

GENOME SEQUENCING

The complete genome sequences of three Japanese *Arisaema* (Araceae) species

Jacob B. Landis¹, Justin Scholten¹, Chelsea D. Specht¹

¹ School of Integrative Plant Science, Section of Plant Biology and the L.H. Bailey Hortorium, Cornell University, Ithaca, NY, 14853, USA

<https://doi.org/10.56179/001c.160257>

Biodiversity Genomes

Here we present three complete genomes of *Arisaema* from Japan based on HiFi sequencing: *A. ehimense*, *A. serratum*, and *A. tosaense*. The assembled genomes have between 6,443 and 62,971 scaffolds. BUSCO scores of 88.6% or higher indicate all three genomes are mostly complete at least with respect to major genes and gene families. The *A. tosaense* genome has a LAI score of 13.92, indicating an assembly of reference quality.

Introduction

The genus *Arisaema* Mart. (Araceae) comprises more than 200 species of herbaceous understory plants distributed primarily across eastern Asia, with Japan representing a major center of diversity (Murata 2011; Ohi-Toma et al. 2016). Species are notable for their complex reproductive biology, including environmentally mediated sex expression (Scholten and Specht 2025) and specialized pollination systems involving fungus gnats and other dipterans through mushroom mimicry (Kakishima et al. 2019). Within Japan, the *Arisaema serratum* complex represents a taxonomically challenging group characterized by morphological variation, overlapping distributions, and frequent reports of putative hybridization (Murata and Ohashi 2009; Scholten et al. 2025). These taxa typically occur in shaded forest understories across elevational gradients, where ecological overlap and reproductive plasticity may facilitate gene flow among species (Murata et al. 2018).

Within this complex, *Arisaema serratum* (Thunb.) Schott is a widespread and morphologically variable species, while *A. tosaense* Makino and related taxa form a geographically structured complex in western Japan (Murata 2011). *Arisaema ehimense* J. Murata & J. Ohno, a narrow endemic species restricted to Shikoku Island, was described as a putative hybrid between *A. serratum* and *A. tosaense* based on morphological intermediacy. Cytogenetic observations indicate that these taxa are diploid ($2n = 28$), and *A. ehimense* exhibits high pollen viability, seed set, and stable population structure leading to its hypothesized status as a homoploid hybrid species (Murata and Ohno 1989). Subsequent allozyme analyses (Maki and Murata 2001) and allele-sharing surveys suggest shared ancestry with both parental taxa, but were inconclusive in resolving its evolutionary origin thus leaving its status as a homoploid hybrid species unresolved (Scholten and Specht 2025).

Despite long-standing interest in hybridization within *Arisaema*, genomic resources for the genus remain limited, constraining efforts to evaluate patterns of gene flow, lineage divergence, and the genetic basis of reproductive isolation. The availability of reference genome sequences for *A. serratum*, *A. tosaense*, and *A. ehimense* provides an important resource for future studies of reticulate evolution in this system. These genomic data will enable detailed investigation of genome structure, patterns of variation, and candidate loci associated with ecological and reproductive traits, and contribute to broader efforts to understand the role of hybridization in plant diversification (Rieseberg 1997; Payseur and Rieseberg 2016).

Methods

Leaf tissue for genome sequencing was collected from natural populations of *Arisaema serratum*, *A. tosaense*, and *A. ehimense* on Shikoku Island, Japan, in May 2024 (Figure 1). The sequenced individual of *A. ehimense* was collected on 10 May 2024 from a coastal roadside population in Ehime Prefecture (191 m elevation). The *A. tosaense* individual was collected on 8 May 2024 from an altitudinal plateau of Mount Saragamine, Kamiukena District (33°43.422' N, 132°53.107' E; 951 m elevation). The *A. serratum* individual was collected on 5 May 2024 from a lowland slope in a Japanese red cedar forest in Kumakogen, Kamiukena District (33°40.236' N, 132°54.750' E; 678 m elevation).

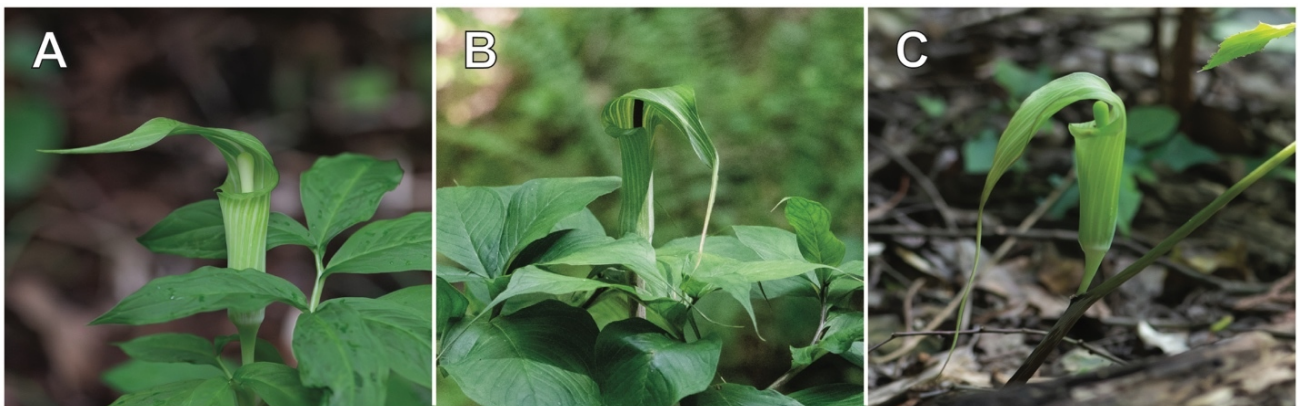


Figure 1. In situ photographs of three *Arisaema* species from Shikoku Island, Japan: (A) *A. serratum*, (B) *A. ehimense*, and (C) *A. tosaense*. Images were taken by J. Scholten.

Fresh leaves of each species were collected in the field and shipped back to Cornell University for data generation and storage at -80°C. Since each individual only had a single leaf or two leaves at the time of collection, there was not enough vegetative material left from the collected sample for sequencing for a herbarium voucher. Therefore, a second individual from the same population was collected to serve as the voucher deposited in the L. H. Bailey Hortorium Herbarium (BH) at Cornell University (Table 1). Frozen leaf material was sent to the Genomics Core Facility at the Icahn School of Medicine Mount Sinai (New York City, USA) for DNA extraction

and PacBio HiFi sequencing. DNA was extracted using the NucleoSpin kit (Machery-Nagel, Dueren, Germany), followed by a BluePippin size selection. Library preparation used the PacBio SMRTbell Prep 3.0 (Pacific Biosciences, California, USA) designed for large genomes. A total of three SMRT cells were sequenced on the PacBio Revio to generate HiFi reads for each species.

Adapter sequences were filtered from the raw sequencing files using HiFiAdapterFilt (Sim et al. 2022) with a minimum length of 44 bp and at least a 97% similarity match. HiFiasm v0.25.0-r726 (Cheng et al. 2021) was used to assemble each genome using eight rounds of assembly cleaning and the purge level set at three. Each assembly was scaffolded using Samba (Zimin and Salzberg 2022) as packaged in MaSuRCA v4.1.3 (Zimin et al. 2013) with the cleaned HiFi reads. The scaffolded assemblies were then sorted largest to smallest, keeping only scaffolds that were longer than 1,000 bp using seqkit v2.7.0 (Shen et al. 2016). The sorted genome was then checked for known contaminants and adapter sequencings using the foreign contaminants screener FCS-adaptor and FCS-GX (Astashyn et al. 2024). Cleaned assemblies were then annotated using Helixer v0.3.6 (Holst et al. 2025), which uses Deep Neural Networks and Hidden Markov Models to predict gene models without the need of generating species specific RNA-Seq data. The resulting gff3 file was used to generate CDS and protein sequence files using gffread v0.9.12 (Pertea and Pertea 2020).

Quality of the genome assemblies were assessed with BUSCO v.5.5.0 (Manni et al. 2021) using the embryophyta odb10 library. The LTR Assembly Index (LAI) was calculated using LTR_retriever v2.9.9 (Ou et al. 2018) to assess the contiguity of the assemblies, with draft assemblies having a value 0-10, reference quality 10-20, and gold standard 20 and above. Predicted genes from Helixer were functionally annotated by comparing the predicted genes against the Uniprot/Swissprot, GO, and KEGG databases with BLASTx v2.1.3.0 for nucleotide sequences, and BLASTp v2.13.0 (Camacho et al. 2009) and HMMER v3.4 (hmmer.org) for protein sequences. Scaffolded genomes have been deposited in NCBI (BioProject PRJNA1443511), while the genomes and annotation files, including coding genes, proteins, and functional annotations are available on Cornell eCommons (Landis et al. 2026).

Results and Data Availability

Total sequencing yield for the three SMRT cells for each species was 115 Gbp for *A. ehimense*, 155 Gbp for *A. serratum*, and 173 Gbp for *A. tosaense*. Average length of HiFi reads were 10,813 bp for *A. ehimense*, 10,848 for *A. serratum*, and 11,854 for *A. tosaense*. Accession information for each sample can be found in [Table 1](#).

Table 1. The three species sequenced and annotated in this study including herbarium voucher number, accession number for the raw data on the Sequence Read Archive, and genome assembly accession number from NCBI.

Taxon	Voucher	SRA	Genome Accession
<i>Arisaema ehimense</i>	J.T.Scholten 472 (BH)	SRR37808430	JBWRXU000000000
<i>Arisaema serratum</i>	J.T.Scholten 466 (BH)	SRR37808429	JBWRXV000000000
<i>Arisaema tosaense</i>	J.T.Scholten 469 (BH)	SRR37808428	JBWRXW000000000

Genome assembly stats indicate that all three assemblies are fragmented and not chromosome-scale, but are mostly complete with respect to gene content based on BUSCO scores. The *A. ehimense* genome was the most fragmented with 62,971 scaffolds and a N50 of 586 Kbp. This species had the largest assembled genome at 8.61 Gbp with a BUSCO score of 88.6% complete, 8.1% fragmented and 3.3% missing (Table 2). Both *A. serratum* and *A. tosaense* were more contiguous with 7,742 and 6,443 contigs respectively as compared to *A. ehimense*, both with N50 values of 3 Mbp. Both the *A. serratum* and *A. tosaense* assemblies also had higher BUSCO scores relative to *A. ehimense* with 93.4% complete (*A. serratum*) and 92.5% complete (*A. tosaense*). The *A. tosaense* genome also had a LAI score of 13.92, indicating that it qualifies as being “reference” quality. The genomes of *A. ehimense* and *A. serratum* failed to generate LAI scores.

Helixer predicted between 55,236 and 62,735 genes across the three genomes (Table 2). Functional annotations of the predicted genes showed that 66.4% (41,659/62,735) of the annotated genes of *A. ehimense*, 69.6% (38,427/55,236) of the annotated genes in *A. serratum*, and 69.7% (39,736/56,937) of the annotated genes of *A. tosaense* had known predicted functions.

Table 2. Genome assembly statistics for each species including total assembly size, number of scaffolds, N50 of scaffolds, the LTR Assembly Index (LAI) value, BUSCO scores, and the number of annotated genes from Helixer. BUSCO scores are shown as the percentage of complete BUSCO loci, the percentage of single copy and duplicated loci, along with fragmented and missing loci.

Taxon	Assembly size (Gbp)	Number of scaffolds	N50 (Mbp)	LAI	BUSCO genome	Number of annotated genes
<i>Arisaema ehimense</i>	8.61	62,971	0.586	–	C:88.6%[S:47.5%,D:41.1%],F:8.1%,M:3.3%,n:1614	62,735
<i>Arisaema serratum</i>	6.27	7,742	3	–	C:93.4%[S:63.9%,D:29.5%],F:4.3%,M:2.3%,n:1614	55,236
<i>Arisaema tosaense</i>	6.20	6,443	3	13.92	C:92.5%[S:62.0%,D:30.5%],F:5.3%,M:2.2%,n:1614	56,937

Funding

This work was supported by funds provided by the Lewis and Clark Fund for Exploration and Research, the Torrey Botanical Society, the Society of Systematic Biologists, the Mario Einaudi Center of International Studies, and the Moore Fund of the L.H. Bailey Hortorium Herbarium. Additionally, the stipend of JS was supported in part by the Tara Atluri '24 Herbarium Engagement Fund.

Acknowledgements

The authors thank Olivia Hullihen and Shun Tanaka for their valuable assistance in preparing the leaf samples during fieldwork; Maya Fridrikh, Irene Salib, and the rest of the Mount Sinai sequencing core for data generation; Jesus Martinez-Gomez for administrative help with genome sequencing; the Cornell BioHPC for computational resources; and Charlie Hale and Zong-Yan Liu for recommending Helixer.

Submitted: April 08, 2026 EDT. Accepted: April 09, 2026 EDT. Published: April 09, 2026 EDT.

REFERENCES

- Astashyn, A., E. S. Tvedte, D. Sweeney, et al. 2024. “Rapid and Sensitive Detection of Genome Contamination at Scale with FCS-GX.” *Genome Biology* 25: 60. <https://doi.org/10.1186/s13059-024-03198-7>.
- Camacho, C., G. Coulouris, V. Avagyan, et al. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10: 421. <https://doi.org/10.1186/1471-2105-10-421>.
- Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. “Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm.” *Nature Methods* 18: 170–75. <https://doi.org/10.1038/s41592-020-01056-5>.
- Holst, F., A. M. Bolger, F. Kindel, et al. 2025. “Helixer: Ab Initio Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model.” *Nature Methods*, 1–8. <https://doi.org/10.1038/s41592-025-02939-1>.
- Kakishima, S., N. Tuno, K. Hosaka, T. Okamoto, T. Ito, and Y. Okuyama. 2019. “A Specialized Deceptive Pollination System Based on Elaborate Mushroom Mimicry.” *bioRxiv* 819136. <https://doi.org/10.1101/819136>.
- Landis, J. B., J. Scholten, and C. D. Specht. 2026. “Data from: The Complete Genome Sequences of Three Japanese *Arisaema* (Araceae) Species.” Cornell University Library eCommons Repository. <https://doi.org/10.7298/gatf-7z66>.
- Maki, M., and J. Murata. 2001. “Allozyme Analysis of the Hybrid Origin of *Arisaema chimense* (Araceae).” *Heredity* 86: 87–93. <https://doi.org/10.1046/j.1365-2540.2001.00813.x>.
- Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38: 4647–54. <https://doi.org/10.1093/molbev/msab199>.
- Murata, J. 2011. *The Picture Book of Plant Systematics in Color: Arisaema in Japan*. Hokuryukan Publishing Co.
- Murata J., and Ohashi H. 2009. “Taxonomic history of *Arisaema serratum* and *A. japonicum*.” *Bunrui* 9: 37–45.
- Murata J., and Ohno J. 1989. “*Arisaema chimense* J. Murata et Ohno (Araceae), a new species from Shikoku, Japan, of putative hybrid origin.” *Shokubutsu kenkyu zasshi. The Journal of Japanese botany* 64: 341–51.
- Murata, J., J. Ohno, T. Koeayashi, and T. Ohi-Toma. 2018. *The Genus Arisaema in Japan*. Hokuryukan Publishing Co.
- Ohi-Toma, T., S. Wu, H. Murata, and J. Murata. 2016. “An Updated Genus-Wide Phylogenetic Analysis of *Arisaema* (Araceae) with Reference to Sections.” *Botanical Journal of the Linnean Society. Linnean Society of London* 182: 100–114. <https://doi.org/10.1111/boj.12459>.
- Ou, S., J. Chen, and N. Jiang. 2018. “Assessing Genome Assembly Quality Using the LTR Assembly Index (LAI).” *Nucleic Acids Research* 46: e126. <https://doi.org/10.1093/nar/gky730>.
- Payseur, B. A., and L. H. Rieseberg. 2016. “A Genomic Perspective on Hybridization and Speciation.” *Molecular Ecology* 25: 2337–60. <https://doi.org/10.1111/mec.13557>.
- Pertea, G., and M. Pertea. 2020. “GFF Utilities: GffRead and GffCompare.” *F1000Research* 9: 304. <https://doi.org/10.12688/f1000research.23297.2>.
- Rieseberg, L. H. 1997. “Hybrid Origins of Plant Species.” *Annual Review of Ecology and Systematics* 28: 359–89. <https://doi.org/10.1146/annurev.ecolsys.28.1.359>.

- Scholten, J., and C. Specht. 2025. "Does Size Matter? The Integrated Roles of Light, Adaptive Sex Expression, and Hybridization in a Widespread *Arisaema* (Araceae) Species from Western Japan." *ARPHA Preprints*, e177788.
- Scholten, J., A. Sprenger, O. Hullihen, and C. D. Specht. 2025. "A Bookend in the *Arisaema japonicum* (Araceae) Taxonomic Debate: Morphological and Genetic Evidence for Synonymization." *Phytotaxa* 725: 1–10. <https://doi.org/10.11646/phytotaxa.725.1.1>.
- Shen, W., S. Le, Y. Li, and F. Hu. 2016. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation." *PloS One* 11: e0163962. <https://doi.org/10.1371/journal.pone.0163962>.
- Sim, S. B., R. L. Corpuz, T. J. Simmonds, and S. M. Geib. 2022. "HiFiAdapterFilt, a Memory Efficient Read Processing Pipeline, Prevents Occurrence of Adapter Sequence in PacBio HiFi Reads and Their Negative Impacts on Genome Assembly." *BMC Genomics* 23: 157.
- Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke. 2013. "The MaSuRCA Genome Assembler." *Bioinformatics (Oxford, England)* 29: 2669–77.
- Zimin, A. V., and S. L. Salzberg. 2022. "The SAMBA Tool Uses Long Reads to Improve the Contiguity of Genome Assemblies." *PLoS Computational Biology* 18: e1009860. <https://doi.org/10.1371/journal.pcbi.1009860>.