Tạp chí Công nghệ Sinh học **14**(1): 1-13, 2016

BÀI TỔNG QUAN

GIẢI MÃ HỆ GEN Ở THỰC VẬT VÀ CÁC LOÀI THUỘC CHI NHÂN SÂM (*PANAX* L.)

Lê Thị Thu Hiền[1], Hugo De Boer[2], Vincent Manzanilla[2], Hà Văn Huân[3], Nông Văn Hải[1]

[1]*Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[2]*Bảo tàng Lịch sử Thiên nhiên, Đại học Oslo, Na Uy*
[3]*Viện Công nghệ sinh học Lâm nghiệp, Đại học Lâm nghiệp*

Ngày nhận bài: 07.3.2016
Ngày nhận đăng: 20.3.2016

TÓM TẮT

Thành công trong việc phát triển các công nghệ giải trình tự gen đã mở ra thời kỳ phát triển mới trong nghiên cứu về khoa học sự sống với rất nhiều kỹ thuật phức tạp và hiện đại đã được phát triển và ứng dụng. Với chiến lược phát triển bền vững, nhiều nước trên thế giới đã đầu tư mạnh cho nghiên cứu giải mã và phân tích hệ gen ở các đối tượng thực vật. Hàng năm, số lượng các loài được giải mã hệ gen tăng lên nhanh chóng. Kết quả đạt được mở ra nhiều cơ hội cho các nghiên cứu cơ bản và ứng dụng, cung cấp dữ liệu cho việc tìm kiếm các chỉ thị phân tử liên quan đến các tính trạng quan trọng và xác định nguồn gen. Ở Việt Nam, nghiên cứu giải mã toàn bộ hoặc một phần hệ gen các loài cây trồng có giá trị chỉ được bắt đầu trong thời gian gần đây trong khuôn khổ các chương trình hợp tác quốc tế. Chi Nhân sâm bao gồm các loài cây rất có giá trị kinh tế với khu vực phân bố hẹp như Sâm vũ diệp (*P. bipinnatifidus* Seem.), Tam thất hoang (*P. stipuleanatus* H.T.Tsai & K.M.Feng) và Sâm Ngọc linh hay còn gọi là Sâm việt nam (*P. vietnamensis* Ha et Grushv.)…Mặc dù là các loài dược liệu quý nhưng những hiểu biết về di truyền phân tử của các loài này còn rất hạn chế. Hiện nay, các nghiên cứu chỉ sử dụng một số chỉ thị phân tử để nhận dạng hay đánh giá đa dạng di truyền nguồn gen. Vì vậy, các nghiên cứu liên quan đến giải mã hệ gen, phát triển bộ mã vạch phân tử góp phần hiểu biết sâu hơn về các đặc tính di truyền phân tử và tiến hóa của loài. Bài viết này sẽ tổng quan một số công trình nghiên cứu về các công nghệ giải trình tự DNA/hệ gen trên thế giới và những ứng dụng trong giải mã hệ gen, hệ gen biểu hiện ở thực vật nói chung và các loài thuộc chi Nhân sâm nói riêng.

Từ khóa: *Chi Nhân sâm, hệ gen, hệ gen biểu hiện, lục lạp, giải trình tự gen thế hệ mới*

MỞ ĐẦU

Trên thế giới, các nghiên cứu về hệ gen thực vật được bắt đầu từ những năm đầu của thế kỷ 21 với công bố về xác định trình tự toàn bộ hệ gen của loài cây mô hình *Arabidopsis thaliana*. Những năm sau đó, hướng nghiên cứu này có những bước phát triển vượt bậc với nghiên cứu giải mã trình tự hệ gen lúa được công bố năm 2005; trình tự hệ gen cây dương năm 2006; toàn bộ trình tự hai kiểu gen cây nho năm 2007; đu đủ chuyển gen năm 2008. Hệ gen của nhiều loài thực vật khác cũng được giải mã (Gupta, Xu, 2008). Chương trình Sáng kiến Hệ gen Thực vật Quốc gia của Hoa Kỳ (National Plant Genome Initiative - NPGI) đã hỗ trợ các dự án giải mã hệ gen thực vật (http://www.whitehouse.gov). Trong thập kỷ tới, các loài thực vật vùng nhiệt đới, nơi có đa dạng sinh học cao, được ưu tiên giải mã hệ gen.

Trên đối tượng cây sâm, chi Nhân sâm (*Panax* L.) là chi gồm nhiều loài cây thuốc có giá trị cao. Trong số đó, có các loài chứa nhiều hợp chất tự nhiên có cấu tạo phân tử khá phức tạp, độc đáo, có hoạt tính tốt và có tác dụng tăng cường thể lực như Nhân sâm (*P. ginseng*) và Sâm Ngọc linh. Cùng với sự phát triển của công nghệ giải trình tự gen thế hệ mới, việc tiến hành các nghiên cứu giải mã hệ gen các loài thuộc chi *Panax* đã được thực hiện trong những năm gần đây. Các công bố chủ yếu tập trung ở các quốc gia có Sâm phân bố và được sử dụng thường xuyên làm dược phẩm như Trung Quốc, Hàn Quốc…Các nghiên cứu giải mã đầu tiên của các loài trong chi *Panax* thường tập trung vào các vùng gen có giá trị trong nhận dạng, hệ gen biểu hiện, xây dựng thư viện của các trình tự biểu hiện (ESTs) phục vụ nghiên cứu chức năng của các gen sau này như biểu hiện gen, marker phân tử, lập bản đồ di truyền...Bài viết này sẽ tổng quan một số công trình

nghiên cứu về các công nghệ giải trình tự DNA/hệ gen trên thế giới và những ứng dụng trong giải mã gen, hệ gen lục lạp, hệ gen biểu hiện ở thực vật nói chung và các loài thuộc chi *Panax* nói riêng.

CÁC CÔNG NGHỆ GIẢI TRÌNH TỰ DNA/HỆ GEN TRÊN THẾ GIỚI

Các hệ thống máy giải trình tự DNA thế hệ đầu tiên

Giữa những năm 70 của thế kỷ 20, Sanger đã đưa ra khái niệm đầu tiên về phương pháp giải trình tự DNA và sau đó, công bố phương pháp cho phép xác định nhanh các trình tự DNA dựa vào hoạt động của DNA polymerase trong quá trình tổng hợp DNA (Sanger, Coulson, 1975). Năm 1977, hai công trình nổi tiếng về giải trình tự DNA đã được công bố: Công trình của Sanger và đồng tác giả về kỹ thuật giải trình tự DNA sử dụng dideoxynucleotide để làm ngừng phản ứng tổng hợp DNA một cách ngẫu nhiên (Phương pháp dideoxy); Công trình của Maxam và Gilbert về kỹ thuật giải trình tự DNA bằng phương pháp hóa học, trong đó các đoạn DNA được đánh dấu, cắt ngẫu nhiên và điện di trên gel polyacrylamid. Hai phòng thí nghiệm đi tiên phong trong việc cho ra đời các thế hệ máy giải trình tự DNA tự động đầu tiên là Caltech (sau này được Applied Biosystems – ABI thương mại hóa), Phòng thí nghiệm Sinh học phân tử châu Âu – EMBL và Pharmacia-Amersham, sau này là General Electric Healthcare - GE (Smith *et al*., 1986; Ansorge *et al*., 1986, 1987). Kết quả của các nghiên cứu xây dựng và cải biến phương pháp cùng việc thương mại hóa các máy giải trình tự DNA những năm sau đó dẫn tới làn sóng ứng dụng rộng rãi các công nghệ này trong cộng đồng khoa học trên khắp thế giới.

Với thiết bị giải mã DNA tự động đầu tiên hoạt động dựa trên nguyên lý của phương pháp Sanger có cải biến (đánh dấu các ddNTP huỳnh quang thay vì phóng xạ), locus hoàn chỉnh của gen mã hóa hypoxanthineguanine phosphoribosyltransferase (HPRT) đã được xác định (Edwards *et al*., 1990). Năm 1996, máy giải trình tự DNA thương mại đầu tiên, ABI Prism 310 sử dụng điện di trên bản gel ra đời. Các thành phần chính của hệ thống bao gồm hệ mao quản, hệ laser, hệ đầu dò và xử lý tín hiệu. Hai năm sau đó, các ống mao quản tự động đã được thay thế cho các bản gel tự đổ tốn công sức trong hệ thống ABI Prism 3700. Máy giải trình tự DNA tự động này đã được sử dụng để giải mã thành công hệ gen người đầu tiên vào năm 2003 với nỗ lực 13 năm cố gắng của các thành viên Dự án hệ gen người và chi phí ước tính vào khoảng 3 tỷ USD. Tuy nhiên, do giá thành cao và tốc độ xử lý chậm nên việc giải trình tự DNA chủ yếu chỉ dừng ở các gen đơn lẻ, thường phục vụ các xét nghiệm chẩn đoán phân tử trong các phòng thí nghiệm y sinh.

Các hệ thống máy giải trình tự DNA thế hệ mới HT-NGS (high-throughput next generation sequencing)

Thành công trong việc giải mã hệ gen người vào năm 2003 đã mở ra một thời kỳ phát triển mới trong nghiên cứu về khoa học sự sống với rất nhiều kỹ thuật mới, phức tạp và hiện đại, được phát triển và ứng dụng. Một trong những công nghệ có sự phát triển mạnh mẽ và tạo ảnh hưởng ở quy mô toàn cầu là công nghệ giải trình tự DNA thế hệ mới, cho phép giải mã hiệu quả và nhanh chóng toàn bộ hệ gen sinh vật (whole genome sequencing). Nhiều công ty thương mại đã cho ra đời các hệ thống máy giải trình tự DNA thế hệ mới dựa trên các công nghệ HT-NGS. Sau Dự án hệ gen người, 454 sequencing đã được đưa ra thị trường vào năm 2005 bởi 454 và năm sau, Solexa cho ra đời Genome Analyzer. Tiếp đó, Agencourt phát triển SOLiD. Đây là ba hệ thống giải trình tự NGS lớn nhất khi đó với các ưu điểm là dung lượng lớn, độ chính xác cao và chi phí giảm. Các công ty này sau đó được mua bởi các hãng: Applied Biosystems/ Agencourt (2006), Roche/ 454 (2007), Illumina/ Solexa (2007). Tới nay, rất nhiều hệ thống máy giải trình tự thế hệ mới đã được phát triển bởi các hãng/ công ty như Applied Biosystem/ SOLiD; Roche/ 454; Illumina/ Solexa, MiSeq, HiSeq; Pacific Biosciences/ RS; Life technologies/ Ion Torrent PGM…và gần đây là Life technologies/ Ion Proton…cho phép giải mã nhanh toàn bộ hệ gen (Shendure, Ji, 2008; Metzker, 2010; Liu *et al*., 2012; Quail *et al*., 2012; Ferrarini *et al*., 2013).

Về nguyên lý, các hệ thống giải trình tự gen thế hệ mới được thực hiện dựa trên việc giải trình tự bằng tổng hợp (sequencing by synthesis, SBS) (Các thế hệ máy Roche/ 454, Life Technologies/ Ion Torrent và Illumina sử dụng) hoặc giải trình tự gắn nối (sequencing by ligation, SBL) (Applied Biosystems/ SOLiD). Công nghệ NGS được phát triển trên cơ sở các kỹ thuật chuẩn bị mẫu, giải trình tự, lắp ráp hệ gen, chú giải và so sánh hệ gen. Trong đó, công đoạn giải trình tự được thực hiện với 3 bước chính bao gồm: Chuẩn bị các đoạn DNA và gắn lên các giá bám; Khuếch đại các đoạn DNA trên giá bám bằng mồi đặc hiệu adapter; Giải trình tự bằng tổng hợp hoặc bằng gắn nối (Schuster, 2008;

Rusk, 2011). Với ưu thế thời gian đọc nhanh, dung lượng lớn (high-throughput), trình tự đọc được rất chính xác, các hệ thống giải trình tự gen thế hệ mới ngày càng được sử dụng rộng rãi (Poehlmann *et al*., 2007; Pettersson *et al*., 2009; Schadt *et al.*, 2010). Sự phát triển mạnh mẽ của công nghệ NGS dẫn tới một cuộc cách mạng trong công nghệ sinh học phân tử nói chung và công nghệ giải trình tự DNA nói riêng. NGS là một bước tiến vượt bậc về công nghệ giải trình tự DNA, cho phép đọc một lượng dữ liệu khổng lồ, từ 8 Gb đến 600 Gb. Nếu như trước đây, việc giải mã toàn bộ hệ gen rất phức tạp, khó khăn, chi phí lớn, thời gian dài thì ngày nay, với sự phát triển của công nghệ NGS, các chương trình, dự án giải mã 1.000 - 10.000 hệ gen người, 1,000 - 10.000 hệ gen động vật, 1.000 hệ gen thực vật có thể được thực hiện. Việc giải trình tự toàn bộ hệ gen của một sinh vật có thể được thực hiện tại nhiều phòng thí nghiệm, trong một thời gian ngắn thay vì phải kéo dài nhiều năm. Công nghệ NGS hiện được ứng dụng chủ yếu trong các dự án lớn nghiên cứu hệ gen người với mục đích phục vụ y học, metagenomics và các nghiên cứu đa hình hệ gen của những loài đã được giải mã toàn bộ hệ gen. Công nghệ này đã và đang tiếp tục phát triển nhằm giải mã mới hệ gen của những loài chưa có hệ gen tham chiếu. Trong lĩnh vực y học, NGS là một công cụ mạnh nhất cho phép phát hiện được các tác nhân gây bệnh, các đột biến với tỷ lệ thấp. Chính vì vậy, giải trình tự DNA thế hệ mới được ứng dụng trong phát hiện và định lượng các đột biến trong ung thư, nghiên cứu chẩn đoán bệnh di truyền…Tuy nhiên, phương pháp giải trình tự truyền thống (phương pháp Sanger) với đoạn đọc dài và độ chính xác cao vẫn tiếp tục được lựa chọn sử dụng trong nhiều trường hợp, đặc biệt ở các dự án quy mô nhỏ với dung lượng giới hạn ở Kb – Mb (Shendure, Ji, 2008).

GIẢI MÃ HỆ GEN NHÂN/ HỆ GEN BIỂU HIỆN CHI NHÂN SÂM

Chi Nhân sâm gồm nhiều loài cây thuốc có giá trị cao, trong đó có các loài chứa nhiều hợp chất tự nhiên có cấu tạo phân tử khá phức tạp, độc đáo, có hoạt tính tốt và có tác dụng tăng cường thể lực như Nhân sâm và Sâm Ngọc linh. Cùng với sự phát triển của công nghệ NGS, việc tiến hành các nghiên cứu giải mã hệ gen các loài thuộc chi Nhân sâm đã được lên kế hoạch đầu tư và tiến hành trong những năm gần đây. Các công bố chủ yếu tập trung ở các quốc gia có Sâm phân bố và được sử dụng thường xuyên làm dược phẩm như Trung Quốc, Hàn Quốc…Các nghiên cứu giải mã đầu tiên thường tập trung vào hệ gen biểu hiện của các loài trong chi Nhân sâm, xây dựng thư viện của các trình tự biểu hiện (ESTs) phục vụ nghiên cứu chức năng của các gen sau này như biểu hiện gen, marker phân tử, lập bản đồ di truyền... Năm 2010, Sun và đồng tác giả tại Viện Phát triển Thảo Dược Bắc Kinh đã công bố nghiên cứu đầu tiên về giải mã hệ gen biểu hiện của chi Nhân sâm trên đối tượng Sâm bắc mỹ (*P. quinquefolius*). Nhóm nghiên cứu đã sử dụng hệ thống đọc trình tự GS FLX Titanium với công nghệ 454 pyrocequencing và thu được hơn 200 nghìn kết quả đọc trình tự chất lượng cao với độ dài trung bình của mỗi kết quả là 427 bp từ thư viện cDNA của rễ cây *P. quinquefolius*. Phân tích kết quả giải trình tự cho thấy nhóm nghiên cứu đã thu được hơn 30.000 trình tự dịch mã riêng biệt, hơn 2/3 số trình tự này đã có mặt trong các ngân hàng trình tự đã biết, đặc biệt có khoảng 4.000 trình tự được xác định là mã hóa cho các enzyme tham gia vào các quá trình hóa sinh của *P. quinquefolius* bao gồm chu trình chuyển hóa đường, các amino acid, năng lượng, chất béo và các hợp chất thứ sinh…Đặc biệt, gần như tất cả các gen mã hóa enzyme đặc hiệu tham gia trực tiếp vào quá trình tổng hợp khung của các ginsenoside của *P. quinquefolius*, hoạt chất quan trọng quy định dược tính của các loại sâm, đều đã được xác định. Thêm vào đó, trong số các trình tự thu được, nhóm nghiên cứu còn xác định được 150 trình tự mã hóa cho các cytochrome P450 (CYP450) và 235 protein chuyển hóa các gốc đường (unique glycosyltranferase- UGT). Dựa vào các trình tự thu được, nhóm còn tiến hành nghiên cứu mức độ biểu hiện của một CYP450 và bốn UGT trên các mô của Sâm bắc mỹ dựa trên Real-time PCR (Sun *et al.*, 2010).

Năm 2011, hệ gen biểu hiện của một loài khác của chi Nhân sâm là *P. notoginseng* đã được công bố (Luo *et al.*, 2011). Tương tự như nghiên cứu năm 2010, nhóm đã tìm ra được 11 gen mã hóa cho enzyme tham gia tổng hợp khung của triterpene saponin, 174 trình tự mã hóa các cytochrome P450 và 242 UGT (unique glycosyltranferase) trên đối tượng rễ *P. notoginseng*. Ngoài ra, nhóm còn tiến hành các nghiên cứu mức độ biểu hiện của dammarenediol synthase (DS) là enzyme quan trọng của quá trình tổng hợp các triterpenes saponin (ginsenosides) và thấy rằng hoạt chất này biểu hiện mạnh nhất ở rễ Tam thất 4 năm tuổi. Các nhân tố điều hòa dịch mã như Myb, homeobox, WRKY, bHLH (basic helix-loop-helix)…và các trình tự lặp lại đơn giản (SSR) từ các trình tự thu được trên *P.*

notoginseng cũng được nghiên cứu (Luo *et al.*, 2011).

Các nghiên cứu giải mã hệ gen biểu hiện của *P. ginseng* cũng được tiến hành một cách đồng thời nhưng với quy mô lớn và toàn diện hơn. Năm 2011, Chen và đồng tác giả công bố thu được gần 32 nghìn trình tự biểu hiện (ESTs) dựa trên việc phân tích trình tự của thư viện cDNA của rễ Nhân sâm 11 năm tuổi bằng hệ thống giải trình tự GS FLX Titanium. Nhóm còn xác định được 9 gen đặc hiệu tham gia vào quá trình tổng hợp ginsengnosides, 133 gen mã hóa các cytochrome P450 và 235 gen mã hóa các glycosyltransferase (Chen *et al.*, 2011). Năm 2013, dựa vào kết quả của 2 lần chạy của hệ thống 454 pyrocequencing, Li và đồng tác giả đã thu được một kết quả phân tích khổng lồ về hệ gen biểu hiện (transcriptome) của *P. ginseng,* nhóm nghiên cứu đã thu được lần lượt 45849, 6172, 4041 và 3273 trình tự mã hóa từ phân tích thư viện cDNA của rễ, thân, lá và hoa của Nhân sâm, phát hiện ra tổng cộng 233 gen mã hóa enzyme tham gia tổng hợp ginsengnoside, 326 gen mã hóa cytochrome P450 và 129 gen mã hóa các glycosyltransferase. Ngoài ra, nhóm còn xác định được 14 trình tự mã hóa các microRNA có thể điều hòa việc tổng hợp protein của khoảng 100 gen đích và hơn 13 nghìn trình tự lặp lại đơn giản (SSRs) (Li *et al.*, 2013).

Các nghiên cứu giải mã hệ gen biểu hiện của chi Nhân sâm cũng được tiến hành nhưng với quy mô nhỏ hơn tại Canada và Hàn Quốc. Năm 2013, Wu và đồng tác giả tại Canada đã công bố hơn 41 nghìn kết quả giải trình tự các ESTs thu được qua việc phân tích các thư viện cDNA của từ rễ Sâm bắc mỹ (*P. quinquefolius*) ở các giai đoạn phát triển khác nhau, trong đó có 3955 trình tự mã hóa cho các gen tham gia tổng hợp các saponin (Wu *et al.*, 2013). Năm 2013, Ramya và đồng tác giả tại Trung tâm nghiên cứu Sâm Hàn Quốc, công bố nghiên cứu về 69 miRNAs bảo thủ dựa trên ngân hàng trình tự ESTs xây dựng trên thư viện cDNA từ hoa, là và rễ của Nhân sâm (*P. ginseng*) (Mathiyalagan *et al.*, 2013). Bảng 1 thống kê các kết quả giải mã trình tự hệ gen ở chi Nhân sâm sử dụng các công nghệ giải trình tự gen thế hệ mới.

Bảng 1. Các kết quả giải trình tự hệ gen các loài thuộc chi Nhân sâm

Chi/ Loài	Xây dựng thư viện	Dữ liệu (ENA)	Hệ thống giải trình tự	Model	Số đoạn đọc	Số lượng base
Panax	Amplicon	SRX576296	Illumina	Illumina HiSeq 2000	4.227.778	422.777.800
Panax	EST	SRX446787	LS454	454 GS FLX Titanium	311.861	163.822.811
Panax	EST	SRX446788	LS454	454 GS FLX Titanium	308.313	165.256.640
Panax	RNA-Seq	SRX480845	LS454	454 GS FLX	534.324	269.464.225
P. ginseng	EST	SRX017443	LS454	454 GS FLX Titanium	217.529	116.027.341
P. ginseng	EST	SRX181253	LS454	454 GS FLX Titanium	637.238	343.780.247
P. ginseng	EST	SRX181258	LS454	454 GS FLX Titanium	565.651	299.965.032
P. ginseng	EST	SRX181262	LS454	454 GS FLX Titanium	634.773	336.608.507
P. ginseng	EST	SRX181263	LS454	454 GS FLX Titanium	598.289	317.475.738
P. ginseng	RNA-Seq	ERX149253	Illumina	Illumina HiSeq 2000	20.192.616	3.634.670.880
P. ginseng	RNA-Seq	ERX149254	Illumina	Illumina HiSeq 2000	19.175.503	3.451.590.540
P. ginseng	RNA-Seq	ERX149263	Illumina	Illumina HiSeq 2000	19.332.200	3.479.796.000

P. ginseng	RNA-Seq	SRX397736	Illumina	Illumina HiSeq 2000	6.168.036	302.233.764
P. ginseng	RNA-Seq	SRX397758	Illumina	Illumina HiSeq 2000	6.168.015	302.232.735
P. ginseng	RNA-Seq	SRX573758	Illumina	Illumina HiSeq 2000	16.610.381	813.908.669
P. ginseng	WGS	SRX521223	Illumina	Illumina HiSeq 2000	2.737.712	553.017.824
P. ginseng	WGS	SRX521225	Illumina	Illumina HiSeq 2000	1.729.226	380.429.720
P. ginseng	WGS	SRX521227	Illumina	Illumina HiSeq 2000	2.424.098	533.301.560
P. ginseng	RNA-Seq	ERX137460	Illumina	Illumina HiSeq 2000	19.929.279	3.587.270.220
P. ginseng cv. Chunpoong	WGS	SRX476093	Illumina	Illumina HiSeq 2000	150.000.000	30.300.000.000
P. ginseng cv. Yunpoong	WGS	SRX481170	Illumina	Illumina HiSeq 2000	3.000.000	606.000.000
P. notoginseng	EST	SRX017444	LS454	454 GS FLX	188.185	99.149.482
P. notoginseng	RNA-Seq	SRX495493	Illumina	Illumina HiSeq 2000	29.402.101	2.940.210.100
P. notoginseng	RNA-Seq	SRX378873	Illumina	Illumina HiSeq 2000	32.629.487	5.873.307.660
P. notoginseng	RNA-Seq	SRX378878	Illumina	Illumina HiSeq 2000	34.062.655	6.131.277.900
P. notoginseng	RNA-Seq	SRX378880	Illumina	Illumina HiSeq 2000	32.520.520	5.853.693.600
P. quinquefolius	EST	SRX012184	LS454	454 GS FLX	209.747	112.585.959
P. quinquefolius	WGS	SRX481173	Illumina	Illumina HiSeq 2000	6.761.830	1.365.889.660

Các cụm từ viết tắt:

WGS	Whole genome sequencing	Giải trình tự toàn bộ hệ gen
EST	Expressed sequence tag	Các đoạn trình tự gen biểu hiện
RNA-Seq	Whole Transcriptome Shotgun Sequencing	Giải trình tự toàn bộ hệ gen biểu hiện
Amplicon	Amplification based selection	Sàng lọc dựa trên nhân bản gen
ENA	European Nucleotide Archive: http://www.ebi.ac.uk/ena	Cơ sở dữ liệu Nucleotide châu Âu

GIẢI MÃ HỆ GEN LỤC LẠP

Hệ gen lục lạp ở thực vật

Lục lạp (chloroplast) là bào quan phổ biến và đóng vai trò quan trọng trong thế giới thực vật, là nơi thực hiện chức năng quang hợp, tạo ra năng lượng cho tế bào. Lục lạp có hệ thống di truyền riêng (có DNA) và hệ tổng hợp protein độc lập (có chứa ribosome, các loại RNA). DNA của lục lạp cũng có cấu tạo giống DNA của prokaryote (vi khuẩn và tảo lam) có cấu trúc vòng, không chứa histon. DNA của lục lạp chứa thông tin mã hóa cho một số protein mà lục lạp tự tổng hợp trên ribosome của mình. Còn các protein khác do tế bào cung cấp. DNA lục lạp là nhân tố di truyền ngoài nhiễm sắc thể. Người ta cho rằng trong quá trình phát sinh chủng loại, lục lạp được hình thành do sự cộng sinh của một loài vi khuẩn lam trong tế bào (Dyall *et al*., 2004).

Hệ gen lục lạp ở các loài thực vật có cấu trúc dạng mạch vòng với kích thước dao động từ 70 – 217 kb, chứa khoảng 130 gen (Sugiura, 1995). Hệ gen lục lạp thực vật bao gồm hai vùng lặp lại đảo chiều (inverted repeats – IPs) ngăn cách bởi hai vùng DNA đặc trưng, vùng bản sao đơn lớn (large single-copy region – LSC) và vùng bản sao đơn nhỏ (small single-copy region – SSC) (Jansen *et al*., 2005). Các nghiên cứu cấu trúc phân tử hệ gen lục lạp ở hầu hết các loài thực vật bậc cao cho thấy tỷ lệ thay thế trong hệ gen lục lạp thực vật thấp hơn rất nhiều so với hệ gen nhân và chúng cũng có mức tái tổ hợp rất thấp, di truyền theo một dòng cha mẹ (Wolfe *et al*., 1987; Ravi *et al*., 2008). Với tốc độ tiến hóa chậm, khá bảo thủ về kích thước, cấu trúc và thành phần gen, đặc biệt giữa các loài trong cùng chi, gen trên lục lạp thường được sử dụng để nhận dạng và đánh giá mối quan hệ di truyền của các loài ở nhiều cấp độ (Olmstead, Palmer, 1994; Ravi *et al*., 2008).

Giải mã hệ gen lục lạp

Trước đây, các nghiên cứu thường tập trung giải mã một vài vùng gen lục lạp của nhiều loài hoặc để giải mã hệ gen hoàn chỉnh sử dụng PCR thông thường, hàng loạt vùng gen ở các locus bảo thủ được nhân bản và lắp ráp. Tuy nhiên, hướng tiếp cận này mất rất nhiều thời gian và khó thực hiện trên nhiều loài. Gần đây, với sự phát triển của các công nghệ NGS, lục lạp/ DNA tổng số được tách chiết và giải mã toàn bộ. Mặc dù các đoạn đọc được khá ngắn, tuy nhiên, với kích thước hệ gen lục lạp không quá lớn và không quá phức tạp so với hệ gen nhân, cùng các công nghệ giải trình tự cũng như lắp ráp, hiệu chỉnh khả thi, số lượng các loài được giải mã lục lạp tiếp tục tăng nhanh. Đến nay, trình tự hệ gen lục lạp hoàn chỉnh của rất nhiều loài thực vật đã được giải mã (http://www.ncbi.nlm.nih.gov/genome). Các phân tích đánh giá trên cơ sở dữ liệu của toàn bộ hệ gen lục lạp trở nên khả thi và đây là công cụ hữu hiệu, là nguồn thông tin quý giá giúp giải quyết các mối quan hệ phát sinh chủng loài, quá trình thích nghi, tìm kiếm các mã vạch phân tử ngắn, đặc trưng... Đến nay rất nhiều hệ gen lục lạp đã được giải mã hoàn chỉnh sử dụng các công nghệ khác nhau và được phân tích so sánh với các dữ liệu hệ gen lục lạp đã công bố để tìm kiếm sự đa dạng, khác biệt giữa các hệ gen, phân tích nguồn gốc tiến hóa.

Ku *et al*., (2013) đã lắp ráp hoàn chỉnh hệ gen lục lạp loài *Catharanthus roseus* (L.) G. Don (họ Apocynaceae) – loài cây thuốc quan trọng được thương mại rộng rãi trên thị trường dược phẩm trị liệu, so sánh với trình tự hệ gen lục lạp của 2 loài khác là *Coffea arabica* (họ Rubiaceae) và *Asclepias syriaca* (họ Apocynaceae). Phân tích cho thấy sự đa hình đáng kể về thành phần gen trong họ Apocynaceae, trong đó có mặt/ không có mặt 3 gen quan trọng (*accD*, *clpP* và *ycf1*), cũng như khác nhau ở các vùng không mang mã (*rps2-rpoC2* và IRb-*ndhF*). Nghiên cứu cũng đã tìm kiếm được 41 chỉ thị SSR đặc trưng cho *S. roseus* phục vụ chọn tạo giống và xây dựng mối quan hệ phát sinh chủng loài. Curci *et al*. (2015) đã giải mã hệ gen lục lạp loài Artichoke sử dụng công nghệ Illumina GAIIx kết hợp giải trình tự toàn bộ hệ gen và thư viện BAC (các đoạn đọc paired-end có kích thước 75 bp). Kết quả, lục lạp với kích thước 152.529 bp đã được giải mã hoàn chỉnh và so sánh với 8 hệ gen lục lạp các loài thuộc họ Asteraceae. Tới 127 chỉ thị SSRs tiềm năng ứng dụng trong các nghiên cứu quần thể ở chi Cynara đã được phát hiện và 8 vùng mang mã chứa nhiều thông tin nhất đã được đánh giá khả năng sử dụng làm mã vạch đặc trưng trong họ này…

Giải mã hệ gen lục lạp một số loài thuộc chi Nhân sâm

Đối với các loài thuộc chi Nhân sâm, Kim và Hee (2004) đã thực hiện giải mã toàn bộ hệ gen lục lạp của Sâm triều tiên (*P. schinseng* Nees) (mã số AY582139). Hệ gen lục lạp là DNA sợi đôi vạch vòng, bao gồm 156.318 bp, chứa một cặp IR (IRa và IRb) với kích thước mỗi vùng lặp là 26.071, ngăn cách bởi vùng LSC có kích thước 86.106 bp và vùng SSC có kích thước 18.070 bp. Hệ gen bao gồm 114 gen (75 gen mã hóa cho các peptide, 30 tRNA gene, 4 rRNA gene và 5 khung đọc mở bảo thủ [*ycf*s]). Mười sáu gen có 1 intron trong khi 2 gen có 2 intron. Nghiên cứu cũng tiến hành so sánh hệ gen lục lạp Sâm triều tiên với hệ gen lục lạp của 17 loài thực vật có mạch nhằm tìm hiểu các mô hình tiến hóa các đoạn trình tự mang mã và không mang mã của gen, cũng như đánh giá quan hệ phát sinh chủng loài dựa trên trình tự hệ gen lục lạp. Dong *et al*. (2013) đã thực hiện giải mã hệ gen lục lạp *P. notoginseng* sử dụng phương pháp PCR và so sánh với *P. ginseng*. Các vùng đa hình nhất được nhận dạng. Trong đó, các vùng có kích thước ngắn nhất với khả năng nhận dạng tương đương so với các vùng có độ dài truyền thống được sử dụng làm mã vạch phân tử ngắn. Kết quả cho thấy, hệ gen lục lạp của *P. tonoginseng* có kích thước 156.387 bp và chỉ có 464 (0,3%) sai khác giữa hai hệ gen. Các vùng intron *rps16* và hai vùng mã hóa gen *ycf1*, *ycf1a* và *ycf1b* là các mã vạch hữu hiệu có thể sử dụng trong nhận dạng các loài thuộc chi Nhân sâm với mức độ phân biệt đạt được tương

ứng là 83,33% (280 bp của *rps16*), 91,67% (60 bp của *ycf1a*) và 100% (100 bp của *ycf1b*). Zhao *et al*. (2015) đã xác định trình tự hệ gen lục lạp 4 mẫu *P. ginseng* C.A. Meyer (*P. ginseng*) – loài dược liệu đặc biệt có giá trị và thường được sử dụng trong các bài thuốc cổ truyền Trung Hoa. Bốn mẫu bao gồm Damaya (DMY), Ermaya (EMY), Gaolinshen (GLS), Yeshanshen (YSS). Trình tự toàn bộ hệ gen lục lạp của DMY, EMY và GLS là 156.354 bp, của YSS là 156.355 bp. Trình tự hệ gen của 3 chủng đầu tương tự nhau, trong khi ở chủng YSS, 1 bp được chèn vào vị trí 5472. Các phân tích hệ gen học so sánh cho thấy thành phần gen, thành phần GC và thứ tự của gen trong DMY tương tự như ở các loài họ hàng, trong khi đa hình trình tự vùng IR thấp hơn. Nghiên cứu cũng thực hiện các đánh giá đa hình các allele hiếm và sự thích ứng với các thay đổi môi trường của hệ gen lục lạp. Binh Nguyen và nhóm nghiên cứu tại Trường Đại học Quốc gia Seoul (2015) đã thông báo kết quả giải mã hệ gen lục lạp của một mẫu Sâm Ngọc linh với kích thước 155.992 bp. Bảng 2 thống kê một số nghiên cứu và kết quả giải trình tự hệ gen lục lạp các loài thuộc chi Nhân sâm trên thế giới.

Bảng 2. Thống kê một số kết quả giải trình tự hệ gen lục lạp các loài thuộc chi Nhân sâm

Tên loài	Mã số		Kích thước	Bộ dữ liệu	Công bố
P. notoginseng	NC_026447	KJ566590	156.387		Dong *et al.* (2014)
P. schinseng	NC_006290	AY582139	156.318		Kim, Hee (2004)
P. ginseng isolate damaya		KC686331	156.354	SRR1251992	Zhao *et al.* (2015)
P. ginseng isolate Ermaya		KC686332	156.354	SRR1252006	Zhao *et al.* (2015)
P. ginseng isolate Gaolishen		KC686333	156.354	SRR1252007	Zhao *et al.* (2015)
P. ginseng		KF431956	156.355	SRR1252008	Zhao *et al.* (2015)
P. vietnamensis		KP036471	155.992		Binh Nguyen *et al*. (2015)
P. notoginseng		KR021381	156.387		Chưa công bố

DỮ LIỆU HỆ GEN VÀ MÃ VẠCH PHÂN TỬ

Mã vạch phân tử ở thực vật dựa trên các đoạn DNA ngắn trong hệ gen lục lạp và hệ gen nhân hiện nay đang có ảnh hưởng rất lớn và phát triển rất nhanh chóng, cho phép phân biệt các loài sử dụng các trình tự DNA. DNA lục lạp ở thực vật, có thể đảm bảo được tốc độ tiến hóa, có thể thao tác dễ dàng, phù hợp làm mã vạch DNA. Tuy nhiên, có thể trong quá trình bảo quản, DNA lục lạp có số lượng bản sao không nhiều, dẫn tới việc không thu được DNA và cần kết hợp với DNA nhân (Kress *et al.*, 2007). Việc xây dựng được bộ mã vạch phân tử xác định các loài, phục vụ bảo tồn quỹ gen và giám sát thương mại có ý nghĩa khoa học và thực tiễn cấp bách. Hiện nay, mã vạch DNA, đóng vai trò như một công cụ trong phân loại học, đã được xây dựng và mỗi ngày số lượng các trình tự mã vạch trong Ngân hàng gen quốc tế (GenBank) đang được tăng lên đáng kể. Trên đối tượng thực vật, các vùng DNA mã vạch được sử dụng để phân loại thường là các trình tự thuộc hệ gen lục lạp và hệ gen nhân…(Kress *et al.*, 2005; Pennisi, 2007; Hollingsworth *et al.*, 2008; Lahaye *et al.*, 2008; Hollingsworth, 2009; Lê Thị Thu Hiền *et al*., 2012).

Để khắc phục các nhược điểm của phương pháp phân loại dựa trên kiểu hình, các phương pháp phân loại dựa trên vật liệu di truyền đã được sử dụng để định loại chính xác hơn các loài trong chi Nhân sâm. Các kỹ thuật này chủ yếu dựa trên PCR như AFLP (Amplified fragment length polimorphism) (Ha *et al.*, 2002); kỹ thuật RADP (Random amplified polimorphic DNA) (Shaw, But, 2007); hay sử dụng các marker EST-SSR (Expression sequence tags - simple sequence repeats) (Zhang *et al.*, 2011)… hoặc các kỹ thuật định dạng phân tử như RFLP (Restriction fragment length polimorphism). Các phương pháp này bước đầu đều đã phát huy hiệu quả và có thể sử dụng để xác định tính chính xác của cây phân loại dựa trên kiểu hình. Ngoài ra, một số nghiên cứu về mã vạch phân tử cũng như giải mã một phần genome các loài thuộc chi này đã được tiến hành (Brozynska *et al*., 2014; Dong *et al*., 2014). Trên thế

giới, việc sử dụng phương pháp mã vạch DNA để phân loại các loài sâm thuộc chi Nhân sâm đã phổ biến và thông dụng từ những năm giữa thập kỷ 90 của thế kỷ trước. Các mã vạch phân tử được sử dụng để phân loại các loài sâm thuộc chi Nhân sâm tương đối nhiều, chúng có thể nằm trong genome nhân như vùng ITS (internal transcribed spacers)*, 18S rRNA;* trong ty thể như *nad1* hoặc nằm trong hệ gen lục lạp như *matK, psbA-trnH, psbK-I, pspM-trnD, rps16, trnC-trnD*…. Trong các chỉ thị này thì vùng ITS*, psbA-trnH* và *trnC-trnD* cho thấy nhiều đa hình đơn nucleotide hơn cả và có thể dùng để xác định loài và phân loại nhóm cho chi Nhân sâm (Komatsu *et al*., 2001; Zhu *et al.*, 2003; Lee, tWen, 2004; Zuo *et al.*, 2011). Năm 1996, Wen và đồng tác giả đã công bố cây phát sinh chủng loại của 12 loài sâm khác nhau thuộc chi Nhân sâm phân bố ở Bắc Mỹ và Đông Á dựa trên trình tự vùng barcode ITS có độ dài từ 606 đến 608 bp gồm vùng ITS1, vùng xen 5,8S và ITS2. Komatsu và đồng tác giả (2001) đã giải trình tự các gen 18S và *matK* nhằm nghiên cứu so sánh đặc điểm di truyền của *P. vietnamensis* và *P. quinquefolium*. Kết quả cho thấy hai loài hoàn toàn tương đồng ở gen 18S và khác nhau ở 10 vị trí trên gen *matK*. Năm 2004, Lee và Wen công bố một barcode khác của chi Nhân sâm là vùng *trnC-trnD* nằm xen giữa hai gen *trnC* và *trnD* trong hệ gen lục lạp. Dựa trên trình tự vùng này kết hợp với ITS, nhóm nghiên cứu đã xây dựng cây phát sinh chủng loại của 18 loài trong chi Nhân sâm và 2 loài thuộc chi *Aralia*. Bên cạnh đó, một số barcode khác như vùng *trnK* kết hợp với vùng 18S rRNA cũng được các nhà khoa học ở đại học dược Toyama sử dụng và xây dựng thành công cây phát sinh chủng loại của 13 loài (Fushimi *et al*., 1996; Zhu *et al.*, 2003). Trong nghiên cứu này, *P. vietnamensis* var. *fuscidiscus* được xác định là một thứ của *P. vietnamensis* Ha et Grushv. có phân bố ở Vân Nam, Trung Quốc và thứ này khác với *P. vietnamensis* ở 4 vị trí trên gen *trnK*. Kết quả phân tích dữ liệu thông tin gen 18S-rRNA và *trnK* cho thấy *P. vietnamensis* Ha et Grushv. và *P. vietnamensis* var. *fuscidiscus* có mối quan hệ di truyền gần gũi và chung nhánh với *P. zingiberensis* có nguồn gốc ở Vân Nam, Trung Quốc. Các công trình này đã chứng minh được việc sử dụng các vùng chỉ thị barcode để phân loại chi Nhân sâm là hoàn toàn khả thi và mở ra cơ hội xây dựng một bộ mã vạch phân tử hoàn chỉnh cho chi này. Các nghiên cứu xây dựng cây phát sinh chủng loại chi Nhân sâm từ đó đến nay đều sử dụng trình tự vùng ITS của các mẫu như một tiêu chuẩn để tham chiếu cũng như kết hợp với các barcode khác để có kết quả toàn diện hơn (Lee, Wen, 2004; Zuo *et al.*, 2011; Chen *et al.*, 2013). Gần đây, vùng ITS2 có độ dài 218-235 bp đã được nhiều nhóm nghiên cứu sử dụng như một barcode chuẩn để phân biệt các loài sâm. Trong công trình công bố gần đây, Ali và đồng tác giả (2012) đã phân loại được 12 loài với vùng trình tự ITS. Trình tự này cũng được Chen và đồng tác giả (2013) nghiên cứu để xây dựng cây phân loại. Trong đó, nhóm nghiên cứu đã chỉ rõ trong vùng trình tự ITS2 tương đối ngắn, các loài sâm khác nhau thuộc chi Nhân sâm sẽ biểu hiện từ 2-3 đa hình đơn nucleotide (SNPs) (Chen *et al.*, 2013).

GIẢI MÃ GEN VÀ HỆ GEN CÁC LOÀI THUỘC CHI NHÂN SÂM Ở VIỆT NAM

Trong khuôn khổ các chương trình khoa học và công nghệ các cấp, khá nhiều đề tài thực hiện giải mã một phần hoặc toàn bộ hệ gen ở một số đối tượng sinh vật đã được thực hiện như: (i) Đề tài "Nghiên cứu giám định gen hài cốt liệt sỹ bằng kỹ thuật gen" (thuộc Chương trình KC.04.23/01-05); (ii) Đề tài "Nghiên cứu giải mã genome ty thể các tộc người Việt Nam và định hướng ứng dụng" (Chương trình KC.04.25/01-05); (iii) Đề tài "Nghiên cứu giải trình tự một phần bộ gen và xây dựng cơ sở dữ liệu genome Tôm sú (*P. monodon*)" (Chương trình Công nghệ sinh học Nông nghiệp – Thủy sản, 2008-2010); (iv) Đề tài "Giải mã, phân tích hệ gen của các chủng virus lở mồm long móng đang lưu hành ở Việt Nam và ứng dụng trong chuẩn đoán và đinh hướng sản xuất vaccine" (Chương trình NAFOSTED, 2010-2012)… (v) Đề tài "Nghiên cứu biến đổi gen, nhiễm sắc thể ở những người có nồng độ dioxin trong máu cao" (Chương trình khoa học và công nghệ trọng điểm cấp Nhà nước, 2011-2015); (vi) Đề tài "Giải trình tự hệ gen loài vi tảo biển dị dưỡng của Việt Nam *Schizochytrium mangrovei* PQ6 (thuộc Chương trình KC.04.20/11-15); (vii) Đề tài "Nghiên cứu giải mã genome một số giống lúa địa phương của Việt Nam" (Chương trình Nghị định thư Việt Nam – Vương quốc Anh)…

Đối với chi Nhân sâm, Việt Nam có các loài mọc tự nhiên rất có giá trị làm thuốc như Sâm vũ diệp (*P. bipinnatifidus*), Tam thất hoang (*P. stipuleanatus*), Sâm Ngọc linh (*Panax vietnamensis* Ha et Grushv.), Sâm lai châu (*Panax vietnamensis* var. *fuscidiscus*). Trong đó, Sâm Ngọc linh là loài đặc biệt có giá trị khoa học và kinh tế. Sâm Ngọc linh được xác định là một cây thuốc quý của Việt Nam với nhiều thành phần saponin, hàm lượng các acid amine, các chất khoáng vi lượng trong củ, lá và

rễ hơn nhiều những loài sâm khác (Sách Đỏ Việt Nam, Phần II. Thực vật, 2007; Lã Đình Mỡi *et al*., 2013; Phan Kế Long *et al*., 2014). Sâm ngọc linh được phát hiện vào năm 1973 và đến năm 1985 mới được công bố là hoàn toàn mới đối với khoa học. Đến nay, Sâm ngọc linh chỉ được phát hiện ở vùng núi Ngọc Linh thuộc hai tỉnh Quảng Nam và Kon Tum, Gia Lai và Lâm Đồng. Ngọc Linh là dãy núi cao thứ hai của Việt Nam, có tọa độ địa lý từ 107°50' – 108°7' kinh tuyến Đông và từ 15°0' – 15°10' vĩ tuyến Bắc, đỉnh cao nhất là Ngọc Linh cao 2598 m. Những điểm vốn trước đây có Sâm Ngọc linh mọc tự nhiên, chủ yếu tập trung ở địa bàn của hai huyện Đăk Tô (tỉnh Kon Tum) và Trà My (tỉnh Quảng Nam). Tuy nhiên, do vùng phân bố hạn chế và việc khai thác quá mức đã khiến Sâm Ngọc linh trở nên cực hiếm trong tự nhiên. Sâm Ngọc linh đã được đưa vào danh lục đỏ của IUCN (2003) và danh sách các loài hạn chế khai thác và sử dụng vì mục đích thương mại (Nghị định 32/2006/NĐ-CP ngày 31 tháng 3 năm 2006 về quản lý thực vật rừng, động vật rừng nguy cấp quý hiếm).

Để bảo tồn và phát triển cây dược liệu quý hiếm này, một số nhóm nghiên cứu đã thực hiện tái sinh và nhân vô tính Sâm Ngọc linh và phân tích hoạt chất saponin. Những thử nghiệm bước đầu đã thu được cây, nhưng cây yếu, khó phát triển ngoài điều kiện tự nhiên (Nguyễn Ngọc Dung, 1995). Những năm gần đây, nhóm nghiên cứu của Dương Tấn Nhựt tại Viện Nghiên cứu khoa học Tây Nguyên thuộc Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã thành công trong nhân vô tính Sâm Ngọc linh có chất lượng cho sản xuất (Dương Tấn Nhựt *et al*., 2010; 2012a, b) trong khuôn khổ các đề tài "Hệ thống nuôi cấy lớp mỏng tế bào trong nghiên cứu chương trình phát sinh hình thái và bảo tồn cây Sâm Ngọc linh"; "Nghiên cứu nhân giống vô tính và sản xuất sinh khối rễ cây Sâm Ngọc linh (*Panax vietnamensis* Ha et Grushv.) (2008-2011)…

Các nghiên cứu về nhận dạng hình thái và sử dụng chỉ thị phân tử các loài thuộc chi Nhân sâm trên cơ sở phân tích một số vùng DNA đã và đang được triển khai thực hiện, tuy nhiên, ở mức độ và quy mô tương đối hạn chế. Việc phân loại chủ yếu dựa trên các đặc điểm hình thái của thân, lá, rễ của cây sâm kết hợp với phân tích các hợp chất saponin (Lã Đình Mỡi *et al*., 2013). Năm 2007, Nguyễn Tập và đồng tác giả đã sử dụng kỹ thuật RAPD để xây dựng cơ sở dữ liệu DNA của một số cây thuốc quý trong đó có Sâm Ngọc linh (Nguyễn Tập *et al*., 2007). Việc sử dụng các mã vạch phân tử đã được áp dụng nhưng chưa phong phú và toàn diện. Các vùng gen mã vạch được sử dụng chủ yếu là vùng *matK* và ITS (Nguyễn Thị Phương Trang *et al*., 2011; Nguyễn Văn Đạt, Trần Thị Phương Anh, 2013). Năm 2011, Nguyễn Thị Phương Trang và đồng tác giả đã công bố phát hiện về một loài sâm mới ở Việt Nam dựa trên phân tích các sai khác trong trình tự vùng ITS so với các loài khác thuộc chi Nhân sâm phân bố ở Đông Á (Nguyễn Thị Phương Trang *et al*., 2011). Vũ Huyền Trang và đồng tác giả (2013) đã nghiên cứu xây dựng mã vạch DNA cho Sâm Ngọc linh trên cơ sở 5 chỉ thị DNA mã vạch *psbA-trnH, matK, trnL, rbcL và* ITS. Nhóm tác giả đã chứng minh trong 5 chỉ thị mã vạch nghiên cứu, *psbA-trnH* là chỉ thị có tiềm năng nhất, cho phép phân biệt Sâm Ngọc linh với các loài sâm khác trên thế giới với độ chính xác cao. Việc xác định được các chỉ thị phân tử cho phép xác định chính xác được loài sâm này đóng vai trò quan trọng trong việc bảo tồn, sản xuất và đảm bảo chất lượng sản phẩm cho thương hiệu Sâm Ngọc linh.

KẾT LUẬN

Hệ thống giải trình tự gen thế hệ mới ra đời là cuộc cách mạng trong đổi mới trong công nghệ, là cơ sở khoa học cho sự phát triển của rất nhiều lĩnh vực liên quan. Đối với các loài thuộc chi Nhân sâm, như đã phân tích ở trên, trên thế giới, nhiều nghiên cứu tập trung giải mã hệ gen biểu hiện, hệ gen lục lạp và các vùng gen có giá trị trong nhận dạng. Trong nước, chỉ có một số nghiên cứu về chỉ thị phân tử được tiến hành và chưa có nghiên cứu thực hiện việc giải mã trình tự hoàn chỉnh hệ gen. Vì vậy, các nghiên cứu liên quan đến việc giải mã hệ gen, xây dựng cơ sở dữ liệu hệ gen của các loài thuộc chi Nhân sâm có ý nghĩa khoa học và thực tiễn. Thông tin về hệ gen được giải mã và các mã vạch phân tử được xác định hỗ trợ cho các nghiên cứu cơ bản và ứng dụng, trong đó có các nghiên cứu về mối quan hệ di truyền, lập bản đồ và xây dựng hệ thống các chỉ thị phân tử, nghiên cứu chức năng gen, hỗ trợ công tác nhận dạng, bảo tồn và khai thác, sử dụng bền vững nguồn gen quý hiếm này.

Lời cảm ơn: *Công trình được thực hiện trong khuôn khổ đề tài: "Giải mã hệ gen lục lạp của Sâm Ngọc linh (Panax vietnamensis* Ha et Grushv.*)", mã số VAST02.01/16-17, thuộc các hướng khoa học công nghệ ưu tiên cấp Viện Hàn lâm Khoa học và Công nghệ Việt Nam và nhiệm vụ: "Tạo cơ sở dữ liệu mã*

vạch ADN cho các loài cây nghiên cứu" thuộc đề tài: "Xây dựng cơ sở dữ liệu mã vạch ADN (DNA barcode) cho một số loài cây lâm nghiệp gỗ lớn, lâm sản ngoài gỗ có giá trị kinh tế" (2014-2017) (Chương trình Trọng điểm Phát triển và Ứng dụng Công nghệ sinh học trong lĩnh vực nông nghiệp và phát triển nông thôn đến năm 2020, Bộ Nông nghiệp và Phát triển Nông thôn).

TÀI LIỆU THAM KHẢO

Nguyễn Ngọc Dung (1995) Nhân giống Sâm Ngọc linh (*Panax vietnamensis* Ha et Grushv.) bằng con đường sinh học. *Nhà xuất bản Nông nghiệp*: 43-100.

Nguyễn Văn Đạt, Trần Thị Phương Anh (2013) Bước đầu nghiên cứu xây dựng khóa định loại các chi trong họ Ngũ gia bì (Araliaceae) ở Việt Nam. *Hội nghị khoa học toàn quốc về Sinh thái và tài nguyên sinh vật lần thứ 5*: 44-51.

Lê Thị Thu Hiền, Hugo de Boer, Nông Văn Hải, Lê Thanh Hương, Nguyễn Mai Hương, Lars Bjork (2012) Mã vạch phân tử DNA và hệ thống dự liệu mã vạch sự sống. *Tạp chí Công nghệ sinh học*, 10(3): 393-405.

Phan Kế Long, Vũ Đình Duy, Phan Kế Lộc, Nguyễn Giang Sơn, Nguyễn Thị Phương Trang, Lê Thị Mai Linh, Lê Thanh Sơn (2014) Nghiên cứu đặc điểm di truyền của các mẫu sâm thu ở Lai Châu trên cơ sở phân tích trình tự nucleotide vùng gen *mat*K và ITS-rRNA. *Tạp chí Công nghệ sinh học* 12(2): 327-337.

Dương Tấn Nhựt, Hoàng Xuân Chiến, Nguyễn Bá Trực, Nguyễn Bá Nam, Trần Xuân Tình, Vũ Quốc Luận, Nguyễn Văn Bình, Vũ Thị Hiền, Trịnh Thị Hương, Nguyễn Cửu Thành Nhân, Lê Nữ Minh Thùy, Lý Thị Mỹ Nga, Thái Thương Hiền, Nguyễn Thành Hải (2010) Nhân giống vô tính cây Sâm Ngọc linh (*Panax vietnamensis* Ha et Grushv.). *Tạp chí Công nghệ sinh học* 8(3B): 1211-1219.

Lã Đình Mỡi, Châu Văn Minh, Trần Văn Sung, Phạm Quốc Long, Phan Văn Kiệm, Trần Huy Thái, Trần Minh Hợi, Ninh Khắc Bản, Lê Mai Hương (2013) Họ Nhân sâm (Araliaceae Juss.) - Nguồn hoạt chất sinh học đa dạng và đầy triển vọng ở Việt Nam. *Hội nghị khoa học toàn quốc về Sinh thái và tài nguyên sinh vật lần thứ 5*: 1152-1158.

Sách Đỏ Việt Nam. Phần II. Thực vật (2007) Nhà xuất bản Khoa học tự nhiên và Công nghệ, Hà Nội.

Nguyễn Tập, Phạm Thanh Huyền, Lê Thanh Sơn, Ngô Đức Phương, Võ Văn Trại, Đinh Đoàn Long, Hoàng Thị Hòa (2007) Sử dụng chỉ thị ADN (RAPD-PCR) trong nghiên cứu đa dạng di truyền và góp phần phân loại một số loài cây thuốc định hướng công tác bảo tồn và tiêu chuẩn hóa dược liệu ở Việt Nam. *Hội nghị Dược liệu toàn quốc lần thứ hai*: 288-301.

Nguyễn Thị Phương Trang, Lê Thanh Sơn, Nguyễn Giang Sơn, Phan Kế Long (2011) Phát hiện về một loài sâm mới *Panax* sp. (Araliaceae) ở Việt Nam. *Tạp chí Dược học*, 10: 59-63.

Vũ Huyền Trang, Hoàng Đăng Hiếu, Chu Hoàng Hà (2013) Nghiên cứu xây dựng mã vạch DNA cho việc phân loại nhận dạng cây Sâm Ngọc linh. *Hội nghị khoa học công nghệ sinh học toàn quốc*: 1100-1104.

Ali MA, Al-Hemaid FM, Lee J, Choudhary RK, Pandey AK, Al-Harbi NA (2012) Assessing nrDNA ITS2 sequence based molecular signature of ginseng for potential in quality control of drug. *Afr J Pharm Pharmacol* 6(39): 2775-2781.

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res* 15(11): 4593-4602.

Ansorge W, Sproat BS, Stegemann J, Schwager C (1986) A non-radioactive automated method for DNA sequence determination. *J Biochem Biophys Methods* 13(6): 315-323.

Becker C, Shutov AD, Nong VH, Senyuk VI, Jung R, Horstmann C, Fischer J, Nielsen NC, Muntz K (1995) Purification, cDNA cloning and characterization of proteinase B, an asparagine-specific endopeptidase from germinating vetch (*Vicia sativa* L.) seeds. *Eur J Biochem* 228(2): 456-462.

Binh N, Kyunghee K, Young-Chang K, Sang-Choon L, Ji ES, Junki L, Nam-Hoon K, Woojong J, Hong-Il C, Tae-Jin Y (2015) The complete chloroplast genome sequence of *Panax vietnamensis* Ha et Grushv (Araliaceae). *Mitocondrial DNA*, DOI:10.3109/19401736.2015.1110810.

Brozynska M, Furtado A, Henry RJ (2014) Direct chloroplast sequencing: Comparison of sequencing platforms and analysis tools for whole chloroplast barcoding. *PLOS One* 9(10): e110387. DOI:10.1371/journal.pone.0110387.

Chen S, Luo H, Li Y, Sun Y, Wu Q, Niu Y, Song J, Lv A, Zhu Y, Sun C, Steinmetz A, Qian Z (2011) 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax* ginseng. *Plant Cell Rep* 30(9): 1593-1601.

Chen X, Liao B, Song J, Pang X, Han J, Chen S (2013) A fast SNP identification and analysis of intraspecific variation in the medicinal *Panax* species based on DNA barcoding. *Gene* 530(1): 39-43.

Curci PL, De Paola D, Danzi D, Vendramin GG, Sonnante G (2015) Complete chloroplast genome of the multifunctional crop globe Artichoke and comparison with other Asteraceae. *PLOS One* 10(3): e0120589. DOI:10.1371/journal.pone.0120589.

Dong W, Liu H, Xu C, Zuo Y, Chen Z, Zhou S (2014) A chloroplast genome strategy for designing taxon specific

DNA mini-barcodes: a case study on ginsengs. *BMC Genetics* 15: 138.

Duong Tan Nhut, Nguyen Phuc Huy, Hoang Xuan Chien, Tran Cong Luan, Bui The Vinh, Lam Bich Thao (2012a) *In vitro* culture of petiole longitudinal thin cell layer explants of Vietnamese ginseng (*Panax vietnamensis* Ha et Grushv.) and preliminary analysis of saponin content. *Int J Appl Biol Pharm*: 178-190.

Dyal SD, Brown MT, Johnson PJ (2004) Ancient invasions: from endosymbiont to organelles. *Science* 304: 253-257.

Edwards A, Voss H, Rice P, Civitello A, Stegermann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W (1990) Automated DNA sequencing of the human HPRT locus. *Genomics* 6(4): 593-608.

Ferrarini M, Cestaro A, Sargent DJ, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, Viola R, Cavalieri D, Velasco R, Cestaro A, Sargent DJ (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics* 14: 670-670.

Fushimi H, Komatsu K, Isobe M, Namba T (1996) 18S ribosomal RNA gene sequences of three *Panax* species and the corresponding ginseng drugs. *Biol Pharm Bull* 19(11): 1530-1532.

Gupta PK, Xu Y (2008) Genomics of major crops and model plant species. *Int J Plant Genomics*. DOI:10.1155/2008/171928.

Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L (2005) Methods for obtaining and analyzing whole chroloplast genome sequences. *Methods Enzymol* 395: 348-384.

Ha WY, Shaw PC, Liu J, Yau FC, Wang J (2002) Authentication of *Panax ginseng* and *Panax quinquefolius* using amplified fragment length polymorphism (AFLP) and directed amplification of minisatellite region DNA (DAMD). *J Agric Food Chem* 50(7): 1871-1875.

Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol Ecol Resour* 9(2): 439-457.

Hollingsworth PM (2008) DNA barcoding plants in biodiversity hot spots: progress and outstanding questions. *Heredity (Edinb)* 101(1): 1-2.

Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res* 11: 247-261.

Komatsu K, Zhu S, Fushimi H, Qui T, Cai S, Kadota S (2001) Phytogenetic analysis based on 18S rRNA gene and *matK* gene sequences of *Panax vietnamensis* and five related species. *Planta Med* 67(5): 461-465.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proc Natl Acad Sci USA* 102(23): 8369-8374.

Ku C, Chung WC, Chen LL, Kuo CH (2013) The complete plastid genome sequence of Madagascar Periwinkle *Catharanthus roseus* (L.) G. Don: Plastid genome evolution, molecular marker identification, and phylogenetic implications in Asterids. *PLOS One* 8(6): e68518. DOI:10.1371/journal.pone.0068518.

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA* 105(8): 2923-2928.

Lee C, Wen J (2004) Phylogeny of *Panax* using chloroplast trnC–trnD intergenic region and the utility of trnC–trnD in interspecific studies of plants. *Mol Phylogenet Evol* 31: 894-903.

Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, Li Y, Wang L, Qian J, Chen S (2013) Transcriptome analysis reveals ginenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax* ginseng C. A, Meyer. *BMC Genomics* 14: 245.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol:* 1-11.

Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J, Niu Y, Cheng X, Xu H, Li C, Liu J, Steinmetz A, Chen S (2011) Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. *BMC Genomics* 12 Suppl 5: S5.

Mathiyalagan R, Subramaniyam S, Natarajan S, Kim YY, Sun MS, Kim SY, Kim Y-J, Yang DC (2013) *Insilico* profiling of microRNAs in Korean ginseng (*Panax ginseng* Meyer). *J Ginseng Res* 37(2): 227-247.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proc Natl Acad Sci USA* 74: 560-564.

Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11, 31-46.

Nhut DT, Vinh BVT, Hien TT, Huy NP, Nam NB, Chien HX (2012b) Effects of spermidine, proline and carbohydrate sources on somatic embryogenesis from main root transverse thin cell layers of Vietnamese ginseng (*Panax vietnamensis* Ha et Grushv.). *Afr J Biotechnol* 11(5): 1084-1091.

Olmstead RG, Palmer JD (1994) Chloroplast DNA systematic: a review of methods and data analysis. *Am J*

Bot 81: 1205-1224.

Pennisi E (2007) Taxonomy. Wanted: A barcode for plants. *Science* 318(5848): 190-191.

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. *Genomics* 93 (2): 105-111.

Poehlmann A, Kuester D, Meyer F, Lippert H, Roessner A, Schneider-Stock R (2007) Kras mutation detection in colorectal cancer using the Pyrosequencing technique. *Pathol Res Pract* 203: 489-497.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illuminaMiSeq sequencers. *BMC Genomics* 13 (1): 341.

Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. *Plant Syst Evol* 271: 101-122.

Rusk N (2011) Torrents of sequence. *Nat Methods* 8: 44.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94: 441-448.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74: 5463-5467.

Schadt EE, Turner S, Kasarskis A (2010) Window into third-generation sequencing. *Hum Mol Genet* 19 (R2): R227-40.

Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Methods* 5(1): 16-18.

Shaw PC, But PPH (2007) Authentication of *Panax* species and their adulterants by random-primed polymerase chain reaction. *Planta Med* 61: 466-469.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135-1145.

Shi C, Hu N, Huang H, Gao J, Zhao YL, Gao LZ (2012) An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLOS One* 7(2): e31468.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Cornell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071): 674-679.

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

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* 11: 262.

Wen J, Zimmer EA (1996) Phylogeny and biogeography of *Panax* L. (the ginseng genus, araliaceae): Inferences from ITS sequences of nuclear ribosomal DNA. *Mol Phylogenet Evol* 6(2): 167-177.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochrondial. Chloroplast, and nuclear DNAs. *Proc Natl Acad Sci USA* 84: 9054-9058.

Wu D, Austin RS, Zhou S, Brown D (2013) The root transcriptome for North American ginseng assembled and profiled across seasonal development. *BMC Genomics* 14: 564.

Zhang J, Yang W, Cui X, Yu H, Jin H, Chen Z, Shen T (2011) Breeding strains of *Panax notoginseng* by using EST-SSR markers. *Zhongguo Zhong Yao Za Zhi* 36(2): 97-101.

Zhao Y, Yin J, Guo H, Zhang Y, Xiao W, Sun C, Wu J, Qu X, Yu J, Wang X, Xiao J (2015) The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. *Front Plant Sci.* DOI: 10.3389/fpls.2014.00696.

Zhu S, Fushimi H, Cai S, Komatsu K (2003) Phylogenetic relationship in the genus *Panax*: Inferred from chloroplast trnK gene and nuclear 18S rRNA gene sequences. *Planta Med* 69(7): 647-653.

Zuo Y, Chen Z, Kondo K, Funamoto T, Wen J, Zhou S (2011) DNA barcoding of *Panax* species. *Planta Med* 77(2): 182-187.

Tạp chí Công nghệ Sinh học **14**(1): 1-13, 2016

GENOME SEQUENCING IN PLANTS AND THE GENUS *PANAX* L.

Le Thi Thu Hien[1,✉], Hugo de Boer[2], Vincent Manzanilla[2], Ha Van Huan[3], Nong Van Hai[1]

[1]*Institute of Genome Research, Vietnam Academy of Science and Technology*
[2]*Natural History Museum, University of Oslo, Norway*
[3]*College of Forestry Biotechnology, Vietnam National University of Forestry*

SUMMARY

Advances in genome sequencing technologies have created a new genomic era of life sciences research worldwide in which a number of modern and sophisticated techniques and tools have been developed and employed. Many countries have invested in plant genome sequencing as part of a sustainable development strategy. Each year, the number of plant genomes and transcriptomes sequenced has increased. The results obtained offer opportunities for fundamental and applied research, provide valuable data for identification of genes or molecular markers linked to traits that are important for selection, cultivation, and/or production. In Vietnam, partial or complete genome sequencing of crops has been recently conducted, primarily as part of international collaborative projects. The genus *Panax* L. (Araliaceae family) is comprised of several species of commercial value with narrow distributions such as *P. bipinnatifidus* Seem., *P. stipuleanatus* H.T.Tsai & K.M.Feng, and *Panax vietnamensis* Ha *et* Grushv. Despite their very important roles in traditional medicine, understanding of their genetic characteristics is still limited. Molecular studies on the genus have, so far, only evaluated limited markers for phylogenetic analysis. Therefore, genome sequencing of these important herbal plants is needed to understand their genetic characteristics, their evolutionary history and the genes and biochemical pathways contributing to medicinally important metabolites. This review summarizes all related genome sequencing technologies including the most recent advances in the last decade and their applications in genome and transcriptome sequencing of plants in general and in the genus *Panax* L. in particular.

Keywords: *next generation sequencing, chloroplast, genome, transcriptome, Panax* L.

✉ *Author for correspondence*: Tel: *+84-4-37918014*; E-mail: *hienlethu@igr.ac.vn*