bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome assembly of the bearded iris *Iris pallida* Lam.

Author list: Robert E. Bruccoleri[1], Edward J. Oakeley[2], Ann Marie E Faust[3], Marc Altorfer[2], Sophie Dessus-Babus[2], David Burckhardt[2], Mevion Oertli[2], Ulrike Naumann[2], Frank Petersen[2], Joanne Wong[2*]

1. Congenomics, LLC, Glastonbury, CT, US

2. Novartis Institutes for BioMedical Research, Novartis Campus, 4056, Basel, Switzerland

3. Novartis Institutes for BioMedical Research, 250 Massachusetts Avenue, Cambridge, MA, US

* Corresponding Author

bruc@congen.com; edward.oakeley@novartis.com; ann_marie.faust@novartis.com; marc.altorfer@novartis.com; sophie.dessusbabus@10xgenomics.com; david.burckhardt@novartis.com; mevion.oertli@novartis.com; ulrike.naumann@novartis.com; frank.petersen@novartis.com; joanne.wong@novartis.com

Keywords

Long-read sequencing, transcriptome

ORCID:

Robert Bruccoleri: 0009-0001-5687-4554

Edward Oakeley: 0000-0001-6226-7470

Ann Marie Faust: 0000-0002-7468-5984

Ulrike Naumann: 0000-0001-7783-7675

Frank Petersen: 0000-0002-2636-0421

Joanne Wong: 0000-0001-6535-3307

Marc Altorfer: 0009-0004-6074-5087

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Irises are perennial plants, representing a large genus with hundreds of species. While cultivated extensively for their ornamental value, commercial interest in irises lies in the secondary metabolites present in their rhizomes. The Dalmatian Iris (*Iris pallida* Lam.) is an ornamental plant that also produces secondary metabolites with potential value to the fragrance and pharmaceutical industries. In addition to providing base notes for the fragrance industry, iris tissues and extracts possess anti-oxidant, anti-inflammatory, and immunomodulatory effects. However, study of these secondary metabolites has been hampered by a lack of genomic information, instead requiring difficult extraction and analysis techniques. Here, we report the genome sequence of *Iris pallida* Lam., generated with Pacific Bioscience long-read sequencing, resulting in a 10.04 Gbp assembly with a scaffold N50 of 14.34 Mbp and 91.8% complete BUSCOs. This reference genome will allow researchers to study the biosynthesis of these secondary metabolites in much greater detail, opening new avenues of investigation for drug discovery and fragrance formulations.

Research area: Genetics and Genomics; Botany; Plant Genetics

Background and Context:

The family *Iridaceae* comprises at least 250 known species and many hybrid cultivars. Traditionally grown for their ornamental value, irises are known to possess bioactive compounds in their tissues, particularly the rhizomes. Researchers have isolated PTP1B inhibitors from *Iris sanguinea* Hornem. ex Donn[1], antioxidant isoflavonoids[2] and the cytotoxic triterpenoid Belamchinenin A[3] from *Iris domestica*, the anti-*Helicobacter pylori* isoflavonoid irigenin from *Iris confusa* Sealy[4], and anti-biofilm extracts from several iris species[5]. These compounds were extracted from their producer organisms, requiring several kilograms of plant material to yield compounds in milligram quantities[1,3]. The bioactive compounds found in iris species are secondary metabolites, produced through multi-enzyme biosynthetic cascades. Genetic engineering and synthetic biology allow researchers to reconstruct valuable biosynthetic pathways in other host organisms, providing the opportunity for large-scale fermentation and extraction, but one requires access

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to the genes in such pathways to reconstruct them for large-scale expression and production of the secondary metabolites.

The Dalmatian iris *Iris pallida* Lam. (Figure 1) is a member of the *Iridaceae* family that produces bioactive compounds. Its triterpenoid iridals have been shown to act as ligands for the RasGRP family of diacylglycerol/phorbol ester receptors[6], and the triterpenoid iripallidal has been shown to inhibit AKT/mTOR and STAT3 signaling pathways in glioblastoma[7]. The closely related *Iris* x *germanica* L. also produces triterpenoids with anti-proliferative activity[8]. Given the wealth of bioactive compounds produced by iris species, we sought to fully sequence the genome of *Iris pallida* Lam. Karyotype analysis of *Iris pallida* revealed a diploid organism with 12 unique chromosomes[9].

Two *Iris* genomes have been published: *Iris sibirica* L. and *Iris virginica* L.[10]. While the approach involved short read sequencing (Illumina 2 x 150 bp paired-end, followed by Spades assembly), longer read platforms such as PacBio and Bionano Genomics are better suited to bridge across repetitive sequences, which account for a substantial portion of eukaryotic genomes and are expected to be common in the case of *Iris* sp. Besides, the chloroplast genome from *Iris speculatrix* Hance was sequenced to understand the phylogeny of the species[11], a large-scale RNA-seq transcriptional profile was generated for *Iris japonica* to investigate winter dormancy patterns[12], and a transcriptomic profiling effort was undertaken in *Iris* x *germanica* L. to understand reblooming mechanisms[13]. The next closest relative to iris with a fully sequenced genome is *Asparagus officinalis* L.[14]. Using Pacific Biosciences (PacBio) long-read sequencing technology, we obtained a full genomic sequence of 10.04 Gb for *Iris pallida* Lam. from leaf tissue. Estimates of other Iris species' genome sizes range from 2-30 Gb[15]; this range is in line with the genome size of *I. pallida* Lam. in this study. From RNA extracted from rhizome and leaf tissues, we again used PacBio sequencing technology to obtain an RNA transcriptional profile of *Iris pallida* Lam.. The genome annotation was completed with PacBio transcripts, and all abundance numbers were obtained from PacBio data.

The genomic sequence and transcript information of *Iris pallida* Lam. will allow researchers to identify enzymes responsible for bioactive compounds, allowing a better understanding of the biosynthetic pathways that generate bioactive compounds in the plant. This genome sequence and transcript data will

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

also allow researchers to understand phylogenetic relationships between irises and other plant species and facilitate DNA and RNA sequencing efforts for other iris species.

Methods

Genome Assembly Sample and Sequencing

For genomic DNA extraction, four 50 mL Falcon centrifuge tubes were each filled with 10 mL of extraction buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl pH 8.0 and 0.2% beta-mercaptoethanol. The tubes were warmed to 60°C in a water bath. A pre-cooled mortar was filled with liquid nitrogen to a depth of 3 cm. Sterile sand was added to a depth of 0.5 cm. Five young iris leaves, approximately 20 cm in length, were cut from the plant and immediately cut into 2 cm lengths and submerged in liquid nitrogen. They were then ground into a fine powder, with additional liquid nitrogen carefully added as needed. Approximately 25% of the iris/sand powder was added to each warm extraction buffer tube and mixed by inversion 3-4 times. The mixture was incubated at 60°C for 30 minutes and mixed by inversion every 5-10 minutes. Then,10 mL of 25:24:1 phenol:chloroform:isoamyl alcohol was added and gently mixed by continuous inversion for 1 minute. The sample was centrifuged at 6000 g for 10 minutes to separate the phases. The upper phase was carefully removed and transferred to a 50 mL phase-lock gel tube (Eppendorf), and 10 mL of 24:1 chloroform:isoamyl alcohol was added and gently mixed by inversion for 1 minute. The sample was centrifuged at 6000 x g for 10 minutes, after which time the aqueous phase above the phase-lock wax was removed and transferred to a fresh tube. An equal volume of isopropanol was added to the tube, and the sample was mixed by inversion until a gelatinous mixture of nucleic acids formed. This nucleic acid mixture was removed with a glass rod and washed three times with ice-cold 70% ethanol. The nucleic acid was allowed to air dry and then dissolved overnight in TE buffer. The sample was assigned the Novartis tracking ID AS_SAM_17_03QT. Library preparation used continuous long read methods for genomic DNA sequencing for the Sequel 1 instrument, as per manufacturer's instructions (PacBio). For optical mapping, a frozen sample of the *Iris pallida* Lam. leaf (GSM-AAB282) was shipped on dry ice to Bionano Genomics (San Diego, CA, US), which then sequenced the leaf genome to generate a genomic optical map.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A summary of sequencing data for this study is listed in **Table 1**. A total of 236 SMRT cells were used to produce 824,109,057,251 bp of genomic sequence data. The average length and N50 values for the PacBio subreads were 7,422 and 17,373 bp, respectively.

Table 1: Summary of sequencing data generated in this study

Sequencing platform	Data type	Tissue used	Raw data (bp)
Pacific Biosciences	RNA sequencing	Leaf	45,352,970,323
Pacific Biosciences	RNA sequencing	Rhizome	74,815,065,674
Pacific Biosciences	DNA sequencing	Leaf	824,109,057,251

Transcriptome Samples and Sequencing

Iris tissue samples (leaf and rhizome) were ground in liquid nitrogen, and RNA was extracted using the RNeasy plant mini kit (Sigma Aldrich) with the Qiashredder procedure. The RNA samples were then treated with the Turbo DNA-Free kit (Thermo Fisher). Library preparation was carried out according to the procedures for Isoform Sequencing (Iso-Seq) using the Clontech SMARTer PCR DNA synthesis kit with BluePippin Size Selection system. The rhizome sample of *Iris pallida* Lam. was sequenced using 10 SMRT cells and generated 74,815,065,674 bp of subread sequences. The leaf sample was sequenced using 11 SMRT cells and generated 45,352,970,323 bp. Each of these datasets was processed using the PacBio Circular Consensus algorithm in version 5.1.0 of SMRT Link using the ccs2 pipeline named sa3_ds_ccs.

Genome Assembly and Annotation

Falcon_unzip[16] was used to assemble the PacBio long read dataset. We used the Conda channels defaults, bioconda, and conda-forge to install pb-assembly, pbmm2, and genomicconsensus, as of July 24, 2019. Version 1.4.2 of falcon-kit and version 1.3.3 of falcon-unzip modules were present in our execution of Falcon_unzip. Falcon_unzip produced two assembly files, the primary contigs and the haplotigs, which are contigs that represent variant genomic sequences that are similar but not identical to the primary contigs.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A second genomic DNA sample was shipped to Bionano Genomics to generate genomic optical maps. The resulting optical maps were used to scaffold the genome assembly using the HybridScaffolding pipeline in the Bionano Genomics Solve package, version 3.2.1_04122018. Then, the PacBio Arrow algorithm implemented in the sl_resequencing2 pipeline present in SMRT-Link-7.0.1 was used to polish the Bionano Genomics hybrid assembly using original PacBio reads. The total elapsed time was 309 hours (final output folder named arrow_iris_20191208). Telomeres were predicted using a Python script from Jana Sperschneider (https://github.com/JanaSperschneider/FindTelomeres).

Transcriptome Assembly and Annotation

The lima program from PacBio (version lima 1.6.1 (commit v1.6.1-1-g77bd658)) was run to identify all CCS sequences with the expected 3' and 5' sequences. Then, the isoseq3 program from PacBio (commit v0.4.0-121-g22a3096*)[17] was used with the cluster option to group the CCS sequences into transcripts. The isoseq3 polish option was used to improve transcript accuracy. Prior to the availability of Isoseq3 software, the PacBio RNA sequence data were analyzed using tools from Isoseq1 and Isoseq2 software distributions from PacBio, but the analysis was limited by the slow run time of the earlier algorithms. We also included any additional transcripts from these older algorithms in our genome annotation if they were improved matches to the genome sequence compared to Isoseq3 transcripts. Using the transcripts, the NCBI blastn algorithm v2.2.8 was used against the Bionano Genomics scaffolded genome to identify probable locations for the corresponding genomic DNA. Then, Exonerate v2.2.0[18] with the cdna2genome model was used to find the likely genomic location as well as putative exons and introns for each transcript. A maximum intron size of 30,000 was used initially for the search, with a minimum match percentage of 80%. For all transcripts where no genomic location was found, a second Exonerate run was attempted, with a maximum intron size of 2,000,000 bp. This two-stage process was used to reduce the total computer run time required to search for all transcripts.

Finally, all predicted exon and intron locations were loaded into a local relational database to facilitate the preparation of a genome submission to NCBI, which included coding region predictions based on the transcript RNA sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

When preparing the genome submission to NCBI, we aimed to identify likely gene products. We used the Blastx[19] algorithm to find reasonable matches of the translated transcriptome sequences against three publicly available plant proteomes: *Asparagus officinalis* L., *Oryza sativa* L. Japonica Group, and *Zea mays* L.. An E score threshold of 0.001 was used.

Because many transcripts aligned to overlapping regions of the genome, we only reported one transcript per region because we were unable to rank multiple aligning transcripts. We used the Exonerate alignments that were restricted to a maximum coverage range on the genome to 1 million base pairs, because a spot check of the very large alignments appeared to be artifactual and obscured smaller groups of alignments. For each region, we chose the transcript whose genomic sequence yielded the longest open reading frame based on the standard genetic code. We reported this coding sequence in the NCBI submission along with the product name from the Blastx search, and we included a note providing the E and bit score of the Blastx alignment along with the Refseq identifier of the plant protein.

Results and Discussion

Genome

Plant genomes are known to be challenging to generate accurate, relatively complete genome assemblies, due to their large size, heterozygosity, and high frequency of repeat sequences. For these reasons, PacBio long-read sequencing technology was used to generate the *Iris pallida* Lam. genome assembly. The total size of the PacBio genome assembly was 10.46 Gbp. To enhance the assembly, we used Bionano Genomics optical mapping, as optical mapping on top of long-read sequencing has proven to be beneficial to produce higher quality plant genome assemblies[20,21]. The total size of the genome assembly after Bionano Genomics scaffolding was 13.49 Gbp (**Table 2**). The additional size of the scaffolded genome is due to differing haplotigs in the phased assembly from Falcon Unzip.
Because primary contigs and haplotigs were included in the scaffolding process, many regions of this genome sequence are nearly duplicated. These near duplications are important to the future analysis of this genome because we cannot determine which allele of a heterozygous gene is functional.

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Statistics for the *Iris pallida* Lam. genome

Date	Total Length (bp)	N50 (bp)	Longest Contig (bp)	Number of contigs	Coverage	Comments
28-Sep-2019	10,460,090,820	583,967	4,430,189	38684	78.8	Falcon_unzip, primary contigs
28-Sep-2019	2,642,332,941	101,281	1,374,365	38684	N/A	Falcon_unzip, haplotigs
24-Dec-2019	13,489,134,452	14,342,615	85,218,729	45374	61.1	Bionano Genomics plus Arrow

BUSCO version 5.4.7[22,23] was run on the Bionano Genomics scaffolded assembly (**Table 3**) using Augustus gene modeling software and using the maize species parameters provided in Augustus. The lineage dataset was eukaryota_odb10, and the BUSCO mode was set to euk_genome_aug. The GC content of the genome is 41.2%. Compared to other plant genomes, the completeness of our assembly is reasonable with regard to genome/transcript alignment and BUSCO scores[24]. A small number of sequences were omitted from publication by NCBI due to short size or discovery of vector sequence contamination. Thus, the number of scaffolds for this genome reported by NCBI is slightly smaller than the number reported in Table 2.

Table 3: Genome completeness evaluated by BUSCO

Count	Percentage of Searched BUSCOs	Description
234	91.8	Complete BUSCOs (C)
40	15.7	Complete and single-copy BUSCOs (S)
194	76.1	Complete and duplicated BUSCOs (D)
2	0.8	Fragmented BUSCOs (F)
19	7.4	Missing BUSCOs (M)
255	100.0	Total BUSCO groups searched

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transcriptome

Rhizome and leaf tissue samples were processed for RNAseq data. Statistics are shown in **Table 4**. For the rhizome sample, 3,032,725 CCS sequences were produced after processing by lima, resulting in 133,484 high-quality transcripts and 6,959 low-quality transcripts. For the leaf sample, 1,910,385 CCS sequences were produced after lima processing, resulting in 91,528 high-quality transcripts and 5,156 low-quality transcripts. Both high- and low-quality transcripts were used in the annotations. The CCS reads after lima processing were deposited into the NCBI Short Read Archive. There were 96,680 transcripts reported for the leaf sample and 140,135 transcripts reported for the rhizome sample. A total of 63,944 transcript identified coding regions were identified.

Table 4: Transcriptome statistics for *Iris pallida*

PacBio subreads by tissue	Total transcripts	Average length (bp)	N50 (bp)
Leaf	96,680	1,472	1,654
Rhizome	140,135	1,641	1,759

Out of 236,815 transcripts determined by the PacBio Isoseq3 method, 212,672 were aligned to the genome using Nucleotide Blast, an alignment percentage of 89.8%. All transcripts that aligned using Blast were then realigned against the *Iris pallida* genome in the vicinity of the genome matched by Blast. The percentage for successful Exonerate alignments was 88.1%. The quality of alignment from transcriptome to the genome was very high. Exonerate computes the number of identical base matches for its alignments, as well as the number of mismatches. For all Exonerate alignments, the ratio of identical base matches to the sum of base matches and mismatches was 98.1%. Given that the individual plant used for RNA isolation was different than the individual plant used for genomic DNA isolation, this result represents an excellent agreement of nucleotide sequences.

After submitting the *Iris pallida* transcript sequences to NCBI, NCBI reported that 230 rhizome sample transcripts contained sequences from species other than *Iris pallida* Lam.. Most of these contaminant sequences were from fungal species, an expected result given that the rhizome tissue sample was

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

removed from soil. These sequences were removed from the final submitted transcriptome and were not used in the annotation of the *Iris pallida* genome that was submitted to NCBI.

Telomeres

In the Bionano Genomics scaffolded assembly, a total of 26 scaffolds had telomere sequences at their ends. These telomere data were included in the NCBI submission. We did not identify any contigs or scaffolds that had telomeres at both ends.

With 12 unique chromosomes[9], *Iris pallida* would be expected to have 24 unique concatenations of telomeres with chromosome end sequences. Given the inclusion of haplotigs into the genome assembly as well as its draft quality, the identification of 26 telomeres in our assembly is consistent with the observed chromosome number.

Data Validation and Quality Control

The genome quality was assessed first by BUSCO analysis. We found 91.8% complete BUSCOs, of which 15.7% were complete and single copy. Second, we assessed quality by mapping RNA transcripts to the genome assembly. Here, we found an 88.1% success rate for Exonerate alignments. Thus, the DNA and RNA data are in strong agreement, indicating high-quality sample collection, data generation, data processing, and data analysis for the *Iris pallida* Lam. genomic assembly.

The leaf and rhizome iris samples were collected from a private garden in Basel, Switzerland. Tissue samples were collected from the same plant specimen. Their utilization in our research is in full compliance with the Nagoya Protocol.

Reuse Potential

Plant genomes and transcriptomes are essential for understanding the secondary metabolite (also known as natural product) biosynthetic pathways that produce these valuable molecules. Natural products are often extracted from producer species without knowledge of their biosynthesis, so industrial-scale production of natural products is hampered by plant availability. Iris species produce natural products in

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

many compound classes[1–5] but until now, no iris genome has been sequenced using PacBio long-read sequencing. The long-read genome assembly and mapped transcriptome of *Iris pallida* Lam. will allow researchers to sequence parts of or complete genomes of other iris species, broadening our understanding of those natural products that are common to iris, and those that are species-specific and responsible for the unique aromas and biological properties of irises. Additionally, identification of iris genes and pathways might aid researchers who study the phylogenetic relationships of plant families.

Data Availability

The genome assembly is available at NCBI under accession JANAVB010000000. The BioProject identifier at NCBI is PRJNA813844. The BioSample accessions are available within the above BioProject description at NCBI. The Isoseq transcript data from the roots and leaves are available at NCBI in the Transcriptome Shotgun Assembly (TSA) Database under accessions, GKDS00000000 and GKDR00000000, respectively. The Circular Consensus reads that were used as input files to the Isoseq algorithm are available in the NCBI Short Read Archive (SRA). The SRA accession for the root transcript reads is SRR22228979 and the SRA accession for the leaf transcript reads is SRR22228019. In addition, all of the subreads from the genome sequencing are available in the SRA under the above BioProject identifier.

List of abbreviations

bp: base pairs

BUSCO: Benchmarking Universal Single Copy Orthologs

CCS: circular consensus sequencing

EDTA: Ethylenediaminetetraacetic Acid

Hornem.: Jens Wilken Hornemann

L.: Carl von Linné

Lam.: Jean-Baptiste de Lamarck

NCBI: National Center for Biotechnology Information

PacBio: Pacific Biosciences

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SMRT: single molecule real time

TE: Tris and EDTA

Competing interests

R.E.B is a paid consultant to the Novartis Institutes for BioMedical Research, Inc. All other authors declare no competing interests.

Funding

This work was funded by the Novartis Institutes for BioMedical Research, Inc.

http://www.crossref.org/fundingdata/

Authors' contributions

R.E.B., A.M.E.F., E.J.O., U.N., F.P., and J.W. contributed to the study design. A.M.E.F., E.J.O, M.A., and J.W. collected and processed iris tissue samples for nucleic acid extraction. D.B., M.A., S.D.B., and M.O. prepared the nucleic acid libraries and performed the sequencing. R.E.B. and E.J.O. analyzed genomic and transcriptomic data including assembly, scaffolding, and polishing the genome. R.E.B. processed, formatted, and submitted genomic and transcriptomic data to NCBI. R.E.B, A.M.E.F., and J.W. wrote and revised the manuscript.

Acknowledgements

The authors would like to acknowledge Kerstin Oelkers for assisting with the sequencing library preparation and Jasmin Hägele for assisting in the watering of the iris plants, Tim Schuhmann for providing analytical support, Brigitta Liechty for providing an *Iris pallida* plant specimen, Maulik Thaker and Horst Hemmerle for contributions to alternate genomic sequencing efforts.

References

1. Yang, J. L., Ha, T. K. Q., Lee, B. W., Kim, J. & Oh, W. K. PTP1B inhibitors from the seeds of Iris sanguinea and their insulin mimetic activities via AMPK and ACC phosphorylation. *Bioorg Med Chem Lett* **27**, 5076–5081 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Wozniak, D., Janda, B., Kapusta, I., Oleszek, W. & Matkowski, A. Antimutagenic and anti-oxidant activities of isoflavonoids from Belamcanda chinensis (L.) DC. *Mutat Res Genet Toxicol Environ Mutagen* **696**, 148–153 (2010).

3. Ni, G., Li, J. Y. & Yu, D. Q. Belamchinenin A, an unprecedented tricyclic-fused triterpenoid with cytotoxicity from Belamcanda chinensis. *Org Biomol Chem* **16**, 3754–3759 (2018).

4. Abdel-Baki, P. M., El-Sherei, M. M., Khaleel, A. E., Abdel-Aziz, M. M. & Okba, M. M. Irigenin, a novel lead from Iris confusa for management of Helicobacter pylori infection with selective COX-2 and HpIMPDH inhibitory potential. *Sci Rep* **12**, (2022).

5. Hoang, L. *et al.* Phytochemical composition and in vitro biological activity of iris spp. (iridaceae): A new source of bioactive constituents for the inhibition of oral bacterial biofilms. *Antibiotics* **9**, 1–24 (2020).

6. Shao, L. *et al.* Iridals are a novel class of ligands for phorbol ester receptors with modest selectivity for the RasGRP receptor subfamily. *J Med Chem* **44**, 3872–3880 (2001).

7. Koul, N., Sharma, V., Dixit, D., Ghosh, S. & Sen, E. *Bicyclic triterpenoid Iripallidal induces apoptosis and inhibits Akt/mTOR pathway in glioma cells*. http://www.biomedcentral.com/1471-2407/10/328 (2010).

8. Halpert, M. *et al.* Rac-dependent doubling of HeLa cell area and impairment of cell migration and cell cycle by compounds from Iris germanica. *Protoplasma* **248**, 785–797 (2011).

9. Mitra, J. Karyotype Analysis of Bearded Iris. *Botanical Gazette* **117**, 265–293 (1956).

10. Chin, K.-J. & Pirro, S. The Complete Genome Sequences of Iris sibirica and Iris virginica (Iridaceae, Asparagales) . *Biodiversity Genomes* (2023) doi:10.56179/001c.72791.

11. Siu, T. Y. *et al.* The complete chloroplast genome of Iris speculatrix Hance, a rare and endangered plant native to Hong Kong. *Mitochondrial DNA B Resour* **7**, 864–866 (2022).

12. Li, D. *et al.* Hybrid RNA Sequencing Strategy for the Dynamic Transcriptomes of Winter Dormancy in an Evergreen Herbaceous Perennial, Iris japonica. *Front Genet* **13**, (2022).

13. Fan, Z. *et al.* To bloom once or more times: The reblooming mechanisms of Iris germanica revealed by transcriptome profiling. *BMC Genomics* **21**, (2020).

14. Harkess, A. *et al.* The asparagus genome sheds light on the origin and evolution of a young y chromosome. *Nat Commun* **8**, (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

15. Kentner, E. K., Arnold, M. L. & Wessler, S. R. *Characterization of High-Copy-Number Retrotransposons From the Large Genomes of the Louisiana Iris Species and Their Use as Molecular Markers*. www.psc.edu/biomed/genedoc.

16. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050–1054 (2016).

17. Leung, S. K. *et al.* Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**, (2021).

18. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, (2005).

19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

20. *White Paper Series Generating Accurate and Contiguous De Novo Genome Assemblies Using Hybrid Scaffolding Bionano Optical Mapping Reveals True Long-Range Structure of the Genome while Reducing Sequencing Costs*. (2020).

21. Li, C., Lin, F., An, D., Wang, W. & Huang, R. Genome sequencing and assembly by long reads in plants. *Genes* vol. 9 Preprint at https://doi.org/10.3390/genes9010006 (2018).

22. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. v. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

23. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**, 543–548 (2018).

24. Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759–1768 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2023.08.29.555454; this version posted August 31, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: Representative specimen of *Iris pallida* (Photo credit: A.M.E.F.)