Tạp chí Công nghệ Sinh học **18**(1): 87-102, 2020

XÂY DỰNG BẢN ĐỒ BỘ GEN LỤC LẠP HOÀN CHỈNH CỦA LOÀI LAN HÀI HỒNG (*Paphiopedilum delenatii* Guillaumin 1924) ĐẶC HỮU VIỆT NAM

Nguyễn Thanh Điềm[1], Lê Thị Lý[2], Nguyễn Hữu Thuần Anh[1], Nguyễn Thành Công[1], Vũ Thị Huyền Trang[1,2,✉]

[1]*Trường Đại học Nguyễn Tất Thành, thành phố Hồ Chí Minh*
[2]*Trường Đại học Quốc tế, Đại học Quốc gia thành phố Hồ Chí Minh*

✉Người chịu trách nhiệm liên lạc. E-mail: vthtrang@ntt.edu.vn

Ngày nhận bài: 22.4.2019
Ngày nhận đăng: 09.7.2019

TÓM TẮT

Lục lạp (chloroplasts) và ty thể (mitochondria) là những bào quan có bộ gen riêng so với bộ gen trong nhân tế bào. Bộ gen lục lạp cung cấp thông tin nghiên cứu về mối quan hệ tiến hóa của các loài, xác định một loài một cách chính xác, cung cấp chỉ thị ứng dụng trong chuyển gen, nhân giống… Nhờ công nghệ giải trình tự thế hệ mới mà việc giải trình tự bộ gen lục lạp dễ dàng hơn. Tuy nhiên quy trình lắp ráp bộ gen lục lạp hiện nay còn khá phức tạp do yêu cầu cần sử dụng nhiều công cụ tin sinh học khác nhau, yêu cầu máy có cấu hình cao, tốn nhiều thời gian. Trong bài viết này, chúng tôi mô tả chi tiết quy trình lắp ráp bộ gen lục lạp hoàn chỉnh của mẫu lan Hài hồng (*Paphiopedilum delenatii*) đồng thời đưa ra một số khảo sát giúp cho việc lắp ráp dễ dàng và độ tin cậy cao. Bộ gen lục lạp loài lan Hài hồng sau khi được lắp ráp có chiều dài 160.955 bp, gồm một vùng sao chép lớn (large single copy region, LSC), một vùng sao chép nhỏ (small single copy region, SSC) được phân tách bởi hai vùng lặp lại đảo ngược. Tổng số gen là 130 gen, GC content là 35,6%. Dữ liệu trình tự đã được đăng kí vào Ngân hàng gen (GenBank) với mã số MK463585. Nghiên cứu này còn đưa ra những thông số tối ưu để lắp ráp bộ gen. Kết quả nghiên cứu không chỉ đóng góp thông tin bộ gen lục lạp hỗ trợ công tác bảo tồn loài lan Hài đặc hữu của Việt Nam mà còn có ý nghĩa trong việc hỗ trợ hướng nghiên cứu lắp ráp bộ gen lục lạp, có thể áp dụng trên nhiều đối tượng khác.

Từ khóa: *Paphiopedilum delenatii*, lắp ráp bộ gen, chú thích bộ gen, bản đồ bộ gen, bộ gen lục lạp

GIỚI THIỆU

Bộ gen lục lạp đã được nghiên cứu rộng rãi trên thực vật. Thông tin bộ gen lục lạp không chỉ được sử dụng trong nghiên cứu nhận diện loài, xác định mối quan hệ giữa các loài, tìm hiểu tiến hóa phân tử mà còn phục vụ việc chuyển gen, nhân giống và thuần hóa cây trồng (Daniell *et al.*, 2016; Xiang *et al.*, 2016; Yeisoo *et al.*, 2017). Việc giải trình tự bộ gen lục lạp gặp nhiều khó khăn khi áp dụng kỹ thuật giải trình tự Sanger (Sanger sequencing), do chỉ thu được các đoạn trình tự ngắn. Tuy nhiên nhờ sự ra đời của công nghệ giải trình tự thế hệ mới (Next Generation Sequencing – NGS) với khả năng xử lí khối lượng dữ liệu khổng lồ với tốc độ nhanh và chi phí giải trình tự ngày càng giảm (Shendure, Ji, 2008) mà việc giải trình tự toàn bộ hệ gen của một loài sinh vật ngày càng phổ biến. Từ đó càng có nhiều công trình nghiên cứu về bộ gen lục lạp được công bố. Tian và đồng tác giả (2018) đã giải và phân tích bộ gen lục lạp của loài *Epipremum aureum.* Các thông tin từ bộ gen lục lạp đã góp phần đáng kể (hoặc không nhỏ) vào việc nhân giống và hỗ trợ chuyển gen của loại cây thuốc này (Tian *et al.*,

2018). Guo và đồng tác giả (2017) đã giải thành công bộ gen lục lạp của loài *Paeonia ostii* giúp tăng năng suất của loại dược liệu này (Guo *et al.*, 2018). Đối tượng Sâm Ngọc Linh, loài nhân sâm quý đặc trưng của Việt Nam cũng đã được giải mã trình tự bộ gen lục lạp dựa trên 4 mẫu loài (02 *Panax vietnamensis*, 01 *P. bipinnatifidus*, 01 *P. stipuleanatus* vào năm 2018 và từ đó phân tích được sự phát sinh chủng loài và xác định được 4 chỉ thị tiềm năng làm mã vạch phân tử cho phân loại nhóm đối tượng này (Manzanilla *et al.*, 2018).

Mặc dù công nghệ NGS đã cải thiện công việc giải trình tự bộ gen lục lạp, các quy trình lắp ráp bộ gen lục lạp còn khá phức tạp và những công trình mô tả một cách chi tiết quy trình này còn hạn chế. Công trình về quy trình lắp ráp bộ gen lục lạp điển hình trên thế giới như: Dự án lắp ráp bộ gen lục lạp từ trình tự DNA tổng số dựa trên tần số K-mer (Izan *et al.*, 2017) của Izan (2017). Dự án này đã đưa ra một quy trình được mô tả chi tiết để lắp ráp bộ gen lục lạp. Riêng ở Việt Nam thì những công bố về giải trình bộ gen lục lạp còn hạn chế. Năm 2015, Huỳnh Phước Hải và công sự đã đưa ra quy trình lắp ráp bộ gen lục lạp theo phương pháp không sử dụng bộ gen tham chiếu và thực nghiệm thành công một số tập dữ liệu như *Arabidopsis thaliana, Oryzasativa indica*, *Sorghum bicolor* từ cơ sở dữ liệu ENA LECA (Huỳnh Phước Hải, Nguyễn Văn Hòa, 2015).

Hiện nay, dữ liệu bộ gen lục lạp được công bố trên GenBank ngày càng nhiều nên có thể dựa trên những trình tự này để lắp ráp bộ gen một cách nhanh chóng, dễ dàng và có độ tin cậy cao. Đây là phương pháp lắp ráp dựa theo trình tự mẫu (homologus modeling). NOVOPlasty là một trong các chương trình chính để thực hiện công việc này. So với các chương trình CLC, SOAPdenovo2, MIRA, MITObim, NOVOPlasty đã được công nhận là có độ chính xác cao, tiết kiệm dung lượng máy và thời gian (Nicolas *et al.*, 2017). Chương trình này đã được áp dụng trong nhiều nghiên cứu như nghiên cứu giải trình tự bộ gen lục lạp *Fagus crenata* của Worth và Liu (2019) (Worth, Liu, 2019), nghiên cứu giải trình tự bộ gen lục lạp *Ailanthus altissima* của Saina và đồng tác giả (2018) (Saina *et al.*, 2018)… Tuy nhiên, những nghiên cứu này không chú trọng việc mô tả cụ thể quy trình. Vì vậy trong nghiên cứu này chúng tôi mô tả chi tiết quy trình lắp ráp và chú thích bộ gen lục lạp hoàn chỉnh đơn giản có thể thực hiện trên máy tính cá nhân với thời gian ngắn và cho kết quả chính xác. Đối tượng thực hiện là loài lan Hài hồng (*Paphiopedilum delenatii*) đặc hữu của Việt Nam được xếp vào loại Cực kỳ nguy cấp (Critically Endangerd – CR) (IUCN, 2018).

VẬT LIỆU VÀ PHƯƠNG PHÁP

Vật liệu

Mẫu lá lan Hài hồng *Paphiopedilum delenatii* được cung cấp và định danh hình thái dựa trên cây có hoa bởi Viện Nghiên cứu Khoa học Tây Nguyên (Đà Lạt).

Tách DNA tổng số

Mẫu lá được thu và rửa sạch bằng cồn 70^{o}. DNA tổng số được tách bằng phương pháp SDS. Thành phần đệm chiết cho tách thủ công gồm 100 mM Tris-HCl, 100 mM EDTA, 250 mM NaCl) với 20% SDS (Ahmed *et al.*, 2009). Mẫu lá được nghiền với 5 µL proteinase K và 3 mL hỗn hợp gồm (9 µL beta-mercaptoethanol và 3 mL dung dịch đệm chiết) ở 65^{o}C, sau đó mẫu được ủ thêm 30 phút ở 65^{o}C để phá vỡ màng tế bào và màng nhân. Protein được biến tính và loại bỏ bằng cách thêm 600 µL hỗn hợp dung dịch phenol: chloroform: isoamine (25:24:1) rồi ly tâm 10000 rpm trong10 phút để thu pha chứa DNA (Ahmed *et al.*, 2009). Ngoài ra tăng độ tinh sạch mẫu, 5 µL RNAse được thêm vào sau đó rồi ủ ở 37^{o}C để loại bỏ RNA đồng thời biến tính protein lần 2 bằng 600 µL hỗn hợp dung dịch chloroform:isoaminetỉ lệ 24:1. DNA được kết tủa bằng dung dịch isopropanol, ủ qua đêm ở –20^{o}C. Ly tâm để thu tủa rồi rửa tủa lần lượt bằng ethanol 70%, 80%, 90%. DNA được bảo quản ở –20^{o}C trong dung dịch TE.

Kiểm tra chất lượng DNA

Chất lượng DNA tổng số cho giải trình tự

NGS cần đạt độ tinh sạch cao tương ứng với $OD_{260/280}$ từ 1,8 - 2,2, không bị nhiễm RNA, DNA ít bị đứt gãy và nồng độ cần trên 20 ng/µL, lượng mẫu ≥300 ng, thể tích mẫu DNA trong EB buffer ≥10µL theo yêu cầu của Công ty GENEWIZ (South Plainfield, NJ, USA).

Độ tinh sạch được kiểm tra bằng máy đo quang phổ NanoDrop 2000 ở các bước sóng 260 và 280. Tính nguyên vẹn và nồng độ của DNA được kiểm tra bằng phương pháp điện di trên gel agarose 0.8% trong dung dịch 50 mL TBE 0,5X rồi soi dưới đèn huỳnh quang, nếu băng sáng đậm, dày, gọn, không bị vệt dài, nằm ở vị trí trên 10 kb thì thể hiện DNA tổng số có nồng độ cao và ít bị đứt gãy. Nồng độ DNA cũng được kiểm tra bằng cả máy đo quang phổ Nanodrop 2000 (Thermo Fisher Scientific Inc.) ở các bước sóng 260 và 280 và máy Quantus E6150 (Promega Inc.). Mẫu DNA tổng số đạt yêu cầu được gửi giải trình tự tại công ty GENEWIZ (South Plainfield, NJ, USA) bằng kỹ thuật Illumina HiSeq.

Kiểm tra chất lượng trình tự thô và lọc bỏ các đoạn trình tự có chất lượng thấp

Chất lượng tín hiệu của dữ liệu trình tự thô được kiểm tra bằng chương trình FastQC version 0.11.8 (Andrews, 2010). Ngưỡng chất lượng cho độ tin cậy cao khi lắp ráp genome được khảo sát dựa theo nhiều chỉ tiêu đánh giá gồm "Per sequence quality scores" (điểm chất lượng trên số lượng trình tự), "Per base sequence quality" (điểm chất lượng trên từng vị trí nucleotide), "Per base N content" (tỉ lệ trình tự chứa base N) và "Adapter content" (tỉ lệ trình tự còn chứa Adapter). Những trình tự có điểm chất lượng dưới ngưỡng mong muốn, những trình tự có tỉ lệ N trên 10% và những trình tự còn Adapter được loại bỏ khỏi dữ liệu bằng phần mềm Prinseq (Schmieder, Edwards, 2011).

Lắp ráp trình tự bộ gen

Chương trình NOVOPlasty 2.7.2 (Nicolas *et al.*, 2017) được vận hành trên nền hệ điều hành Ubuntu 18.04 thuê trên máy chủ Google Cloud Platform 16 GB RAM để lắp ráp các đoạn trình tự thô (read) thành các contig, đến lượt các contig lại tiếp tục được lắp ráp để thành trình tự bộ gen hoàn chỉnh. Genome range (khoảng ước lượng chiều dài của bộ gen) được thiết lập là 150000 – 170000 bp (căn cứ theo chiều dài các genome tham khảo - Bảng 1). Read length (chiều dài của các trình tự thô) được thiết lập là 150 bp dựa vào kết quả thống kê chiều dài các trình tự thô (read) (Hình 2B).

Bảng 1. Chiều dài bộ gen lục lạp hoàn chỉnh của một số loài lan Hài tham khảo từ NCBI (https://www.ncbi.nlm.nih.gov/nucleotide).

P. armeniacum (KT388109.1)	162,682 bp
P. niveum (NC_026776.1)	159,108 bp
P. dianthum (NC_036958.1)	154,699 bp

Các thông số cần được thiết lập khác bao gồm ngưỡng trình tự đạt chất lượng (Phred quality score), Insert size (chiều dài đoạn nằm giữa 2 adapter ở 2 đầu đoạn trình tự thô), K-mer (chuỗi con K-mer), trình tự genome mẫu (Reference sequence, viết tắt là Refseq), một đoạn trình tự đặc thù (seed). Để kiểm tra thông số tối ưu cho kết quả trình tự genome có độ chính xác và tin cậy cao, từng thông số này lần lượt được khảo sát. Trong mỗi trường hợp các thông số còn lại được thiết lập ở chế độ mặc định (default /auto) (Bảng 2).

Phần mềm Prinseq (Schmieder, Edwards, 2011) được sử dụng để loại bỏ các dữ liệu không nằm trong khung giá trị khảo sát.

Chú thích bộ gen

Chương trình Geseq (https://chlorobox.mpimp-golm.mpg.de/geseq.html) được sử dụng để chú thích tên, vị trí, cấu trúc của các gen trong bộ gen. Thuộc tính DNA được thiết lập là "dạng vòng". Nguồn gốc trình tự (source sequence) được thiết lập là "plastid". Chiều dài, chiều trình tự, trật tự gen được kiểm tra tính chính xác bằng cách so sánh với dữ liệu chú thích bộ gen lục lạp hoàn chỉnh của một số genome tham khảo trên ngân hàng GenBank, đó là *P. armeniacum* (KT388109.1), *P. dianthum* (NC_036958.1) và *P. Niveum* (NC_026776.1). Công cụ BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) được sử dụng để thực hiện việc kiểm

tra này đồng thời để xuất file dữ liệu chú thích dưới định dạng GenBank.

Vẽ bản đồ bộ gen

Dữ liệu trình tự đã chú thích gen dưới định dạng GenBank được đưa vào chương trình OGDraw (https://chlorobox.mpimpgolm.mpg.de/OGDraw.html) để vẽ và xuất bản đồ bộ gen ở định dạng ảnh, thể hiện màu sắc và tên gọi các gen khác nhau trong bộ gen.

Bảng 2. Khảo sát các thông số được thiết lập cho quá trình lắp ráp trình tự bộ gen.

Khảo sát	Phred quality score	K-mer	Insert size	Seed	Refseq
Quality	≥ 39 ≥ 30 ≥ 20	39 (default)	Auto	*rbc*L- *P. armeniacum*	*P. armeniacum*
Insert size	≥ 20	39 (default)	290 295 300 350 Auto	*rbc*L- *P. armeniacum*	*P. armeniacum*
K-mer	≥ 20	39 35 30 25 20	Auto	*rbc*L- *P. armeniacum*	*P. armeniacum*
Refseq, seed	≥ 20	39 (default)	Auto	*rbc*L- *P. armeniacum* *rbc*L - *P.niveum* *rbc*L - *P.dianthum* matK - *P. armeniacum* Complete chloroplast genome *Dendrobium nobile* *rbc*L- *Dendrobium nobile*	*P. armeniacum* *P. niveum* *P. dianthum* *P. armeniacum* - -

KẾT QUẢ

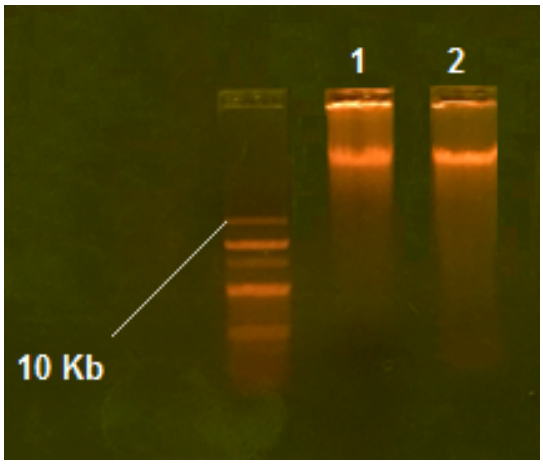
Tách DNA tổng số

Kết quả đo độ tinh sạch của cả 2 mẫu tách đều đạt yêu cầu trong khoảng 1.8 -2.2 (Bảng 3). Băng DNA điện di cho vạch sáng đậm rõ nét, ít bị vệt dài (Hình 1) thể hiện nồng độ và độ nguyên vẹn rất cao. Các mẫu đều đạt đủ chất lượng để gửi giải trình tự.

Điều đáng chú ý là nồng độ DNA đo bằng Nanodrop thể hiện cao hơn đo bằng Quantus hơn 2 lần. Nanodrop là máy đo quang phổ phổ biến khi khuếch đại các đoạn trình tự DNA ngắn, hoặc để giải trình tự Sanger. Quantus thì chi phí hóa chất cao hơn nên ít phổ biến. Tuy nhiên, đây là máy đo tín hiệu huỳnh quang với độ nhạy cao khi định lượng axit nucleic sẽ giúp kiểm soát nồng độ DNA ban đầu, được đề nghị sử dụng cho kiểm tra nồng độ DNA cho các phản ứng giải trình tự NGS (Lienhard, Schäffer, 2019).

Bảng 3. Kết quả đo OD và nồng độ bằng máy đo Nanodrop và Quantus.

Quy trình tách chiết	Mẫu DNA	A260/280	Nồng độ DNA (ng/µL)		Thể tích (µL)	Hàm lượng mẫu (ng)
			Đo bằng máy Nanodrop	Đo bằng máy Quantus		
SDS	1	1.85	250	110	25	2750 - 6250
	2	2.12	359	125	25	3125 - 8975



Hình 1. Kết quả điện di trên gel agarose 0.8% của 2 mẫu DNA tổng số và thang DNA

Kiểm tra chất lượng trình tự thô

Bộ dữ liệu trình tự thô thu được gồm cả 2 chiều là chiều xuôi (forward) và chiều ngược (reverse). Việc kiểm tra chất lượng trình tự thô được thực hiện trên trình tự cả 2 chiều để tăng độ tin cậy khi liên ứng (consensus) trình tự 2 chiều thành một trình tự thống nhất. Kết quả kiểm tra chất lượng bằng phần mềm FastQC được thể hiện ở Hình 2.

Tổng số trình tự thô (read) thu được ở mỗi chiều là 11.635.039 đoạn, tỉ lệ GC 35%. Chiều dài của các đoạn trình tự nằm trong khoảng 149-151 bp, trong đó các đoạn trình tự có chiều dài 150 bp chiếm đa số (Hình 2B). Tỉ lệ nucleotide N của cả 2 file trình tự trên tổng số base đều có giá trị 0% (Hình 2A). Tỉ lệ phần trăm adapter ở cả 2 file trình tự chiếm 1-3%, xuất hiện chủ yếu ở vị trí base 110-136 (Hình 2C). Chất lượng trình tự xét theo từng vị trí base của các trình tự hầu hết đều nằm trong ngưỡng màu xanh với điểm chất lượng từ 32 trở lên, ngoại trừ một đoạn ngắn ở cuối trình tựchiều ngược có giá trị rơi vào khu vực màu cam. Đường giá trị trung bình (màu xanh) đều trên 38 điểm (Hình 2D). Điểm chất lượng trình tự (Phred score) của phần lớn trình tự đều đạt từ 38-40 và ở cả 2 dữ liệu không có trình tự nào chất lượng thấp hơn 19 (Hình 2E). Mức độ lặp lại trình tự trong cả bộ chiều xuôi và chiều ngược ở mức 1-2 và phần trăm trình tự còn lại sau khi đã loại bỏ các đoạn lặp lại chiếm 92,17% (Hình 2F).

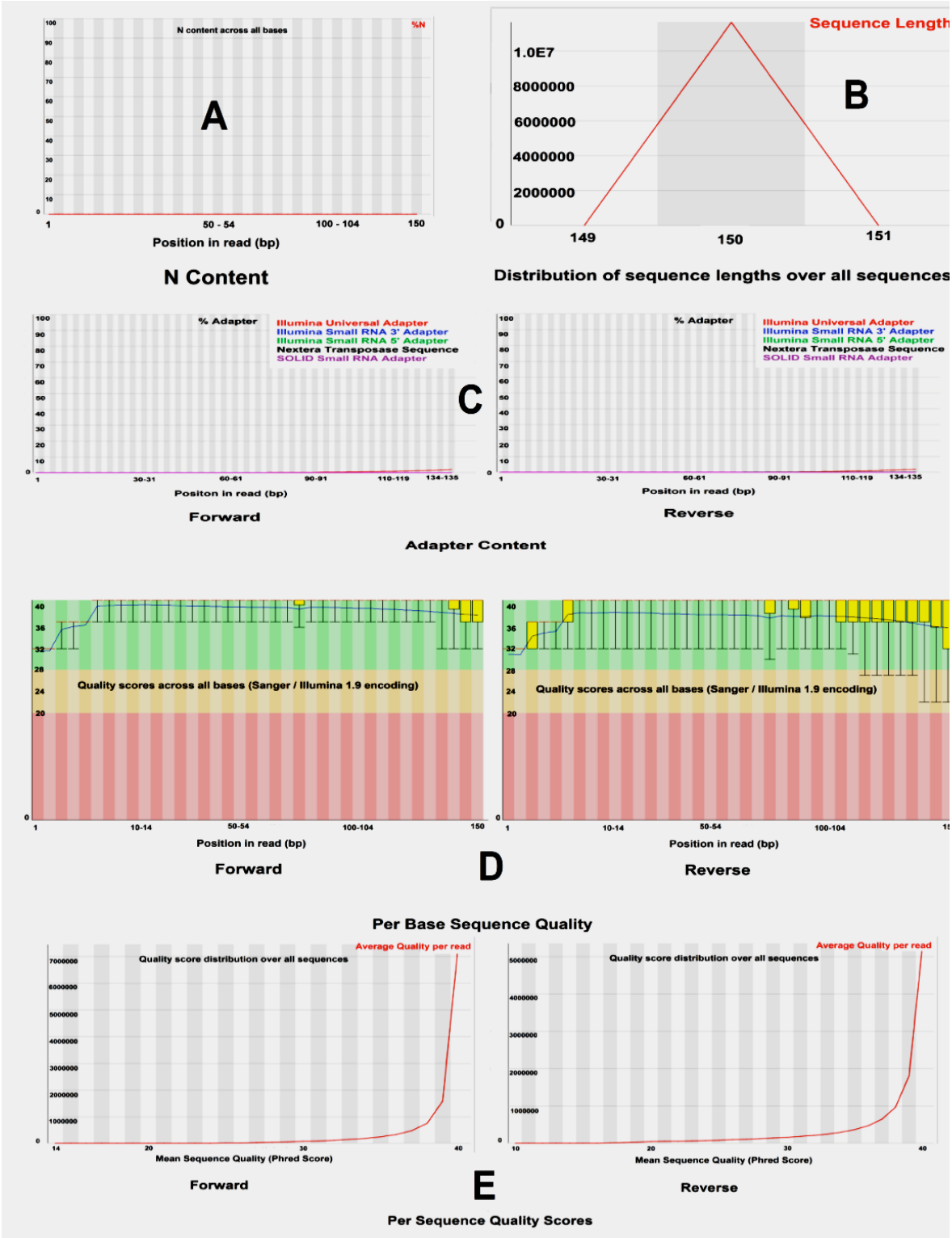
Lắp ráp bộ gen

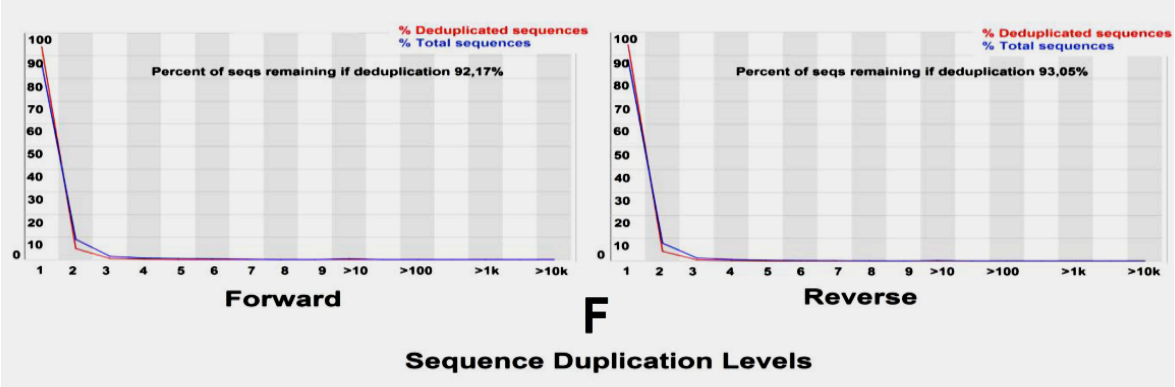
Ở giá trị K-mer 20 (với Phred quality score: 39, Insert zise: auto, Seed: gen *rbc*L của Pap*hiopedilum armeniacum*, Refseq: bộ gen lục lạp của *Paphiopedilum armeniacum*), chương trình xuất ra 5 đoạn contig với đoạn lớn nhất dài 90.573 bp, kết quả chiều dài genome lắp ráp được là 160.924 bp, độ bao phủ trình tự 923 lần (Bảng 4). Ngoại trừ trường hợp này, các kết quả khảo sát còn lại đều cho ra 3 contig gồm 1 contig dài và 2 contig ngắn. Mặc dù chiều dài các contig trong các trường hợp không giống nhau hoàn toàn, kết quả chiều dài genome đều thu được là 160.955 bp, độ bao phủ trình tự đạt từ 612-871 lần (Bảng 4). Độ bao phủ tuy thấp hơn so với trường hợp K-mer 20, chiều dài genome thu được lại dài hơn 32 nucleotide.

Kiểm tra tính chính xác cấu trúc bộ gen

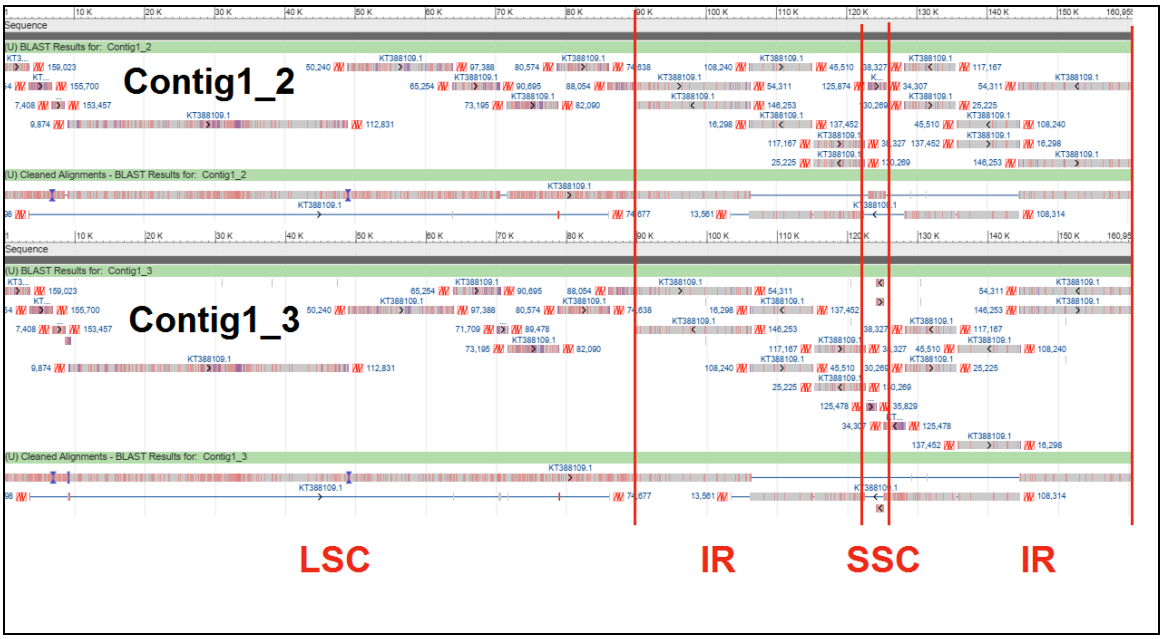
Thành công của việc lắp ráp tạo ra được 2 kết quả bộ gen vòng hoàn chỉnh trong đó một vòng gen do sự kết hợp sắp gióng cột từ Contig 1+2 và vòng gen kia do Contig 1+3 tạo ra. Cấu trúc bộ gen gồm vùng sao chép lớn (LSC, dài 90.365 bp) và vùng sao chép nhỏ (SSC, dài 2.550 bp) được phân tách bằng một cặp vùng lặp lại đảo ngược (IR, dài 34.020 bp cho mỗi

vùng). Hai vòng gen DNA lục lạp này khi được BLAST với nhau thì có độ tương đồng 100% và có chiều dài bộ gen cũng bằng nhau 160.955 bp, tuy nhiên vùng SSC ngược chiều nhau.





Hình 2. Kết quả kiểm tra chất lượng trình tự thô từ FastQC (A: Phần trăm N, B: Chiều dài trình tự, C: Phần trăm adapter, D: Chất lượng Base, E: Chất lượng trình tự, F: Mức độ lặp lại của trình tự).



Hình 3. Kết quả BLAST 2 trình tự DNA với trình tự bộ gen refseq *P. armeniacum* (KT388109.1). Ghi chú: các đoạn dài ngắn màu xám thể hiện sự tương đồng (match) nucleotide, giữa các đoạn xám này có các sọc nhỏ màu đỏ thể hiện các vị trí nucleotide biến dị di truyền (variation).

Bằng cách truy cập Ngân hàng gen sử dụng BLAST mỗi bộ gen với trình tự mẫu *P. armeniacum* (KT388109.1) trên NCBI, chúng tôi xác định được chiều của 2 vùng SSC và LSC ở vòng gen do Contig 1+3 tạo ra ngược chiều nhau, còn chiều của vùng SSC ở vòng gen do Contig 1+2 tạo ra cùng chiều với vùng LSC của chính nó đồng thời cũng cùng chiều với vùng SSC của bộ gen refseq *P. armeniacum* (Hình 3). Cấu trúc hai vùng single copy cùng chiều với nhau cũng đã được báo cáo trong các nghiên cứu trước đây (Li *et al.*, 2018). Từ đó chúng tôi chọn trình tự tạo từ Contig 1+2 làm dữ liệu để thực hiện chú thích bộ gen.

Bảng 4. Kết quả lắp ráp bộ gen.

Thông số khảo sát	Giá trị khảo sát	Kết quả		
		Contig	Chiều dài bộ gen (genome size)	Độ bao phủ (genome coverage)
Phred quality score	>=39	Contig 01 : 158405 bp Contig 02 : 3724 bp Contig 03 : 3585 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	612
	>=30	Contig 01 : 158405 bp Contig 02 : 3546 bp Contig 03 : 3691 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	794
	>=20	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3693 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
Insert size	290	Contig 01 : 158405 bp Contig 02 : 3541 bp Contig 03 : 3684 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	295	Contig 01 : 158405 bp Contig 02 : 3541 bp Contig 03 : 3691 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	300	Contig 01 : 158405 bp Contig 02 : 3541 bp Contig 03 : 3696 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	350	Contig 01 : 158405 bp Contig 02 : 3746 bp Contig 03 : 3541 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	auto	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3693 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
K-mer	39	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3693 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	35	Contig 01 : 157486 bp Contig 02 : 4507 bp Contig 03 : 4653 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	815
	30	Contig 01 : 158405 bp Contig 02 : 3572 bp Contig 03 : 3718 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	818
	25	Contig 01 : 93614 bp Contig 02 : 68502 bp Contig 03 : 68356 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	871
	20	Contig 01 : 90573 bp Contig 02 : 66880 bp Contig 03 : 4427 bp Contig 04 : 67031 bp Contig 05: 4427 bp	Contig 1+2+3 : 160924 bp Contig 1+4+5 : 160924 bp	923
- Seed - Refseq	-gen rbcL của Paphiopedilum armeniacum - bộ gen lục lạp của Paphiopedilum armeniacum	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3693 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	-gen rbcL của Paphiopedilum niveum - bộ gen lục lạp của Paphiopedilum niveum	Contig 01 : 158405 bp Contig 02 : 3537 bp Contig 03 : 3703 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	-gen rbcL của Paphiopedilum dianthum - bộ gen lục lạp của Paphiopedilum dianthum	Contig 01 : 158406 bp Contig 02 : 3545 bp Contig 03 : 3694 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805
	- gen matK của *Paphiopedilum armeniacum* - bộ gen lục lạp của *Paphiopedilum armeniacum*	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3692 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	804
	- bộ gen lục lạp của *Dendrobium nobile* - không có Refseq	Contig 01 : 158405 bp Contig 02 : 3548 bp Contig 03 : 3691 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	804
	-gen rbcL của Dendrobium nobile -không có Refseq	Contig 01 : 158405 bp Contig 02 : 3547 bp Contig 03 : 3693 bp	Contig 1+2 : 160955 bp Contig 1+3 : 160955 bp	805

Chú thích bộ gen

Chương trình Geseq được sử dụng để thực hiện chú thích tên, vị trí và cấu trúc các gen trong bộ gen, với trình tự mẫu được thiết lập là *P. armeniacum* (KT388109.1). Bộ gen lục lạp hoàn chỉnh của *P. delenatii* sau khi được lắp ráp có chiều dài 160.955 bp và có tỉ lệ GC 35,6%. Tỉ lệ GC cũng có sự thay đổi giữa vùng LSC, SSC và IRs. Trong đó vùng IRs có tỉ lệ GC cao hơn hẳn (40%)so với vùng SSC (29%), LSC (33%).

Bộ gen lục lạp của *P. delenatii* có tổng cộng 130 gen gồm 77 gen mã hóa protein, 39 gen mã hóa tRNA, 8 gen mã hóa rRNA (Bảng 5).

Bảng 5. Danh sách các gen trong bộ gen lục lạp *P. delenatii.*

Classification of Genes		Name of Gennes	Number
RNA genes	Ribosomal RNAs	*rrn*4.5(x2), *rrn*5(x2), *rrn*16(x2), *rrn*23(x2)	8
	Transfer RNAs	*trn*A_UGC(x2), *trn*C_GCA, *trn*D_GUC, *trn*E_UUC, *trn*F_GAA, *trn*fM_CAU, *trn*G_GCC, *trn*G_UCC, *trn*H_GUG(x2), *trn*I_CAU(x2), *trn*I_GAU(x2), *trn*K_UUU, *trn*L_CAA(x2), *trn*L_UAA, *trn*L_UAG(x2), *trn*M_CAU, *trn*N_GUU(x2), *trn*P_UGG, *trn*Q_UUG, *trn*R_ACG(x2), *trn*R_UCU, *trn*S_GCU, *trn*S_GGA, *trn*S_UGA, *trn*T_GGU, *trn*T_UGU, *trn*V_GAC(x2), *trn*V_UAC, *trn*W_CCA, *trn*Y_GUA	39
Protein-coding genes	Photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J	5
	Photosystem II	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z	15
	Cytochrome	*pet*A, *pet*B, *pet*D, *pet*G, *pet*L, *pet*N	6
	ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I	6
	Rubisco	*rbc*L	1
	NADH dehydrogenease - like complex	*ndh*B(x2), *ndh*C, *ndh*D, *ndh*J, *ndh*K	6
	Ribosomal proteins - small units	*rps*11, *rps*12(x2), *rps*14, *rps*15(x2), *rps*16, *rps*18, *rps*19(x2), *rps*2, *rps*3, *rps*4, *rps*7(x2), *rps*8	16
	Ribosomal proteins - large units	*rpl*14, *rpl*16, *rpl*2(x2), *rpl*20, *rpl*22, *rpl*23(x2), *rpl*32(x2), *rpl*33, *rpl*36	12
	RNA polymerase	*rpo*A, *rpo*B, *rpo*C1, *rpo*C2	4
	Miscellaneous	*acc*D, *ccs*A, *cem*A, *clp*P, *inf*A, *mat*K	6
	Hypothetical chloroplast reading frames (ycf)	*ycf*1(x2), *ycf*2(x2), *ycf*3, *ycf*4	6
Total			130

Vẽ bản đồ gen

Dữ liệu chú thích bộ gen được đưa vào chương trình OGDraw để vẽ bản đồ bộ gen. Hình ảnh trực quan thể hiện bộ gen lục lạp dạng vòng khép kín, vòng tròn trong ghi chú các vùng SSC, LSC, IR. Vòng tròn ngoài thể hiện rõ vị trí, thứ tự, độ dài các đoạn gen. Màu sắc gen khác nhau theo nhóm gen được chú thích ở góc trái bên dưới Hình 4. Các gen nằm bên ngoài vòng tròn được phiên mã theo chiều kim đồng hồ, trong khi các gen nằm bên trong vòng

tròn được phiên mã ngược chiều kim đồng hồ. Màu xám đậm tương ứng với tỉ lệ phần trăm GC, màu xám nhạt tương ứng với tỉ lệ phần trăm AT.

Hình 4. Bản đồ bộ gen lục lạp hoàn chỉnh của loài lan Hài hồng *Paphiopedilum delenatii.*

THẢO LUẬN

Kiểm tra chất lượng trình tự thô

Nucleotide N là những nucleotide mơ hồ không xác định được (James, 2001) trong quá trình giải trình tự tự động từ đó sẽ làm ảnh hưởng đến kết quả lắp ráp bộ gen nên cần được loại bỏ nếu có. Trong nghiên cứu này, dữ liệu thu được có tỉ lệ Nucleotide N là 0%, nghĩa các nucleotide đều được xác định rõ ràng. Trong quá trình giải trình tự bằng kỹ thuật Illumina, các đoạn DNA được cắt nhỏ từ DNA tổng số cần được gắn với các chuỗi tiếp hợp (adapter) là một đoạn trình tự ngắn vào đầu của DNA nhằm hỗ trợ cho việc bắt cặp mồi để thực hiện phản ứng PCR khuếch đại trình tự. Sau đó các adapter sẽ được cắt rời khỏi các đoạn DNA (Levy E, Myers M, 2016). Nếu adapter còn sót lại trên 10% trong dữ liệu trình tự (Andrews, 2010) thì sẽ ảnh hưởng đến chất lượng giải trình tự và kết quả lắp ráp bộ gen. Trong dữ liệu nghiên cứu, tỉ lệ phần trăm adapter ở cả 2 file trình tự vào khoảng 1-3% (Hình 2C), việc này không ảnh hưởng đáng kể đến chất lượng trình tự. Các kết quả cho thấy chất lượng trình tự thô rất tốt và đạt độ tin cậy cao khi lắp ráp genome.

Giá trị chất lượng các base trong kết quả đánh giá được bố trí thành các ngưỡng màu xanh, màu cam và màu hồng. Màu xanh là các giá trị rất tốt, màu cam là các giá trị chấp nhận được, và màu hồng các các giá trị không tốt. Sau khi kiểm tra các trình tự đều có giá trị nằm trong ngưỡng màu xanh (Hình 2D) thể hiện chất lượng trình tự rất cao ở các vị trí base xuyên suốt chiều dài trình tự. Phred score là thông số thể hiện chất lượng trung bình của việc nhận diện nucleotide qua quá trình giải trình tự DNA . Kết quả cho thấy chất lượng trung bình các trình tự thô thu được (Quality score distribution all sequences) rất cao khi đối chiếu tương ứng với tỉ lệ chính xác đạt 99,99% (Bảng 6).

Một thư viện trình tự chất lượng và có độ bao phủ cao khi mức độ lặp lại của mỗi trình tự thấp và trình tự đa dạng. Hiện tượng các trình tự lặp lại với số lượng lớn là do quá trình khuếch đại quá mức trong giai đoạn tạo thư viện (Andrews, 2010). Dựa vào 2 đường trong biểu đồ Duplicate Sequence, đường màu xanh thể hiện phần trăm trình tự lặp lại trong tổng số trình tự ban đầu, đường màu đỏ thể hiện phần trăm trình tự lặp lại sau khi đã loại bỏ các đoạn lặp), một file có giá trị tốt nếu cả 2 đường càng nằm về phía bên trái của biểu đồ chứng tỏ là mức độ lặp lại càng thấp. Kết quả đánh giá mức độ lặp lại của 2 file trình tự cho thấy mức độ lặp lại của trình tự thấp khi cả 2 đường đều nằm về phía góc trái của biểu đồ với mức độ lặp lại ở mức 1-2 và phần trăm trình tự còn lại sau khi đã loại bỏ các đoạn lặp lại chiếm 92,17% (Hình 2F). Do cả 2 file trình tự đều có chất lượng tốt nên chúng tôi không thực hiện tiếp bước lọc bỏ bớt trình tự. Toàn bộ thông tin trình tự sau khi được kiểm tra đạt chất lượng được tiếp tục sử dụng để lắp ráp bộ gen.

Bảng 6. Đánh giá sự tương quan giữa điểm chất lượng và tỉ lệ chính xác (Kwon *et al.*, 2013)

Điểm chất lượng (Phred Quality Score)	Tỉ lệ số nucleotide bị sai (Probability of incorrect base call)	Tỉ lệ chính xác (Base call accuracy)
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%
50	1/100000	99.999%

Ngưỡng chất lượng trình tự làm dữ liệu đầu vào cho việc lắp ráp

Theo lý thuyết, các trình tự không đủ độ chính xác cần được loại bỏ trước khi lắp ráp bộ gen để tránh bị nhiễu thông tin, dẫn đến việc lắp ráp không thành công hoặc thiếu chính xác. Do điểm chất lượng trình tự thô qua kiểm tra đều nằm trong khoảng từ 19 tới 40 (Hình 2E) nên chúng tôi chia 3 mức giá trị khảo sát là >=39, >=30, >=20 (Bảng 4). Kết quả chiều dài bộ gen đều giống nhau có thể giải thích là do số lượng trình tự có điểm chất lượng dưới 19 và dưới 30 chiếm số lượng không đáng kể (Hình 2E) nên không ảnh hưởng nhiều đến việc lắp ráp contig. Tuy nhiên, điểm chất lượng càng cao thì có độ bao phủ trình tự sau khi lắp ráp càng thấp do số lượng trình tự đầu vào (input sequence) ít hơn (Bảng 4). Độ bao phủ là số lần lặp lại của trình tự toàn bộ gen, cũng là một thông số đo lường chất lượng của việc lắp ráp, số lượng này càng lớn độ tin cậy càng cao. Do đó, trong nghiên cứu này những trình tự có chất lượng đạt từ 20 trở lên đều được sử dụng làm dữ liệu cho quá trình lắp ráp genome để đạt được mức bao phủ cao nhất, dù trong trường hợp cụ thể này, cả 3 trình tự genome thu được đều đồng nhất 100%.

Chiều dài chuỗi con K-mer

Một trong những nguyên tắc của lắp ráp bộ gen là xác định đoạn trình tự chồng lắp (overlap) tương đồng để ghép nối với nhau thành các đoạn dài hơn. Cơ sở của việc này là thuật toán sắp gióng cột (alignment). Tuy nhiên, trình tự DNA thường là quá dài để thực hiện

việc sắp gióng cột hiệu quả. Do đó các thuật toán sắp gióng cột thường sẽ chia trình tự ban đầu thành từng đoạn ngắn để dễ bắt cặp tương đồng rồi từ điểm bắt cặp đó so sánh tiếp tương đồng nucleotide về 2 phía. Những đoạn ngắn này được gọi là chuỗi con K-mer (Sohn và Nam, 2018) Chuỗi con trong giải trình tự NGS này được khuyến cáo là dài không quá 39 bp. Chuỗi con quá dài sẽ khó tìm đoạn tương đồng, chuỗi con quá ngắn sẽ dẫn đến đoạn tương đồng quá nhiều mà độ tin cậy thấp. Do đó các giá trị K-mer được chọn để khảo sát hiệu quả lắp ráp là 20, 25, 30, 35 và 39 (Bảng 4).

Quá trình lắp ráp bộ gen gồm 2 giai đoạn là lắp ráp các đoạn trình tự thô ngắn thành các đoạn dài gọi là contig, sau đó contig được lắp ráp lần nữa để tạo thành genome hoàn chỉnh.Số lượng contig nên từ 2-3 là tốt nhất (Nicolas *et al.*, 2017). Trường hợp K-mer 20 tạo ra đến 5 contig, nhưng trong trình tự bộ gen hoàn chỉnh chúng tôi phát hiện có một vài khoảng trống (gap) là các nucleotide không xác định được sau khi lắp ráp hoàn thành. Chiều dài hoàn chỉnh của genome trong trường hợp này ngắn hơn 32 bp so với kết quả ở các trường hợp có 3 contig.

Trình tự bộ gen mẫu (refseq) và trình tự hạt giống (seed)

Để thực hiện lắp ráp một bộ gen mới dựa trên một bộ gen mẫu đã biết (phương pháp homologus modeling), chương trình NOVOPlasty cần có một trình tự genome hoàn chỉnh và để làm bộ gen mẫu (refseq) và một trình tự hạt giống (seed) cũng để làm mẫu vị trí bắt đầu cho việc đối chiếu trình tự.

Bộ gen mẫu (refseq) có độ tương đồng với loài nghiên cứu càng cao thì kết quả lắp ráp càng chính xác và độ tin cậy cao. Hiện nay chỉ mới có trình tự bộ gen hoàn chỉnh của 3 loài cùng chi lan Hài được công bố trên Ngân hàng gen là *P. armeniacum, P. niveum* và *P. dianthum*. Cả ba loài đều có quan hệ rất gần với loài nghiên cứu, trong đó *P. armeniacum* là loài gần nhất do được phân loại cùng tổ (section) với loài nghiên cứu *P. delenatii* dựa theo hình thái.

Trình tự hạt giống (seed) thường là một đoạn trình tự ngắn, được chương trình sử dụng làm xuất phát điểm cho toàn bộ quá trình lắp ráp bộ gen. Do đó, seed thường phải có độ bảo tồn cao để đảm bảo độ tương đồng ổn định với loài mới. Seed có thể thuộc bộ gen bào quan của chính loài đó hay loài khác trong chi. Ngoài ra, trong trường hợp không tìm được trình tự của loài có mối quan hệ gần với loài được lắp ráp bộ gen, seed cũng có thể là trình tự bộ gen bào quan của một loài xa hơn. Chương trình NOVOPlasty đề nghị sử dụng hạt giống là trình tự gen *rbc*L (Nicolas *et al.*, 2017). Đây là gen mã hóa cho protein RUBP (Ribulose 1,5-bisphosphate), được xác định là trình tự có độ bảo tồn cao ở cấp độ trên chi (Bafeel *et al.*, 2012), phù hợp với yêu cầu của NOVOPlasty. Mặc dù vậy, gen matK cũng vẫn cho kết quả tin cậy cao và hoàn toàn có thể thay thế *rbc*L. Không những vậy, phép thử không dùng Refseq cũng cho kết quả tối ưu, ngay cả với trình tự hạt giống (bộ gen lục lạp của *Dendrobium nobile*) khác chi và khác xa hơn về mặt di truyền. Thậm chí trình tự hạt giống (gen *rbc*L của *Dendrobium nobile*) chỉ cần là một đoạn gen rất ngắn của chi khác vẫn có thể áp dụng. Kết quả khảo sát này có ý nghĩa khẳng định tính khả thi của việc lắp ráp bộ gen ngay cả ở các cá thể mà chưa có trình tự tương đồng gần để tham khảo.

Chú thích bộ gen

Độ tương đồng trình tự giữa loài nghiên cứu *P. delenatii* và loài tham khảo *P. armeniacum* là 97,84%. Tỉ lệ GC của bộ gen lục lạp *P. denlenatii* và *P. armeniacum* có giá trị khá giống nhau là 35,6% và 35,4% (Bảng 8) nằm trong khoảng tỉ lệ GC% trung bình ở thực vật là 33,6-47,5% (Smarda *et al.*, 2012). Hiện tượng này được hình thành do quá trình sao chép và xảy ra lỗi trong sửa chữa DNA (Talat, Wang, 2015), DNA polymerase ở lục lạp có xu hướng kết hợp sai A, T thay vì G và C (Howe *et al.*, 2003). Tỉ lệ GC vùng IRs (40%) cao hơn so với vùng SSC (29%), LSC (33%) là do vùng IR chứa các gen rRNA (*rrn*4.5, *rrn*5, *rrn*23, *rrn*16) và một số vùng mã hóa (Talat, Wang, 2015). Dựa vào tỉ lệ GC có thể biết được sự đa dạng của bộ gen từ đó phân tích được mối quan hệ tiến hóa của các loài (Smarda *et al.*, 2014).

Trong cấu trúc của bộ gen lục lạp thì các vùng sao chép đơn có khả năng đột biến điểm cao gấp 2,3 lần so với vùng IR (Shaw *et al.*, 2007). Do đó vùng sao chép đơn thường được nghiên cứu nhiều hơn (Shaw *et al.*, 2007). Tuy nhiên vùng IR chứa các gen lặp lại (gen mã hóa ribosome, một số gen tRNA, gen mã hóa protein) có vai trò quan trọng trong việc duy trì sự sắp xếp các gen của DNA lục lạp (Václav *et al.*, 2018).

Trong bộ gen lục lạp của thực vật trên cạn và tảo lục có thể phân các gen thành 2 nhóm chính: những gen liên quan đến biểu hiện gen và những gen liên quan đến quá trình quang hợp (Sugiura, 1995). Trong nhóm gen liên quan đến quá trình quang hợp các gen *psa*, *psb*, *pet*, *atp* mã hóa lần lượt hệ thống quang hóa I (Photosytem I - PSI), hệ thống quang hóa II (Photosytem II - PSII), cytochrome, ATP sythase đều có vai trò quan trọng trong quá trình quang hợp. Trong đó PSI giúp tạo ra ATP, PSII tạo ra NADH, ATP, O_2 cho cây (Nelson, Yocum, 2006). NADH dehydrogenase là loại enzyme có vai trò quan trọng trong chuỗi vận chuyển điện tử trong quá trình hô hấp của ty thể. Tuy nhiên, trong lục lạp thì có các gen *ndh* mã hóa cho NADH dehydrogenase-like complexcó vai trò tương tự NADH dehydrogenase. NDH có vai trò vận chuyển điện tử của lục lạp (Ifuku *et al.*, 2011; Nelson, Yocum, 2006). Ngoài ra còn có một số gen khác *rps*, *rpl* mã hóa cho protein của ribosome, *rpo* mã hóa RNA polymerase. Như vậy, việc chú thích bộ gen lục lạp mang lại những thông tin quan trọng về các gen, cấu trúc, trình tự, vị trí của chúng nhờ đó góp phần cho những công tác nghiên cứu sau này.

Bảng 6. So sánh bộ gen *P. delenatii* và *P. armeniacum.*

	***P. delenatii* (MK463585)**	***P. armeniacum* (KT388109.1)**
Chiều dài bộ gen (bp)	160.955 bp	162.682 bp
Chiều dài IR (bp)	34.020 bp	67.072 bp
Chiều dài LSC (bp)	90.365 bp	91.942 bp
Chiều dài SSC (bp)	2.550 bp	3.668 bp
GC content (%)	35,6%	35,4%
GC content của IR (%)	40%	39%
GC content của LSC (%)	33%	32,6%
GC content của SSC (%)	29%	31%
Tổng số gen (bao gồm các gen lặp)	130(23)	131(24)
Số CDS (bao gồm các gen lặp)	77(9)	79(11)
Số gen rRNA (bao gồm các gen lặp)	8(4)	8(4)
Số gen tRNA (bao gồm các gen lặp)	39(9)	38(8)

Bản đồ bộ gen

Việc lập bản đồ là một bước quan trọng trong nghiên cứu giải trình tự bộ gen. Trình tự và bản đồ bộ gen đều mang lại một cái nhìn tổng quát về bộ gen, nhưng bản đồ bộ gen thì ít chi tiết hơn trình tự bộ gen. Trình tự bộ gen sẽ cho biết vị trí chính xác từng nucleotide trong DNA, trong khi đó bản đồ gen chỉ thể hiện vị trí các mốc trong bộ gen (Craig, 2003). Trong bản đồ bộ gen thì các vị trí GCA, CCC, CATTT, GAA được xem là một vị trí, trong khi đó mỗi vị trí nucleotide trong trình tự bộ gen được xem là một vị trí. Từ đó cho thấy bản đồ bộ gen là sự thể hiện tóm tắt lại toàn bộ trình tự bộ gen. Việc lập bản đồ bộ gen cho thấy thông tin các gen trên bản đồ giúp các nhà khoa học dễ hình dung trực quan ở mức độ tổng quát về toàn bộ bộ gen, giúp các nhà khoa học phát hiện ra các gen mới hay đặc điểm mới của bộ gen.

Hiện tại, dữ liệu genome lục lạp này đang được tiếp tục phân tích để tìm kiếm các thông tin hữu ích như đánh giá độ đa dạng của các vùng trình tự tiềm năng làm mã vạch DNA, phân tích các vùng trình tự lặp lại (repeat) và các vùng vệ tinh (microsatellite) hiện diện trong genome phục vụ đánh giá đa dạng di truyền và nhận diện phân tử, đồng thời phân tích phát sinh chủng loài từ bộ genome lục lạp.

KẾT LUẬN

Nghiên cứu đã mô tả chi tiết quy trình lắp ráp và chú thích bộ gen lục lạp hoàn chỉnh của loài lan Hài hồng (*Paphiopedilum delenatii*) đặc hữu của Việt Nam. Kết quả genome là cơ sở để phân tích các dữ liệu khác phục vụ nghiên cứu và ứng dụng trên đối tượng này. Đồng thời quy trình được đề xuất trong nghiên cứu có thể dễ dàng thực hiện trên máy tính cá nhân với thời gian ngắn, cho kết quả chính xác và có thể được áp dụng rộng rãi cho nhiều đối tượng thực vật khác.

Lời cảm ơn: *Các tác giả xin chân thành cảm ơn Quỹ Phát triển Khoa học và Công nghệ Đại học Nguyễn Tất Thành thông qua đề tài mã số 2019.01.27/HĐ-KHCN cho nghiên cứu này.*

TÀI LIỆU THAM KHẢO

Ahmed I, Islam M, Arshad W, Mannan A, Ahmad W, Mirza B (2009) High-quality plant DNA extraction for PCR: an easy approach. *J Appl Genet* 50(2): 105-7.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc*.

Bafeel S, Alaklabi A, Arif I, Khan H, Alfarhan A, Ahamed A, Thomas J, Bakir M (2012) Ribulose-1,5-biphosphate carboxylase (rbcL) gene sequence and random amplification of polymorphic DNA (RAPD) profile of regionally endangered tree species Coptosperma graveolens subsp. arabicum (S. Moore) Degreef. *Plant OMICS* 5: 285-290.

Craig J V (2003) *Genome Map*. Retrieved from http://www.genomenewsnetwork.org/resources/what s_a_genome/Chp3_1.shtml?fbclid=IwAR0wwaneDHuQLOVSNuafB9rLrrfCzvflRw_tnNUi0yYb5vsh8veTi_yYviY

Daniell H, Lin C S, Yu M, Chang W J (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* 17(1): 134.

Guo S, Guo L, Zhao W, Xu J, Li Y, Zhang X, Shen X, Wu M, Hou X (2018) Complete chloroplast genome sequence and phylogenetic analysis of Paeonia ostii. *Molecules* 23(2).

Howe C J, Barbrook A C, Koumandou V L, Nisbet R E R, Symington H A, Wightman T F (2003) Evolution of the chloroplast genome. *Philos Trans R Soc Lond B Biol Sci* 358(1429): 99-107.

Huỳnh Phước Hải, Nguyễn Văn Hòa (2015) Quy trình lắp ráp bộ gien Chloroplast. *Tạp chí Khoa học Trường Đại học Cần Thơ*: 9-16.

Ifuku K, Endo T, Shikanai T, Aro E M (2011) Structure of the chloroplast NADH dehydrogenase-like complex: nomenclature for nuclear-encoded subunits. *Plant Cell Physiol* 52(9): 1560-8.

Izan S, Esselink D, Visser R G F, Smulders M J M, Borm T (2017) De Novo assembly of complete chloroplast genomes from non-model species based on a K-mer frequency-based selection of chloroplast reads from total DNA sequences. *Front Plant Sci* 8: 1271.

James T (2001) *Beginning Perl for Bioinformatics*. O'Reilly & Associates, Inc., Sebastopol, California, USA.

Sohn J I, Nam J W (2018) The present and future of de novo whole-genome assembly. *Brief Bioinform* 19(1): 23-40.

Kwon S, Park S, Lee B, Yoon S (2013) In-depth analysis of interrelation between quality scores and real errors in Illumina reads. *Conf Proc IEEE Eng Med Biol Soc* 2013: 635-8.

Levy E S, Myers M R (2016) Advancements in Next-Generation Sequencing. *Annual review of genomics and human genetics* 17.

Li Y, Zhang J, Li L, Gao L, Xu J, Yang M (2018) Structural and comparative analysis of the complete chloroplast genome of Pyrus hopeiensis-"Wild plants with a tiny population"-and three other Pyrus species. *Int J Mol Sci* 19(10): 3262.

Lienhard A, Schäffer S (2019) Extracting the invisible: obtaining high quality DNA is a challenging task in small arthropods. *PeerJ* 7: e6753-e6753.

Manzanilla V, Kool A, Nguyen Nhat L, Nong Van H, Le Thi Thu H, de Boer H J (2018) Phylogenomics and barcoding of Panax: toward the identification of ginseng species. *BMC Evolutionary Biology* 18(1): 44.

Nelson N, Yocum F C (2006) Structure and function of photosystem I and II. *Annu Rev Plant Biol* 57: 521-65.

Nicolas D, Patrick M, Guillaume S (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res* 45(4): e18.

Saina J K, Li Z Z, Gichira A W, Liao Y Y (2018) The complete chloroplast genome sequence of tree of Heaven (Ailanthus altissima (Mill.) (Sapindales: Simaroubaceae), an important pantropical tree. *Int J Mol Sci* 19(4).

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6): 863-864.

Shaw J, Lickey E B, Schilling E E, Small R L (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am J Bot* 94(3): 275-88.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26(10): 1135-45.

Smarda P, Bures P, Horová L, Leitch I J, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová O (2014) Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci U S A* 111(39): E4096.

Smarda P, Bures P, Smerda J, Horova L (2012) Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytol* 193(2): 513-21.

Sugiura M (1995) The chloroplast genome. *Essays Biochem* 30: 49-57.

Talat F, Wang K (2015) Comparative Bioinformatics analysis of the chloroplast genomes of a wild diploid Gossypium and two cultivated Allotetraploid Species. *Iran J Biotechnol* 13(3): 47-56.

Tian N, Han L, Chen C, Wang Z (2018) The complete chloroplast genome sequence of Epipremnum aureum and its comparative analysis among eight Araceae species. *PLOS ONE* 13: e0192956.

Václav B, Jiří L, Bartas M, Fojta M (2018) Complex analyses of short Inverted Repeats in all sequenced chloroplast DNAs. *Biomed Res Int* 2018: 10.

Worth J R P, Liu L (2019) The complete chloroplast genome of Fagus crenata (subgenus Fagus) and comparison with F. engleriana (subgenus Engleriana). *PeerJ* 7: e7026.

Xiang l, Su Y, Li X, Xue G, Wang Q, Shi J, Wang L, Chen S (2016) Identification of Fritillariae bulbus from adulterants using ITS2 regions. *Plant Gene* 7.

Yeisoo Y, Hyun Oh L, Joong Hyoun C, Han Yong P, Soo-Cheul Y (2017) The complete chloroplast genome sequence of Oryza sativa aus-type variety Nagina-22 (Poaceae). *Mitochondrial DNA Part B* 2(2): 819-820.

CONSTRUCTION OF COMPLETE CHLOROPLAST GENOME OF THE ENDEMIC SPECIES *PAPHIOPEDILUM DELENATII* GUILLAUMIN (1924) OF VIETNAM

Nguyen Thanh Diem[1], Ly Le[2], Nguyen Huu Thuan Anh[1], Nguyen Thanh Cong[1], Vu Thi Huyen Trang[1,2,✉]

[1]Nguyen Tat Thanh University, Ho Chi Minh City
[2]International University, Ho Chi Minh National University

SUMMARY

Chloroplasts and mitochondria are organelles that have their own genome in a cell. The chloroplast genome provides information on the evolutionary relationship and species identification,

valuable markers for transgenic plants, and cloning plants, *etc*. The application of Next Generation Sequencing has improved the chloroplast genome sequencing. However, the assembly process of chloroplast genome is quite complicated due to the need of different complex bioinformatics tools, high configuration computer and laborous. Here we configured the process of assembling the chloroplast genome of *Paphiopedilum delenatii*. The assembled chloroplast genome was 160,955 bp in length, including a large and a small single copy region (LSC, SSC) separated by a pair of inverted repeats (IR). Total genes were 130 genes, GC content is 35.6%. Genome data was mapped and registered in GenBank under accession number MK463585. The optimal parameters for genome assembling were recommended. This study not only provided information for conservation of the Vietnam endemic *Paphiopedilum delenatii* species but also supported the genome assemble researches which could be applied on other subjects.

Keywords: *Paphiopedilum delenatii*, genome assembling, genome annotation, gen map, chloroplast genome