frontiers | Frontiers in Plant Science

TYPE Mini Review
PUBLISHED 17 September 2024
DOI 10.3389/fpls.2024.1474658

OPEN ACCESS

EDITED BY
Kai-Hua Jia,
Shandong Academy of Agricultural Sciences, China

REVIEWED BY
Lei Liang,
Chinese Academy of Agricultural Sciences, China

*CORRESPONDENCE
Bingguang Xiao
xiaobgsubmission@126.com
Enhui Shen
enhuishen@zju.edu.cn

RECEIVED 02 August 2024
ACCEPTED 30 August 2024
PUBLISHED 17 September 2024

CITATION
Tong Z, Huang Y, Zhu Q-H, Fan L, Xiao B and Shen E (2024) Retrospect and prospect of *Nicotiana tabacum* genome sequencing. *Front. Plant Sci.* 15:1474658. doi: 10.3389/fpls.2024.1474658

Retrospect and prospect of *Nicotiana tabacum* genome sequencing

Zhijun Tong[1], Yujie Huang[2], Qian-Hao Zhu[3], Longjiang Fan[2], Bingguang Xiao[1]* and Enhui Shen[2,4]*

[1]Key Laboratory of Tobacco Biotechnological Breeding, Yunnan Academy of Tobacco Agricultural Sciences, Kunming, China, [2]Institute of Crop Sciences, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China, [3]Black Mountain Laboratories, Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture and Food, Canberra, ACT, Australia, [4]The Rural Development Academy, Zhejiang University, Hangzhou, China

Investigating plant genomes offers crucial foundational resources for exploring various aspects of plant biology and applications, such as functional genomics and breeding practices. With the development in sequencing and assembly technology, several *Nicotiana tabacum* genomes have been published. In this paper, we reviewed the progress on *N. tabacum* genome assembly and quality, from the initial draft genomes to the recent high-quality chromosome-level assemblies. The application of long-read sequencing, optical mapping, and Hi-C technologies has significantly improved the contiguity and completeness of *N. tabacum* genome assemblies, with the latest assemblies having a contig N50 size over 50 Mb. Despite these advancements, further improvements are still required and possible, particularly on the development of pan-genome and telomere-to-telomere (T2T) genomes. These new genomes will capture the genomic diversity and variations among different *N. tabacum* cultivars and species, and provide a comprehensive view of the *N. tabacum* genome structure and gene content, so to deepen our understanding of the *N. tabacum* genome and facilitate precise breeding and functional genomics.

KEYWORDS

***N. tabacum*, genome sequencing, pan-genome, telomere-to-telomere genome, retrospect and prospect**

The genus *Nicotiana*, one of the six largest genera in the family Solanaceae, contains more than 80 species, including 49 distributed to America and 25 to Australia (Berbeć and Doroszewska, 2020; Chase et al., 2003; Knapp et al., 2004). Among them, common tobacco (*Nicotiana tabacum* L.) is acknowledged as one of the most crucial non-food crops globally. The significant economic and agricultural impact of *N. tabacum* is evident through its cultivation across vast areas, with its primary producers being China, Brazil, India, and the USA (Audrine, 2020; Battey et al., 2020). In addition to its economic value, *N. tabacum* has also become a model organism for the studies of plant biology and genetics due to its relatively short growth cycle, biochemical complexity, and ease of genetic manipulation (Gebhardt, 2016). Thereby, deciphering *N. tabacum* genome would offer crucial foundational resources for functional genomics studies and molecular breeding of tobacco itself and for facilitating functional genomics of other plants with *N. tabacum* as a model species.

N. tabacum is an allopolyploid (2n=4x=48) species and is evolved from the interspecific hybridization event between *N. sylvestris* (S-genome; 2n=2x=24) and *N. tomentosiformis* (T-genome; 2n=2x=24) occurred about 200,000 years ago (Leitch et al., 2008). Assembling a high quality *N. tabacum* genome sequence is challenging due to the high proportion (>70%) of repeat sequences and the closely related homologous sequences derived from its two progenitor species (Renny-Byfield et al., 2011). Owing to the rapid advancements in sequencing technology and the refinement of assembly algorithms in the last two decades (Xie et al., 2024), several *N. tabacum* genomes have been published since 2013. These genomes have significantly facilitated comprehensive genetic studies of *N. tabacum*, enabling researchers to have a better understanding of the complexity of the *N. tabacum* genome and its implications for agriculture and biotechnology.

Herein, we reviewed the assemblies and quality of the published genomes of *N. tabacum* and its two progenitors and proposed the strategies for further improvement and utilization of *N. tabacum* genome (Figure 1). Although the first plant genome, i.e., that of *Arabidopsis thaliana*, was published in 2000 (The Arabidopsis Genome Initiative, 2000), no common tobacco genome was available until 2013 when the draft genome sequences of two tobacco-related progenitor species were published (Sierro et al., 2013). Those two assemblies were generated using Illumina short reads, covering 83.3% (*N. sylvestris*) and 71.7% (*N. tomentosiformis*) of their estimated genome sizes (2.68 Gb and 2.36 Gb, respectively). Both assemblies have an N50 size of approximately 80 kilobases (kb) (Figure 1A). The availability of these two genome assemblies boosted assembling of the allopolyploid *N. tabacum* genome, because the same group published the first draft genomes of three *N. tabacum* cultivars (K326, TN90, and BX) in 2014 (Sierro et al., 2014). Compared to its progenitors, these three *N. tabacum* assemblies had a significantly improved N50 size (345 kb, 351 kb, and 386 kb, respectively) although the genome coverage was still around 82% (Figure 1A). But the quality of this version of the *N. tabacum* genome was still far behind that of other plant species generated at the same period of time, likely due to the complexity of the *N. tabacum* genome. By combining with BioNano optical mapping, the first chromosome-level genome of *N. tabacum* was published in 2017 (Edwards et al., 2017). However, only 64% of the genome assembly could be anchored to chromosomal locations and the contig N50 size was 335 kb that still needed to be improved. Assembling these genomes have mainly relied on the next generation sequencing (NGS) technology, the genomes contained many gaps which could not be filled by the short reads alone produced by NGS. The shortcoming of the short reads can be overcome by the long reads and ultra-long reads, ranging from 200 kb to potentially unlimited lengths, generated by the third-generation sequencing (TGS) technology (including PacBio and Nanopore) (Van Dijk et al., 2023). TGS together with other innovations, such as high-throughput chromosome conformation capture (Hi-C) provided platforms and tools for generation of high-quality and gap-less genomes. As a result, several high-quality genomes of *N. tabacum* and its two progenitors have been assembled in the last two years (Figure 1A) (Sierro et al., 2024; Wang et al., 2024; Zan et al., 2023). Compared to the previous draft genomes of *N. tabacum* cv. K326, the new K326 assembly had a contig N50 size of ~11.8 megabases (Mb), a significant increase from previous ~350 kb (Edwards et al., 2017; Sierro et al., 2024). Meanwhile, the contig N50 size of the two progenitors of *N. tabacum* also reached 15.0 Mb (*N. sylvestris*) and 10.6 Mb (*N. tomentosiformis*) (Sierro et al., 2024). In addition, the genomes of two more *N. tabacum* cultivars, 'ZY300' used for producing flue-cured tobacco in China and 'SR1' typically used for producing cigars, have also been recently published for the first time (Wang et al., 2024). With the application of high-fidelity (HIFI) reads generated by the PacBio circular consensus sequencing (CCS) method, the contig N50 size of the cultivar 'SR1' reached 56.1 Mb (Wang et al., 2024). By comparing the quality and the technologies used in assembling of the *N. tabacum* genomes reported in 2023 and 2024, it is obvious that the genomes assembled with longer read lengths have a higher level of completeness, for instance, 97.6% of the total assembly of K326 could be anchored to chromosomes (Sierro et al., 2024), and the genomes assembled with CCS have a longer contig N50 size and a higher accuracy (Figure 1A). Generation of the high-quality *N. tabacum* genomes would greatly expand opportunities in both breeding and functional genomics research of the crop.

Despite the tremendous progress on sequencing and assembling *N. tabacum* genomes, there is still much to be done in order to fully decipher *N. tabacum* genome. Learning the progress and experience on genome assembling in other plant species, we propose two broad directions for the further development of *N. tabacum* genome (Figure 1B). The first is to build a tobacco pan-genome. The pan-genome of a species represents the set of all DNA sequence diversity within the species. Pan-genome studies in other plant species (e.g., *Arabidopsis*, rice, maize, barley, wheat, cotton, tomato, and potato) have revealed high genomic variability and diversity among different individuals and demonstrated the great capability of using pan-genome in evolutionary and functional genomics studies (Sherman and Salzberg, 2020; Shi et al., 2023). Several distinct types of *N. tabacum*, including flue-cured, burley, oriental, and cigar types, were domesticated and have been systematically improved through extensive breeding programs (Lu et al., 2013). Constructing a pan-genome of these types of tobacco cultivars would discover their genomic variations and provide a better reference for identifying the genetic components and their

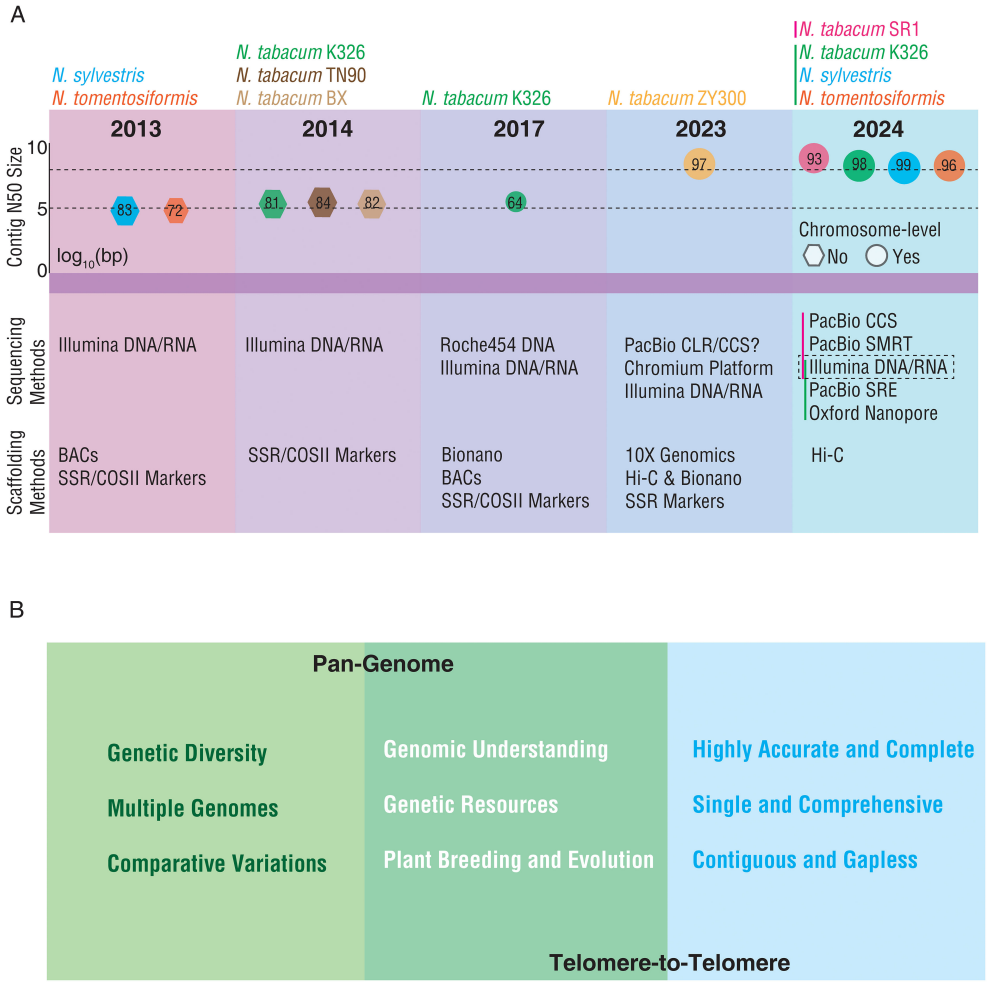

FIGURE 1
Retrospect and prospect of *N. tabacum* genome sequencing. **(A)** The information of the published genomes of *N. tabacum* and its two progenitors. The upper panel presents the $\log_{10}$ (contig N50 size) of the published *N.* tabacum and its two progenitors' genomes. The lower panel denotes the sequencing and scaffolding methods applied in assembling of each genome. The size of hexagons denotes the proportion of the estimated genome size covered by the assembly. The size of circles denotes the proportion of the assembly anchored to chromosomes. The red and green vertical bar denote the approaches adopted by the two studies (Wang et al., 2024 and Sierro et al., 2024). **(B)** The directions for the future development of *N.* tabacum genome. The left panel presents the scope and focus for pan-genome research. The right panel presents the characteristics of T2T research. The middle panel presents the expected outcomes of pan-genome and T2T genome.

associated molecular mechanisms underlying critical agronomic traits, such as yield, disease resistance, flavor profile, and nicotine content, to guide the breeding practices of these traits. Besides, wild relatives of tobacco would expand genetic diversity and confer a plenty of genes to survive in tough conditions. Thus, incorporating wild resources into *N. tabacum* materials has become one of modern breeding strategies (Xu et al., 2017). Several wild species in the genus *Nicotiana* have also published genomes, which can be taken into consideration in the construction of pangenome for *N. tabacum*. The most conspicuous one is *N. benthamiana*, a model organism in plant research, and four groups have published five versions of genomes with the contig N50 ranged from 89 kb to 54 Mb (Ko et al., 2024; Kurotani et al., 2023; Ranawaka et al., 2023; Wu et al., 2023). Most of the remaining ones including *N. attenuate*, *N. knightiana*, *N. longiflora*, *N. obtusifolia*, *N. otophora*, *N. paniculate*, *N. rustica* and *N. undulata* were not at chromosome-level (Supplementary Figure S1) (Sierro et al., 2014, 2018; Xu et al., 2017). Therefore, more efforts are needed to improve the quality of related genomes in the future. The second is to generate a telomere-to-telomere genome (T2T) *N. tabacum* genome, meaning a gapless and highly accurate assembly of entire *N. tabacum* chromosomes (Navrátilová et al., 2022; Nurk et al., 2022). The T2T genome is essential to identify the genetic make-ups of important agronomic traits, particularly the components in the dark matter regions, so to have a comprehensively understanding of the biological processes associated with the traits of interest and to finally promote precise breeding (Deng et al., 2022; Li and Durbin, 2024). Achieving these two goals will enable us to deepen understanding of the *N. tabacum* genome and of the genetic and molecular basis contributing to the distinct features observed in different types of *N. tabacum* resources, and finally to promote studies on the evolution of the species and custom designed breeding.

Author contributions

ZT: Data curation, Writing – original draft. YH: Investigation, Writing – original draft. Q-HZ: Writing – review & editing. LF: Writing – review & editing. BX: Funding acquisition, Writing – review & editing. ES: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper was funded with support from the National Natural Science Foundation of China (grant number 31860411), the China National Tobacco Company (110202101002(JY-02), 110202101038 (JY-15)) and the Yunnan Tobacco Company (2020530000241009, 2021530000241013, 2022530000241003, 2024530000241001).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1474658/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The timescale tree of published genomes in the genus *Nicotiana*. The green hexagons and red circles denote the assembly level of related species. The tree was constructed by the online platform TIMETREE5 (Hedges et al., 2015).

References

Audrine, P. (2020). *A policy perspective on tobacco farming and public health in Indonesia*. Jakarta, Indonesia: Center for Indonesian Policy Studies.

Battey, J. N. D., Sierro, N., and Ivanov, N. V. (2020). "Characterizing the genome of nicotiana tabacum," in *The tobacco plant genome*. Eds. N. V. Ivanov, N. Sierro and M. C. Peitsch (Neuchâtel, Switzerland: Springer Cham), 51–57.

Berbeć, A., and Doroszewska, T. (2020). "The use of nicotiana species in tobacco improvement," in *The tobacco plant genome*. Eds. N. V. Ivanov, N. Sierro and M. C. Peitsch (Neuchâtel, Switzerland: Springer Cham), 101–146.

Chase, M. W., Knapp, S., Cox, A. V., Clarkson, J. J., Butsko, Y., Joseph, J., et al. (2003). Molecular systematics, GISH and the origin of hybrid taxa in Nicotiana (Solanaceae). *Ann. Bot.* 92, 107–127. doi: 10.1093/aob/mcg087

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., et al. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* 15, 1268–1284. doi: 10.1016/j.molp.2022.06.010

Edwards, K. D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A. D., Bombarely, A., et al. (2017). A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* 18, 448. doi: 10.1186/s12864-017-3791-6

Gebhardt, C. (2016). The historical role of species from the Solanaceae plant family in genetic research. *Theor. Appl. Genet.* 129, 2281–2294. doi: 10.1007/s00122-016-2804-1

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification article fast track. *Mol. Biol. Evol.* 32, 835–845. doi: 10.1093/molbev/msv037

Knapp, S., Chase, M. W., and Clarkson, J. J. (2004). Nomenclatural changes and a new sectional classification in Nicotiana (Solanaceae). *Taxon* 53, 73–82. doi: 10.2307/4135490

Ko, S., Lee, S., Koo, H., Seo, H., and Yu, J. (2024). High-quality chromosome-level genome assembly of Nicotiana benthamiana. *Sci. Data* 11, 386. doi: 10.1038/s41597-024-03232-0

Kurotani, K., Hirakawa, H., Shirasawa, K., Tanizawa, Y., Nakamura, Y., Isobe, S., et al. (2023). Genome sequence and analysis of nicotiana benthamiana, the model plant for interactions between organisms. *Plant Cell Physiol.* 64, 248–257. doi: 10.1093/pcp/pcac168

Leitch, I. J., Hanson, L., Lim, K. Y., Kovarik, A., Chase, M. W., Clarkson, J. J., et al. (2008). The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). *Ann. Bot.* 101, 805–814. doi: 10.1093/aob/mcm326

Li, H., and Durbin, R. (2024). Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* 25, 658–670 doi: 10.1038/s41576-024-00718-w

Lu, X., Gui, Y., Xiao, B., Li, Y., Tong, Z., Liu, Y., et al. (2013). Development of DArT markers for a linkage map of flue-cured tobacco. Chinese Sci. Bull. 58, 641–648. doi: 10.1007/s11434-012-5453-z

Navrátilová, P., Toegelová, H., Tulpová, Z., Kuo, Y. T., Stein, N., Doležel, J., et al. (2022). Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. *Plant Biotechnol. J.* 20, 1373–1386. doi: 10.1111/pbi.13816

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* (80-.) 376, 44–53. doi: 10.1126/science.abj6987

Ranawaka, B., An, J., Lorenc, M. T., Jung, H., Sulli, M., Aprea, G., et al. (2023). A multi-omic Nicotiana benthamiana resource for fundamental research and biotechnology. *Nat. Plants* 9, 1558–1571. doi: 10.1038/s41477-023-01489-8

Renny-Byfield, S., Chester, M., Kovaík, A., Le Comber, S. C., Grandbastien, M. A., Deloger, M., et al. (2011). Next generation sequencing reveals genome downsizing in allotetraploid nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.* 28, 2843–2854. doi: 10.1093/molbev/msr112

Sherman, R. M., and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nat. Rev. Genet.* 21, 243–254. doi: 10.1038/s41576-020-0210-7

Shi, J., Tian, Z., Lai, J., and Huang, X. (2023). Plant pan-genomics and its applications. *Mol. Plant* 16, 168–186. doi: 10.1016/j.molp.2022.12.009

Sierro, N., Auberson, M., Dulize, R., and Ivanov, N. V. (2024). Chromosome-level genome assemblies of Nicotiana tabacum, Nicotiana sylvestris, and Nicotiana tomentosiformis. *Sci. Data* 11, 135. doi: 10.1038/s41597-024-02965-2

Sierro, N., Battey, J. N. D., Bovet, L., Liedschulte, V., Ouadi, S., Thomas, J., et al. (2018). The impact of genome evolution on the allotetraploid Nicotiana rustica – an intriguing story of enhanced alkaloid production. *BMC Genomics* 19, 855. doi: 10.1186/s12864-018-5241-5

Sierro, N., Battey, J. N. D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., et al. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* 5, 3833. doi: 10.1038/ncomms4833

Sierro, N., Battey, J. N. D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., et al. (2013). Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biol.* 14, R60. doi: 10.1186/gb-2013-14-6-r60

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the fowering plant Arabidopsis thaliana. *Nature* 408, 796–815. doi: 10.1038/35048692

Van Dijk, E. L., Naquin, D., Gorrichon, K., Jaszczyszyn, Y., Ouazahrou, R., Thermes, C., et al. (2023). Genomics in the long-read sequencing era. *Trends Genet.* 39, 649–671. doi: 10.1016/j.tig.2023.04.006

Wang, J., Zhang, Q., Tung, J., Zhang, X., Liu, D., Deng, Y., et al. (2024). High-quality assembled and annotated genomes of Nicotiana tabacum and Nicotiana benthamiana reveal chromosome evolution and changes in defense arsenals. *Mol. Plant* 17, 423–437. doi: 10.1016/j.molp.2024.01.008

Wu, Y., Li, D., Hu, Y., Buckler, E. S., and Huang, S. (2023). Article Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* 186, 2313–2328. doi: 10.1016/j.cell.2023.04.008

Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., et al. (2024). Technology-enabled great leap in deciphering plant genomes. *Nat. Plants* 10, 551–566. doi: 10.1038/s41477-024-01655-6

Xu, S., Brockmöller, T., Navarro-quezada, A., Kuhl, H., Gase, K., and Ling, Z. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc. Natl. Acad. Sci.* 114, 6133–6138. doi: 10.1073/pnas.1700073114

Zan, Y., Chen, S., Ren, M., Liu, G., Liu, Y., Si, H., et al. (2023). The allotetraploid Nicotiana tabacum genome and GenBank genomics highlight the genomic features, genetic diversity and regulation of morphological, metabolic and disease-resistance traits. . *BioRxiv*. doi: 10.1101/2023.02.21.529366