

Công nghệ giải trình tự thế hệ mới:

TỔNG QUAN VỀ CÁC XU HƯỚNG VÀ ỨNG DỤNG

Nguyễn Đăng Tôn, Nông Văn Hải

Viện Nghiên cứu hệ gen

Công nghệ giải trình tự thế hệ mới (Next Generation Sequencing - NGS) đã có mặt và phát triển từ 10 năm nay. Với ưu thế có giá thành rất thấp và thời gian đọc nhanh, các công nghệ NGS được ứng dụng rộng rãi trong nhiều lĩnh vực, từ nghiên cứu cơ bản cho đến các nghiên cứu về chẩn đoán lâm sàng, hệ gen học trong nông nghiệp và khoa học hình sự... Trong khuôn khổ bài viết, các tác giả đưa ra cái nhìn tổng quan về các hướng ứng dụng của công nghệ NGS và thảo luận về những ưu, nhược điểm của mỗi công nghệ.

Mở đầu

Vào những năm 70 của thế kỷ trước, phương pháp giải trình tự ADN bằng chuỗi tổng hợp hoặc kỹ thuật tách các đoạn ADN đã được Sanger và cộng sự phát minh. Phương pháp này dễ dàng tự động hóa, không dùng các chất độc hại nên được sử dụng phổ biến cho giải trình tự ADN từ những năm 80 cho đến nay. Trình tự bộ gen người đầu tiên đã được giải mã bằng phương pháp Sanger vào năm 2004 với sự hợp tác của 15 quốc gia do Mỹ đứng đầu, đã tiêu tốn rất nhiều thời gian và nguồn lực. Câu hỏi được đặt ra là, làm thế nào để có thể rút ngắn thời gian và giảm chi phí giải trình tự toàn bộ hệ gen. Với lý do này, Viện Nghiên cứu hệ gen người quốc gia (NHGRI - Hoa Kỳ) đã khởi động chương trình đầu tư với mục tiêu giảm chi phí giải mã hệ gen người xuống 1.000 USD trong 10 năm. Đây là động lực thúc đẩy sự phát triển và thương mại hóa các công nghệ NGS. Các phương pháp giải trình tự này có 3 đặc điểm cải tiến chính: dựa vào thư viện NGS mà không cần nhân dòng các đoạn ADN, hàng ngàn cho tới hàng triệu phản ứng giải trình tự được thực hiện cùng lúc, kết quả giải trình tự được xác định trực tiếp không cần thông qua điện di. Dưới đây là tóm tắt một số công nghệ NGS đang được

ứng dụng rộng rãi hiện nay.

Một số công nghệ NGS

Công nghệ giải trình tự pyrosequencing 454

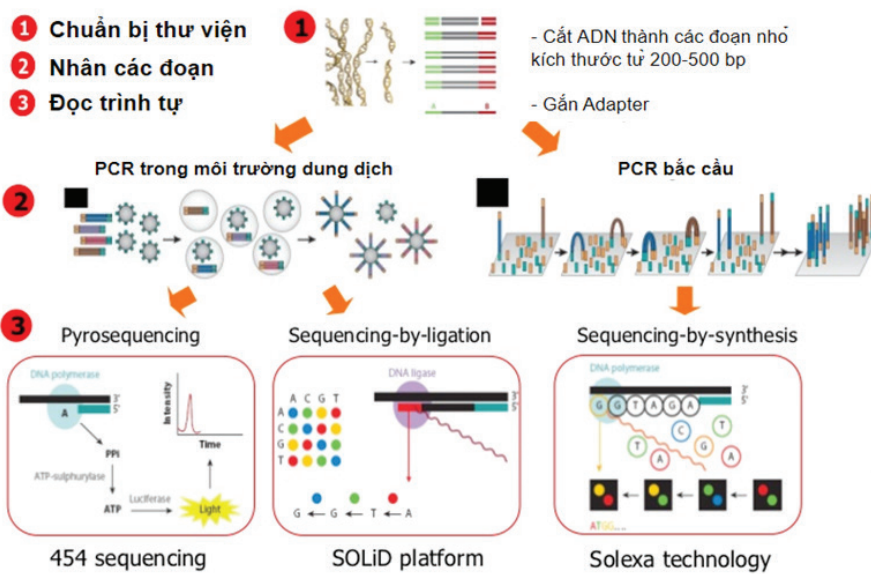
Pyrosequencing 454 do 2 nhà khoa học Nyren và Ronaghi của Viện Kỹ thuật Stockholm (Thụy Điển) phát minh, và được phát triển bởi Công ty 454 Life Science, là một hệ thống giải trình tự ADN 2 bước có độ tương đồng cao với dung lượng lớn hơn rất nhiều so với hệ thống giải trình tự Sanger. Kỹ thuật này dựa trên nguyên lý “giải trình tự bằng tổng hợp” bao gồm: Khởi động một sợi ADN đã được giải trình tự và giải trình tự sợi bổ sung bằng phản ứng của enzyme. Đây là hệ thống có thể khuếch đại một số lượng lớn các đoạn ADN trong các giếng picotiter. Nguyên lý “giải trình tự bằng việc tổng hợp” cũng dựa trên việc nhận biết các pyrophosphate (PPi) được giải phóng trong quá trình gắn nucleotide, tạo ra một tín hiệu ánh sáng, hiệu quả hơn kỹ thuật kết thúc chuỗi bằng dideoxynucleotide.

Công nghệ pyrosequencing có tính linh hoạt cao khi liên kết cặp primer, có khả năng phân tích được nhiều đột biến, dữ liệu và thông tin trình tự tập hợp được đầy đủ. Kỹ thuật này có tính nhạy cao hơn hẳn so với phương pháp truyền thống, độ chính

xác lên tới 99,9% với các đoạn 200 base và 99% với các đoạn 400 base. Hệ thống giải trình tự được 400-600 triệu bp trong vòng 10 giờ, giúp giảm giá thành đáng kể so với khi sử dụng phương pháp Sanger để giải trình tự một số lượng lớn ADN. Công nghệ này cùng với các kỹ thuật giải trình tự thế hệ mới khác có nhiều ảnh hưởng đến hoạt động nghiên cứu có lượng dữ liệu đầu vào lớn, và yêu cầu xử lý trong thời gian ngắn, nhờ đó đã tạo ra nhiều đột phá trong các lĩnh vực: công nghệ sinh học, sinh học pháp y, hệ thống học và y học.

Công nghệ giải trình tự SBS (sequencing by synthesis)

Công nghệ giải trình tự bằng tổng hợp SBS sử dụng 4 nucleotide đánh dấu huỳnh quang để giải trình tự hàng chục triệu cluster đồng thời trên bề mặt flow-cell. Trong mỗi chu trình giải trình tự, một deoxynucleoside triphosphate đánh dấu (dNTP) được thêm vào chuỗi acid nucleic. Nhãn huỳnh quang của nucleotide đóng vai trò như một khóa dùng phản ứng polymer hóa, do đó sau khi mỗi dNTP được tích hợp, dye huỳnh quang được ghi lại để xác định nucleotide và sau đó bị cắt bỏ để tổng hợp nucleotide tiếp theo. Vì cả 4 loại dNTP gắn khóa dùng tổng hợp thuận nghịch có mặt đồng thời dưới dạng đơn phân tử,



Hình 1. Sơ đồ tổng quát các bước giải trình tự ADN bằng NGS.

- Chuẩn bị mẫu: ADN được cắt thành các đoạn nhỏ có kích thước từ 200 đến 500 bp, gắn adapter phù hợp; 2. Các đoạn ADN sau khi gắn adapter được tiến hành làm giàu bằng PCR (Polymerase chain reaction) trong môi trường dung dịch (emulsion PCR) đối với công nghệ giải trình tự trên máy đọc trình tự 454 của Roche và máy đọc trình tự SOLiD của Life Technologies hoặc PCR bắc cầu (bridge PCR) theo công nghệ Solexa của Illumina; 3. Đọc trình tự các đoạn ADN đã được làm giàu bằng các công nghệ tương ứng

sự cạnh tranh ngẫu nhiên giúp giảm thiểu việc tổng hợp mất cân đối. Việc xác định các base dựa vào cường độ tín hiệu đo được trong mỗi chu trình, giúp giảm thiểu các sai số thô so với công nghệ khác. Kết quả cuối cùng là trình tự được đọc từng base một với độ chính xác cao, loại bỏ được các lỗi do đặc thù trình tự, cho phép quá trình giải trình tự mạnh mẽ trên toàn bộ genome, bao gồm các trình tự lặp lại.

Phương pháp giải trình tự SBS có khả năng xử lý đồng thời số lượng cực lớn các đoạn cần đọc trình tự với độ bao phủ đồng nhất, giúp kết quả xác định biến dị di truyền có độ tin cậy cao. Khối lượng mẫu lớn cho phép sử dụng các công cụ thống kê và cho điểm, tương tự như các phương pháp truyền thống trong việc xác định thể đồng hợp, dị hợp và phân biệt các lỗi giải trình tự. Cụ thể, mỗi base đọc thô có một điểm số chất lượng giúp phần mềm có thể xác định sự sai khác và cho điểm số tin cậy.

Công nghệ giải trình tự SBS là quy trình một chiều, mang tính tự động hóa cao, yêu cầu thời gian và thao tác ít nhất. Với khả năng tạo ra hàng Gb dữ liệu ADN trong một lần chạy, ngay cả các genome lớn của các động vật có vú cũng có thể giải trình tự trong một vài tuần thay vì một

vài năm như trước kia.

Công nghệ giải trình tự SOLiD

Công nghệ giải trình tự SOLiD được thiết kế dựa trên nguyên lý ghép nối, toàn bộ genome được cắt thành các đoạn ADN ngắn, sau đó từng đoạn được gắn với các adapter rồi cố định vào các hạt từ. Quá trình giải trình tự được thực hiện thông qua các oligonucleotide được đánh dấu huỳnh quang. Đầu tiên các primer được ghép cặp thông qua liên kết bổ sung với các đoạn adapter, quá trình tổng hợp được thực hiện như sau: 4 loại oligonucleotide (mỗi đoạn gồm 8 base) có đánh dấu huỳnh quang được gắn vào các vị trí tiếp theo bắt đầu từ vị trí 5'P của mỗi. Sau mỗi một lượt gắn của các đoạn oligonucleotide thì một tín hiệu huỳnh quang được ghi nhận, tiếp đến nhờ hoạt tính của enzyme, các nucleotide bắt đầu từ số 6 bị loại ra, kèm theo đó nhãn huỳnh quang được loại bỏ để lộ ra vị trí 5'P, quá trình gắn được tiếp tục thực hiện cho tới hết chiều dài của đoạn ADN. Kết thúc quá trình thứ nhất, một primer mới tiếp tục được gắn với sợi khuôn ở vị trí tịnh tiến 1 nucleotide về phía trước so với vị trí gắn của primer đầu tiên và tiếp tục quá trình thu nhận tín hiệu thông qua các phản ứng ghép nối. Quá trình được lặp lại 5 lần, mỗi lần sử dụng một mỗi và tịnh tiến 1 nucleotide về phía trước

Nhìn ra thế giới

so với vị trí gắn mỗi trước đó.

Dữ liệu xuất ra của hệ thống có thể lên tới 60 Gb với khoảng hơn 1 tỷ đoạn được đọc sau mỗi lượt chạy. Kích thước và độ tin cậy của mỗi đoạn ADN được đánh giá là tương đương so với hệ thống của Illumina, trong khi đó chi phí thiết bị của công nghệ SOLiD thấp hơn đáng kể so với các hệ thống khác cùng thế hệ.

Công nghệ giải trình tự Nanoball sequencing

Nanoball sequencing là một công nghệ mới, tiên tiến, được sử dụng để giải trình tự toàn bộ hệ gen của một sinh vật dựa trên quá trình sao chép của vi khuẩn. Các đoạn ADN được khuếch đại rồi cuộn lại thành các nanoball. Quá trình giải trình tự thông qua kỹ thuật nanoball gồm có: i) Tinh sạch ADN tổng, tiếp đó sử dụng enzyme để biến chúng thành các đoạn ADN có kích thước từ 400 đến 500 bp, sau đó gắn adapter vào các đoạn ADN, rồi cuộn chúng lại thành các nanoball; ii) Các đoạn ADN được sao chép theo cơ chế sao chép ADN dạng vòng của vi khuẩn, sau đó các nanoball được chuyển lên một flow cell có chứa các mẫu dò có gắn huỳnh quang rồi gắn với các trình tự nucleotide đặc hiệu; iii) Tín hiệu huỳnh quang thu được tại mỗi điểm đặc hiệu được ghi lại thông qua một máy ảnh có độ phân giải cao và được phân tích thông qua các công cụ phân tích tin sinh học. Cuối cùng, các dữ liệu về gen được so sánh, lắp ráp và xác định trình tự.

Công nghệ giải trình tự SMRT (Single Molecular Real-Time)

SMRT là công nghệ giải trình tự tiên tiến nhất hiện nay. Nếu tất cả các công nghệ giải trình tự nêu trên đều dựa trên phương pháp cơ bản do Sanger phát minh ra năm 1977, gọi là "phương pháp gián đoạn chuỗi" (chain-termination method), thì SMRT xác định trình tự ADN theo một cách hơi khác: Nó "quan sát" quá trình tổng hợp một chuỗi ADN tự nhiên bằng ADN polymerase đơn lẻ, các tín hiệu từ

nucleotide được đánh dấu phosphate sẽ được phát hiện theo thời gian thực giúp xác định chính xác nucleotide nào trong 4 loại nucleotide đang được gắn vào mạch, do đó khi đoạn ADN được sao chép xong thì máy cũng xác định xong trình tự đoạn ADN đó. Công nghệ này vô cùng hữu ích cho các ứng dụng giải trình tự toàn bộ hệ gen của các sinh vật chưa có thông tin về hệ gen (de novo genome sequencing).

Các hướng ứng dụng NGS

Giải trình tự ADN

Giải trình tự toàn bộ hệ gen (Whole genome sequencing - WGS) và giải trình tự hệ gen biểu hiện (Whole exome sequencing - WES) là 2 xu hướng chính ứng dụng công nghệ NGS vào giải trình tự ADN. WGS/WES cho phép tìm kiếm và xác định các biến dị di truyền SNP, InDel, CNV, SV... có trong cá thể, quần thể đối với những loài đã có hệ gen tham chiếu. Việc phát hiện các biến dị di truyền giúp phát hiện các biến dị gây ra bệnh di truyền đã biết, từ đó tiến hành sàng lọc ung thư/bệnh di truyền sớm hoặc theo dõi đáp ứng điều trị; cung cấp các biến dị cho nghiên cứu GWAS, chú giải...; đánh giá đa dạng quần thể bằng cách xác định tần số allele...

Những ưu điểm vượt trội về công suất và giá thành đã khiến WGS trở thành phương pháp thích hợp nhất cho nhiều mục đích nghiên cứu. WGS ngày càng được sử dụng nhiều hơn cho các nghiên cứu giải mã như khoa học pháp y, di truyền nông nghiệp và các chẩn đoán lâm sàng (chẩn đoán bệnh di truyền là một ví dụ). Một trong những ứng dụng lâm sàng quan trọng khác là giải trình tự các chủng gây bệnh và các bệnh truyền nhiễm quan tâm.

Đối với nhiều ứng dụng, việc giải trình tự toàn bộ hệ gen là không thực tế và không cần thiết. Do đó, giải trình tự hệ gen biểu hiện (WES) là phương án tiếp cận hợp lý đối với những vùng gen quan tâm, ví dụ Exon chỉ chiếm 2% trong hệ gen người nhưng biến đổi

trong đó lại gây nên 85% các căn bệnh đã biết. Vì lý do này, WES đã được sử dụng cho nghiên cứu lâm sàng trong những năm gần đây, cho ra nhiều công cụ chẩn đoán hứa hẹn sẽ làm thay đổi đáng kể dịch vụ y tế và chăm sóc sức khỏe.

Một hướng tiếp cận cao hơn mới phát triển trong thời gian gần đây là giải trình tự các gen quan tâm đã được khuếch đại bằng PCR. Phương pháp này phù hợp với các kiểm tra bệnh, tập trung vào một lượng giới hạn các bệnh có liên quan đến các biến thể hoặc ứng dụng trong giải trình tự gen 16S của rARN từ các loài khác nhau. Đây là phương pháp được sử dụng rộng rãi trong hệ thống học và phát sinh chủng loại, cụ thể là giữa các mẫu đa dạng về di truyền. Phương pháp này dùng để đánh giá sự đa dạng của vi khuẩn trong môi trường, giúp các nhà nghiên cứu phân tích được hệ vi sinh vật từ các mẫu khó hoặc không thể nghiên cứu được.

Giải trình tự ARN (ARN-seq)

Các bản phiên mã của sinh vật nhân chuẩn rất phức tạp và nhiều gen tạo ra các bản đối mã. Trên cơ sở các công nghệ NGS, nhiều quy trình đặc hiệu cho ARN-seq đã được phát triển, quy trình đầu tiên xuất hiện năm 2008. Các quy trình này cho phép xác định các bản đối mã điều hòa có thể đóng vai trò quan trọng trong các chức năng sinh học. Phân tích các bản phiên mã hiện nay cũng có thể thực hiện được ở mức độ các tế bào đơn lẻ dựa trên các phương pháp chuẩn bị mẫu đã được cải tiến. Phân tích toàn bộ các bản phiên mã của các tế bào đơn lẻ đã cho thấy có thể có những bản phiên mã không hoàn toàn giống nhau giữa các tế bào giống nhau. Một phương pháp mới được công bố năm 2014 có tên là giải trình tự ARN huỳnh quang tại chỗ (FISSEQ), có thể nghiên cứu về toàn bộ các bản phiên mã của các tế bào đơn lẻ, đồng thời xác định được vị trí chính xác của mỗi bản phiên mã trong tế bào.

Hạn chế của phương pháp ARN-seq truyền thống là việc xác định mức độ ổn định của ARN không phản ánh được hoạt động phiên mã hoặc tốc độ sinh tổng hợp protein. Vài năm trước đây, phương pháp NET-seq ra đời giúp chúng ta nhìn được các bản phiên mã ở độ phân giải từng nucleotide thông qua giải trình tự đặc hiệu từng bản sao mới được tạo nên. Phương pháp này dựa trên kết tủa miễn dịch của ARN polymerase kết hợp với giải trình tự đầu 3' của các ARNs mới tổng hợp được đồng kết tủa. NET-seq là phương pháp thay thế cho ARNP ChIP-seq với độ phân giải cao hơn và giữ lại thông tin về mạch ARN.

Một phương pháp khác là ribo-seq đã được sử dụng để xác định bản đồ dịch mã ribosome trên ARN thông tin, phương pháp này kết hợp giữa dấu vết của nuclease và giải trình tự 28-30 nucleotide trong vùng được bảo vệ với ribosome của bản phiên mã. Kỹ thuật này được ứng dụng trong các nghiên cứu về điều hòa dịch mã của gen [60] và cơ chế sinh tổng hợp protein.

Nhìn chung, sự xuất hiện của NGS đã cho phép các nhà khoa học nghiên cứu về các hệ thống sinh học ở mức độ chưa từng đạt được từ trước tới nay. Cùng với sự phát triển của công nghệ, các phương thức chuẩn bị mẫu và công cụ phân tích dữ liệu cũng được tăng cường mạnh mẽ. Do đó, NGS đã trở thành công nghệ chủ chốt trong nghiên cứu cơ bản và nhanh chóng trở thành công cụ cho mọi nghiên cứu về di truyền học và các nghiên cứu liên quan khác. Với sự phát triển mạnh mẽ của công nghệ, trong những năm tới đây, sẽ có hàng triệu bộ gen người được giải mã, tạo ra bộ dữ liệu gen khổng lồ so với những gì đang có. Vì vậy, vấn đề đạo đức cần được đặc biệt quan tâm, đồng thời cần đưa ra những phương pháp hiệu quả trong việc lưu trữ và phân tích dữ liệu khi tốc độ dữ liệu được tạo ra ngày càng lớn.