

**BỘ GIÁO DỤC
VÀ ĐÀO TẠO**

**VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM**

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ



Phạm Mai Hương

**GIẢI TRÌNH TỰ VÀ NGHIÊN CỨU ĐẶC ĐIỂM HỆ GEN LỤC
LẠP CỦA CÂY XÀ CĂN BA VÌ (*Ophiorrhiza baviensis*) BẰNG
CÔNG NGHỆ GIẢI TRÌNH TỰ THỂ HỆ MỚI PACBIO SMRT**

LUẬN VĂN THẠC SĨ NGÀNH SINH HỌC

PHẠM MAI HƯƠNG

SINH HỌC THỰC NGHIỆM

2023

Hà Nội - 2023

**BỘ GIÁO DỤC
VÀ ĐÀO TẠO**

**VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM**

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ



Phạm Mai Hương

**GIẢI TRÌNH TỰ VÀ NGHIÊN CỨU ĐẶC ĐIỂM HỆ GEN LỤC LẠP
CỦA CÂY XÀ CĂN BA VÌ (*Ophiorrhiza baviensis*) BẰNG CÔNG NGHỆ
GIẢI TRÌNH TỰ THỂ HỆ MỚI PACBIO SMRT**

Chuyên ngành: Sinh học thực nghiệm

Mã số: 8420114

LUẬN VĂN THẠC SĨ NGÀNH SINH HỌC

**NGƯỜI HƯỚNG DẪN KHOA HỌC :
GS.TS. CHU HOÀNG HÀ**

Hà Nội – 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận văn này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận văn là trung thực nếu sai tôi hoàn toàn chịu trách nhiệm trước pháp luật.

Tác giả

Phạm Mai Hương

LỜI CẢM ƠN

Lời đầu tiên, tôi xin chân thành cảm ơn thầy hướng dẫn, GS.TS. Chu Hoàng Hà, đã tận tình hướng dẫn, chỉ bảo và luôn có sự phản hồi tỉ mỉ trong thời gian nhanh nhất trong suốt thời gian qua, nhằm giúp tôi có thể hoàn thành luận văn này.

Tôi xin cảm ơn lãnh đạo và các nhân viên tại Phòng thí nghiệm trọng điểm Công nghệ Gen và Trung tâm Giám định ADN, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, đã giúp đỡ tôi có thêm nhiều kiến thức và kinh nghiệm trong mọi bước tiến hành luận văn.

Tôi cũng xin được cảm ơn Viện Hàn lâm Khoa học và Công nghệ Việt Nam và các thành viên trong đề tài “Giải trình tự và nghiên cứu đặc điểm hệ gen lục lạp của cây dược liệu thuộc loài Xà căn ba vì (*Ophiorrhiza baviensis*) bằng công nghệ giải trình tự thế hệ mới Pacbio SMRT sequencing, nhằm phân loại và bảo tồn nguồn gen”, với mã số đề tài: CSCL08.02/22-22, đã giúp đỡ tôi đạt được những kết quả trong luận văn này.

Bên cạnh đó, tôi xin gửi lời cảm ơn đến ban Lãnh đạo, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ để luận văn được hoàn thành.

Cuối cùng, tôi muốn gửi lời cảm ơn tới bố mẹ tôi, tới gia đình và bạn bè - những người đã hết sức ủng hộ, giúp đỡ và động viên tôi trong suốt quá trình học tập đã qua.

MỤC LỤC

MỞ ĐẦU.....	1
Chương 1. TỔNG QUAN NGHIÊN CỨU	3
1.1. Đặc điểm chung và phân bố của loài Xà căn ba vì.....	3
1.2. Tình hình nghiên cứu về cây Xà căn ba vì trên thế giới.....	5
1.3. Tình hình nghiên cứu về cây Xà căn ba vì trong nước.....	6
1.4. Định danh Xà căn ba vì bằng chỉ thị phân tử	9
1.5. Giải trình tự thể hệ mới và ứng dụng trong nghiên cứu bảo tồn nguồn gen và phân loại thực vật	11
1.5.1. Giải trình tự thể hệ mới	11
1.5.2. Ứng dụng của NGS trong nghiên cứu bảo tồn nguồn gen và phân loại thực vật	14
Chương 2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU	18
2.1. Đối tượng nghiên cứu.....	18
2.2. Phương pháp nghiên cứu.....	18
2.2.1. Tách chiết DNA tổng số của mẫu thực vật	18
2.2.2. Tạo thư viện và giải trình tự	19
2.2.3. Lắp ráp hệ gen lục lạp	19
2.2.4. Chú giải hệ gen lục lạp.....	20
2.2.5. So sánh hệ gen lục lạp và xây dựng cây phát sinh chủng loại	20
Chương 3. KẾT QUẢ VÀ THẢO LUẬN	22
3.1. Kết quả tách chiết và lưu trữ DNA tổng số của mẫu thực vật.....	22
3.2. Kết quả giải trình tự hệ gen lục lạp bằng công nghệ giải trình tự Pacbio ...	23
3.3. Kết quả lắp ráp hệ gen	25
3.4. Kết quả chú giải hệ gen lục lạp	26
3.5. Kết quả so sánh hệ gen lục lạp và xây dựng cây phát sinh chủng loại.....	33
3.5.1. Kết quả so sánh hệ gen lục lạp	33
3.5.2. Kết quả phân tích phát sinh loài	38
KẾT LUẬN VÀ KIẾN NGHỊ.....	41
Kết luận	41
Kiến nghị	41
DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ.....	42
DANH MỤC TÀI LIỆU THAM KHẢO	43
PHỤ LỤC.....	47

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ CÁI VIẾT TẮT

STT	Tên viết tắt	Tên đầy đủ
1	bp	Basepair
2	CCS	Circular consensus sequencing
3	CLR	Continuous long read
4	CNS	Conserved noncoding sequences
5	CPT	Camptothecin
6	DNA	Deoxyribonucleic acid
7	dNTP	Deoxyribonucleotide triphosphate
8	dsDNA	Double-stranded DNA
9	ETS	External transcribed spacer
10	HGAP	Hierarchical Genome Assembly Process
11	HR-ESI-MS	High-resolution electrospray ionisation mass spectra
12	IGS	Intergenic spacer
13	IR	Inverted repeat
14	ITS	Internal transcribed spacer
15	LPS	Lipopolysaccharide
16	LSC	Large single copy
17	NBCI	National Center for Biotechnology Information
18	NGS	Next generation sequencing
19	NMR	Nuclear magnetic resonance
20	NO	Nitric oxide
21	<i>O.</i>	<i>Ophiorrhiza</i>
22	PacBio	Pacific BioSciences
23	Pi	Nucleotide diversity
24	RNA	Ribonucleic Acid
25	RSCU	Relative synonymous codon usage
26	SGS	Sanger Sequencing
27	SMRT	Single-molecule real-time sequencing
28	sp.	Species

29	SSC	Small single copy
30	ssDNA	Single-stranded DNA
31	SSR	Microsatellite, simple sequence repeats
32	XCBV	Xà căn ba vì
33	ZMW	Zero-mode waveguide

DANH MỤC CÁC BẢNG

Bảng 1.1. Hoạt tính sinh học của các hợp chất khai thác từ cây Xà cã ba vì.	8
Bảng 3.1. Nồng độ DNA tổng số đo bằng nanodrop.	23
Bảng 3. 3. Tóm tắt thông tin lắp ráp và chú giải hệ gen lục lạp Xà cã ba vì.	27
Bảng 3.4. Thành phần gen của hệ gen lục lạp Xà cã ba vì.	28
Bảng 3.4. Tần suất sử dụng codon cho các gen mã hóa protein trên hệ gen lục lạp Xà cã ba vì.....	33

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Hình 1.1. Cây Xà căn ba vì (<i>Ophiorrhiza baviensis</i>).....	4
Hình 1.2. Công thức hóa học của các hợp chất từ cây Xà căn ba vì.....	7
Hình 2.1. (A) Môi trường sống của Xà căn ba vì; (B) Chùm quả của cây Xà căn ba vì.....	18
Hình 3.1. Ảnh điện di trên gel agarose 0.8%.....	22
Hình 3.2. Phân bố độ dài (A) và chất lượng (B) đoạn đọc.	25
Hình 3.3. Bản đồ hệ gen lục lạp loài Xà căn ba vì ở Việt Nam.....	26
Hình 3.4. Phân tích các lần lặp lại trình tự đơn của hệ gen lục lạp Xà căn ba vì.	31
Hình 3.5. Phân tích trình tự lặp lại dài trên quy mô bộ gen lục lạp của loài Xà căn ba vì.....	32
Hình 3.6. Biểu đồ nhận dạng so sánh bộ gen lục lạp của ba loài Xà căn.	36
Hình 3.7. Phân tích so sánh các giá trị đa dạng nucleotide giữa ba trình tự bộ gen lục lạp của các loài Xà căn.....	36
Hình 3.8. So sánh các vị trí tiếp giáp các vùng cấu trúc giữa ba bộ gen lục lạp.....	38
Hình 3.9. Cây phát sinh loài Maximum Likelihood dựa trên các trình tự gen <i>rps16</i> và vùng nối gen <i>trnL-trnF</i>	39

MỞ ĐẦU

Lục lạp là một bào quan thiết yếu trong tế bào thực vật hoặc vi sinh vật quang hợp, là nơi sản sinh ra năng lượng nuôi sống tế bào qua hoạt động quang hợp. Mỗi lục lạp có chứa các ribosome riêng và một hệ gen tách biệt với hệ gen nhân của tế bào với kích thước trong khoảng 20 - 120kb. Bởi vì kích thước hệ gen lục lạp nhỏ, đơn giản hơn so với hệ gen nhân, nên lục lạp thường được là đích giải trình tự đầu tiên. Trong khi đó, trình tự hệ gen lục lạp cũng được sử dụng rộng rãi trong phân tích tiến hóa, barcoding và meta-barcoding, lại chỉ chứa khoảng 100-120 gene mã hóa protein. Cho đến thời điểm hiện tại, trên cơ sở dữ liệu Trung tâm thông tin về công nghệ sinh học quốc gia NCBI Genbank có khoảng hơn 1000 hệ gen lục lạp của các loài thực vật. Tuy nhiên, con số này là rất nhỏ so với sự đa dạng thực vật hiện có trên hành tinh, từ đó đặt ra tiềm năng và sự cần thiết phải thu thập và lưu trữ trình tự của các loài này. Đối với loài dược liệu như Xà cẩu ba vì, thì tiềm năng khai thác và sự cần thiết phải phân loại một cách có hệ thống lại càng cần thiết. Thông tin về đặc điểm sinh thái và hệ gen của loài này vô cùng hạn chế, chỉ có 4 trình tự của Xà cẩu ba vì bao gồm gen *rps16* (#MH626923.1), vùng nối gen *trnL-trnF* (#MH626989.1), *ETS* (#MH626743.1) và *ITS* (#MH626804.1) trên cơ sở dữ liệu genbank của Trung tâm Thông tin Công nghệ sinh học Quốc gia (Hoa Kỳ) (NCBI). Mỗi trình tự chỉ có kích thước dưới 1000 bp, đều thuộc hệ gen lục lạp. Như vậy, có thể thấy sự cần thiết phải có một nghiên cứu trên toàn bộ hệ gen lục lạp của loài Xà cẩu ba vì cho công tác phân loại, đánh giá đa dạng và nghiên cứu đặc điểm hệ gen lục lạp, làm cơ sở cho công tác bảo tồn và nghiên cứu mở rộng về sau. Với kích thước ước tính của hệ gen lục lạp của các loài Xà cẩu là khoảng 154 kb, tiềm năng khai thác thông tin genome trên hệ gen lục lạp này là rất lớn, hứa hẹn cung cấp nhiều thông tin khoa học quan trọng.

Hiện nay, công nghệ giải trình tự PacBio cũng đã được ứng dụng để giải trình tự hệ gen lục lạp, và đã có nghiên cứu chứng minh cho khả năng vượt trội của PacBio khi lắp ráp *de novo* với độ chính xác 99%, và khi tăng độ lặp lại độ chính xác có thể lên đến trên 99,9%. Cho đến nay đã có rất nhiều công trình sử dụng công nghệ PacBio để giải trình tự hệ gen lục lạp, đặc biệt là các loài có tính ứng dụng cao như các loài dược liệu. Trong lĩnh vực nghiên cứu hệ gen, trong nước ta chưa có bất kỳ công bố nào liên quan đến khảo sát hệ gen nhân và hệ gen lục lạp của các loài thuộc chi Xà cẩu. Xuất phát từ tình hình thực tiễn và sự cần thiết của nghiên cứu,

chúng tôi tiến hành đề tài “*Giải trình tự và nghiên cứu đặc điểm hệ gen lục lạp của cây Xà cấn ba vì (Ophiorrhiza baviensis) bằng công nghệ giải trình tự thế hệ mới Pacbio SMRT*”

Chương 1. TỔNG QUAN NGHIÊN CỨU

1.1. ĐẶC ĐIỂM CHUNG VÀ PHÂN BỐ CỦA LOÀI XÀ CĂN BA VÌ

Xà căn là một nhóm các loài thuộc chi *Ophiorrhiza* là một chi thực vật lớn có hoa trong họ Thiến thảo (Rubiaceae), bao gồm khoảng 400 loài trên thế giới và 13 loài ở Việt Nam [1]. Các loài thực vật thuộc chi này là bộ cây thân thảo một năm hoặc lâu năm, một số ít khác lại là cây bụi phụ. Mặc dù chi thực vật này có tính đơn ngành rõ ràng dựa trên hình dạng quả nang, việc định danh ở cấp độ loài đôi khi rất khó khăn do sự biến đổi hình thái cao của chúng và hầu hết các loài rất khó phân biệt do thiếu kiến thức về hình dạng hoa của chúng [2–4]. Định danh sai hoặc nhầm lẫn với các dạng holotype trở thành vấn đề chính trong quá trình phân loại các loài thực vật thuộc chi này. Xà căn ba vì (XCBV) hay cây dẹt Ba Vì (danh pháp *Ophiorrhiza baviensis*) là một loài trong họ Thiến thảo [5]. Loài này được cho là trùng khớp với loài khác có danh pháp *Ophiorrhiza alatiflora*. Xà căn ba vì có đặc điểm là cây thân thảo hoặc cây bụi phụ, cao đến 50 cm, sống lâu năm, mọc thẳng hoặc leo bám; thân cây trơn nhẵn, nhánh mọc dày đặc dần lên phía trên. Cuống lá dài khoảng 0,5–2 cm, đôi khi dài đến 5 cm; lá có hình phiến giấy hoặc hình trứng thuôn dài; đỉnh có nhiều gai, có lông hình lược liềm ở trục dọc theo các gân lá; gân phụ từ 5–13 đôi. Cụm hoa tụ lại, nhiều hoa; cuống hoa dài khoảng 1–4 cm, có màu đỏ đậm hoặc màu đỏ tía. Đài hoa mọc đối xứng dày đặc. Tràng hoa màu trắng hồng, hình ống, mặt ngoài có lông tơ. Quả nang mitriform, 2,5–4 × 8–10 mm, sáng bóng. Cây ra hoa từ tháng 3 đến tháng 5; ra quả vào khoảng tháng 5 đến tháng 10 (Hình 1.1).

Cây XCBV phân bố từ Tây Nam Trung Quốc (Vân Nam) đến miền Bắc Việt Nam (Cao Bằng, Hà Nội, Ninh Bình và Phú Thọ) và miền Nam Việt Nam (Kon Tum) với tổng diện tích ước tính là hơn 3000 km² với số lượng hơn 10.000 cây. Nó mọc ở những nơi ẩm ướt của sườn núi hoặc ven suối, dưới những khu rừng lá rộng ẩm ướt, ở độ cao 800–1500 m. Đây là khu vực thuộc vùng khí hậu nhiệt đới ẩm gió mùa, với điều kiện tự nhiên thuận lợi cho sự phát triển của nhiều loài thực vật phong phú và quý hiếm. Trong số đó có rất nhiều loài là đặc hữu, được sử dụng trong các bài thuốc dân gian từ lâu đời. Đáng chú ý là gần một nửa số quần thể được tìm thấy trong các khu bảo tồn thiên nhiên hoặc các công viên, ví dụ, Vườn Quốc

gia Cúc Phương ở Việt Nam và Khu bảo tồn thiên nhiên quốc gia Laojunshan ở Trung Quốc.



Hình 1.1. Cây Xà căn ba vì (*Ophiorrhiza baviensis*). A. Holotype của *O. alatiflora* H.S. Lo var. *trichoneura* H.S. Lo; B. Hình thái chung; C. Cụm hoa ở mặt bên; D. Mặt bên của cụm hoa; E. Tràng hoa dài; F. Tràng hoa kiểu ngắn. Tỷ lệ = 1 cm [5].

1.2. TÌNH HÌNH NGHIÊN CỨU VỀ CÂY XÀ CĂN BA VÌ TRÊN THẾ GIỚI

Mặc dù chi Xà căn có khoảng 400 loài, tuy nhiên, có lẽ vì số lượng loài lớn và thuộc đối tượng ít quan tâm (Least concern) nên ít nghiên cứu tập trung đến các loài trong chi này. Phần lớn các nghiên cứu đều chỉ liên quan đến thành phần hóa học và phân loại hình thái [5–7]. Một số loài Xà căn được biết đến từ lâu đời với ứng dụng trong y học cổ truyền như Xà căn thảo *Herba Ophiorrhiza Japonicae*, Xà căn Quảng Châu - *Ophiorrhiza cantoniensis Hance* ở Trung Quốc, được sử dụng để điều trị viêm, đau, ung thư và nhiễm trùng do vi khuẩn và virus. Hơn nữa, các loài Xà Căn có khả năng chữa lành vết rạn nứt, viêm miệng, loét và vết thương [8, 9], đồng thời hoạt động như một chất chống oxy hóa [10], thuốc chống ho và thay thế giảm đau [11]. Chúng cũng được áp dụng để điều trị các trường hợp bệnh dạ dày, bệnh phong và vô kinh, bên cạnh việc sở hữu các đặc tính an thần và nhuận tràng thu được từ chiết xuất vỏ rễ của chúng [9]. Trên thực tế, Xà Căn đậu (*Ophiorrhiza mungos L.*) được biết đến với cái tên cụ thể là 'rễ rắn' do nó được biết đến như một phương pháp điều trị vết rạn nứt.

Trong y học hiện đại, các loài Xà Căn rất phổ biến do đặc tính chống ung thư của camptothecin (CPT) cấu thành của chúng, nhờ vào khả năng ức chế topoisomerase-1 của axit deoxyribonucleic (DNA). Tuy nhiên, việc sử dụng chúng trong điều trị các bệnh khác nhau có thể không giống nhau giữa các trường phái điều trị khác nhau. Ví dụ, người Tanchangya ở Bangladesh sử dụng bột nhão của *O. rugosa* var. *prostrata* (D.Don) Deb & Mondal để trị mụn nhọt, những người thuộc bộ lạc Mama pha trà từ lá của nó để trị đau nhức cơ thể hoặc ép lấy nước uống trị tiêu chảy, trong khi bộ tộc Chakma chữa đau tai bằng cách đắp lá đã phơi khô nghiền nát lên da [12]. Các loài Xà Căn rõ ràng rất giàu các phân tử có hoạt tính sinh học, mang lại tác dụng dược lý vượt trội vì chúng có thể được sử dụng để điều trị vô số bệnh từ nhẹ đến mãn tính.

Về khả năng sản xuất CPT, hợp chất này được tìm thấy ở cây Xà Căn đậu từ năm 1985 [13]. Các nghiên cứu về hóa thực vật kéo dài bốn thập kỷ qua đã dẫn đến việc phân lập gần 100 chất chuyển hóa thứ cấp, chủ yếu là alkaloid và anthraquinon, từ các loài Xà Căn khác nhau. Các chất chuyển hóa thứ cấp chính được phân lập từ chi Xà Căn là ancaloit (49), anthraquinon (20), triterpenoit (8), diterpenes (1), sesquiterpenes (3), monoterenes (1), steroid (6), flavonoid (2), coumarin (1), iridoids (6) và axit phenolic (2). Các chất chuyển hóa chính như xanthophylls (1),

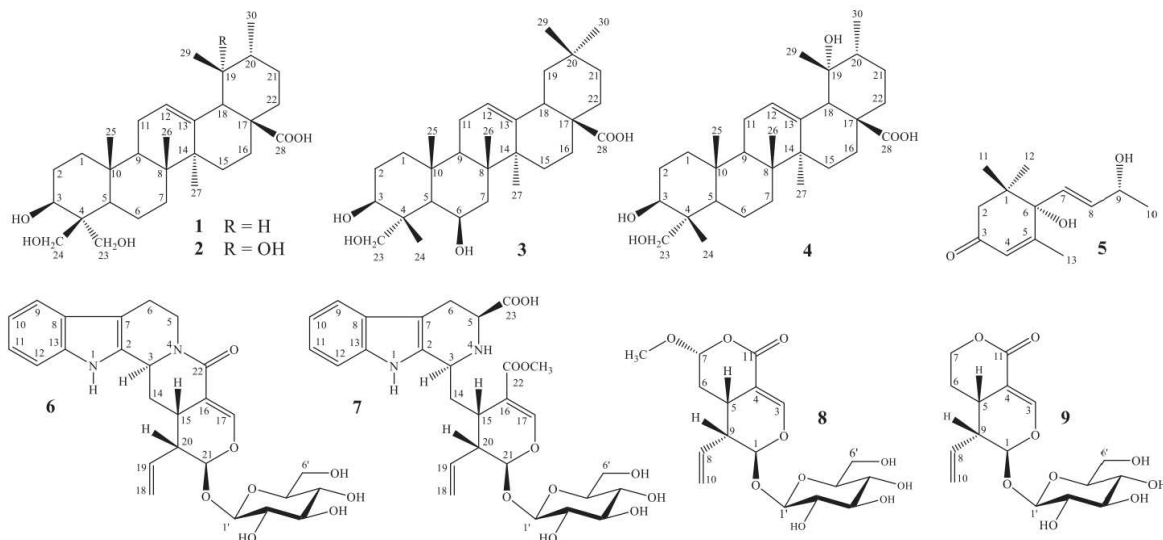
pheophytins (2) và axit béo (3) cũng được báo cáo từ một số loài Xà Cắn. Trong số đó, *Ophiorrhiza mungos* và *Ophiorrhiza mungos* var. *angustifolia* cho thấy hàm lượng CPT cao, trong khi một số loài/giống Xà Cắn cho thấy mức CPT bằng không hoặc không phát hiện được. Các loài Xà Cắn, chủ yếu là *Ophiorrhiza pumila*, được tái sinh thông qua hệ thống nuôi cấy mô cho thấy sự tăng hàm lượng CPT [6]. Mặc dù chứa nhiều hợp chất thứ cấp có ích, đặc biệt là CPT và được sử dụng trong các bài thuốc dân gian lâu đời, tuy nhiên, các nghiên cứu có hệ thống về phân loại, tên gọi, dược tính hay công tác thống kê vùng phân bố và bảo tồn của các loài thuộc chi Xà cắn vẫn còn nhiều thiếu sót và chưa được quan tâm.

Về chi Xà Cắn nói chung, cho đến hiện tại chỉ có hai trình tự hệ gen lục lạp hoàn chỉnh của hai loài *O. pumila* (#MW528277.1) và *O. densa* (#MW683127.1), cùng với 1 phần trình tự hệ gen lục lạp của loài *O. mungos* voucher Bremer 3301 (#KY378702.1) trên cơ sở dữ liệu genbank của Trung tâm Thông tin Công nghệ sinh học Quốc gia (Hoa Kỳ) (NCBI). Về loài XCBV nói riêng, thông tin về đặc điểm sinh thái và hệ gen của loài này vô cùng hạn chế, chỉ có 4 trình tự của XCBV bao gồm trình tự nằm trên vùng nối gen *trnL-trnF* (#MH626989.1), gen *rps16* (#MH626923.1), *ETS* (External transcribed spacer, #MH626743.1) và *ITS* (#MH626804.1). Mỗi trình tự chỉ có kích thước dưới 1000 bp, đều thuộc hệ gen lục lạp. Con số này là quá nhỏ đối với loài dược liệu như XCBV, từ đó đặt ra tiềm năng và sự cần thiết phải thu thập và lưu trữ trình tự của loài này.

1.3. TÌNH HÌNH NGHIÊN CỨU VỀ CÂY XÀ CẮN BA VÌ TRONG NƯỚC

Trong y học cổ truyền Việt Nam, một số loài Xà cắn như cây Xà cắn đậu được sử dụng với tác dụng bổ gan, mật, ngoài ra còn dùng chữa rắn cắn. Công bố của nhóm tác giả Cường và cộng sự vào năm 2019, là nghiên cứu đầu tiên ở Việt Nam cũng như trên thế giới về thành phần hóa học và hoạt tính sinh học của cây XCBV [7]. Nghiên cứu đã chứng minh một triterpene loại ursane mới, axit 3 β , 23,24-trihydroxyurs-12-en-28-oic (1), cùng với tám hợp chất đã biết (2-9) được tạo thành từ các phần trên không của loài cây này (Hình 1.2). Trong số đó, các hợp chất 2-5 lần đầu tiên được tìm thấy từ chi Xà cắn, trong khi các hợp chất 6-9 lần đầu tiên được công bố. Cấu trúc của những chất này đã được làm sáng tỏ bằng các phân tích HR-ESI-MS (High-resolution electrospray ionisation mass spectra - Khối phổ ion hóa phun tĩnh điện phân giải cao) và quang phổ NMR (Nuclear magnetic resonance - Cộng hưởng từ hạt nhân), cũng như so sánh với những công bố trước đó. Hơn nữa,

tất cả các hợp chất phân lập được đánh giá về các hoạt tính gây độc tế bào chống lại MCF-7, Hela, KB, A549 và SK-LU-1 các dòng tế bào ung thư và ảnh hưởng của chúng đối với việc sản xuất NO do LPS gây ra.



Hình 1.2. Công thức hóa học của các hợp chất từ cây Xà cẩu ba vì [7].

Kết quả hiển thị trong Bảng 1.1 cho thấy rằng hợp chất 1, 3 và 4 thể hiện độc tính tế bào đối với tất cả năm dòng tế bào có giá trị IC_{50} dao động từ 37,89 đến 79,6 $\mu\text{g}/\text{mL}$. Các nghiên cứu trước đây đã báo cáo về độc tính tế bào của hợp chất 3 và 4 đối với các dòng tế bào khác. Hợp chất 3 được phát hiện có hoạt tính gây độc tế bào chống lại các dòng tế bào NCI-H460, HepG-2, MCF-7, HL-60, HCT-16 với giá trị IC_{50} là 11,8 đến 77,66 μM , trong khi hợp chất 4 cũng được báo cáo là có biểu hiện độc tính tế bào đối với các dòng tế bào Daoy, Hep-2, HT-29, MCF-7 với giá trị IC_{50}/EC_{50} từ 9,5 đến 29,43 μM [14–17]. Các hợp chất 2, 5-9 không có hoạt tính chống lại tất cả năm dòng tế bào ung thư được thử nghiệm có $IC_{50} > 100 \mu\text{g}/\text{mL}$. Ngoài ra, các hợp chất 1-9 được đánh giá về khả năng ức chế sản xuất NO (Nitric oxide) trong các tế bào RAW264.7 được kích thích bởi LPS (Lipopolysaccharide) (L-NMMA được sử dụng làm đối chứng dương). Kết quả cũng chỉ ra rằng các hợp chất 3–5 và 7–9 cho thấy tác dụng ức chế với giá trị IC_{50} nằm trong khoảng từ 58,25 đến 93,73 $\mu\text{g}/\text{mL}$. Hợp chất 1, 2 và 6 không hiển thị hoạt động với $IC_{50} > 100 \mu\text{g}/\text{mL}$.

Bảng 1.1. Hoạt tính sinh học của các hợp chất khai thác từ cây Xà cẩu ba vì [7].

Hợp chất	Hoạt tính gây độc tế bào					Sản xuất ức chế NO	
	IC ₅₀ (µg/ml)					Hợp chất	IC ₅₀ (µg/ml)
	MCF7	Hela	KB	A549	SK-LU-1		
1	62.18 ±5.39	57.02 ±6.24	79.60 ±7.46	59.50 ±6.84	74.09 ±6.63	1	>100
2	>100	>100	>100	>100	>100	2	>100
3	64.31 ±6.14	57.13 ±5.04	47.73 ±3.15	68.3 3±5.22	60.02 ±3.95	3	58.25 ±6.49
4	37.89 ±2.78	48.22 ±4.98	38.15 ±0.03	46.77 ±5.98	44.09 ±3.90	4	58.72 ±6.51
5	>100	>100	>100	>100	>100	5	80.59 ±4.19
6	>100	>100	>100	>100	>100	6	>100
7	>100	>100	>100	>100	>100	7	68.91 ±2.75
8	>100	>100	>100	>100	>100	8	88.54 ±3.38
9	>100	>100	>100	>100	>100	9	93.73 ±5.29
Ellipticine	0.42 ±0.04	0.41 ±0.02	0.45 ±0.03	0.50 ±0.04	0.36 ±0.02	L- NMMA	7.10 ±068

Mặc dù chứa nhiều hợp chất thứ cấp có ích và được sử dụng trong các bài thuốc dân gian lâu đời, tuy nhiên, các nghiên cứu có hệ thống về phân loại của loài XCBV vẫn còn nhiều thiếu sót và chưa được quan tâm. Cho đến nay, các nghiên cứu ứng dụng các loài cây dược liệu bản địa tại Việt Nam vẫn gặp khó khăn trong việc phân loại để nhận biết chính xác các loài cây được sử dụng. Các phương pháp định danh hình thái đã được áp dụng, tuy nhiên, chưa mang lại hiệu quả do các tiêu chuẩn phân biệt thường dựa trên hình thái bên ngoài của cây như thân, lá, hoa, và quả. Điều này có thể gây lên sự nhầm lẫn trong quá trình phân loại do hình thái của các loài cây trong cùng một chi có độ tương đồng rất cao. Cách giải quyết triệt để cho vấn đề này đó là sử dụng các chỉ thị phân tử, cách tiếp cận này sẽ mang lại kết quả chính xác tuyệt đối trong việc phân loại ở cấp độ loài. Trong lĩnh vực nghiên cứu hệ gen, cho đến hiện tại, trong nước chưa có bất kỳ công bố nào liên quan đến khảo sát hệ gen nhân và hệ gen lục lạp của các loài thuộc chi Xà căn. Do hệ gen thực vật có kích thước khá lớn và tồn nhiều tài nguyên để có thể giải trình tự toàn bộ hệ gen của một loài cây, vì thế, giải trình tự hệ gen lục lạp sẽ là một cách tiếp cận hiệu quả hơn khi ứng dụng trong lĩnh vực phân loại. Bởi vì kích thước hệ gen lục lạp nhỏ trong khoảng 20 - 120kb và đơn giản hơn so với hệ gen nhân, nên lục lạp thường được là đích giải trình tự đầu tiên. Bên cạnh đó, trình tự hệ gen lục lạp cũng được sử dụng rộng rãi trong phân tích tiến hóa, barcoding và meta-barcoding, lại chỉ chứa khoảng 100-120 gene mã hóa protein. Như vậy, có thể thấy sự cần thiết phải có một nghiên cứu trên toàn bộ hệ gen lục lạp của loài XCBV cho công tác phân loại, đánh giá đa dạng và nghiên cứu đặc điểm hệ gen lục lạp, làm cơ sở cho công tác bảo tồn và nghiên cứu mở rộng về sau. Với kích thước ước tính của hệ gen lục lạp của các loài Xà căn là khoảng 154 kb, tiềm năng khai thác thông tin genome trên hệ gen lục lạp này là rất lớn, hứa hẹn cung cấp nhiều thông tin khoa học quan trọng.

1.4. ĐỊNH DANH XÀ CĂN BA VÌ BẢNG CHỈ THỊ PHÂN TỬ

Đi liền với sự phát triển của công nghệ giải trình tự và việc mở rộng các ứng dụng của chỉ thị phân tử đã phát triển hệ thống phân loại các loài sinh vật dựa trên trình tự nucleotide của chúng. Đối với thực vật, ngoài phân loại dựa trên hình thái và đặc điểm sinh trưởng, phát triển, thì việc phân loại dựa trên trình tự nucleotide đóng vai trò rất quan trọng, cho phép các nhà quản lý hay các nhà nghiên cứu tiến hành phân loại loài hiệu quả. Quá trình phân loại thực vật dựa trên trình tự DNA

hay thuật ngữ DNA barcoding là việc sử dụng các trình tự đặc thù trong hệ gen của sinh vật nhằm xác định đến bậc phân loại loài của sinh vật đó [18]. Việc phân loại cho phép xây dựng cơ sở dữ liệu có hệ thống nhằm tìm hiểu, bảo tồn và đánh giá sự đa dạng sinh học của các vùng sinh cảnh khác nhau trên Trái Đất. Đối với thực vật trên cạn, hệ thống chỉ thị phân tử (DNA barcoding) dựa trên trình tự hai gen *rbcL* và *matK*. Hai gen này nằm trên hệ gen lục lạp và để có một cơ sở dữ liệu tốt thì các loài thực vật phải được gán một hồ sơ về trình tự hai gen *rbcL* và *matK*. Việc sử dụng các chỉ thị phân tử trong giới thực vật lại không được chấp nhận từ sớm mà phải những năm trở lại đây với được sử dụng rộng rãi. Do đó, có rất nhiều loài còn thiếu thông tin và trình tự phân loại. Sau khi tìm kiếm mở rộng nhiều vùng gen trên ty thể, lục lạp và gen nhân thì có 4 vùng gen ưu tiên được sử dụng rộng rãi để phân loại thực vật đó là *rbcL*, *matK*, *trnH-psbA* và *ITS*. Sử dụng các chỉ thị phân tử cho phép phân loại loài từ tất cả các giai đoạn phát triển thông thường của một loài thực vật như quả, hạt, mầm, cây trưởng thành đực hay cái, hoặc mẫu thực vật có trong phân của loài động vật ăn thực vật. Do đó, DNA barcoding trở thành công cụ hữu hiệu cho công tác phân loại. Quá trình phân loại dựa trên DNA nhìn chung bao gồm 2 bước chính là: 1) xây dựng thư viện trình tự DNA của các loài đã biết và 2) so sánh và ghép trình tự của loài chưa biết với trình tự có trong thư viện. Bước đầu tiên yêu cầu các nhà phân loại lựa chọn và thu thập một hoặc một vài cá thể trên mỗi loài để làm mẫu tham chiếu trong thư viện. Mẫu có thể là mẫu mô lấy từ chính các bộ sưu tập thực vật trong thư viện hoặc được thu trực tiếp từ cây ngoài môi trường sống của chúng. Quá trình thu mẫu phải đi kèm với việc gắn tag đi kèm thông tin về hình thái. Đây là những cơ sở quan trọng nhằm bổ sung cho quá trình phân loại [19].

Một khi thư viện DNA được hoàn thiện thì có thể sử dụng để xác định cho các mẫu cần phân loại khác. Tuy nhiên, việc phân loại dựa trên một phần gen cục bộ cũng có những hạn chế và hiệu suất phân biệt đến loài là khác nhau giữa các chỉ thực vật. Thêm vào đó, việc thiếu cơ sở dữ liệu trình tự, nghĩa là thiếu trình tự tham chiếu cho bước đầu định danh sẽ dẫn đến hạn chế, cản trở phân loại. Đối với loài XCBV, thực tế là chưa có công trình nghiên cứu cụ thể nào về phân loại của loài này một cách toàn diện và có hệ thống. Trong một nghiên cứu tổng quát loài thuộc chi Xà căn thì loài gần gũi nhất với XCBV là loài Xà căn đậu (*O. mungos*) nằm cùng một nhánh với loài *O. elmeri* và *Spiradiclis bifida* với giá trị bootstrap cao [20]. Bên cạnh đó, bằng trình tự trên vùng gen *ndhF-rps16-trnT-F* thì XCBV tạo

thành nhánh nhóm với các loài *O. hayatana-az37*, *O. japonica-az05*, *O. kwangsiensis-ba56*. Tuy nhiên, các nhánh này không có dạng nhánh đôi, cho thấy mức độ phân loại thấp. Khi sử dụng thêm trình tự vùng *ITS* thì loài XCBV tạo thành nhánh đôi với loài *O. hayatana-cz08*. Điều này cho thấy, việc sử dụng càng đầy đủ các vùng gen thì phân loại càng hiệu quả.

1.5. GIẢI TRÌNH TỰ THỂ HỆ MỚI VÀ ỨNG DỤNG TRONG NGHIÊN CỨU BẢO TỒN NGUỒN GEN VÀ PHÂN LOẠI THỰC VẬT

1.5.1. Giải trình tự thể hệ mới

Các công nghệ giải trình tự đầu tiên được phát triển vào năm 1977 bởi Sanger cùng đồng sự [21] từ Đại học Cambridge được trao giải Nobel hóa học năm 1980 và Maxam AM cùng Gilbert WA [22] từ Đại học Harvard. Khám phá của họ đã mở ra cánh cửa để nghiên cứu mã di truyền của các sinh vật và mang lại nguồn cảm hứng cho các nhà nghiên cứu trong việc phát triển công nghệ giải trình tự nhanh hơn và hiệu quả hơn [23]. Trong đó công nghệ giải trình tự Sanger (Sanger Sequencing - SGS) đã trở thành kỹ thuật được áp dụng nhiều nhất vì hiệu quả cao và độ phóng xạ thấp [24], được tự động hóa để có hiệu suất cao hơn.

Trình tự bộ gen người đầu tiên đã được giải mã bằng phương pháp Sanger vào năm 2004 đã tiêu tốn rất nhiều thời gian và nguồn lực. Do vậy, cần tìm ra các phương pháp có thể rút ngắn thời gian và giảm chi phí giải trình tự toàn bộ hệ gen. Đây là động lực thúc đẩy sự phát triển và thương mại hóa các công nghệ giải trình tự thế hệ mới (Next generation sequencing - NGS) [25]. Công nghệ NGS cho phép phân tích song song hàng loạt với dữ liệu lớn từ nhiều mẫu với chi phí ít hơn [26]. Các công nghệ NGS có thể giải trình tự song song hàng triệu đến hàng tỷ đoạn đọc trong một lần chạy và thời gian cần thiết để tạo ra các đoạn đọc có kích thước GigaBase chỉ là vài ngày hoặc vài giờ, tốt hơn so với giải trình tự thế hệ đầu tiên như giải trình tự Sanger. Tuy nhiên, NGS không có khả năng đọc chuỗi DNA hoàn chỉnh của bộ gen, chúng bị giới hạn trong việc giải trình tự các đoạn DNA nhỏ và phải qua hàng triệu đoạn đọc. Giới hạn này vẫn là một điểm tiêu cực đặc biệt đối với các dự án lắp ráp bộ gen vì nó đòi hỏi tài nguyên máy tính cao [23].

Các công nghệ NGS tiếp tục được cải thiện và số lượng trình tự tăng lên trong những năm qua. Các công nghệ giải trình tự thế hệ thứ hai là các công nghệ giải trình tự mới được phát triển sau thế hệ thứ nhất, chúng có đặc điểm là cần chuẩn bị các thư viện giải trình tự khuếch đại trước khi bắt đầu giải trình tự các

dòng DNA khuếch đại và có những công nghệ giải trình tự thế hệ thứ ba là những công nghệ giải trình tự mới xuất hiện gần đây, ngược lại với thế hệ thứ hai, những công nghệ này được phân loại là Công nghệ giải trình tự đơn phân tử (Single Molecule Sequencing Technology) vì chúng có thể giải trình tự một phân tử đơn lẻ mà không cần thiết phải tạo các thư viện khuếch đại và có khả năng tạo ra các lần đọc dài hơn với chi phí thấp hơn nhiều và trong thời gian ngắn hơn.

Giải pháp cho một thế hệ giải trình tự thứ ba được phát triển và đưa ra thị trường bởi Pacific BioSciences (PacBio). Phương pháp giải trình tự đơn phân tử thời gian thực (SMRT- Single-molecule real-time) cho đoạn đọc dài hơn và tốc độ đọc nhanh hơn các phương pháp giải trình tự thế hệ thứ hai, giúp giải quyết các đoạn đọc khó và các đoạn gen methyl hóa, giải mã cấu trúc bậc hai của DNA và RNA, phát hiện điểm sai khác của gen mà không bị hạn chế bởi lỗi đọc trình tự, và quan trọng hơn hết là giúp lắp ráp *de novo* các bộ gen có kích cỡ và độ phức tạp vượt quá khả năng phân tích của SGS [27].

Công nghệ giải trình tự PacBio

NGS đã mang lại những cải tiến lớn so với giải trình tự Sanger, nhưng những hạn chế của chúng, đặc biệt là độ dài đoạn đọc ngắn, khiến chúng kém phù hợp với một số đối tượng nghiên cứu, bao gồm lắp ráp và xác định vùng gen phức tạp, đồng dạng gen và phát hiện methyl hóa. SMRT được phát triển bởi Pacific BioSciences, cung cấp một phương pháp thay thế để khắc phục những hạn chế này [27].

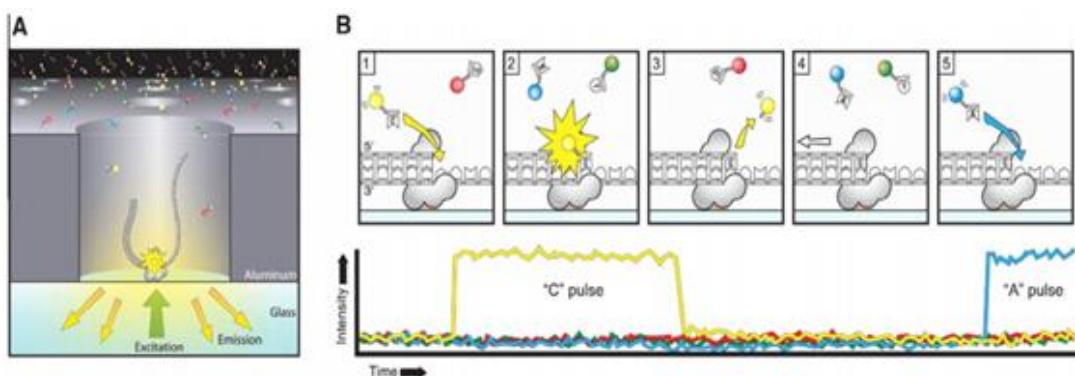
SMRT xác định trình tự DNA bằng cách “quan sát” sự tổng hợp các chuỗi DNA bằng cách tạo ra DNA polymerase đơn lẻ. Bốn loại nucleotide (A, T, G, C) có đánh dấu phosphate tạo tín hiệu được gắn vào mạch để xác định chính xác loại nucleotide trong thời gian thực. Trong khi các đoạn DNA sao chép thì phần mềm tin sinh học hoạt động song song xác định trình tự DNA. Hai quá trình này được hoàn thành cùng lúc.

Phản ứng tổng hợp DNA xảy ra với một lượng thể tích rất nhỏ. Thành phần phản ứng bao gồm: DNA mẫu, DNA polymerase, bốn loại nucleotide gắn gốc huỳnh quang phát ra các màu sắc khác nhau khi bị kích thích bằng tia laser. Bốn loại nucleotide này thực hiện phản ứng tổng hợp DNA như bình thường, tức là nó không gây ngừng quá trình phản ứng. Khi phản ứng tổng hợp DNA xảy ra, thiết bị giải trình tự chiếu tia laser vào vùng phản ứng, tia laser sẽ kích thích nucleotide tương ứng đang được gắn vào mạch phát ra ánh sáng với màu đặc trưng. Màu này

được máy ghi lại và sau đó chuyển thành ký hiệu A, T, G hay C. Khi phản ứng tổng hợp DNA hoàn thành thì việc giải trình tự cũng xong.

Giải trình tự PacBio nắm bắt thông tin trình tự trong quá trình sao chép của phân tử DNA khuôn. Khuôn mẫu, được gọi là SMRTbell, là một DNA hình tròn khép kín, sợi đơn, được tạo ra bằng cách nối các adaptors hình kẹp tóc vào cả hai đầu của phân tử DNA sợi kép (dsDNA- double-stranded DNA) đích. Sợi khuôn SMRTbell là kết quả của một giao thức có thể lựa chọn kích cỡ, trong đó các mảnh khuôn quá lớn hoặc quá nhỏ sẽ bị loại bỏ để đảm bảo quá trình giải trình tự đạt hiệu quả.

Để quan sát quá trình tổng hợp, cấu trúc nanophotonic được sử dụng, được gọi là zero-mode waveguide (ZMW). Mỗi ZMW có đường kính khoảng 70nm và sâu khoảng 100nm, là nơi xảy ra phản ứng tổng hợp DNA. Việc cố định nơi xảy ra phản ứng cho phép phát hiện từng dNTP có gắn huỳnh quang mặc dù nồng độ dNTP được dán nhãn tương đối cao, từ 0,1 đến 10 μM , được DNA polymerase tổng hợp nhanh, chính xác và có quy trình. Quy trình chế tạo ZMW gần đây đã được cải tiến, dẫn đến năng suất cao hơn của các thiết bị thích hợp cho giải trình tự SMRT. Pacific Biosciences đã phát minh ra SMRT chip, là một bản cứng có chứa hàng ngàn khoang ZMW [28].



Hình 1.3. Nguyên tắc giải trình tự DNA đơn phân tử thời gian thực [20].

Mỗi nucleotide được gắn với một gốc phát huỳnh quang riêng biệt có khả năng phát ra ánh sáng màu sắc khác nhau ở các bước sóng khác nhau khi được kích thích bằng tia laser (Hình 1.3). Các gốc phát huỳnh quang này được gắn với nhóm -NH của nucleotide và sẽ bị DNA polymerase loại bỏ để nối nucleotide này với nhóm -OH của nucleotide tiếp theo khi phản ứng sao chép DNA xảy ra. Phần gốc phát huỳnh quang bị cắt ra này sẽ nhanh chóng bị khuếch tán ra ngoài khu vực hoạt động của DNA polymerase. Như vậy thì sau khi gắn xong 1 nucleotide mới, chuỗi

DNA mới tạo thành sẽ là chuỗi bình thường, ko phát huỳnh quang, và sẵn sàng cho phản ứng gắn tiếp theo [27, 28].

Ưu điểm vượt trội của công nghệ PacBio là độ dài đoạn đọc. Trong khi hệ thống khởi nguồn PacBio RS II với bộ hóa chất thế hệ đầu tiên C1 có khả năng tạo những đoạn đọc khoảng 1500 bp thì hệ thống Sequel II hiện nay có thể tạo các đoạn đọc có độ dài trung bình trên 35 kilobase (kb), với chỉ số N50 hơn 50 kb (tức hơn một nửa dữ liệu là các đoạn đọc có độ dài lớn hơn 50 kb), và độ dài đoạn đọc tối đa lớn hơn 175 kb, với dữ liệu data trên mỗi chip là 160Gb cho hệ gen vi khuẩn. Thêm vào đó, do giải trình tự PacBio diễn ra trong thời gian thực (real time) nên dựa trên những thay đổi động lực học của xung ánh sáng, sự biến đổi của các base như methyl hóa có thể được phát hiện.

1.5.2. Ứng dụng của NGS trong nghiên cứu bảo tồn nguồn gen và phân loại thực vật

Tài nguyên di truyền sinh vật là vật liệu ban đầu để lai tạo giống mới và là hạt nhân của đa dạng sinh học, vì thế nó giữ vai trò rất quan trọng trong chiến lược phát triển nông nghiệp của mỗi quốc gia. Với nhận thức đó, Việt Nam đã sớm xây dựng hệ thống văn bản quy phạm pháp luật áp dụng trong quản lý bảo tồn nguồn gen. Mặc dù còn nhiều hạn chế, cho đến nay, khoa học và công nghệ đã cho thấy sự đóng góp đáng kể trong lĩnh vực lưu giữ, bảo tồn và khai thác phát triển nguồn gen, và phát triển kinh tế – xã hội của đất nước.

Ngoài những nhiệm vụ bảo tồn thì các nhiệm vụ ứng dụng công nghệ sinh học trong đánh giá di truyền nguồn gen, hay khai thác và phát triển nguồn gen đã được triển khai và ngày càng đóng góp thiết thực cho các hoạt động nghiên cứu trong các lĩnh vực kinh tế – kỹ thuật quan trọng của đất nước. Chương trình quốc gia về bảo tồn và sử dụng bền vững nguồn gen phần nào đã đáp ứng mục tiêu ứng dụng khoa học và công nghệ để nâng cao hiệu quả của các công tác bảo tồn; đồng thời sử dụng hiệu quả và bền vững các nguồn gen sinh vật để phát triển kinh tế - xã hội; cũng như bảo vệ môi trường và quốc phòng - an ninh; đặc biệt là các đối tượng nguồn gen bản địa, quý, hiếm, đặc hữu có giá trị kinh tế và giá trị khoa học cao [29]. Chương trình cũng hình thành được mạng lưới nguồn gen quốc gia với các tổ chức nghiên cứu đầu mối chuyên ngành (vi sinh vật, động vật, thực vật, thủy sản và dược liệu) đủ mạnh; tối ưu hoá nguồn nhân lực và cơ sở vật chất kỹ thuật cho các tổ chức trong Mạng lưới quỹ gen; tạo lập cơ sở dữ liệu nguồn gen quốc gia phục vụ

công tác bảo tồn, sử dụng bền vững nguồn gen, và xây dựng hệ thống cơ sở dữ liệu quỹ gen quốc gia.

Hiện nay, việc lưu giữ bảo quản chuyên chỗ các nguồn gen cây trồng nông nghiệp đang được thực hiện tại 23 đơn vị thuộc hệ thống. Các hình thức bảo quản chính là ngân hàng gen *in vitro*, ngân hàng gen hạt và ngân hàng gen đồng ruộng. Ngân hàng gen *in vitro* đã bảo quản 200 giống cây rừng, ngân hàng gen hạt giống đã bảo tồn được 1.000 giống của 35 loài cây có hạt, và khu lưu trữ giống bảo quản 850 giống của 20 loài cây. Một số nguồn gen đặc biệt quý, khó có khả năng tái sinh tự nhiên đã được nghiên cứu bảo tồn *in vitro* trong phòng thí nghiệm. Bảo tồn hạt giống dược liệu bao gồm 174 mẫu hạt giống của 143 loài, trong đó 62 loài đã được đánh giá thời gian bảo quản an toàn trong kho lạnh ngắn hạn. Bên cạnh đó, hệ thống đã xây dựng được quy trình bảo tồn chuyên chỗ, giữ 730 loài cây thuốc cần bảo tồn theo 4 cấp độ; đánh giá khả năng lưu giữ trong kho lạnh của 150 loài cây thuốc [30].

Một trong những cách tiếp cận có tiềm năng và thông dụng nhất trong việc bảo tồn nguồn gen và phân loại thực vật có thể kể đến là ứng dụng công nghệ giải trình tự thế hệ mới (NGS). Vài năm gần đây đã chứng kiến những tiến bộ mang tính cách mạng trong công nghệ giải trình tự DNA với sự ra đời của các kỹ thuật NGS. Các phương pháp NGS hiện cho phép giải trình tự hàng triệu bazơ chỉ trong một lần chạy, với chi phí chỉ bằng một phần nhỏ so với giải trình tự Sanger truyền thống. Công nghệ NGS đã có những ứng dụng nổi bật trong sinh học thực vật bao gồm các kỹ thuật trong lĩnh vực phát triển chỉ thị phân tử, lai và lai nhập nội, điều tra phiên mã, nghiên cứu phát sinh loài, sinh thái, di truyền đa bội, và các ứng dụng cho các bộ sưu tập ngân hàng gen lớn.

Khi NGS tiếp tục được cải thiện với độ sâu giải trình tự cao hơn, giảm chi phí và mở rộng ứng dụng cho nhiều dự án từ sinh thái học đến nhân giống có sự hỗ trợ của các chỉ thị phân tử, các thách thức tính toán cũng tăng lên tương ứng. Việc tạo ra 180 triệu đoạn đọc đã trở nên đơn giản, nhưng phải làm gì với độ sâu dữ liệu như vậy là một thách thức. Thách thức đối với dữ liệu NGS còn phức tạp hơn bởi thực tế là mỗi nền tảng giải trình tự đều đưa ra một loạt thách thức riêng đối với việc lắp ráp và phân tích. Một cách tiếp cận để vượt qua thách thức này là sử dụng các phần mềm có sẵn để kiểm soát chất lượng, lắp ráp và phân tích định lượng của trình tự NGS [31–33]. Năm 2009, tạp chí Bioinformatics đã dành toàn bộ một số báo cho các công cụ và thuật toán tin sinh học đã được phát triển cho các thử thách

phân tích trình tự thể hệ mới [34]. Các công cụ và chương trình tin sinh học này liên tục phát triển và cải tiến để bắt kịp với các tiến bộ kỹ thuật NGS, với phần mềm mới luôn được tạo ra.

Trong khi nhiều gói phần mềm (package) ban đầu có sẵn chạy bằng câu lệnh trong môi trường UNIX, một số gói đã xuất hiện trên thị trường cho phép phát triển các pipeline để phân tích hoặc cho phép một nhà khoa học sử dụng các pipeline tính toán hiện có với giao diện thân thiện với người dùng. Nhiều nền tảng trong số này kết hợp các thuật toán đã được phát triển để giải quyết các thách thức của việc lập bản đồ các đoạn đọc thô với bộ gen tham chiếu hoặc thực hiện lắp ráp *de novo* trong trường hợp không có bộ gen tham chiếu. Một trong những nền tảng như vậy là Galaxy [35, 36]. Galaxy là một nền tảng mã nguồn mở hoàn toàn cho phép một nhà khoa học tạo pipeline phân tích tùy chỉnh hoặc sử dụng pipeline của nhà phát triển khác để phân tích. Nền tảng này cho phép người dùng kiểm soát chất lượng dữ liệu, phân tích thống kê, và trực quan hóa kết quả đầu ra.

Trước khi có công nghệ NGS, việc giải quyết một loạt các câu hỏi ở cấp độ hệ gen bị hạn chế đối với những nghiên cứu trên các sinh vật mô hình sở hữu bộ gen lớn (hoặc họ hàng gần của chúng), từ thư viện các chỉ thị được giải trình tự cho đến toàn bộ trình tự bộ gen. Giải trình tự có mục tiêu đề cập đến một loạt các công nghệ được thiết kế để cô lập các vùng gen cụ thể cho NGS. Phần gen được giảm thiểu của mẫu trình tự được nhắm mục tiêu cụ thể cho phép ghép các phản ứng và đơn giản hóa đáng kể việc phân tích. Hiện nay, công nghệ giải trình tự PacBio cũng đã được ứng dụng để giải trình tự hệ gen lục lạp, và đã có nghiên cứu chứng minh cho khả năng vượt trội của PacBio khi lắp ráp *de novo* với độ chính xác 99%, và khi tăng độ lặp lại độ chính xác có thể lên đến trên 99,9%. Cho đến nay đã có rất nhiều công trình sử dụng công nghệ PacBio để giải trình tự hệ gen lục lạp, đặc biệt là các loài có tính ứng dụng cao như các loài dược liệu.

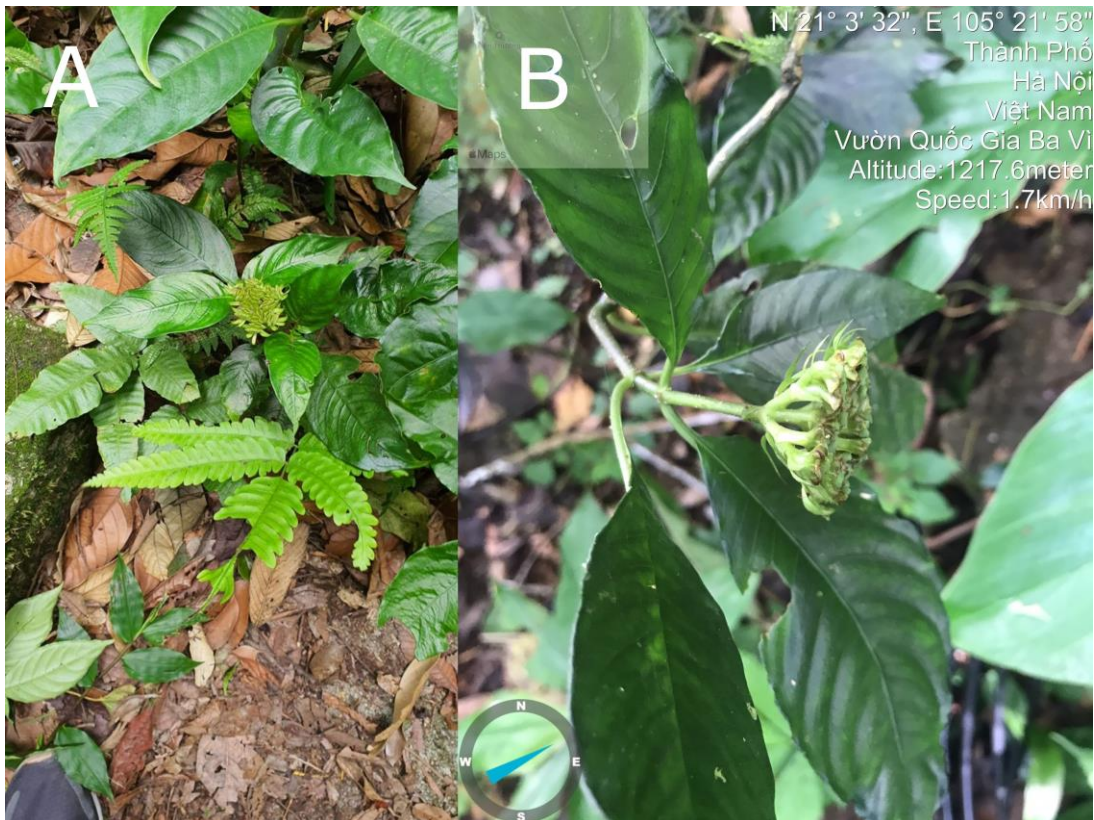
Steele và cộng sự, điều tra việc sử dụng NGS trong phân tích phát sinh loài của hai dòng cây đơn tính, Asparagales và cỏ (grass), sử dụng nền tảng Illumina (80–120bp/đoạn đọc) [37]. Họ đưa ra quan điểm rằng ngay cả dữ liệu có độ che phủ thấp, không nhằm mục đích tập hợp các trình tự hệ gen nhân hoàn chỉnh, cũng có thể cung cấp trình tự bộ gen của các vùng sao chép cao (plastids, ti thể, DNA ribosome nhân) đủ tốt để cung cấp các tập hợp chất lượng cao. Những vùng này có thể cung cấp một lượng lớn các thông tin về phát sinh loài để tạo ra những đơn vị phân loại có liên quan chặt chẽ hơn so với các phân tích phát sinh loài trước đây.

Kết quả không phụ thuộc vào kích thước bộ gen, lượng plastid có trong DNA tổng số (được xác định bằng giá trị PCR Ct thời gian thực), hoặc sự có mặt của các trình tự tham chiếu có sẵn để lắp ráp. Chi phí tạo dữ liệu thấp hơn đáng kể và tiết kiệm được nhiều thời gian trong phòng thí nghiệm. Ngoài ra, có lẽ 90% dữ liệu từ hệ gen nhân vẫn chưa được phân tích và là nguồn tài nguyên có giá trị tiềm năng để phân tích các tập hợp trình tự lặp lại.

Chương 2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. ĐỐI TƯỢNG NGHIÊN CỨU

Mẫu thực vật tươi (Xacan01) được thu thập tại Vườn Quốc gia Ba Vì thuộc Xã Tân Lĩnh, Huyện Ba Vì, Thành phố Hà Nội (N21°3'32", E105°21'58", độ cao 1217,6 mét) (Hình 2.1). Mẫu được để trong túi sạch và bảo quản ở nhiệt độ 4°C. Mẫu thực vật được định danh hình thái và mẫu tiêu bản được lưu trữ tại Viện Sinh thái và Tài nguyên sinh vật- Viện Hàn lâm Khoa học và Công nghệ Việt Nam.



Hình 2.1. (A) Môi trường sống của Xà căn ba vì; (B) Chùm quả của cây Xà căn ba vì. Chụp bởi: Trần Thu Hoài.

2.2. PHƯƠNG PHÁP NGHIÊN CỨU

2.2.1. Tách chiết DNA tổng số của mẫu thực vật

0,1g lá được nghiền bằng bi và nitor lỏng trong máy nghiền. Sau đó, mẫu lá này được thêm 400ul dung dịch ly giải lục lạp Complet buffer (Chloroplast isolation 1X, DTT 1M, BSA 10%- chloroplast isolation kit ab234623- Abcam-Mỹ) để làm giàu nhằm tăng nồng độ DNA lục lạp và ủ 15 phút trong đá. Hỗn hợp sau đó được ly tâm ở 12.000 vòng/phút trong 5 phút và loại bỏ dịch. DNA tổng số của mẫu thu

được tách chiết bằng bộ kit Exgene TM Plant SV mini (lot. No 11722E09032-Geneall-Hàn Quốc) theo hướng dẫn của nhà sản xuất, sử dụng các mẫu được làm giàu (bao gồm cả DNA nhân và DNA lục lạp). Chất lượng và nồng độ DNA được đánh giá bằng điện di trên gel agarose 0.8%, máy Nanodrop 2000 (Thermo) và Qubit lặp lại hai lần.

2.2.2. Tạo thư viện và giải trình tự

DNA tổng số đã tách chiết được phân mảnh thành các đoạn, sau đó được sửa chữa các hỏng hóc từ quá trình phân mảnh và sửa hai đầu 5', 3' bằng bộ kit SMRTbell Damage Repair Kit – SPv3 (Pacific Biosciences - PacBio) trước khi gắn với adapter của PacBio. Các sản phẩm không gắn adapter sẽ bị loại bỏ bởi enzyme Exo III và Exo VII. Thư viện được làm sạch bằng hạt từ Ampure PB (Beckman Coulter), và được kiểm tra độ dài cũng như nồng độ bằng Bioanalyzer 2100. Sau đó thư viện được làm sạch và chọn kích thước bằng Blue Pippin (SageScience) với nồng độ gel 0,75% để lọc ra các đoạn DNA thư viện có độ dài từ 20kb trở lên. Thư viện được kiểm tra lần cuối về kích cỡ và độ phân mảnh với Bioanalyzer 2100 trước khi đưa lên SMRT Cell (PacBio).

Thư viện sau khi chuẩn bị được gắn với polymerase và tinh sạch bằng bộ kit Sequel Binding and Internal Ctrl Kit 3.0 (PacBio) và SMRTbell Clean Up Column v2 Kit-Dif (PacBio) theo quy trình được tạo bởi phần mềm Sample Setup có trong SMRTLink portal phiên bản 5.1.

2.2.3. Lắp ráp hệ gen lục lạp

DNA tổng số đã được giải trình tự bằng cách sử dụng công nghệ giải trình tự PacBio. Các trình tự có nguồn gốc từ bộ gen lục lạp (cp) được xác định thông qua chương trình pbmm2 bằng cách sử dụng bộ gen cp tham chiếu của loài *Ophiorrhiza* thu được từ cơ sở dữ liệu Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>) [38]. Sau đó, Phần mềm Quy trình lắp ráp bộ gen phân cấp phiên bản 4 (Hierarchical Genome Assembly Process - HGAP) đã được sử dụng để lắp ráp bộ gen lục lạp XCBV [39]. Quy trình làm việc của HGAP bao gồm các bước sau: (1) Chọn đoạn đọc trình tự dài nhất làm bộ dữ liệu trình tự hạt giống (seed). (2) Sử dụng từng trình tự hạt giống làm tham chiếu để chọn lọc các đoạn đọc ngắn hơn và lắp ráp sơ bộ các đoạn đọc ngắn thông qua quy trình đồng thuận (consensus). (3) Lắp ráp tổng thể bằng cách sử dụng bộ lắp ráp sơ bộ kết hợp với các đoạn đọc dài. (4) Tinh chỉnh

bản lắp ráp bằng cách sử dụng tất cả dữ liệu đọc ban đầu để tạo ra trình tự consensus cuối cùng đại diện cho bộ gen.

2.2.4. Chú giải hệ gen lục lạp

Các gen mã hóa protein, rRNA và tRNA được chú thích bởi công cụ Geseq [40]. Phần mềm tRNAscan-SE, phiên bản 2.0 đã được áp dụng để xác minh các gen tRNA với các thông số mặc định [41]. Công cụ OrganellarGenomeDRAW (OGDRAW) phiên bản 1.3.142 được chọn để minh họa bản đồ gen [42]. Các trình tự lặp lại được tìm kiếm bằng cách sử dụng hai cách tiếp cận. Công cụ tìm trình tự lặp lại đơn giản (simple sequence repeats, SSR) dựa trên nền tảng web MISA được sử dụng để phát hiện các microsatellites, với các thông số được cài đặt như sau: 10 đơn vị lặp lại cho mono-, 5 đơn vị lặp lại cho di-, 4 đơn vị lặp lại cho tri- và 3 đơn vị lặp lại cho tetra-, penta- và các hexa-nucleotide SSR [43]. Trong số các SSR của mỗi loại, việc so sánh kích thước của các SSR được sử dụng để đếm các SSR đa hình. Ngoài ra, các trình tự lặp lại dài trong hệ gen lục lạp được khảo sát bằng công cụ REPuter với các thông số được thiết lập như sau: kích thước lặp lại tối thiểu là 20 bp, khoảng cách hamming 3 kb và độ tương đồng trình tự 90% trở lên [44].

2.2.5. So sánh hệ gen lục lạp và xây dựng cây phát sinh chủng loại

Để so sánh hệ gen lục lạp cây XCBV, chúng tôi sẽ thu thập các bộ gen lục lạp có sẵn của chi Xà căn từ cơ sở dữ liệu GenBank [38]. Cấu trúc bộ gen tổng thể, kích thước bộ gen, thành phần gen và các trình tự lặp lại trên các bộ gen sẽ được so sánh. Toàn bộ chuỗi plastome của các bộ gen lục lạp Xà căn được căn chỉnh (alignment) bằng công cụ MAFFT và được hiển thị trực quan với chế độ LAGAN trong công cụ mVISTA [45]. Đối với biểu đồ mVISTA, chúng tôi đã sử dụng bộ dữ liệu chú giải gen chức năng của hệ gen lục lạp của đề tài làm tham chiếu. Irscope được sử dụng để hiển thị trực quan và so sánh vùng tiếp giáp của các vùng sao chép đơn lớn (Large single copy - LSC), sao chép đơn nhỏ (Small single copy - SSC) và vùng lặp lại đảo ngược (Inverted repeat - IR) giữa các bộ gen [46]. Chúng tôi cũng xác định tần suất sử dụng codon (Codon usage bias) và phân kỳ trình tự giữa các loài Xà căn thông qua tính toán phân tích độ đa dạng nucleotide (Pi) giữa các bộ gen lục lạp trong phần mềm DNASP phiên bản 6.12.03 [47]. Đối với phân tích phân kỳ trình tự, chúng tôi áp dụng kích thước cửa sổ là 600 bp với kích thước bước 200 bp.

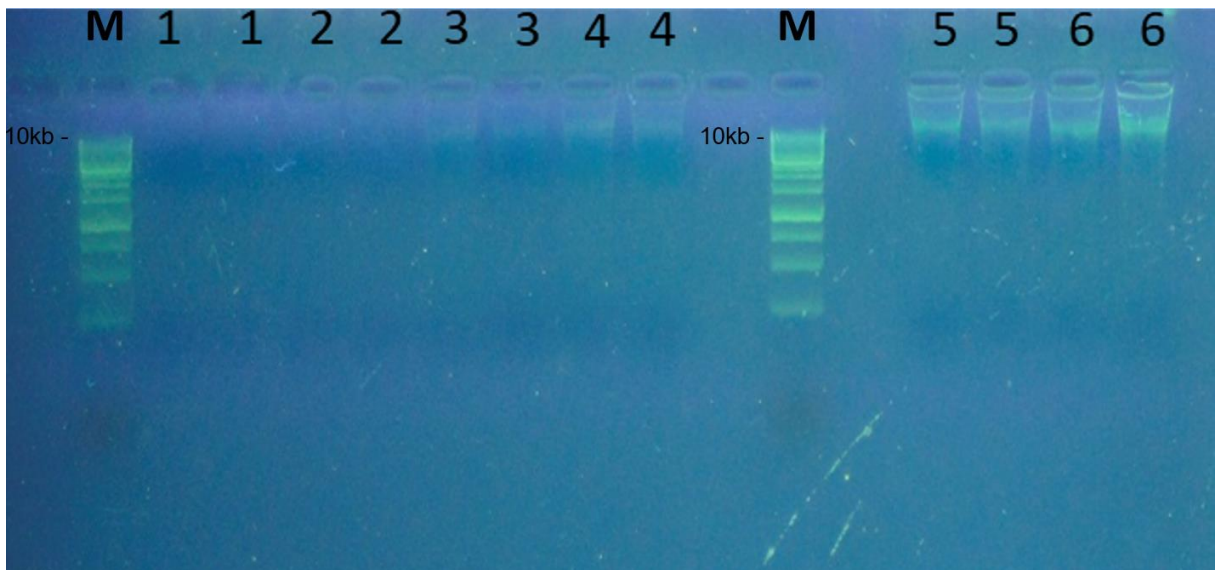
Trình tự kết hợp giữa gen *rps16* và vùng nối gen *trnL-trnF* của các loài Xà căn cùng với các thành viên khác của họ Thiến thảo từ cơ sở dữ liệu Genbank đã được sử dụng để xác định mối quan hệ phát sinh loài của XCBV. Các trình tự này được căn chỉnh bằng phần mềm MAFFT trước khi xây dựng cây phát sinh loài có khả năng tối đa (maximum likelihood tree) được xây dựng bằng FastTree với 1000 bootstrap và được trực quan hoá bằng phần mềm Figtree 1.4.4 [48, 49].

Chương 3. KẾT QUẢ VÀ THẢO LUẬN

3.1. KẾT QUẢ TÁCH CHIẾT VÀ LƯU TRỮ DNA TỔNG SỐ CỦA MẪU THỰC VẬT

Bộ kit ly giải lục lạp của Abcam đưa ra phương án để ly giải lục lạp của rất nhiều loài thực vật. Lục lạp thu được có thể sử dụng trong các nghiên cứu liên quan đến quá trình quang hợp và như vật liệu đầu để nghiên cứu về màng lục lạp, protein, DNA và RNA lục lạp. Mẫu lá được rửa sạch và bảo quản ở nhiệt độ 4°C ít nhất 10 ngày trước khi tách chiết vì việc giữ mẫu ở nhiệt độ này giúp làm giảm đáng kể độ nhớt của mẫu và nồng độ polysaccharide của mẫu.

Lượng mẫu thực vật (lá) yêu cầu cho thí nghiệm này khá lớn (10-20g /lần tách chiết), trong khuôn khổ đề tài, chúng tôi đã sử dụng gần hết số lượng mẫu thu về để tách chiết lục lạp của mẫu Xacan01, tuy nhiên vẫn chưa thành công. Vì lý do đó, chúng tôi đã thử và áp dụng phương pháp làm giàu lục lạp trên mẫu DNA tổng số.



Hình 3.1. Ảnh điện di trên gel agarose 0.8%. M: Ladder 10kb; 1-4: Mẫu tách DNA tổng số; 5-6: Mẫu làm giàu lục lạp.

Hình 3.1 cho thấy kết quả tách chiết của sáu mẫu lá Xacan01. Với các mẫu 1, 2, 3, và 4, DNA tổng số được tách bằng bộ kit Plant SV Mini của Geneall- Hàn quốc theo hướng dẫn của nhà sản xuất. Mẫu 5 và 6 là các mẫu được làm ly giải lục lạp trước khi tách DNA tổng số. Nồng độ và chất lượng của các mẫu được kiểm tra

bằng đo mật độ quang bằng máy Nanodrop 2000 (Thermo), sai số giữa hai lần đo thấp và giá trị trung bình được thể hiện trong Bảng 3.1.

Bảng 3.1. Nồng độ DNA tổng số đo bằng nanodrop.

	1	2	3	4	5	6
Nồng độ (ng/μl)	44,4	53,2	72,4	109,4	98,9	103,15
A260/280	0,8	0,92	1,08	1,07	1,17	1,08

Mẫu 5, là mẫu có xử lý ly giải lục lạp bước đầu, có chỉ số A260/280 cao nhất so với các mẫu còn lại nên được chọn để đi đo Qubit. Kết quả đo nồng độ DNA bằng Qubit là 90,96 ng/ μ l. Tuy chất lượng mẫu chưa được tốt, bDNA trên ảnh điện di trên gel agarose cho thấy nhiều đoạn DNA đứt gãy, chỉ số A260/280 thấp, nhưng lượng DNA là đủ để giải trình tự đoạn dài Pacbio. Vì vậy nhóm nghiên cứu đã quyết định vẫn tiếp tục giải trình tự đoạn dài Pacbio trên mẫu này.

3.2. KẾT QUẢ GIẢI TRÌNH TỰ HỆ GEN LỤC LẠP BẰNG CÔNG NGHỆ GIẢI TRÌNH TỰ PACBIO

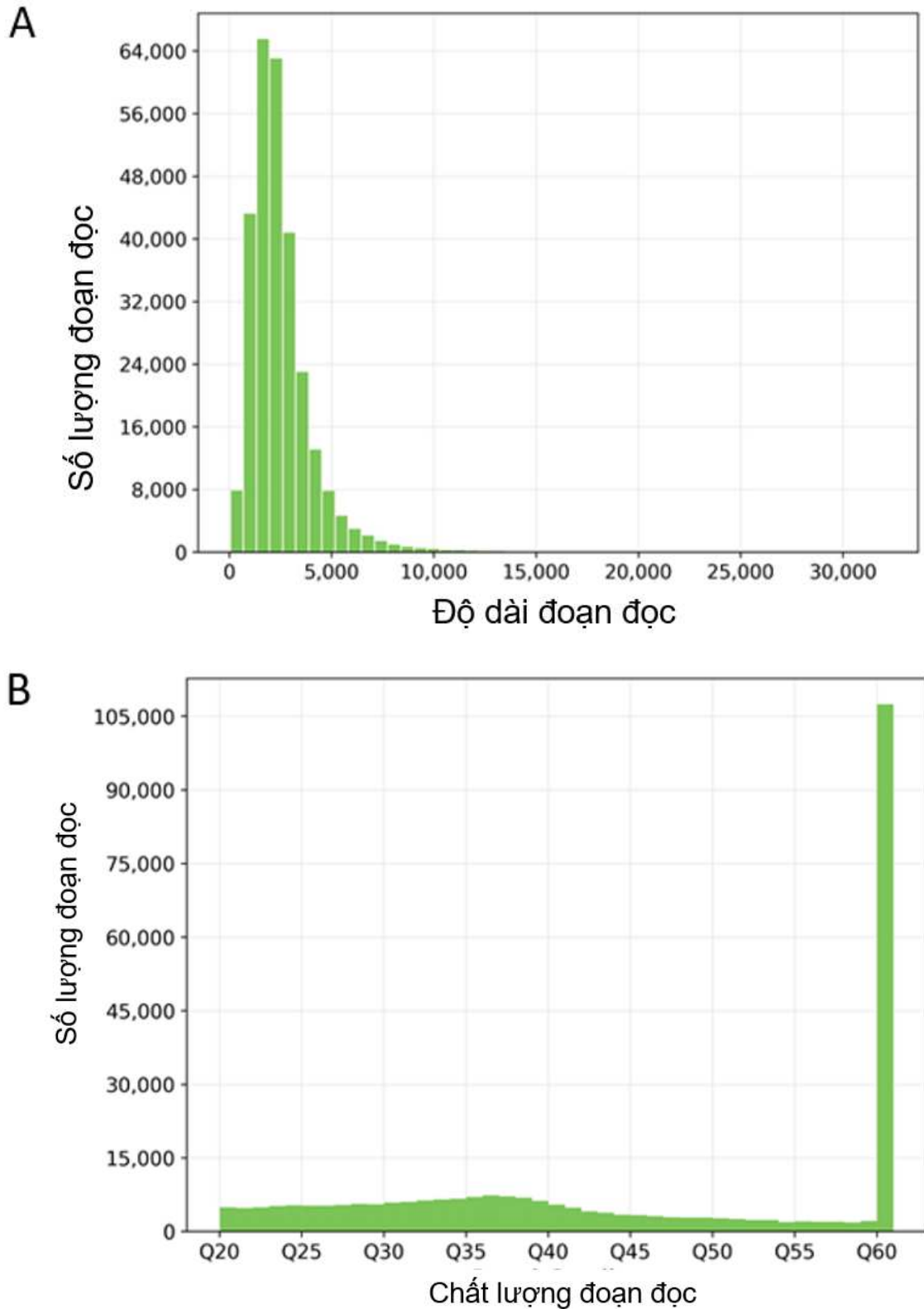
Những tiến bộ trong công nghệ giải trình tự đã cho phép các nhà nghiên cứu giải trình tự DNA dễ dàng hơn và giảm được các chi phí. Các nghiên cứu phát triển công nghệ cho đến nay tập trung vào việc giải trình tự nhiều đoạn đọc ngắn hoặc các đoạn đọc dài hơn nhưng với số lượng ít hơn. Về nguyên tắc, giải trình tự đoạn đọc dài đã có thể thực hiện được với các công nghệ giải trình tự thế hệ thứ ba PacBio và Oxford Nanopore. Tuy nhiên, các công nghệ thế hệ thứ ba này có độ chính xác trình tự kém chỉ 90% (Q10), so với các thông số từ công nghệ Illumina với 99,9% (Q30) [50, 51]. Tuy nhiên, công nghệ PacBio có thể đạt được tỷ lệ lỗi trình tự tương đương với Illumina thông qua một phương pháp gọi là trình tự đồng thuận vòng tròn (CCS - circular consensus sequencing) [52, 53]. Phương pháp CCS của PacBio tạo ra một mẫu “SMRTbell” bằng cách gắn các adapter ssDNA vào dsDNA đích, cho phép polymerase giải trình tự trên từng sợi của dsDNA đích nhiều lần. Quá trình này dẫn đến một đoạn đọc dài liên tục (CLR - continuous long read) bao gồm nhiều đoạn đọc con của trình tự mục tiêu [52].

Để cải thiện độ chính xác của trình tự lên tới 99%, trong nghiên cứu này, chúng tôi đã triển khai phương pháp giải trình tự PacBio CCS với các thông tin được thể hiện trong Bảng phụ lục 1. Phương pháp này cho phép chúng tôi thu được đầy đủ độ dài trình tự tối đa của công nghệ PacBio CCS mà không ảnh hưởng đến chất lượng trình tự. Tổng cộng 28.402.467.862 bp dữ liệu trình tự thô đã được tạo ra với độ dài đoạn đọc trung bình là 1.938 bp, kích thước N50 là 2.412 bp (Bảng 3.2).

Bảng 3.2. Tóm tắt chất lượng giải trình tự.

Chỉ số	Xacan01
Tổng số trình tự thô (bp)	28.402.467.862
Độ dài đoạn đọc trung bình (bp)	1.938
Subread N50 (bp)	2.412
Số đoạn đọc Q20	855.782
Tổng số trình tự thô Q20 (bp)	2.556.497.101
Độ dài đoạn đọc trung bình Q20 (bp)	2.987
Chất lượng đoạn đọc trung bình Q20	Q45
Độ che phủ	158X

Vì DNA tổng số tách chiết được từ mẫu lá của cây XCBV đã được sử dụng để giải trình tự nên trước khi lắp ráp, cần có thêm một bước lọc các đoạn đọc có nguồn gốc từ lặp thể của cây XCBV. Do đó, khoảng 9% số đoạn đọc thô với bộ lọc chất lượng từ Q20 thuộc bộ gen lục lạp XCBV đã được lọc ra bằng cách mapping các trình tự thô với hệ gen lục lạp tham chiếu bằng công cụ pbmm2. Sau khi lọc, đoạn đọc lớn nhất có độ dài lên đến hơn 64 Mbp (trung bình: 2.987 bp), chất lượng đoạn đọc trong khoảng từ Q20 đến Q60 (trung bình: Q45) với độ che phủ 158X (Bảng 3.2, Hình 3.2). Kết quả giải trình tự này cho thấy chất lượng tốt để đưa vào quy trình lắp ráp hệ gen lục lạp trong bước tiếp theo.

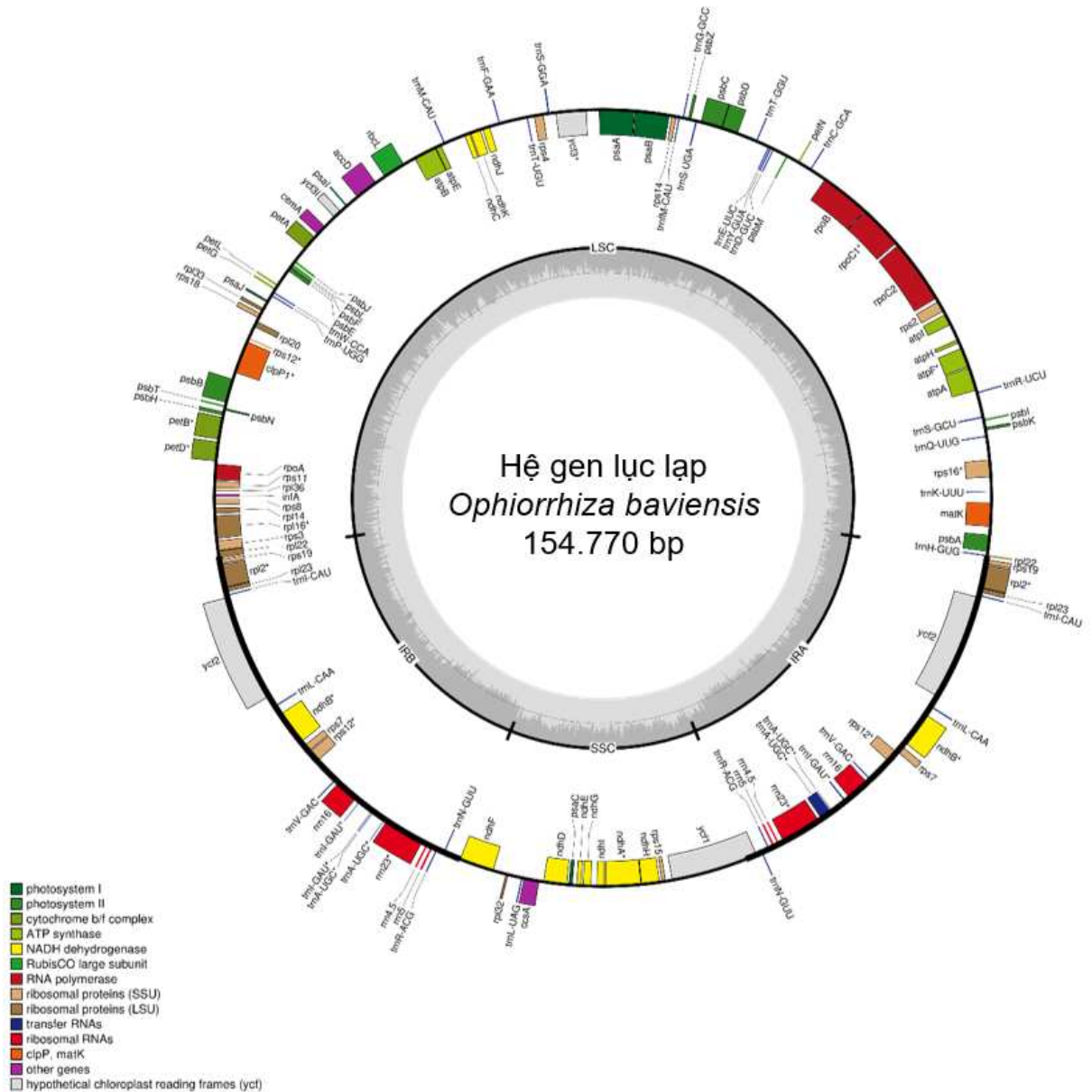


Hình 3.2. Phân bố độ dài (A) và chất lượng (B) đoạn đọc.

3.3. KẾT QUẢ LẮP RÁP HỆ GEN

Trình tự sau khi lắp ráp cho thấy kích thước bộ gen lục lạp là 154.770 bp (Hình 3.3) và tỷ lệ phần trăm của hàm lượng GC là 37,6%. Như đã báo cáo trong

hầu hết các bộ gen lục lạp của thực vật hạt kín, lục thể XCBV được lắp ráp bao gồm đầy đủ bốn cấu trúc điển hình bao gồm các vùng, LSC (84.626 bp), SSC (18.574 bp) và một cặp lặp lại đảo ngược (IR 25.685 bp).



Hình 3.3. Bản đồ hệ gen lục lạp loài Xà căn ba vì ở Việt Nam. Các gen hiển thị bên trong vòng tròn được phiên mã theo chiều kim đồng hồ, trong khi các gen bên ngoài được phiên mã ngược chiều kim đồng hồ. Vòng tròn bên trong màu xám nhạt hiển thị nội dung AT, màu xám đậm tương ứng với thành phần GC.

3.4. KẾT QUẢ CHÚ GIẢI HỆ GEN LỤC LẠP

Chú giải hệ gen lục lạp

Bảng 3.2. Tóm tắt thông tin lắp ráp và chú giải hệ gen lục lạp Xà căn ba vì.

Đặc điểm bộ gen	XCBV
Kích thước bộ gen (bp)	154.770
Kích thước vùng LSC (bp)	84.826
Kích thước vùng SSC (bp)	18.574
Kích thước vùng IR (bp)	25.685
GC (%)	37,6
Số lượng gen	128
Số lượng gen mã hóa protein	87
Số lượng gen mã hóa tRNA	33
Số lượng gen mã hóa rRNA	8

Kết quả chú giải từ GeSeq và tRNAscan-SE cho thấy hệ gen lục lạp của XCBV sở hữu tổng cộng 128 gen, trong đó, có 87 gen mã hóa protein, 33 gen tRNA và 8 gen rRNA (16S, 23S, 5S và 4,5S) (Bảng 3.3). Các mô hình gen chú giải được phân loại thành ba nhóm chính dựa trên chức năng của chúng (Bảng 3.6). Về loại gen liên quan đến quang hợp, có 44 gen mã hóa các tiểu đơn vị của ATP synthase, phức hợp cytochrom, hệ thống quang điện tử I và II, NADPH dehydrogenase, cùng với tiểu đơn vị lớn của Rubisco liên quan đến chuỗi vận chuyển điện tử quang hợp. 76 gen khác thuộc nhóm chức năng liên quan đến quá trình phiên mã và dịch mã. Phần lớn là gen tRNA, và những gen khác là gen rRNA và gen mã hóa RNA polymerase phụ thuộc DNA, các tiểu đơn vị của ribosome và protein ribosome. Chín gen còn lại được phân loại trong danh mục các gen khác, bao gồm năm gen có chức năng liên quan tới quá trình xử lý RNA (*matK*), tổng hợp cytochrom loại c (*ccsA*), tổng hợp axit béo (*accD*), chuyển hóa carbon (*cemA*) và phân giải protein

(*clpP*). Ngoài ra, bốn gen mã hóa các khung đọc được bảo tồn (*ycf1*, *ycf2* và *ycf3*) cũng được chú thích trong hệ gen lục lạp này.

Bảng 3.3. Thành phần gen của hệ gen lục lạp Xà cấn ba vì.

Phân loại	Các nhóm gen	Gen
Quang hợp	Tiểu đơn vị của ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF^a</i> , <i>atpH</i> , <i>atpI</i>
	Tiểu đơn vị của NADH-dehydrogenase	<i>ndhA^a</i> , <i>ndhB</i> (×2) ^a , <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Tiểu đơn vị của phức hợp cytochrome b/f	<i>petL</i> , <i>petB^a</i> , <i>petG</i> , <i>petA</i> , <i>petDa</i> , <i>petN</i>
	Tiểu đơn vị của photosystem I	<i>psaJ</i> , <i>psaC</i> , <i>psaA</i> , <i>psaI</i> , <i>psaB</i>
	Tiểu đơn vị của photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i>
	Tiểu đơn vị của rubisco	<i>rbcL</i>
Phiên mã và dịch mã	Tiểu đơn vị lớn của ribosome	<i>rpl14</i> , <i>rpl16a</i> , <i>rpl2</i> (×2) ^a , <i>rpl20</i> , <i>rpl22</i> , <i>rpl23</i> (×2), <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
	RNA polymerase phụ thuộc DNA	<i>rpoB</i> , <i>rpoA</i> , <i>rpoC1^a</i> , <i>rpoC2</i>
	Tiểu đơn vị nhỏ của protein ribosome	<i>rps11</i> , <i>rps12</i> (×2) ^a , <i>rps14</i> , <i>rps15</i> , <i>rps16a</i> , <i>rps18</i> , <i>rps19</i> (×2), <i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> (×2), <i>rps8</i>

	rRNA	<i>rrn23S</i> (×2), <i>rrn16S</i> (×2), <i>rrn5S</i> (×2), <i>rrn4.5S</i> (×2)
	tRNA	<i>trnA-UGC</i> (×2) ^a , <i>trnC-GCA</i> , <i>trnD-GUC</i> , <i>trnE-UUC</i> , <i>trnF-GAA</i> , <i>trnG-GCCa</i> , <i>trnH-GUG</i> , <i>trnI-GAU</i> (×2) ^a , <i>trnL-CAA</i> (×2), <i>trnL-UAG</i> , <i>trnN-GUU</i> (×2), <i>trnP-UGG</i> , <i>trnQ-UUG</i> , <i>trnR-ACG</i> (×2), <i>trnR-UCU</i> , <i>trnS-GCU</i> , <i>trnS-GGA</i> , <i>trnS-UGA</i> , <i>trnT-GGU</i> , <i>trnT-UGU</i> , <i>trnV-GAC</i> (×2), <i>trnW-CCA</i> , <i>trnY-GUA</i>
	Yếu tố bắt đầu dịch mã	<i>infA</i>
Các gen khác	Tiêu đơn vị của acetyl-CoA-carboxylase (tổng hợp axit béo)	<i>accD</i>
	Gen tổng hợp cytochrom loại c	<i>ccsA</i>
	Protein màng bao (chuyển hóa carbon)	<i>cemA</i>
	Protease	<i>clpP^b</i>
	Maturase (xử lý RNA)	<i>matK</i>
	Các khung đọc mở được bảo tồn	<i>ycf1</i> , <i>ycf2</i> (×2), <i>ycf3^b</i>

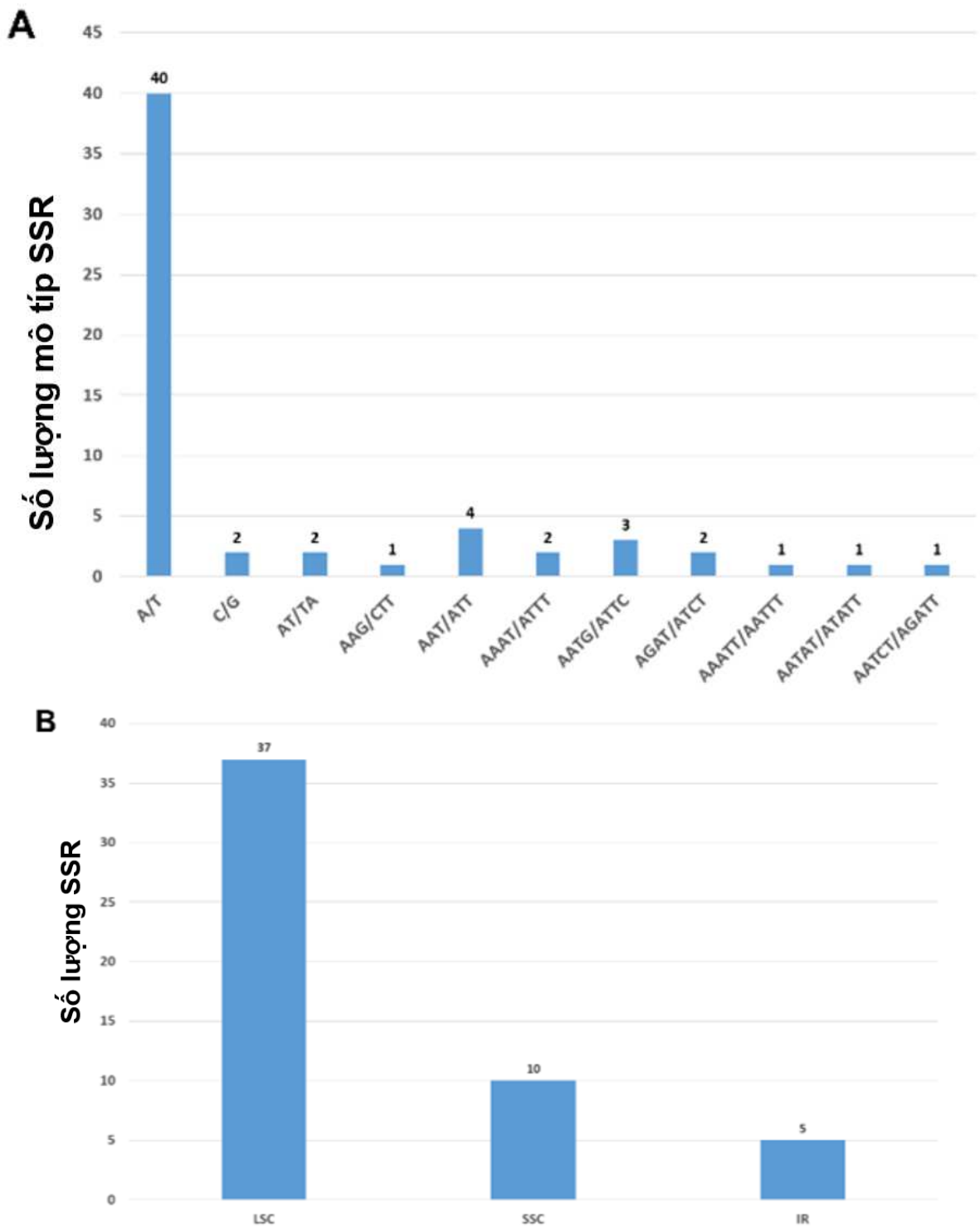
Các gen được đánh dấu là gen có một (a) hoặc đôi (b) intron và các gen có lặp (×2).

Mặt khác, mỗi vùng IR của bộ gen lục lạp XCBV được chú giải bao gồm 18 gen (tất cả 4 gen rRNA, 7 gen tRNA, 1 gen mã hóa protein NADH-dehydrogenase, 4 gen mã hóa protein ribosome và 2 gen khác). Có 19 gen chứa intron, trong đó có 17 gen (*atpF*, *petB*, *petD*, *rpl2* (×2), *rpl16*, *ndhA*, *ndhB* (×2), *rpoC1*, *rps12*, *rps16*, *trnA-UGC* (×2), *trnG -GCC* và *trnI-GAU* (×2)) với một intron và hai gen (*ycf3*, *clpP*) với hai intron (Bảng 3.4).

Hệ gen lục lạp của thực vật hạt kín thường có thành phần và hàm lượng gen được bảo tồn cao với 127-134 gen được tìm thấy. Hệ gen lục lạp XCBV được nghiên cứu đã cho thấy bốn vùng cấu trúc điển hình và kích thước như dự kiến (~ 154 kb) đối với thực vật hạt kín. Phân tích chú giải chức năng gen cho thấy kết quả tương tự như các đặc điểm di truyền đã được công bố trước kia của hệ gen lục lạp hạt kín. Số lượng gen có trong hệ gen lục lạp của XCBV là 128, trong đó, có 17 gen bao gồm một hoặc hai intron. Ngoài ra, việc mất intron ở hai gen *petB* và *petD* đã được quan sát thấy trong bộ gen XCBV. Intron đóng một vai trò quan trọng trong quy định biểu hiện gen. Một số nghiên cứu gần đây đã tiết lộ sự mất một vài gen hoặc intron trong hệ gen lục lạp [19–21], trong đó, sự mất intron của gen *petB* và *petD* đã được báo cáo ở nhiều thực vật hạt kín [22].

Trình tự lặp lại

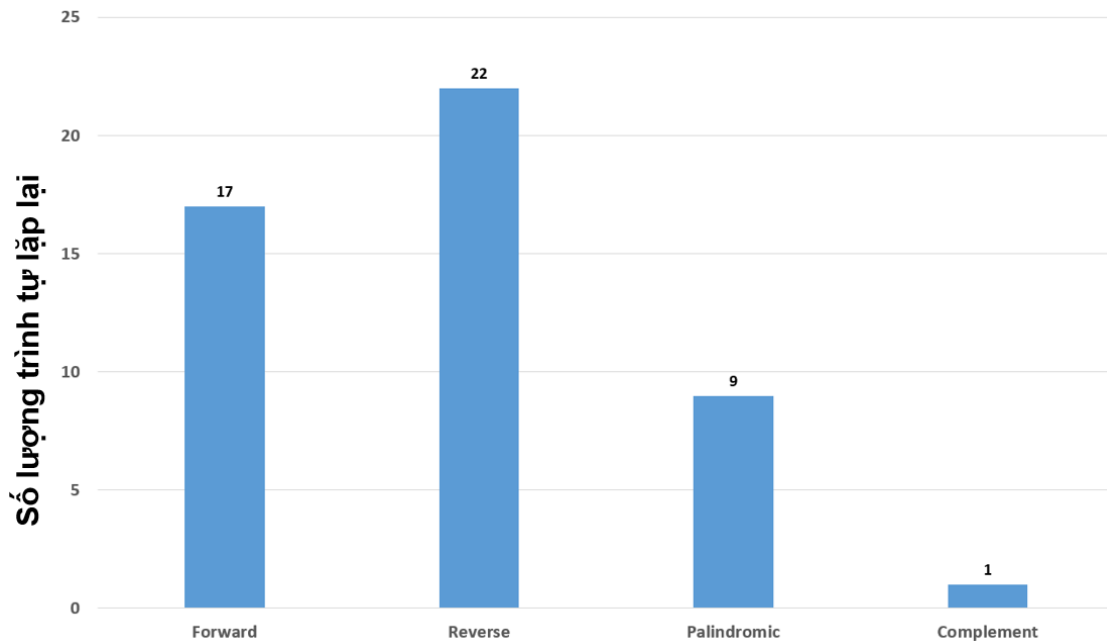
Tổng cộng có 59 trình tự lặp lại đơn giản (simple sequence repeats – SSR, microsatellites) đã được tìm thấy trong bộ gen lục lạp XCBV thông qua công cụ web MISA. Hầu hết tất cả các trình tự lặp lại được sàng lọc đều là các trình tự lặp lại mono- (bao gồm A/T và C) với kích thước dao động từ 10 đến 16 bp (Hình 3.4A). Hai di-, năm tri-, bảy tetra- và ba penta-nucleotide SSR đã được tìm thấy trong lục lạp thể của *O. baviensis* Drake (Hình 3.4B). Những trình tự lặp lại này đóng một vai trò quan trọng trong việc tạo ra các chỉ thị di truyền ở loài XCBV, có thể được áp dụng để đánh giá sự đa dạng ở cấp độ loài trong các nghiên cứu sinh thái học và phát sinh loài.



Hình 3.4. Phân tích các lần lặp lại trình tự đơn của hệ gen lục lạp Xà căn ba vì. (A) Số lượng motif trình tự SSR được xác định; (B) Tần suất của các loại lặp lại trong các vùng LSC, SSC và IR.

Hệ gen lục lạp của XCBV được chú giải bao gồm 49 trình tự lặp lại dài (long repeat) bao gồm 9 trình tự lặp lại dạng palindromic cùng với 12 trình tự lặp lại forward và 22 trình tự lặp lại reverse. Chỉ có một trình tự lặp lại dạng complement

được tìm thấy (Hình 3.5). Kích thước của các trình tự lặp lại được phát hiện nằm trong khoảng từ 20 đến 58 bp, trong khi phần lớn trình tự lặp lại (67%) ngắn hơn 30 bp. Hầu hết các trình tự lặp lại dài đều nằm trong gen *ycf* và vùng đệm giữa các gen (intergenic spacer - IGS).



Hình 3.5. Phân tích trình tự lặp lại dài trên quy mô bộ gen lục lạp của loài Xà cấn ba vì.

Ngoài hai vùng IR, 49 trình tự lặp lại ngắn đã được tìm thấy trong các vùng mã hóa và không mã hóa của plastome XCBV. Hệ gen lục lạp XCBV bao gồm nhiều trình tự lặp lại dài, được coi là chỉ thị sinh học của các điểm đột biến [23,24]. Số trình tự lặp lại tương tự với số liệu của các loài khác thuộc họ Thiến thảo [25,26]. Các trình tự lặp lại đã được báo cáo là có liên quan chặt chẽ đến việc tái cấu trúc plastome ở thực vật hạt kín và có thể được coi là tín hiệu nhận biết sự tái tổ hợp vì khả năng tạo ra các cấu trúc thứ cấp của nó. Tuy nhiên, trong nghiên cứu này các trình tự lặp lại có thể chưa đủ để chứng minh được sự tái tổ hợp giữa các plastome. Ở thực vật bậc cao, SSR được xác định là chỉ thị phân tử quan trọng để điều tra sự biến đổi của quần thể do tính di truyền đơn dòng của chúng và thường được sử dụng để đánh giá sự đa dạng di truyền và cấu trúc quần thể trong các nghiên cứu tiến hóa [27–29]. Tổng cộng, 59 SSR đã được sàng lọc trong bộ gen lục lạp XCBV với phần trăm A/T cao. Những trình tự lặp lại này đóng một vai trò quan trọng trong việc thiết kế các chỉ thị di truyền phân tử ở các loài XCBV, có thể được

áp dụng để đánh giá sự biến đổi ở cấp độ loài trong các nghiên cứu phát sinh loài và sinh thái học.

3.5. KẾT QUẢ SO SÁNH HỆ GEN LỤC LẠP VÀ XÂY DỰNG CÂY PHÁT SINH CHŨNG LOẠI

3.5.1. Kết quả so sánh hệ gen lục lập

Tần suất sử dụng mã di truyền của 64 gen mã hóa protein đã được đánh giá giữa ba bộ gen lục lập: XCBV và hai loài Xà căn có sẵn khác. Tổng số mã di truyền được tìm thấy trong các vùng mã hóa là 51.517, trong đó, mã kết thúc bằng A- và U- được tìm thấy thường xuyên hơn so với G/C- (Bảng 3.7). Leucine là mã di truyền được tìm thấy phổ biến nhất trong số 20 axit amin với tỷ lệ 10,46% (5.068 mã di truyền), theo sau là serine với 9,95% (4.817 mã di truyền). Trong khi đó, hiếm thấy nhất là tryptophan với tổng số 681 mã di truyền chiếm khoảng 1,4%. Có 31 mã di truyền cho thấy codon usage bias (RSCU < 1) trong khi 32 mã di truyền khác được quan sát là thường xuyên hơn mức sử dụng dự kiến ở trạng thái cân bằng (RSCU > 1) (Bảng 3.7). Ngoài ra, tần suất sử dụng cho các mã mở đầu AUG và UGG (methionine và tryptophan) không có bias (RSCU = 1).

Bảng 3.4. Tần suất sử dụng mã di truyền cho các gen mã hóa protein trên hệ gen lục lập Xà căn ba vì.

Codon	AA	ObsFreq	RCSU
UAA	*	1259	1,22
UAG	*	825	0,80
UGA	*	1004	0,98
GCU	A	446	1,23
GCC	A	351	0,97
GCA	A	401	1,11
GCG	A	250	0,69
UGU	C	679	1,20
UGC	C	449	0,80
GAU	D	1012	1,42
GAC	D	413	0,58
GAA	E	1337	1,43

GAG	E	537	0,57
UUU	F	2212	1,20
UUC	F	1481	0,80
GGU	G	540	0,96
GGC	G	383	0,68
GGA	G	747	1,33
GGG	G	577	1,03
CAU	H	880	1,36
CAC	H	414	0,64
AUU	I	1830	1,22
AUC	I	1205	0,80
AUA	I	1471	0,98
AAA	K	2050	1,35
AAG	K	982	0,65
UUA	L	1040	1,23
UUG	L	1095	1,30
CUU	L	1063	1,26
CUC	L	653	0,77
CUA	L	737	0,87
CUG	L	480	0,57
AUG	M	856	1,00
AAU	N	1779	1,38
AAC	N	800	0,62
CCU	P	611	1,03
CCC	P	618	1,04
CCA	P	726	1,23
CCG	P	414	0,70
CAA	Q	987	1,40
CAG	Q	420	0,60
CGU	R	376	0,67
CGC	R	280	0,50
CGA	R	576	1,02

CGG	R	420	0,75
AGA	R	1093	1,94
AGG	R	627	1,12
UCU	S	1113	1,39
UCC	S	982	1,22
UCA	S	824	1,03
UCG	S	622	0,77
AGU	S	747	0,93
AGC	S	529	0,66
ACU	T	668	1,13
ACC	T	651	1,10
ACA	T	647	1,09
ACG	T	406	0,68
GUU	V	784	1,38
GUC	V	411	0,72
GUA	V	682	1,20
GUG	V	402	0,71
UGG	W	681	1,00
UAU	Y	1345	1,36
UAC	Y	637	0,64

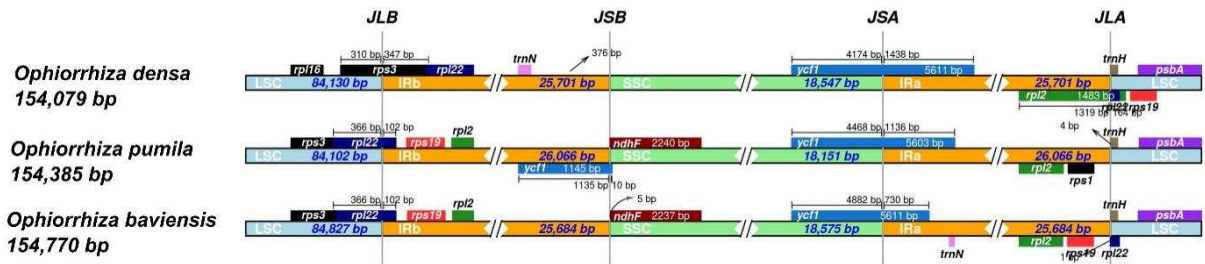
* Mã kết thúc.

Để mô tả sự khác biệt về bộ gen, phân tích tỷ lệ tương đồng trình tự đã được thực hiện giữa ba loài Xà căn với dữ liệu chú giải chức năng của XCBV sử dụng làm tham chiếu. So sánh bằng công cụ mVISTA cho thấy mức độ giống nhau giữa ba loài khá cao và có một số vùng biến đổi trình tự (Hình 3.6). Kết quả cho thấy tần suất phân kỳ cao hơn ở vùng LSC và SSC so với vùng IR. Ngoài ra, các vùng mã hóa của ba bộ gen lục lạp cho thấy tính bảo tồn cao, trong khi đó, phần lớn các biến dị được phát hiện trong các trình tự không mã hóa được bảo tồn (conserved noncoding sequences - CNS) (Hình 3.6). Trong số các trình tự mã hóa protein, các gen thể hiện sự khác biệt nhiều trong ba hệ gen lục lạp bao gồm *matK*, *rpoC2*, *rpoB*, *clpP*, *rpl16*, *ndhF*, *ndhA*, và *ycf1*.

Các phân tích so sánh giữa hệ gen lục lạp XCBV và hai loài Xà căn có sẵn đã được thực hiện để khám phá cấu trúc plastome. Thành phần gen và mô hình sử dụng codon cho thấy độ bảo tồn cao, có thể áp dụng cho các nghiên cứu di truyền quần thể và phát sinh loài. Hơn nữa, ba hệ gen lục lạp Xà căn ít biến dị hơn trong các vùng mã hóa so với các vùng không mã hóa, điều này phù hợp với mô hình phổ biến được tìm thấy ở hầu hết các thực vật hạt kín [30] (Hình 3.6). Tần suất sử dụng codon có liên quan chặt chẽ đến biểu hiện gen và ảnh hưởng đến mức độ biểu hiện mRNA và protein trong bộ gen [31–33]. Axit amin phổ biến nhất trong Xà căn là Leucine (Leu), cũng thường được phát hiện trong các thực vật hạt kín khác. Sự tương đồng cao trong thành phần codon có thể chỉ ra rằng các loài Xà căn này đã trải qua áp lực môi trường tương tự trong suốt quá trình tiến hóa của chúng. Các hệ gen lục lạp Xà căn chỉ ra rằng giá trị RSCU của hầu hết các codon kết thúc bằng A/U đều lớn hơn 1, điều này có thể do phần trăm hàm lượng A/T cao hơn trong hệ gen. Ngoài ra, chúng tôi đã nghiên cứu rằng trình tự gen *ycf1* cùng với năm đoạn đệm giữa các gen (IGS), bao gồm các vùng *petA-psbJ*, *trnH-GUG-psbA*, *trnS-GCU-trnR-UCU*, *psbM-trnD-GUC*, và *ndhC-trnM-CAU*, có giá trị đa dạng nucleotide tương đối cao ($Pi > 0,015$). Các vùng phân kỳ này có thể được nghiên cứu để cung cấp các chỉ thị phân tử cho DNA barcoding và nghiên cứu phát sinh loài ở chi Xà căn.

Ranh giới IR/LSC và IR/SSC của ba bộ gen lục lạp Xà căn được so sánh bằng công cụ IRscope. Nhìn chung, kết quả chỉ ra rằng kích thước vùng, thành phần và hàm lượng gen có sự tương đồng cao giữa hệ gen lục lạp của XCBV và *O. pumila* (Hình 3.8). Mặt khác, hệ gen lục lạp *O. densa* cho thấy một vài sự khác biệt với hai loài Xà căn trên. Kích thước của các vùng IR dao động từ 25.684 bp (XCBV) đến 26.066 bp (*O. pumila*), kích thước vùng IR của *O. densa* là 25.701 bp. Gen *rpl22* nằm trong vùng LSC có đoạn chồng lấp dài 102 bp với vùng IRb đối với XCBV và *O. pumila*, trong khi ở hệ gen *O. densa* thì đoạn chồng lấp dài 347 bp của gen *rps3* được tìm thấy trong vùng ranh giới này. Ngoại trừ *O. densa*, gen *ndhF* được phát hiện trên ranh giới vùng SSC và IRb của hai hệ gen lục lạp còn lại. Đường ranh giới giữa IRa và SSC được tìm thấy trong gen *ycf1* với các phần đuôi lần lượt dài 1438, 1316 và 730 bp của gen này nằm trong vùng IRa của *O. densa*, *O. pumila* và XCBV tương ứng (Hình 3.8). Vùng ranh giới giữa IRa và LSC chỉ ra sự hiện diện của gen *trnH* ở chiều thuận của cả ba loài và gen *rpl22* ở chiều ngược của

XCBV và *O. densa*. Kết quả phân tích IR cho thấy sự co lại hoặc mở rộng vùng IR của ba loài.



Hình 3.8. So sánh các vị trí tiếp giáp các vùng cấu trúc giữa ba bộ gen lục lạp. JLB (vùng ranh giới IRb/LSC), JSB (vùng ranh giới IRb/SSC), JSA (vùng ranh giới IRa/SSC), JLA (vùng ranh giới IRa/LSC).

Các vùng cấu trúc của hệ gen lục lạp ở thực vật trên cạn hay bị thay đổi về chiều dài trong quá trình tiến hóa, dẫn đến sự xuất hiện của nhiều đặc điểm trong vùng ranh giới [34]. Sự mở rộng và thu hẹp ranh giới giữa vùng IR và vùng sao chép đơn (SC) là nguyên nhân chính gây ra sự thay đổi kích thước của bộ gen lục lạp và gây ảnh hưởng đến tốc độ tiến hóa [35,36]. Nghiên cứu của chúng tôi cho thấy tập hợp các gen nằm trong vùng ranh giới của loài Xà căn bao gồm *rpl22*, *rps19*, *ndhF*, *ycf1* và *trnH*. Bên cạnh đó, một số sự sắp xếp lại gen đáng chú ý đã được quan sát thấy trong plastome của loài *O. densa*. Đó là sự hiện diện của gen *rps3* tại vùng JLB thay vì gen thường thấy là *rpl22*, sự mở rộng của gen *rpl2* sang JLA và sự vắng mặt của gen *rps19* trong các vùng IR. Sự mở rộng và co lại cũng như sự biến đổi tại điểm nối của các vùng SC-IR đã cho thấy rằng tổ chức gen ở các vùng IR có thể chứng minh khoảng cách tiến hóa giữa các loài ở một mức độ nào đó.

3.5.2. Kết quả phân tích phát sinh loài

Số lượng trình tự sẵn có của XCBV trên cơ sở dữ liệu của Ngân hàng gen, đặc biệt là thuộc về bộ gen lục lạp, rất hạn chế (chỉ có gen *rps16* và trình tự spacer DNA giữa hai gen *trnL-trnF*). Do đó, chúng tôi đã trích xuất các trình tự này từ bộ gen lục lạp XCBV đã lắp ráp và sử dụng chúng để tiếp cận mối quan hệ phát sinh loài của loài XCBV được nghiên cứu. Hình 3.9 cho thấy độ phân giải phát sinh gen dựa trên tổ hợp trình tự giữa gen *rps6* và vùng đệm giữa các gen *trnL-trnF* với giá trị bootstrap cao lên đến 92% giữa loài XCBV trong nghiên cứu này và trình tự

tham chiếu thuộc XCBV voucher Averyanov & al. VH940 (AAU) (#MH626923.1). Với giá trị bootstrap là 100%, tất cả tám loài thuộc chi Xà căn được nhóm lại và tách ra riêng biệt với hai loài *Xanthophytum*. Trong trường hợp xác định barcoding trong họ Thiến thảo, các trình tự *rps16-trnL-F* kết hợp mang lại khả năng phân tích phát sinh loài cao.



Hình 3.9. Cây phát sinh loài Maximum Likelihood dựa trên các trình tự gen *rps16* và vùng nối gen *trnL-trnF*. Các con số trên các nhánh biểu thị tỷ lệ phần trăm bootstrap sau 1000 lần lặp lại trong quá trình xây dựng cây phân loại. Loài được điều tra trong nghiên cứu này được tô màu đỏ.

Phần lớn các cấp độ phân loại của các kết nối phát sinh loài thực vật đã được chứng minh bằng cách sử dụng bộ gen lục lạp hoàn chỉnh và gen mã hóa protein [37,38]. Nghiên cứu hiện tại cung cấp phân tích phát sinh loài của chi Xà căn dựa trên các trình tự kết hợp giữa gen *rps16* và spacer DNA *trnL-F* kết hợp. Nghiên cứu trước đây của Razafimandimbison và Rydin đã chứng minh rằng XCBV đã được chứng minh là có một mối quan hệ gần gũi với hai loài *O. japonica* và *O. hayatana* [39]. Trong nghiên cứu này, cây phát sinh loài dựa trên sự kết hợp của gen *rps16* và trình tự đệm liên gen *trnL-F* cho thấy mối quan hệ chặt chẽ giữa loài được nghiên cứu và voucher Averyanov & al. VH940 (AAU). Cách tiếp cận này cho thấy hiệu quả trong việc phân loại các bậc phân loại thấp hơn trong họ Thiến thảo. Hơn nữa, sự kết hợp của các chỉ thị này có thể dẫn đến việc phân loại loài tốt hơn so với kết quả từ một gen duy nhất [39]. Nghiên cứu này sẽ giúp làm rõ vị trí tiến hóa của XCBV trong chi Xà căn, cũng như cung cấp dữ liệu hệ gen lục lạp có thể áp dụng

để nghiên cứu sâu hơn về nguồn gốc và sự đa dạng của họ Thiến thảo. Nhìn chung, nghiên cứu phát sinh loài dựa trên bộ gen lục lạp XCBV của chúng tôi đã thành công trong việc phát hiện ra các mối liên hệ trong chi Xà cấn.

KẾT LUẬN VÀ KIẾN NGHỊ

KẾT LUẬN

Mẫu Xà căn ba vì thu tại vườn Quốc gia Ba Vì đã được tách chiết thành công với chất lượng đủ để thực hiện giải trình tự hệ gen.

Bộ gen lục lạp được giải trình tự cho tổng cộng 28.402.467.862 bp đoạn đọc thô, trong đó, khoảng 9% với bộ lọc chất lượng từ Q20 thuộc bộ gen lục lạp XCBV.

Quá trình lắp ráp đã tạo ra bộ gen lục lạp có kích thước 154.770 bp bao gồm đầy đủ bốn vùng cấu trúc điển hình, và tỷ lệ phần trăm của hàm lượng GC là 37,6%. Kết quả chú giải cho thấy hệ gen lục lạp của XCBV sở hữu tổng cộng 128 gen, trong đó, có 87 gen mã hóa protein, 33 gen tRNA và 8 gen rRNA. Theo kết quả so sánh giữa XCBV với hai lạp thể thuộc chi Xà căn đã được công bố khác, cấu trúc và thành phần gen thể hiện sự tương đồng cao và phân tích các vùng tiếp giáp SC-IR cho thấy sự mở rộng và thu hẹp của các vùng IR.

Cây phát sinh loài chỉ ra mối quan hệ chặt chẽ giữa loài được nghiên cứu và voucher Averyanov & al. VH940 (AAU). Nghiên cứu này cung cấp tiềm năng sử dụng bộ gen lục lạp để tăng cường phân loại loài và bảo tồn nguồn gen trong quá trình nghiên cứu sâu hơn về họ Thiến thảo.

KIẾN NGHỊ

Ứng dụng kết quả nghiên cứu với thông tin về các đặc điểm hệ gen lục lạp của loài Xà căn ba vì trong phân tích tiến hóa, barcoding và meta-barcoding, làm cơ sở cho công tác bảo tồn và nghiên cứu mở rộng về sau.

DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ

Pham, M.H.; Tran, T.H.; Le, T.D.; Le, T.L.; Hoang, H.; Chu, H.H. The Complete Chloroplast Genome of An *Ophiorrhiza baviensis* Drake Species Reveals Its Molecular Structure, Comparative, and Phylogenetic Relationships. *Genes* 2023, 14, 227. <https://doi.org/10.3390/genes14010227>

DANH MỤC TÀI LIỆU THAM KHẢO

1. John A., Rajan R., Baby S., 2018, Secondary metabolites from *Ophiorrhiza*. 08, . <https://doi.org/10.2174/2210315508666180515104735>
2. Wu L., 2017, Notes on *Ophiorrhiza hispida* (Rubiaceae) from China. *J. Trop. Subtrop. Bot.* 25:597
3. Lei W., S. H., YU Y-L., 2017, The taxonomic identity of *Ophiorrhiza rarior* and *O. mycetiifolia* (Rubiaceae). *Phytotaxa*, 299, pp. 261–266. <https://doi.org/10.11646/phytotaxa.299.2.10>
4. Chen T., Taylor C., 2011, *Ophiorrhiza*. In: Press S (ed) *Flora of China*, Wu, Z.Y., Beijing & Missouri Botanical Garden Press, St. Louis, pp 258–282
5. Lei W., Liu W-J., Nguyen KS., 2019, Revision of three taxa of *Ophiorrhiza* (Rubiaceae) from China. *Phytotaxa*, 387, pp. 129–139. <https://doi.org/10.11646/phytotaxa.387.2.5>
6. Asano T., Sudo H., Yamazaki M., Saito K., 2006, Camptothecin production in cell cultures of *Ophiorrhiza* species. *Med. Plant Biotechnol.* 451–467
7. Viet Cuong LC., Anh LT., Huu Dat TT., Anh TTP., Lien LQ., et al., 2021, Cytotoxic and anti-inflammatory activities of secondary metabolites from *Ophiorrhiza baviensis* growing in Thua Thien Hue, Vietnam. *Nat Prod Res*, 35, pp. 4218–4224. <https://doi.org/10.1080/14786419.2019.1693564>
8. Krishnan SA., Dileepkumar R., Nair AS., Oommen O V., 2014, Studies on neutralizing effect of *Ophiorrhiza mungos* root extract against *Daboia russelii* venom. *J Ethnopharmacol*, 151, pp. 543–547. <https://doi.org/https://doi.org/10.1016/j.jep.2013.11.010>
9. Midhu CK., Hima S., Binoy J., Satheeshkumar K., 2019, Influence of incubation period on callus tissues for plant regeneration in *Ophiorrhiza pectinata* Arn. through somatic embryogenesis. *Proc Natl Acad Sci India Sect B Biol Sci*, 89, pp. 1439–1446. <https://doi.org/10.1007/s40011-018-01061-x>
10. Martins D., Nunez C V., 2015, Secondary metabolites from Rubiaceae species. *Molecules* 20:13422–13495
11. C Varghese S., K P D., Rajan R., M.V K., Gopalakrishnan R., et al., 2012, A new record of *Ophiorrhiza trichocarpon* Blume (Rubiaceae: Ophiorrhizeae) from Western Ghats, India: Another Source Plant of Camptothecin. *J Sci Res*, 4, . <https://doi.org/10.3329/jsr.v4i2.9378>
12. Adnan M., Nazim Uddin Chy M., Mostafa Kamal ATM., Azad MOK., Paul A., et al., 2019, Investigation of the biological activities and characterization of bioactive constituents of *Ophiorrhiza rugosa* var. *prostrata* (D.Don) & Mondal leaves through in vivo, in vitro, and in silico approaches. *Molecules*, 24, . <https://doi.org/10.3390/molecules24071367>
13. Hsiang YH., Hertzberg R., Hecht S., Liu LF., 1985, Camptothecin induces protein-linked DNA breaks via mammalian DNA topoisomerase I. *J Biol Chem*, 260, pp. 14873–14878

14. Lee T-H., Juang S-H., Hsu F., Wu C., 2005, Triterpene acids from the leaves of *Planchonella duclitan* (Blanco) Bakhuizen. *J Chinese Chem Soc*, 52, .
<https://doi.org/10.1002/jccs.200500184>
15. Quang TH., Ngan NTT., Minh C Van., Kiem P Van., Boo H-J., et al., 2012, Cytotoxic triterpene saponins from the stem bark of *Kalopanax pictus*. *Phytochem Lett*, 5, pp. 177–182. <https://doi.org/https://doi.org/10.1016/j.phytol.2011.12.005>
16. Thang TD., Kuo P-C., Yu C-S., Shen Y-C., Hoa LTM., et al., 2011, Chemical constituents of the leaves of *Glochidion obliquum* and their bioactivity. *Arch Pharm Res*, 34, pp. 383–389. <https://doi.org/10.1007/s12272-011-0305-y>
17. Luo Y., Xu Q-L., Dong L-M., Zhou Z-Y., Chen Y-C., et al., 2015, A new ursane and a new oleanane triterpene acids from the whole plant of *Spermacoce latifolia*. *Phytochem Lett*, 11, pp. 127–131. <https://doi.org/https://doi.org/10.1016/j.phytol.2014.12.005>
18. de Vere N., Rich T., Trinder S., Long C., 2015, DNA barcoding for plants. *Methods Mol Biol*, 1245, pp. 101–118. https://doi.org/10.1007/978-1-4939-1966-6_8
19. Kress WJ., 2017, Plant DNA barcodes: Applications today and in the future. *J Syst Evol*, 55, pp. 291–307. <https://doi.org/https://doi.org/10.1111/jse.12254>
20. Razafimandimbison SG., Rydin C., 2019, Molecular-based assessments of tribal and generic limits and relationships in Rubiaceae (Gentianales): Polyphyly of Pomazoteae and paraphyly of Ophiorrhizeae and Ophiorrhiza. *Taxon*, 68, pp. 72–91. <https://doi.org/https://doi.org/10.1002/tax.12023>
21. Sanger F., Nicklen S., Coulson AR., 1977, DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74, pp. 5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>
22. Maxam AM., Gilbert W., 1977, A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74, pp. 560–564. <https://doi.org/10.1073/pnas.74.2.560>
23. Kchouk M., Gibrat J-F., Elloumi M., 2017, Generations of sequencing technologies: From first to next generation. *Biol Med*, 09, .
<https://doi.org/10.4172/0974-8369.1000395>
24. Pareek CS., Smoczynski R., Tretyn A., 2011, Sequencing technologies and genome sequencing. *J Appl Genet*, 52, pp. 413–435. <https://doi.org/10.1007/s13353-011-0057-x>
25. Nguyen DT., 2018, Giải trình tự thế hệ mới: tổng quan về các xu hướng và ứng dụng. <https://sinhhocvietnam.com/giai-trinh-tu-the-he-moi-tong-quan-ve-cac-xu-huong-va-ung-dung/>
26. Shendure J., Ji H., 2008, Next-generation DNA sequencing. *Nat Biotechnol*, 26, pp. 1135–1145. <https://doi.org/10.1038/nbt1486>
27. Rhoads A., Au KF., 2015, PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, 13, pp. 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>
28. Anonymous., 2016, Large genome assembly with PacBio long reads.

- <https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Large-Genome-Assembly-with-PacBio-Long-Reads>
29. Cục bảo tồn đa dạng sinh học., 2018, Bảo tồn và khai thác phát triển nguồn gen ở Việt Nam
 30. Bộ Nông nghiệp và Phát triển Nông thôn., 2016, Khoa học và công nghệ với bảo tồn, khai thác, và phát triển nguồn gen
 31. Horner DS., Pavesi G., Castrignanò T., De Meo PD., Liuni S., et al., 2010, Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform*, 11, pp. 181–197. <https://doi.org/10.1093/bib/bbp046>
 32. Metzker ML., 2010, Sequencing technologies — the next generation. *Nat Rev Genet*, 11, pp. 31–46. <https://doi.org/10.1038/nrg2626>
 33. Miller JR., Koren S., Sutton G., 2010, Assembly algorithms for next-generation sequencing data. *Genomics*, 95, pp. 315–327. <https://doi.org/10.1016/j.ygeno.2010.03.001>
 34. Bateman A., Quackenbush J., 2009, Bioinformatics for next generation sequencing. *Bioinformatics*, 25, pp. 429. <https://doi.org/10.1093/bioinformatics/btp037>
 35. Giardine B., Riemer C., Hardison RC., Burhans R., Elnitski L., et al., 2005, Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15, pp. 1451–1455. <https://doi.org/10.1101/gr.4086505>
 36. Goecks J., Nekrutenko A., Taylor J., Team TG., 2010, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11, pp. R86. <https://doi.org/10.1186/gb-2010-11-8-r86>
 37. Steele PR., Hertweck KL., Mayfield D., McKain MR., Leebens-Mack J., et al., 2012, Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *Am J Bot*, 99, pp. 330–348. <https://doi.org/10.3732/ajb.1100491>
 38. Coordinators NR., 2013, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 41, pp. D8–D20. <https://doi.org/10.1093/nar/gks1189>
 39. Chin C-S., Alexander DH., Marks P., Klammer AA., Drake J., et al., 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, 10, pp. 563–569. <https://doi.org/10.1038/nmeth.2474>
 40. Tillich M., Lehwark P., Pellizzer T., Ulbricht-Jones ES., Fischer A., et al., 2017, GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*, 45, pp. W6–W11. <https://doi.org/10.1093/nar/gkx391>
 41. Chan PP., Lin BY., Mak AJ., Lowe TM., 2021, tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*, 49, pp. 9077–9096. <https://doi.org/10.1093/nar/gkab688>

42. Lohse M., Drechsel O., Bock R., 2007, OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet*, 52, pp. 267–274. <https://doi.org/10.1007/s00294-007-0161-y>
43. Beier S., Thiel T., Münch T., Scholz U., Mascher M., 2017, MISA-web: a web server for microsatellite prediction. *Bioinformatics*, 33, pp. 2583–2585. <https://doi.org/10.1093/bioinformatics/btx198>
44. Kurtz S., Schleiermacher C., 1999, REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, 15, pp. 426–427. <https://doi.org/10.1093/bioinformatics/15.5.426>
45. Frazer KA., Pachter L., Poliakov A., Rubin EM., Dubchak I., 2004, VISTA: computational tools for comparative genomics. *Nucleic Acids Res*, 32, pp. W273–9. <https://doi.org/10.1093/nar/gkh458>
46. Amiryousefi A., Hyvönen J., Poczai P., 2018, IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*, 34, pp. 3030–3031. <https://doi.org/10.1093/bioinformatics/bty220>
47. Rozas J., Ferrer-Mata A., Sánchez-DelBarrio JC., Guirao-Rico S., Librado P., et al., 2017, DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol*, 34, pp. 3299–3302. <https://doi.org/10.1093/molbev/msx248>
48. Katoh K., Rozewicki J., Yamada KD., 2019, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*, 20, pp. 1160–1166. <https://doi.org/10.1093/bib/bbx108>
49. Price M., Dehal P., Arkin A., 2010, FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, pp. e9490. <https://doi.org/10.1371/journal.pone.0009490>
50. Fox EJ., Reid-Bayliss KS., Emond MJ., Loeb LA., 2014, Accuracy of next generation sequencing platforms. *Next Gener Seq Appl*, 1, . <https://doi.org/10.4172/jngsa.1000106>
51. Weirather JL., de Cesare M., Wang Y., Piazza P., Sebastiano V., et al., 2017, Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6, pp. 100. <https://doi.org/10.12688/f1000research.10571.2>
52. Travers KJ., Chin C-S., Rank DR., Eid JS., Turner SW., 2010, A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*, 38, pp. e159. <https://doi.org/10.1093/nar/gkq543>
53. Wenger AM., Peluso P., Rowell WJ., Chang P-C., Hall RJ., et al., 2019, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, 37, pp. 1155–1162. <https://doi.org/10.1038/s41587-019-0217-9>

PHỤ LỤC

Bảng 1. Tóm tắt thông tin thông tin chạy mẫu giải trình tự.

Thông số	Lần chạy ngày 12.11.2022 17:55
Tên mẫu	Xacan01
Loại thiết bị	Sequel
Số thiết bị	54241
Thể tích mẫu được sử dụng	16 μ L
Số lượng SMRTCells	1
Nồng độ	1.0 ng/ μ L 0.22 nM
Primer	Sequencing Primer v4
Loading	Diffusion
Chế độ	CCS Reads
Binding Kit	Sequel® Binding Kit 3.0
Cleanup	Có
AMPure Cleanup Anticipated Yield	50%
Nồng độ trên Plate	11 pM
Nồng độ Primer	19.9 nM
Nồng độ Template	0.1 nM
Nồng độ Polymerase	14.8 nM
Nồng độ Template	0.1 nM
Thể tích Pipetting tối thiểu	1 μ L

Hà Nội, ngày 18 tháng 8 năm 2022

PHIẾU XÁC ĐỊNH TÊN KHOA HỌC

I. Thông tin mẫu gửi

- Đơn vị/ Người gửi mẫu: TS. Trần Thu Hoài – Viện Công nghệ sinh học, Viện Hàn Lâm khoa học công nghệ Việt Nam
- Địa chỉ: Số 18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội.
- Ký hiệu mẫu: Xacan01
- Tình trạng mẫu: Mẫu tiêu bản tươi.
- Số lượng tiêu bản: 03.
- Nơi lấy mẫu: Vườn Quốc gia Ba Vì.
- Thời gian lấy mẫu: 14/8/2022
- Thời gian gửi mẫu: 15/8/2022

II. Người và phương pháp xác định

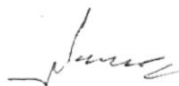
- Người xác định: TS. Nguyễn Sinh Khang
- Đơn vị: Phòng Tài nguyên thực vật, Viện Sinh thái và Tài nguyên sinh vật
- Phương pháp xác định:
 - + Xác định mẫu vật dựa trên phương pháp hình thái so sánh.
 - + Các tài liệu sử dụng chính: Thực vật chí Việt Nam, Cây cỏ Việt Nam (tập 1-3), Danh lục thực vật Việt Nam (tập 2,3), <http://www.theplantlist.org/>;

III. Kết quả xác định

- Ký hiệu mẫu: Xacan01; Tên mẫu: Xà Cắn Ba Vì, Cây dẹt Ba Vì
- Tên Khoa học: *Ophiorrhiza baviensis* Drake.
- Họ thực vật: Rubiaceae

(Kết quả xác định tên khoa học chỉ có giá trị đối với mẫu được gửi đến và được lưu tại phòng Tài nguyên thực vật)

Người xác định





TS. Nguyễn Sinh Khang



Nguyễn Văn Sinh

Article

The Complete Chloroplast Genome of An *Ophiorrhiza baviensis* Drake Species Reveals Its Molecular Structure, Comparative, and Phylogenetic Relationships

Mai Huong Pham ¹ , Thu Hoai Tran ¹, Thi Dung Le ¹, Tung Lam Le ¹ , Ha Hoang ¹ and Hoang Ha Chu ^{1,2,*}¹ Institute of Biotechnology (IBT), Vietnam Academy of Science & Technology (VAST), Hanoi 100000, Vietnam² Faculty of Biotechnology, Graduate University of Science and Technology, VAST, Hanoi 100000, Vietnam

* Correspondence: chuhoangha@ibt.ac.vn

Abstract: *Ophiorrhiza baviensis* Drake, a flowering medical plant in the Rubiaceae, exists uncertainly within the *Ophiorrhiza* genus' evolutionary relationships. For the first time, the whole chloroplast (cp) genome of an *O. baviensis* Drake species was sequenced and annotated. Our findings demonstrate that the complete cp genome of *O. baviensis* is 154,770 bp in size, encoding a total of 128 genes, including 87 protein-coding genes, 8 rRNAs, and 33 tRNAs. A total of 59 SSRs were screened in the studied cp genome, along with six highly variable loci, which can be applied to generate significant molecular markers for the *Ophiorrhiza* genus. The comparative analysis of the *O. baviensis* cp genome with two published others of the *Ophiorrhiza* genus revealed a high similarity; however, there were some notable gene rearrangements in the *O. densa* plastome. The maximum likelihood phylogenetic trees were constructed based on the concatenation of the *rps16* gene and the *trnL-trnF* intergenic spacer sequence, indicating a close relationship between the studied *O. baviensis* and other *Ophiorrhiza*. This study will provide a theoretical molecular basis for identifying *O. baviensis* Drake, as well as species of the *Ophiorrhiza* genus, and contribute to shedding light on the chloroplast genome evolution of Rubiaceae.



Citation: Pham, M.H.; Tran, T.H.; Le, T.D.; Le, T.L.; Hoang, H.; Chu, H.H. The Complete Chloroplast Genome of An *Ophiorrhiza baviensis* Drake Species Reveals Its Molecular Structure, Comparative, and Phylogenetic Relationships. *Genes* **2023**, *14*, 227. <https://doi.org/10.3390/genes14010227>

Academic Editors: Wajid Zaman and Hakim Manghwar

Received: 30 November 2022

Revised: 19 December 2022

Accepted: 7 January 2023

Published: 15 January 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: *Ophiorrhiza baviensis*; chloroplast genome; comparative analysis; phylogeny

1. Introduction

The chloroplast (cp) is an essential organelle in photosynthetic plant and microbial cells that produces energy to feed the cell through photosynthesis [1]. Each chloroplast contains its own ribosomes and a separate genome from the cell's nuclear genome, ranging in size from 20 to 160 kilobase pairs (kp). The cp genome is uniparentally inherited with a quadripartite structure consisting of one large single-copy (LSC) region, one small single-copy (SSC) region, and two inverted repeat regions (IRs) of the same length [2]. As a result of the small size of the cp genome, which contains only around 100 to 120 protein coding genes, chloroplasts are often the first target for sequencing in evolutionary analysis, barcoding, and meta-barcoding [2]. In the NCBI Genbank database at present, there are more than 1000 cp genomes of plant species. However, this number is very small compared to the existing plant diversity on the planet, which raises the need to collect and store sequences of uncharacterized species.

For medicinal plants such as *Ophiorrhiza baviensis*, the potential for exploitation and the need for systematic classification are even more essential. *O. baviensis* is a species of flowering plant in the Rubiaceae family, first described scientifically by Drake in 1895, and re-identified by Wu et al. [3]. Information on the ecology and genomic characteristics of this species is extremely limited, with only four sequences of *O. baviensis*—the gene junctions *trnL-trnF* (#MH626989.1), *rps16* (#MH626923.1), the external transcribed spacer (ETS) (#MH626743.1), and *ITS* (#MH626804.1)—available on the Genbank database of the National Center for Biotechnology Information (USA) (NCBI). Each sequence is less than

1000 base pairs (bp) in size, only two of which belong to the chloroplast genome. Thus, there is a need to study the entire chloroplast genome of *O. baviensis* species for taxonomy and diversity assessment, as well as chloroplast genome characterization, conservation, and future research. With an estimated chloroplast genome size of 154 kb, the potential for exploiting genomic information on the *O. baviensis* chloroplast genome is very large.

Recently, PacBio sequencing technology has been applied to sequence cp genomes, and there have been studies demonstrating the superior ability of PacBio in de novo assembly with 99% accuracy; moreover, as the repeatability increases, this can exceed 99.9% [4]. PacBio sequencing is also a great technology in resolving gaps in rRNA, i.e., internal transcribed spacer (ITS) regions and the surrounding regions to obtain accurate molecular biology information for species identification. For the first time, we report a new complete chloroplast genome of *O. Baviensis* Drake from Vietnam and compare it with previously published *Ophiorrhiza* complete chloroplast genome data to evaluate the genome organization, phylogenetic relationships, and conserved genetic resources.

2. Materials and Methods

2.1. Sample Collection and Chloroplast Genome Sequencing

O. baviensis samples were collected in Ba Vi National Park, Hanoi, Vietnam in August 2022 (code number: Xacan 01), 1217.6 m, 21°3'32" N; 105°4'58" E (Figure 1). The voucher specimens were placed in the herbarium of the Institute of Ecology and Biological Resources (HN), Hanoi, Vietnam. Fresh leaves with the same code number were used to extract genomic DNA.

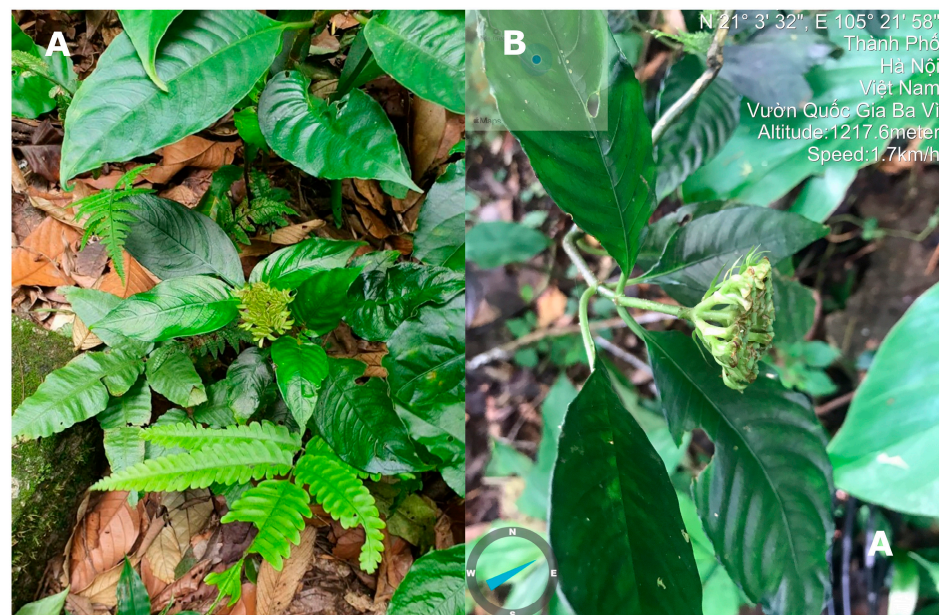


Figure 1. *O. baviensis* Drake. (A) Habitat; (B) Morphological characteristic of infructescence in side view; Photos by Thu Hoai Tran.

2.2. DNA Extraction and Chloroplast Genome Sequencing

We treated samples prior to extraction with the Chloroplast Isolation Kit (ab234623-Abcam, Cambridge, UK) for cp enrichment to increase the cpDNA concentration. The total DNA was extracted by the GeneAll®Exgene™ Plant SV mini kit using the enriched samples (including both genomic DNA and cp DNA). The extracted DNA integrity was evaluated by electrophoresis on a 0.8% agarose gel for 45 minutes at 120 V, and the DNA concentration was measured by Nanodrop 2000 (Thermo, Waltham, MA, USA) and Qubit 2.0 devices to ensure quality for library preparation and sequencing on the Pacbio system according to the manufacturer's instructions.

The total DNA was fragmented and the DNA damage from fragmentation, as well as the 5'/3' ends, underwent repair using the SMRTbell Damage Repair Kit SPv3 (#100-992-200, Pacific Biosciences, Menlo Park, CA, USA) before being attached to PacBio adapters. Products without adapters are rejected by the Exo III and Exo VII enzymes. The SMRTbell library was cleaned with Ampure PB beads (Beckman Coulter, Brea, CA, USA) and checked for length and concentration using the Bioanalyzer 2100. Subsequently, it was cleaned and sized using BluePippin (SageScience, Beverly, MA, USA) with a gel concentration of 0.75% to filter out library DNA fragments above 20 kb in length. The library was lastly checked for size and fragmentation with the Bioanalyzer 2100 before loading to the SMRT Cell (#101-008-000, PacBio).

The prepared library was loaded on one chip and sequenced on a PacBio SEQUEL system at the National Key Laboratory for Gene Technology, Institution of Biotechnology (Hanoi, Vietnam). SMRTbell library was attached with polymerase and purified using the Sequel Binding Internal Ctrl Kit 2.0 (#101-400-900, PacBio) and the SMRTbell Clean Up Column v2 Kit-Dif (101-184-100, PacBio) according to the procedure generated by the Sample Setup software included in the SMRTLink portal version 5.1.

2.3. Genome Assembly and Annotation

Total DNA was sequenced using the PacBio platform. Sequences derived from the cp genome were identified through the pbmm2 program using the cp genome of the reference *Ophiorrhiza* species (accession number: NC_057496.1) obtained from the Genbank database [5]. Then, the Hierarchical Genome Assembly Process version 4 (HGAP4) software was used to assemble the cp genome [6]. Protein-coding genes and RNA were annotated by the GeSeq webtool [7], while tRNAscan-SE software version 2.0 was applied to verify the tRNA genes [8]. The OrganellarGenomeDRAW (OGDRAW) web-tool was selected to generate the circular gene map [9]. Repeat elements were identified using two approaches. The web-based MISA finder was used for detecting microsatellites in nucleotide sequences, with the following parameters: 10 repeats for mono-, 5 for di-, 4 for tri-, and 3 for tetra-, penta-, and hexa-nucleotide SSRs [10]. Size comparison of the SSRs among the SSRs of each type was used to count polymorphic SSRs. The size and pattern of repeats in the cp genome were identified using the REPuter with the following set of parameters: minimum repeat size 20 bp, hamming distance 3 kb, and 90% or more sequence similarity [11].

2.4. Genome Comparison and Phylogenetic Identification

For cp genome comparison, we collected available cp genomes of *Ophiorrhiza* species (*O. pumila*—NC_057496.1 and *O. densa*—NC_058252.1) from the GenBank database (<https://www.ncbi.nlm.nih.gov/genbank/>, accessed on 15 November 2022). The overall genome structure, gene content, genome size, and number of repeats across the genomes were compared. The entire cp genome sequences of the *Ophiorrhiza* species were aligned through MAFFT software with default parameters and visualized in the mVISTA webtool with the LAGAN mode [12]. We used the annotated cp genome of the project as the reference genome in the mVISTA diagram. Subsequently, Irscope was used to visualize and compare the contiguous region between the large and small single-copy, along with the inverted repeat regions of the genomes. We also examined codon usage bias and sequence divergence via computational nucleotide diversity (Π) analysis among cp genomes in DnaSP software version 6.12.03 [13]. For the sequence divergence analysis, we applied a window size of 600 bp with a step size of 200 bp.

A concatenation of the *rps16* gene and *trnL-trnF* intergenic spacer sequences from the *Ophiorrhiza* species and two *Xanthophytum* species of the Rubiaceae family from the Genbank database was used to identify the phylogenetic relationships of the studied *O. Baviensis* Drake. The nucleotide sequences were aligned with MAFFT software with default parameters [14] before the maximum likelihood (GTR+CAT model) phylogenetic tree was constructed using FastTree [15] with a 1000 bootstrap and visualized by FigTree software version 1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>, accessed on 1 July 2021).

3. Results

3.1. Chloroplast Genome Assembly and Annotation

Using the PacBio SEQUEL I system, 28,402,467,862 bp of raw sequence data were generated with a mean read length of 1938 bp, an N50 contig size of 2412 bp, and approximately 9% of the raw reads belonging to the *O. baviensis* cp genome with 158 × coverage. The resequencing assembly resulted in a circular cp genome size of 154,770 bp (Figure 2), and the percentage of GC content was 37.6%. As reported in most angiosperm cp genomes, the assembled *O. baviensis* Drake plastome demonstrated the typical quadripartite structure consisting of four regions, LSC (84,626 bp), SSC (18,574 bp), and a pair of inverted repeats (IRs 25,685 bp).

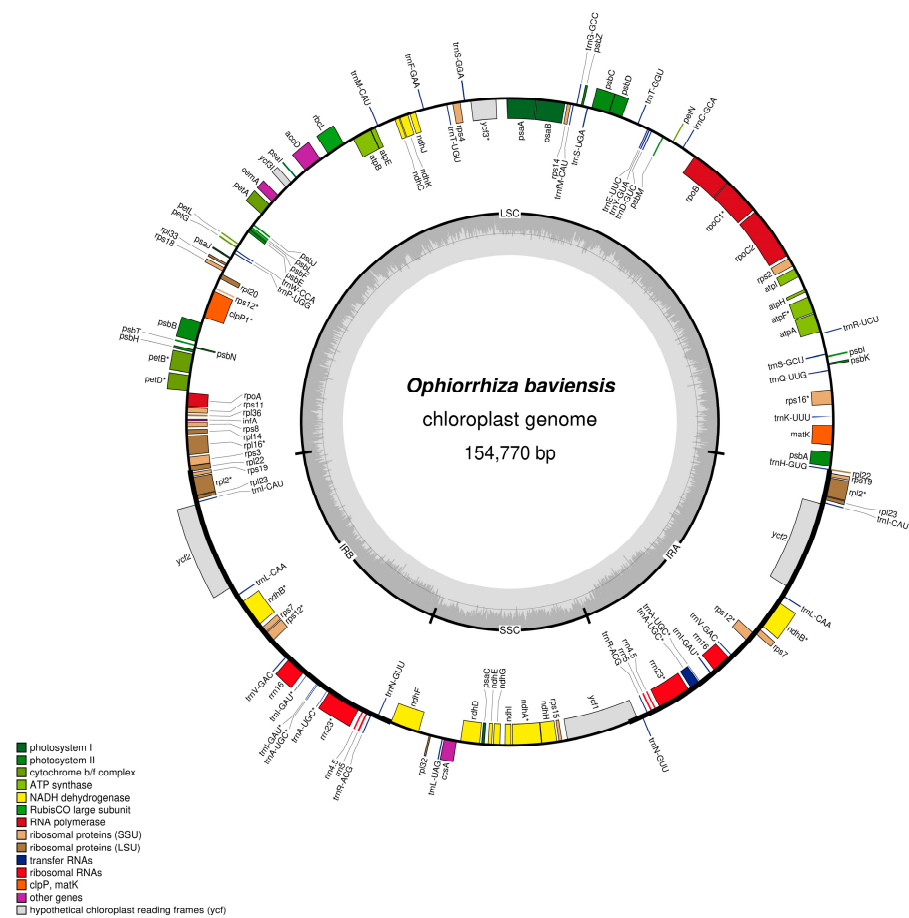


Figure 2. Chloroplast genome map of *O. baviensis* Drake in Vietnam. Genes shown inside the circle are transcribed clockwise, whereas genes outside are transcribed counterclockwise. The light gray inner circle shows the AT content, the dark gray corresponds to the GC content.

In addition, the annotation results from GeSeq and tRNAscan-SE revealed that the *O. baviensis* Drake cp genome possessed a total of 128 genes, of which there were 87 protein-coding genes, 33 tRNA genes, and 8 rRNA genes (16S, 23S, 5S, and 4.5S) (Table 1). The annotated gene models were assigned into three major groups based on their functions. Regarding the photosynthesis-related gene category, there were 44 genes encoding the subunits of ATP synthase, cytochrome complex, photosystem I and II, and putative NADPH dehydrogenase, along with the large subunit of Rubisco related to the photosynthetic electron transport chain. The other 76 genes were functionally characterized in the transcription and translation processes. The majority were tRNA genes, and the others were rRNA genes and genes encoding DNA-dependent RNA polymerase, the subunits of the ribosome, and ribosome proteins. The remaining nine genes were classified in the category of other genes, consisting of five genes with reported functions in RNA processing (*matK*),

c-type cytochrome synthesis (*ccsA*), fatty acid synthesis (*accD*), carbon metabolism (*cemA*), and proteolysis (*clpP*). In addition, four genes encoding the conserved reading frames (*ycf1*, *ycf2*, and *ycf3*) were also annotated in the cp genome.

Table 1. Summary of the chloroplast genome of *O. baviensis* Drake species.

Genome Features	<i>O. baviensis</i> Drake
Genome size (bp)	154,770 bp
LSC size (bp)	84,826
SSC size (bp)	18,574
IR size (bp)	25,685
GC content (%)	37.6
No. of genes	128
No. of PCGs	87
No. of tRNA	33
No. of rRNA	8

Otherwise, each IR region of the *O. baviensis* cp genome was annotated to comprise 18 genes (all 4 rRNA genes, 7 tRNA genes, 1 NADH-dehydrogenase protein-coding gene, 4 ribosomal protein-coding genes, and 2 other genes). There were 17 cp genes that harbored introns, among which 15 genes (*atpF*, *rpl2* ($\times 2$), *rpl16*, *ndhA*, *ndhB* ($\times 2$), *rpoC1*, *rps12*, *rps16*, *trnA-UGC* ($\times 2$), *trnG-GCC*, and *trnI-GAU* ($\times 2$)) contained a single intron, while two genes (*ycf3*, *clpP*) had double introns (Table 2).

Table 2. Gene composition of *O. baviensis* Drake chloroplast genome.

Category of Genes	Group of Genes	Name of Genes
Photosynthesis	Subunits of ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpFa</i> , <i>atpH</i> , <i>atpI</i>
	Subunits of NADH-dehydrogenase	<i>ndhAa</i> , <i>ndhB</i> ($\times 2$) <i>a</i> , <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Subunits of cytochrome b/f complex	<i>petL</i> , <i>petB</i> , <i>petG</i> , <i>petA</i> , <i>petD</i> , <i>petN</i>
	Subunits of photosystem I	<i>psaJ</i> , <i>psaC</i> , <i>psaA</i> , <i>psaI</i> , <i>psaB</i>
	Subunits of photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i>
	Subunit of rubisco	<i>rbcL</i>
	Large subunit of ribosome	<i>rpl14</i> , <i>rpl16a</i> , <i>rpl2</i> ($\times 2$) <i>a</i> , <i>rpl20</i> , <i>rpl22</i> , <i>rpl23</i> ($\times 2$), <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
Transcription and translation	DNA-dependent RNA polymerase	<i>rpoB</i> , <i>rpoA</i> , <i>rpoC1a</i> , <i>rpoC2</i>
	Small subunit of ribosomal proteins	<i>rps11</i> , <i>rps12</i> ($\times 2$) <i>a</i> , <i>rps14</i> , <i>rps15</i> , <i>rps16a</i> , <i>rps18</i> , <i>rps19</i> ($\times 2$), <i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> ($\times 2$), <i>rps8</i>
	rRNA genes	<i>rrn23S</i> ($\times 2$), <i>rrn16S</i> ($\times 2$), <i>rrn5S</i> ($\times 2$), <i>rrn4.5S</i> ($\times 2$)
	tRNA genes	<i>trnA-UGC</i> ($\times 2$) <i>a</i> , <i>trnC-GCA</i> , <i>trnD-GUC</i> , <i>trnE-UUC</i> , <i>trnF-GAA</i> , <i>trnG-GCCa</i> , <i>trnH-GUG</i> , <i>trnI-GAU</i> ($\times 2$) <i>a</i> , <i>trnL-CAA</i> ($\times 2$), <i>trnL-UAG</i> , <i>trnN-GUU</i> ($\times 2$), <i>trnP-UGG</i> , <i>trnQ-UUG</i> , <i>trnR-ACG</i> ($\times 2$), <i>trnR-UCU</i> , <i>trnS-GCU</i> , <i>trnS-GGA</i> , <i>trnS-UGA</i> , <i>trnT-GGU</i> , <i>trnT-UGU</i> , <i>trnV-GAC</i> ($\times 2$), <i>trnW-CCA</i> , <i>trnY-GUA</i>
Other genes	Translational initiation factor	<i>infA</i>
	Subunit of acetyl-CoA-carboxylase (fatty acid synthesis)	<i>accD</i>
	c-type cytochrome synthesis gene	<i>ccsA</i>
	Envelope membrane protein (carbon metabolism)	<i>cemA</i>
	Protease	<i>clpPb</i>
	Maturase (RNA processing)	<i>matK</i>
Conserved open reading frames	<i>ycf1</i> , <i>ycf2</i> ($\times 2$), <i>ycf3b</i>	

Genes marked with the sign are the gene with a single (a) or double (b) introns and duplicated genes ($\times 2$).

3.2. Repeat Sequences and Codon Analysis

A total of 59 simple sequence repeats (SSRs) were investigated in the *O. baviensis* Drake chloroplast genome via the MISA web-tool. Almost all of the screened repeats were mono repeats (composed of A/T and C) with the size ranging from 10 to 16 bp (Figure 3A). Two di-, five tri-, seven tetra-, and three penta-nucleotide SSRs were found in the *O. baviensis* Drake plastid. A total of 53 SSRs were classified as simple based SSRs and the six remaining SSRs presented in a compound formation. The majority of SSR types were discovered in the LSC, while the IR regions included the smallest number of SSRs (Figure 3B).

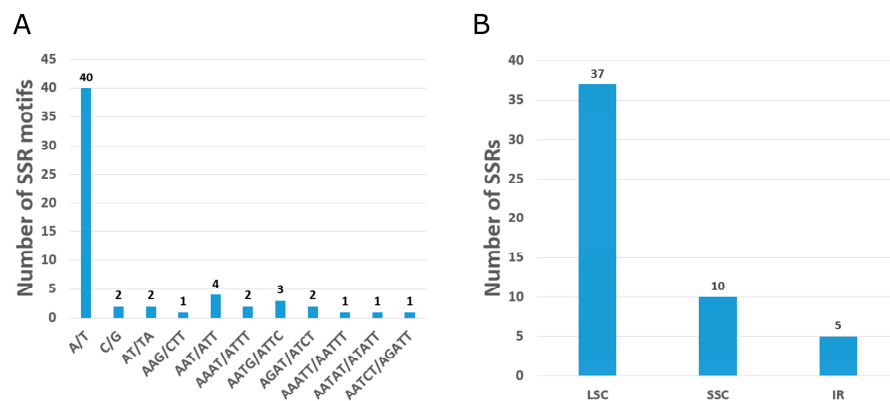


Figure 3. Analysis of single sequence repeats (SSRs) of the *O. baviensis* Drake chloroplast genome. (A) Number of identified SSR sequence motifs; (B) Frequency of repeat types in LSC, SSC, and IR regions.

The cp genome of *O. baviensis* Drake was annotated to possess 49 long repeats including 9 palindromic repeats, along with 12 forward and 22 reverse repeats. There was only one complement repeat (Figure 4). The unit size of the detected repeats ranged from 20 to 58 bp, while a majority of the repeat size (67%) was shorter than 30 bp.

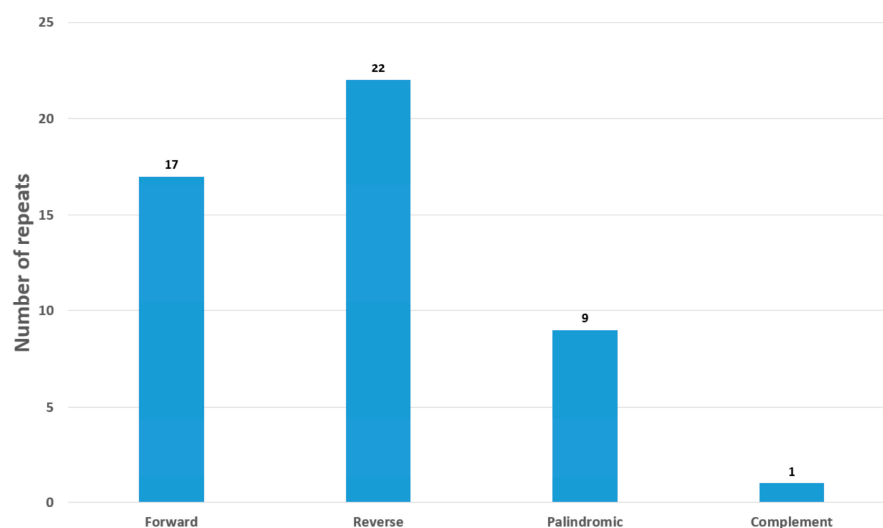


Figure 4. Repeat analysis of *O. baviensis* Drake chloroplast genome.

The codon usage frequency of 64 protein-coding genes was evaluated for three cp genomes: *O. baviensis* Drake and two other available *Ophiorrhiza* species. The total number of codons found in the coding regions was 51,517, while the A- and U-ending were found more frequently than the G/C-ending (Table 3). Leucine was the most prevalent among the 20 amino acids with a percentage of 10.46% (5068 codons), followed by serine with 9.95% (4817 codons). Meanwhile, the rarest was tryptophan with a total of 681 codons accounting

for approximately 1.4%. A total of 30 codons exhibited the codon usage bias ($RSCU < 1$), while 32 codons were observed to be more frequent than the expected usage at equilibrium ($RSCU > 1$) (Table 3). The usage frequency for the start codons AUG and UGG (methionine and tryptophan) exhibited no bias ($RSCU = 1$).

Table 3. Relative synonymous codon usage (RSCU) for protein-coding genes in *O. baviensis*.

Codon	AA	Frequency	RCSU	Codon	AA	Frequency	RCSU	Codon	AA	Frequency	RCSU
UAA	*	1259	1.22	AUC	I	1205	0.80	CGG	R	420	0.75
UAG	*	825	0.80	AUA	I	1471	0.98	AGA	R	1093	1.94
UGA	*	1004	0.98	AAA	K	2050	1.35	AGG	R	627	1.12
GCU	A	446	1.23	AAG	K	982	0.65	UCU	S	1113	1.39
GCC	A	351	0.97	UUA	L	1040	1.23	UCC	S	982	1.22
GCA	A	401	1.11	UUG	L	1095	1.30	UCA	S	824	1.03
GCG	A	250	0.69	CUU	L	1063	1.26	UCG	S	622	0.77
UGU	C	679	1.20	CUC	L	653	0.77	AGU	S	747	0.93
UGC	C	449	0.80	CUA	L	737	0.87	AGC	S	529	0.66
GAU	D	1012	1.42	CUG	L	480	0.57	ACU	T	668	1.13
GAC	D	413	0.58	AUG	M	856	1.00	ACC	T	651	1.10
GAA	E	1337	1.43	AAU	N	1779	1.38	ACA	T	647	1.09
GAG	E	537	0.57	AAC	N	800	0.62	ACG	T	406	0.68
UUU	F	2212	1.20	CCU	P	611	1.03	GUU	V	784	1.38
UUC	F	1481	0.80	CCC	P	618	1.04	GUC	V	411	0.72
GGU	G	540	0.96	CCA	P	726	1.23	GUA	V	682	1.20
GGC	G	383	0.68	CCG	P	414	0.70	GUG	V	402	0.71
GGA	G	747	1.33	CAA	Q	987	1.40	UGG	W	681	1.00
GGG	G	577	1.03	CAG	Q	420	0.60	UAU	Y	1345	1.36
CAU	H	880	1.36	CGU	R	376	0.67	UAC	Y	637	0.64
CAC	H	414	0.64	CGC	R	280	0.50				
AUU	I	1830	1.22	CGA	R	576	1.02				

* Stop codon.

3.3. Chloroplast Genome Comparison

To characterize genomic divergence, the percentage of sequence identity was evaluated for three *Ophiorrhiza* species with the functional annotation of *O. baviensis* Drake as a reference. The comparison using the mVISTA program revealed that the gene organization among the three species was highly similar and there were several regions of sequence variation (Figure 5). The results exhibited a higher frequency of divergence in the LSC and SSC regions than in the IR regions. Moreover, the coding regions of the three cp genomes were observed to be more conserved, whereas a majority of the detected variations were screened in the conserved non-coding sequences (CNS). Among the protein-coding gene sequences, the highly disparate genes consisted of *matK*, *rpoC2*, *rpoB*, *clpP*, *rpl16*, *ndhF*, *ndhA*, and *ycf1*.

The sliding window analysis indicated that the average polymorphism information (Pi) values of the LSC (Pi = 0.005635) and SSC (Pi = 0.007472) regions were greater than that of the IR (Pi = 0.001285) regions, which showed that most of the variations were located in the LSC and SSC regions (Figure 6). Of the three *Ophiorrhiza* species, the average value of nucleotide diversity (Pi) was 0.00441.

3.4. IR Contraction and Expansion in the Chloroplast Genome

The IR/LSC and IR/SSC boundaries of three *Ophiorrhiza* cp genomes were compared using the IRscope program. Overall, the results indicated that the region size, gene organization, and gene content showed a high similarity among the cp genome of *O. baviensis* and *O. pumila* (Figure 7). On the other hand, the *O. densa* cp genome showed several variants with the two abovementioned *Ophiorrhiza* species. The size of IR regions ranged from 25,684 bp (*O. baviensis* Drake) to 26,066 bp (*O. pumila*), and the size of IR of *O. densa* was 25,701 bp. The *rpl22* gene was located within the LSC region with a 102 bp overlap with

the IRb for *O. baviensis* and *O. pumila*, while *O. densa* showed a 347 bp overlap of the *rps3* gene in this boundary. Apart from *O. densa*, the *ndhF* gene was detected on the boundary of the SSC and IRb region. The border across IRa and SSC was found in the *ycf1* gene with 1438, 1316, and 730 bp tail sections of the gene placed in the IRa of *O. densa*, *O. pumila*, and *O. baviensis*, respectively (Figure 7). The IRa and LSC boundary showed the presence of the *trnH* gene in the forward strand of all three species and the *rpl22* gene in the reverse strand of *O. baviensis* and *O. densa*. The results of the IR analysis indicated extensive contraction and expansion of the IR regions in the three species.

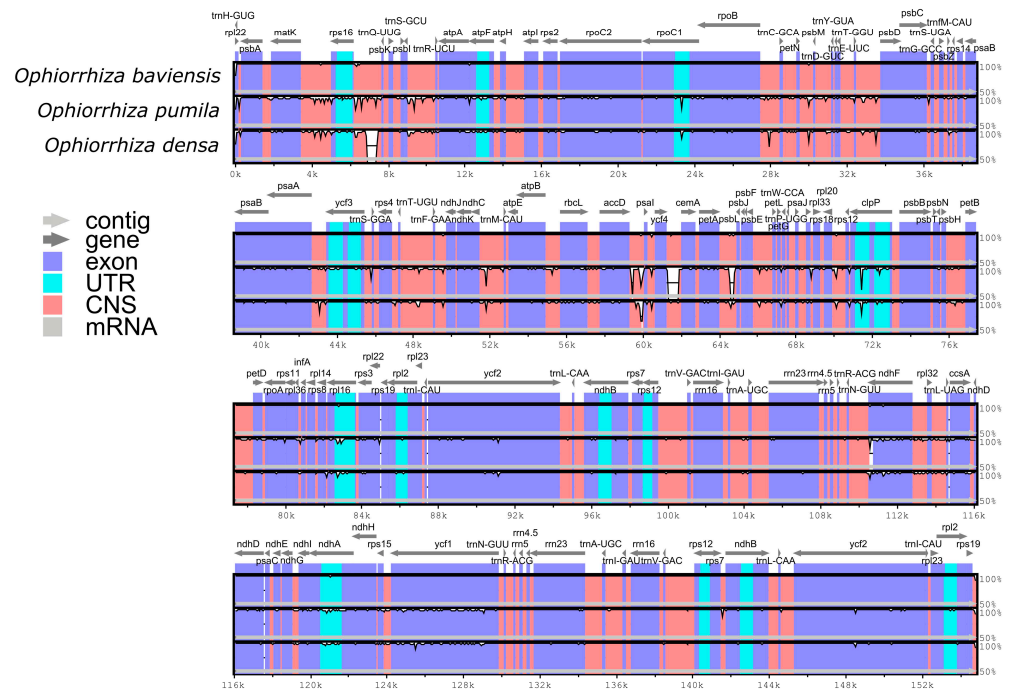


Figure 5. Complete chloroplast genome alignments of the three *Ophiorrhiza* species. The horizontal axis indicates the coordinates within the chloroplast genome. The vertical scale indicates the percent identity within 50–100%. Annotated genes are displayed along the top.

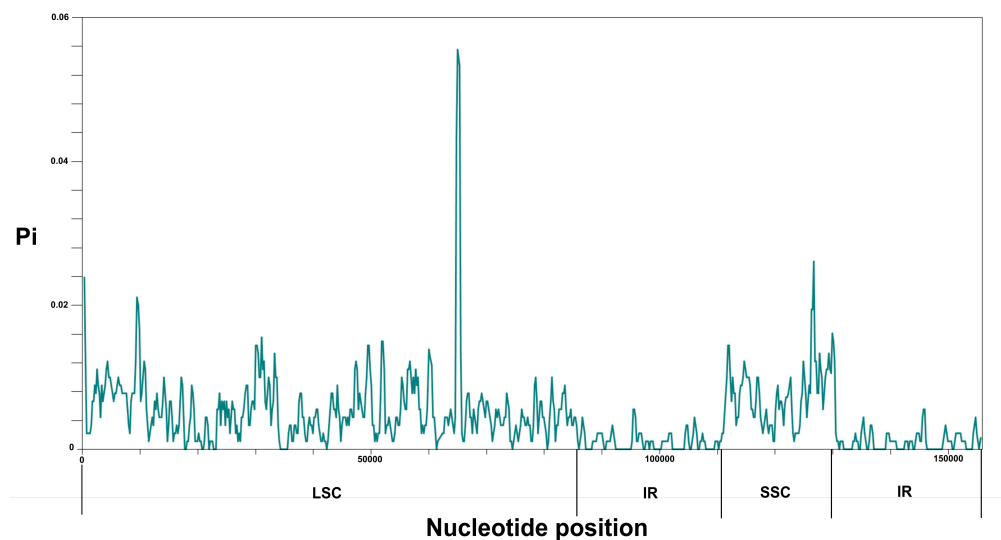


Figure 6. Nucleotide diversity (Pi) values among the three *Ophiorrhiza* species. X-axis: the position in the genome; Y-axis: Pi value. Pi, polymorphism information.

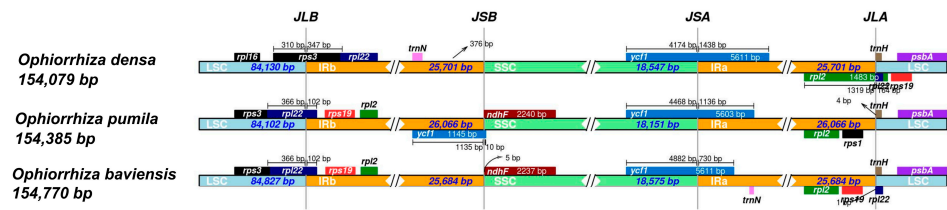


Figure 7. Comparison of LSC, IR, and SSC junction positions among the three *Ophiorrhiza* chloroplast genomes. JLB (junction IRb/LSC), JSB (junction IRb/SSC), JSA (junction IRa/SSC), JLA (junction IRa/LSC).

3.5. Phylogenetic Inference

The number of available sequences of *O. baviensis* on the Genbank databases, especially belonging to the cp genome, is limited (only the *rps16* gene and the *trnL-trnF* intergenic spacer). Therefore, we extracted these sequences from the assembled cp genome and used them to access the phylogenetic relationship of the studied *O. baviensis* at the species level. Figure 8 shows the phylogenetic resolution based on the concatenated sequence between the *rps16* gene and the *trnL-trnF* intergenic spacer with a high bootstrap value of 92% between the studied *O. baviensis* Drake and the reference *O. baviensis* voucher Averyanov & al. VH940 (AAU) (Accession number: MH626923.1). With a bootstrap value of 100%, all eight *Ophiorrhiza* species were grouped separately from the two *Xanthophytum* species as an outgroup. In the case of barcoding among the Rubiaceae family, the combined *rps16-trnL-F* intergenic spacer sequences provided a high capacity for phylogenetic resolution.

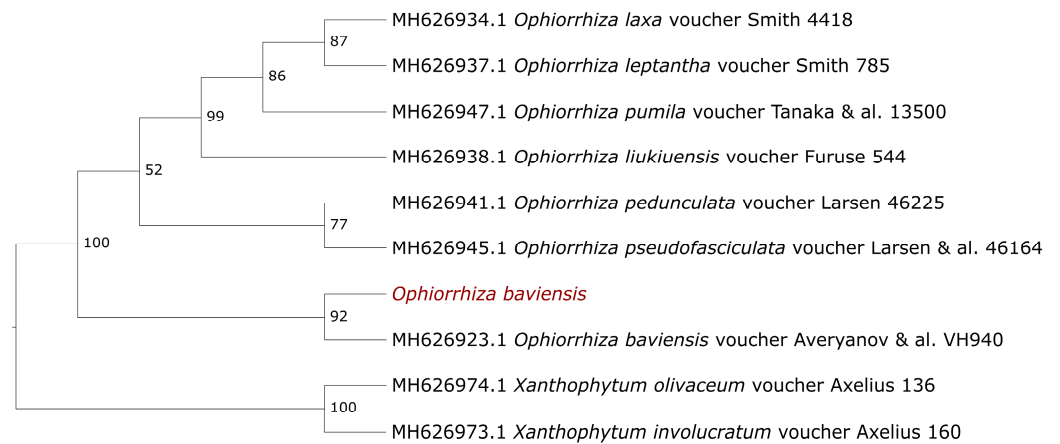


Figure 8. The maximum likelihood phylogenetic trees based on the concatenated sequences between the *rps16* genes and the *trnL-trnF* intergenic spacer. Numbers on the branches indicate bootstrap percentage after 1000 replications in constructing the tree. The species investigated in this study are colored in red.

4. Discussion

Rubiaceae is a family of flowering plants containing 620 genera with approximately 13,500 species over the world, which makes it the fourth-largest angiosperm family. Over 300 cp genomes in the Rubiaceae family have been published in the Genbank database until now, only three of which belong to *Ophiorrhiza*. The genus *Ophiorrhiza* consists of about 200–300 species mainly distributed in humid tropical forests from East India to the Western Pacific, and from South China to Northern Australia [16,17]. Bioactive compounds from this family, such as quinine, emetine, caffeine, and camptothecin are of major pharmaceutical importance; thus, many species in the genus *Ophiorrhiza* are of interest [18]. In the present study, we sequenced and annotated the entire cp genome of a Vietnamese medicinal plant.

Angiosperm cp genomes have a highly conserved gene order and gene content with 127–134 genes found across the chloroplast genomes. The analyzed *O. baviensis* cp genomes

demonstrated the typical quadripartite structure and showed the expected size range (~154 kb) for *Ophiorrhiza* species and the conserved gene contents. Our gene annotation results were similar to the genetic properties of angiosperm chloroplast genomes. The number of genes present in the cp genome from *O. baviensis* was 128, of which, 17 genes included one or two introns. In addition, the deletions of the *petB* and *petD* introns were observed in the studied *O. baviensis* cp genome, which also occurred in *O. pumila* species. Introns play an important role in gene expression regulation. Recent research has revealed gene or intron loss in chloroplast genomes [19–21], among which *petB* and *petD* intron loss was reported in many angiosperms [22].

In addition to two copies of IR regions, 49 small repeats were found to be located within coding and non-coding regions of the *O. baviensis* plastome. The cp genome includes numerous dispersed repeats, which are supposed to be biomarkers of mutational hotspots [23,24]. The repeat number is similar to the data of other species belonging to the Rubiaceae family [25,26]. Repeats are closely related to angiosperm plastome reconstruction and can be assumed as recognition signals of recombination because of their potential to generate secondary structures. In this study, the similar number of repeats in comparison with previous estimates might not demonstrate inter- and intra-specific plastome recombination. In higher plants, SSRs are identified as crucial molecular markers for the investigations of population variation due to their distinct uniparental inheritance, and they are commonly used to evaluate genetic diversity and population structure in evolutionary studies [27–29]. In total, 59 SSRs were screened in the *O. baviensis* cp genomes with strong A/T bias. These repeats play a significant role for generating genetic markers in *O. baviensis* species, which may be applied to assess the variation at the intraspecific level in phylogenetic and ecological studies.

Comparative analyses on *O. baviensis* and two available *Ophiorrhiza* cp genomes were implemented to explore the plastome structure in the taxa. The cp genome size of the three *Ophiorrhiza* ranged from 154,079 bp (*O. densa*) to 154,770 bp (*O. baviensis*), the figure for *O. pumila* was 154,385 bp. Gene organization and codon usage patterns exhibited high conservation, which could be applicable for further population genetics and phylogenetic studies. Moreover, the three *Ophiorrhiza* cp genomes were less variable in their coding regions than in their noncoding regions, which is consistent with the common pattern in most angiosperms [30] (Figure 5). Codon usage preference is closely related to gene expression and can affect the level of mRNA and proteins in the genome [31–33]. The most prevalent amino acid in the *Ophiorrhiza* was leucine (Leu), which has also been commonly detected in the other angiosperms. The high similarity in codon usage may indicate that these *Ophiorrhiza* species underwent similar environmental pressure through their evolutionary processes. The *Ophiorrhiza* cp genomes indicated that the RSCU values of most codons ending in A/U were greater than 1, which may be caused by a bias toward a high A/T ratio in composition. Additionally, we investigated that the partial sequences of the *ycf1* gene along with five intergenic spacers (IGSs), including *petA-psbJ*, *trnH-GUG-psbA*, *trnS-GCU-trnR-UCU*, *psbM-trnD-GUC*, and *ndhC-trnM-CAU*, had relatively high nucleotide diversity values ($P_i > 0.015$). These divergence regions could be studied to provide molecular markers for DNA barcoding and phylogenetic research in *Ophiorrhiza*.

While the three plastomes showed an approximate similarity in genome size, the size of the structural regions exhibited significant differences in a detailed comparison of junction sites (Figure 7). The regions of the cp genome frequently undergo length variations during the evolution of terrestrial plants, which leads to the emergence of many boundary features [34]. The expansion and contraction of the boundaries between IRs and the single-copy (SC) regions are the primary causes of the size change in cp genomes and influence the evolution rate of cp genomes [35,36]. Our finding revealed that the boundary-gene set of the *Ophiorrhiza* species included *rpl22*, *rps19*, *ndhF*, *ycf1*, and *trnH*. Several notable gene rearrangements were observed in the *O. densa* plastome; these were the presence of the *rps3* gene at the JLB instead of the *rpl22* gene, the expansion of the *rpl2* gene to the JLA, and the absence of the *rps19* gene in the IR regions. Expansion and contraction, as well as variation,

at the junction of the SC–IR regions were characterized, suggesting that gene organization in the IR regions can report the distance between species to some extent.

The majority of taxonomic levels of plant phylogenetic connections have been demonstrated using complete chloroplast genomes and protein-coding genes [37,38]. The current study provides the phylogeny of the *Ophiorrhiza* genus based on the combined *rps16-trnL-F* intergenic spacer sequences. The previous study of Razafimandimbison and Rydin demonstrated that *O. baviensis* had been resolved as a sister relationship with *O. japonica* and *O. hayatana* [39]. In terms of species classification, the phylogenetic tree based on the concatenation of the *rps16* gene and the *trnL-F* intergenic spacer indicated the close relationship between the studied plant and the *O. baviensis* voucher Averyanov & al. VH940 (AAU) with a high bootstrap value of 92%. This approach showed effectiveness in the classification of the lower taxonomic levels among the Rubiaceae family. Further, the combination of these barcodes can lead to better species classification compared to the results from a single gene [39]. This study will help to clarify the evolutionary position of *O. baviensis* in the *Ophiorrhiza* genus, as well as offering applicable cp genome data for further research into the genesis and diversification of the Rubiaceae family. Overall, our phylogenetic investigation of the *O. baviensis* cp genome was successful in discovering the intragenetic connections within the *Ophiorrhiza* genus.

5. Conclusions

In this study, the first complete chloroplast genome of an *O. baviensis* Drake species from Vietnam was characterized and compared with two other published *Ophiorrhiza* plastomes. The assembly resulted in a whole cp genome of 154,770 bp in size. According to the comparative result, the structure and gene content of three *Ophiorrhiza* cp genomes exhibited a high similarity, and the SC-IR junction analysis revealed the expansion and contraction of IR regions. Additionally, the phylogenetic tree indicated close relationships between our novel cp genome sequence and other *Ophiorrhiza* species. This study provides the potential to employ cp genomes for enhancing species classification and genetic source conservation during further study of the Rubiaceae family.

Author Contributions: Conceptualization, H.H.C.; sampling, T.H.T.; methodology, T.H.T., H.H. and T.D.L.; software, M.H.P.; validation, T.L.L.; formal analysis, M.H.P.; data curation, T.L.L.; writing—original draft preparation, M.H.P.; writing—review and editing, H.H.C.; visualization, M.H.P.; supervision, H.H.C. and T.H.T.; project administration, H.H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project of the Vietnam Academy of Science and Technology (VAST): “Sequencing and characterizing the chloroplast genome of an *Ophiorrhiza baviensis* species by PacBio SMRT next-generation sequencing technology for genetic classification and conservation” (project no. CSCL08.02/22-22).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This complete chloroplast genome of *O. Baviensis* Drake has been deposited at DDBJ/ENA/GenBank under the accession number OP902221.

Acknowledgments: We thank Khang Sinh Nguyen, researcher at the Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology for authenticating the taxonomic identification of the plant samples.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Neuhaus, H.E.; Emes, M.J. Nonphotosynthetic Metabolism in Plastids. *Annu. Rev. Plant Biol.* **2000**, *51*, 111. [[CrossRef](#)]
2. Bendich, A.J. Circular Chloroplast Chromosomes: The Grand Illusion. *Plant Cell* **2004**, *16*, 1661–1666. [[CrossRef](#)]
3. Lei, W.; Liu, W.-J.; Nguyen, K.S. Revision of Three Taxa of *Ophiorrhiza* (Rubiaceae) from China. *Phytotaxa* **2019**, *387*, 129–139. [[CrossRef](#)]

4. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289. [[CrossRef](#)] [[PubMed](#)]
5. PacificBiosciences. Pbbmm2: A Minimap2 Frontend for PacBio Native Data Formats. Available online: <https://github.com/PacificBiosciences/pbbmm2> (accessed on 10 January 2021).
6. Chin, C.-S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; et al. Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data. *Nat. Methods* **2013**, *10*, 563–569. [[CrossRef](#)]
7. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq—Versatile and Accurate Annotation of Organelle Genomes. *Nucleic Acids Res.* **2017**, *45*, W6–W11. [[CrossRef](#)] [[PubMed](#)]
8. Chan, P.P.; Lin, B.Y.; Mak, A.J.; Lowe, T.M. TRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. *Nucleic Acids Res.* **2021**, *49*, 9077–9096. [[CrossRef](#)] [[PubMed](#)]
9. Lohse, M.; Drechsel, O.; Bock, R. OrganellarGenomeDRAW (OGDRAW): A Tool for the Easy Generation of High-Quality Custom Graphical Maps of Plastid and Mitochondrial Genomes. *Curr. Genet.* **2007**, *52*, 267–274. [[CrossRef](#)] [[PubMed](#)]
10. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-Web: A Web Server for Microsatellite Prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [[CrossRef](#)] [[PubMed](#)]
11. Kurtz, S.; Schleiermacher, C. REPuter: Fast Computation of Maximal Repeats in Complete Genomes. *Bioinformatics* **1999**, *15*, 426–427. [[CrossRef](#)]
12. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. VISTA: Computational Tools for Comparative Genomics. *Nucleic Acids Res.* **2004**, *32*, W273–W279. [[CrossRef](#)]
13. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [[CrossRef](#)] [[PubMed](#)]
14. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization. *Brief. Bioinform.* **2019**, *20*, 1160–1166. [[CrossRef](#)] [[PubMed](#)]
15. Price, M.; Dehal, P.; Arkin, A. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **2010**, *5*, e9490. [[CrossRef](#)]
16. Chen, T.; Taylor, C. Ophiorrhiza. In *Flora of China*; Press, S., Ed.; Beijing & Missouri Botanical Garden Press: St. Louis, MO, USA, 2011; Volume 19, pp. 258–282.
17. Lei, W.; Tan, Y.; Hareesh, V.S.; Liu, Q. Ophiorrhiza Macrocarpa (Rubiaceae), a New Viviparous Species from Yunnan, Southwestern China. *Nord. J. Bot.* **2018**, *36*, njb-01637. [[CrossRef](#)]
18. Hamzah, A.S. Isolation, Characterization and Biological Activities of Chemical Constituents of Ophiorrhiza and Hedyotis Species. Ph.D. Dissertation, Universiti Pertanian Malaysia, Serdang, Malaysia, 1994.
19. GAO, L.; SU, Y.-J.; WANG, T. Plastid Genome Sequencing, Comparative Genomics, and Phylogenomics: Current Status and Prospects. *J. Syst. Evol.* **2010**, *48*, 77–93. [[CrossRef](#)]
20. Frailey, D.C.; Chaluvadi, S.R.; Vaughn, J.N.; Coatney, C.G.; Bennetzen, J.L. Gene Loss and Genome Rearrangement in the Plastids of Five Hemiparasites in the Family Orobanchaceae. *BMC Plant Biol.* **2018**, *18*, 30. [[CrossRef](#)]
21. Oyebanji, O.; Zhang, R.; Chen, S.-Y.; Yi, T.-S. New Insights Into the Plastome Evolution of the Millettoid/Phaseoloid Clade (Papilionoideae, Leguminosae). *Front. Plant Sci.* **2020**, *11*, 151. [[CrossRef](#)]
22. Li, X.; Li, Y.; Sylvester, S.P.; Zang, M.; El-Kassaby, Y.A.; Fang, Y. Evolutionary Patterns of Nucleotide Substitution Rates in Plastid Genomes of Quercus. *Ecol. Evol.* **2021**, *11*, 13401–13414. [[CrossRef](#)]
23. Abdullah; Mehmood, F.; Shahzadi, I.; Ali, Z.; Islam, M.; Naeem, M.; Mirza, B.; Lockhart, P.J.; Ahmed, I.; Waheed, M.T. Correlations among Oligonucleotide Repeats, Nucleotide Substitutions, and Insertion–Deletion Mutations in Chloroplast Genomes of Plant Family Malvaceae. *J. Syst. Evol.* **2021**, *59*, 388–402. [[CrossRef](#)]
24. Liu, Q.; Li, X.; Li, M.; Xu, W.; Schwarzacher, T.; Heslop-Harrison, J.S. Comparative Chloroplast Genome Analyses of Avena: Insights into Evolutionary Dynamics and Phylogeny. *BMC Plant Biol.* **2020**, *20*, 406. [[CrossRef](#)] [[PubMed](#)]
25. Ly, S.N.; Garavito, A.; De Block, P.; Asselman, P.; Guyeux, C.; Charr, J.-C.; Janssens, S.; Mouly, A.; Hamon, P.; Guyot, R. Chloroplast Genomes of Rubiaceae: Comparative Genomics and Molecular Phylogeny in Subfamily Ixoroideae. *PLoS ONE* **2020**, *15*, e0232295. [[CrossRef](#)] [[PubMed](#)]
26. Amenu, S.G.; Wei, N.; Wu, L.; Oyebanji, O.; Hu, G.; Zhou, Y.; Wang, Q. Phylogenomic and Comparative Analyses of Coffeaeae Alliance (Rubiaceae): Deep Insights into Phylogenetic Relationships and Plastome Evolution. *BMC Plant Biol.* **2022**, *22*, 88. [[CrossRef](#)] [[PubMed](#)]
27. Varshney, R.K.; Sigmund, R.; Börner, A.; Korzun, V.; Stein, N.; Sorrells, M.E.; Langridge, P.; Graner, A. Interspecific Transferability and Comparative Mapping of Barley EST-SSR Markers in Wheat, Rye and Rice. *Plant Sci.* **2005**, *168*, 195–202. [[CrossRef](#)]
28. Dong, W.; Liu, H.; Xu, C.; Zuo, Y.; Chen, Z.; Zhou, S. A Chloroplast Genomic Strategy for Designing Taxon Specific DNA Mini-Barcodes: A Case Study on Ginsengs. *BMC Genet.* **2014**, *15*, 138. [[CrossRef](#)]
29. Provan, J.; Powell, W.; Hollingsworth, P.M. Chloroplast Microsatellites: New Tools for Studies in Plant Ecology and Evolution. *Trends Ecol. Evol.* **2001**, *16*, 142–147. [[CrossRef](#)] [[PubMed](#)]
30. Yang, C.-H.; Liu, X.; Cui, Y.-X.; Nie, L.-P.; Lin, Y.-L.; Wei, X.-P.; Wang, Y.; Yao, H. Molecular Structure and Phylogenetic Analyses of the Complete Chloroplast Genomes of Three Original Species of Pyrrosiae Folium. *Chin. J. Nat. Med.* **2020**, *18*, 573–581. [[CrossRef](#)]
31. Zhou, M.; Guo, J.; Cha, J.; Chae, M.; Chen, S.; Barral, J.M.; Sachs, M.S.; Liu, Y. Non-Optimal Codon Usage Affects Expression, Structure and Function of Clock Protein FRQ. *Nature* **2013**, *495*, 111–115. [[CrossRef](#)]

32. Somaratne, Y.; Guan, D.-L.; Wang, W.-Q.; Zhao, L.; Xu, S.-Q. The Complete Chloroplast Genomes of Two *Lespedeza* Species: Insights into Codon Usage Bias, RNA Editing Sites, and Phylogenetic Relationships in Desmodieae (Fabaceae: Papilionoideae). *Plants* **2020**, *9*, 51. [[CrossRef](#)]
33. Lyu, X.; Liu, Y. Nonoptimal Codon Usage Is Critical for Protein Structure and Function of the Master General Amino Acid Control Regulator CPC-1. *Mbio* **2020**, *11*, e02605-20. [[CrossRef](#)]
34. Ding, S.; Dong, X.; Yang, J.; Guo, C.; Cao, B.; Guo, Y.; Hu, G. Complete Chloroplast Genome of *Clethra Fargesii* Franch., an Original Sympetalous Plant from Central China: Comparative Analysis, Adaptive Evolution, and Phylogenetic Relationships. *Forests* **2021**, *12*, 441. [[CrossRef](#)]
35. Kim, K.-J.; Lee, H.-L. Complete Chloroplast Genome Sequences from Korean Ginseng (*Panax Schinseng* Nees) and Comparative Analysis of Sequence Evolution among 17 Vascular Plants. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **2004**, *11*, 247–261. [[CrossRef](#)] [[PubMed](#)]
36. Zhang, H.; Li, C.; Miao, H.; Xiong, S. Insights from the Complete Chloroplast Genome into the Evolution of *Sesamum indicum* L. *PLoS ONE* **2013**, *8*, e80508. [[CrossRef](#)]
37. De Las Rivas, J.; Lozano, J.J.; Ortiz, A.R. Comparative Analysis of Chloroplast Genomes: Functional Annotation, Genome-Based Phylogeny, and Deduced Evolutionary Patterns. *Genome Res.* **2002**, *12*, 567–583. [[CrossRef](#)]
38. Moore, M.J.; Bell, C.D.; Soltis, P.S.; Soltis, D.E. Using Plastid Genome-Scale Data to Resolve Enigmatic Relationships among Basal Angiosperms. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19363–19368. [[CrossRef](#)] [[PubMed](#)]
39. Razafimandimbison, S.G.; Rydin, C. Molecular-Based Assessments of Tribal and Generic Limits and Relationships in Rubiaceae (Gentianales): Polyphyly of Pomazoteae and Paraphyly of Ophiorrhizeae and Ophiorrhiza. *Taxon* **2019**, *68*, 72–91. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.