ĐÀO QUANG HÀ

SINH HỌC THỰC NGHIỆM

2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Đào Quang Hà

ỨNG DỤNG TIN SINH HỌC TRONG PHÂN TÍCH HỆ PHIÊN MÃ CÂY SÂM NGỌC LINH (*Panax vietnamensis* Ha et Grushv.) 4 NĂM TUỔI

LUẬN VĂN THẠC SĨ
SINH HỌC THỰC NGHIỆM

Hà Nội - 2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Đào Quang Hà

ỨNG DỤNG TIN SINH HỌC TRONG PHÂN TÍCH HỆ PHIÊN MÃ CÂY SÂM NGỌC LINH (*Panax vietnamensis* Ha et Grushv.) 4 NĂM TUỔI

Chuyên ngành: Sinh học thực nghiệm
Mã số: 8420114

LUẬN VĂN THẠC SĨ
NGÀNH SINH HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC:
1. TS. Nguyễn Tường Vân
2. PGS. TS. Lê Thị Thu Hiền

Hà Nội - 2022

LỜI CAM ĐOAN

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

Tác giả luận văn

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến toàn bộ Quý Thầy Cô của Học viện Khoa học và Công nghệ vì đã truyền tải những kiến thức quý báu, những kinh nghiệm đầy thực tế cho em. Mặc dù đại dịch COVID-19 gây ra nhiều khó khăn cho quá trình học tập, sự tâm huyết và những truyền đạt tận tâm từ Quý Thầy Cô đã giúp em rất nhiều trong suốt hai năm học vừa qua.

Với tất cả sự trân trọng và quý mến, em xin gửi lời cảm ơn sâu sắc tới PGS. TS. Lê Thị Thu Hiền, Phó Viện trưởng Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Cảm ơn cô đã hỗ trợ em vào những thời điểm khó khăn nhất, ở bên em ngay cả những lúc em đã định không tiếp tục theo đuổi việc học. Em cảm thấy vô cùng may mắn khi có được sự giúp đỡ của cô để có thể hoàn thành công việc nghiên cứu.

Em cũng vô cùng biết ơn sự giúp đỡ của TS. Nguyễn Tường Vân, cán bộ nghiên cứu Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Để có thể hoàn thành luận văn này, sự hỗ trợ và giúp đỡ của cô Vân đặc biệt có ý nghĩa. Em thực sự cảm ơn cô đã động viên, giúp em tiếp tục hoàn thiện luận văn này.

Ngoài ra, em cũng xin được gửi lời cảm ơn tới các thành viên của Phòng Đa dạng sinh học hệ gen, Viện Nghiên cứu hệ gen, đặc biệt cảm ơn ThS. Lưu Hàn Ly, ThS. Phạm Lê Bích Hằng, CN. Vũ Thị Trinh và CN. Nguyễn Thị Bích Ngọc, vì đã hỗ trợ em rất nhiều trong quá trình nghiên cứu và học hỏi, để em có thể hoàn thiện luận văn này; cảm ơn ThS. Nguyễn Nhật Linh vì những hướng dẫn quý giá vào những ngày đầu em tiếp cận với hướng nghiên cứu tin sinh học; cảm ơn TS. Vũ Tuấn Nam vì những hỗ trợ trong quá trình nghiên cứu, cũng như những kinh nghiệm quý báu trong cuộc sống.

Nghiên cứu này được thực hiện với sự hỗ trợ từ đề tài "Giải trình tự và phân tích hệ phiên mã (transcriptome) ở sâm Ngọc Linh (*Panax vietnamensis* Ha et Grushv.).

Sự giúp đỡ, động viên và chia sẻ của gia đình, các thầy cô và bạn bè hỗ trợ em rất nhiều trong quá trình hoàn thành luận văn này. Em vô cùng trân trọng và biết ơn tất cả mọi người!

Hà Nội, ngày 20 tháng 12 năm 2022
Tác giả luận văn

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Chữ viết tắt	Tên đầy đủ (tiếng Anh)	Tên đầy đủ (tiếng Việt)
BLAST	Basic local alignment search tool	Công cụ tìm kiếm và so sánh các trình tự tương đồng
Bp	Base pair	Cặp base
DNA	Deoxyribonucleic acid	Acid deoxyribonucleic
Đtg	*et al*.	Đồng tác giả
EST	Expressed sequence tag	Các đoạn trình tự gen biểu hiện
FastQC	Fast quality check	Công cụ kiểm tra đánh giá chất lượng dữ liệu
IUCN	International Union for Conservation of Nature and Natural Resources	Liên minh Quốc tế Bảo tồn thiên nhiên và Tài nguyên thiên nhiên
Kb	Kilo base	Kilo base (1.000 bp)
NCBI	National Center for Biotechnology Information	Trung tâm Tin sinh học Quốc gia
RNA-Seq	Whole transcriptome shotgun sequencing	Giải trình tự toàn bộ hệ phiên mã
ORF	Open reading frame	Khung đọc mở
SSR	Simple sequence repeat	Trình tự lặp lại đơn giản

DANH MỤC HÌNH

DANH MỤC BẢNG

MỞ ĐẦU

Với nền y học cổ truyền trên thế giới nói chung và ở Việt Nam nói riêng, từ lâu sâm đã là nguồn dược liệu quý giá. Ngày nay, tên gọi sâm được dùng phổ biến để chỉ một số loài thực vật, trong đó chủ yếu là các loài thuộc chi *Panax* L. và *Eleutherococcus* Maxim thuộc họ Araliaceae. Giá trị kinh tế của sâm về mặt thương mại được ước tính vượt 2,1 tỷ đô la Mỹ trên quy mô toàn cầu [1]. Các loài sâm được sử dụng phổ biến nhất trong chi này là sâm Hàn Quốc (*Panax ginseng*), sâm Mỹ (*Panax quinquefolius*), sâm Trung Quốc (*Panax notoginseng*), sâm Nhật Bản (*Panax japonicus*), sâm Himalaya (*Panax pseudoginseng*) và sâm Việt Nam (*Panax vietnamensis* Ha et Grushv.) [2]. Nhiều loài sâm đang bị đe dọa trong tự nhiên, do đó các sáng kiến phục vụ phát triển hiệu quả và bền vững là đặc biệt cần thiết để bảo tồn đa dạng sinh học ở cả cấp độ loài và cấp độ hệ gen, đồng thời đáp ứng nhu cầu về sản phẩm dược liệu. *P. vietnamensis*, hay còn được biết đến với tên gọi sâm Ngọc Linh, là loài dược liệu đặc hữu ở nước ta. Sâm Ngọc Linh có giá trị cao về khoa học và kinh tế, tuy nhiên các nghiên cứu còn hạn chế. Cũng giống như các loài thuộc chi *Panax*, sâm Ngọc Linh mang các hợp chất triterpene saponin tự nhiên, hay còn được biết đến với tên gọi là ginsenoside. Những hợp chất này có hoạt tính tốt, đồng thời mang tác dụng chống suy nhược cơ thể, oxy hóa, đái tháo đường, ung thư, tăng cường thể lực, bảo vệ hệ thần kinh. Do việc khai thác quá mức cũng như phân bố tương đối hẹp, sâm Ngọc Linh trở nên khó phát triển trong môi trường tự nhiên và đã được thêm vào Danh lục đỏ của Liên minh Quốc tế Bảo tồn thiên nhiên và Tài nguyên thiên nhiên (IUCN) và Sách đỏ Việt Nam (2007) [3]. Do đó, những nghiên cứu di truyền cũng như đặc tính hóa học và dược tính của sâm Ngọc Linh đã và đang được thực hiện, góp phần phát triển loài thực vật quý hiếm này ở nước ta.

Với mục đích bổ sung thêm những thông tin hữu ích về hệ phiên mã sâm Ngọc Linh, qua đó hỗ trợ đánh giá nguồn gen, giúp bảo tồn, khai thác và sử

dụng bền vững sâm Ngọc Linh ở Việt Nam, nghiên cứu "**Ứng dụng tin sinh học trong phân tích hệ phiên mã cây sâm Ngọc Linh (*Panax vietnamensis* Ha et Grushv.) 4 năm tuổi**" đã được thực hiện với mục tiêu và những nội dung cụ thể như sau:

1 MỤC TIÊU

Nghiên cứu đặt mục tiêu phân tích được kết quả giải trình tự và xây dựng được cơ sở dữ liệu của hệ phiên mã đặc hiệu mô (lá, thân rễ) ở sâm Ngọc Linh 4 năm tuổi.

2 NỘI DUNG

- Giải trình tự hệ phiên mã đặc hiệu mô thân rễ và lá của sâm Ngọc Linh 4 năm tuổi;
- Phân tích, lắp ráp, hiệu chỉnh trình tự các hệ phiên mã mô lá và thân rễ;
- Chú giải hệ phiên mã mô thân rễ và lá.

CHƯƠNG 1. TỔNG QUAN TÀI LIỆU

1.1 GIỚI THIỆU VỀ CHI *PANAX* VÀ SÂM NGỌC LINH

Chi *Panax* L., hay chi Sâm thuộc họ Araliaceae (Ngũ gia bì/ Nhân sâm), bộ Apiales (Hoa tán), là một trong những chi bao gồm nhiều loài dược liệu quý. *Panax* phân bố tự nhiên chủ yếu ở một số quốc gia thuộc khu vực Bắc Mỹ, châu Á, đặc biệt là vùng Đông Á. Nhiều loài *Panax* chứa các ginsenoside, hay còn được biết tới là các hợp chất tự nhiên triterpene saponin. Các hợp chất này có cấu tạo phân tử phức tạp; có lợi cho sức khỏe con người: bảo vệ hệ thần kinh, chống đái tháo đường, oxy hóa, ung thư và tăng cường thể lực [3, 4]. Do vùng phân bố hạn chế và giá trị kinh tế cùng nhu cầu rất cao nên nhiều loài thuộc chi *Panax* đã bị khai thác quá nhiều và khan hiếm trong môi trường tự nhiên. Nhiều quốc gia cũng đã triển khai các chương trình khai thác, bảo tồn cũng như sử dụng nguồn gen bền vững đặc biệt quý này, chẳng hạn như sâm Trung Quốc, Hàn Quốc và Mỹ được trồng thương mại và phổ biến nhất [5]. Gần đây, 12 loài và 2 thứ thuộc chi này đã được mô tả (Bảng 1.1) [5].

Bảng 1.1 Danh sách các loài trong chi *Panax* trên thế giới

Tên khoa học	Khu vực địa lý
P. bipinnatifidus Seem.	
P. bipinnatifidus var. angustifolius	
P. bipinnatifidus var. bipinnatifidus	
P. ginseng C.A.Mey.	Hàn Quốc (Nhân sâm)
P. japonicus (T.Nees) C.A.Mey.	Nhật Bản
P. notoginseng (Burkill) F.H.Chen	Trung Quốc
P. pseudoginseng Wall.	Himalaya

Tên khoa học	Khu vực địa lý
P. quinquefolius L.	Mỹ
P. sokpayensis Shiva K.Sharma et Pandit	
P. stipuleanatus H.T.Tsai et K.M.Feng	
P. trifolium L.	
P. vietnamensis Ha et Grushv.	Việt Nam
P. wangianus S.C.Sun	
P. zingiberensis C.Y.Wu et Feng	

Ở Việt Nam, các nghiên cứu về xác định thành phần và phân loại loài cho thấy, chi *Panax* gồm 3 loài mọc trong tự nhiên và có thể sử dụng trong dược liệu, bao gồm *P. vietnamensis* Ha et Grushv. (sâm Ngọc Linh), *P. bipinnatifidus* Seem. (sâm Vũ diệp) và *P. stipuleanatus* Tsai et Feng (Tam thất hoang) [6, 7].

Dù được tìm thấy lần đầu tiên vào năm 1973, phải tới năm 1985, Sâm Ngọc Linh mới được chính thức công nhận là loài thực vật mới [7]. Cây sâm này có các tên gọi khác nhau như sâm khu Năm (K5), sâm Việt Nam, trúc tiết sâm (hay sâm đốt trúc), củ Ngải rọm hay cây Thuốc giấu. Sâm Ngọc Linh mọc dày thành đám dưới tán rừng dọc theo các suối ẩm trên đất nhiều mùn, tập trung ở khu vực Ngọc Linh (thuộc tỉnh Kon Tum và Quảng Nam). Đây là là dãy núi cao thứ hai ở nước ta (đỉnh cao nhất cao 2.598 m). Hai thứ khác của *P. vietnamensis* được phát hiện gần đây là *P. vietnamensis* var. *fuscidiscus* K.Komatsu, S.Zhu et S.Q.Cai (sâm Lai Châu) phân bố ở Lai Châu, *P. vietnamensis* Ha et Grushvitzky var. *langbianensis* N.V.Duy, V.T.Tran & L.N.Trieu (sâm Lang Bian) được phát hiện ở vùng núi Langbian, Lâm Đồng. Ngoài ra, *Panax* sp. (*Panax* sp. Puxailaileng) được phát hiện phân bố ở vùng núi Puxailaileng, tỉnh Nghệ An và được công nhận là một thứ của Ngọc Linh, tuy nhiên cần các có nghiên cứu sâu

hơn [9]. Tương tự với nhiều loài thuộc chi *Panax*, sâm Ngọc Linh rất hiếm khi được tìm thấy trong môi trường tự nhiên do khai thác quá mức cũng như phân bố hạn chế, có mặt trong Sách đỏ Việt Nam [6].

Bảng 1. mô tả một số đặc điểm hình thái của các thứ/loài thuộc chi *Panax* ở nước ta (bao gồm sâm Ngọc Linh, sâm Vũ diệp, sâm Nghệ An, sâm Lai Châu, Tam thất hoang) [6, 8, 9].

Bảng 1.2 Một số đặc điểm hình thái của các thứ/loài thuộc chi *Panax* ở nước ta sử dụng trong định loại hình thái các mẫu thu thập

Đặc điểm	Sâm Ngọc Linh	Sâm Lai Châu	Sâm Nghệ An	Sâm Vũ diệp	Tam thất hoang
Chiều cao cây (cm)	30-110	40-80		30-100	25-75
Rễ	Có các đốt và có thể phân nhánh, nằm ngang (đường kính khoảng 1 tới 2 cm)	Thân rễ mập, nạc, hợp trục, nằm hơi chếch hoặc ngang, không rễ phụ (đường kính: khoảng 1,5-2,2 cm)	Thân rễ mọc bò ngang, mang nhiều rễ nhánh và củ, (đường kính: 1-2 cm)	Thân rễ lớn, phương ngang, thường ở trên mặt đất, (đường kính khoảng 1,5 tới 3,5 cm)	Rễ mập, nằm ngang, thường ít phân nhánh, một số chỗ lõm do vết thân, (đường kính khoảng 1,5 tới 3 cm)
Thân (mang lá)	Phụ thuốc số lượng nhánh ở rễ, phần mang lá có thể có từ 1 tới 5 thân	Thân đơn độc, nhẵn, mọc thẳng đứng, vỏ màu lục, cao 0,3-0,7 m, phần giữa xốp khi tươi và rỗng khi khô	Thân kí sinh, thẳng đứng, nhỏ, có đường kính thân 4-8 mm, màu xanh lục hoặc hơi tím,	Phụ thuộc số lượng nhánh của rễ, phần mang lá có từ 1 tới 3 thân, đường kính thân: 0,3-0,6 cm	Thường có 1 thân, ít khi 2 hoặc 4 thân. Mọc thẳng, nhẵn, đường kính từ 0,3 tới 0,6 cm

Đặc điểm	Sâm Ngọc Linh	Sâm Lai Châu	Sâm Nghệ An	Sâm Vũ diệp	Tam thất hoang
Lá	Dạng kép hình chân vịt, mọc vòng (3 tới 5 lá), dạng bầu dục, nhọn và thuôn 2 đầu, kích thước khoảng 6 tới 14 x 2,3 tới 4 cm, mép có thêm răng cưa	Lá mọc vòng 3 tới 6 lá ở đỉnh thân, dạng kép hình chân vịt, mang 5 tới 7 lá chét mỏng, dạng màng. Cuống lá dài khoảng 7 tới 14 cm.	Lá kép chân vịt mọc vòng với 3 tới 5 nhánh lá. Cuống lá kép: 6-12 mm, 5 lá, lá ở chính giữa lớn nhất (12-15 cm). Lá hình bầu dục, mép răng cưa, chóp nhọn, lông gai ở hai mặt lá	Lá kép chân vịt, thường gồm 3 cái mọc vòng ở ngọn. Lá chét (từ 3 tới 5), có răng cưa	Lá kép hình chân vịt (khoảng 1 tới 3 lá), mọc vòng ở ngọn, cuống khoảng 5 tới 10 cm. Lá cuống khá ngắn, phiến hình hoặc mác thuôn, nhọn 2 đầu, kích thước 5x2 -13x4 cm, mép và cửa thùy có rang cưa, lông ở gân lá phía trên
Cụm hoa	Tán đơn hoặc kép (có thể có 1 - 2 tán phụ), mọc ở ngọn, chiều dài của cuống 15-30 cm, cao vượt tán do dài hơn cuống lá	Mọc đơn ở giữa vòng và đỉnh thân. Cuống dài khoảng 25 cm. Cụm hoa là tán, gần hình cầu, đường kính đến từ 2,5 tới 4 cm, có 70 tới 100 hoa (hoặc hơn)	Tán đơn, dưới các lá thẳng với thân, cuống tán hoa 10 tới 20 cm (có thể kèm 1 tới 4 tán phụ hay hoặc 1 hoa riêng lẻ). Có 60-100 hoa trong một tán.	Tán mọc ngọn, đơn, cuống 5-510 cm, cụm hoa có 20 tới 90 hoa, cuống mảnh, dài khoảng 1 tới 1,5 cm	Tán đơn, mọc ngọn (ít khi có tán phụ nhỏ). Cuống cụm cao tương đương hoặc hơn tán lá, dài 5 tới 10 cm. Hoa trên một tán: 30-80
Hoa	Cuống ngắn, trắng xanh, với 5 đài hoa, 5 nhị và 5	Hoa vàng nhạt hoặc lục, rộng 3-4 mm. Bầu ở dưới, gồm 2 lá noãn,	Cuống hoa ngắn 1 tới 1,5 cm, lá đài 5, cánh hoa 5, màu vàng nhạt,	Màu vàng xanh, gồm 5 nhị, 5 cánh, 5 lá đài nhỏ. Đầu vòi	Hoavàng xanh, 5 lá đài khá nhỏ, có 5 cánh, 5 nhị hoa, bầu có 2 ô, đầu nhụy

Đặc điểm	Sâm Ngọc Linh	Sâm Lai Châu	Sâm Nghệ An	Sâm Vũ diệp	Tam thất hoang
	cánh. Bầu hai ô, vòi nhụy chẻ chia đôi ở đầu	hợp với vòi nhụy hình thành bầu, chẻ đôi	nhị 5, bầu 1 ô với 1 vòi nhụy	nhụy chẻ chia đôi. Bầu có 2 ô.	chẻ làm đôi.
Quả	Quả hình cầu, mọng (đường kính khoảng 0,5 tới 0,6 cm). Chín đỏ và thường xuất hiện chấm đen ở đỉnh.		Tập trung ở trung tâm của tán, dài 0,8-1cm, rộng 0,5-0,6 cm.	Quả hình cầu hoặc cầu dẹt, đường kính 0,6-1,2 cm, có màu đỏ khi chín	Hình cầu dẹt, mọng, đường kính khoảng 0,6 tới 1,2 cm và thời điểm chín có màu đỏ
Hạt	1 - 2 hạt, hạt nhỏ, hơi tròn hoặc gần giống hình thận. Không nhẵn			Hạt 2, hơi tròn và có màu xám trắng, vỏ khá cứng và có rốn hạt	1 - 2 hạt, khá giống với hạt đậu, màu xám trắng, vỏ cứng và có rốn hạt
Mùa hoa, quả	Mùa hoa vào tháng 4-5, quả 6-9			Hoa tháng 4-5 và quả tháng 5-9 (10)	Hoa tháng 4-5. Quả tháng 5-9 (10)

Với giá trị khoa học và kinh tế rất cao, trong thời gian qua, Kon Tum và Quảng Nam đã tích cực phát triển và bảo tồn sâm Ngọc Linh. Việc trồng sâm Ngọc Linh trên vùng bán sinh trưởng hoặc sinh trưởng tự nhiên cho kết quả cây sâm phát triển và sinh trưởng tương đối tốt. Từng hộ gia đình được giao sâm giống để quản lý và chăm sóc. Nếu được tập trung phát triển, cách làm này sẽ

mang lại hiệu quả thiết thực và góp phần xóa đói giảm nghèo cho các hộ đồng bào dân tộc thiểu số [10].

1.2 GIẢI TRÌNH TỰ DNA THẾ HỆ MỚI TRÊN THẾ GIỚI

Sự thành công trong nghiên cứu giải trình tự hệ gen của con người đã mang đến một thời kỳ phát triển mạnh mẽ đối với nghiên cứu khoa học. Các công nghệ giải trình tự mới, phức tạp và hiệu quả hơn lần lượt ra đời, ứng dụng cho nghiên cứu và phát triển khoa học. Trong đó, một công nghệ có sức ảnh hưởng mạnh mẽ ở quy mô toàn cầu chính là giải trình tự gen thế hệ mới (NGS - next generation sequencing). Công nghệ này cho phép giải trình tự nhanh chóng và chính xác, được sử dụng cho nhiều ứng dụng như giải trình tự toàn bộ hệ gen sinh vật (whole genome sequencing), hệ phiên mã transcriptome (RNA-seq) hay hệ gen biểu hiện (whole exome sequencing). Tiếp nối từ Dự án giải trình tự hệ gen người, vào năm 2005, 454 sequencing đã ra mắt thị trường. Năm 2006, Genome Analyzer được cho ra đời bởi Solexa và SOLiD được phát triển bởi Agencourt. Các hệ thống này có các ưu điểm là độ chính xác rất cao, giá thành hạ thấp, đồng thời có dung lượng rất lớn so với trước đó. Các tổ chức đó sau này được mua lại bởi Applied Biosystems/ Agencourt vào năm 2006), Roche/ 454 và Illumina/ Solexa vào năm 2007. Đến nay, nhiều hệ thống máy NGS đã được phát triển, có thể kể tới Roche/ 454; Illumina NextSeq, MiSeq, NovaSeq, HiSeq,; Pacific Biosciences/ RS; Applied Biosystem/ SOLiD; Life technologies/ Ion Torrent PGM, Life technologies/ Ion Proton… nhờ đó giúp giải trình tự cho toàn bộ hệ gen một cách chính xác và nhanh chóng [11, 12, 13, 14, 15].

Các hệ thống giải trình tự thế hệ mới trên hoạt động dựa vào nguyên tắc gắn nối hoặc tổng hợp. Công nghệ NGS được cải tiến trên cơ sở các bước như sau: (i) chuẩn bị mẫu đầu vào, (ii) giải trình tự, (iii) lắp ráp hệ gen, (iv) chú giải và so sánh. Công đoạn giải trình tự bao gồm các bước như sau: (i) chuẩn bị và gắn lên giá bám các đoạn DNA; (ii) khuếch đại các DNA trên nhờ mồi đặc hiệu; (iii) giải trình tự bằng gắn nối hoặc tổng hợp [16, 17]. Với ưu thế đọc nhanh

chóng, trình tự đọc được xử lý với dung lượng rất lớn và hiệu quả cao, các hệ thống NGS ngày càng được ứng dụng rộng rãi [18, 19, 20].

1.3 NGHIÊN CỨU GIẢI TRÌNH TỰ HỆ PHIÊN MÃ Ở CÁC LOÀI THUỘC CHI *PANAX*

Bên cạnh sự phát triển mạnh mẽ của NGS, việc nghiên cứu giải trình tự hệ phiên mã cho các loài ở chi *Panax* đã và đang được tiến hành trong thời gian gần đây. Thống kê cho thấy, các nghiên cứu hiện tại chủ yếu tập trung ở một số khu vực, quốc gia có phân bố cây sâm và thường xuyên được sử dụng làm dược phẩm, chẳng hạn như Hàn Quốc hay Trung Quốc.

Trong năm 2010, tại Viện Phát triển thảo dược Bắc Kinh, Sun và đtg đã công bố nghiên cứu đầu tiên về giải trình tự hệ phiên mã của *P. quinquefolius*. Nhóm tác giả sử dụng hệ thống GS FLX Titanium, sử dụng công nghệ 454 pyrosequencing, qua đó thu được khoảng 200.000 đoạn đọc chất lượng tốt, có độ dài mỗi đoạn đọc trung bình của thư viện cDNA rễ lên tới 427 bp. Kết quả, trên 30.000 unigene đã được xác định và khoảng 2/3 số trình tự này có thể tìm thấy từ các ngân hàng trình tự [21].

Năm 2011, Luo và đtg đã công bố hệ phiên mã của *P. notoginseng* thuộc chi *Panax* [22]. Các nhân tố điều hòa dịch mã (WRKY, Myb, bHLH, homeobox) và các trình tự SSR ở *P. notoginseng* cũng được nghiên cứu từ dữ liệu hệ phiên mã.

Dựa trên số liệu từ 2 lần chạy hệ thống 454 pyrosequencing, vào năm 2013, Li và đtg đã thu được 45.849, 6.172, 4.041 và 3.273 trình tự mã hóa từ thư viện cDNA lần lượt của rễ, thân, lá và hoa của *P. ginseng* [23]. Nghiên cứu của Cao và đtg (2015) đã phát hiện được 2 unigene mã hóa geranyl diphosphate synthase (GPS) ở rễ sâm *P. ginseng* 4 năm tuổi [24]. Dựa trên các trình tự hệ gen tham chiếu của *P. ginseng* giống "Chunpoong" đã được công bố, các bộ dữ liệu hệ phiên mã và chú giải chức năng, Jayakodi và đtg đã xây dựng cơ sở dữ liệu về hệ gen của *P. ginseng* tại http://ginsengdb.snu.ac.kr/ [25]. Đây là nền

tảng truy cập mở đầu tiên cung cấp dữ liệu toàn diện của *P. ginseng*, là nguồn tài nguyên quý giá cho nhiều lĩnh vực nghiên cứu liên quan đến *P. ginseng* và các loài khác thuộc bộ Apiales cũng như cho cộng đồng các nhà khoa học nghiên cứu thực vật nói chung.

Trên đối tượng sâm Nhật Bản *P. japonicus*, Rai và đtg (2016) đã tiến hành phân tích trình tự hệ phiên mã, đồng thời so sánh các gen tiềm năng liên quan đến sinh tổng hợp các saponin với các loài sâm khác [26]. Kết quả giải trình tự RNA bằng hệ thống Illumina thu được 135.235 unigene. Trên đối tượng *P. vietnamensis,* Vu và đtg (2020) đã công bố kết quả giải trình tự hệ phiên mã của *P. vietnamensis*, chú giải các unigene sử dụng 7 cơ sở dữ liệu bao gồm Clusters of Orthologous Groups (COG), GO, KEGG, Clusters of Orthologous Groups (KOG), Pfam, Swissprot và NR. Mẫu phân tích là mẫu mô hỗn hợp của các mẫu sâm không xác định tuổi và tỉ lệ chú giải đạt 35,49% với 31.686 unigene được phát hiện [27].

1.4 VAI TRÒ CỦA CÁC CÔNG CỤ TIN SINH TRONG PHÂN TÍCH TRÌNH TỰ HỆ PHIÊN MÃ

Các công nghệ giải trình tự thế hệ mới cho phép xác định dữ liệu lớn về gen/hệ gen trong thời gian ngắn. Kể từ khi các kỹ thuật giải trình tự hiện đại ra đời, việc xác định trình tự nucleic acid đã trở thành một công cụ phổ biến và thiết yếu trong mọi lĩnh vực khoa học sinh học. Tương ứng, lĩnh vực tin sinh học là trung tâm của việc giải thích và ứng dụng dữ liệu sinh học này. Sử dụng các phương pháp toán học và thống kê được triển khai bởi nhiều ngôn ngữ lập trình, các công cụ tin sinh học tổ chức, phân tích và giải thích thông tin sinh học ở cấp độ phân tử, tế bào và gen. Tin sinh học, sử dụng các công cụ tính toán, toán học và thống kê để thu thập, sắp xếp và phân tích dữ liệu trình tự gen lớn và phức tạp cũng như dữ liệu sinh học liên quan. Sức mạnh tổng hợp của NGS và tin sinh học rất quan trọng đối với phân tích, xử lý và đánh giá kết quả giải trình tự

hệ phiên mã, trong đó bao gồm ba bước chính: (1) kiểm soát chất lượng và chuẩn bị dữ liệu, (2) lắp ráp bộ gen; và (3) phân tích sau lắp ráp.

Các công nghệ giải trình tự hiện đại có thể tạo ra một số lượng lớn các lần đọc trình tự trong một thí nghiệm. Tuy nhiên, chưa có công nghệ giải trình tự nào là hoàn hảo và mỗi công cụ sẽ tạo ra các loại và số lượng lỗi khác nhau, ví dụ các nucleotide được xác định không chính xác. Tỉ lệ base được xác định sai này sẽ phụ thuộc vào giới hạn kỹ thuật của từng nền tảng giải trình tự. Việc xác định và loại trừ các loại lỗi có thể ảnh hưởng đến việc phân tích tiếp theo là rất cần thiết. Do đó, kiểm soát chất lượng trình tự là bước đầu tiên thiết yếu trong phân tích, giúp tiết kiệm thời gian và chất lượng của các bước sau này. Việc kiểm soát chất lượng cơ sở dữ liệu thô sau giải trình tự thường sử dụng công cụ FastQC [28] để đánh giá các tiêu chí như: chất lượng trình tự base, tỉ lệ adapter, điểm chất lượng trình tự, tỉ lệ base, tỉ lệ G-C, phân bố độ dài trình tự, tỉ lệ lặp lại của các trình tự, các trình tự biểu hiện quá mức và công cụ Trimmomatic [29] để loại bỏ các trình tự ngắn, ví dụ, dưới 36 base hoặc có điểm chất lượng thấp.

Sau khi có được bộ dữ liệu đã xử lý và làm sạch, các bước lắp ráp, phân đoạn trình tự thành unigene, dự đoán khung đọc mở và ước lượng độ phong phú đều yêu cầu bắt buộc các công cụ tin sinh tương ứng để giải quyết. Các công cụ hiện nay cho phép việc lắp ráp hệ gen *de novo,* giúp tạo các trình tự tham chiếu, ngay cả đối với hệ gen phức tạp hoặc đa bội, cung cấp thông tin hữu ích để lập bản đồ hệ gen của các sinh vật mới hoặc hoàn thiện hệ gen của các sinh vật đã biết, làm rõ các vùng rất giống nhau hoặc lặp đi lặp lại để lắp ráp *de novo* chính xác, cũng như xác định các biến thể cấu trúc và sắp xếp lại, chẳng hạn như xóa, đảo ngược hoặc chuyển vị.

Các công cụ tin sinh cũng đóng vai trò quan trọng trong quá trình chú giải các đoạn đọc đã lắp ráp. Chú giải hệ gen là quá trình lấy thông tin cấu trúc và chức năng của protein hoặc gen từ dữ liệu thô bằng cách sử dụng các phương pháp tin sinh để phân tích, so sánh, ước tính. Sau khi hệ gen được giải trình tự

và lắp ráp, các đoạn đọc cần được chú giải để có được những thông tin hợp lý hơn về các đặc điểm cấu trúc và vai trò chức năng của nó. Sử dụng các phương pháp chú giải gen, có thể dự đoán các gen hoặc protein thuộc hệ gen cụ thể. Chú giải chức năng của các gen hoặc protein mới này có thể được thực hiện bằng cách tìm kiếm sự giống nhau của chúng với các trình tự đã được xác định bằng thực nghiệm và có sẵn trong cơ sở dữ liệu.

Thông tin trình tự hệ gen được lưu trữ trong một số định dạng tệp như FASTA, GFF3 và GENBANK. Có nhiều định dạng tệp khác nhau để trình bày trình tự, cấu trúc và thông tin con đường liên quan đến gen và protein, đồng thời cơ sở để chọn và tải xuống một tệp cụ thể có sẵn trên cơ sở dữ liệu trực tuyến. Các cơ sở dữ liệu để thực hiện chú giải cũng đa dạng và phụ thuộc vào mục đích nghiên cứu. Một số cơ sở dữ liệu phổ biến bao gồm GenBank, EMBL; GO, KEGG, UniProtKB/Swiss-Prot, Pfam. Tương ứng với mỗi cơ sở dữ liệu, các công cụ tin sinh cũng được tối ưu thuật toán để khai thác dữ liệu trình tự gen/protein, đưa ra các thông tin chi tiết như kích thước, độ bao phủ, tỉ lệ tương đồng, số lượng gap, độ tin cậy, vị trí khung đọc mở... cũng như tham khảo các tài liệu nghiên cứu trước đó để làm tiền đề cho việc so sánh, đối chiếu.

CHƯƠNG 2. VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1 VẬT LIỆU

2.1.1 Các mẫu sâm nghiên cứu

Nghiên cứu sử dụng 6 mẫu sâm Ngọc Linh 4 năm tuổi (bao gồm 3 mẫu thân rễ C4.1, C4.2, C4.3 và 3 mẫu lá L4.1, L4.2, L4.3) thu thập tại Trại Sâm gốc Tắk Ngo, thuộc Trung tâm Sâm Ngọc Linh (Nam Trà My, Quảng Nam). Một số mẫu sâm Ngọc Linh 4 năm tuổi phục vụ phân tích hệ phiên mã được trình bày trên Hình 2.1.

Mẫu C4.1 Mẫu L4.1

Hình 2.1 Hình ảnh một số mẫu sâm Ngọc Linh thu thập

2.1.2 Hóa chất, thiết bị

Các hóa chất sử dụng được đảm bảo độ tinh khiết và chất lượng cho các thí nghiệm sinh học phân tử và có nguồn gốc từ các hãng như Thermo Fisher Scientific (Hoa Kỳ), Illumina (Hoa Kỳ), Qiagen (Hoa Kỳ)… Các máy móc, thiết bị chính được sử dụng thuộc Viện Hàn lâm Khoa học và Công nghệ Việt Nam và Bảo tàng Lịch sử Tự nhiên, Đại học Oslo, Na Uy. Ngoài ra, nghiên cứu sử

dụng các chương trình phần mềm chuyên dụng miễn phí hoặc được cấp phép tại Viện Nghiên cứu hệ gen thuộc Viện Hàn lâm Khoa học và Công nghệ Việt Nam và Bảo tàng Lịch sử Tự nhiên, Đại học Oslo, Na Uy.

2.2 PHƯƠNG PHÁP NGHIÊN CỨU

2.2.1 Thu mẫu

2.2.1.1 Phương pháp thu mẫu và bảo quản mẫu nhằm giải trình tự hệ phiên mã

Mẫu mô thân rễ và mô lá của cây sâm Ngọc Linh 4 năm tuổi được thu và bảo quản riêng biệt. Các cây được rửa sạch bằng nước cất đã khử trùng, thấm khô, cắt nhỏ và bảo quản trong dung dịch RNAlater tự pha. Mẫu mô trong RNAlater được giữ ở 4^{o}C trong 24 giờ trước khi bảo quản thời gian dài trong tủ lạnh sâu -80^{o}C.

2.2.1.2 Phương pháp tách chiết RNA tổng số

RNA tổng số của các mẫu lá, thân rễ sâm Ngọc Linh được tách chiết và tinh sạch sử dụng RNeasy Plant Mini Kit (Qiagen, Hoa Kỳ) theo hướng dẫn của nhà sản xuất.

RNA tổng số của các mẫu lá và thân rễ sâm Ngọc Linh được kiểm tra, đánh giá chất lượng bằng điện di trên gel agarose 0,8%. Để xác định nồng độ và độ sạch, các mẫu RNA được phân tích bằng máy quang phổ Biospetrometer (Eppendorf, Đức). RNA sau tách chiết và tinh sạch được bảo quản ở -80^{o}C cho đến khi được sử dụng cho các thí nghiệm tiếp theo.

2.2.1.3 Phương pháp tổng hợp cDNA và chuẩn bị thư viện cDNA

Các đoạn mRNA sẽ được phiên mã ngược tạo thành các đoạn cDNA theo hướng dẫn của nhà sản xuất. Thư viện được chuẩn bị sử dụng TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, Hoa Kỳ).

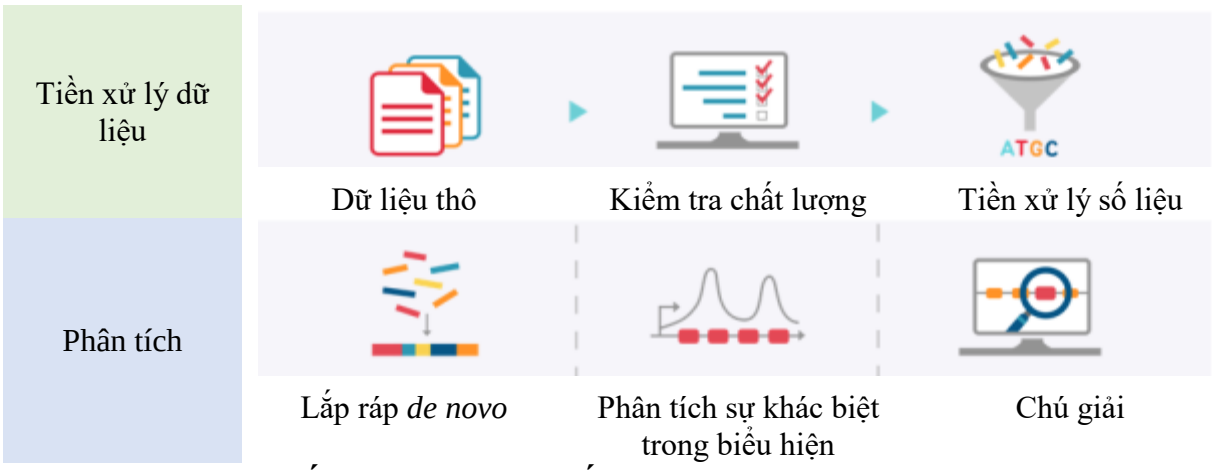
2.2.1.4 Giải trình tự hệ phiên mã

Các thư viện cDNA đạt yêu cầu chất lượng được sử dụng để tiến hành

giải trình tự hệ phiên mã bằng hệ thống NGS của Illumina.

2.2.2 Phân tích và lắp ráp trình tự các hệ phiên mã

Dữ liệu thô sau khi giải trình tự được kiểm tra chất lượng và tiền xử lý dữ liệu, sau đó trải qua các bước lắp ráp *de novo*, phân tích sự khác biệt trong biểu hiện và chú giải transcriptome (Hình 2.).



Hình 2.2 Một số bước phân tích, lắp ráp và chú giải trình tự hệ phiên mã

2.2.2.1 Kiểm tra chất lượng dữ liệu thô sau khi giải trình tự và tiền xử lý số liệu

Chất lượng các đoạn đọc thô sau khi giải trình tự được đánh giá lại, kiểm tra chất lượng bằng phần mềm FastQC v0.11.9. Phần mềm Trimmomatic v0.36 được sử dụng để tiền xử lý số liệu nhằm loại bỏ các trình tự là các adapter và các base chất lượng kém.

2.2.2.2 Lắp ráp *de novo* các đoạn đọc

Các đoạn đọc sau bước chọn lọc chất lượng sẽ được lắp ráp sử dụng phần mềm Trinity trinityrnaseq_r20140717 [30]. Các thông số được sử dụng trong lắp ráp *de novo* các đoạn đọc của mẫu phân tích đều được sử dụng theo mặc định của phần mềm. Trong quá trình lắp ráp, các đoạn đọc được chuẩn hóa *in silico* với độ bao phủ lớn nhất là 50.

2.2.2.3 Phân nhóm các đoạn trình tự thành các unigene (Clustering)

Đối với các gen đã được lắp ráp, contig dài nhất được lọc và chia thành các trình tự phiên mã sử dụng phần mềm CD-HIT-EST v4.6 [31, 32]. Các thông số được sử dụng trong phân tích đều là các thông số mặc định của phần mềm với ngưỡng giá trị phân nhóm là 0,9 tương đương độ tương đồng 90%, kích thước word là 8 base, sử dụng fast mode, loại bỏ các unigene với kích thước nhỏ hơn 200 bp... Các trình tự phiên mã sau lắp ráp được coi là các unigene.

2.2.2.4 Dự đoán khung đọc mở (ORF)

Các unigene đã thu được từ bước trên được sử dụng làm nguyên liệu đầu vào cho quá trình dự đoán khung đọc mở (Open Reading Frame - ORF). Phần mềm TransDecoder v3.0.1 [33] được sử dụng để tìm kiếm và phát hiện những vùng mã hóa trong trình tự hệ phiên mã. ORF được phát hiện bằng cách tìm kiếm các codon mở đầu và kết thúc. Các ORF được lựa chọn với kích thước trên 100 amino acid. Trong số đó, 500 ORF có kích thước lớn nhất được lựa chọn và một mô hình Markov bậc 5 đặc hiệu khung đọc được sử dụng. Các ORF sau đó sẽ được tính điểm và so sánh để tìm ra các vùng mã hóa. Ngoài ra, các ORF có kích thước tối thiểu lớn hơn 900 base cũng được lựa chọn.

2.2.2.5 Ước lượng độ phong phú

Bước ước lượng độ phong phú của các unigene được đánh giá sử dụng phần mềm RSEM v1.2.29 [34]. Các đoạn đọc sau khi lọc chất lượng được sử dụng để tính toán ước lượng dựa trên hệ phiên mã đã lắp ráp sử dụng phần mềm Bowtie tích hợp trong RSEM. Nhờ đó, RSEM có thể ước tính biểu hiện gen và tính toán độ phong phú của các unigene dựa trên số lượng các đoạn đọc. Các thông số được sử dụng trong ước lượng độ phong phú đều là các giá trị mặc định của RSEM.

2.2.3 Chú giải chức năng hệ phiên mã

2.2.3.1 Chú giải hệ phiên mã dựa trên cơ sở dữ liệu GO

Quá trình chú giải các unigene trên cơ sở dữ liệu GO được thực hiện bằng công cụ BLASTX trong phần mềm DIAMOND với ngưỡng E-value = 1.0E-5. Quá trình chú giải sẽ đưa ra mã GO chứa đựng thông tin về trình tự và chức năng của các gen/ protein tham chiếu. Các mã chú giải GO được phân thành 3 nhóm chính dựa trên chức năng bao gồm các gen liên quan đến các quá trình sinh học (BP), thành phần tế bào (CC) và chức năng phân tử (MF). Trong 3 nhóm này, mỗi unigene sẽ được phân loại chi tiết hơn về nhóm chức năng.

2.2.3.2 Chú giải dựa trên cơ sở dữ liệu EggNOG

Chú giải các unigene trên cơ sở dữ liệu EggNOG bao gồm việc tìm kiếm, xác định các trình tự tương đồng trên Eukaryotic Orthologous Groups (KOGs), Cluster of Orthologous Groups (COGs) và Non-supervised Orthologous Groups (NOGs). Công cụ BLASTX trong phần mềm DIAMOND cũng được sử dụng để thực hiện chú giải các unigene dựa trên ngưỡng E-value là 1.0E-5 [35]. Các unigene đã chú giải sẽ được sử dụng để tìm kiếm chú giải trên các nhóm tương đồng tương ứng trên cơ sở dữ liệu EggNOG.

2.2.3.3 Chú giải dựa trên cơ sở dữ liệu NT và NR của NCBI

Các unigene được chú giải sử dụng công cụ BLASTN của NCBI [36]. Giá trị ngưỡng E value được sử dụng trong quá trình chú giải là 1.0E-5. Cơ sở dữ liệu NT cung cấp thông tin về trình tự các nucleic acid tham chiếu. Cơ sở dữ liệu NR chứa thông tin về trình tự các chuỗi amino acid tham chiếu.

2.2.3.4 Chú giải dựa trên cơ sở dữ liệu KEGG

Để chú giải chức năng của các unigene thu được trong nghiên cứu dựa trên cơ sở dữ liệu KEGG, công cụ BLASTX của phần mềm DIAMOND được sử dụng với ngưỡng giá trị E-value là 1.0E-5. Việc tìm kiếm trình tự tương đồng được thực hiện bằng phương pháp Bi-directional Best Hit (BBH). Kết quả tìm kiếm và chú giải là các mã KO chứa các thông tin về protein và các con đường chuyển hóa, tổng hợp tham chiếu liên quan.

2.2.3.5 Chú giải dựa trên cơ sở dữ liệu UniProt

Quá trình chú giải các unigene trên cơ sở dữ liệu UniProt được thực hiện bằng công cụ BLASTX trong phần mềm DIAMOND với ngưỡng E-value = 1.0E-5. Các trình tự unigene chú giải sẽ được tìm kiếm tương đồng trên hệ thông thông tin về chức năng của các protein tham chiếu UniProt KnowledgeBase (UniProtKB). Cơ sở dữ liệu này cũng cung cấp thông tin về tên, trình tự amino acid, dữ liệu phân loại hoặc tham khảo... thông qua mã UniProtKB.

2.2.3.6 Chú giải dựa trên cơ sở dữ liệu Pfam

Việc chú giải các unigene trên cơ sở dữ liệu Pfam được thực hiện bằng công cụ BLASTX trong phần mềm DIAMOND với ngưỡng E-value = 1.0E-5. Chú giải trên Pfam được thực hiện thông qua so sánh sắp hàng trình tự và sử dụng các mô hình Markov ẩn (Hidden Markov models, HMMs). Quá trình chú giải sẽ cung cấp thông tin về các domain hoạt động, nhờ đó dự đoán chức năng của các protein.

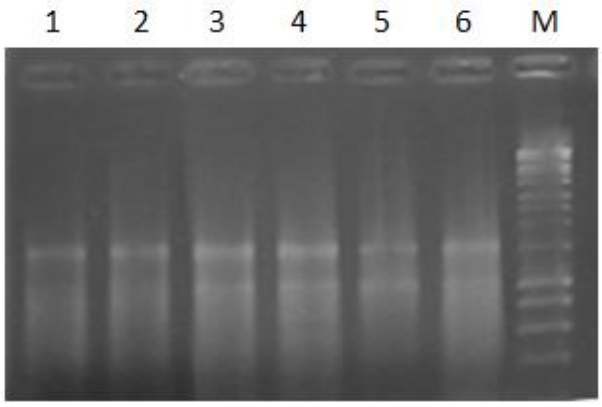
CHƯƠNG 3. KẾT QUẢ VÀ THẢO LUẬN

3.1 GIẢI TRÌNH TỰ HỆ PHIÊN MÃ ĐẶC HIỆU MÔ LÁ VÀ THÂN RỄ SÂM NGỌC LINH 4 NĂM TUỔI

3.1.1 Kết quả tách chiết, tinh sạch RNA tổng số và xây dựng các thư viện cDNA

RNA tổng số ở các mẫu lá và thân rễ của sâm Ngọc Linh được tách chiết, tinh sạch và đánh giá chất lượng bằng điện di trên gel agarose 8% (

Hình 3.1). Kết quả cho thấy, các mẫu RNA thu được có chất lượng tương đối tốt, có sự xuất hiện của các băng ribosome RNA với kích thước khoảng 1,5-2 kb. Để tinh sạch một số mẫu RNA, DNase được sử dụng nhằm loại bỏ DNA tổng số.



Hình 3.1 Kiểm tra kết quả tách chiết và tinh sạch RNA tổng số trên gel agarose

M: Thang chuẩn DNA 1 kb; trong đó, số thứ tự tương ứng với các mẫu: 1-3: C4.1-C4.3, 4-6: L4.1-L4.3

Kết quả kiểm tra nồng độ và độ sạch của RNA (Bảng 3.1) cho thấy, các sản phẩm có chỉ số A260/A280 thể hiện sự tinh sạch của mẫu dao động từ 1,90-2,15 chứng tỏ các mẫu RNA tách chiết được đủ điều kiện để tiến hành những thí nghiệm tiếp theo. Các RNA sau đó được phân thành những đoạn có kích thước

khoảng 450-550 bp sử dụng máy M220 Focused Ultrasonicator, đáp ứng các yêu cầu và được sử dụng để tổng hợp và xây dựng thư viện cDNA phục vụ giải trình tự gen thế hệ mới.

Bảng 3.1 Nồng độ và độ sạch (A260/A280) của các mẫu RNA sâm Ngọc Linh sau tách chiết và tinh sạch

STT	Tên mẫu	Nồng độ (ng/µl)	A260/A280
1	L4.1	1.010,0	1,96
2	L4.2	792,20	2,05
3	L4.3	1.468,3	2,08
4	C4.1	207,10	2,11
5	C4.2	482,40	2,11
6	C4.3	470,80	1,90

3.1.2 Kết quả giải trình tự hệ phiên mã của mô lá và mô thân rễ sâm Ngọc Linh 4 năm tuổi

Các thư viện được sử dụng cho giải trình tự thế hệ mới Illumina. Để đánh giá chất lượng của các đoạn đọc thô thu được sau khi giải trình tự, phần mềm FastQC được sử dụng để phân tích các chỉ số quan trọng của bộ dữ liệu như số lượng, kích thước của các đoạn đọc, % GC, điểm chất lượng trung bình (Q) của các trình tự theo từng base của các đoạn đọc, mức độ lặp của các đoạn đọc... Điểm chất lượng được biểu diễn dưới dạng biểu đồ hộp, trong đó hộp màu vàng thể hiện 50% số lượng phân bố các điểm chất lượng (gọi là khoảng interquartile), đường màu xanh chỉ giá trị trung bình. Điểm chất lượng trung bình 20 nghĩa là 99% độ chính xác và các đoạn đọc vượt qua điểm 20 được cho là chất lượng tốt. Điểm chất lượng trung bình từ 28 trở lên được đánh giá là chất lượng đọc rất tốt và đáng tin cậy. Điểm chất lượng trung bình bằng 30 tương đương với độ tin cậy 99,9% [37]. Thông tin thống kê cơ bản của kết quả giải trình tự các thư viện cDNA (Bảng 3.2) cho thấy đã thu được một số lượng lớn

các đoạn đọc từ các mẫu mô lá, thân rễ của sâm Ngọc Linh 4 năm tuổi với chất lượng khá cao, có độ tin cậy và đạt yêu cầu cho những phân tích tiếp theo.

Bảng 3.2 Thống kê các bộ dữ liệu thô khi giải trình tự các thư viện cDNA

Mẫu	Tổng số đoạn đọc	Tổng số base đọc	GC (%)	Q20 (%)	Q30 (%)
L4.1	89.888.850	9.078.773.850	43,33	98,89	96,25
L4.2	75.082.240	7.583.306.240	44,53	98,95	96,43
L4.3	66.477.494	6.714.226.894	43,15	98,20	94,39
C4.1	62.126.214	6.274.747.614	44,06	98,89	96,26
C4.2	78.869.048	7.925.373.848	42,70	97,14	92,04
C4.3	63.378.710	6.401.249.710	42,28	98,79	96,04
Tổng	435.822.556	43.977.678.156			

3.2 PHÂN TÍCH VÀ LẮP RÁP CÁC HỆ PHIÊN MÃ

3.2.1 Kết quả kiểm tra chất lượng các đoạn đọc sau khi trimming

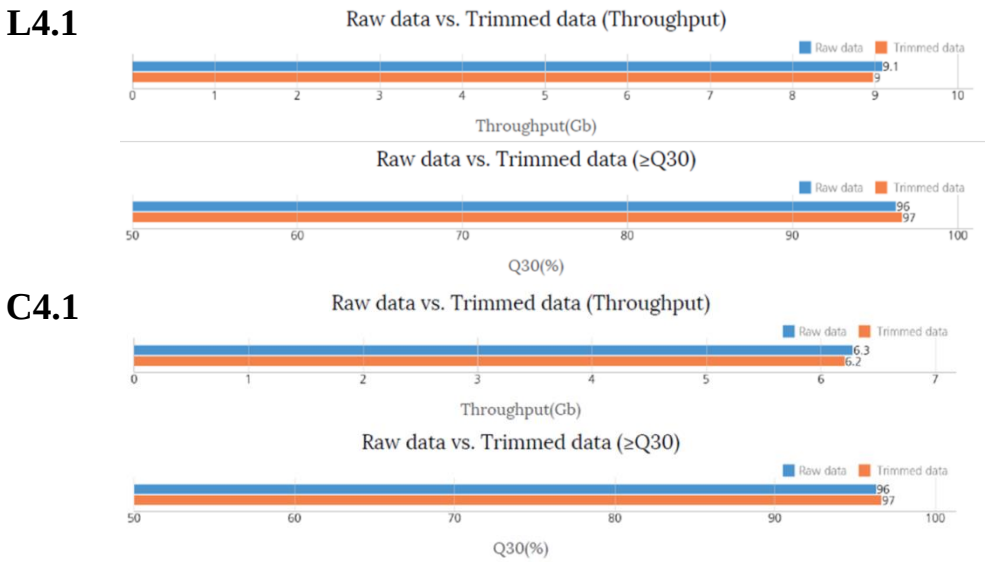
Các bộ dữ liệu thô, sau khi được kiểm tra chất lượng bằng FastQC, tiếp tục được xử lý nhằm loại bỏ trình tự adapter và các base có điểm chất lượng thấp hơn ba từ hai đầu sử dụng Trimmomatic. Theo đó, các đoạn đọc có độ dài ngắn hơn 36 bp sẽ được loại bỏ để tạo ra dữ liệu sau khi trimming (Bảng 3.3). Kết quả, 429.930.834 các đoạn đọc tương ứng với 43.228.319.306 base đã thu được.

Bảng 3.3 Thống kê các bộ dữ liệu thu được sau khi trimming

Mẫu	Tổng số đoạn đọc	Tổng số base đọc	GC (%)	Q20 (%)	Q30 (%)
L4.1	89.141.174	8.975.651.153	43,34	99,12	96,61
L4.2	74.486.684	7.496.686.220	44,53	99,17	96,77
L4.3	65.523.252	6.583.474.260	43,17	98,64	95,01
C4.1	61.627.462	6.203.290.729	44,07	99,12	96,62

C4.2	76.350.630	7.648.769.112	42,74	97,88	93,14
C4.3	62.801.632	6.320.447.832	42,29	99,06	96,45
Tổng	429.930.834	43.228.319.306			

Sau khi trimming, điểm chất lượng trung bình của các trình tự theo từng base của các đoạn đọc cũng được đánh giá lại. So sánh chất lượng đoạn đọc của dữ liệu thô và dữ liệu sau khi trimming cho thấy dữ liệu giải trình tự thô ban đầu có chất lượng khá tốt và bước trimming để loại bỏ những phần không đạt yêu cầu giúp cho kết quả giải trình tự các đoạn đọc được tốt hơn (Hình 3.2). Kết quả lọc chất lượng giúp giảm số lượng đoạn đọc chất lượng thấp thông qua loại bỏ những vùng trình tự kém chất lượng trong các đoạn đọc. Điều này khiến cho số lượng đoạn đọc thu được sau bước lọc giảm nhưng chất lượng của các đoạn đọc lại tăng. Nhìn chung, phân tích điểm chất lượng trung bình của các đoạn đọc chỉ ra kết quả giải trình tự gen thế hệ mới với các mẫu cDNA của mô lá và mô thân rễ sâm Ngọc Linh thu được là tương đối tốt và đủ điều kiện cho lắp ráp và chú giải hệ phiên mã.



Hình 3.2 So sánh dữ liệu thô và dữ liệu sau trimming ở mẫu mô lá (L4.1) và thân rễ (C4.1) của sâm Ngọc Linh 4 năm tuổi

3.2.2 Kết quả lắp ráp *de novo* các hệ phiên mã

Quá trình phân tích dữ liệu trình tự hệ phiên mã của sâm Ngọc Linh bắt đầu bằng lắp ráp *de novo* các đoạn đọc trình tự đã chọn lọc chất lượng để tạo ra các contig là những đoạn trình tự có kích thước lớn hơn. Ở bước đầu tiên, phần mềm Trinity được sử dụng để chia nhỏ dữ liệu trình tự nhằm phân tích độc lập bằng biểu đồ de Brujin tương ứng với các gen hay locus. Theo đó, các đoạn đọc được lắp ráp gối lên nhau để nối thành những phân đoạn dài hơn mà không chứa các gap hay N. Quá trình lắp ráp *de novo* được thực hiện thông qua 3 module phần mềm riêng biệt của Trinity là Inchworm, Chrysalis và Butterfly. Sau quá trình lắp ráp, phần mềm Trinity sẽ tạo ra một tệp "Trinity.fasta" chứa thông tin về trình tự phiên mã. Các đoạn trình tự thuộc cùng một locus sẽ được phân nhóm thành các cluster dựa trên độ tương đồng về trình tự. Những cluster phiên mã này được tạm coi là các gen.

Bảng 3.4 thể hiện kết quả thống kê của toàn bộ các contig được lắp ráp ban đầu từ các đoạn đọc thu được sau quá trình lọc chất lượng bao gồm số lượng gen, số lượng bản phiên mã (transcript), tỉ lệ GC, chỉ số N50, kích thước trung bình của contig và tổng số base được lắp ráp. Dữ liệu cho thấy kết quả lắp ráp hệ phiên mã của các mẫu mô lá và thân rễ sâm ở 4 năm tuổi có sự khác nhau về số lượng gen (47.870-70.526), số lượng bản phiên mã (69.638-106.276) và số lượng base lắp ráp (44.043.555-95.505.500). Cụ thể, các mẫu mô lá sâm Ngọc Linh 4 năm tuổi khác nhau về số lượng gen (68.454-81.480), số lượng bản phiên mã (99.609-121.554) và số lượng base lắp ráp (66.828.420-95.505.500); các mẫu thân rễ sâm Ngọc Linh 4 năm tuổi khác nhau về số lượng gen (47.870-63.268), số lượng bản phiên mã (69.638-100.651) và số lượng base lắp ráp (44.043.555-69.878.264). Kết quả lắp ráp chi tiết các đoạn đọc của các mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi được trình bày trên Bảng 3.5. Các chỉ số như số lượng gen, số lượng bản phiên mã, tỉ lệ GC, chỉ số N90-N10, kích thước các contig và số lượng base lắp ráp được thống kê cho toàn bộ các contig thu được và cho riêng contig có isoform với kích thước lớn nhất sau lắp ráp. Trong

đó, chỉ số N50 thể hiện cho ít nhất 50% các nucleotide được lắp ráp được tìm thấy trong các contig có chiều dài nhỏ nhất và cho isoform dài nhất tương ứng với các giá trị 993, 915, 1.074 của các mẫu mô lá sâm Ngọc Linh 4 năm tuổi; 861, 927, 816 của các mẫu thân rễ sâm Ngọc Linh 4 năm tuổi.

Bảng 3.4 Kết quả thống kê của contig được lắp ráp đầu tiên

Mẫu	Số lượng gen	Số lượng transcript	GC (%)	N50 (bp)	Kích thước trung bình của contig (bp)	Tổng số base được lắp ráp (bp)
L4.1	70.526	106.276	40,59	1.127	728,62	77.434.527
L4.2	68.454	99.609	40,89	1.033	670,91	66.828.420
L4.3	81.480	121.554	40,08	1.325	785,70	95.505.500
C4.1	47.870	69.638	40,20	870	632,46	44.043.555
C4.2	63.268	100.651	39,01	1.023	694,26	69.878.264
C4.3	59.717	88.228	38,86	828	613,73	54.148.380

Bảng 3.5 Kết quả thống kê quá trình lắp ráp các đoạn đọc ở mẫu mô lá (L4.1) và thân rễ (C4.1)

Mẫu	Thông số lắp ráp	Tất cả các contig của transcript	Chỉ isoform dài nhất/"gen"
L4.1	Tổng số trinity "gen"	70.526	70.526
	Tổng số trinity transcript	106.276	70.526
	GC (%)	40,59	40,80
	N90	300	258
	N80	458	347
	N70	665	481
	N60	888	699
	N50	1.127	993
	N40	1.374	1.312
	N30	1.645	1.642
	N20	1.989	2.025
	N10	2.547	2.618

	Độ dài contig lớn nhất	7.794	7.794
	Độ dài contig nhỏ nhất	201	201
	Median của độ dài contig	460,0	356,0
	Độ dài contig trung bình	728,62	623,76
	Tổng số base được lắp ráp	77.434.527	43.990.999
C4.1	Tổng số trinity "gen"	47.870	47.870
	Tổng số trinity transcript	69.638	47.870
	GC (%)	40,20	40,55
	N90	291	263
	N80	430	361
	N70	587	512
	N60	736	697
	N50	870	861
	N40	1.008	1.015
	N30	1.147	1.161
	N20	1.315	1.336
	N10	1.568	1.590
	Độ dài contig lớn nhất	7.854	7.854
	Độ dài contig nhỏ nhất	201	201
	Median của độ dài contig	494,0	402,0
	Độ dài contig trung bình	632,46	589,77
	Tổng số base được lắp ráp	44.043.555	28.232.335

3.2.3 Kết quả phân nhóm các đoạn trình tự thành các unigene

Các đoạn đọc sau khi lắp ráp *de novo* được tiến hành phân nhóm để tìm ra các unigene sử dụng CD-HIT-EST. Đây là một thuật toán bắt đầu với trình tự đầu vào có kích thước lớn nhất được chọn làm nhóm đại diện đầu tiên và phân chia những trình tự còn lại thành trình tự đại diện hay dư thừa dựa trên độ tương đồng với các đại diện đang xét. Độ tương đồng trình tự được tính toán dựa trên số lượng các word chung bằng cách sử dụng word indexing và bảng đếm để lọc ra những so sánh trình tự không cần thiết và tính độ tương đồng. Các bản phiên mã có kích thước lớn nhất thuộc cùng một locus sau khi được phân nhóm sẽ được coi là các contig unigene. Bảng 3.6 thống kê các contig unigene sau khi phân nhóm bao gồm số lượng gen, số lượng bản phiên mã, tỉ lệ GC, chỉ số N50,

kích thước trung bình của contig và tổng số base được lắp ráp. Các dữ liệu cho thấy, các mẫu mô khác nhau về số lượng gen (46.034-67.882), số lượng bản phiên mã (46.034-67.882) và số lượng base lắp ráp (27.641.662-50.440.507). Sự chênh lệch này chủ yếu do ảnh hưởng của số lượng đoạn đọc cũng như số lượng và kích thước các contig được lắp ráp.

Bảng 3.6 Kết quả thống kê của contig unigene

Mẫu	Số lượng gen	Số lượng transcript	GC (%)	N50 (bp)	Kích thước trung bình của contig (bp)	Tổng số base được lắp ráp (bp)
L4.1	67.882	67.882	40,73	1.022	635,07	43.109.582
L4.2	66.796	66.796	41,00	932	594,70	39.723.558
L4.3	77.381	77.381	39,97	1.120	651,85	50.440.507
C4.1	46.034	46.034	40,49	876	600,46	27.641.662
C4.2	59.818	59.818	38,90	961	615,62	36.825.282
C4.3	57.892	57.892	39,04	828	583,11	33.757.379

3.2.4 Kết quả dự đoán khung đọc mở

Bảng 3.7 thể hiện dữ liệu thống kê kết quả dự đoán ORF của các mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi. Kết quả thống kê cho thấy, đối với 3 mẫu L4.1, L4.2 và L4.3, có khoảng 33,04-37,01% các unigene được dự đoán ORF. Trong số đó, có 92,65-96,17% được dự đoán có 1 ORF và 3,83-7,35% có nhiều ORF. Trong số 22.091-25.713 ORF được dự đoán có 32,68%-48,4% ORF hoàn chỉnh, 25,45-34,38% ORF thiếu bộ ba mở đầu, 5,76-7,72% ORF thiếu bộ ba kết thúc và 18,71-27,18% ORF thiếu cả 2 loại này. Đối với 3 mẫu C4.1, C4.2 và C4.3, có khoảng 23,22-36,18% các unigene được dự đoán ORF. Trong số đó, có 95,61-97,95% được dự đoán có 1 ORF và 2,05-4,39% có nhiều ORF. Trong số 14.513-20.695 ORF được dự đoán có 22,22-36,53% ORF hoàn chỉnh, 46,65-

65,9% ORF thiếu bộ ba mở đầu, 1,73-4,09% ORF thiếu bộ ba kết thúc và 10,15-12,72% ORF thiếu cả 2 loại này. Kết quả, không có sự khác nhau nhiều giữa các mẫu dựa trên tỉ lệ tương đối giữa số lượng unigene và ORF.

Bảng 3.7 Kết quả thống kê dự đoán ORF của các mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

Mẫu	Số lượng unigen	Unigen dự đoán là ORF	Unigen dự đoán là đơn ORF	Unigen dự đoán là đa ORF	Số lượng ORF	Gen hoàn chỉnh	Một phần/ vùng mang mã	Một phần của đầu 5'	Một phần của đầu 3'
L4.1	67.882	22.426 (33,04%)	21.293 (94,95%)	1.133 (5,05%)	23.619	9.311 (39,42%)	5.659 (23,96%)	6.826 (28,9%)	1.823 (7,72%)
L4.2	66.796	24.722 (37,01%)	23.775 (96,17%)	947 (3,83%)	25.713	8.403 (32,68%)	6.989 (27,18%)	8.840 (34,38%)	1.481 (5,76%)
L4.3	77.381	20.458 (26,44%)	18.954 (92,65%)	1.504 (7,35%)	22.091	10.692 (48,4%)	4.134 (18,71%)	5.622 (25,45%)	1.643 (7,44%)
C4.1	46.034	16.653 (36,18%)	16.272 (97,71%)	381 (2,29%)	17.044	4.967 (29,14%)	1.939 (11,38%)	9.650 (56,62%)	488 (2,86%)
C4.2	59.818	13.887 (23,22%)	13.277 (95,61%)	610 (4,39%)	14.513	5.302 (36,53%)	1.846 (12,72%)	6.771 (46,65%)	594 (4,09%)
C4.3	57.892	20.270 (35,01%)	19,855 (97,95%)	415 (2,05%)	20.695	4.599 (22,22%)	2.101 (10,15%)	13.638 (65,9%)	357 (1,73%)

3.2.5 Kết quả ước lượng độ phong phú

Các đoạn đọc sau lọc chất lượng ở các mẫu mô lá và mô thân rễ sâm Ngọc Linh 4 năm tuổi được so sánh với trình tự tham chiếu của chính các mẫu này đã lắp ráp sử dụng phần mềm Bowtie. Để phân tích sự khác biệt biểu hiện gen, độ phong phú của các unigene giữa các mẫu được ước tính thông qua số lượng đoạn đọc sử dụng thuật toán RSEM. Bảng 3.8 thể hiện tỉ lệ lắp ráp các đoạn đọc của các mẫu mô lá và thân rễ sâm Ngọc Linh. Kết quả phân tích cho thấy, có 66,92-72,76% các đoạn đọc được lắp ráp vào hệ phiên mã tham chiếu

của chính các mẫu này. Điều này cho thấy có 27,24-33,08% các đoạn đọc không được lắp ráp vào hệ gen tham chiếu.

Bảng 3.8 Tỉ lệ lắp ráp các đoạn đọc của mẫu mô lá và thân rễ sâm

Mẫu	Số lượng đoạn đọc được xử lý	Số lượng đoạn đọc được lắp ráp	Số lượng đoạn đọc không được lắp ráp
L4.1	89.141.174	59.649.098 (66,92%)	29.942.076 (33,08%)
L4.2	74.486.684	50.129.670 (67,30%)	24.357.014 (32,70%)
L4.3	65.523.252	45.102.590 (68,83%)	20.420.662 (31,17%)
C4.1	61.627.462	44.405.868 (72,06%)	17.221.594 (27,94%)
C4.2	76.350.630	52.520.512 (68,79%)	23.830.118 (31,21%)
C4.3	62.801.632	45.693.642 (72,76%)	17.107.990 (27,24%)

3.3 CHÚ GIẢI HỆ PHIÊN MÃ CỦA CÁC MÔ SÂM NGỌC LINH 4 NĂM TUỔI

3.3.1 Chú giải hệ phiên mã mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

Các unigene thu được ở các mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi được chú giải sử dụng các cơ sở dữ liệu Gene Ontology (GO), Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (EggNOG), NCBI Nucleotide (NT), NCBI non-redundant Protein (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Universal Protein Resource (UniProt) và Pfam.

3.3.1.1 Kết quả chú giải dựa trên cơ sở dữ liệu GO

Kết quả chú giải các transcript của hệ phiên mã mẫu mô lá và thân rễ sâm Ngọc Linh dựa trên sự tương đồng về mặt trình tự trên cơ sở dữ liệu GO sẽ tương ứng với các mã GO chứa thông tin về trình tự và chức năng của các gen hoặc các sản phẩm của gen tham chiếu. Các transcript đã chú giải được phân chia theo chức năng thành 3 nhóm chính: Chu trình sinh học (biological process

- BP), thành phần tế bào (cellular component - CC) và chức năng phân tử (molecular function - MF). Trong nghiên cứu này, kết quả chú giải chức năng cho tỉ lệ các transcript thuộc nhóm BP chiếm 13,80-18,09% ở mô lá cây 4 năm tuổi và 14,05-17,53% ở mô thân rễ cây 4 năm tuổi. Các unigene thuộc nhóm CC chiếm từ 15,15-19,37% ở mô lá cây 4 năm tuổi và 15,75-19,50% ở mô thân rễ cây 4 năm tuổi. Tỉ lệ này của các transcript thuộc nhóm MF là 13,41-17,35% ở mô lá cây 4 năm tuổi và 14,11-17,50% ở mô thân rễ cây 4 năm tuổi. Số lượng unigene không được chú giải chiếm từ 45,19-57,64% ở mô lá cây 4 năm tuổi và 45,46-56,08% ở mô thân rễ cây 4 năm tuổi. Có thể thấy, tỉ lệ các unigene không được chú giải tương đối cao. Điều này có thể giải thích do sự hạn chế về số lượng các gen/ sản phẩm của gen tham chiếu cho các loài thuộc chi *Panax* trên cơ sở dữ liệu. Hình 3.3 minh họa thông tin phân nhóm của các mã GO thu được sau quá trình chú giải mẫu mô L4.1 sâm 4 năm tuổi.

3.3.1.2 Kết quả chú giải dựa trên cơ sở dữ liệu EggNOG

Thông tin về tỉ lệ phân nhóm dựa theo chức năng của các unigene đã chú giải ở mẫu mô lá L4.1 dựa trên cơ sở dữ liệu EggNOG được minh họa trên Hình 3.4. Các unigene được chú giải trên cơ sở dữ liệu EggNOG được chia làm 3 nhóm chính: (1) Nhóm các gen có chức năng trong các quá trình sinh học và lưu trữ thông tin (information storage and processing); (2) Nhóm các gen liên quan đến chu trình và tín hiệu tế bào (cellular processes and signaling); (3) Nhóm các gen liên quan đến trao đổi chất (metabolism). Tuy nhiên, tỉ lệ các unigene không được chú giải vẫn tương đối cao do hạn chế về thông tin của các loài thuộc chi *Panax* trên cơ sở dữ liệu EggNOG.

3.3.1.3 Kết quả chú giải dựa trên cơ sở dữ liệu NT và NR của NCBI

Dữ liệu về trình tự nucleotide và trình tự amino acid trên cơ sở dữ liệu NT và NR của NCBI được sử dụng làm tham chiếu để chú giải các hệ phiên mã. Quá trình tìm kiếm tương đồng các unigene của hệ phiên mã sâm Ngọc Linh đưa ra kết quả là các mã NCBI đại diện tương ứng với các unigene so sánh, chứa

thông tin về trình tự, tên chú giải, kích thước, độ bao phủ, tỉ lệ tương đồng, số lượng gap, độ tin cậy, vị trí khung đọc mở... (Bảng 3.).

3.3.1.4 Kết quả chú giải dựa trên cơ sở dữ liệu KEGG

Cơ sở dữ liệu KEGG chứa chủ yếu thông tin về trình tự, chức năng và các pathway liên quan của các protein tham chiếu. Bảng 3. trình bày kết quả chú giải hệ phiên mã của mẫu mô lá và thân rễ sâm Ngọc Linh đại diện. Hình 3.6 mô tả thông tin phân nhóm của các mã GO thu được sau quá trình chú giải các mẫu. Tương ứng với các unigene là các mã KEGG chứa thông tin về trình tự, tên chú giải, pathway và các thông tin so sánh khác như kích thước, độ bao phủ, tỉ lệ tương đồng, số lượng gap, độ tin cậy, vị trí khung đọc mở...

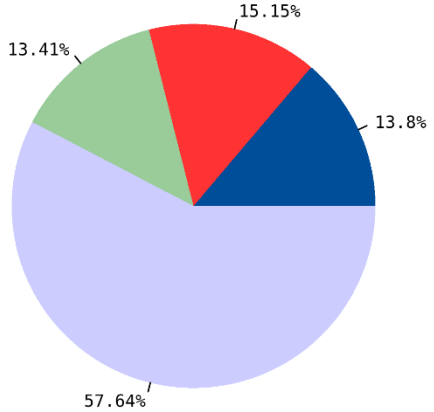
3.3.1.5 Kết quả chú giải dựa trên cơ sở dữ liệu UniProt

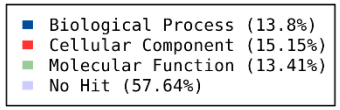
Dữ liệu về trình tự amino acid / protein trên cơ sở dữ liệu UniProt được sử dụng làm tham chiếu cho quá trình chú giải hệ phiên mã mẫu mô lá và thân rễ sâm Ngọc Linh. Bảng 3.1 trình bày kết quả chú giải hệ phiên mã đại diện. Tương ứng với các unigene là các mã UniProtKB chứa thông tin về trình tự và các thông tin so sánh khác như kích thước, độ bao phủ, tỉ lệ tương đồng, số lượng gap, độ tin cậy, vị trí khung đọc mở...

3.3.1.6 Kết quả chú giải dựa trên cơ sở dữ liệu Pfam

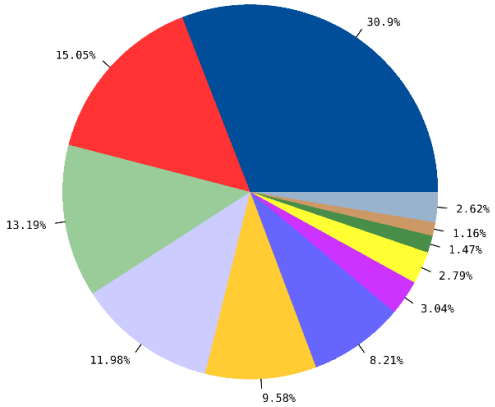
Cơ sở dữ liệu Pfam chứa thông tin về trình tự, chức năng và thông tin về các domain của các protein tham chiếu. Quá trình chú giải các unigene của hệ phiên mã mẫu mô lá và thân rễ sâm Ngọc Linh đưa ra kết quả là các mã chú giải Pfam tương ứng với các unigene so sánh, chứa thông tin về trình tự, tên chú giải và các thông tin so sánh khác như kích thước, độ bao phủ, tỉ lệ tương đồng, số lượng gap, độ tin cậy, vị trí khung đọc mở... (Bảng 3.11).

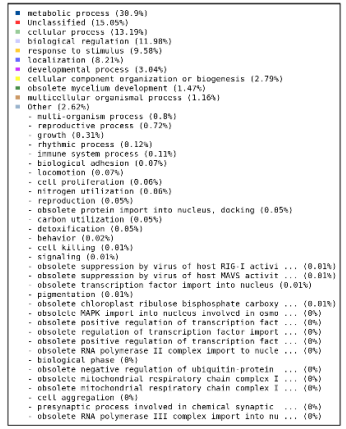
L4.1 GO Analysis Result



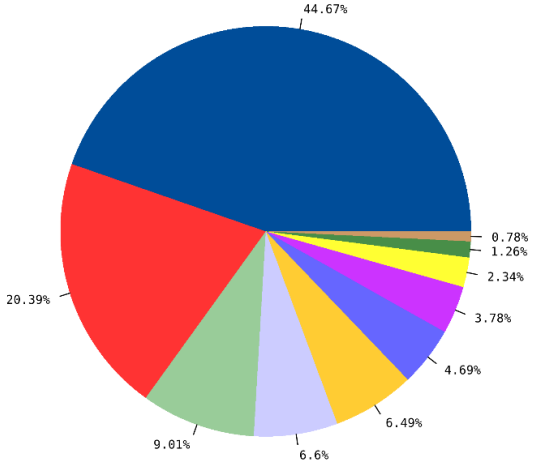


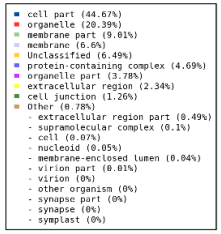
Biological Process



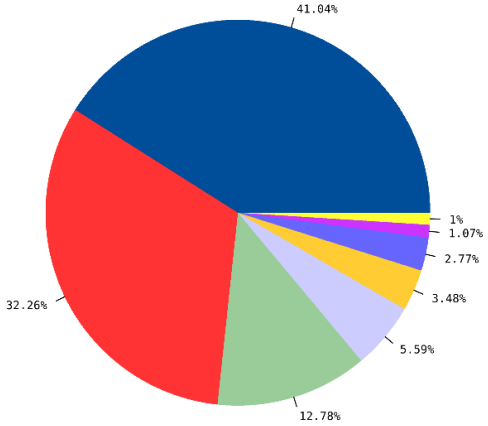


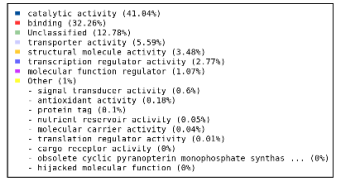
Cellular Component

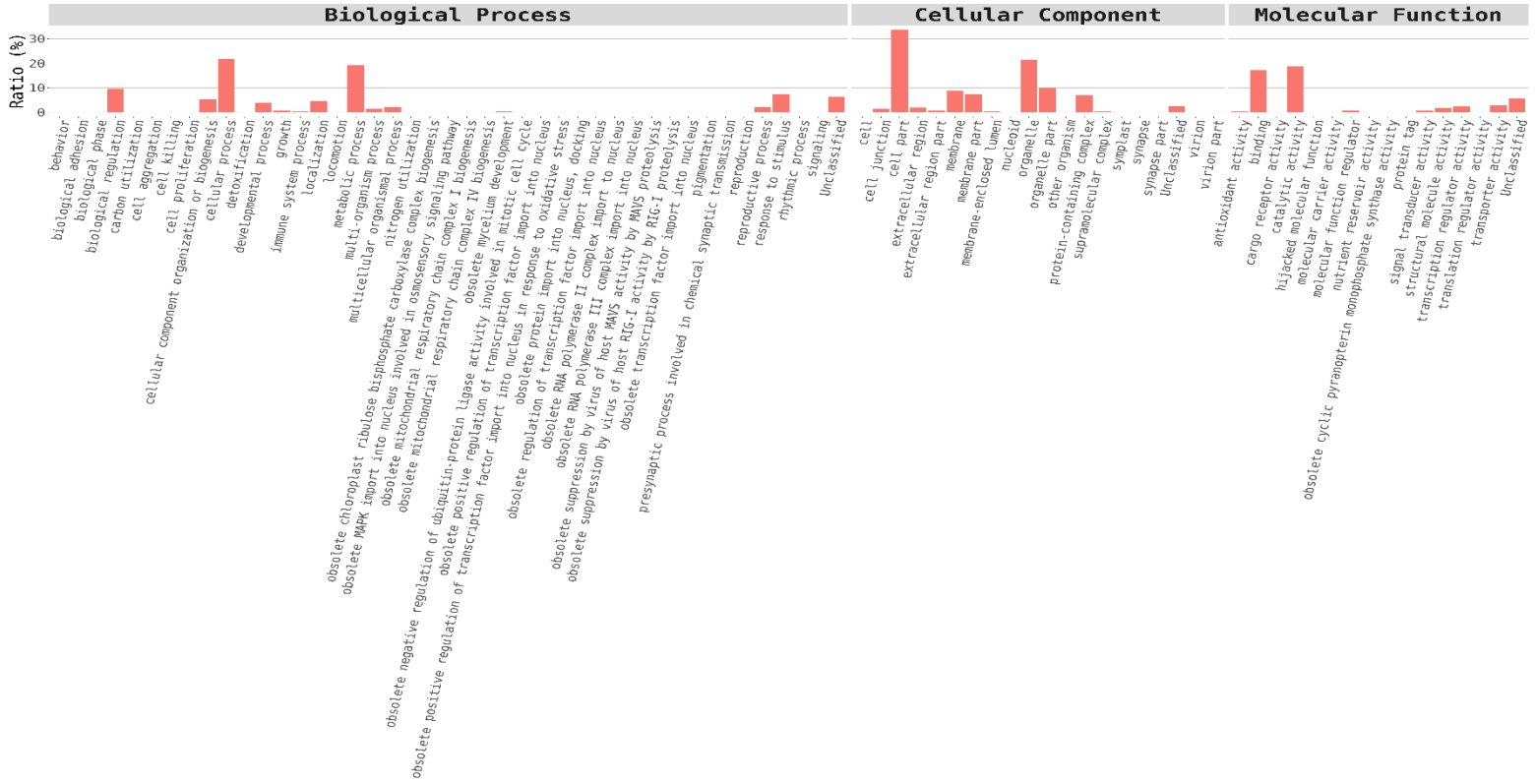




Molecular Function

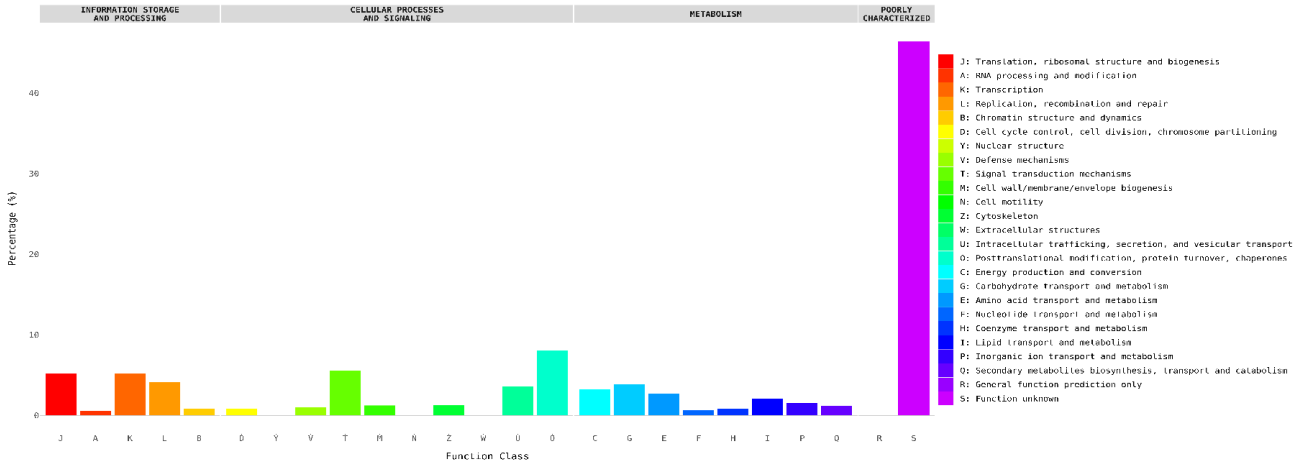






Hình 3.3. Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) dựa trên cơ sở dữ liệu GO

L4.1



Hình 3.4. Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) dựa trên cơ sở dữ liệu EggNOG

Bảng 3.9 Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) và thân rễ (C4.1) sâm Ngọc Linh 4 năm tuổi dựa trên cơ sở dữ liệu NT/NR của NCBI

L4.1

NT

Query_ID	D1_5_L_Read_Count	D1_5_L_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Gaps	Gaps (%)	Query_Strand	Target_Strand
c3_g1_i1	2	1.19	267	1	267	100	XM_017392762 PREDICTED: Daucus carota	1763	1064	1330	15.14	233.8	2E-57	223/270	82	6/270	2	Plus	Plus
c4_g1_i1	5963	198.71	1636	151	1372	74.69	XM_017382511 PREDICTED: Daucus carota	1556	172	1396	78.73	957.7	0	996/1231	80	15/1231	1	Plus	Plus
c8_g2_i1	12	2.14	464	6	324	68.75	XM_022040345 PREDICTED: Carica papaya	6655	2111	2429	4.79	180.2	4E-41	246/320	76	2/320	0	Plus	Plus
c9_g1_i1	8	1.21	511	63	344	55.19	XM_017366427 PREDICTED: Daucus carota	1448	413	691	19.27	274.4	2E-69	240/284	84	7/284	2	Plus	Plus
c10_g1_i1	17	6.37	319	265	309	14.11	AM460859 Vitis vinifera contig VV78X1937	5921	4068	4112	0.76	69.4	6E-08	42/45	93	0/45	0	Plus	Plus
c11_g1_i1	1373	59.39	1305	268	1305	79.54	DQ487109 Panax ginseng dehydrin 4 (Dhn	1076	61	1071	93.96	1476.6	0	980/1055	92	61/1055	5	Plus	Plus
c17_g1_i1	190	11.15	1013	8	781	76.41	XM_017369103 PREDICTED: Daucus carota	884	1	762	86.2	549.6	6E-152	626/783	79	30/783	3	Plus	Plus
c23_g1_i1	25	1.66	919	18	205	20.46	KX887330 Bupleurum falcatum mitochonc	463792	307012	307199	0.04	276.3	1E-69	175/188	93	0/188	0	Plus	Plus
c27_g1_i2	102	4.99	1176	85	935	72.36	XM_002533412 PREDICTED: Ricinus commu	1401	266	1115	60.67	379.7	1E-100	652/867	75	33/867	3	Plus	Plus
c31_g1_i1	6575	323.48	1171	64	769	60.29	XM_016613923 PREDICTED: Nicotiana taba	1139	125	833	62.25	508.9	1E-139	568/712	79	9/712	1	Plus	Plus
c33_g1_i1	1394	36.32	2038	305	1760	71.44	XM_017391671 PREDICTED: Daucus carota	2104	428	1883	69.2	1626.2	0	1264/1456	86	0/1456	0	Plus	Plus
c41_g1_i1	7	1.24	465	21	463	95.27	XM_017385836 PREDICTED: Daucus carota	7339	856	1298	6.04	409.2	5E-110	370/444	83	2/444	0	Plus	Plus
c54_g1_i1	5880	174.08	1817	63	1775	94.28	GU997668 Panax notoginseng clone 00637	1709	2	1706	99.77	2911.4	0	1672/1717	97	16/1717	0	Plus	Plus
c89_g1_i1	714	37.80	1102	266	316	4.63	XM_009360239 PREDICTED: Pyrus x bretsch	1250	427	476	4	71.3	6E-08	47/51	92	1/51	1	Plus	Plus
c151_g1_i	8059	182.97	2310	213	1645	62.03	XM_017386090 PREDICTED: Daucus carota	1614	129	1561	88.79	1085.1	0	1161/1443	80	20/1443	1	Plus	Plus
c165_g1_i	8	1.36	478	1	478	100	KJ906036 Synthetic construct Homo sapie	1485	533	1010	32.19	883.8	0	478/478	100	0/478	0	Plus	Plus
c197_g1_i	0	0.00	235	78	140	26.81	XM_017398333 PREDICTED: Daucus carota	2169	30	93	2.95	67.6	1E-07	55/64	85	1/64	1	Plus	Plus
c207_g1_i	8	1.66	426	1	187	43.9	XM_007593504 Colletotrichum fioriniae P.	864	678	864	21.64	346.4	3E-91	187/187	100	0/187	0	Plus	Plus
c214_g1_i	6680	160.87	2189	328	1620	59.07	XM_017372767 PREDICTED: Daucus carota	2014	395	1684	64.05	1293.8	0	1100/1297	84	11/1297	0	Plus	Plus

NR

Query_ID	D1_5_L_Read_Count	D1_5_L_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c3_g1_i1	2	1.19	267	1	268	100.37	PSS17981.1 Hippocampus abunda	401	272	360	22.19	154.8	1E-34	72/89	80	82/89	92	0/89	0	+1
c4_g1_i1	5963	198.71	1636	55	1366	80.2	XP_018841021.1 PREDICTED: nudix	447	4	447	99.33	662.1	1E-186	334/444	75	369/444	83	7/444	1	+1
c8_g1_i1	16	2.10	561	1	562	100.18	XP_017241313.1 PREDICTED: golgin	2231	726	912	8.38	250.8	3E-63	126/187	67	159/187	85	0/187	0	+1
c8_g2_i1	12	2.14	464	1	463	99.78	CDP01183.1 unnamed protein proc	1950	551	703	7.85	216.1	8E-53	110/154	71	129/154	83	1/154	0	+1
c9_g1_i1	8	1.21	511	89	347	50.68	XP_017221915.1 PREDICTED: uncha	377	67	152	22.81	156	1E-34	72/86	83	83/86	96	0/86	0	+2
c10_g1_i1	17	6.37	319	316	25	91.54	CBI37296.3 unnamed protein prod	3048	401	508	3.54	78.2	2E-11	51/114	44	60/114	52	23/114	20	-2
c11_g1_i1	1373	59.39	1305	377	956	44.44	WP_079403015.1 hypothetical prot	218	1	186	85.32	206.5	2E-49	178/193	92	179/193	92	7/193	3	+2
c12_g1_i2	589	39.55	910	126	648	57.47	KZM89260.1 hypothetical protein D	663	514	663	22.62	191	5E-45	101/178	56	124/178	69	32/178	17	+3
c15_g1_i1	4120	88.68	2426	1887	2109	9.19	EYU42196.1 hypothetical protein M	82	10	82	89.02	117.9	1E-22	60/74	81	64/74	86	1/74	1	+3
c17_g1_i1	190	11.15	1013	338	782	43.93	XP_017224592.1 PREDICTED: thylak	239	91	238	61.92	282	2E-72	134/148	90	143/148	96	0/148	0	+2
c19_g1_i1	273	15.70	1030	175	988	79.03	PSS16223.1 hypothetical protein C	288	1	282	97.92	130.2	1E-26	108/296	36	143/296	48	39/296	13	+1
c20_g1_i2	35	1.83	1113	1112	11	99.01	ACB28472.1 polyprotein, partial [A	953	569	928	37.78	430.6	5E-117	202/367	55	264/367	71	7/367	1	-3
c23_g1_i1	25	1.66	919	410	554	15.78	XP_018809152.1 PREDICTED: uncha	664	598	645	7.23	70.9	8E-09	32/48	66	38/48	79	0/48	0	+2
c27_g1_i2	102	4.99	1176	43	928	75.34	XP_021643121.1 zinc finger CCCH d	295	1	295	100	435.3	2E-118	219/299	73	244/299	81	8/299	2	+1
c31_g1_i1	6575	323.48	1171	50	764	61.06	KZM89147.1 hypothetical protein D	548	354	545	35.04	289.7	1E-74	154/238	64	171/238	71	46/238	19	+2
c33_g1_i1	1394	36.32	2038	294	1758	71.88	XP_017247160.1 PREDICTED: phosp	496	9	496	98.39	932.6	7E-268	447/488	91	472/488	96	0/488	0	+3
c41_g1_i1	7	1.24	465	7	466	98.92	CBI19108.3 unnamed protein prod	1038	179	333	14.93	244.6	2E-61	121/155	78	136/155	87	2/155	1	+1
c45_g1_i1	4	2.86	251	6	243	94.82	PSR89453.1 DnaJ subfamily B mem	803	227	304	9.71	83.6	3E-13	46/81	56	53/81	65	5/81	6	+3

C4.1

NT

Query_ID	B4_2_C_Read_Count	B4_2_C_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Gaps	Gaps (%)	Query_Strand	Target_Strand
c16_g1_i1	4	2.53	263	52	223	65.4	XM_007034091 PREDICTED: Theobron	2459	384	565	7.4	100.8	2E-17	142/183	77	12/183	6	Plus	Plus
c149_g1_i	1	1.44	211	2	119	55.92	XM_018979921 PREDICTED: Juglans r	3689	2441	2324	3.14	86.1	4E-13	96/120	80	4/120	3	Plus	Minus
c151_g1_i	0	0.00	215	11	215	95.35	XM_011045820 PREDICTED: Populus e	2140	968	1172	9.58	219	3E-53	177/206	85	2/206	0	Plus	Plus
c189_g1_i	17	2.01	586	37	325	49.32	KP940485 Neobartsia pedicularoide	99157	76583	76871	0.29	490.5	2E-134	281/289	97	0/289	0	Plus	Plus
c194_g1_i	2538	69.44	1878	546	642	5.17	XM_006291836 Capsella rubella hyp	688	405	501	14.1	108.2	8E-19	84/97	86	0/97	0	Plus	Plus
c200_g1_i	2167	94.50	1252	197	943	59.66	XM_017388501 PREDICTED: Daucus c	1308	267	1004	56.42	728.7	0	631/747	84	9/747	1	Plus	Plus
c216_g1_i	740	34.22	1192	149	927	65.35	XM_017373555 PREDICTED: Daucus c	2782	1719	2494	27.89	660.4	0	642/782	82	9/782	1	Plus	Plus
c217_g1_i	11	2.52	398	13	183	42.96	XM_018288017 Pochonia chlamydos	453	199	369	37.75	106.4	6E-19	134/172	77	2/172	1	Plus	Plus
c218_g1_i	5	1.36	366	96	158	17.21	XM_017844099 PREDICTED: Phoenix d	2033	1506	1568	3.1	78.7	1E-10	56/63	88	0/63	0	Plus	Plus
c220_g1_i	872	55.28	923	300	624	35.21	XM_017401611 PREDICTED: Daucus c	603	10	334	53.9	329.8	8E-86	276/325	84	0/325	0	Plus	Plus
c226_g1_i	9	1.88	417	233	417	44.36	XM_016834865 PREDICTED: Gossypiu	2087	367	554	9.01	100.8	3E-17	144/188	76	3/188	1	Plus	Plus
c227_g1_i	2	1.98	231	2	228	98.27	XM_017402715 PREDICTED: Daucus c	2000	546	772	11.35	265.2	5E-67	199/227	87	0/227	0	Plus	Plus
c238_g1_i	0	0.00	207	44	207	79.23	NG_008317 Homo sapiens fibroblas	687920	386323	386486	0.02	215.3	4E-52	149/165	90	2/165	1	Plus	Plus
c247_g1_i	33	3.81	596	3	244	40.6	XM_003632139 PREDICTED: Vitis vinif	1857	1306	1547	13.03	220.9	3E-53	202/243	83	2/243	0	Plus	Plus
c248_g1_i	546	28.75	1071	1	634	59.2	XM_017402327 PREDICTED: Daucus c	3182	2243	2875	19.89	649.3	0	540/634	85	1/634	0	Plus	Plus
c261_g1_i	9577	363.20	1410	194	1190	70.71	XM_017394140 PREDICTED: Daucus c	1353	188	1175	73.02	913.4	0	833/999	83	13/999	1	Plus	Plus
c268_g1_i	3030	319.31	634	257	460	32.18	XM_017361521 PREDICTED: Daucus c	816	395	597	24.88	243	7E-60	180/204	88	1/204	0	Plus	Plus
c273_g1_i	2445	258.25	633	108	427	50.55	XM_017387943 PREDICTED: Daucus c	637	77	397	50.39	414.8	1E-111	289/321	90	1/321	0	Plus	Plus
c274_g1_i	1306	78.46	963	179	705	54.72	XM_017394616 PREDICTED: Daucus c	1121	285	826	48.35	449.9	6E-122	446/543	82	17/543	3	Plus	Plus
c278_g1_i	54	5.11	684	1	242	35.38	XM_016724819 PREDICTED: Capsicum	2526	1697	1938	9.58	171	4E-38	195/245	79	6/245	2	Plus	Plus
c282_g2_i	6	10.52	202	1	142	70.3	XM_007773320 Coniophora puteana	1650	350	494	8.79	126.7	2E-25	122/147	82	7/147	4	Plus	Plus

NR

Query_ID	B4_2_C_Read_Count	B4_2_C_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c16_g1_i1	4	2.53	263	13	262	95.06	XP_021290198.1 transcription fa	578	1	93	16.09	114.4	2E-22	58/93	62	74/93	79	10/93	10	+1
c149_g1_i	1	1.44	211	190	4	88.63	KZM95709.1 hypothetical protei	985	596	657	6.29	77	3E-11	39/62	62	44/62	70	0/62	0	-2
c151_g1_i	0	0.00	215	3	216	99.53	XP_010086965.1 non-specific ph	526	215	285	13.5	146	5E-32	66/71	92	69/71	97	0/71	0	+3
c186_g1_i	10	4.54	295	1	295	100	XP_012442288.1 PREDICTED: unc	425	155	249	22.35	76.6	5E-11	37/98	37	59/98	60	3/98	3	+1
c189_g1_i	17	2.01	586	110	242	22.7	POE44573.1 hypothetical protei	116	73	116	37.93	94	6E-16	44/44	100	44/44	100	0/44	0	+2
c190_g1_i	15	1.70	603	284	533	41.46	KZN00512.1 hypothetical protei	209	108	207	47.85	59.3	0.00002	37/100	37	49/100	49	17/100	17	+2
c194_g1_i	2538	69.44	1878	138	648	27.21	XP_017217350.1 PREDICTED: prol	207	3	163	77.78	155.2	7E-34	89/171	52	116/171	67	11/171	6	+3
c200_g1_i	2167	94.50	1252	200	941	59.27	XP_017243990.1 PREDICTED: WU	280	37	280	87.14	406	1E-109	209/247	84	226/247	91	3/247	1	+2
c204_g1_i	15	4.10	365	2	188	51.23	OMO66499.1 Ctr copper transpo	145	72	134	43.45	89.7	7E-15	44/63	69	52/63	82	1/63	1	+2
c216_g1_i	740	34.22	1192	2	1040	87.16	XP_017229044.1 PREDICTED: tRN	867	512	867	41.06	398.7	2E-107	223/361	61	253/361	70	20/361	5	+2
c217_g1_i	11	2.52	398	1	265	66.58	CZR59742.1 related to UPF0591	150	63	150	58.67	168.3	2E-38	83/88	94	87/88	98	0/88	0	+1
c218_g1_i	5	1.36	366	1	145	39.62	CDP12353.1 unnamed protein p	463	416	463	10.37	88.6	1E-14	38/48	79	42/48	87	0/48	0	+1
c220_g1_i	872	55.28	923	305	623	34.56	XP_017257073.1 PREDICTED: unc	106	1	106	100	181.4	4E-42	89/106	83	99/106	93	0/106	0	+2
c226_g1_i	9	1.88	417	25	418	94.48	XP_021686325.1 preprotein tran	543	1	127	23.39	138.3	2E-29	73/132	55	94/132	71	6/132	4	+1
c227_g1_i	2	1.98	231	2	230	99.13	XP_017258202.1 PREDICTED: nuc	546	95	170	13.92	142.9	4E-31	71/76	93	74/76	97	0/76	0	+2
c238_g1_i	0	0.00	207	39	207	81.64	XP_011812390.1 PREDICTED: LOW	147	20	75	38.1	75.9	6E-11	39/56	69	41/56	73	0/56	0	+3
c246_g1_i	78	22.07	359	3	282	77.99	KZM92961.1 hypothetical protei	556	468	556	16.01	132.9	7E-28	66/94	70	76/94	80	6/94	6	+3
c247_g1_i	33	3.81	596	3	237	39.43	XP_018814447.1 PREDICTED: ribu	473	359	436	16.49	145.6	2E-31	69/78	88	75/78	96	0/78	0	+3
c248_g1_i	546	28.75	1071	2	644	60.04	XP_010656678.1 PREDICTED: SAR	944	731	944	22.67	352.4	2E-93	171/214	79	189/214	88	0/214	0	+2
c260_g1_i	1573	108.71	863	84	561	55.39	XP_017218034.1 PREDICTED: patl	158	1	158	100	136.3	1E-28	66/159	41	102/159	64	1/159	0	+3
c261_g1_i	9577	363.20	1410	197	1190	70.5	AHL39156.1 class III peroxidase	333	1	332	99.7	574.3	3E-160	288/332	86	307/332	92	1/332	0	+2

Bảng 3.10 Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) và thân rễ (L4.1) sâm Ngọc Linh 4 năm tuổi dựa trên cơ sở dữ liệu KEGG của NCBI

L4.1

Query_ID	D1_5_L_Read_Count	D1_5_L_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Accession	KO_ID	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identities	Identities (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c3_g1_i1	2	1.19	267	1	268	100.37	dcr:108219351 hippocampu	dcr:108219351	.	446	311	399	19.96	149.4	3E-34	68/89	76	78/89	87	0/89	0	+1
c4_g1_i1	5963	198.71	1636	148	1366	74.51	dcr:108224069 nudix hydrol	dcr:108224069	K03426	428	21	428	95.33	661	2E-187	324/409	79	360/409	88	4/409	0	+1
c8_g1_i1	16	2.10	561	1	562	100.18	dcr:108214054 golgin subfa	dcr:108214054	.	2231	726	912	8.38	250.8	2E-64	126/187	67	159/187	85	0/187	0	+1
c8_g2_i1	12	2.14	464	1	463	99.78	vvi:100255792 centromere-a	vvi:100255792	.	1913	567	724	8.26	197.2	2E-48	102/158	64	126/158	79	4/158	2	+1
c9_g1_i1	8	1.21	511	89	347	50.68	dcr:108198653 uncharacteri	dcr:108198653	.	377	67	152	22.81	156	7E-36	72/86	83	83/86	96	0/86	0	+2
c10_g1_i1	17	6.37	319	316	25	91.54	mdm:103405261 uncharacte	mdm:103405261	.	208	96	203	51.92	59.7	4E-07	41/112	36	56/112	50	19/112	16	-2
c11_g1_i1	1373	59.39	1305	764	1031	20.54	lang:109354969 phosphopr	lang:109354969	.	220	125	210	39.09	89.7	2E-15	54/92	58	64/92	69	9/92	9	+2
c12_g1_i2	589	39.55	910	126	648	57.47	dcr:108195282 uncharacteri	dcr:108195282	K17362	153	4	153	98.04	191	4E-46	101/178	56	124/178	69	32/178	17	+3
c15_g1_i1	4120	88.68	2426	1935	2109	7.21	psom:113345610 protein FA	psom:113345	.	182	125	182	31.87	107.5	1E-20	54/58	93	55/58	94	0/58	0	+3
c17_g1_i1	190	11.15	1013	338	782	43.93	dcr:108200841 thylakoid lu	dcr:108200841	.	239	91	238	61.92	282	2E-73	134/148	90	143/148	96	0/148	0	+2
c19_g1_i1	273	15.70	1030	187	988	77.86	hbr:110641378 uncharacteri	hbr:110641378	.	291	18	289	93.47	114.4	5E-23	93/285	32	133/285	46	31/285	10	+1
c20_g1_i2	35	1.83	1113	1112	11	99.01	zma:111589729 uncharacte	zma:11158972	.	1567	981	1343	23.17	414.5	2E-113	198/367	53	259/367	70	4/367	1	-3
c23_g1_i1	25	1.66	919	410	554	15.78	jre:108982253 uncharacteri	jre:108982253	.	664	598	645	7.23	70.9	5E-10	32/48	66	38/48	79	0/48	0	+2
c27_g1_i2	102	4.99	1176	43	928	75.34	hbr:110637365 zinc finger C	hbr:110637365	.	295	1	295	100	435.3	1E-119	219/299	73	244/299	81	8/299	2	+1
c31_g1_i1	6575	323.48	1171	68	764	59.52	dcr:108206054 elongation f	dcr:108206054	K03232	232	1	232	100	289.3	1E-75	194/232	83	213/232	91	0/232	0	+2
c33_g1_i1	1394	36.32	2038	294	1758	71.88	dcr:108218638 phosphoeth	dcr:108218638	K05929	496	9	496	98.39	932.6	5E-269	447/488	91	472/488	96	0/488	0	+3
c41_g1_i1	7	1.24	465	7	466	98.92	vvi:100255792 centromere-a	vvi:100255792	.	1913	179	333	8.1	244.6	1E-62	121/155	78	136/155	87	2/155	1	+1
c45_g1_i1	4	2.86	251	6	240	93.63	cit:102614074 uncharacteri	cit:102614074	.	806	195	269	9.31	69.3	4E-10	36/78	46	46/78	58	3/78	3	+3
c54_g1_i1	5880	174.08	1817	171	1629	80.3	dcr:108219710 fatty acid hy	dcr:108219710	K10528	488	2	488	99.8	742.7	6E-212	358/487	73	414/487	85	1/487	0	+3
c89_g1_i1	714	37.80	1102	477	873	36.03	dzi:111285486 probable E3	dzi:111285486	.	191	56	186	68.59	162.5	2E-37	84/134	62	100/134	74	5/134	3	+3
c117_g1	3	1.47	286	111	255	50.7	dcr:108223494 primary amir	dcr:108223494	K00276	729	1	48	6.58	49.7	0.0004	24/48	50	31/48	64	0/48	0	+3

C4.1

Query_ID	B4_2_C_Read_Cou	B4_2_C_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Accession	KO_ID	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identities	Identities (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c16_g1_i	4	2.53	263	13	262	95.06	tcc:18602600 transcription factor bHLH	tcc:18602600	.	578	1	93	16.09	111.7	6E-23	57/93	61	69/93	74	10/93	10	+1
c149_g1_	1	1.44	211	190	4	88.63	dcr:108222275 U-box domain-contain	dcr:108222275	.	1035	646	707	5.99	77	1E-12	39/62	62	44/62	70	0/62	0	-2
c151_g1_	0	0.00	215	3	216	99.53	pvu:PHAVU_009G165500g hypothetic	pvu:PHAVU_009G	K01114	515	204	274	13.79	145.6	3E-33	65/71	91	69/71	97	0/71	0	+3
c186_g1_	10	4.54	295	1	295	100	gra:105767338 uncharacterized LOC10	gra:105767338	.	425	155	249	22.35	76.6	3E-12	37/98	37	59/98	60	3/98	3	+1
c194_g1_	2538	69.44	1878	138	648	27.21	dcr:108194925 probable calcium-bindi	dcr:108194925	.	207	3	163	77.78	155.2	4E-35	89/171	52	116/171	67	11/171	6	+3
c200_g1_	2167	94.50	1252	200	941	59.27	dcr:108215890 WUSCHEL-related hom	dcr:108215890	.	280	37	280	87.14	406	8E-111	209/247	84	226/247	91	3/247	1	+2
c204_g1_	15	4.10	365	2	188	51.23	jcu:105646677 copper transporter 4	jcu:105646677	K14686	160	84	146	39.38	88.2	1E-15	45/63	71	52/63	82	1/63	1	+2
c216_g1_	740	34.22	1192	2	1040	87.16	dcr:108204226 tRNA (cytosine(34)-C(5	dcr:108204226	K15334	867	512	867	41.06	398.7	1E-108	223/361	61	253/361	70	20/361	5	+2
c217_g1_	11	2.52	398	1	265	66.58	mbe:MBM_08106 hypothetical protein	mbe:MBM_08106	.	150	63	150	58.67	166.8	3E-39	82/88	93	87/88	98	0/88	0	+1
c218_g1_	5	1.36	366	1	145	39.62	mus:103994883 F-box/kelch-repeat pr	mus:103994883	.	434	387	434	11.06	82	8E-14	32/48	66	42/48	87	0/48	0	+1
c220_g1_	872	55.28	923	305	623	34.56	dcr:108226599 uncharacterized protei	dcr:108226599	.	106	1	106	100	181.4	2E-43	89/106	83	99/106	93	0/106	0	+2
c226_g1_	9	1.88	417	25	418	94.48	hbr:110669133 preprotein translocase s	hbr:110669133	K10956	543	1	127	23.39	138.3	1E-30	73/132	55	94/132	71	6/132	4	+1
c227_g1_	2	1.98	231	2	230	99.13	dcr:108227525 nucleobase-ascorbate	dcr:108227525	K14611	546	95	170	13.92	142.9	2E-32	71/76	93	74/76	97	0/76	0	+2
c228_g1_	0	0.00	228	9	168	70.18	dcr:108192885 peptidyl-prolyl cis-trans	dcr:108192885	K14826	499	280	332	10.62	51.2	9E-05	27/53	50	33/53	62	0/53	0	+3
c238_g1_	0	0.00	207	38	206	81.64	nle:100591060 MTX3; metaxin 3	nle:100591060	K17776	315	117	172	17.78	71.6	6E-11	35/56	62	37/56	66	0/56	0	+2
c246_g1_	78	22.07	359	3	282	77.99	dcr:108223737 protein NRT1/ PTR FAM	dcr:108223737	.	607	519	607	14.66	132.9	4E-29	66/94	70	76/94	80	6/94	6	+3
c247_g1_	33	3.81	596	3	237	39.43	jre:108986327 ribulose bisphosphate c	jre:108986327	.	473	359	436	16.49	145.6	9E-33	69/78	88	75/78	96	0/78	0	+3
c248_g1_	546	28.75	1071	2	644	60.04	vvi:100266959 SART-1 family protein DI	vvi:100266959	K11984	944	731	944	22.67	352.4	9E-95	171/214	79	189/214	88	0/214	0	+2
c260_g1_	1573	108.71	863	84	561	55.39	dcr:108195566 pathogenesis-related p	dcr:108195566	.	158	1	158	100	136.3	8E-30	66/159	41	102/159	64	1/159	0	+3
c261_g1_	9577	363.20	1410	197	1190	70.5	pop:7495509 peroxidase 72	pop:7495509	K00430	333	1	332	99.7	572.4	7E-161	287/332	86	306/332	92	1/332	0	+2
c266_g1_	1856	52.70	1817	1	1522	83.76	dcr:108197577 U-box domain-containi	dcr:108197577	.	689	179	689	74.17	651.7	1E-184	347/512	67	413/512	80	6/512	1	+1
c268_g1_	3030	319.31	634	95	449	55.99	dcr:108194558 uncharacterized protei	dcr:108194558	.	127	1	127	100	186.4	5E-45	98/128	76	104/128	81	11/128	8	+2
c270_g1_	56	5.07	706	6	459	64.31	vvi:104879748 myosin-binding protein	vvi:104879748	.	778	611	767	20.18	108.2	2E-21	64/159	40	95/159	59	10/159	6	+3

Bảng 3.11 Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) và thân rễ (L4.1) sâm Ngọc Linh 4 năm tuổi dựa trên cơ sở dữ liệu UniProt của NCBI

L4.1

Query_ID	D1_5_L_Read_Count	D1_5_L_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c4_g1_i1	5963	198.71	1636	58	1366	80.01	sp|Q94A82|NUD19_ARATH Nudix hydrolase 1	438	4	438	99.32	594	2E-168	297/442	67	358/442	80	13/442	2	+1
c9_g1_i1	8	1.21	511	128	347	43.05	sp|Q700D9|MYBF_ARATH Putative Myb famil	255	13	85	28.63	106.3	3E-22	49/73	67	60/73	82	0/73	0	+2
c11_g1_i1	1373	59.39	1305	761	1013	19.39	sp|Q6ESR4|DHN1_ORYSJ Dehydrin DHN1 OS=	290	171	269	34.14	66.2	1E-09	47/100	47	50/100	50	17/100	17	+2
c17_g1_i1	190	11.15	1013	338	785	44.23	sp|P82658|TL19_ARATH Thylakoid lumenal 1	229	80	228	65.07	250.4	3E-65	116/149	77	132/149	88	0/149	0	+2
c19_g1_i1	273	15.70	1030	772	976	19.9	sp|Q9FMY4|Y5316_ARATH Uncharacterized p	271	184	265	30.26	63.5	5E-09	30/82	36	50/82	60	14/82	17	+1
c20_g1_i2	35	1.83	1113	1109	20	97.93	sp|Q7LHG5|YI31B_YEAST Transposon Ty3-I G	1498	966	1344	25.3	197.6	2E-49	129/386	33	188/386	48	30/386	7	-3
c23_g1_i1	25	1.66	919	434	575	15.45	sp|Q94HW2|POLR1_ARATH Retrovirus-relate	1466	1003	1049	3.21	50.4	0.00004	22/47	46	31/47	65	0/47	0	+2
c27_g1_i2	102	4.99	1176	55	922	73.81	sp|Q7F8R0|C3H14_ORYSJ Zinc finger CCCH do	300	6	297	97.33	328.2	1E-88	168/297	56	205/297	69	13/297	4	+1
c31_g1_i1	6575	323.48	1171	68	764	59.52	sp|P93447|EF1D_PIMBR Elongation factor 1-d	226	1	226	100	284.6	2E-75	198/232	85	217/232	93	6/232	2	+2
c33_g1_i1	1394	36.32	2038	300	1758	71.59	sp|Q944H0|PEAM2_ARATH Phosphomethylet	491	6	491	98.98	865.9	3E-250	406/486	83	457/486	94	0/486	0	+3
c54_g1_i1	5880	174.08	1817	240	1626	76.33	sp|K4CF70|HPL_SOLLC Fatty acid hydroperoxi	476	11	475	97.69	662.1	6E-189	321/466	68	389/466	83	5/466	1	+3
c89_g1_i1	714	37.80	1102	477	882	36.84	sp|Q852U6|RHY1A_ARATH Probable E3 ubiqu	250	99	247	59.6	77.8	3E-13	53/153	34	69/153	45	22/153	14	+3
c133_g1_i	0	0.00	203	3	204	99.51	sp|P38972|PUR4_YEAST Phosphoribosylformy	1358	1290	1356	4.93	91.3	4E-18	39/67	58	48/67	71	0/67	0	+3
c146_g1_i	1	0.99	228	1	223	97.81	sp|O74192|DAK_PICPA Dihydroxyacetone kin	608	301	368	11.18	60.8	7E-09	31/74	41	48/74	64	6/74	8	+1
c151_g1_i	8059	182.97	2310	247	1678	61.99	sp|Q9FKQ1|DTX27_ARATH Protein DETOXIFIC	486	11	481	96.91	689.1	6E-197	334/477	70	396/477	83	6/477	1	+1
c153_g1_i	1	1.10	221	2	107	47.96	sp|P38995|ATU2_YEAST Copper-transporting	1004	922	956	3.49	49.3	0.00002	24/35	68	27/35	77	0/35	0	+2
c158_g1_i	3	3.15	224	32	215	82.14	sp|P04158|SC1_SCHCO Fruiting body protein	109	1	63	57.8	45.8	0.0002	25/64	39	37/64	57	4/64	6	+2
c165_g1_i	8	1.36	478	2	479	100	sp|P81947|TBA1B_BOVIN Tubulin alpha-1B c	451	157	315	35.25	325.5	3E-88	159/159	100	159/159	100	0/159	0	+2

C4.1

Query_ID	B4_2_C_Read_Count	B4_2_C_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identites	Identites (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c151_g1_i	0	0.00	215	6	216	98.14	sp|Q8H965|NPC6_ARATH	520	212	281	13.46	133.7	8E-31	59/70	84	65/70	92	0/70	0	+3
c194_g1_i	2538	69.44	1878	150	780	33.6	sp|Q93Z27|CML45_ARATH	204	6	204	97.55	97.4	6E-19	77/210	36	101/210	48	11/210	5	+3
c200_g1_i	2167	94.50	1252	203	935	58.55	sp|Q5QMM3|WOX8_ORYS	267	26	266	90.26	264.2	2E-69	137/247	55	179/247	72	9/247	3	+2
c204_g1_i	15	4.10	365	8	185	48.77	sp|Q8SAA5|COPT4_ARATH	145	81	140	41.38	68.2	7E-11	32/60	53	46/60	76	1/60	1	+2
c217_g1_i	11	2.52	398	4	250	62.06	sp|Q9UTI9|YKZ2_SCHPO U	148	65	146	55.41	62	6E-09	31/82	37	51/82	62	0/82	0	+1
c218_g1_i	5	1.36	366	1	145	39.62	sp|Q8L736|SKI11_ARATH I	467	419	467	10.49	70.9	1E-11	31/49	63	38/49	77	1/49	2	+1
c226_g1_i	9	1.88	417	28	418	93.76	sp|Q38885|SCY1_ARATH P	551	1	134	24.32	96.7	2E-19	61/135	45	86/135	63	6/135	4	+1
c227_g1_i	2	1.98	231	2	230	99.13	sp|Q8GZD4|NAT3_ARATH	551	99	174	13.79	110.9	6E-24	50/76	65	66/76	86	0/76	0	+2
c238_g1_i	0	0.00	207	33	204	83.09	sp|Q6UX73|CP089_HUMAI	402	345	402	14.43	65.1	3E-10	35/58	60	41/58	70	1/58	1	+3
c246_g1_i	78	22.07	359	6	165	44.57	sp|Q9M817|PTR6_ARATH	607	527	584	9.56	82.8	3E-15	33/58	56	46/58	79	5/58	8	+3
c247_g1_i	33	3.81	596	3	237	39.43	sp|Q40281|RCA_MALDO R	437	360	437	17.85	141	1E-32	66/78	84	73/78	93	0/78	0	+3
c248_g1_i	546	28.75	1071	89	641	51.63	sp|Q9LFE0|DOT2_ARATH S	820	623	819	24.02	317.4	2E-85	159/197	80	174/197	88	13/197	6	+2
c260_g1_i	1573	108.71	863	84	561	55.39	sp|P27538|PR2_PETCR Pat	158	1	158	100	136	7E-31	69/159	43	101/159	63	1/159	0	+3
c261_g1_i	9577	363.20	1410	197	1190	70.5	sp|Q9FJZ9|PER72_ARATH	336	1	333	99.11	504.6	1E-141	243/333	72	290/333	87	2/333	0	+2
c266_g1_i	1856	52.70	1817	22	1519	82.44	sp|O80742|PUB19_ARATH	686	189	684	72.3	403.3	5E-111	231/506	45	325/506	64	17/506	3	+1
c271_g1_i	115	14.09	573	151	388	41.54	sp|Q9FN30|TAT2_ARATH I	414	3	81	19.08	89.7	4E-17	49/82	59	59/82	71	6/82	7	+1
c273_g1_i	2445	258.25	633	114	438	51.34	sp|O49636|MPC4_ARATH	108	1	108	100	205.3	7E-52	96/108	88	100/108	92	0/108	0	+3
c274_g1_i	1306	78.46	963	188	698	53.06	sp|Q9SHE9|LBD4_ARATH	172	1	172	100	244.6	1E-63	123/173	71	142/173	82	4/173	2	+2
c278_g1_i	54	5.11	684	3	267	38.74	sp|Q94BY4|OFT35_ARATH	652	559	646	13.5	159.5	5E-38	71/88	80	79/88	89	0/88	0	+3

Bảng 3.12 Kết quả chú giải và phân nhóm các gen thuộc hệ phiên mã của mẫu mô lá (L4.1) và thân rễ (L4.1) sâm Ngọc Linh 4 năm tuổi dựa trên cơ sở dữ liệu Pfam của NCBI

L4.1

Query_ID	D1_5_L_Read_Cou	D1_5_L_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Pfam_ID	Pfam_Symbol	Pfam_Description	CLAN_ID	CLAN_Name	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identities	Identities (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c3_g1_i1	2	1.19	267	1	211	79.03	E0CQV8_VITVI/10-382 E0CG	PF07690	MFS_1	Major Fac	CL0015	MFS	373	304	373	18.77	117.9	5E-25	53/70	75	61/70	87	0/70	0	+1
c4_g1_i1	5963	198.71	1636	772	1153	23.35	F6HH73_VITVI/208-334 F6H	PF00293	NUDIX	NUDIX do	CL0261	NUDIX	127	1	127	100	247.7	3E-63	117/127	92	124/127	97	0/127	0	+1
c9_g1_i1	8	1.21	511	164	320	30.72	K4D1Y8_SOLLC/63-114 K4D	PF00249	Myb_DNA	Myb-like	CL0123	HTH	52	1	52	100	101.3	9E-20	45/52	86	50/52	96	0/52	0	+2
c11_g1_i1	1373	59.39	1305	806	1013	15.94	B4G1H1_MAIZE/200-269 B4C	PF00257	Dehydrin	Dehydrin			70	2	70	98.57	63.9	4E-08	34/69	49	37/69	53	0/69	0	+2
c12_g1_i	589	39.55	910	264	576	34.4	B9RQJ7_RICCO/58-134 B9R	PF03061	4HBT	Thioester	CL0050	HotDog	77	1	77	100	91.3	2E-16	52/104	50	64/104	61	27/104	25	+3
c17_g1_i	190	11.15	1013	338	776	43.34	I1MJM9_SOYBN/70-236 I1MJ	PF01789	PsbP	PsbP			167	22	167	87.43	264.2	2E-68	123/146	84	137/146	93	0/146	0	+2
c20_g1_i	35	1.83	1113	479	131	31.36	T1LIH7_TRIUA/944-1059 T1LI	PF00665	rve	Integrase	CL0219	RNase_H	116	1	116	100	164.1	2E-38	74/116	63	93/116	80	0/116	0	-3
c23_g1_i	25	1.66	919	410	569	17.41	Q94LQ7_ORYSJ/550-793 Q	PF07727	RVT_2	Reverse t	CL0027	RdRP	244	13	65	21.72	60.8	2E-07	28/53	52	37/53	69	0/53	0	+2
c27_g1_i	102	4.99	1176	544	736	16.41	M1C6S2_SOLTU/169-233 M1	PF00013	KH_1	KH doma	CL0007	KH	65	1	64	98.46	114.8	2E-23	57/64	89	60/64	93	0/64	0	+1
c31_g1_i	6575	323.48	1171	503	764	22.37	C5WWP8_SORBI/139-226 C5	PF00736	EF1_GNE	EF-1 gua			88	2	88	98.86	157.9	2E-36	76/87	87	84/87	96	0/87	0	+2
c33_g1_i	1394	36.32	2038	1074	1680	29.78	M0RZC4_MUSAM/248-443 M	PF13489	Methyltra	Methyltra	CL0063	NADP_R	196	1	196	100	334.3	2E-89	161/202	79	182/202	90	6/202	2	+3
c54_g1_i	5880	174.08	1817	486	1599	61.31	I1LU31_SOYBN/96-467 I1LU3	PF00067	p450	Cytochro			372	1	370	99.46	501.9	8E-140	232/371	62	307/371	82	1/371	0	+3
c89_g1_i	714	37.80	1102	741	873	12.07	M1D124_SOLTU/145-188 M1D	PF13639	zf-RING_	Ring finge	CL0229	RING	44	1	44	100	86.3	6E-15	33/44	75	38/44	86	0/44	0	+3
c133_g1	0	0.00	203	3	204	99.51	B2WLR0_PYRTR/1097-1365	PF13507	GATase_	CobB/Co	CL0014	Glutamin	269	203	268	24.54	125.9	1E-27	57/67	85	61/67	91	1/67	1	+3
c145_g1	1	1.30	212	89	212	58.49	H6CC03_EXODN/117-305 H6	PF08325	WLM	WLM dom	CL0126	Peptidas	189	1	41	21.69	70.5	7E-11	32/41	78	36/41	87	0/41	0	+2
c146_g1	1	0.99	228	1	184	80.7	K1XP63_MARBU/21-370 K1X	PF02733	Dak1	Dak1 dom	CL0245	EDD	350	289	349	17.43	83.2	1E-14	41/61	67	47/61	77	0/61	0	+1
c151_g1	8059	182.97	2310	1036	1519	20.95	A0A087GEB7_ARAAL/268-4	PF01554	MatE	MatE	CL0222	MviN_MA	163	1	161	98.77	283.5	6E-74	136/161	84	153/161	95	0/161	0	+1
c158_g1	3	3.15	224	104	221	52.68	S7RN66_GLOTA/26-104 S7R	PF01185	Hydroph	Fungal hy			79	1	38	48.1	47.8	0.0005	23/39	58	29/39	74	1/39	2	+2
c165_g1	8	1.36	478	2	176	36.61	G3RQF7_GORGO/3-214 G3I	PF00091	Tubulin	Tubulin/F	CL0566	Tubulin	212	155	212	27.36	121.7	6E-26	58/58	100	58/58	100	0/58	0	+2
c173_g1	1	1.46	206	4	160	76.21	D8RE90_SELML/762-842 D8	PF10431	ClpB_D2	C-termin			81	30	81	64.2	67.8	4E-10	30/52	57	43/52	82	0/52	0	+1
c207_g1	8	1.66	426	2	131	30.52	H1V1W5_COLHI/1-87 H1V1W5	PF03661	UPF0121	Uncharac			87	45	87	49.43	87	1E-15	41/43	95	42/43	97	0/43	0	+2

C4.1

Query_ID	B4_2_C_Read_Cou	B4_2_C_FPKM	Query_Length	Query_Start	Query_End	Query_Coverage	Target_Description	Pfam_ID	Pfam_Symbol	Pfam_Description	CLAN_ID	CLAN_Name	Target_Length	Target_Start	Target_End	Target_Coverage	Bit	e-value	Identities	Identities (%)	Positives	Positives (%)	Gaps	Gaps (%)	Query_Frame
c151_g1	0	0.00	215	3	216	99.53	I1JXK3_SOYBN/29-387 I1JX						359	176	246	19.78	142.9	1E-32	65/71	91	69/71	97	0/71	0	+3
c194_g1	2538	69.44	1878	567	651	4.53	B9T0T1_RICCO/98-164 B9T						67	1	28	41.79	51.2	0.0004	24/28	85	26/28	92	0/28	0	+3
c200_g1	2167	94.50	1252	389	575	14.94	M0ZMN1_SOLTU/94-155 M0						62	1	62	100	128.6	1E-27	60/62	96	62/62	100	0/62	0	+2
c204_g1	15	4.10	365	8	170	44.66	B9HC68_POPTR/39-138 B9						100	47	100	54	86.7	2E-15	42/54	77	46/54	85	0/54	0	+2
c217_g1	11	2.52	398	1	235	59.05	K1W9H9_MARBU/12-140 K1						129	52	129	60.47	145.6	3E-33	74/78	94	77/78	98	0/78	0	+1
c220_g1	872	55.28	923	311	503	20.91	F6HFT6_VITVI/3-66 F6HFT6						64	1	64	100	108.6	1E-21	52/64	81	57/64	89	0/64	0	+2
c227_g1	2	1.98	231	2	230	99.13	A0A022QCM0_ERYGU/51-4						408	45	120	18.63	127.1	7E-28	60/76	78	71/76	93	0/76	0	+2
c238_g1	0	0.00	207	7	202	94.69	G1R2H9_NOMLE/659-932 G						334	118	182	19.46	48.5	0.0003	30/69	43	41/69	59	8/69	11	+1
c246_g1	78	22.07	359	57	144	24.51	M4DXC2_BRARP/544-810 M						267	1	29	10.86	50.4	0.0001	16/29	55	27/29	93	0/29	0	+3
c248_g1	546	28.75	1071	2	368	34.27	G7I9G8_MEDTR/168-780 G7						613	487	613	20.72	167.2	3E-39	83/127	65	97/127	76	5/127	3	+2
c260_g1	1573	108.71	863	84	549	54	M1A057_SOLTU/1-155 M1A0						155	1	155	100	97.8	2E-18	50/155	32	85/155	54	0/155	0	+3
c261_g1	9577	363.20	1410	335	1082	53.05	B9S693_RICCO/46-294 B9						249	1	249	100	453.8	2E-125	225/249	90	236/249	94	0/249	0	+2
c266_g1	1856	52.70	1817	1063	1507	24.49	B9SH76_RICCO/468-615 B9						148	1	148	100	170.6	4E-40	92/148	62	113/148	76	0/148	0	+1
c268_g1	3030	319.31	634	308	443	21.45	B9RK87_RICCO/78-122 B9F						45	1	45	100	84.7	1E-14	38/45	84	42/45	93	0/45	0	+2
c271_g1	115	14.09	573	241	388	25.83	B9SBC0_RICCO/37-403 B9						367	1	49	13.35	80.1	2E-13	34/49	69	45/49	91	0/49	0	+1
c273_g1	2445	258.25	633	126	438	49.45	M0TQJ0_MUSAM/5-108 M0						104	1	104	100	199.9	2E-49	90/104	86	98/104	94	0/104	0	+3
c274_g1	1306	78.46	963	224	521	30.94	B9GK49_POPTR/13-111 B9C						99	1	99	100	188.3	1E-45	91/99	91	93/99	93	0/99	0	+2
c282_g1	2	3.59	201	1	202	100.5	B0DBJ8_LACBS/2-99 B0DE						98	21	87	68.37	128.6	2E-28	65/67	97	67/67	100	0/67	0	+1
c282_g2	6	10.52	202	15	120	52.48	S7PXT5_GLOTA/89-123 S7						35	1	35	100	65.1	3E-09	31/35	88	33/35	94	0/35	0	+3
c284_g1	35	2.53	833	6	576	68.55	A0A022QYP8_ERYGU/33-2						252	47	236	75.4	292.4	5E-77	133/190	70	158/190	83	0/190	0	+3
c288_g1	478	34.06	843	2	593	70.23	B9INR7_POPTR/311-562 B9						252	56	252	78.17	320.5	2E-85	162/197	82	176/197	89	0/197	0	+2
c293_g1	128	7.58	974	212	290	8.11	I1KVG6_SOYBN/15-42 I1KV						28	3	28	92.86	50.8	0.0003	24/26	92	24/26	92	0/26	0	+2
c294_g1	29	3.14	623	210	378	27.13	B9S5W2_RICCO/27-83 B9S						57	1	57	100	107.5	1E-21	43/57	75	50/57	87	1/57	1	+3

3.3.2 Kết quả chú giải các unigene của mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

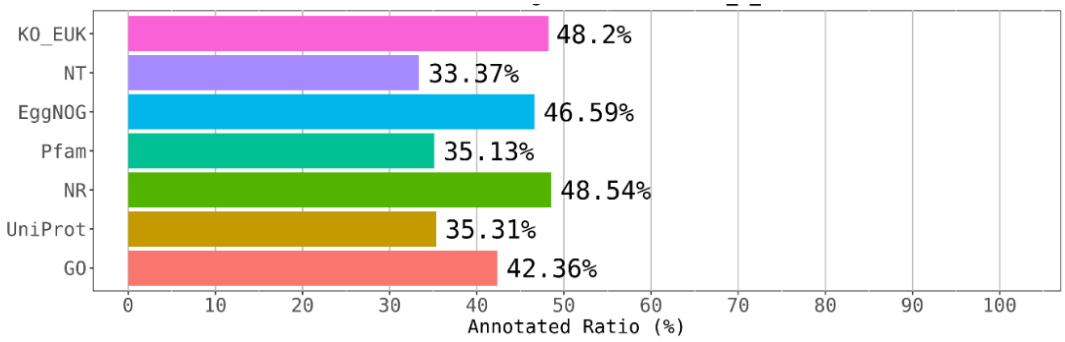
Quá trình chú giải hệ phiên mã mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi được thống kê trong Bảng 3.13. Kết quả tổng hợp cho thấy tỉ lệ các unigene được chú giải trên tổng số các unigene thu được chiếm 51,72-64,81% ở mẫu mô lá 4 tuổi và 52,87-66,32% ở mẫu thân rễ 4 tuổi.

Bảng 3.13 Tỉ lệ các unigene được chú giải trong hệ phiên mã các mẫu mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

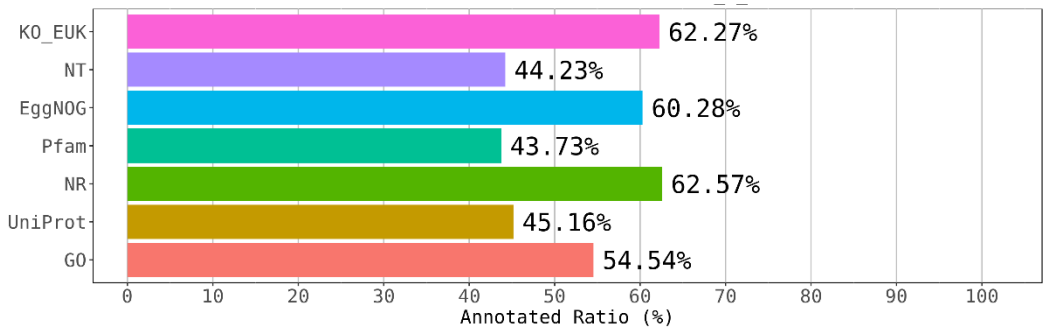
Mẫu	Số lượng unigene	Số lượng unigene được chú giải	Tỉ lệ chú giải (%)
L4.1	77.381	40.021	51,72
L4.2	66.796	43.289	64,81
L4.3	61.600	36.847	59,82
C4.1	46.034	30.528	66,32
C4.2	57.892	34.218	59,11
C4.3	59.818	31.624	52,87

Thông tin về tỉ lệ các unigene được chú giải ở từng mẫu trên từng cơ sở dữ liệu cũng được trình bày trong Hình 3.5. So sánh tỉ lệ chú giải thành công giữa các cơ sở dữ liệu nhận thấy KEGG, EggNOG và NR luôn đạt tỉ lệ chú giải thành công cao nhất, sau đó là GO, UniProt, Pfam và thấp nhất là NT.

L4.1



C4.1



Hình 3.5 Tỉ lệ các unigene của hệ phiên mã mẫu mô lá (L4.1) và thân rễ (C4.1) được chú giải trên các cơ sở dữ liệu

3.3.3 Tổng hợp, phân tích và so sánh dữ liệu hệ phiên mã ở mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

Dữ liệu hệ phiên mã các mẫu đơn được chuẩn hóa để có thể so sánh đối chiếu. Kết quả so sánh tỉ lệ chú giải cụ thể của unigene sâm Ngọc Linh 4 năm tuổi dựa trên các cơ sở dữ liệu khác nhau được tổng hợp trên

Bảng 3.4. Số lượng unigene được chú giải chức năng bởi ít nhất một cơ sở dữ liệu trong số các cơ sở dữ liệu GO, EggNOG, NT, NR, KEGG, UniProt và Pfam ở mẫu sâm 4 năm tuổi là 60,85%. Tỉ lệ chú giải cao nhất cũng thuộc về cơ sở dữ liệu KEGG với tỉ lệ 57,95%. Trong khi đó, cơ sở dữ liệu có khả năng chú giải số unigene thấp nhất thuộc về Uniprot với 45,40%.

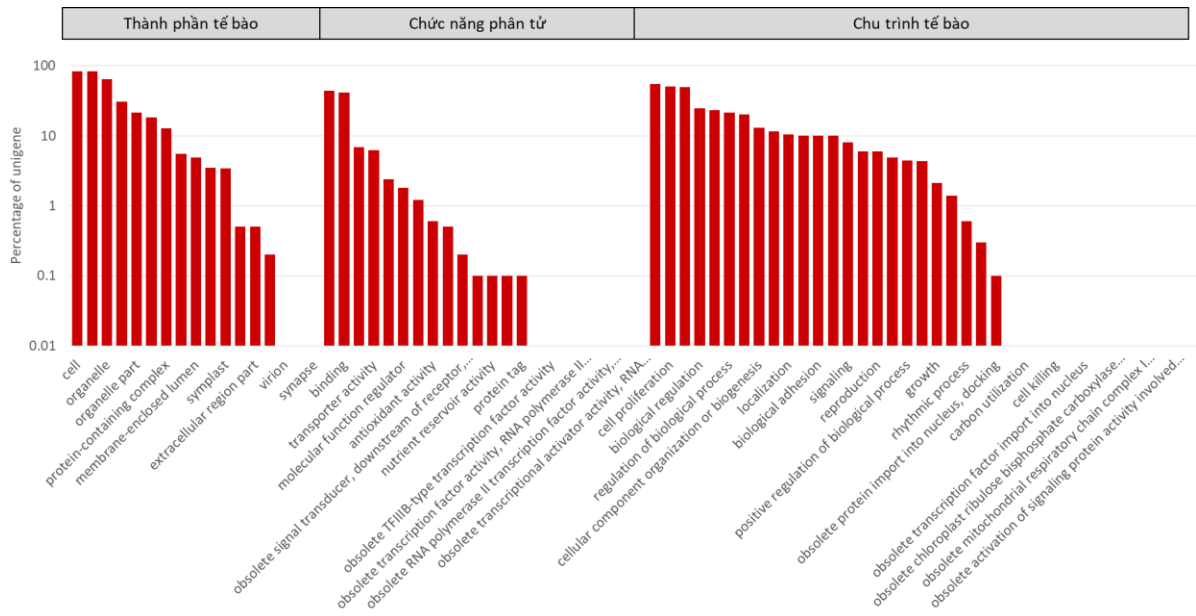
Bảng 3.14 Khả năng chú giải các unigene của hệ phiên mã sâm Ngọc Linh 4 năm tuổi trên các cơ sở dữ liệu khác nhau

Cơ sở dữ liệu	GO	Uniprot	NR	EggNOG	KEGG	NT	Pfam	Tổng
Số unigene	22.900	19.510	24.816	24.305	24.901	19.690	19.529	26.149
Tỉ lệ chú giải (%)	53,29	**45,40**	57,75	56,56	**57,95**	45,82	45,45	**60,85**

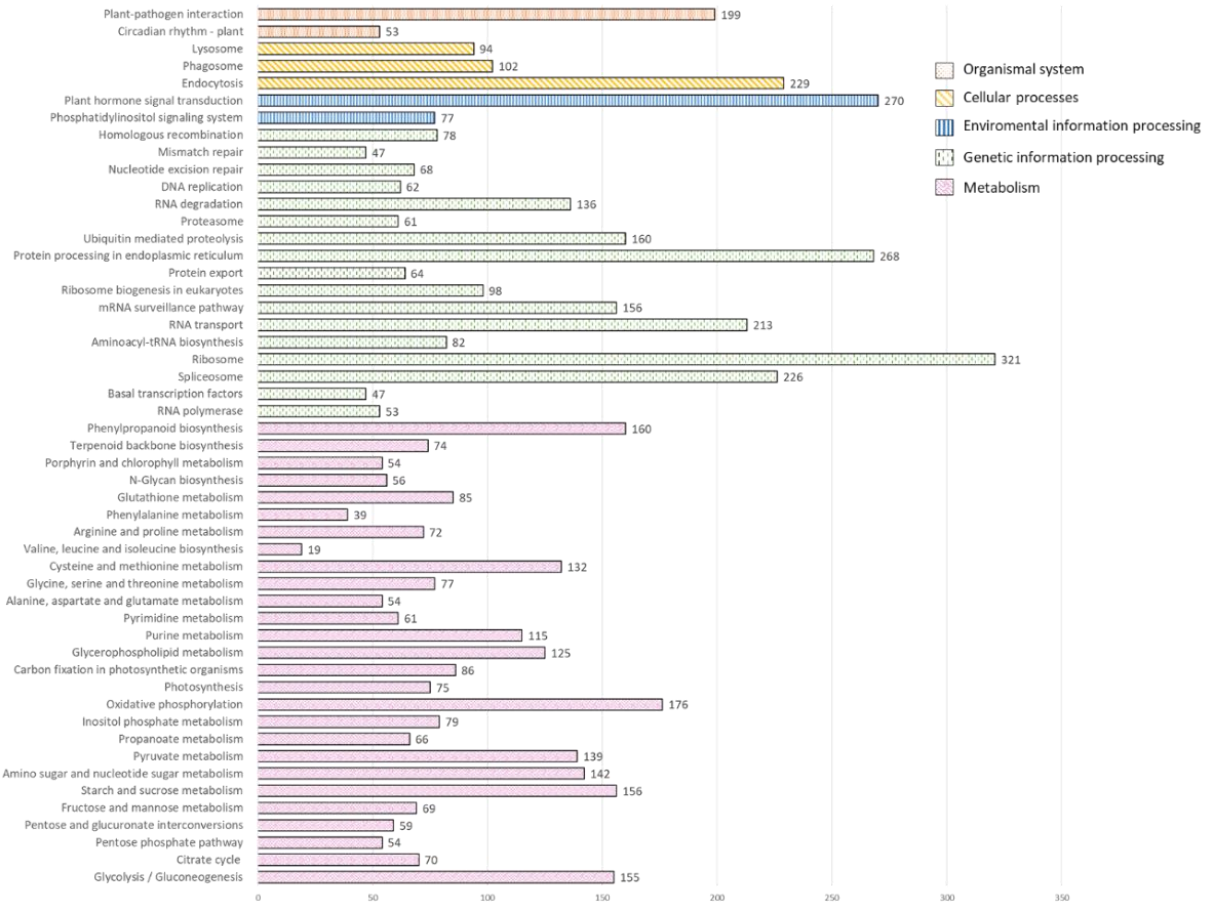
Với cơ sở dữ liệu GO, tất cả unigene được phân loại dựa trên 3 nhóm chức năng lớn và các nhóm nhỏ hơn. Đối với mẫu sâm Ngọc Linh 4 năm tuổi, tổng số 22.900 unigene được chú giải với 3 nhóm lớn CC, BP và MF có tỉ lệ lần

lượt là 47,81%, 36,96% và 15,23%. Trong nhóm CC, các unigene thuộc phân nhóm "tế bào/thành phần tế bào" chiếm tỉ lệ cao nhất với hơn 83%, tiếp đến là "bào quan" và "màng tế bào". Trong nhóm BP, chủ yếu là các unigene mã cho protein tham gia "chu trình tế bào" và "quá trình trao đổi chất" với tỉ lệ lần lượt là 54,6% và 49,0%. Phần lớn unigene nằm trong nhóm MF được phân loại vào phân nhóm "hoạt động xúc tác" với 10.096 unigene và "liên kết" với 9.522 unigene (Hình 3.6).

Với cơ sở dữ liệu KEGG, chức năng unigene liên quan đến các con đường chuyển hóa trong sinh vật được đánh giá. Ở hệ phiên mã sâm 4 năm tuổi, 24.901 unigene đã được chú giải và chia vào 211 con đường chuyển hóa trong sinh vật (Hình 3.7). Các unigene được chú giải này sẽ là cơ sở cho các nghiên cứu tiếp theo về các hợp chất thứ cấp quan trọng ở cây sâm Ngọc Linh Việt Nam.

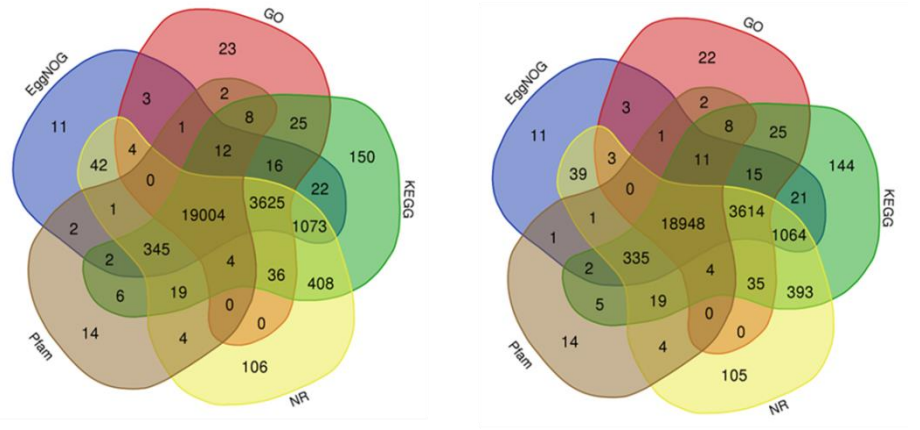


Hình 3.6 Chú giải GO cho unigene sâm Ngọc Linh 4 năm tuổi



Hình 3.7 Chú giải KEGG của hệ phiên mã sâm Ngọc Linh 4 năm tuổi

Hệ phiên mã của mô lá và mô thân rễ sâm Ngọc Linh 4 năm tuổi được so sánh dựa trên kết quả chú giải bởi các cơ sở dữ liệu GO, EggNOG, NR, KEGG và Pfam (Hình 3.8). Cơ sở dữ liệu có khả năng chú giải cao nhất là KEGG và cơ sở dữ liệu có khả năng chú giải thấp nhất trong 5 cơ sở dữ liệu trên là EggNOG.



Hình 3.8 Biểu đồ venn biểu diễn số unigene được chú giải bởi các cơ sở dữ liệu của hệ phiên mã mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

Tiếp theo, sự có mặt của các transcript mang ưu thế ở hệ phiên mã mỗi loại mô sâm 4 năm tuổi được phân tích. Bảng 3.15 cho thấy một số transcript mã hóa cho các protein tác động tới khả năng chống chịu bệnh, chống stress hay sinh trưởng và già hóa đều được tìm thấy ở hai hệ phiên mã sâm 4 năm tuổi. Tuy nhiên, sự biểu hiện của chúng khi được quan sát thấy ở lá thấp hơn ở thân rễ (dựa vào số lượng read count). Gen biểu hiện mạnh nhất ở lá là "Senescence-associated protein" và cao thứ hai ở rễ. Điều này có thể do thời điểm thu mẫu vào đầu thu, thời gian bắt đầu chuyển sang mùa rụng lá. Một số gen mã cho protein liên quan đến yếu tố gây bệnh, catalase_like hay metallothionein_like được tăng cường biểu hiện vào thời gian này [38], cũng được thấy biểu hiện cao ở lá. Ngoài ra, gen mã cho protein liên quan lục lạp (choloroplast) có sự biểu hiện cao ở lá - loại mô đặc trưng chứa lục lạp.

Bảng 3.15 Danh sách transcript chiếm ưu thế trong hệ phiên mã của mô lá và thân rễ sâm Ngọc Linh 4 năm tuổi

TT	Unigene ID	Accession no.	Chú giải	Read count
Mô lá				
1	c698379_g1_i1	mtr:MTR_0055s0050	Senescence-associated protein	1.600.684
2	c801000_g7_i2	MF377623	*P. vietnamensis* chloroplast	420.495
3	c494428_g1_i1	DQ384524	*P. ginseng* GBR5-like protein mRNA	393.661
4	c242479_g3_i1	dcr:108218502	Catalase-like	201.600
5	c801000_g2_i1	MF377623	*P. vietnamensis* chloroplast	190.935
6	c854330_g1_i1	dcr:108220108	Metallothionein-like protein type 3	150.255
7	c795392_g4_i3	dcr:108214401	Heat shock cognate 70 kDa protein 2-like	138.584
8	c492933_g1_i1	KJ741402	*P. notoginseng* pathogenesis-related protein (PR10-1) mRNA	110.463
9	c801571_g4_i1	DQ384525	*P. ginseng* ADP-ribosylation factor-like protein mRNA	86.711
10	c789616_g1_i1	dcr:108212952	Uncharacterized protein	73.169
Mô thân rễ				
1	c797413_g2_i1	KC751542	*P. notoginseng* RNase-like major storage protein	2.195.014
2	c698379_g1_i1	mtr:MTR_0055s0050	Senescence-associated protein	870.817
3	c801571_g4_i1	DQ384525	*P. ginseng* ADP-ribosylation factor-like protein	270.295
4	c795392_g4_i3	dcr:108214401	Heat shock cognate 70 kDa protein 2-like	113.164
5	c800014_g1_i2	dcr:108218481	Delta(12)-fatty-acid desaturase FAD2-like	97.079
6	c840150_g1_i1	qsu:112036655	Pathogenesis-related protein 1-like	94.124

7	c791685_g1_i2	pop:7494466	18.1 kDa class I heat shock protein	83.537
8	c696765_g1_i1	dcr:108224592	Protein translation factor SUI1 homolog	77.054
9	c739909_g1_i1	dcr:108218482	Delta(12)-fatty-acid desaturase FAD2-like	75.405
10	c494485_g1_i1	dcr:108192288	22.0 kDa heat shock protein-like	71.853

Kết quả từ bảng trên cũng cho thấy, một số unigene mã cho protein tham gia quá trình già hóa, protein heat shock và protein liên quan yếu tố gây bệnh đều xuất hiện nhiều ở hệ phiên mã mô lá và mô thân rễ. Về hệ phiên mã mô thân rễ sâm 4 năm tuổi, đứng đầu trong danh sách các transcript chiếm ưu thế là unigene mã cho "RNase-like major storage protein". Ngoài một số unigene mã cho protein heat shock hay liên quan yếu tố gây bệnh tương tự ở lá, có thể quan sát được một số unigene khác mã cho protein "ADP-ribosylation factor-like protein", "Delta(12)-fatty-acid desaturase FAD2-like" và "Protein translation factor SUI1 homolog" biểu hiện cao ở mô thân rễ. "Delta(12)-fatty-acid desaturase FAD2-like" có hai đại diện unigene (vị trí 5 và 9). Họ gen *FAD2-like* mã cho các enzyme bao gồm hydrolase, epoxygenase, conjugase và acetylenase góp phần đa dạng hóa các loại chất béo [39].

Tóm lại, sử dụng các cơ sở dữ liệu GO, EggNOG, NT, NR, KEGG, UniProt và Pfam, các unigene thu được ở các mẫu mô thân rễ và lá của sâm Ngọc Linh 4 năm tuổi đã được chú giải với tỉ lệ đạt từ 60,88-66,29%. KEGG, EggNOG và NR có tỉ lệ chú giải thành công cao nhất, tiếp đó là GO, UniProt, Pfam và thấp nhất là NT. Với GO, các unigene được phân vào 3 nhóm chức năng lớn và 52 nhóm nhỏ. Với NR, các trình tự gen sâm tương đồng với trình tự của 534 loài sinh vật. Với KEGG, 24.901 unigene đã được chú giải và chia vào 211 con đường chuyển hóa trong sinh vật; trong đó, 74 unigene thuộc con đường sinh tổng hợp khung terpenoid đã được chú giải, là cơ sở cho các nghiên cứu tiếp theo về các hoạt chất thứ cấp mang giá trị cao ở cây sâm Ngọc Linh.

Ngoài ra, các unigene thuộc các hệ phiên mã mô thân rễ và lá sâm có sự tương đồng về tỉ lệ giữa các nhóm chức năng. Ở hệ phiên mã, các transcript có ưu thế thường liên quan tới các protein chống chịu bệnh, sinh trưởng, già hóa và chống stress.

Việc có được cái nhìn bao quát về mức độ biểu hiện của sâm Ngọc Linh ở các thời điểm, lứa tuổi khác nhau (cụ thể trong nghiên cứu này là 4 năm tuổi), hoặc các mô khác nhau (ở đây là mô lá và mô thân rễ), cho phép cân nhắc để tối ưu các yếu tố sinh trưởng, canh tác, thời điểm thu hoạch…, qua đó có thể thu được hàm lượng các hợp chất thứ cấp quan trọng cao nhất. Do đó, kết quả của nghiên cứu này là mảnh ghép quan trọng để có thể hiểu rõ hơn các con đường tổng hợp, mức độ biểu hiện khác biệt ở hai nhóm mô khác nhau ở 4 năm tuổi, cũng như là tiền đề cho các nghiên cứu tiếp theo ở lứa tuổi/nhóm mô khác; bên cạnh đó, có định hướng nuôi trồng, chăm sóc và thu hoạch sâm vào thời điểm phù hợp và có hiệu quả cao nhất. Các kết quả thu được cũng góp phần xác định quyền sở hữu quốc gia về nguồn gen đặc hữu, từ đó có thể khai thác để đánh giá và nâng cao chất lượng nguồn gen sâm Ngọc Linh Việt Nam thông qua công nghệ sinh học.

CHƯƠNG 4. KẾT LUẬN VÀ KIẾN NGHỊ

4.1 KẾT LUẬN

- Các hệ phiên mã của các mẫu mô thân rễ và lá sâm Ngọc Linh 4 năm tuổi được lắp ráp từ 429.930.834 các đoạn đọc, tương ứng với 43.228.319.306 base sau khi lọc chất lượng. Hệ phiên mã của các mẫu mô khác nhau về số lượng gen, số lượng bản phiên mã và số lượng base lắp ráp.

- Các bộ dữ liệu hệ phiên mã được phân nhóm thành các unigene, tìm kiếm, dự đoán ORF và ước lượng độ phong phú. Đối với hệ phiên mã sâm Ngọc Linh 4 năm tuổi, khoảng 23,22-37,01% các unigene được dự đoán là ORF. Tổng 66,92-72,76% các đoạn đọc của hệ phiên mã mẫu mô lá và thân rễ sâm Ngọc Linh được lắp ráp vào hệ phiên mã tham chiếu.

- Các unigene được chú giải với tỉ lệ đạt từ 51,72-64,81%, sử dụng các cơ sở dữ liệu GO, EggNOG, NT, NR, KEGG, UniProt và Pfam. KEGG, EggNOG và NR có tỉ lệ chú giải thành công cao nhất, tiếp đó là GO, UniProt, Pfam và thấp nhất là NT. Với GO, các unigene được phân chia vào 3 nhóm chức năng lớn và 52 nhóm nhỏ. Với KEGG, các unigene được chia vào 211 con đường chuyển hóa. Với NR, các trình tự gen sâm Ngọc Linh tương đồng với trình tự của 534 loài sinh vật.

- Các unigene thuộc các hệ phiên mã mô lá và thân rễ sâm có sự tương đồng về tỉ lệ giữa các nhóm chức năng. Các đoạn transcript mang ưu thế ở các hệ phiên mã có liên quan đến các protein chống stress, chống chịu bệnh, già hóa hay sinh trưởng.

4.2 KIẾN NGHỊ

Dữ liệu thu được từ nghiên cứu này là cơ sở thực tiễn để tiếp tục mở rộng các nghiên cứu xây dựng dữ liệu toàn diện về hệ phiên mã của sâm Ngọc

Linh ở các độ tuổi, góp phần cung cấp thông tin đầy đủ và hữu ích về hệ phiên mã; xa hơn là các gen chức năng liên quan đến các tính trạng quan trọng, phục vụ công tác đánh giá, so sánh nguồn gen quý của sâm Ngọc Linh ở các lứa tuổi khác nhau; hỗ trợ bảo tồn, chọn tạo, quản lý và nhân giống sâm Ngọc Linh bằng công nghệ sinh học.

TÀI LIỆU THAM KHẢO

1. Baeg IH, So SH (2013). The world ginseng market and the ginseng (Korea). *J Ginseng Res* 37(1): 1.
2. Yang MS, Wu MY (2016). Chinese ginseng. *Nutraceuticals* 50: 693-705.
3. Jung J, Lee NK, Paik HD (2017). Bioconversion, health benefits, and application of ginseng and red ginseng in dairy products. *Food Sci Biotechnol* 26: 1155-1168.
4. Patel S, Rauf A (2017). Adaptogenic herb ginseng (*Panax*) as medical food: status quo and future prospects. *Biomed Pharmacother* 85: 120-127.
5. Shin BK, Kwon SW, Park JH (2015). Chemical diversity of ginseng saponins from *Panax ginseng*. *J Ginseng Res* 39(4): 287-298.
6. Sách Đỏ Việt Nam. Phần II. Thực vật (2007). *Nhà xuất bản Khoa học tự nhiên và Công nghệ, Hà Nội.*
7. Ha TD, Grushvitzky IV (1985). A new species of the genus *Panax* (Araliaceae) from Vietnam. *Bot Zhurn* 70: 519-522.
8. Nguyễn Tập (2005). Các loài thuộc chi *Panax* ở Việt Nam. *Tạp chí Dược liệu* 10(3): 71-76.
9. Trần Ngọc Lân, Nguyễn Tiến Dũng, Nguyễn Thị Thu, Lê Thị Thu Hiền, Ngô Hoàng Linh, Nguyễn Đức Nam, Trần Quốc Thành, Hoàng Nghĩa Nhạc, Phùng Văn Hào (2016). Kết quả nghiên cứu về loài sâm Puxailaileng ở vùng núi cao tỉnh Nghệ An. *Tạp chí Khoa học-Công nghệ Nghệ An* 12: 7-11.
10. Lương Đức Toàn (2018). Tính chất, chất lượng đặc thù của sâm củ và điều kiện tự nhiên vùng trồng sâm Ngọc Linh tỉnh Kon Tum. Kỷ yếu Hội nghị Đầu tư và phát triển sâm Ngọc Linh Kon Tum và các dược liệu

khác. Bộ Y tế, UBND Tỉnh Kon Tum, Bộ Nông nghiệp và Phát triển Nông thôn. Kon Tum.

11. Shendure J, Ji H (2008). Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135-1145.
12. Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46.
13. 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.
14. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.
15. 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.
16. Schuster SC (2008). Next-generation sequencing transforms today's biology. *Nat Methods* 5(1): 16-18.
17. Rusk N (2011). Torrents of sequence. *Nat Methods* 8: 44.
18. 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.
19. Pettersson E, Lundeberg J, Ahmadian A (2009). Generations of sequencing technologies. *Genomics* 93 (2): 105-111.
20. Schadt EE, Turner S, Kasarskis A (2010). Window into third-generation sequencing. *Hum Mol Genet* 19 (R2): R227-40.

21. 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.

22. 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* 5: S5.

23. Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323.

24. Cao H, Nuruzzaman M, Xiu H, Huang J, Wu K, Chen X, Li J, Wang L, Jeong JH, Park SJ, Yang F, Luo J, Luo Z (2015). Transcriptome analysis of methyl jasmonate-elicited *Panax ginseng* adventitious roots to discover putative ginsenoside biosynthesis and transport genes. *Int J Mol Sci* 16(2): 3035-3057.

25. Jayakodi M, Lee SC, Lee YS, Park HS, Kim NH, Jang W, Yang TJ (2015). Comprehensive analysis of *Panax ginseng* root transcriptomes. *BMC Plant Biol* 15(1): 1-12.

26. Rai A, Yamazaki M, Takahashi H, Nakamura M, Kojoma M, Suzuki H, Saito K (2016). RNA-seq transcriptome analysis of *Panax japonicus*, and its comparison with other *Panax* species to identify potential genes involved in the saponins biosynthesis. *Front Plant Sci* 7: 481.

27. Vu DD, Shah SNM, Pham MP, Bui VT, Nguyen MT, Nguyen TPT (2020). *De novo* assembly and transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* Ha et Grushv., including the development of EST-SSR markers for population genetics. *BMC Plant Biol* 20: 358.

28. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wignett S (2012). FastQC. A quality control tool for high throughput sequence data. Babraham Institute, Cambridge, United Kingdom.

29. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 29(7): 644-652.

31. Li W, Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13): 1658-1659.

32. Fu L, Niu B, Zhu Z, Wu S, Li W (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23): 3150-3152.

33. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8): 1494-512.

34. Buchfink B, Xie C, Huson DH. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1): 59-60.

35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215(3): 403-410.

36. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Pachter L (2010). Transcript assembly and abundance estimation

from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat Biotechnol* 28(5): 511.

37. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38: 1767-1771.
38. Gepstein S, Sabehi G, Carp MJ, Hajouj T, Nesher MFO, Yariv I, Bassani M (2003). Large-scale identification of leaf senescence-associated genes. *Plant J* 36(5): 629-642.
39. Dyer JM, Chapital DC, Kuan JCW, Mullen RT, Turner C, McKeon TA, Pepperman AB (2002). Molecular analysis of a bifunctional fatty acid conjugase/desaturase from tung. Implications for the evolution of plant fatty acid diversity. *Plant Physiol* 130(4): 2027-2038.