Brief Communication

Towards comprehensive integration and curation of chloroplast genomes

Zhongyi Hua[1,†] (iD), Dongmei Tian[2,3,4,†], Chao Jiang[1,†], Shuhui Song[2,3,4,5,†], Ziyuan Chen[1], Yuyang Zhao[1], Yan Jin[1], Luqi Huang[1,*], Zhang Zhang[2,3,4,5,*] and Yuan Yuan[1,*] (iD)

[1]National Resource Center for Chinese Materia Medica, Chinese Academy of Chinese Medical Sciences (CACMS), Beijing, China

[2]China National Center for Bioinformation, Beijing, China

[3]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[4]University of Chinese Academy of Sciences, Beijing, China

[5]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

Dear Editor,

Chloroplasts are semi-autonomous genetic organelles that contain their own DNA. Since the first chloroplast genome was sequenced in 1986, chloroplast genomes have been extensively utilized as fundamental tools in plant phylogenetics and genetically modified to produce protein drugs, especially in the fight against COVID-19 (Daniell *et al.*, 2021, 2022). More chloroplast genome sequences could not only help us learn more about plant diversity and evolution (Daniell *et al.*, 2016), but they could also help chloroplast biotechnological applications by codon optimization and identifying non-conserved intergenic spacer regions and regulatory sequences that are needed for genetic engineering (Daniell *et al.*, 2021). Consequently, the chloroplast genomes of numerous plant species, particularly economically significant crops, have been continuously sequenced. Powered by high-throughput sequencing, over 7000 plant chloroplast genomes have been deposited in the National Center for Biotechnology Information (NCBI) organelle genome database, of which over 50% have been sequenced in the last 3 years (1082, 1175, and 1539 were sequenced in 2019, 2020, and 2021, respectively). With the accumulation of data, inaccurate taxonomic information (Locatelli *et al.*, 2020), disunity of genomic terms (Abeysooriya *et al.*, 2021), and other attendant problems have emerged, provoking significant challenges in employing chloroplast genomes. Many efforts have been made, but current databases still suffer from a lack of comprehensiveness and data curation, as well as incomplete data collection. Most existing curated databases are taxon-specific (e.g., cpGDB for spermatophytes (Singh *et al.*, 2020) and OGDA for algae (Liu *et al.*, 2020)) or limited to certain data types (e.g., ChloroMitoSSRDB for simple sequence repeats [SSRs; Sablok *et al.*, 2015]), which could be further improved by incorporating more comprehensive data. Additionally, each published organelle genome database covers only a fraction of chloroplasts, and thousands of chloroplast genomes are still dispersed in different nucleotide databases. Therefore, there is an urgent need to establish an integrated portal with a comprehensive collection and curation of chloroplast genomes.

Here, we developed the Chloroplast Genome Information Resource (CGIR), an integrated platform (https://ngdc.cncb.ac.cn/cgir) comprising 19 388 chloroplast genome assemblies and their corresponding meta-information (Figure 1a). The CGIR comprises five modules: (1) genomes, (2) genes, (3) SSRs, (4) barcodes, and (5) DNA signature sequences (DSSs; Figure 1b). The 'Genomes' module displayed 19 388 chloroplast assemblies from 11 946 different species (Figure 1c). Noticeably, among all assemblies, we sequenced 1170 assemblies from 718 species, of which the chloroplast genomes from 307 species were reported for the first time, including one family (Juncaginaceae) and 53 genera. In addition to boosting the number of sequenced species, newly added assemblies allow a group of species to have chloroplasts from many individuals. Compared to the NCBI Organelle Genome Database and CpGDB, with only one component for each species, multiple assemblies with explicit taxonomic information can provide more information for plant phylogeny. The taxonomic information of assemblies was curated in accordance with *The Catalogue of Life Checklist 2021* to eliminate disunity across different databases and contributors (Figure 1d). Functional information on plant species was integrated into the CGIR according to the *World Checklist of Useful Plant Species*. The 'Genes' module contains information on genes as well as their associated coding DNA sequence (CDS) and protein sequence (Figure 1e). To ensure a high-quality dataset, we first unified gene names by curating incorrect capitalization, spelling mistakes, extra characters in gene names, and synonymous gene names (Figure 1f). More importantly, not only was uniformity achieved, but corrections were also made. For example, the gene NADH–ubiquinone oxidoreductase chain 6 (*nad6*) should be encoded in the mitochondrial genome. However, this gene was observed in some chloroplast genome annotations, such as *Bulbophyllum reptans* (GenBank accession: NC_058531.1). By manual curation, we confirmed that *nad6* in NC_058531.1 was *ndhG* (Figure 1f).

Please cite this article as: Hua, Z., Tian, D., Jiang, C., Song, S., Chen, Z., Zhao, Y., Jin, Y., Huang, L., Zhang, Z. and Yuan, Y. (2022) Towards comprehensive integration and curation of chloroplast genomes. *Plant Biotechnol J.*, https://doi.org/10.1111/pbi.13923.
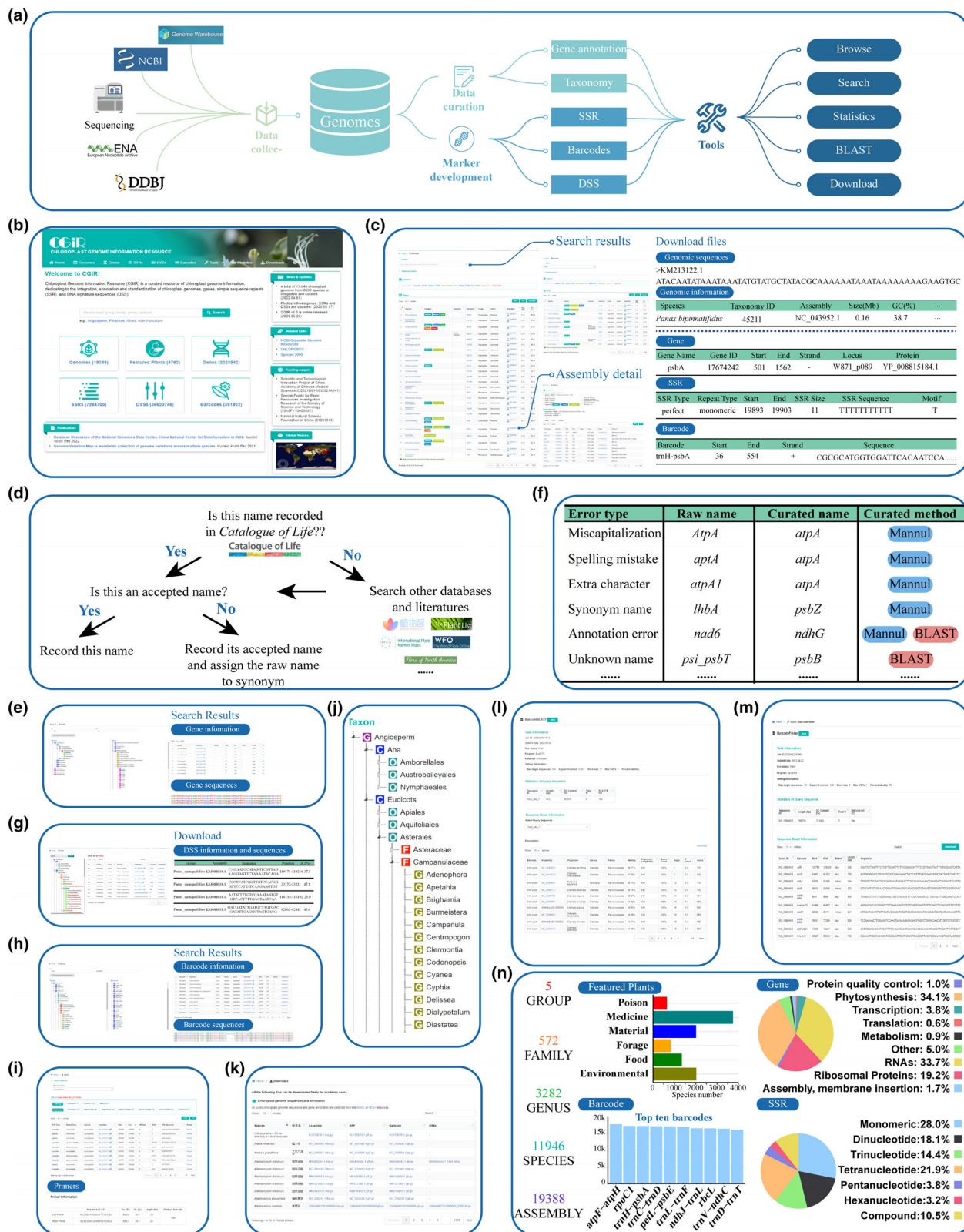
Figure 1 Architecture of the CGIR. (a) Design and construction of the CGIR, (b) the CGIR homepage, (c) the Genome module, (d) the curation model of taxonomic information, (e) the Gene module, (f) the curation model of gene annotation, (g) the Barcode module, (h) the DSS module, (i) the SSR module, (j) the Taxonomy tree view, (k) the Download module, (l) BarcodeBlast, (m) BarcodeFinder, and (n) the statistics of CGIR.

To better utilize these chloroplast genomes, the remaining three modules contained three commonly used DNA markers developed based on chloroplast genomes. The 'Barcodes' module contains DNA barcodes extracted from 29 different loci using the electronic PCR approach (Figure 1g), making the CGIR an excellent complement to traditional DNA barcode databases (e.g., Barcode of Life Data System [BOLD; Ratnasingham and Hebert, 2007]), which are mainly from *rbcL* and *matK* loci. The 'DSS' module contains the candidate DSSs from all species with more than one chloroplast assembly deposited in the CGIR (Figure 1h). DSS is a species-level marker that can be used as a complement to conventional DNA markers (Hua *et al*., 2022). The 'SSR' module comprises 7 284 705 SSRs and their associated primers (Figure 1i), far exceeding that of any other plastid SSR database.

In addition, the CGIR provides various methods for viewing, searching, and downloading data. To help users find the genome of a certain taxon, the 'Genomes' module allows users to search by species name, as well as class, order, genus, and family names. The synonyms are also listed in the search results, enabling researchers to determine whether to use these assemblies (Figure 1c). Because chloroplast data are always used in inter-species comparisons, the CGIR also provides a taxonomy tree view for users who are concerned with specific aspects of chloroplast data (e.g., *rbcL* gene, CDS sequences) in higher taxa (Figure 1j). Using this view, users can browse, search, and retrieve gene, barcode, and DSS data at any taxonomic level. A separate download module is also provided for easy data downloads (Figure 1k). Additionally, the 'BarcodeBLAST' tool allows users to search their barcode sequences against those deposited in the CGIR using BLAST (Figure 1l), and the 'BarcodeFinder' tool can help users to identify barcode regions in their uploaded chloroplast sequences (Figure 1m).

In general, the integration of high-throughput sequencing, public genomic resources, and careful manual curation guaranteed both the quantity and quality of chloroplast data in the CGIR, making it the largest comprehensive chloroplast repository available (Figure 1n). The CGIR will be a valuable resource for researchers working on phylogenetics and chloroplast genetic engineering. The curated taxonomy information and molecular markers are of tremendous value to plant phylogenetics; the labelled featured plants and corrected gene information will assist researchers in identifying suitable research objects and locating intergenic spacer regions, both of which are necessary for designing chloroplast engineering vectors (Daniell *et al*., 2021). In future, the CGIR will be continuously updated to incorporate more types of data.

Acknowledgements

Conflict of interests

The authors declare no conflict of interest.

Author contributions

Y.Y., S.S., Z.Z., and L.H. designed the project. Z.H. and C.J. wrote the manuscript. Z.C., Y.Z., C.J., and Y.J. collected the samples. Z.H., Z.C., and D.T. curated the data. D.T., S.S., and Z.Z. constructed the database. All authors read and approved the manuscript.

References

Abeysooriya, M., Soria, M., Kasu, M.S. and Ziemann, M. (2021) Gene name errors: lessons not learned. *PLoS Comput. Biol.* **17**, e1008984.

Daniell, H., Lin, C.-S., Yu, M. and Chang, W.-J. (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* **17**, 134.

Daniell, H., Jin, S., Zhu, X.-G., Gitzendanner, M.A., Soltis, D.E. and Soltis, P.S. (2021) Green giant—a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. *Plant Biotechnol. J.* **19**, 430–447.

Daniell, H., Nair, S.K., Guan, H., Guo, Y., Kulchar, R.J., Torres, M.D.T., Shahed-al-Mahmud, M. *et al*. (2022) Debulking different Corona (SARS-COV-2 delta, omicron, OC43) and influenza (H1N1, H3N2) virus strains by plant viral trap proteins in chewing gums to decrease infection and transmission. *Biomaterials*, **288**, 121671.

Hua, Z., Jiang, C., Song, S., Tian, D., Chen, Z., Jin, Y., Zhao, Y. *et al*. (2022) Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence. *Mol. Ecol. Resour.* **2022**, 1–12. https://doi.org/10.1111/1755-0998.13697

Liu, T., Cui, Y., Jia, X., Zhang, J., Li, R., Yu, Y., Jia, S. *et al*. (2020) OGDA: a comprehensive organelle genome database for algae. *Database*, **2020**, 10.

Locatelli, N.S., McIntyre, P.B., Therkildsen, N.O. and Baetscher, D.S. (2020) GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proc. Natl. Acad. Sci. USA*, **117**, 32211–32212.

Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). *Mol. Ecol. Notes*, **7**, 355–364.

Sablok, G., Padma Raju, G.V., Mudunuri, S.B., Prabha, R., Singh, D.P., Baev, V. *et al*. (2015) ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection. *Database*, **2015**, bav084.

Singh, B.P., Kumar, A., Kaur, H., Singh, H. and Nagpal, A.K. (2020) CpGDB: a comprehensive database of chloroplast genomes. *Bioinformation*, **16**, 171–175.