

Phân tích *in-silico* xác định và mô tả họ gen kháng *NBS-LRR* ở chuối *Musa acuminata*

Trần Đức Trung*, Tạ Hồng Linh

Viện Khoa học Nông nghiệp Việt Nam

Ngày nhận bài 1/9/2021; ngày chuyển phản biện 6/9/2021; ngày nhận phản biện 7/10/2021; ngày chấp nhận đăng 12/10/2021

Tóm tắt:

Bệnh hại do vi khuẩn, virus là mối đe dọa nghiêm trọng đối với ngành sản xuất chuối trên thế giới nói chung, Việt Nam nói riêng. Ngoài các biện pháp canh tác, việc khai thác và ứng dụng nguồn gen kháng - phần lớn thuộc họ gen *NBS-LRR* (Nucleotide binding site - Leucine rich repeat) trong chọn tạo, cải thiện giống chuối là phương thức bền vững để đối phó với dịch bệnh. Hệ gen loài chuối *Musa acuminata* được giải mã tạo tiền đề cho các nghiên cứu di truyền chuyên sâu, bao gồm xác định họ gen kháng *NBS-LRR*. Trong nghiên cứu này, các công cụ tin sinh học đã được sử dụng để xác định, mô tả cấu trúc và quá trình tiến hóa của 97 gen *NBS-LRR*. Ngoại trừ gen *Ma_NBS_083* thuộc phân họ *RNL*, các gen kháng còn lại ở chuối *M. acuminata* thuộc phân họ *CNL* và hầu hết là gen chứa một exon. 66 gen *NBS-LRR* (68%) được định vị trên 11 nhiễm sắc thể (NST), 38 gen trong số đó (39%) hình thành các cụm gen. Các gen *NBS-LRR* của chuối được chia thành 7 phân nhóm dựa trên sự đa dạng về trình tự và số lượng, cấu trúc các motif bảo thủ đặc trưng cho các vùng chức năng. Kết quả này cung cấp cơ sở khoa học cho các nghiên cứu đánh giá biểu hiện gen kháng và khai thác ứng dụng trong chọn tạo giống chuối.

Từ khóa: cây phát sinh, gen kháng *NBS-LRR*, MEME motif, *Musa acuminata*, phân tích *in-silico*, tiến hóa.

Chỉ số phân loại: 4.6

Giới thiệu

Trải qua quá trình tiến hóa với môi trường tác biến động liên tục từ các tác nhân gây bệnh (virus, vi khuẩn, nấm, sâu bệnh...), giới thực vật đã kiến tạo hệ thống miễn dịch hiệu quả. Tuy nhiên, do không sở hữu hệ miễn dịch thích nghi (adaptive immune system) hay các tế bào miễn dịch lưu động (mobile immune cell) như ở động vật nên sự đáp ứng miễn dịch ở thực vật hoàn toàn phụ thuộc vào hệ thống miễn dịch nội sinh (innate immune system), bao gồm miễn dịch bề mặt tế bào và nội bào. Đáp ứng miễn dịch bề mặt tế bào được khởi động dựa trên cơ chế nhận diện các phân tử ngoại lai (có nguồn gốc từ tác nhân gây bệnh với đặc tính bảo thủ về tiến hóa) bằng các thụ thể nhận diện khuôn mẫu (Pattern recognition receptors - PRRs). Trong khi đó, đáp ứng miễn dịch nội bào được kích hoạt thông qua cơ chế nhận diện (trực tiếp hoặc gián tiếp) các protein từ tác nhân gây bệnh (Avirulence - Avr) xâm nhập nội bào của protein kháng (R) [1, 2] theo cơ chế đặc hiệu đã được Flor mô tả qua mô hình “gen đối gen” [3].

Hầu hết, các protein R ở thực vật được mã hóa bởi họ gen *NBS-LRR* đặc trưng bởi các vùng gắn nucleotide và vùng lặp giàu leucine. Trải qua quá trình tiến hóa, họ gen *NBS-LRR* đã phân thành 3 phân họ khác biệt bởi vùng chức năng đặc trưng tại đầu N của protein được mã hóa, có thụ

thể tương tự là Toll/interleukin-1 (Toll/interleukin 1 receptor like - TIR), vùng xoắn kép (Coiled coil - CC) hoặc vùng kháng bệnh phấn trắng số 8 (Resistance to powdery mildew 8 - RPW8) [4]. Theo đó, các phân họ được định danh là *TIR-NBS-LRR* (*TNL*), *CC-NBS-LRR* (*CNL*) và *RPW8-NBS-LRR* (*RNL*). Các protein *TNL* và *CNL* có chức năng như là cảm biến phát hiện tác nhân gây bệnh thông qua sự bắt gắp hoặc kiểm soát trạng thái các protein tác hiệu của vùng *LRR*, qua đó kích hoạt chuỗi tín hiệu và khởi động đáp ứng miễn dịch của tế bào [5]. Trong khi đó, phân họ *RNL* (bao gồm hai nhóm được đặt tên theo gen chức năng là *ADRI* và *NRG1*) tham gia vận hành quá trình truyền tín hiệu thay vì phát hiện tác nhân gây bệnh [6]. Trong cấu trúc protein *NBS-LRR*, *NBS* là vùng chức năng đặc trưng, dài khoảng 300 axit amin (aa), được cấu trúc bởi 8 motif có mức độ bảo thủ biến thiên giữa các phân họ là P-loop, RNBS-A, Kinase2, RNBS-B, RNBS-C, GLPL, RNBS-D và MHDV. Vùng *LRR* ở đầu C đặc trưng bởi motif LxxLxLxxNxL(s/t)GxLP liên quan đến cấp độ biểu hiện tính kháng của gen *NBS-LRR* [7].

Kế thừa kết quả các nghiên cứu giải mã hệ gen và ứng dụng tin sinh học, họ gen kháng *NBS-LRR* đã được xác định ở nhiều loài thực vật [8-11]. Theo đó, số lượng gen *NBS-LRR* được xác định từ khoảng 50 gen ở đu đủ [12] đến 1.015 gen ở táo [13] và phân bố không đồng đều, thường được

*Tác giả liên hệ: Email: ductrung83@gmail.com

In-silico* genome-wide identification and characterisation of *NBS-LRR* gene family in banana *Musa acuminata

Duc Trung Tran*, Hong Linh Ta

Vietnam Academy of Agricultural Sciences

Received 1 September 2021; accepted 12 October 2021

Abstract:

Various common diseases caused by bacteria and viruses are increasing threats to the production of bananas worldwide. Besides the practical application of cultivation techniques, intensive exploitation of resistant genes, most of which belong to the *NBS-LRR* (Nucleotide-binding site - Leucine-rich repeat) gene family, in fortifying banana's disease resistance, has been considered as a sustainable approach. The recently published genome assembly of *Musa acuminata*, a species of the *Musa* genus native to Southern Asia, has facilitated the genome-wide identification and characterisation of *NBS-LRR* genes. In the present study, an *in-silico* approach was employed to identify and characterise 97 *NBS-LRR* genes and their evolution in *M. acuminata* genome. Except *Ma_NBS_083* belonging to the *RNL* sub-family, other identified genes are members of the sub-family *CNL* and most of them contain only one exon. Sixty-six (66) *NBS-LRR* genes (68%) were mapped on 11 banana chromosomes, among them 38 genes (39%) located in clusters. Phylogenetic analysis in combination with conserved motif identification classified the banana *NBS-LRR* genes into seven groups. This study provides novel insight into the *NBS-LRR* gene family in bananas as a potential resource for further research in functional evaluation and utilisation of resistant genes.

Keywords: evolution, *in-silico* analysis, MEME motifs, *Musa acuminata*, *NBS-LRR* genes, phylogenetics.

Classification number: 4.6

tìm thấy trong các cụm cluster trên các NST [14]. Việc duy trì số lượng lớn gen *NBS-LRR* như vậy phản ánh quá trình đấu tranh sinh học lâu dài ở mức độ phân tử thông qua quá trình tiến hóa giữa thực vật và các tác nhân gây bệnh. Hiểu rõ phân bố trong hệ gen và đường hướng tiến hóa của gen *NBS-LRR* sẽ góp phần quan trọng trong nghiên cứu cải thiện tính kháng cây trồng phục vụ nông nghiệp [7].

Chuối (*Musa* spp.) là cây trồng có giá trị kinh tế và phổ biến ở các vùng nhiệt đới, á nhiệt đới ẩm. Được thuần hóa từ hơn 7.000 năm trước ở khu vực Đông Nam Á, các giống chuối hiện nay là kết quả của quá trình lai khác loài (chủ yếu dựa trên *M. acuminata* và *M. balbisiana*) trong chi chuối *Musa*, được củng cố liên tục và trong thời gian dài bởi quá trình chọn lọc, nhân giống vô tính những dòng nhị bội, đa bội để đáp ứng nhu cầu của con người. Giống Cavendish chiếm hơn một nửa sản lượng chuối thế giới là dòng chuối tam bội vô tính được chọn lọc bởi quá trình này [15]. Chính vì đặc điểm trên, sâu bệnh hại (đặc biệt là các bệnh do virus, vi khuẩn, nấm) ngày càng trở nên thích nghi và dần phá vỡ tính kháng (vốn bị giới hạn do nền di truyền kém đa dạng của chuối) đã trở thành hiểm họa hiện hữu với ngành sản xuất chuối trên thế giới [16]. Thông qua khai thác cơ sở dữ liệu (CSDL) trình tự hệ gen của chuối *M. acuminata* [17], nghiên cứu này sẽ xác định và dự đoán cấu trúc, hướng tiến hóa của họ gen kháng *NBS-LRR*, qua đó góp phần tạo lập cơ sở di truyền phục vụ công tác cải thiện tính kháng bệnh ở chuối.

Vật liệu và phương pháp nghiên cứu

Vật liệu nghiên cứu

Trình tự hệ gen và hệ phiên mã của chuối *M. acuminata* được thu thập từ CSDL Banana genome hub (<https://banana-genome-hub.southgreen.fr/>). Hệ gen chuối được giải mã có kích thước 450,6 Mbp (tương đương 75,8% kích thước hệ gen được ước lượng bằng phương pháp Flow cytometry), bao gồm 11 NST giả định và 7.248 scaffold không được lắp ráp.

Họ protein NB-ARC đặc trưng cho các gen *NBS-LRR* thực vật thu nhập từ CSDL Pfam (định danh PF00931).

Phương pháp nghiên cứu

Họ protein NB-ARC được sử dụng làm trình tự tham vấn trong phân tích so sánh tBLASTn (e-value 1e-20) với CSDL hệ gen chuối (đã loại bỏ các scaffold có kích thước <1.000 bp) để xác định và trích xuất các Blast hit không trùng lặp. Bằng công cụ FGENSEH, các gen và protein tương ứng trên các Blast hit được dự đoán thông qua so sánh với CSDL gen của lớp thực vật một lá mầm. Cấu trúc của các protein dự đoán được sàng lọc thông qua so sánh với cơ sở dữ liệu

InterProScan. Chỉ những protein mang vùng chức năng NB-ARC và được mã hóa bởi gen nguyên vẹn mới được sử dụng cho các phân tích cấu trúc tiếp theo.

Cấu trúc của từng gen/protein *NBS-LRR* được chú giải bằng các công cụ: trình tự và cấu trúc exon-intron được xác định từ kết quả dự đoán gen FGENESH [18], vùng xoắn CC được xác định bằng MARCOIL [19], vùng lặp *LRR* được chú giải bằng LRRfinder2.0 [20], các motif bảo thủ của các protein *NBS-LRR* được xác định bằng MEME [21].

Tương đồng trình tự các gen và protein *NBS-LRR* của chuỗi được xác định bằng thuật toán MUSCLE, cây phân loại được xây dựng bằng phương pháp MLM dựa trên mô hình của Whelan và Goldman với giá trị tin cậy bootstrap 500 bằng phần mềm Geneious V11. Mô hình và chiều hướng tiến hóa của các protein *NBS-LRR* được xác định thông qua tỷ lệ giữa đột biến không đồng nghĩa với đột biến đồng nghĩa (Ka/Ks) và các gen *NBS-LRR* được định vị trên các NST chuỗi bằng công cụ TBTools [22].

Kết quả nghiên cứu

Số lượng và phân loại các gen mã hóa *NBS-LRR*

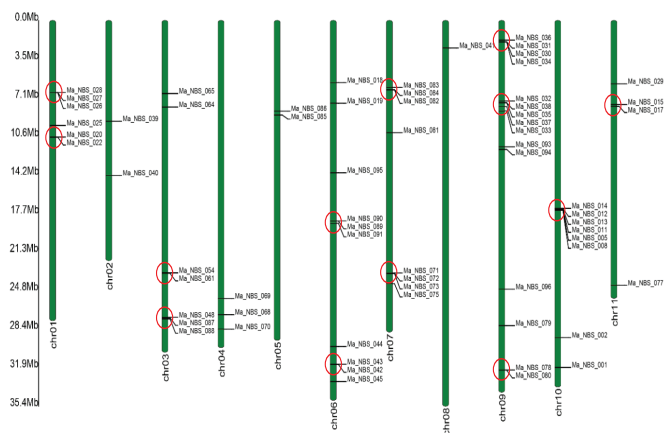
Bằng việc sử dụng họ protein NB-ARC làm trình tự tham vấn để sàng lọc đầy đủ các gen mã hóa *NBS-LRR* tiềm năng, 246 trình tự mã hóa protein NBS đã được phát hiện trên toàn hệ gen chuỗi *M. acuminata* V2.0. Tiếp theo đó, phân tích dự đoán gen trên các vùng trình tự này đã xác định được 97 gen không trùng lặp và nguyên vẹn mã hóa protein NBS (ký hiệu từ *Ma_NBS_001* đến *Ma_NBS_097*). Thông qua bước phân tích chú giải, trình tự và cấu trúc các vùng chức năng đặc trưng (vùng xoắn CC, vùng lặp *LRR* và các vùng chức năng khác) được xác định. Theo đó, 96 gen *NBS-LRR* ở *M. acuminata* thuộc phân họ *CNL* và được phân thành 4 nhóm, bao gồm 44 gen *CC-NBS-LRR*, 9 gen *CC-NBS*, 33 gen *NBS-LRR* và 10 gen *NBS*. Đáng chú ý, đã xác định được 1 gen mã hóa protein *RPW8-NBS-LRR* thuộc phân họ *RNL* (bảng 1).

Bảng 1. Thống kê số lượng gen *NBS-LRR* thuộc các phân họ và phân bố trên các NST trong hệ gen chuỗi *M. acuminata*.

Phân họ	Số lượng												
	Tổng	NST 1	NST 2	NST 3	NST 4	NST 5	NST 6	NST 7	NST 8	NST 9	NST 10	NST 11	Scaffolds
Phân họ <i>CNL</i>													
<i>CC-NBS-LRR</i>	44	2	1	5	2	2	3	4	-	5	2	2	16
<i>CC-NBS</i>	9	-	-	1	-	-	-	-	-	2	3	-	3
<i>NBS-LRR</i>	33	2	1	1	1	-	7	3	1	6	3	2	6
<i>NBS</i>	10	2	-	-	-	-	-	-	-	2	-	-	6
Phân họ <i>RNL</i>													
<i>RPW8-NBS-LRR</i>	1	-	-	-	-	-	-	1	-	-	-	-	-
Tổng số	97	6	2	7	3	2	10	8	1	15	8	4	31

Cấu trúc và sự phân bố của các gen *NBS-LRR* trên các NST

Các gen *NBS-LRR* trong hệ gen chuỗi *M. acuminata* có chiều dài trung bình 4.560 bp, biến thiên từ 690 (*Ma_NBS_028*) đến 13.566 bp (*Ma_NBS_094*), số lượng exon trung bình là 2,03. Có 61 (62,9%) gen *NBS-LRR* chỉ có 1 exon. Các gen *NBS-LRR* phân bố trên cả 11 NST giả định. Tuy nhiên, chỉ có 66/97 gen *NBS-LRR* (68%) được định vị trên các NST, 31 gen còn lại (32%) nằm trên các scaffold không được lắp ráp vào các NST. Các gen *NBS-LRR* không phân bố rải rác trên các NST, tập trung nhiều nhất trên các NST 9 (15 gen) và 6 (10 gen) hình thành các cụm gen. Căn cứ vào tiêu chí xác định các cụm gen của Schranz và cs (2006) [23] (khoảng cách giữa 2 gen *NBS-LRR* liên kề không quá 200 kb và không có hơn 8 gen khác) cho thấy, có 38 (39%) gen *NBS-LRR* được định vị trong 13 cụm gen và 35 gen *NBS-LRR* còn lại được xác định là gen đơn (hình 1).



Hình 1. Phân bố trên 11 NST giả định của 66 gen *NBS-LRR* (38 gen được định vị trong 13 cụm gen, vòng tròn đỏ).

Mô hình motif bảo thủ và quan hệ phát sinh giữa các protein *NBS-LRR*

20 motif bảo thủ có kích thước từ 11 (motif 17) đến 72 aa (motif 9) đã được xác định bằng công cụ MEME trên 97 protein *NBS-LRR* là chỉ dấu cho sự đa dạng về cấu trúc của các protein này. Theo vị trí trên trình tự protein, vùng xoắn

kép CC tại đầu N của 53 protein được đại diện bởi các motif 16, 5 và 15. Vùng NBS đặc trưng của toàn bộ protein *NBS-LRR* có kích thước từ 66 đến 351 aa được đại diện bởi các motif 1, 10, 6, 7, 17, 3, 18, 2, 13, 8 và 12. Trong khi đó, các motif 4, 20, 19, 14, 9 và 11 đại diện cho vùng lặp LRR tại đầu C của các protein. Thông qua so sánh với CSDL Pfam, xác định được 10/20 motif bảo thủ tương đồng cao với các vùng chức năng đặc trưng của họ gen *NBS-LRR* thực vật (bảng 2).

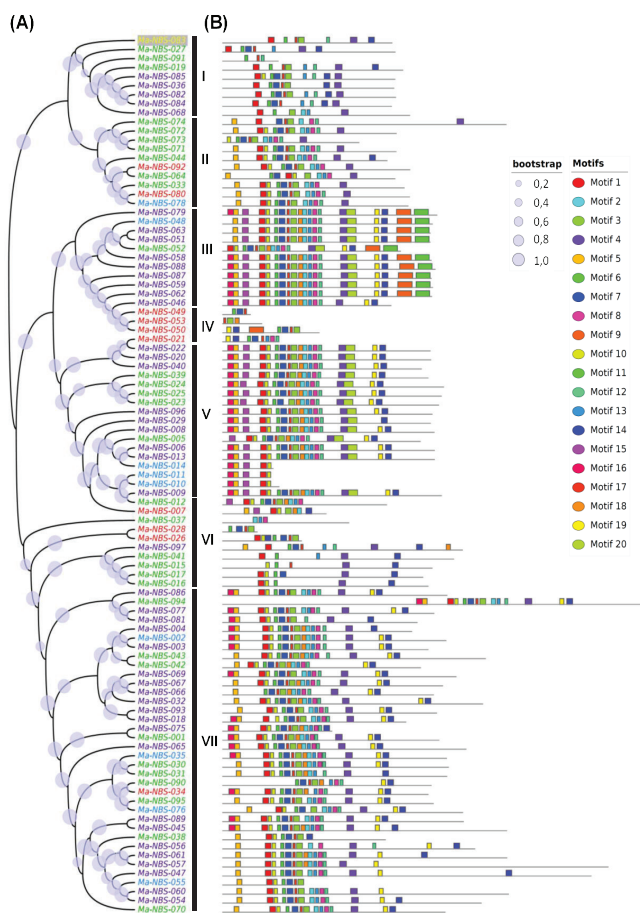
Nhằm đánh giá mô hình cấu trúc và quan hệ tiến hóa giữa các gen *NBS-LRR* ở chuỗi, cây phân loại dựa trên trình tự đầy đủ của 97 protein đã được xây dựng. Theo đó, hầu hết các protein *NBS-LRR* thuộc các nhóm định danh dựa trên chú giải các vùng chức năng thể hiện quan hệ tiến hóa

gần gũi với mức độ tin cậy cao (hình 2A). Kết hợp với mô hình cấu trúc dựa trên số lượng và tổ chức của các motif bảo thủ, các protein *NBS-LRR* cho thấy sự đa dạng về cấu trúc và được phân thành 7 phân nhóm chính (hình 2B). Các phân nhóm I (9 protein), IV (4 protein) và VI (10 protein) có đặc điểm cấu trúc không đặc trưng cho protein *NBS-LRR* do thiếu các motif bảo thủ cho vùng CC ở đầu N, trong khi các motif thành phần vùng NBS và LRR ở đầu C không đồng nhất. *Ma NBS 083* là gen mã hóa protein *RPW8-NBS-LRR* duy nhất của phân họ *RNL* tìm thấy trong hệ gen *M. acuminata* thuộc nhóm này. Trong khi đó, cấu trúc của 10 protein *NBS-LRR* trong phân nhóm II đã có mặt motif 5 đặc trưng cho vùng Rx ở đầu N, hầu hết các motif cấu trúc của vùng NBS và motif 4 của vùng LRR ở đầu C. Cấu trúc xoắn

Bảng 2. Thông tin 10 motif bảo thủ được xác định bởi công cụ MEME tương đồng với các vùng chức năng đã biết của protein R trên CSDL Pfam.

Tên motif và kích thước	MEME logo	Số protein mang motif	Vùng tương đồng trên CSDL Pfam		
			Họ protein	Mô tả	E-value
Motif 1 (29 aa)		86	NB-ARC	Vùng NBS trên các gen kháng <i>NBS-LRR</i>	4,6e-08
Motif 3 (29 aa)		83	NB-ARC	Vùng NBS trên các gen kháng <i>NBS-LRR</i>	5,9e-08
Motif 4 (33 aa)		79	LRR_4	Vùng chứa motif lặp LRR_4 (2 copies)	7,6e-08
Motif 5 (24 aa)		72	Rx_N	Vùng Rx đặc trưng ở đầu N của các protein kháng ở thực vật	4,7e-09
Motif 6 (15 aa)		90	NB-ARC	Vùng NBS trên các gen kháng <i>NBS-LRR</i>	0,00045
Motif 7 (29 aa)		84	NB-ARC	Vùng NBS trên các gen kháng <i>NBS-LRR</i>	0,0021
Motif 9 (72 aa)		11	LRR_4	Vùng chứa motif lặp LRR_4 (2 copies)	0,0042
Motif 15 (29 aa)		29	Matriline_ccoil	Cấu trúc xoắn kép CC đặc trưng	0,008
Motif 16 (29 aa)		39	Rx_N	Vùng Rx đặc trưng ở đầu N của các protein kháng ở thực vật	7,7e-09
Motif 20 (46 aa)		24	LRR_8	Leucine rich repeat	0,0043

CC không có mặt ở các protein trong phân nhóm này. Phân nhóm III bao gồm 11 protein có cấu trúc đầy đủ nhất với sự có mặt của 20 motif bảo thủ đại diện cho các vùng chức năng của protein *NBS-LRR*. 2 motif 9 và 11 bảo thủ cho vùng LRR là điểm đặc trưng phân biệt phân nhóm III với các protein còn lại trong hệ gen chuối. Cấu trúc của phân nhóm V (17 protein) tương đồng với phân nhóm III, ngoại trừ sự vắng mặt của các motif 9 và 11 đặc trưng cho vùng LRR đầu C. 3 protein thuộc phân nhóm CC-*NBS* (mặc dù thiếu 9/11 motif vùng NBS và các motif vùng LRR) được xếp vào phân nhóm này bởi mức độ tương đồng cao cả về trình tự và cấu trúc vùng CC, NBS với các protein còn lại trong nhóm. 36 protein của phân nhóm VII có cấu trúc đa dạng hơn với đặc điểm chung là thiếu motif 15 đặc trưng cho vùng xoắn kép CC và các motif 20, 9 và 11 đặc trưng cho vùng LRR ở đầu C (hình 2B).



Hình 2. Quan hệ phân loại, cấu trúc và các motif bảo thủ của các gen *NBS-LRR* ở chuối *M. acuminata*. (A) Cây phân loại dựa trên trình tự đầy đủ của 97 protein *NBS-LRR*. Giá trị bootstrap thể hiện mức độ tin cậy của từng nhánh trong cây phân loại (20-100% xác suất). Tên các protein được đánh dấu màu theo phân loại gen *NBS-LRR* dựa trên chú giải các vùng chức năng tại bảng 1; (B) Thành phần và tổ chức của các motif bảo thủ trong cấu trúc các protein *NBS-LRR*. Các motif được ký hiệu từ 1 đến 20 và phân biệt bởi các màu sắc khác nhau. Các protein được biểu thị tỷ lệ với chiều dài thực tế.

Gen lặp và mô hình tiền hóa của các gen *NBS-LRR*

Quá trình phân tán của một họ gen trong hệ gen nhờ các sự kiện sao chép diễn ra trên quy mô toàn hệ gen hoặc một vùng cụ thể trên NST. Trong hệ gen chuối, thông qua so sánh tương đồng trình tự đã phát hiện 28 cặp gen *NBS-LRR* lặp có mức độ tương đồng >80%. Trong số này, 20 cặp được xác định là gen lặp liên tiếp (TD) dựa vào tiêu chí khoảng cách giữa 2 gen dưới 200 kb [23]. 8 cặp gen còn lại có khoảng cách trên 200 kb hoặc nằm trên các NST khác nhau nên được phân loại là gen lặp phân tán (SD) (bảng 3). Các cặp TD phân bố không đồng đều trên các NST của chuối, bao gồm các nhóm lặp của 2, 3 hoặc 5 gen *NBS-LRR* trên các NST số 1, 3, 7, 9, 10 và 11. Trong khi đó, các SD được phát hiện trên NST số 9 (bao gồm 2 gen) và nhóm 4 gen trên NST số 10 và 11 (hình 3).

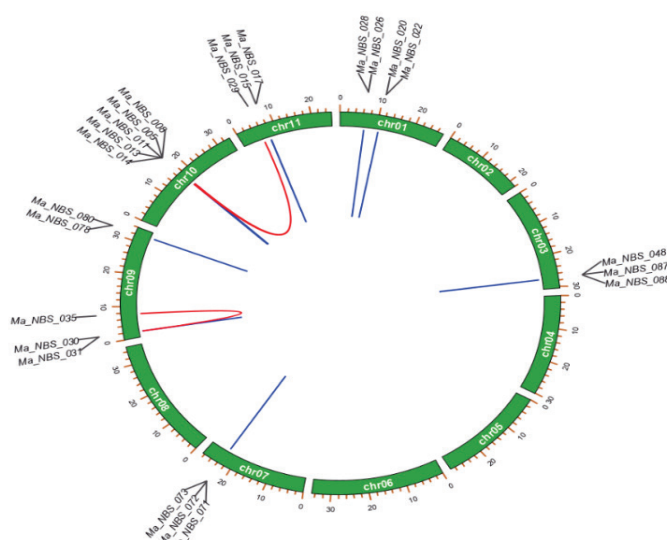
Bảng 3. Mô hình các cặp gen *NBS-LRR* lặp và chiều hướng chọn lọc trong quá trình tiến hóa của hệ gen chuối.

TT	Cặp gen	Mô hình	Ka/Ks	Mya
1	<i>Ma_NBS_028</i> <i>Ma_NBS_026</i>	TD	0,46	9,02
2	<i>Ma_NBS_020</i> <i>Ma_NBS_022</i>	TD	0,48	3,95
3	<i>Ma_NBS_048</i> <i>Ma_NBS_087</i>	TD	0,80	4,68
4	<i>Ma_NBS_048</i> <i>Ma_NBS_088</i>	TD	0,72	4,94
5	<i>Ma_NBS_088</i> <i>Ma_NBS_087</i>	TD	0,67	4,85
6	<i>Ma_NBS_071</i> <i>Ma_NBS_072</i>	TD	0,72	2,98
7	<i>Ma_NBS_073</i> <i>Ma_NBS_071</i>	TD	1,01	2,17
8	<i>Ma_NBS_073</i> <i>Ma_NBS_072</i>	TD	1,16	1,76
9	<i>Ma_NBS_031</i> <i>Ma_NBS_035</i>	SD	0,23	35,59
10	<i>Ma_NBS_030</i> <i>Ma_NBS_031</i>	TD	0,23	36,23
11	<i>Ma_NBS_030</i> <i>Ma_NBS_035</i>	SD	0,29	32,33
12	<i>Ma_NBS_078</i> <i>Ma_NBS_080</i>	TD	0,79	4,58
13	<i>Ma_NBS_014</i> <i>Ma_NBS_005</i>	TD	0,97	2,27
14	<i>Ma_NBS_014</i> <i>Ma_NBS_008</i>	TD	0,31	14,61
15	<i>Ma_NBS_014</i> <i>Ma_NBS_005</i>	TD	0,97	2,27
16	<i>Ma_NBS_014</i> <i>Ma_NBS_008</i>	TD	0,31	14,61
17	<i>Ma_NBS_014</i> <i>Ma_NBS_011</i>	TD	0,27	20,07
18	<i>Ma_NBS_014</i> <i>Ma_NBS_013</i>	TD	0,71	2,17
19	<i>Ma_NBS_014</i> <i>Ma_NBS_029</i>	SD	0,29	27,83
20	<i>Ma_NBS_011</i> <i>Ma_NBS_005</i>	TD	0,26	20,10
21	<i>Ma_NBS_011</i> <i>Ma_NBS_008</i>	TD	0,36	5,38
22	<i>Ma_NBS_005</i> <i>Ma_NBS_008</i>	TD	0,42	5,07
23	<i>Ma_NBS_005</i> <i>Ma_NBS_013</i>	TD	0,44	3,80
24	<i>Ma_NBS_008</i> <i>Ma_NBS_013</i>	TD	0,43	5,56
25	<i>Ma_NBS_029</i> <i>Ma_NBS_005</i>	SD	0,27	29,36
26	<i>Ma_NBS_029</i> <i>Ma_NBS_008</i>	SD	0,31	27,03
27	<i>Ma_NBS_029</i> <i>Ma_NBS_011</i>	SD	0,27	27,77
28	<i>Ma_NBS_015</i> <i>Ma_NBS_017</i>	TD	0,60	4,39

Mya: triệu năm trước.

Các đột biến liên quan đến một hay một vài nucleotide trên trình tự một gen có thể tạo ra sự thay đổi (đột biến không đồng nghĩa) hoặc không làm thay đổi (đột biến đồng nghĩa) trình tự protein mà gen đó mã hóa. Tỷ lệ *Ka/Ks* của một gen mã hóa protein phản ánh mức độ và chiều hướng của áp lực chọn lọc mà gen đó đang trải qua trong hệ gen. Tỷ lệ *Ka/Ks* < 1 cho thấy một cặp gen đang trải qua quá trình thanh lọc nhằm

loại bỏ những đột biến, trong khi đó $Ka/Ks > 1$ là chỉ dấu của quá trình chọn lọc thích nghi của đột biến [24]. Kết quả xác định chiều hướng chọn lọc của các cặp gen lặp cho thấy, 26 cặp gen *NBS-LRR* trong hệ gen chuối đang trải qua quá trình thanh lọc ($Ka/Ks < 1$) ở các mức độ khác nhau. Thời gian phân kỳ trong hệ gen của cặp gen lặp này ước tính từ 2,17 đến 36,23 triệu năm trước. Trong khi đó, chỉ có 2 cặp gen (*Ma_NBS_073* và *Ma_NBS_071*, *Ma_NBS_073* và *Ma_NBS_072*) đang trải qua quá trình chọn lọc thích nghi ($Ka/Ks > 1$). Thời gian phát sinh của hai cặp gen lặp này khá gần so với các cặp gen còn lại, ước tính khoảng 2 triệu năm trước (bảng 3).



Hình 3. Phân bố của các nhóm gen *NBS-LRR* lặp trong hệ gen *M. acuminata*. Đường nối màu xanh biểu thị TD, đường nối màu đỏ thể hiện SD. Vị trí tương đối của các gen có thể được xác định thông qua ký hiệu và kích thước của các NST.

Bàn luận

M. acuminata là 1 trong 2 loài thuộc chi *Musa* (cùng với *M. balbisiana*) và là tổ tiên của các giống chuối lai lưỡng bội, đa bội phổ biến trong sản xuất hiện nay [25]. Việc phân tích hệ gen *M. acuminata* không chỉ cung cấp luận cứ cho quá trình phát sinh, tiến hóa của các loài thuộc chi *Musa* mà quan trọng hơn là tạo cơ sở khoa học phục vụ cho quá trình cải tạo các giống chuối hiện nay. *NBS-LRR* là họ gen quan trọng quy định tính kháng sâu bệnh hại và đã được xác định ở nhiều loài thực vật thông qua các cách tiếp cận khác nhau [12, 13, 26, 27]. Trong nghiên cứu này, số lượng, cấu trúc, sự phân bố và mô hình tiến hóa của họ gen *NBS-LRR* đã được mô tả trên quy mô toàn hệ gen (phiên bản V2.0) của loài chuối *M. acuminata* [17].

Tổng số 97 gen mã hóa protein *NBS-LRR* đã được xác định trên hệ gen chuối thông qua kết hợp công cụ so sánh trình tự tBLASTn nhằm rà soát các vùng gen tương đồng cao và dự đoán gen FGENESH để xác định chính xác các gen hoàn chỉnh (với trình tự từ vị trí TATA-box của promoter đến vùng PolyA và không chứa các khoảng trống) [28]. Số lượng gen *NBS-LRR* hoàn chỉnh trong hệ gen chuối được xác định đầy đủ hơn so với các công bố trước đây của Martin và cs (2016) [17] (90

gen mã hóa *NBS-LRR*) khi áp dụng thuật toán chú giải hệ gen; Chang và cs (2020) [29] (94 gen mã hóa *NBS-LRR*) dựa trên sàng lọc CSDL chú giải protein của hệ gen *M. acuminata* v2.0. Theo kết quả phân tích mô tả cấu trúc, hầu hết các gen *NBS-LRR* ở chuối thuộc phân họ *CNL* tương tự với thành phần gen kháng ở các cây một lá mầm khác [4]. Đáng chú ý, đã xác định được gen *Ma_NBS-083* duy nhất thuộc nhóm *ADR1* của phân họ *RNL* chưa được phát hiện trong các nghiên cứu trước (chỉ bao gồm các gen thuộc phân họ *CNL*) [17, 29]. Bên cạnh đó, chỉ có 66/97 gen *NBS-LRR* hoàn chỉnh, ít hơn so với con số 100% gen theo nghiên cứu của Chang và cs (2020) [29] định vị trên 11 NST của hệ gen *M. acuminata* V2.0. Sự khác biệt trên do nghiên cứu này sử dụng dữ liệu trình tự hệ gen đầy đủ (bao gồm 11 NST giả định cùng các scaffold chưa được lắp ráp hoàn thiện) và áp dụng ngưỡng chính xác cao (e-value 1e-20) cho các phân tích so sánh trình tự, do đó đã loại bỏ nhiều gen ứng viên có mức độ tương đồng trình tự thấp. Trong khi đó, nghiên cứu của Chang và cs (2020) [29] chỉ khai thác thông tin dự đoán protein đã được chú giải trên các NST giả định hệ gen *M. acuminata* V2.0. Như vậy, nghiên cứu này đã bổ sung thông tin về số lượng cũng như thành phần của họ gen *NBS-LRR* ở chuối *M. acuminata*.

Các protein *NBS-LRR* ở *M. acuminata* có sự đa dạng về trình tự và cấu trúc. Phân tích phân loại kết hợp với số lượng và tổ chức của 20 motif bảo thủ được xác định bởi công cụ MEME đã phân 97 protein *NBS-LRR* thành 7 phân nhóm. Vùng NBS là dấu hiệu đặc trưng cho toàn bộ protein của 7 phân nhóm. Tuy nhiên, so với các phân nhóm còn lại, các protein của phân nhóm I, IV và VI có sự biến động rất rõ ràng về số lượng và cấu trúc các motif bảo thủ của vùng NBS, nhất là các motif bảo thủ tương đồng với vùng NBS tham chiếu trong Pfam (hình 2 và bảng 2). Nếu coi vùng NBS là tiêu chuẩn để dự đoán chức năng của các gen *NBS-LRR* thì điều này là chỉ dấu cho thấy các protein thuộc phân nhóm I, IV và VI ở trong hệ gen chuối hoặc là kết quả của sự thanh lọc các gen dư thừa, hoặc thể hiện sự chuyên biệt hóa cấu trúc của các gen này trong quá trình tiến hóa hệ gen chuối [8]. Các nghiên cứu sâu hơn về biểu hiện gen sẽ giúp làm sáng tỏ giả thuyết này.

Quá trình nhân lên và phân tán của họ gen kháng *NBS-LRR* trong hệ gen không chỉ giúp duy trì số lượng gen kháng cho quá trình biến đổi liên tục đáp ứng mô hình “gen đối gen”, đảm bảo tính kháng của thực vật trước sự tác động không ngừng của tác nhân gây bệnh, mà còn là dấu hiệu quan trọng của sự vận động, tiến hóa của hệ gen thực vật. Thông qua đánh giá mức độ tương đồng và tính toán tỷ số Ka/Ks của các gen *NBS-LRR*, thời điểm phát sinh và phân tán các cặp gen *NBS-LRR* lặp xa nhất có thể ước tính là khoảng 36,23 triệu năm trước, phù hợp với thời điểm tiến hóa phát sinh của chi *Musa* (~50 triệu năm trước) [30]. Bên cạnh đó, 16/26 cặp gen lặp có thời gian phát sinh gần hơn, ước tính trong khoảng 5 triệu năm trước và hầu hết trong quá trình thanh lọc (bảng 3). Đây là minh chứng cho quá trình chọn lọc diễn ra liên tục trong thời gian dài nhằm duy trì tính bảo thủ cấu trúc của gen *NBS-LRR* trong tiến hóa hệ gen chuối. Bên cạnh đó, việc chỉ có 2 cặp gen lặp có thời gian phân kỳ gần đây và đang trong quá trình chọn lọc thích nghi có thể

là chỉ dấu cho số ít những đột biến mới phát sinh và có lợi cho cây chuối. Kết quả này cũng đồng thời phản ánh thực tế vốn gen cây chuối bị thu hẹp do lịch sử hàng nghìn năm thuần hóa và chọn lọc của con người.

Có đến 31/97 gen *NBS-LRR* chưa được định vị trên các NST do thông tin hệ gen *M. acuminata* V2.0 chưa hoàn chỉnh. Do đó, kết quả xác định gen, mô tả cấu trúc protein và đánh giá chiều hướng chọn lọc của các gen *NBS-LRR* trong nghiên cứu này chưa phản ánh đầy đủ nhất bức tranh tiến hóa của họ gen này trong hệ gen chuối. Việc hoàn thiện hệ gen lắp ráp của chuối (*M. acuminata* V4.0) trong thời gian tới [31] sẽ tạo cơ hội cho các phân tích đầy đủ hơn về họ gen này. Mặc dù vậy, các kết quả đạt được đã góp phần làm sáng tỏ những khía cạnh liên quan đến họ gen *NBS-LRR*, tạo nền tảng cho các nghiên cứu khai thác ứng dụng trong chọn tạo và cải thiện tính kháng sâu bệnh ở chuối.

KẾT LUẬN

Trong nghiên cứu này, 97 gen *NBS-LRR* của chuối *M. acuminata* đã được xác định và chia vào 7 phân nhóm đa dạng về số lượng và cấu trúc các motif bảo thủ đại diện cho các vùng chức năng đặc trưng của họ gen *NBS-LRR*. Mô hình tiến hóa các gen *NBS-LRR* cho thấy, quá trình thanh lọc các gen lặp đóng vai trò chủ đạo trong tiến hóa hệ gen chuối, góp phần duy trì sự bảo thủ cấu trúc của họ gen *NBS-LRR*. Kết quả của nghiên cứu này là nguồn thông tin hữu ích cho các nghiên cứu đánh giá biểu hiện gen kháng, phát triển chỉ thị phân tử phục vụ chọn tạo và cải thiện tính kháng sâu bệnh hại chuối bằng chỉ thị phân tử.

LỜI CẢM ƠN

Nghiên cứu này được thực hiện trong khuôn khổ đề tài “Nghiên cứu xác định gen tiềm năng kháng bệnh hại trên cây chuối (*Musa* spp.) thông qua phân tích tin sinh học” do Viện Khoa học và Nông nghiệp Việt Nam tài trợ. Các tác giả xin chân thành cảm ơn.

TÀI LIỆU THAM KHẢO

- [1] A.R. Bentham, et al. (2020), “A molecular roadmap to the plant immune system”, *Journal of Biological Chemistry*, **295**(44), pp.14916-14935.
- [2] A. Corrion, B. Day (2015), *Pathogen Resistance Signalling in Plants*, John Wiley & Sons.
- [3] H.H. Flor (1942), “Inheritance of pathogenicity in *Melampsora lini*”, *Phytopathology*, **32**, pp.653-669.
- [4] Z.Q. Shao, et al. (2019), “Revisiting the origin of plant *NBS-LRR* genes”, *Trends in Plant Science*, **24**(1), pp.9-12.
- [5] E.J. Andersen, et al. (2018), “Disease resistance mechanisms in plants”, *Genes*, **9**(7), DOI: 10.3390/genes9070339.
- [6] Z. Wu, et al. (2019), “Differential regulation of TNL-mediated immune signaling by redundant helper CNLs”, *New Phytologist*, **222**(2), pp.938-953.
- [7] V. Pandolfi, et al. (2017), “Resistance (R) genes: applications and prospects for plant biotechnology and breeding”, *Curr. Protein Pept. Sci.*, **18**(4), pp.323-334.
- [8] Z.Q. Shao, et al. (2016), “Large-scale analyses of angiosperm nucleotide binding site - Leucine rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns”, *Plant Physiology*, **170**(4), pp.2095-2109.
- [9] F. Zheng, et al. (2016), “Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family”, *BMC Genomics*, **17**, DOI 10.1186/s12864-016-2736-9.

- [10] H. Wei, et al. (2020), “Genomic organization and comparative phylogenetic analysis of *NBS-LRR* resistance gene family in *Solanum pimpinellifolium* and *Arabidopsis thaliana*”, *Evolutionary Bioinformatics*, **16**, pp.1-13.
- [11] P. Li, et al. (2016), “RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants”, *BMC Genomics*, **17**(1), DOI 10.1186/s12864-016-3197-x.
- [12] B.W. Porter, et al. (2009), “Genome-wide analysis of *Carica papaya* reveals a small NBS resistance gene family”, *Molecular Genetics and Genomics*, **281**(6), pp.609-626.
- [13] P. Arya, et al. (2014), “Genome-wide identification and expression analysis of NBS - encoding genes in *Malus x Domestica* and expansion of NBS genes family in Rosaceae”, *PLOS ONE*, **9**(9), DOI: 10.1371/journal.pone.0107987.
- [14] H. Mizuno, et al. (2020), “Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice”, *Sci. Rep.*, **10**(1), DOI: 10.1038/s41598-020-57729-w.
- [15] M. Busche, et al. (2020), “Genome sequencing of *Musa acuminata* dwarf Cavendish reveals a duplication of a large segment of chromosome 2”, *G3: Genes - Genomes - Genetics*, **10**(1), pp.37-42.
- [16] X. Perrier, et al. (2011), “Multidisciplinary perspectives on banana (*Musa* spp.) domestication”, *Proceedings of the National Academy of Sciences*, **108**(28), pp.11311-11318.
- [17] G. Martin, et al. (2016), “Improvement of the banana “*Musa acuminata*” reference sequence using NGS data and semi-automated bioinformatics methods”, *BMC Genomics*, **17**(1), DOI 10.1186/s12864-016-2579-4.
- [18] V. Solovyev, et al. (2006), “Automatic annotation of eukaryotic genes, pseudogenes and promoters”, *Genome Biology*, **7**(1), DOI:10.1186/gb-2006-7-s1-s10.
- [19] M. Delorenzi, T. Speed (2002), “An HMM model for coiled-coil domains and a comparison with PSSM-based predictions”, *Bioinformatics*, **18**(4), pp.617-625.
- [20] V. Offord, D. Werling (2013), “LRRfinder2.0: a webserver for the prediction of leucine-rich repeats”, *Innate Immun.*, **19**(4), pp.398-402.
- [21] T.L. Bailey, et al. (2009), “MEME suite: tools for motif discovery and searching”, *Nucleic Acids Research*, **37**, pp.W202-W208.
- [22] C. Chen, et al. (2020), “TBtools: an integrative toolkit developed for interactive analyses of big biological data”, *Molecular Plant*, **13**(8), pp.1194-1202.
- [23] M.E. Schranz, et al. (2006), “The ABC’s of comparative genomics in the Brassicaceae: building blocks of crucifer genomes”, *Trends Plant Sci.*, **11**(11), pp.535-542.
- [24] Z. Yang, J.P. Bielawski (2020), “Statistical methods for detecting molecular adaptation”, *Trends Ecol. Evol.*, **15**(12), pp.496-503.
- [25] Z. Wang, et al. (2019), “*Musa balbisiana* genome reveals subgenome evolution and functional divergence”, *Nature Plants*, **5**(8), pp.810-821.
- [26] T.P. Frazier, et al. (2016), “Identification, characterization, and gene expression analysis of nucleotide binding site (NB)-type resistance gene homologues in switchgrass”, *BMC Genomics*, **17**(1), DOI 10.1186/s12864-016-3201-5.
- [27] R. Lozano, et al. (2015), “Identification and distribution of the *NBS-LRR* gene family in the Cassava genome”, *BMC Genomics*, **16**, DOI: 10.1186/s12864-015-1554-9.
- [28] G.F. Ejigu, J. Jung (2020), “Review on the computational genome annotation of sequences obtained by next generation sequencing”, *Biology*, **9**(9), DOI: 10.3390/biology9090295.
- [29] W. Chang, et al. (2020), “*NBS-LRR* gene family in banana (*Musa acuminata*): genome-wide identification and responses to *Fusarium oxysporum* f. sp. cubense race 1 and tropical race 4”, *European Journal of Plant Pathology*, **157**, pp.549-563.
- [30] P. Christelova, et al. (2011), “A multi gene sequence-based phylogeny of the Musaceae (banana) family”, *BMC Evol. Biol.*, **11**, DOI: 10.1186/1471-2148-11-103.
- [31] C. Belser, et al. (2021), “Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing”, *Commun. Biol.*, **4**, DOI: 10.1038/s42003-021-02559-3.