab, sequencing, and assembly processes, and represents a key contribution to BGE's outputs

Author information

Members of the Genoscope Sequencing Technical Team are listed here: https://doi.org/10.5281/zenodo.146114

Acknowledgements

We would like to acknowledge the assembly reviewer, Tyler Alioto, from the Centro Nacional de Análisis Genómico (CNAG). The authors acknowledge the support of the Freiburg Galaxy Team: Saim Momin and Björn Grüning, Bioinformatics, University of Freiburg (Germany), funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

De Panis D: **ERGA-BGE genome report ANNOT analyses**. WorkflowHub, 2024a.
Publisher Full Text

De Panis D: **ERGA-BGE genome report ASM analyses (one-asm HiFi + HiC)**. WorkflowHub, 2024b.
Publisher Full Text

Gruber AR, Lorenz R, Bernhart SH, *et al.*: **The Vienna RNA websuite.** *Nucleic Acids Res.* 2008; **36**(Web Server issue): W70–W74.
PubMed Abstract | Publisher Full Text | Free Full Text

Kalvari I, Nawrocki EP, Argasinska J, *et al.*: **Non-coding RNA analysis using the Rfam database.** *Curr Protoc Bioinformatics.* 2018; **62**(1): e51.
PubMed Abstract | Publisher Full Text | Free Full Text

Kozomara A, Birgaoanu M, Griffiths-Jones S: **miRBase: from microRNA sequences to function.** *Nucleic Acids Res.* 2019; **47**(D1): D155–D162.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.**

Bioinformatics. 2018; **34**(18): 3094–3100.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Manzo E, Nillies S, Gutiérrez DG, *et al.*: **Species delimitation in *Xanthium* sect. *Acanthoxanthium* (Heliantheae, Asteraceae) and the neglected species *Xanthium argenteum*.** *TAXON.* 2024; **73**(5): 1251–1271.
Publisher Full Text

Mazzoni CJ, Ciofi C, Waterhouse RM: **Biodiversity: an atlas of European reference genomes.** *Nature.* 2023; **619**(7969): 252.
PubMed Abstract | **Publisher Full Text**

Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics.* 2013; **29**(22): 2933–2935.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

POWO: **Plants of the World Online | Kew Science**. Plants of the World Online, 2024.
Reference Source

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

The UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Tomasello S: **How many names for a beloved genus? – Coalescent-based species delimitation in *Xanthium* L. (Ambrosiinae, Asteraceae).** *Mol Phylogenet Evol.* 2018; **127**: 135–145.
PubMed Abstract | **Publisher Full Text**

Tomasello S, Stuessy TF, Oberprieler C, *et al.*: **Ragweeds and relatives: molecular phylogenetics of Ambrosiinae (Asteraceae).** *Mol Phylogenet Evol.* 2019; **130**: 104–114.
PubMed Abstract | **Publisher Full Text**

Workman R, Timp W, Fedak R, *et al.*: **High molecular weight DNA extraction from Recalcitrant Plant Species for third generation sequencing V.1.** 2018.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 12 January 2026

https://doi.org/10.21956/openreseurope.23683.r66304

© **2026 Escudero M et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Marcial Escudero
University of Seville, Seville, Andalusia, Spain
Rogelio Sánchez-Villegas
Molecular Biology and Biochemical Engineering, Pablo de Olavide University, Seville, Andalusia, Spain
Inés Gómez-Ramos
Plant Biology and Ecology, University of Seville, Seville, Seville, Spain

This paper presents the genome assembly of *Xanthium orientale* subsp. *italicum*, a species from which there is no previous assembly. The assembly reaches the chromosome level with an adequate quality up to standard with other assemblies in this journal. The methods of both the DNA and rna extraction, hi-C protocol, sequencing and bioinformatic analysis are exposed with sufficient detail, ensuring that replication is possible. But some attention should be put into references in the methods, regarding pipeline and protocol citation (the details of the corrections are specified below).

Regarding the results, the annotation results and the comparison between the two annotation evaluation methods, BUSCO and the OMAr, is interesting, innovative and clearly stated. However, to align with standards in this journal, we suggest supplementing the manuscript with two additional summary tables.. The first should describe the overall assembly using metrics comparable to those in Table 1 of Ruhsam et al. (2025), including assembly name, assembly accession, assembly level for primary assembly, Span, number of contigs, number of scaffolds, N50 of contig and scaffolds, *k*-mer completeness, QV, BUSCO and percentage of the assembly mapped to chromosomes and organelles. The second table should catalog information regarding the chromosomal pseudomolecules (INSDC accession, Name, Length (Mb) and GC%), similar to table 3 in Ruhsam et al. (2025). The inclusion of these tables, especially the second one with the INSDC accession of the pseudomolecules, would make the dataset more accessible. In addition, the BlobToolKit GC-coverage plot would be appreciated as it is normally included in this kind of papers but if GC content is included within the chromosomal pseudomolecule table it is not strictly necessary. Also, the number of pseudomolecules should be clearly stated in the text of the results section even if it was previously mentioned in the abstract, on top of what percentage of the

genome is included within those pseudomolecules.

Overall, the paper adds a valuable contribution with the assembly of this species and genome annotation. The annotation evaluation through the two databases, BUSCO and OMArK, is innovative and a valuable contribution by itself. Both the assembly and the annotation appear to give satisfactory results. However, more results of the assembly should be given.

Minor comments

- The last sentence of the first paragraph of the Introduction ("Its original distribution range is most probably South-Western North America") would benefit from the inclusion of a reference supporting this claim.
- In genetic information, it would be appreciated more information from the ancestral taxa from which the information was taken.
- Cite the protocol from DNA/RNA extraction in protocols.io properly, in an article format: (Vacherie et al, 2022) in the text and then adding the complete citation in the references.
- In the first sentence of the library preparation and sequencing section, please substitute "manufacturers' instructions" with "manufacturer instructions".
- Cite the genome assembly pipeline properly like an article, in the text as (Istace et al. 2024) and then add the complete citation to the references.

https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1200.2

- The sizes of the Hi-C and PacBio data should be included in the results section not in Methods. Also, the PacBio mean coverage should be indicated.

References

1. Istace, B., Aury, J.-M., & Belser, C. (2024). GALOP - Genome Assembly using Long reads Pipeline. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1200.2.
2. Ruhsam M, Griffiths A, Royal Botanic Garden Edinburgh Genome Acquisition Lab et al. (2025). The genome sequence of Bottle Sedge, Carex rostrata Stokes. Wellcome Open Res 10:132. https://doi.org/10.12688/wellcomeopenres.23874.1.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biodiversity Genomics, Sytematics and Evolution

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 29 December 2025

https://doi.org/10.21956/openreseurope.23683.r65860

© 2025 D'Agostino N. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Nunzio D'Agostino
University of Naples Federico II, Portici, Italy

The manuscript by Tomasello et al. presents the high-quality reference genome for Xanthium orientale subsp. italicum, a wind-pollinated species of the Asteraceae with a complex taxonomy. By providing this genomic resource, the study offers an essential foundation for investigating the genetic basis of invasiveness, fitness differences among congeners, and for advancing taxonomic and evolutionary research.
The manuscript is generally well-written and clearly structured. However, several inconsistencies and missing methodological details are present, which should be addressed to ensure clarity, reproducibility, and accurate interpretation of the results.
Abstract: The phrase "the entirety of the genome sequence was assembled" is misleading and should be revised. Because this is not a telomere-to-telomere (T2T) assembly, the word "entirety" overstates the completeness of the result. A more accurate formulation is simply "the genome sequence was assembled." This avoids implying full coverage or gapless assembly and better reflects the actual scope of the work.

Keywords: in scientific writing, genus, species, and subspecies names should always be italicized

Main text:
The statement "Its original distribution range is most probably South-Western North America" requires a citation to support the claim. Additionally, the phrasing could be improved for clarity and precision. Since the distribution refers to historical range, it would be more accurate to use past tense, e.g., "was most likely in southwestern North America". Note that "southwestern" should not be capitalized or hyphenated.

The statement "It was identified by ST based on morphology and DNA barcoding on 13 October 2020" lacks crucial methodological details. For transparency and reproducibility, the authors should specify: Which DNA barcode marker(s) were used and which bioinformatics procedure followed for species identification.

The statement "Sampling was performed by extracting leaf samples, which were immediately frozen in liquid nitrogen by Eleonora Manzo" is unusual for the main text. Typically, the names of individuals performing specific tasks are reported in the Author Contributions section, not in the Methods. This applies equally to anyone who performed sampling or other experimental

procedures (see earlier in the main text).

Typo here: “scaffolded intotheir respective locations”.

The statement “The mitochondrial and chloroplast genomes were assembled as a single contig using Oatk v1.0 and included in the released assembly” is unclear and potentially misleading. Earlier in the Methods, the authors indicate that plant nuclei were isolated prior to library preparation, which implies that cytoplasmic organelle DNA should have been largely excluded from the sequencing libraries. It is therefore unclear how mitochondrial and chloroplast genomes could have been assembled from these nuclear-enriched samples.
The authors should clarify: (i) Whether organelle DNA was present in the raw sequencing data; (ii) How the organelle genomes were assembled despite nuclear isolation; (iii) Any additional methods or steps used to capture or assemble the mitochondrial and chloroplast genomes. Without this clarification, the statement may be confusing to readers and could misrepresent the assembly strategy.

The phrase “Gaps in the annotation” is unclear and unusual. It is not evident whether the authors are referring to missing gene annotations, regions of low confidence, unannotated genomic segments, or sequencing/assembly gaps. The authors should clarify exactly what they mean and consider rephrasing to a more precise and conventional description.

The statement “At each locus, data were aggregated and consolidated” is vague and lacks essential methodological detail. It is unclear how the aggregation and consolidation were performed, and which software, tools, or algorithms were used for this process.
The term “homology data” is somewhat imprecise and could be misleading, as it implies inferred evolutionary relationships rather than direct sequence similarity. It would be more accurate to describe these data as “similarity-based data” or “data derived from sequence similarity”, which better reflects the methodology used.

The statement “The genome assembly has a total length of 2,257,535,428 bp in 32 scaffolds including the mitochondrial and chloroplast genomes... with a GC content of 37.03%” is misleading. It is not appropriate to include cytoplasmic organelle genomes in the calculation of the average nuclear GC content, because organelle genomes often have distinct GC compositions compared to the nuclear genome. The authors should clarify whether the reported GC content refers solely to the nuclear genome or includes organellar sequences. If organelle sequences were included, it would be more accurate to report the nuclear genome GC content separately and provide the mitochondrial and chloroplast GC content as additional, distinct values. This distinction is important for accurate genomic characterization and for meaningful comparisons with other species.

The manuscript would benefit from a dedicated section that explicitly discusses the potential for reuse of the dataset presented. Such a section should highlight possible applications, including comparative genomics, functional studies, evolutionary analyses, and, as suggested by the authors, investigations into the genetic basis of invasiveness and fitness differences among congeners. Providing guidance on how other researchers might exploit the data for these purposes would enhance the transparency, reproducibility, and overall utility of this genomic resource for the broader scientific community.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
No

Are the datasets clearly presented in a useable and accessible format?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant Bioinformatics and genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.