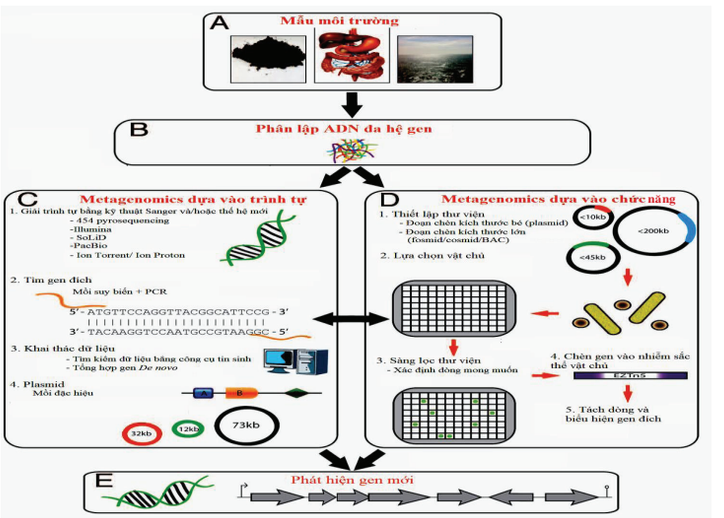
METAGENOMICS VÀ VIỆC KHAI THÁC TIỀM NĂNG ĐA DẠNG SINH HỌC NGUỒN GEN VI SINH VẬT CỦA VIỆT NAM

PGS.TS PHẠM CÔNG HOẠT
Bộ KH&CN
TS PHÙNG THU NGUYỆT
Viện Hàn lâm KH&CN Việt Nam
TS TRẦN NGỌC HÙNG
Bộ Nông nghiệp và Phát triển nông thôn

Metagenomics (nghiên cứu đa hệ gen) đã tạo nên những tiến bộ vượt bậc trong việc nghiên cứu về sinh thái học, tiến hóa và đa dạng vi sinh vật. Đây là định hướng nghiên cứu mới, quan trọng đã trở thành chiến lược trong phát triển kinh tế bền vững, an ninh quốc phòng, bảo vệ sức khỏe và môi trường của nhiều quốc gia trên thế giới. Ở Việt Nam, metagenomics bước đầu được triển khai để giải các bài toán có tính chất tổng thể, nhằm xây dựng cơ sở dữ liệu về đa dạng sinh học của các hệ vi sinh vật có giá trị.

Metagenomics

Metagenomics là phương pháp nghiên cứu (phân tích) đa hệ gen (metagenome) của tất cả các vi sinh vật thu nhận trực tiếp từ mẫu môi trường tự nhiên mà không thông qua nuôi cấy. Thực chất, metagenomics là thuật ngữ được sử dụng để mô tả một lĩnh vực nghiên cứu khoa học và các kỹ thuật cho phép phân tích toàn thể vi sinh vật sống trong bất kỳ môi trường tự nhiên nào. Metagenomics bắt nguồn từ ý tưởng tách dòng DNA trực tiếp từ các mẫu môi trường của Pace (Đại học Colorado - Hoa Kỳ) vào năm 1991. Năm 1995, Healy và cộng sự (Đại học Florida - Hoa Kỳ) đã xây dựng thành công một thư viện DNA đa hệ gen từ hỗn hợp các sinh vật trong cỏ khô và tìm được 14 dòng có hoạt tính cellulase từ thư viện. Sau thành công của Healy và cộng sự, rất nhiều thư viện DNA metagenome của vi sinh vật trong các môi trường sống đã được xây dựng nhằm khai thác gen cũng như nghiên cứu sự đa dạng vi sinh vật. Sau hơn 20 năm phát triển, Metagenomics đã trở thành một công cụ mạnh mẽ được sử dụng rất rộng rãi để nghiên cứu sự đa dạng của vi sinh vật.



Hình 1: các phương pháp tiếp cận chính của metagenomics để tìm gen mới

Lịch sử phát triển và ứng dụng metagenomics

Thuật ngữ metagenomics lần đầu tiên được sử dụng và công bố vào năm 1985 xuất phát từ ý tưởng nghiên cứu phân tích thông tin di truyền của một hệ sinh vật. Hiện nay, metagenomics được mở rộng sử dụng dựa vào các công nghệ hiện đại để nghiên cứu các loài vi sinh vật trong các phức hợp vi sinh vật thu trực tiếp từ môi trường tự nhiên mà không cần nuôi cấy. Các nhà nghiên cứu về vi sinh vật cho biết, số lượng vi sinh vật tồn tại tự nhiên là vô cùng lớn và rất phong phú về thành phần loài song phần lớn chúng lại không thể nuôi cấy trên môi trường nhân tạo, do đó khó

(thậm chí là không thể) xác định trình tự DNA cũng như nghiên cứu phân tích chúng. Những nghiên cứu đầu tiên về metagenomics đã tập trung phân tích trình tự 16S rRNA được tìm thấy không thuộc về bất kỳ loài nào đã được phân lập và nghiên cứu trước đây. Điều này cho thấy, đã có rất nhiều loài vi sinh vật bị bỏ sót, không được nghiên cứu. Với kết quả phân tích trình tự các đoạn 16S rRNA thu trực tiếp từ môi trường tự nhiên cho thấy, nếu chỉ sử dụng phương pháp nuôi truyền thống, chúng ta chỉ có thể nghiên cứu được 1% số lượng các loài vi sinh vật tồn tại trong mẫu tự nhiên nói chung; 0,001-0,1% các loài vi sinh vật biển; 0,25% các loài vi sinh vật nước ngọt; 0,25% vi sinh vật trong trầm tích; 0,3% vi sinh vật đất. Một nghiên cứu đã minh chứng cho thông tin này: năm 2002, Mya Breitbart (Đại học South Florida - Hoa Kỳ) và Forest Rohwer (Đại học San Diego State - Hoa Kỳ) cùng các đồng nghiệp đã sử dụng công nghệ metagenomics cho thấy, trong 200 lít nước biển có chứa hơn 5.000 loại virus khác nhau. Torsvik (Đại học Bergen - Na Uy), Turnbaugh (Đại học Harvard - Hoa Kỳ), Tringe (Viện Nghiên cứu Genome - Hoa Kỳ) và cộng sự cũng sử dụng công nghệ metagenomics cho thấy, có trên 1 triệu loại virus khác nhau cho mỗi kg trầm tích biển (bao gồm cả thực khuẩn thể). Công nghệ metagenomics ra đời đã khắc phục được những hạn chế của các phương pháp truyền thống, hướng sự tập trung nghiên cứu vào các vi sinh vật chưa được chú ý đến hoặc chưa biết đến. Bằng cách phân lập và nghiên cứu trực tiếp genome của toàn bộ các vi sinh vật trong một môi trường nghiên cứu, ta có thể có thông tin di truyền của hệ vi sinh vật ở đó mà không cần phải phân lập và nuôi cấy từng tế bào riêng lẻ.

Phương pháp tiếp cận tìm gen mới bằng metagenomics

Metagenomics nghiên cứu đa hệ gen quần xã vi sinh vật thông qua ba bước: 1) tách chiết DNA đa hệ gen của vi sinh vật trong mẫu thu thập; 2) thiết lập thư viện DNA đa hệ gen hoặc giải trình tự DNA đa hệ gen và 3) sàng lọc hoặc phân lập gen mong muốn. Metagenomics tiếp cận đa hệ gen theo hai phương pháp chính là phân lập gen dựa trên việc thiết lập thư viện DNA đa hệ gen và khai thác trình tự DNA, phân lập gen dựa trên dữ liệu giải trình tự DNA đa hệ gen. Trong đó, cách tiếp cận dựa vào chức năng đang ngày càng phổ biến.

Phân lập gen từ thư viện DNA của đa hệ gen

Trong giai đoạn đầu phát triển công nghệ metagenomics, cách tiếp cận đa hệ gen là phân lập gen dựa vào thư viện DNA đa hệ gen. Theo cách này, trước hết, DNA đa hệ gen được tách chiết trực tiếp từ mẫu môi trường. Bước tiếp theo là toàn bộ DNA đa hệ gen sẽ được phân cắt bằng enzyme hạn chế để tạo thành các đoạn DNA có kích thước đủ để chứa được trọn vẹn trình tự của gen mong muốn. Sau đó, các đoạn DNA này được gắn vào vector thích hợp và chuyển vào chủng vi sinh vật chủ. Với số lượng dòng đủ lớn, thư viện có thể chứa được toàn bộ các gen của đa hệ gen. Các dòng biểu hiện protein ngoại lai sau đó sẽ được sàng lọc hoạt tính (ví dụ như sản xuất vitamin, chất kháng kháng sinh, enzyme...) trên môi trường có cơ chất đặc hiệu. Từ đó các dòng mang gen mã hóa cho tính trạng mong muốn sẽ được lựa chọn, giải trình tự để thu được trình tự gen. Nhiều cellulose đã được phát hiện nhờ phương pháp này (từ thư viện DNA đa hệ gen của vi sinh vật trong dạ cỏ trâu có 61 dòng khác nhau có hoạt tính cellulose đã được phân lập, trong đó 13 dòng có hoạt tính endoglucanase; từ thư viện DNA hệ gen của vi sinh vật sống trong chất thải của nhà máy giấy có 2 dòng có hoạt tính endoglucanase, 3 dòng có hoạt tính exoglucanase và 2 dòng có hoạt tính β- glucosidase đã được phân lập…). Tuy nhiên, việc phân lập gen dựa trên việc sàng lọc thư viện DNA đa hệ gen trên môi trường có cơ chất thường tốn rất nhiều thời gian và công sức do phải sàng lọc một khối lượng lớn các dòng trong thư viện. Hơn nữa, cách tiếp cận này còn yêu cầu số lượng dòng thư viện phải rất lớn và chất lượng thư viện phải cao. Ngày nay, với sự phát triển vượt bậc của khoa học và công nghệ, việc giải trình tự toàn bộ hệ gen đã trở nên khả thi hơn (tiết kiệm thời gian và kinh phí hơn). Do đó, metagenomics khai thác gen cũng như đánh giá sự đa dạng vi sinh vật đều dựa trên trình tự DNA đa hệ gen của vi sinh vật.

Khai thác và phân lập gen từ dữ liệu trình tự DNA đa hệ gen

Trong những năm gần đây, kỹ thuật giải trình tự thông lượng cao (High Throughput Sequencing - HTS) được ứng dụng rộng rãi trong việc giải trình tự DNA đa hệ gen của quần xã vi sinh vật sống trong môi trường sống nhất định. Điểm khác biệt quan trọng giữa HTS và kỹ thuật giải trình tự Sanger truyền thống là dung lượng. Trong khi một Sanger điển hình tạo ra được 102 trình tự (với độ dài 600-900 bp) thì HTS (ví dụ 454 và Illumina) có thể sinh ra 106-109 trình tự (với độ dài 100-700 bp) cho mỗi lần chạy. Với ưu điểm giải trình tự nhanh và chính xác, kỹ thuật giải trình tự HTS được sử dụng rất phổ biến để giải trình tự DNA đa hệ gen của vi sinh vật trong môi trường sống. DNA đa hệ gen sau khi được giải trình tự là một hỗn hợp các đoạn trình tự riêng rẽ, vì vậy, các trình tự này sẽ được sắp xếp lại và được xử lý bằng các phần mềm chuyên dụng như

SOAPdenovo, BWA, FragGeneScan, MetageneMark, MetageneAnnotator (MGA)/Metagene, Orphelia… để có được trình tự DNA đa hệ gen hoàn chỉnh. Trình tự hoàn chỉnh này sẽ được dự đoán chức năng sinh học và đơn vị phân loại dựa vào mức độ tương đồng của trình tự với nhiều cơ sở dữ liệu tham khảo có sẵn sẽ đưa ra thông tin về chức năng của gen, như KEGG (Kyoto Encyclopedia of Genes and Genomes, là cơ sở dữ liệu trực tuyến liên quan đến hệ gen, các con đường enzyme và các sản phẩm sinh học), eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups, là cơ sở dữ liệu chứa các nhóm orthologous), COG/KOG (COG - Clusters of Orthologous Group, là cơ sở dữ liệu protein của sinh vật nhân sơ, nhân chuẩn đơn bào; KOG - eukaryotic orthologous groups, là cơ sở dữ liệu từ 7 hệ gen sinh vật nhân chuẩn: 3 loài động vật, 1 loài thực vật, *Arabidopsis thaliana*, 2 loài nấm và các ký sinh trùng nội bào), PFAM và TIGRFAM (là cơ sở dữ liệu các họ protein). Tuy nhiên, không có cơ sở dữ liệu tham khảo nào chứa đầy đủ các thông tin về chức năng sinh học, về loài… Vì vậy, để đạt được kết quả tốt nhất thì tất cả các cơ sở dữ liệu cần được sử dụng.

Ứng dụng của metagenomics trong khai thác gen mới và đánh giá sự đa dạng vi sinh vật

Trong nhiều môi trường, số vi sinh vật không thể nuôi cấy được bằng công nghệ chiếm khoảng 99%. Vì vậy, các phương pháp nghiên cứu sự đa dạng di truyền của vi sinh vật, cấu trúc quần thể và vai trò sinh thái của các vi sinh vật không thông qua nuôi cấy vi sinh vật là rất cần thiết. Metagenomics, thông qua việc giải trình tự toàn bộ DNA đa hệ gen, là một phương pháp hiệu quả cho phép nghiên cứu sự đa dạng loài cũng như khai thác các enzyme với hoạt tính xúc tác sinh học mới của các vi khuẩn không nuôi cấy được từ hệ sinh thái trong các môi trường tự nhiên. Chính vì vậy, số lượng các nghiên cứu đa hệ gen của khu hệ vi sinh vật dựa trên việc giải mã toàn bộ hệ gen bằng máy giải trình tự thế hệ mới không ngừng gia tăng kể từ khi DNA đa hệ gen đầu tiên của các vi sinh vật sống trong hệ thống thoát nước của mỏ axit được giải trình tự. Cho đến nay, nhiều DNA đa hệ gen của vi sinh vật ở các môi trường khác nhau đã được đánh giá đa dạng vi sinh vật cũng như đa dạng di truyền như môi trường nước biển, ruột người, dạ cỏ bò, phân compost và đất... Phân tích trình tự DNA đa hệ gen vi sinh vật của 33 mẫu đất lấy từ đất trồng cỏ, đất rừng, đất sa mạc, đất Arctic và đất rừng ngập mặn cho thấy, vi sinh vật có mặt trong đất gồm 11 ngành và 53 chi. Trong đó, Proteobacteria là ngành chiếm tỷ lệ cao nhất trong quần xã vi sinh vật của đất (trừ mẫu đất sa mạc). Ở đất sa mạc, cả 2 ngành Proteobacteria và Actinobacteria chiếm ưu thế: 30% Proteobacteria và 29% Actinobacteria. Ngoài ra, Firmicutes và Bacteroidetes là hai ngành chính chiếm ưu thế ở quần xã vi sinh vật đường ruột người không có mặt thường xuyên trong các quần xã vi sinh vật đất. Bằng phương pháp giải trình tự gen trên các thiết bị thế hệ mới của Illumina Inc/Genome Analysis, DNA đa hệ gen của toàn bộ vi sinh vật có trong 124 mẫu phân người châu Âu đã được giải trình tự. Kết quả xử lý và phân tích trình tự cho thấy, kích thước DNA đa hệ gen lên đến 576,7 gigabase (Gb), có 3,3 triệu gen vi sinh vật (lớn hơn gần 150 lần so với số gen của 1 hệ gen người hoàn chỉnh) đã được xác định đặc điểm. Trong đó, hơn 90% gen là của vi khuẩn, phần còn lại chủ yếu là gen của vi khuẩn cổ, chỉ 1% là gen của sinh vật nhân chuẩn và virus. Kết quả phân tích cũng cho thấy, hơn 99% gen vi khuẩn thuộc trong nhóm 1.000 và 1.150 loài vi khuẩn phổ biến và mỗi cơ thể người có ít nhất 160 loài như vậy. Ngoài ra, cũng rất nhiều enzyme chuyển hóa lignocelluose được tìm thấy từ các hệ tự nhiên khác như trong dạ cỏ trâu bò hay mẫu nước hồ và đất rừng ngập mặn. Việc tìm kiếm các enzyme mới mã hóa cho hệ enzyme cellulase chuyển hóa lignocellulose từ một số hệ mini sinh thái bằng phương pháp metagenomic hiện đang có xu hướng phát triển, giúp khám phá ra nhiều enzyme mới với đặc tính quý, giảm giá thành trong quá trình ứng dụng công nghiệp và đáp ứng nhu cầu về việc sản xuất năng lượng sạch trong tương lai.

Định hướng nghiên cứu, ứng dụng metagenomics trên thế giới và việc triển khai tại Việt Nam

Metagenomics đã tạo nên những tiến bộ vượt bậc trong sinh thái học, tiến hóa và đa dạng vi sinh vật. Đây là định hướng mới, quan trọng và trở thành chiến lược trong phát triển kinh tế bền vững, an ninh quốc phòng, bảo vệ sức khỏe và môi trường trên thế giới hiện nay. Metagenomics là công cụ mới, tổ hợp rất nhiều kỹ thuật sinh học kết hợp tin - sinh học để phân tích sàng lọc, xây dựng thư viện metagenomics, quản lý và khai thác cho môi trường tự nhiên, cơ thể người, động vật, thực vật không thông qua nuôi cấy.

Metagenomics được ứng dụng và mang lại hiệu quả cho phát triển kinh tế, an sinh xã hội và môi trường, như trong lĩnh vực khoa học trái đất: phát triển mô hình genome trên cơ sở sinh thái vi sinh vật nhằm mô tả và dự báo các quá trình môi trường toàn cầu, thay đổi khí hậu và sự bền vững của trái đất. Trong lĩnh vực khoa học sự sống: các học thuyết mới trình độ tiên tiến dự báo trước các năng lực sinh học quần thể vi sinh vật

gốc, sinh thái và sự tiến hóa. Trong lĩnh vực y - sinh học: xác định rõ mức độ toàn cầu sự đóng góp của hệ vi sinh vật cho sức khỏe, phát hiện và điều trị bệnh cho các cá thể, cộng đồng người, động vật, thực vật; phát triển các phương pháp điều trị mới trên cơ sở các kiến thức metagenomics. Trong lĩnh vực năng lượng: phát triển hệ vi sinh vật và các quá trình cho nguồn năng lượng sinh học mới có hiệu quả kinh tế, môi trường bền vững hơn và có khả năng phục hồi nhanh với các biến động xấu của trái đất. Trong lĩnh vực môi trường: phát triển các công cụ để quan trắc môi trường khi có sự cố ở tất cả các mức độ, từ thay đổi khí hậu đến rò rỉ khí đốt, hóa chất, dầu từ kho chứa và các phương pháp vi sinh vật cơ bản được gọi là phương pháp xanh phục vụ việc duy trì các hệ sinh thái.

Để tiếp cận được với những thành tựu của thế giới, tạo điều kiện khai thác hiệu quả nguồn gen vi sinh vật phong phú và đa dạng của Việt Nam ứng dụng vào nhiều lĩnh vực thực tiễn khác nhau, ngày 23.4.2014, Bộ trưởng Bộ Khoa học và Công nghệ đã ban hành Quyết định số 826/QĐ-BKHCN phê duyệt Danh mục đặt hàng dự án khoa học và công nghệ cấp quốc gia "Nghiên cứu metagenome của vi sinh vật từ một số môi trường đặc thù nhằm tìm kiếm các gen, enzyme, chất xúc tác sinh học mới để sản xuất các chế phẩm sinh học phục vụ đời sống, bảo vệ sức khỏe con người và môi trường" với mục tiêu đặt ra là: sử dụng metagenomics để khai thác các vật liệu di truyền mới mã hóa các enzyme, các chất có hoạt tính sinh học, các chất xúc tác, các chất kháng u... có tiềm năng công nghệ trong sản xuất thuốc, tạo kit chẩn đoán bệnh, phát triển cây trồng, vật nuôi có giá trị kinh tế, góp phần bảo vệ sức khỏe cộng đồng, phát triển kinh tế bền vững, bảo vệ môi trường.

Dự án khoa học và công nghệ trên đã được đặt hàng và giao cho Viện Hàn lâm Khoa học và Công nghệ Việt Nam chủ trì với 5 định hướng nghiên cứu chính gồm: 1) Nghiên cứu metagenome của vi sinh vật vùng đất ô nhiễm chất diệt cỏ/dioxin nhằm tìm kiếm các gen mới có khả năng phân hủy dioxin; 2) Nghiên cứu metagenome của vi sinh vật đất vùng rễ một số đại diện cây trồng ở Việt Nam: cây thuốc có củ (cây nghệ), cây công nghiệp (cà phê, lạc) nhằm tăng năng suất và chất lượng cây trồng; 3) Nghiên cứu metagenome của ba hệ mini sinh thái tiềm năng nhằm khai thác các gen mới mã hóa hệ enzyme chuyển hóa hiệu quả lignocellulose; 4) Nghiên cứu metagenome của vi sinh vật trong các đầm nuôi tôm, góp phần tạo cơ sở khoa học để phát triển nghề nuôi tôm ở Việt Nam; 5) Nghiên cứu metagenome của vi sinh vật liên kết hải miên tại biển miền Trung Việt Nam nhằm phát hiện và sàng lọc các chất hoạt tính sinh học mới.

Ở Việt Nam, đây là những nghiên cứu đầu tiên ứng dụng công nghệ metagenomics để giải các bài toán có tính chất tổng thể, cấp thiết, mở ra một cách nhìn toàn diện về một đối tượng cụ thể nói riêng cũng như các mắt xích trong hệ sinh thái nói chung, từ đó xây dựng cơ sở dữ liệu về đa dạng sinh học của các hệ vi sinh vật có giá trị, làm cơ sở cho việc khai thác và ứng dụng trong các nghiên cứu về sinh học hệ thống sau này.

Những nghiên cứu đánh giá đa dạng và tiềm năng di truyền của vi sinh vật từ hệ sinh thái của Việt Nam sẽ góp phần đáng kể vào sự phát triển kinh tế dựa trên nền tảng sinh học bền vững của đất nước. Kết quả nghiên cứu thu được có thể thay đổi phương thức canh tác để đạt hiệu quả cao nhất nhằm phát triển bền vững nền nông nghiệp cũng như nguồn cây nguyên liệu phục vụ phát triển công nghiệp dược của Việt Nam trong tương lai ✍

Tài liệu tham khảo

1. Talebnia F., Karakashev D., Angelidaki I. Production of bioethanol from wheat straw: an overview on pretreatment, hydrolysis and fermentation. Bioresource Technology 2010, 101(13):4744-4753.

2. Nguyen T.A.D., Kim K.R., Kim M.S., Sim S.J. Thermophilic hydrogen fermentation from Korean rice straw by Thermotoga neapolitana. International Journal of Hydrogen Energy 2010, 35(24):13392-13398.

3. Gao L., Yang H., Wang X., Huang Z., Ishii M., Igarashi Y., Cui Z. Rice straw fermentation using lactic acid bacteria. Bioresource technology 2008, 99(8):2742-2748.

4. Gírio F., Fonseca C., Carvalheiro F., Duarte L., Marques S., Bogel-Lukasik R. Hemicelluloses for fuel ethanol: a review. Bioresource technology 2010, 101(13):4775-4800.

5. Saha B.C., Cotta M.A. Comparison of pretreatment strategies for enzymatic saccharification and fermentation of barley straw to ethanol. New biotechnology 2010, 27(1):10-16.

6. Lo Y.C., Saratale G.D., Chen W.M., Bai M.D., Chang J.S. Isolation of cellulose-hydrolytic bacteria and applications of the cellulolytic enzymes for cellulosic biohydrogen production. Enzyme and Microbial Technology 2009, 44(6):417-425.

7. Rajendhran J., Gunasekaran P. Strategies for accessing soil metagenome for desired applications. Biotechnology advances 2008, 26(6):576-590.