THÀNH TỰU MỚI TRONG GIẢI MÃ HỆ GEN THỰC VẬT

Chu Đức Hà[1], Nguyễn Thị Duyên[2], Phạm Phương Thu[2], La Việt Hồng[2], Lê Huy Hàm[1, 3], Phạm Xuân Hội[1], Trần Phan Lam Sơn[4]

[1]Viện Di truyền Nông nghiệp, VAAS

[2]Trường Đại học Sư phạm Hà Nội 2

[3]Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

[4]Trung tâm Khoa học Tài nguyên Bền vững RIKEN, Nhật Bản

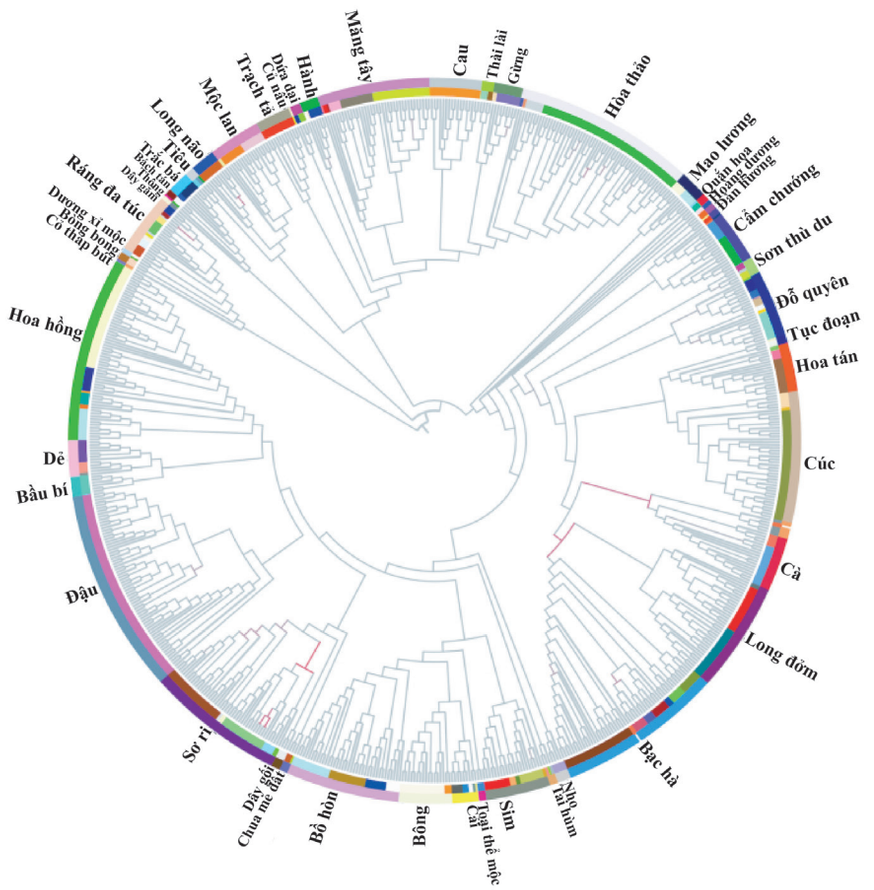
Mới đây, nỗ lực của các nhà khoa học đã được ghi nhận trong việc giải mã thành công trình tự hệ gen của 689 loài thực vật bậc cao ở Trung Quốc - một trong những nghiên cứu dữ liệu lớn đầu tiên trên thế giới được tiến hành trên đối tượng thực vật. Kết quả nghiên cứu là những tiền đề quan trọng cho việc nhận dạng các loài thực vật mới bằng công nghệ ADN mã vạch cũng như cung cấp những dữ liệu quan trọng về một số gen tiềm năng nhằm cải thiện tính di truyền ở cây trồng. Trong bài viết này, các tác giả tóm lược các kết quả chính của dự án giải mã hệ gen thực vật của các nhà khoa học Trung Quốc. từ đó đề xuất một số hướng nhằm khai thác tối đa những thành tựu này phục vụ nghiên cứu.

Thành công của công nghệ giải trình tự thế hệ mới đã hỗ trợ đắc lực cho giới khoa học trong việc khám phá vật chất di truyền của hầu hết các loài sinh vật, từ đó có thể tiếp cận gần hơn đến cơ chế tiến hóa của toàn bộ sinh giới (Earth BioGenome, https://www.earthbiogenome.org/). Mặt khác, thông tin di truyền của loài có thể cung cấp những dữ liệu quan trọng về một số gen tiềm năng nhằm cải thiện tính di truyền ở các loài sinh vật, bao gồm cây trồng [1]. Hiện nay, hơn 391.000 loài thực vật đã được phát hiện và ghi nhận trên trái đất, tuy nhiên chỉ có khoảng 350 loài, hầu hết là cây trồng cạn, cây mô hình và các loài hoang dại mới được giải mã hệ gen gần đây [2]. Trong nỗ lực nhằm làm sáng tỏ bức tranh về toàn bộ giới thực vật, các nhà khoa học Trung Quốc đã thực hiện dự án "Giải mã 10.000 hệ gen thực vật" (10,000 Plant Genomes Project - 10KP) [3]. Kết quả đầu tiên của dự án là đã thu thập và giải mã được cho 689 loài thực vật tại một khu vực diện tích rộng lớn trong Vườn Bách thảo Ruili (Trung Quốc). Đây là những tiền đề quan trọng cho việc nhận dạng các loài mới bằng công nghệ ADN mã vạch (DNA barcoding).

Những kết quả chính của dự án "Giải mã 10.000 hệ gen thực vật" ở Trung Quốc

Trong dự án này, hơn 1.000 mẫu lá cây, đại diện cho 689 loài thực vật bậc cao đã được thu thập tại Vườn Bách thảo Ruili (Vân Nam, Trung Quốc) trong chỉ giới địa lý từ 97°38'47" đến 98°05'57" Bắc, và từ 23°52'42" đến 24°09'20" Đông, với độ cao từ 738 đến 1.200 m so với mực nước biển. Toàn bộ công đoạn tách ADN tổng số và giải trình tự toàn hệ gen sau đó được thực hiện tại Viện Gen Bắc Kinh (Beijing Genomics Institute, https://www.bgi.com/global/). Đây được xem là trung tâm giải mã gen lớn nhất thế giới hiện nay [4], dẫn đầu trong lĩnh vực giải trình tự hệ gen động vật (bao gồm cả loài người) [5], thực vật [3, 6] và vi sinh vật [7].

Các nhà khoa học Trung Quốc đã phân loại toàn bộ lượng mẫu thu được thành 137 họ và 47 bộ [2] (hình 1). Trong số đó, một số lượng lớn các mẫu được nhận dạng hình thái và xếp vào họ Đậu (Fabaceae): 71 loài, Hòa thảo (Poaceae): 45 loài và Cúc (Asteraceae): 37 loài [2]. Dựa trên kết quả giải trình tự hệ gen lục lạp, cây phân loại đã được thiết lập thành công dựa theo thuật toán



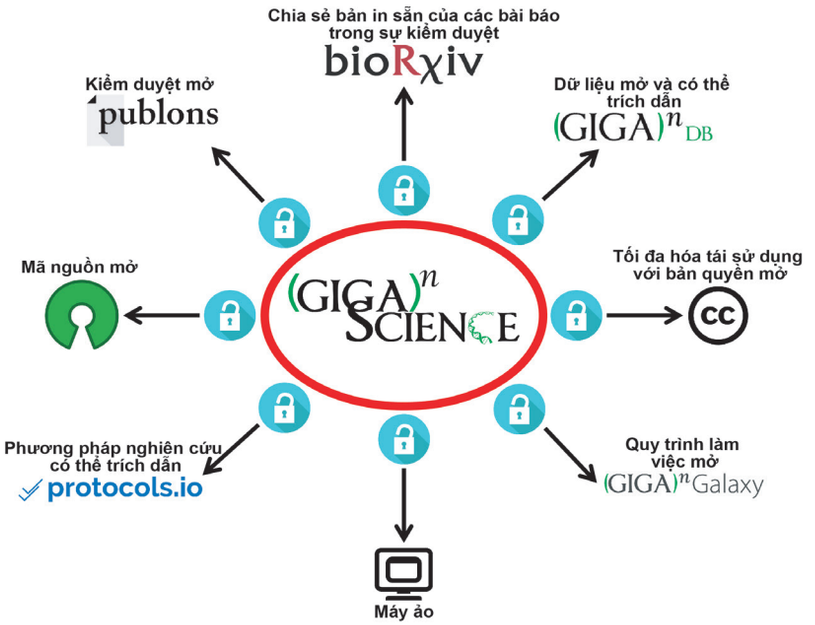
Hình 1. Cây phát sinh giữa các loài trong 47 bộ thu thập tại Vườn Bách thảo Ruili.

Maximum Likelihood, từ đó cho phép xác định và đưa ra giả thuyết về mức độ quan hệ gần gũi giữa các loài và giữa 47 bộ (hình 1). Ví dụ, một số nhánh chính có thể được ghi nhận trên cây phân loại như bộ Đậu (Fabales), Hoa hồng (Rosales), Hòa thảo (Poales) và Sơ ri (Malpighiales). Bên cạnh đó, hình 1 cũng cho thấy, bộ Đậu và bộ Dây gối (Celastrales) có quan hệ gần gũi với bộ Sơ ri hơn so với bộ Chua me đất (Oxalidales) với giá trị bootstrap = 100%. Ngoài ra, một nhóm gồm nhiều bộ thực vật một lá mầm được xếp cùng trong nhánh lớn, gồm bộ Hành (Liliales), Măng tây (Asparagales), Hòa thảo, Cau (Arecales), Thài lài (Commelinales), Gừng (Zingiberales), Củ nâu (Dioscoreales) và Dứa dại (Pandanales), với nhánh xuất hiện sớm nhất là Trạch tả (Alismatales), tương tự như ghi nhận trong nghiên cứu trước đây [8]. Mặt khác, mối quan hệ giữa một số bộ trong cây phát sinh vẫn chưa rõ ràng, như giữa bộ Long đởm (Gentianales), Bạc hà (Lamiales) và Cà (Solanales) (hình 1) [9, 10].

Một trong những khía cạnh được quan tâm trong dự án này là các kết quả về phân tích kích thước hệ gen của các loài. Việc tính toán kích thước hệ gen của toàn bộ 689 loài thực vật đã được so sánh với một số kết quả giải mã hệ gen của một số loài thực vật trước đó và đều thu được những sự đồng thuận. Họ có biến động về hệ gen lớn nhất là Hoàng đàn (*Cupressaceae*), với loài có hệ gen nhỏ nhất chỉ 0,18 Gb [cây Thông mụ - *Cunninghamia lanceolata* (Lamb.) Hook. var. *lanceolata*] và loài có kích thước hệ gen lớn nhất lên tới 10,26 Gb [cây Tùng bonsai - *Juniperus pingii* var. wilsonii (Rehder) Silba]. Bên cạnh đó, kích thước hệ gen lục lạp của các loài này dao động từ 113.621 đến 183.602 bp [2].

Đề xuất một số hướng khai thác dữ liệu phục vụ nghiên cứu

Toàn bộ thông tin, bao gồm dữ liệu trình tự thô, bản giải mã hệ gen lục lạp và hệ gen nhân của tất cả 689 loài thực vật được sắp xếp và lưu giữ trên cơ sở dữ liệu GigaDB của GigaScience (http://dx.doi.org/10.5524/100502). Cần phải nói thêm, GigaScience (chỉ số ảnh hưởng năm 2017 = 7,267) là một tạp chí mở (open access) tập trung vào các nghiên cứu về dữ liệu lớn (big data) trong khoa học sự sống và y sinh. Với mục đích cách mạng hóa trong việc xuất bản các bài báo khoa học, GigaScience cho phép công khai toàn bộ dữ liệu tin sinh để các nhà khoa học có thể khai thác và tái sử dụng thông tin theo từng mục đích nghiên cứu (hình 2). Có thể thấy rằng, với kho dữ liệu khổng lồ thu được từ dự án giải trình tự hệ gen thực vật, các nhà khoa học trên toàn thế giới có thể khai thác, xử lý và phân tích được rất nhiều vấn đề. Dưới đây là một số đề xuất của chúng tôi nhằm khai thác tối đa những dữ liệu này



Hình 2. Ưu điểm vượt trội của Tạp chí GigaScience.

phục vụ mục đích nghiên cứu.

Thứ nhất, các dữ liệu thô về trình tự toàn hệ gen có thể mở ra những phân tích về mặt tiến hóa của các gen mục tiêu cũng như cho phép chúng ta tìm hiểu những khía cạnh cụ thể của quá trình tiến hóa hệ gen ở thực vật, bao gồm cơ chế tiến hóa của các đoạn lặp, hiện tượng đa bội hóa và hiện tượng lặp toàn hệ gen.

Thứ hai, những dữ liệu giải trình tự này bước đầu có thể được sử dụng làm hệ tham chiếu cho việc khám phá hệ gen của các loài họ hàng trong tương lai cũng như tham khảo cho bản chú giải hệ gen tiếp theo của loài. Hơn nữa, đây còn là một trong những nghiên cứu dữ liệu lớn đầu tiên trên thế giới được tiến hành trên đối tượng thực vật, là một phần của dự án 10KP. Vì vậy, dự án còn cung cấp những kinh nghiệm quý báu về các phương pháp thu thập, lấy mẫu thực vật, phân tích và giải trình tự hệ gen cũng như quản lý dữ liệu loài trên một quy mô lớn. Đây là một tiền đề quan trọng cho dự án giải trình tự toàn bộ sinh giới (Earth BioGenome Project, https://www.earthbiogenome.org/) đang được tiến hành.

Thứ ba, các kết quả trên có thể được sử dụng để phát triển một phương pháp nhận dạng loài mới dựa trên những dữ liệu giải trình tự hoặc ảnh mô tả hình thái, đồng thời giải quyết mối quan hệ họ hàng giữa các loài dựa trên kết quả giải trình tự toàn hệ gen.

Cuối cùng, Việt Nam với thảm thực vật đa dạng cần phải xem xét một cách toàn diện về việc thu thập và giải trình tự toàn bộ loài đặc hữu, từ đó tạo điều kiện cho công tác bảo tồn, phục tráng cũng như lưu giữ những nguồn gen thực vật quý. Các phương pháp nghiên cứu (thu thập mẫu, tách chiết ADN, giải trình tự, phân tích dữ liệu và quản lý thông tin) cần được xem xét một cách toàn diện dựa trên những kinh nghiệm thu được từ dự án 10KP nói riêng, những dự án "dữ liệu lớn" trong sinh học nói chung

TÀI LIỆU THAM KHẢO

[1] E. Pennisi (2011), "Plant biology, green genomes", *Science*, **332(6036)**, pp.1372-1375.

[2] H. Liu, et al. (2019), "Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden", *GigaScience*, **8(4)**, pp.giz007.

[3] S. Cheng, et al. (2018), "10KP: A phylodiverse genome sequencing plan", *GigaScience*, **7(3)**, pp.1-9.

[4] A. McCarthy (2013), "BGI Americas: commercializing next-generation sequencing", *Chem. Biol.*, **20(6)**, pp.743-744.

[5] J. Huang, et al. (2017), "A reference human genome dataset of the BGISEQ-500 sequencer", *GigaScience*, **6(5)**, pp.1-9.

[6] X. He, J. Wang (2007), "Bgi-Ris V2", *Methods Mol. Biol.*, **406**, pp.275-299.

[7] D. Cyranoski (2012), "Chinese genomics giant BGI plots commercial path", *Nat. Biotechnol.*, **30(12)**, pp.1159-1160.

[8] M.W. Chase (2004), "Monocot relationships: an overview", *Am. J. Bot.*, **91(10)**, pp.1645-1655.

[9] K. Bremer, et al. (2001), "A phylogenetic analysis of 100+ genera and 50+ families of euasterids based on morphological and molecular data with notes on possible higher level morphological synapomorphies", *Plant Syst. Evol.*, **229(3-4)**, pp.137-169.

[10] N. Refulio-Rodriguez, R. Olmstead (2014), "Phylogeny of Lamiidae", *Am. J. Bot.*, **101(2)**, pp.287-299.