Genomes & islands & evolution: Oh my!

M. Renee Bellinger

U.S. Geological Survey, Pacific Island Ecosystems Research Center, P.O. Box 44, Hawai'i National Park, Hawai'i, USA, *mbellinger@usgs.gov*

DOI: http://dx.doi.org/10.53875/capitulum.03.1.04

ABSTRACT

A central question in evolutionary biology is how lineages quickly diversify to occupy different ecological niches, along with determining genomic factors that facilitate evolutionary change. Isolated, oceanic archipelagos are famous for adaptive radiations characterized by endemic, species-rich clades with substantial ecological variation, yet genome resources key to determining eco-evo processes are generally lacking. Here I present a comparison of the number of genome reference assemblies available (as of May 31, 2023) for three major eukaryotic lineages, briefly describe genome sequencing and benchmarking strategies, and highlight as a case study a genome assembly project for *Bidens hawaiensis* (Koʻokoʻolau, Asteraceae or Compositae; Coreopsidae), a member of a hexaploid Hawaiian plant adaptive radiation. The total number of plant genome references (1,394) was found to substantially lag the total number of genome references for animal (6,003) and fungi (4,400). Improvements to the quality of de novo assembled genomes are fueled by second- and third-generation long-read sequencing advancements, among other sequencing approaches. In conjunction, strategies to improve genome contiguity include optical maps, Hi-C chromatin capture, or trio binning. Continual improvements to genome sequencing and assembly algorithms have brought within reach telomere-to-telomere genome assemblies, albeit this level of sequencing has to date only been achieved in a few cases. With improvements in sequencing techniques and per-base pair costs that continue to trend downward, the number of high-quality genomes is anticipated to continue to increase, leading to the filling in of taxonomic gaps and sampling of groups of taxa from under sampled geographic areas. Increasing the number of plant genome resources available for the study of island endemism could help to shed light on genome-phenome relationships and genome characteristics that have produced the stunning biological diversity that we now observe across the globe.

Keywords: *Bidens*, Compositae, genome assembly, genome benchmarking, island endemism, next-generation sequencing platforms, repetitive content

INTRODUCTION

Plant genome assemblies are foundational to elucidating evolutionary histories, taxonomic boundaries, and genetic underpinnings to functional trait diversification and adaptive processes. The field of genomics is growing by leaps and bounds in concert with rapid advancements in sequencing technologies and computational power. In conjunction with those advances, the per-base cost for obtaining DNA sequences has plummeted and new tools are continually emerging to produce increasingly higher quality genomes, to the point that it is now possible to generate "telomere-to-telomere" (T2T) genome assemblies (McCartney et al., 2022).

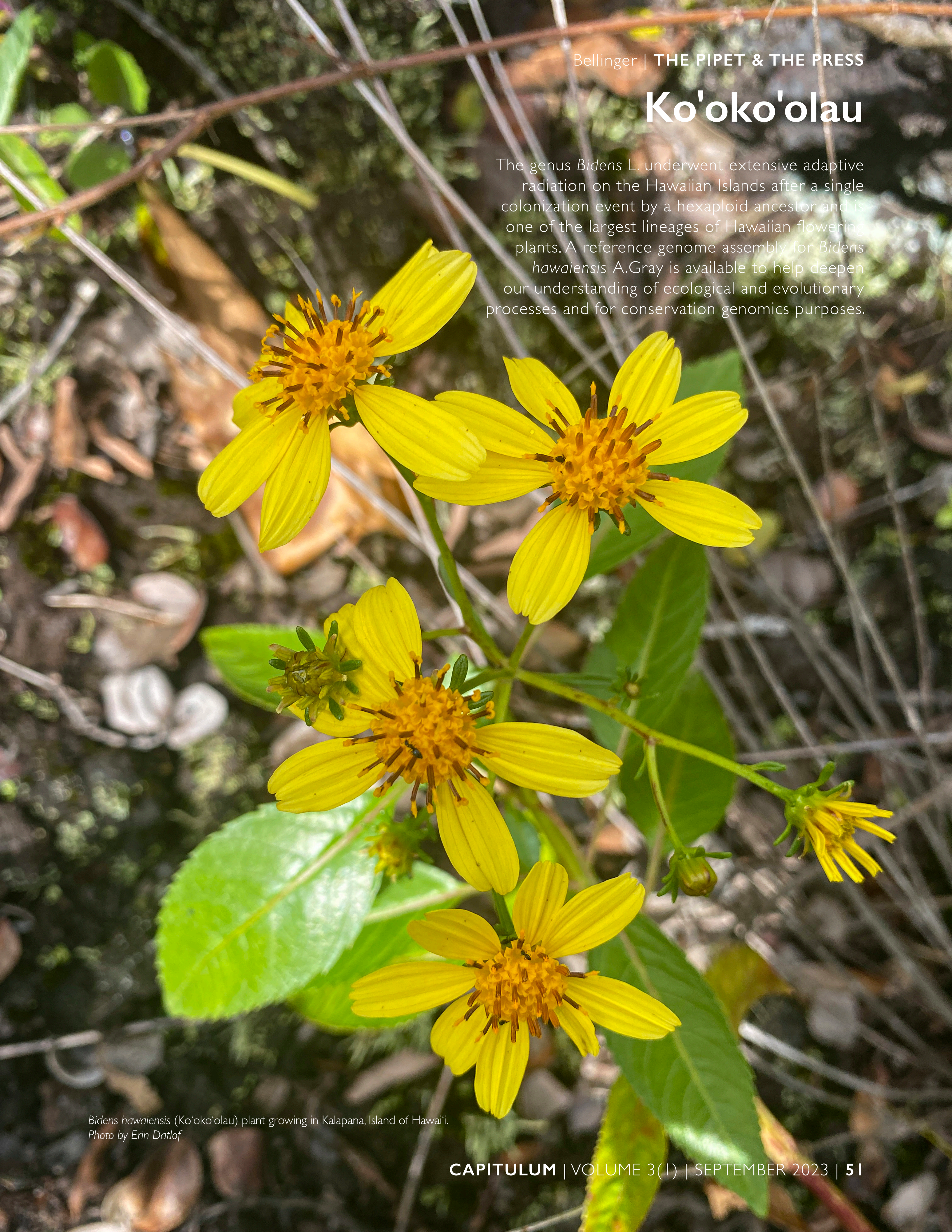
STATUS OF REFERENCE GENOME ASSEMBLIES: PLANTS, ANIMALS, AND FUNGI

Although advancements in sequencing technologies have led to the rapid accumulation of reference genome assemblies -- defined as the highest quality genome sequence available for a single species -- the current number of genome assemblies available for green plants (Kingdom Viriplantitae), at 1,394 genomes, lags behind the number of genomes sequenced for other major Opisthokonta Kingdoms Metazoa (animals) and Fungi, at 6,003

VERSION OF RECORD FIRST PUBLISHED ONLINE ON 4 SEPTEMBER 2023

Ko'oko'olau

The genus *Bidens* L. underwent extensive adaptive radiation on the Hawaiian Islands after a single colonization event by a hexaploid ancestor and is one of the largest lineages of Hawaiian flowering plants. A reference genome assembly for *Bidens hawaiensis* A.Gray is available to help deepen our understanding of ecological and evolutionary processes and for conservation genomics purposes.



Bidens hawaiensis (Koʻokoʻolau) plant growing in Kalapana, Island of Hawaiʻi.
Photo by Erin Datlof

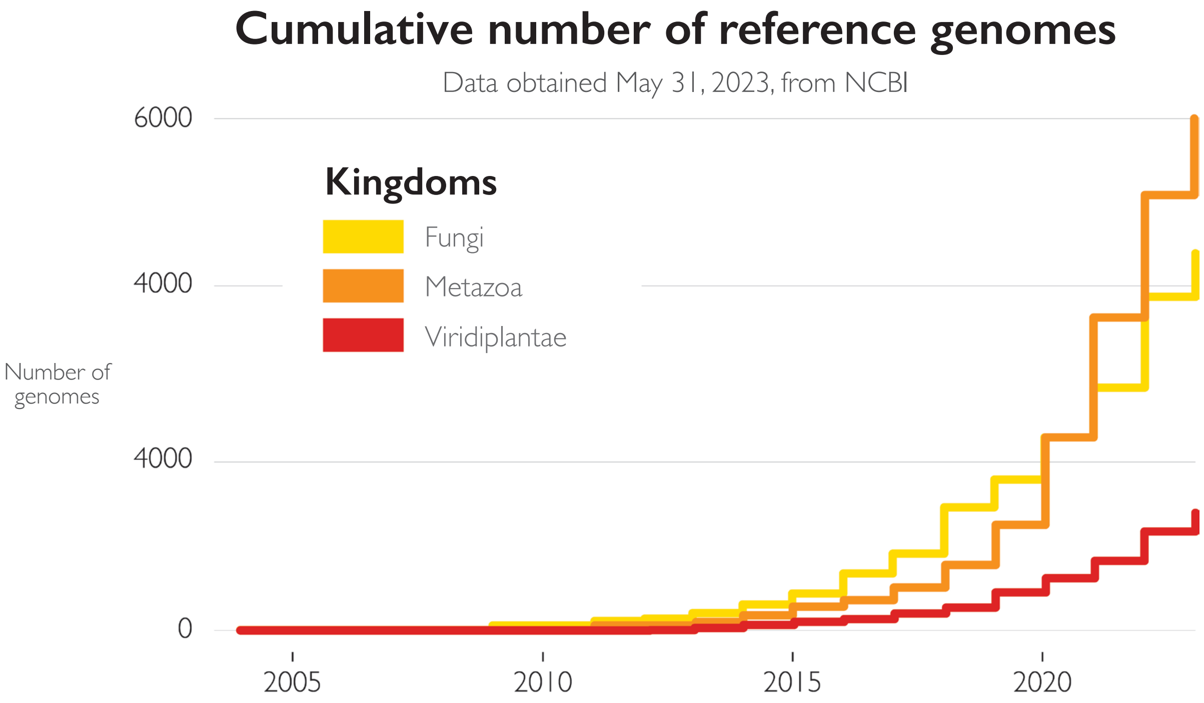


Figure 1. The cumulative number of reference genomes available by year for Kingdoms Metazoa (animals), Fungi, and Viriplantae (green plants). The reference genome is listed only once and is categorized by the most recent year of release.

and 4,400 genomes, respectively (Figure 2, download date May 31, 2023, data from the National Center for Biotechnology Information (NCBI). The gap between the cumulative number of plant and animal genomes has widened considerably over the past few years (Figure 1). Yet despite that gap, the proportion of green plants and animal genomes subjected to iterative genome improvements has remained similar between those two groups (Figure 2).

Particular to Asteraceae (or Compositae), the number of readily available reference genomes includes 43 species belonging to 29 genera (Figure 2; NCBI data download date May 31, 2023), which is a relatively small number considering the enormous size of this family, having 25,000+ named species and at least 1,700 genera (Mandel et al., 2019). The number of reference genomes available for Compositae has more than tripled since 2021, the point at which colleagues and I surveyed and benchmarked all publicly available (reasonably high-quality) Compositae genomes, n = 12, for comparison to a genome we assembled for ko'oko'olau, *Bidens hawaiensis* A. Gray (Bellinger et al., 2022), a single-island endemic and member of a Hawaiian adaptive radiation. With the relatively low number of Compositae genome assemblies available it is unsurprising that few genomes of endemic island taxa have been sequenced (but see Bellinger et al., 2022 and Cerca et al., 2022), consistent with the assertion by Cerca et al., (2023) that the application of genomic tools to understand the evolution of oceanic island organisms is still in its infancy.

TOWARDS PRODUCING GENOME REFERENCE ASSEMBLIES

While embarking on a genome sequencing project one might vet the suitability of available approaches by surveying which technology is in widest use and reasons why, and powers and pitfalls of particular DNA sequencing platforms. Although there is no one-size fits all approach for genome sequencing, Pacific Biosystems (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing platforms have recently been utilized to produce landmark T2T quality, gap-free genomes,

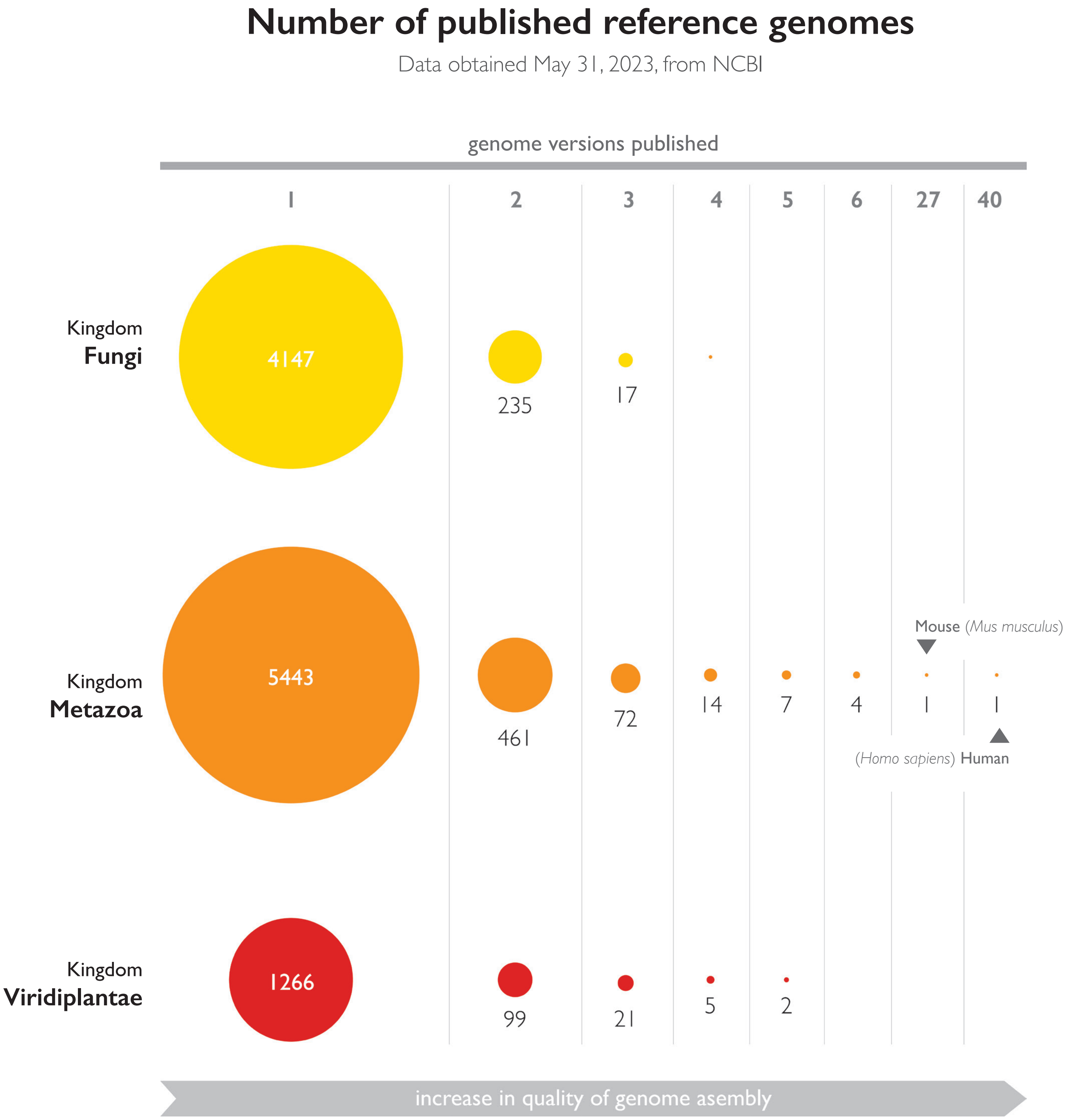


Figure 2. The number of published reference genomes available for Kingdom Viridiplantae, (green plants) and clade group Opisthokonta Cavalier-Smith 1987 Kingdom Metazoa (animals) and Kingdom Fungi. Data obtained May 31, 2023, from NCBI. The genome iteration column indicates how many genome versions have been published to date, which signals the level of polish and improvements to genome assembly quality. Also shown are the proportion of genomes by version.

especially when used in combination (McCartney et al., 2022, Rautiainen et al., 2023). The HiFi sequencing approach produces highly accurate reads (>99.5% accuracy), which overcomes a limitation inherent to PacBio noisy long-reads that are prone to high levels of sequencing error (up to 15%). The HiFi read accuracy is achieved by circularizing sheared DNA (length 15,000-20,000 bp), repeatedly sequencing the circularized DNA, and then performing a read-error correction step. Another notable advance in sequencing is the increase in ONT raw-read accuracy, having reached 99%, with an average long-read length of 100 Kb (Marx, 2023) on certain platforms and upper bound reported as 2.73 Mb (Payne et al., 2019). The use of long-reads in genome assembly can allow for reading through repetitive regions of chromosomes that would otherwise cause assembly fragmentation. Additionally, ONT and PacBio long-reads (but not HiFi reads) can provide information on methylation patterns that might be of interest to understanding epigenetic signals related to inheritance of functional traits.

Limitations to long-read sequencing protocols include the requirement of fresh-tissue for extraction of high-molecular weight (HMW) DNA, and some sequencing protocols require a relatively large amount of HMW DNA for library preparation. DNA sourced from dried specimens is not suitable for long-read sequencing because the drying and preservation process leads to DNA degradation and fragmentation (McAssey et al., 2023). With regards to DNA input amounts, the standard PacBio workflow requires at least 3 µg of DNA input per 1 Gb of genome length (PacBio, 2022). For some organisms, obtaining that amount of DNA from a single individual is not possible. However, advancements in workflows such as the low-DNA input HiFi library protocol enables generating long-read (~15 Kb) sequences from as little as 300 ng to 3 µg of DNA starting material, with a genome assembly size limit of 1 Gb per single molecule real time (SMRT) cell – noting that use of additional SMRT cells can scale to produce larger genome assemblies (PacBio, 2022). An alternative to long-read platforms is linked-read sequencing, typically performed on a short-read platform such as Illumina, which can be successfully applied to assemble long-reads from HMW DNA extracted from minimal tissue inputs and that performs relatively well even for highly heterozygous genomes (Helmkampf et al., 2019).

GENOME FEATURES AND ASSEMBLY BENCHMARKING

Several techniques can be utilized to evaluate genome features and benchmark the quality of a genome assembly. The assembled genome size can be compared to a haploid (or monoploid) size estimates from flow-cytometry (1C values) (e.g., Bellinger et al., 2022, Cerca et al., 2022) or through reference-free k-mer spectrum analysis (e.g., Ranallo-Benavidez et al., 2020). In simplest terms, kmer-spectrum analysis amounts to breaking DNA sequence data into short fragments (somewhere from 13 to 33 base pairs in length), tallying their frequencies, and modeling their complexity. The k-mer spectrum analysis can also be applied to estimate major genome characteristics such as heterozygosity and repeat content (Vurture et al., 2017), and can detect polyploid events, as was indicated for the hexaploid *B. hawaiensis* genome (Bellinger et al., 2022) using the polyploid-aware mixture model implemented in Genomescope v2 (Ranallo-Benavidez et al., 2020). Furthermore, k-mer spectrum analysis can be leveraged to identify subgenomes in cases where chromosome sequences are assigned to homeolog pairs, based on an approach developed by Cerca et al. (2022) for the tetraploid, critically endangered, Galápagos-endemic species *Scalesia atractyloides* Arnot. In that study, a hierarchical clustering algorithm grouped chromosomes into clusters (subgenomes) based on uneven representation of 'fossil transposable elements' that were actively replicating while the two subgenomes were separated, thus leading the authors to conclude the *Scalesia* Arn. ex Lindl. radiation is of allopolyploid origin. Another common approach for evaluating the quality of a genome assembly is to characterize the recovery of benchmarking universal single-copy orthologs (BUSCOs, Simão et al., 2015) through searches of genes contained within highly curated single-copy ortholog databases tailored to several major taxonomic lineages (OrthoDB, Kriventseva et al., 2019). Expectedly, genome assemblies reconstructed from HiFi and/or long reads tended to provide higher recovery of single-copy orthologs and have fewer missing or partial genes (Bellinger et al., 2022).

The gold-standard for producing a genome reference is to assemble an error-free, chromosome-level, gap free genome. Genome assembly contiguity is evaluated by the number and length of contiguous assembled sequences (contigs) or scaffolds, the latter being contigs ordered by their locations on chromosomes, even if not assigned to chromosomes, *per se*. The quality of an assembly, even those described as "chromosome resolved," thus requires consideration of contig lengths and the number of NNNN breaks that denote sections of unresolved DNA sequences. Assembly contiguity can be hindered by the quality of the DNA inputs, sequencing technique (long- versus short-reads), assembly strategy, and the repetitive content of the genome (Bennetzen et al., 2014). Regarding the latter, plant genomes can possess extremely high or low repeat content, even within the same family. For example, the haploid-resolved 6.8 Gb genome assembly for *Glebionis coronaria* (L.) Tzvelev (crown daisy) shows a transposable element content of ~93% (Wang et al., 2022). In contrast, *Erigeron canadensis* L. (horseweed), with a much smaller genome (~426 Mb), shows an extremely low repeat content, at 6.25% (Peng et al., 2014). Repetitive elements can cause assembly fragmentation, especially when reads do not traverse genomic intervals that span the entire length of the repeat, leading to an assembly break. On the other hand, the contiguity of a genome assembly can be improved by incorporating optical mapping or high-throughput chromosome conformation capture "Hi-C" or "Omni-C" information (e.g., Yuan et al., 2020, Zhang et al., 2019). These approaches utilize varying combinations of restriction enzymes (or for Omni-C a sequence-independent endonuclease) and short-read sequencing strategies to map DNA reads that are in 3-dimensional proximity based on chromatin packing, which enables joining contigs that would otherwise go unplaced on scaffolds or chromosomes. An alternative, or additional, strategy to improve genome contiguity is to use trio-binning to assign reads to parental genomes and phase the genome into maternal and paternal haplotypes (Cheng et al., 2021).

CASE STUDY

The koʻokoʻolau (*Bidens hawaiensis* A.Gray) reference genome assembly recently produced by myself and colleagues was motivated by a desire to create a genomic resource for this Hawaiian endemic adaptive radiation, which may serve as a model system for understanding eco-morphological diversification and the evolutionary genomics of explosive plant diversifications within insular systems (Bellinger et al., 2022). We assembled the *B. hawaiensis* large (estimated 7.4 Gb) and highly complex, hexaploid genome (base number of 12 chromosomes, 2n = 6x = 72, Ballard 1986) using HiFi sequences obtained from only two PacBio flow cells sequenced on a PacBio Sequel II and high molecular weight (HMW) DNA extracts. Those two cells produced 9.4 million raw sequences (850 Gb raw data), which yielded 3.83 million HiFi sequences having an average size of 15.1 kb and N50 length of 13.5 kb. With only HiFi reads and modest sequencing depth (~15x per monoploid genome), our assembly was comparatively contiguous relative to all other Compositae genomes published at the time, despite the plant's hexaploid status, large genome size, and high repeat content, at ~70%. Additionally, among the Compositae genomes we quality benchmarked for completeness, the BUSCO recoveries were >90% for 8 of the 12 genomes, with *B. hawaiensis* at 96.6%, second only to lettuce (*Lactuca sativa*, haploid genome size of 2.1 Gb), at 97.2% (refer to Bellinger et al., 2022 for details). Further improvements to the *B. hawaiensis* genome can be achieved by polishing with long-reads, optical mapping, or Hi-C/Omni-C scaffolding protocols (Zhang et al., 2019, Gladman et al., 2023).

CONCLUSIONS

Although the total number of publicly available genome assemblies has markedly increased over the past decade, the number of plant genome reference assemblies lags the number of genome assemblies available for other major eukaryotic lineages. Relatively few plant genomes are available to serve as references for the study of island endemics belonging to adaptively radiated clades. Although few in number, the availability of Compositae genomes has more than tripled in just two years, an increase perhaps fueled by decreasing costs of sequencing and the now routine use of third generation, long-read sequencing platforms that are capable of sequencing large, highly complex genomes. To help fill knowledge gaps, and for conservation purposes, colleagues and I recently published a *B. hawaiensis* genome assembly to contribute to the understanding of ecologically and evolutionarily driven morphological diversification within this highly polymorphic clade (Bellinger et al., 2022). This genome resource, along with others, can assist with determining the genetic basis of functional traits involved in eco-morphological diversification and processes that lead to high levels of island endemism.

METHODS

Genome statistics were obtained by extracting reference genome meta-data from the National Center for Biotechnology Information (NCBI) database using the Datasets and Dataformats command line tools v 15.1.0 (Sayers et al., 2021; download date: May 31, 2023). The taxonomic assignments of genomes followed the NCBI taxonomy database (Schoch et al., 2020).

DISCLAIMERS

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

ACKNOWLEDGMENTS

I thank two anonymous reviewers and the Capitulum editorial team for providing comments and suggestions for manuscript improvement, Erin Datlof for sharing photos of *Bidens hawaiensis*, Lauren Urie for assistance filling in Supplemental Table information, the NCBI Datasets and CGR Team for provision of accessible tools and curation of NCBI resources, and Matt Knope and Erin Datlof for stimulating conversations related to Hawaiian Bidens and plant evolution. The technical support and advanced computing resources from University of Hawaii Information Technology Services – Cyberinfrastructure, funded in part by the National Science Foundation MRI award # 1920304, are gratefully acknowledged.

LITERATURE CITED

Ballard, R. 1986. *Bidens pilosa* complex (Asteraceae) in North and Central America. *Am. J. Bot.* 73: 1452–1465.

Bellinger, M.R., Datlof, E.M., Selph, K.E., Gallaher, T.J. and Knope, M.L. 2022. A genome for *Bidens hawaiensis*: a member of a hexaploid Hawaiian plant adaptive radiation. *J. of Heredity*. 113: 205-214.

Bennetzen, J.L. & Wang, H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol. 65*: 505-530.

Cerca, J., Cotoras, D.D., Bieker, V.C., De-Kayne, R., Vargas, P., Fernández-Mazuecos, M., López-Delgado, J., White, O., Stervander, M., Geneva, A.J. & Andino, J.E.G. 2023. Evolutionary genomics of oceanic island radiations. *Trends Ecol. Evol.* 38: 631-642.

Cerca, J., Petersen, B., Lazaro-Guevara, J.M., Rivera-Colón, A., Birkeland, S., Vizueta, J., Li, S., Li, Q., Loureiro, J., Kosawang, C. and Díaz, P.J. 2022. The genomic basis of the plant island syndrome in Darwin's giant daisies. *Nature Comm*. 13: 3729.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18: 170-175.

Gladman, N., Goodwin, S., Chougule, K., McCombie, W.R. and Ware, D. 2023. Era of gapless plant genomes: Innovations in sequencing and mapping technologies revolutionize genomics and breeding. *Curr. Opin. Biotechnol.* 79: p102886.

Helmkampf, M., Bellinger, M.R., Geib, S.M., Sim, S.B. & Takabayashi, M. 2019. Draft genome of the rice coral *Montipora capitata* obtained from linked-read sequencing. *Genome Biol. Evol*. 11: 2045-2054.

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A. & Zdobnov, E.M. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47(D1): D807-D811.

Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E. & Funk, V.A. 2019. A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proc. Natl. Acad. Sci. U.S.A.* 116: 14083-14088.

Marx, V. 2023. Method of the year: long-read sequencing. *Nat. Methods 20*: 6-11.

McAssey, E.V., Downs, C., Yorkston, M., Morden, C. and Heyduk, K. 2023. A comparison of freezer-stored DNA and herbarium tissue samples for chloroplast assembly and genome skimming. *Appl. Plant Sci*. 11: e11527.

Mc Cartney, A.M., Shafin, K., Alonge, M., Bzikadze, A.V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G.A. & Miga, K.H. 2022. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Meth.* 19: 687-695.

PacBio. 2022. Considerations for using the low and ultra-low DNA input workflows for whole genome sequencing. Application note, url https://www.pacb.com/wp-content/uploads/Application-Note-Considerations-for-Using-the-Low-and-Ultra-Low-DNA-Input-Workflows-for-Whole-Genome-Sequencing.pdf

Peng, Y., Lai, Z., Lane, T., Nageswara-Rao, M., Okada, M., Jasieniuk, M., O'geen, H., Kim, R.W., Sammons, R.D., Rieseberg, L.H. and Stewart Jr, C.N. 2014. De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. *Plant Physiol*. 166: 1241-1254.

Payne, A., Holmes, N., Rakyan, V. and Loose, M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. 35: 2193-2198.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, & M.C. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Comm.* 11: 1432.

Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M. & Koren, S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* 1-9.

Sayers, E.W., Beck, J., Bolton, E.E., Bourexis, D., Brister, J.R., Canese, K., Comeau, D.C., Funk, K., Kim, S., Klimke, W. & Marchler-Bauer, A. 2021. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 49: D10.

Schoch C.L., Ciufo S., Domrachev M., Hotton C.L., Kannan S., Khovanskaya R., Leipe D., Mcveigh R., O'Neill K., Robbertse B., Sharma S., Soussov V., Sullivan J.P., Sun L., Turner S., Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). baaa062.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210-3212.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. and Schatz, M.C. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. *33*: 2202-2204.

Wang, S., Wang, A., Wang, H., Jiang, F., Xu, D. and Fan, W. 2022. Chromosome-level genome of a leaf vegetable *Glebionis coronaria* provides insights into the biosynthesis of monoterpenoids contributing to its special aroma. *DNA Research*. 29: p.dsac036.

Yuan, Y., Chung, C.Y.L. and Chan, T.F. 2020. Advances in optical mapping for genomic research. *Comput. Struct. Biotechnol. J.* 18: 2051-2062.

Zhang, X., Zhang, S., Zhao, Q., Ming, R. and Tang, H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* 5: 833-845.