Tạp chí Công nghệ Sinh học **19**(3): 393-410, 2021

BÀI TỔNG QUAN

NGHIÊN CỨU PHÁT TRIỂN DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT VÀ ĐỊNH HƯỚNG ỨNG DỤNG

Lê Thị Thu Hiền[1,2,✉], Nguyễn Tường Vân[3], Kim Thị Phương Oanh[1,2], Nguyễn Đăng Tôn[1,2], Huỳnh Thị Thu Huệ[1,2], Nguyễn Thùy Dương[1,2], Phạm Lê Bích Hằng[1], Nguyễn Hải Hà[1,2]

[1]*Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[2]*Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*
[3]*Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

✉Người chịu trách nhiệm liên lạc. E-mail: hienlethu@igr.ac.vn; hienlethu@igr.vast.vn

Ngày nhận bài: 14.12.2020
Ngày nhận đăng: 18.3.2021

TÓM TẮT

Nghiên cứu và phân tích dữ liệu lớn về hệ gen sinh vật được ứng dụng trong nhiều lĩnh vực và có tác động lớn đến đời sống xã hội trên quy mô toàn cầu. Nhờ sự ra đời của các công nghệ giải trình tự gen thế hệ mới, hệ gen sinh vật có thể nhanh chóng được xác định. Nhiều quốc gia đã chú trọng đến thúc đẩy và đầu tư cho các hoạt động nghiên cứu và ứng dụng dữ liệu hệ gen. Các dự án lớn về hệ gen người, động vật, thực vật, vi sinh vật đã và đang được mạng lưới các nhà khoa học thuộc chuyên ngành công nghệ gen, tin sinh học, sinh học tính toán, tự động hóa, trí tuệ nhân tạo thuộc các tổ chức khoa học công nghệ quốc gia hoặc nhiều quốc gia, độc lập hoặc hợp tác triển khai thực hiện. Những nguồn dữ liệu khổng lồ được xây dựng, lưu trữ, quản lý và khai thác hiệu quả. Việt Nam đã ưu tiên đầu tư và phát triển hướng nghiên cứu hệ gen thông qua thành lập các đơn vị chuyên trách cũng như triển khai nghiên cứu hệ gen người và các sinh vật đặc hữu của Việt Nam. Bài viết này tổng quan về: Các công nghệ sử dụng để tạo ra dữ liệu lớn về hệ gen; Một số dự án nghiên cứu và xây dựng cơ sở dữ liệu lớn về hệ gen trên thế giới; Nghiên cứu phát triển dữ liệu lớn về hệ gen ở một số quốc gia và ở Việt Nam; Khai thác và ứng dụng dữ liệu lớn về hệ gen trong các lĩnh vực y dược học phục vụ chăm sóc sức khỏe con người, nông - lâm nghiệp, an toàn thực phẩm và môi trường.

Từ khóa: dữ liệu lớn về hệ gen, giải trình tự gen thế hệ mới, hệ gen, hệ gen biểu hiện, hệ gen phiên mã

MỞ ĐẦU

Hệ gen (genome) của mỗi cá thể sinh vật chứa đựng tất cả thông tin di truyền cần thiết cho sự hình thành, phát triển và hoạt động của sinh vật đó. Trong những năm gần đây, tiến bộ của khoa học kỹ thuật đã cho phép con người số hóa được hệ gen của muôn loài và lưu trữ trong các cơ sở dữ liệu lớn (big data). Hiện nay, nghiên cứu và khai thác dữ liệu toàn bộ hoặc một phần hệ gen của một cá thể sinh vật hoặc nhiều cá thể trong quần thể là một lĩnh vực khoa học và công nghệ mới, có rất nhiều tiềm năng ứng dụng và vai trò quan trọng do tác động tích cực và sâu rộng trong nhiều lĩnh vực của đời sống xã hội trên quy mô toàn cầu.

Những quốc gia phát triển, nơi có tiềm lực và điều kiện tiếp cận các công nghệ tiên tiến, đã rất chú trọng đến thúc đẩy các hoạt động nghiên cứu và ứng dụng dữ liệu hệ gen của các loài sinh vật. Những nguồn dữ liệu khổng lồ và rất phức tạp được xây dựng, lưu trữ, quản lý và khai thác hiệu quả nhờ nỗ lực và sự hợp tác của mạng lưới các nhà khoa học và chuyên gia thuộc nhiều chuyên ngành như công nghệ gen, tin sinh học, sinh học tính toán, tự động hóa, trí tuệ nhân tạo đến từ các viện/trung tâm nghiên cứu, trường đại học, các

công ty, tổ chức quốc tế. Những nguồn dữ liệu này được phân tích và sử dụng để tạo ra các sản phẩm khoa học công nghệ có tính ứng dụng cao trong nhiều lĩnh vực từ y dược học phục vụ chăm sóc sức khỏe con người, tới nông - lâm nghiệp, an toàn thực phẩm, môi trường.

Việt Nam, với một nền kinh tế đang phát triển và hướng vào hội nhập quốc tế, đã ưu tiên đầu tư và phát triển hướng khoa học công nghệ chuyên sâu này thông qua thành lập các trung tâm/đơn vị chuyên trách cũng như triển khai nghiên cứu hệ gen người và các sinh vật đặc hữu của Việt Nam.

Trong khuôn khổ bài viết này, việc nghiên cứu xây dựng và khai thác dữ liệu hệ gen trên thế giới cũng như ở Việt Nam được tìm hiểu, trong đó tập trung tổng quan về: (1) Các công nghệ sử dụng để tạo ra dữ liệu lớn về hệ gen sinh vật; (2) Một số dự án nghiên cứu và xây dựng cơ sở dữ liệu lớn về hệ gen sinh vật trên thế giới; (3) Khai thác và ứng dụng dữ liệu lớn về hệ gen sinh vật; (4) Nghiên cứu phát triển dữ liệu lớn về hệ gen sinh vật ở một số quốc gia tiêu biểu; (5) Nghiên cứu phát triển dữ liệu về hệ gen sinh vật ở Việt Nam; (6) Kết luận.

CÁC CÔNG NGHỆ SỬ DỤNG ĐỂ TẠO RA DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT

Phương pháp xác định trình tự gen đầu tiên đã được Sanger và nhóm nghiên cứu công bố năm 1977. Những năm sau đó, nhiều phương pháp cải biến cùng các hệ thống xác định trình tự gen tự động ra đời đã dẫn tới làn sóng ứng dụng rộng rãi các công nghệ giải trình tự gen trong cộng đồng khoa học trên thế giới.

Năm 2005, công nghệ xác định trình tự gen thế hệ mới (next generation sequencing - NGS) đã ra đời. Rất nhiều hệ thống máy đã được phát triển bởi các hãng như Applied Biosystem/SOLiD; Roche/454; Illumina/Solexa; Pacific Biosciences/RS; Life technologies/Ion PGM, Life technologies/Ion Proton (Shendure, Ji, 2008; Metzker, 2010; Liu *et al*., 2012; Quail *et al*., 2012; Ferrarini *et al*., 2013). Với ưu thế về thời gian, dung lượng, độ chính xác, các công nghệ NGS ngày càng được sử dụng rộng rãi trong nghiên cứu tương quan toàn bộ hệ gen (genome-wide association studies - GWAS), xác định trình tự toàn bộ hệ gen (whole genome sequencing - WGS), hệ gen biểu hiện (whole exome sequencing - WES) hay hệ gen phiên mã transcriptome (RNA-seq)…và có tầm ảnh hưởng rất mạnh ở quy mô toàn cầu, cho phép tạo ra một lượng dữ liệu khổng lồ (Pettersson *et al*., 2009).

Tuy nhiên, để giải quyết khó khăn với những hệ gen có độ phức tạp cao, các đoạn lặp dài hay có số lượng bản sao và cấu trúc đa dạng, công nghệ xác định trình tự gen thế hệ thứ ba (3G) với các đoạn đọc kích thước lớn đã ra đời và gồm hai loại: xác định trình tự tổng hợp (synthetic sequencing) dựa trên công nghệ xác định trình tự các đoạn đọc ngắn để lắp ráp thành các đoạn trình tự dài *in silico* và xác định trình tự thời gian thực đơn phân tử (single-molecular real-time sequencing, SMRT) (Schadt *et al.*, 2010). Hiện nay, phổ biến nhất là hệ thống Illumina, Ion Torrent, hệ thống SMRT PacBio (Pacific Biosciences) xác định trình tự tổng hợp các đoạn dài và hệ thống dựa trên vi giọt của 10X Genomics và MinION (Oxford Nanopore Technologies) (Goodwin *et al.*, 2016).

Thế hệ thứ tư, xác định trình tự mRNA *in situ* (đọc trình tự acid nucleic trực tiếp trong mô hoặc tế bào) được công bố năm 2015, đã mở ra một hướng đi mới cho phân tích biểu hiện gen, tìm kiếm các chỉ thị sinh học, chẩn đoán và phân loại bệnh nhân trong điều trị ung thư.

MỘT SỐ DỰ ÁN NGHIÊN CỨU VÀ XÂY DỰNG CƠ SỞ DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT TRÊN THẾ GIỚI

Khác với những công nghệ giải trình tự gen thế hệ đầu tiên, việc xác định được trình tự toàn bộ hệ gen rất phức tạp, đòi hỏi sự tham gia của rất nhiều nhà khoa học, với chi phí lớn và kéo dài nhiều năm thì nhờ sự ra đời của các công nghệ mới, nhiều phòng thí nghiệm có thể xác định trình tự toàn bộ hệ gen sinh vật trong một thời gian ngắn. Các dự án giải trình tự hệ gen người, động vật, thực vật, vi sinh vật ở quy mô lớn đã và đang được các tổ chức khoa học công nghệ ở

nhiều quốc gia, độc lập hoặc hợp tác triển khai thực hiện. Thông tin khổng lồ về hệ gen được lưu trữ và quản lý tại các trung tâm quốc tế và quốc gia về sinh học tính toán và tin sinh học, từ đó khai thác ứng dụng trong rất nhiều lĩnh vực quan trọng của đời sống xã hội.

Dự án hệ gen người (1990-2003)

Các cơ quan khoa học của nhiều nước, dẫn đầu là Viện Sức khỏe quốc gia và Bộ Năng lượng của Hoa Kỳ đã hợp tác thực hiện Dự án trong 13 năm, với chi phí 3-4 tỷ USD. Năm 1999, Công ty tư nhân về công nghệ sinh học Celera Genomics của Hoa Kỳ cũng triển khai Dự án xác định trình tự hệ gen người. Năm 2001, "bản nháp" trình tự hệ gen người (khoảng 3 tỷ bp) đã được 2 nhóm đồng thời công bố (IHGSC, 2001; Venter *et al*., 2001). Dữ liệu trình tự hoàn chỉnh của hệ gen người được lưu trữ trên cơ sở dữ liệu của Viện Nghiên cứu hệ gen người quốc gia (Hoa Kỳ), cho phép các nhà khoa học trên toàn cầu truy cập phục vụ các nghiên cứu y sinh (www.genome.gov).

Dự án 1.000 hệ gen người (2008-2015)

Nhằm xác định kiểu gen và các đa hình di truyền với tần suất xuất hiện tối thiểu là 1% trong quần thể người nghiên cứu, dự án đầu tiên xác định trình tự hệ gen trên quy mô lớn tới 1.000 cá thể đã được cộng đồng khoa học quốc tế thực hiện và dữ liệu của dự án đã được chia sẻ miễn phí cho cộng đồng khoa học trên toàn cầu (https://www.internationalgenome.org/; Birney, Soranzo, 2015).

Dự án 100.000 hệ gen người (2012-2018)

Chính phủ Vương quốc Anh đã tiến hành Dự án giải trình tự toàn bộ 100.000 hệ gen của các bệnh nhân từ Dịch vụ Y tế quốc gia bị mắc bệnh hiếm hoặc ung thư. Các kết quả khám bệnh và dữ liệu hệ gen thu được từ Dự án năm 2018 là nền tảng phát triển dịch vụ y học hệ gen - phương thức chăm sóc, chẩn đoán và điều trị tiên tiến cho các bệnh nhân (https://www.genomicsengland.co.uk/).

Dự án 10.000 hệ gen động vật có xương sống (2009)

Dự án được thực hiện bởi mạng lưới các nhà sinh học và hệ gen học nhằm xác định và phân tích trình tự toàn bộ hệ gen của 10.000 loài động vật có xương sống góp phần tìm hiểu sự phức tạp của sự sống các loài động vật thông qua những thay đổi ở mức độ gen. Đây là một phần quan trọng của Dự án quốc tế về hệ gen động vật có xương sống, hướng tới giải trình tự 66.000 loài (https://genome10k.soe.ucsc.edu/).

Dự án quốc tế về hệ gen động vật có xương sống

Mục tiêu của dự án là xác định trình tự hoàn chỉnh với chất lượng cao và chú giải hệ gen của tất cả 66.000 loài động vật có xương sống trên trái đất phục vụ các nghiên cứu cơ bản về sinh học, bệnh học và bảo tồn. Dự án đã công bố 15 hệ gen tham chiếu chất lượng cao của 14 loài đại diện cho các lớp: động vật có vú, chim, bò sát, lưỡng cư và cá. Các dữ liệu gen được lưu trữ và chia sẻ cho cộng đồng khoa học thông qua hệ thống dữ liệu hệ gen mở Genome Ark - một thư viện số mới được xây dựng bởi Mạng lưới G10K-VGP với sự tham gia của hơn 150 chuyên gia đến từ 12 quốc gia, trên 50 viện nghiên cứu, trường đại học, công ty, phục vụ nhận dạng và bảo tồn nguồn gen của các loài có nguy cơ tuyệt chủng (https://vertebrategenomesproject.org/).

Dự án 1.000 hệ gen phiên mã và phát sinh chủng loại của thực vật

Trong khuôn khổ của Chương trình xác định trình tự 1.000 hệ gen phiên mã thực vật, 1.124 loài đại diện cho sự đa dạng của thực vật đã được giải trình tự hệ gen phiên mã phục vụ các nghiên cứu về tiến hóa ở thực vật (One Thousand Plant Transcriptomes Initiative, 2019).

Dự án 10.000 hệ gen thực vật (2017-2022)

Dự án nhằm xây dựng dữ liệu lớn về hệ gen thực vật phục vụ các nghiên cứu tiến hóa. Các tổ chức tài trợ chính bao gồm Viện Nghiên cứu hệ gen Bắc Kinh ở Thâm Quyến (Beijing Genome Institute - BGI-Thâm Quyến) và Ngân hàng Gen quốc gia Trung Quốc (China National Gene Bank - CNGB). Dự án này là một phần quan trọng của Dự án Hệ gen sinh vật toàn cầu (Earth BioGenome Project - EBP), với mục tiêu thu

được các trình tự thô của ít nhất 1,5 triệu loài sinh vật nhân thực (https://db.cngb.org/10kp/).

Dự án 1 triệu hệ gen vi sinh vật

Viện Nghiên cứu hệ gen Bắc Kinh (www.genomics.cn) hợp tác với các viện nghiên cứu, trường đại học, công ty đầu ngành ở Trung Quốc triển khai dự án giải trình tự hệ gen vi sinh vật nhằm tìm hiểu nguồn gen vi sinh vật đa dạng của quốc gia (https://en.genomics.cn/en-project-wswyj-1778.html).

Dự án 100.000 hệ gen mầm bệnh vi sinh vật

Bắt đầu từ 2012, Dự án do Bart Weimer (Trường Đại học California, Davis, Hoa Kỳ) khởi xướng và phối hợp với Cục Quản lý Thực phẩm và Dược phẩm Hoa Kỳ đặt mục tiêu giải trình tự hệ gen của 100.000 vi sinh vật gây bệnh thực phẩm và tạo cơ sở dữ liệu hệ gen, phục vụ chăm sóc sức khỏe cộng đồng (https://100kgenomes.org/).

Dự án Hệ gen sinh vật toàn cầu

Với sự tham gia của mạng lưới chuyên gia quốc tế đến từ nhiều quốc gia và vùng lãnh thổ như Liên minh châu Âu, Hoa Kỳ, Australia, Nhật Bản, Trung Quốc, Brazil, Canada, Nam Phi, Dự án nhằm giải trình tự, lưu trữ và phân tích hệ gen của tất cả sinh vật nhân thực trên trái đất phục vụ nghiên cứu đa dạng sinh học (https://www.earthbiogenome.org/org).

KHAI THÁC VÀ ỨNG DỤNG DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT

Các công nghệ NGS hiện được ứng dụng rộng rãi trong nhiều dự án lớn nhằm nghiên cứu và xây dựng cơ sở dữ liệu hệ gen người và các sinh vật khác. Công nghệ này đã và đang tiếp tục phát triển, có những ảnh hưởng sâu rộng trong lĩnh vực sinh học phân tử và công nghiệp sinh học như cải tiến các công cụ tạo sinh vật biến đổi gen, phát triển nhiên liệu sinh học, thay đổi phương thức nuôi trồng, phát triển dược phẩm điều trị ung thư và các loại bệnh khác. Các dữ liệu hệ gen có được từ GWAS, WGS, WES, GBS… được ứng dụng trong rất nhiều ngành quan trọng, từ y dược học, nông - lâm nghiệp, tới an toàn thực phẩm, môi trường...

Trong lĩnh vực y dược học

NGS là một công cụ mạnh nhất cho phép phát hiện được các đột biến có tần suất xuất hiện thấp, các biến thể di truyền là các tác nhân gây bệnh di truyền đơn gen, bệnh phức tạp do đa gen, ung thư... Hiện nay, các dữ liệu trình tự toàn bộ hệ gen người ngày càng đóng vai trò quan trọng trong phát hiện các bệnh di truyền, xác định mối liên quan giữa ung thư và nguyên nhân gây bệnh, thúc đẩy nghiên cứu và ứng dụng y học chính xác trong chẩn đoán lâm sàng và điều trị, hỗ trợ kiểm soát bệnh, đáp ứng với thuốc, xác định các vi sinh vật gây bệnh truyền nhiễm ở người phục vụ chẩn đoán và sản xuất vaccine, phân tích so sánh ở mức độ hệ gen, nghiên cứu lịch sử di truyền, nguồn gốc tiến hóa của các chủng tộc, các quần thể người… (Wu *et al*., 2016; Bah *et al*., 2018; Nông Văn Hải, 2019).

Đối với các bệnh di truyền Mendel (những bệnh di truyền chủ yếu ở người gây ra bởi sự rối loạn của gen đơn), cơ sở dữ liệu lớn nhất OMIM cung cấp thông tin về khoảng 7.000 bệnh khác nhau, trong đó có khoảng 3.500 các rối loạn di truyền không rõ nguyên nhân (http://omim.org). Theo cách tiếp cận truyền thống, các gen là nguyên nhân gây bệnh di truyền được định vị dựa trên các phân tích liên kết, trong đó xác định các biến thể di truyền giữa hàng trăm vùng gen ứng viên và kiểu hình hay trạng thái bị bệnh. Sau đó, các gen này được giải trình tự sử dụng công nghệ Sanger và đánh giá sự biến đổi của trình tự (Botstain *et al*., 2003). Phương pháp này cho phép phát hiện được các gen là nguyên nhân gây ra một số bệnh và thường được sử dụng để phân tích từng đoạn gen đơn và hiệu chỉnh, đánh giá các biến thể di truyền được phát hiện từ công nghệ NGS. Hạn chế của phương pháp là cần nhiều thời gian cũng như nhân lực để phân tích gen lớn hay phân tích đồng thời nhiều gen (Ku *et al*., 2011). Trong những trường hợp này, cách tiếp cận hiệu quả và phổ biến hơn là khai thác dữ liệu giải trình tự hệ gen WGS hay WES và xác định các biến thể di truyền của các bệnh Mendel, trong đó có nhiều bệnh hiếm (Roach *et al*., 2010, Bamshad *et al*., 2011; Chitty *et al*., 2015). Với số

lượng hệ gen được xác định trình tự ngày một nhiều, ví dụ, Dự án 1.000-10.000 hệ gen người và các dự án khác, những thông tin về hệ gen, các đa hình di truyền ở người, tần suất xuất hiện các đa hình ngày càng được hiểu rõ và khai thác ứng dụng, phát triển các kit chẩn đoán...

Đối với các bệnh phức tạp hay di truyền đa nhân tố chịu ảnh hưởng bởi nhiều hơn một gen, phương pháp GWAS thường được sử dụng để phân tích nhiều vị trí trên hệ gen ở nhiều cá thể khác nhau của nhóm bệnh và nhóm chứng, xác định các kiểu gen có tương quan với bệnh. Hàng ngàn đa hình liên quan đến bệnh hoặc các tính trạng đã được xác định thông qua GWAS. Vì vậy, GWAS có thể được khai thác trong chăm sóc sức khỏe, cung cấp cho các cá nhân thông tin về rủi ro phát sinh bệnh. Dữ liệu GWAS về kiểu gen và kiểu hình của các loại bệnh (Database of Genotype and Phenotype - dbGaP) được lưu trữ trên cơ sở dữ liệu của Trung tâm Thông tin Công nghệ sinh học Quốc gia Hoa Kỳ (National Center for Biotechnology Information - NCBI). Cộng đồng các nhà khoa học trên toàn cầu có thể truy cập tại https://www.ncbi.nlm.nih.gov/gap/.

Là một loại bệnh do biến đổi gen phức tạp, hàng năm ung thư là nguyên nhân gây tử vong cho rất nhiều bệnh nhân trên thế giới. Nhiều tổ chức quốc tế đã rất quan tâm xác định nguyên nhân gây ung thư sử dụng các dữ liệu trình tự WES, như ung thư dạ dày (Wang *et al*., 2011), ung thư tiền liệt tuyến (Barbieri *et al*., 2012). Cơ sở dữ liệu COSMIC hiện nay là nơi tích hợp và lưu trữ nhiều nhất các đột biến tế bào sinh dưỡng được phát hiện từ hàng triệu mẫu bệnh nhân mắc ung thư. Đến 3/2021, số lượng đột biến được lưu trữ trên COSMIC là 10 triệu (https://cancer.sanger.ac.uk/cosmic). Ngoài ra, Hiệp hội Hệ gen ung thư quốc tế (International Cancer Genome Consortium - ICGC) nghiên cứu sự thay đổi gen ở nhiều loại ung thư khác nhau và xây dựng cơ sở dữ liệu toàn diện về các đột biến gen xuất hiện ở các khối u của hơn 50 loại và phân loại ung thư khác nhau (https://dcc.icgc.org/). Số lượng hệ gen ở các loại ung thư được xác định tăng dần thông qua phân tích trình tự hệ gen của các bệnh nhân ở quy mô lớn, các đột biến thuộc vùng gen không mang mã cũng được phát hiện ở nhiều loại ung thư (Weinhold *et al*., 2014).

Dữ liệu về hệ gen người còn được sử dụng trong phân tích mối tương quan di truyền giữa đa hình các vùng điều khiển được xem là tác nhân gây nên các bệnh ở người và mức độ biểu hiện của gen. Thông qua việc phân tích WGS hay GWAS và chú giải chức năng của hệ gen, tất cả các đa hình tồn tại trong hệ gen được phát hiện và là dữ liệu nguồn để phân tích các đa hình trên vùng điều khiển (Wu *et al*., 2016). Những năm gần đây, nhiều nghiên cứu tập trung đánh giá các locus liên quan với bệnh từ việc khai thác các dữ liệu GWAS. Nghiên cứu lập bản đồ các gen liên quan các tính trạng số lượng (quantitative trait loci - QTL) dựa trên dữ liệu WGS cũng được sử dụng phổ biến (Lappalainen *et al*., 2013). So với dữ liệu GWAS, dữ liệu WGS cho phép phát hiện nhiều đa hình trên hệ gen hơn, tương ứng hỗ trợ việc xác định mối tương quan di truyền hiệu quả hơn. Do dữ liệu cần xử lý rất lớn nên gần đây, các công cụ tăng tốc độ xử lý dữ liệu WGS đã được xây dựng (Chiang *et al*., 2014).

Đối với y học chính xác và dự đoán, dữ liệu hệ gen cũng được khai thác ứng dụng rất hiệu quả. Kiểu gen của từng cá nhân có thể được xác định từ dữ liệu hệ gen WGS hay WES... So sánh với thông tin đã công bố hoặc từ các cơ sở dữ liệu bệnh đã biết, các chuyên gia có thể biết được sự biểu hiện của các tính trạng và nguy cơ mắc một số bệnh. Những dự đoán bệnh sớm cho từng bệnh nhân cụ thể dựa trên thông tin di truyền của chính họ, đã giúp bác sĩ áp dụng cá thể hóa trong chẩn đoán và điều trị (Biesecker, 2013). Nhóm nghiên cứu tại Trung Quốc đã xây dựng cơ sở dữ liệu dbWGFP tổng hợp gần 8,58 tỷ các đa hình đơn nucleotide (SNP) dựa trên thông tin của WGS hay WES và dự đoán chức năng của chúng (dbWGFP: http://bioinfo.au.tsinghua.edu.cn/dbwgfp). Một ví dụ về ứng dụng của y học chính xác và dự đoán là việc lựa chọn thuốc phù hợp cho bệnh nhân với hiệu quả điều trị tối đa và hạn chế rủi ro gây ra bởi tác dụng phụ của thuốc ở mức tối thiểu, hoặc đưa ra liệu pháp riêng giúp từng bệnh nhân nhanh chóng hồi phục (Bellmunt *et al*., 2015). Ngày nay, y học chính xác hay y học cá thể hóa đang trở thành phương pháp tiên

tiến, hiện đại và phát triển rất mạnh trên toàn cầu.

Trong lĩnh vực nông - lâm nghiệp

Hơn một thập kỷ trở lại đây, các nghiên cứu hệ gen động vật, thực vật và vi sinh vật có những bước phát triển rất mạnh nhờ sử dụng nhiều công nghệ mới như WGS, RNA-seq, RAD-seq, xác định kiểu gen thông qua giải trình tự (genotyping by sequencing - GBS), microarray. Dữ liệu từ hệ gen tham chiếu, hệ gen phiên mã của các loài cho phép phát hiện chính xác với số lượng rất lớn các kiểu gen, xác định chức năng, vai trò điều khiển và mức độ biểu hiện của gen, nghiên cứu sự chống chịu của cây trồng, vật nuôi với các tác động của môi trường, tìm kiếm các chỉ thị phân tử liên quan đến các tính trạng hoặc bệnh cây trồng, vật nuôi phục vụ các chương trình chọn tạo giống chất lượng... (Kim *et al*., 2020; You *et al*., 2020). Đến 30/5/2021, 3.019 loài động vật, 701 loài thực vật, 30.478 loài vi khuẩn đã được giải trình tự hệ gen và lưu trữ trên cơ sở dữ liệu của NCBI (www.ncbi.nlm.nih.gov/genome/).

Đối với công tác chọn tạo giống năng suất, chất lượng và chống chịu được các tác nhân sinh học và phi sinh học, dữ liệu về hệ gen là nguồn thông tin hữu ích, mở ra những triển vọng mới trong phát triển các chỉ thị phân tử ứng dụng trong chọn tạo giống (marker assisted selection - MAS), cho phép xác định những vùng gen hay những gen quy định hoặc liên quan đến tính trạng quan tâm. Khác với phương pháp chọn tạo giống truyền thống phải đánh giá kiểu hình của một quần thể lớn và cả phả hệ nhằm phát hiện những cá thể chứa gen mục tiêu, quy trình chọn giống mới sử dụng chỉ thị phân tử chỉ tập trung vào những cá thể riêng biệt mang các chỉ thị liên kết với các gen quy định tính trạng quan tâm như sinh trưởng, kháng bệnh, chống chịu các điều kiện bất lợi của môi trường (hạn, mặn, lạnh, nhiễm bệnh…). Ở mức độ cao hơn, thông tin về hệ gen sẽ được sử dụng trong phương pháp chọn tạo giống có sự trợ giúp của gen (genome selection - GS) (Xue, 2020). Cụ thể, dữ liệu hệ gen tham chiếu với độ chính xác cao được sử dụng trong các nghiên cứu cấu trúc và chức năng của gen, hỗ trợ lắp ráp và chú giải các hệ gen của các loài tương tự, phát hiện số lượng lớn các chỉ thị phân tử và các gen mục tiêu, cũng như xác định các đa hình di truyền. Dữ liệu hệ gen phiên mã được khai thác để đánh giá sự biểu hiện gen ở các mô, các giai đoạn phát triển, trong các điều kiện sinh lý, bệnh lý và môi trường khác nhau nhằm xác định cơ chế phân tử, chức năng của các gen mục tiêu liên quan đến tính kháng với các điều kiện bất lợi sinh học và phi sinh học, tìm kiếm các chỉ thị phân tử phục vụ chọn tạo giống (Vlk, Řepková, 2017; Sudhagar *et al*., 2018). Ví dụ, sử dụng công nghệ RNA-seq, Garnica và đồng tác giả (2013) đã nghiên cứu mầm bệnh *Puccinia striiformis* gây hại nghiêm trọng cho lúa mì và xác định các gen liên quan phục vụ chọn tạo giống lúa mì kháng bệnh. Tang và đồng tác giả (2013) phân tích hệ gen phiên mã của cây bạch dương *Populus euphratica* ở các vùng khô hạn hoặc nửa khô hạn nhằm tìm kiếm các gen liên quan đến tính chịu hạn. Hệ gen của đậu tương đã được khai thác để khám phá chức năng của các nhân tố điều khiển NAC đặc hiệu thực vật trong quá trình phát triển và mất nước của cây (Le *et al*. 2011). Trong nghiên cứu tương tác giữa mầm bệnh và cây chủ, công nghệ SMRT đã được ứng dụng để giải trình tự hệ gen của vi khuẩn *Xanthomonas oryzae* và hệ gen phiên mã của cây lúa *Oryza sativa* (Wilkins *et al*., 2015). Phân tích hệ gen phiên mã của cá *Sparus aurata* cho phép xác định được 63.880 trình tự mang mã của 21.384 gen, trong đó có các gen liên quan đến sinh trưởng, tiêu hóa và phản ứng miễn dịch với ký sinh trùng (Calduch-Giner *et al*., 2013). Liu và đồng tác giả (2015) đã xác định được 18 chỉ thị SNP liên quan đến tính trạng kháng bệnh nhiễm khuẩn nước lạnh trên 7.849 SNP ở cá hồi vân.

Trong công tác quản lý dịch bệnh

Đối với công tác quản lý dịch bệnh ở người, cây trồng, vật nuôi, dữ liệu NGS góp phần phát hiện mầm bệnh, đặc biệt là các bệnh do vi sinh vật gây ra, phương thức lây truyền của tác nhân, nguy cơ bùng phát, qua đó kiểm soát sự xuất hiện và xác định cơ chế, nguồn lây lan của bệnh cũng như phát triển các phương pháp điều trị (Van Borm *et al*., 2014; Lefterova *et al*., 2015; Hadidi

et al., 2016; Berry *et al*., 2020; Chen *et al*., 2021; Shahid *et al*., 2021). Coronavirus mới SARS-CoV-2 (gây bệnh Covid-19) đã gây ra đại dịch trên toàn cầu với khả năng lây lan rất cao. Do sự phát triển rất nhanh của dịch bệnh, việc xác định trình tự gen thông qua NGS và khai thác dữ liệu hệ gen đóng vai trò quan trọng ở nhiều khía cạnh, góp phần cung cấp thông tin về nguồn gốc và cơ chế lây nhiễm của SARS-CoV-2 ở người. Các công nghệ giải trình tự metagenome và giải trình tự tế bào đơn cũng được áp dụng để nghiên cứu các rối loạn về vi sinh vật đường ruột và di truyền miễn dịch của bệnh nhân COVID-19 (Chen *et al*., 2021). Việc áp dụng các kỹ thuật giải trình tự này có thể có ý nghĩa trong việc tìm kiếm các vật chủ SARS-CoV-2 trung gian mới nhằm ngăn chặn sự lây truyền giữa các loài. Các thông tin này sẽ hỗ trợ phát triển phương pháp chẩn đoán SARS-CoV-2 và tìm kiếm phương thức điều trị mới. Della và đồng tác giả (2020) đã phát hiện các chủng virus Y và đánh giá hiệu quả phát hiện virus cùng các kiểu gen của virus gây bệnh trên khoai tây sử dụng công nghệ giải trình tự gen 3G nanopore. Cũng bằng công nghệ này, Fellers và đồng tác giả (2019) đã phát hiện các bệnh do virus ở lúa mì. Biek và đồng tác giả (2012) đã nghiên cứu sự lây truyền *Mycobacterium bovis* ở gia súc và các ổ bệnh trong tự nhiên sử dụng dữ liệu WGS của 31 mẫu thu thập từ 5 nông trại. NGS là công cụ hỗ trợ hiệu quả cho cuộc chiến của con người chống lại các trường hợp khẩn cấp về sức khỏe cộng đồng, dịch bệnh ở cây trồng, vật nuôi trong tương lai.

Trong lĩnh vực an toàn thực phẩm

Với các phương pháp truyền thống, để phát hiện và nhận dạng các mầm bệnh trong thực phẩm bị ô nhiễm cần tiến hành rất nhiều thử nghiệm, trong khi các kỹ thuật NGS cho phép phát hiện nhanh và đồng thời các mầm bệnh chỉ trong một lần chạy hay một phản ứng. Dữ liệu hệ gen của 100.000 vi sinh vật gây bệnh thực phẩm làn nguồn thông tin hữu ích trực tiếp hỗ trợ chăm sóc sức khỏe cộng đồng, phát hiện các mầm bệnh và sự bùng phát dịch bệnh, giúp truy xuất nguồn gốc mầm bệnh và phát triển các phương pháp chẩn đoán nhanh hơn (https://100kgenomes.org/). Lefébure và đồng tác giả (2010) đã sử dụng công nghệ NGS để nghiên cứu sự phức tạp của hệ gen và sự chuyển gen ngang của hai loài vi khuẩn *Campylobacter* spp gây ngộ độc thực phẩm. Mellmann và đồng tác giả (2011) đã sử dụng công nghệ NGS để nghiên cứu hệ gen vi khuẩn đường ruột *Escherichia coli* O104:H4 gây ngộ độc thực phẩm và bùng phát dịch ở người.

Trong lĩnh vực môi trường

Các nhà khoa học về sinh vật hoang dã đã kết hợp các nghiên cứu về sinh thái, tiến hóa và hệ gen học để khai thác các dữ liệu lớn về hệ gen, phục vụ nghiên cứu phát sinh chủng loại, phân tích mối quan hệ giữa vật chủ và mầm bệnh, phát hiện các con đường lây nhiễm, phát triển thuốc phòng trị bệnh, bảo tồn các hệ sinh thái (Tan *et al*., 2019). Sự bùng phát dịch và sự lây nhiễm các mầm bệnh có thể dẫn đến sự suy giảm nghiêm trọng của hệ sinh thái. Dữ liệu hệ gen là công cụ hiệu quả được sử dụng để giám sát, phát hiện và giảm thiểu tác động của mầm bệnh đến các quần thể sinh vật trong tự nhiên (Fitak *et al*., 2019). Ví dụ, năm 2011, nhiều chim két được phát hiện đã chết ở hai thành phố là Mannheim và Heidelberg nước Đức, dẫn đến sự suy giảm nghiêm trọng của chim két ở hai thành phố này và các vùng lân cận. Becker và đồng tác giả (2012) đã nghiên cứu và xác định virus Usutu gây bệnh cùng sự phát tán của mầm bệnh ở 6 loài chim két hoang dã và nuôi nhốt ở Đức. Đối với các mẫu môi trường, dữ liệu hệ gen cho phép khám phá đa dạng vi sinh vật không thông qua nuôi cấy, hiểu biết về các hệ thống sinh học phức tạp từ mức độ cá thể, đến quần thể và quần xã, sự tương tác của các loài trong môi trường cộng sinh và cạnh tranh (Joly, Faure, 2015).

Như vậy, có thể thấy những nghiên cứu về hệ gen và khai thác dữ liệu lớn của hệ gen đang là lĩnh vực khoa học công nghệ mới, phát triển rất nhanh, mạnh và sâu rộng ở nhiều quốc gia trên thế giới. Đây là cuộc cách mạng trong đổi mới công nghệ, là cơ sở khoa học cho sự phát triển bền vững của rất nhiều ngành liên quan.

NGHIÊN CỨU PHÁT TRIỂN DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT Ở MỘT SỐ QUỐC GIA TIÊU BIỂU

Trong gần hai thập kỷ trở lại đây, song song

với sự phát triển rất mạnh của các công nghệ giải trình tự gen thế hệ mới, hướng nghiên cứu cơ bản nhằm xác định trình tự toàn bộ hệ gen các loài sinh vật, xây dựng và khai thác ứng dụng dữ liệu lớn về hệ gen được sự quan tâm của rất nhiều quốc gia, khu vực trên thế giới và có những bước tiến vượt bậc. Từ 2013, chính phủ của hơn 14 quốc gia đã đầu tư trên 4 tỷ USD để triển khai các chương trình y học - hệ gen quốc gia, tập trung chủ yếu vào các bệnh hiếm và ung thư, hay tiến hành các dự án nghiên cứu hệ gen trong quần thể (Stark *et al*., 2019). Dự đoán đến 2025, trên 60 triệu bệnh nhân sẽ có trình tự hệ gen của riêng mình phục vụ các hoạt động chăm sóc sức khỏe cá nhân (Birney *et al*., 2017) và công nghệ NGS cùng hệ gen học, với các dữ liệu giải trình tự hàng triệu hệ gen, sẽ trở thành lĩnh vực công nghệ đột phá, làm thay đổi xã hội và đem lại lợi ích kinh tế rất lớn với hàng nghìn tỷ USD mỗi năm (https://www.mckinsey.com/). Mỗi quốc gia có những cách tiếp cận và đang ở những giai đoạn khác nhau trên con đường xây dựng và khai thác dữ liệu hệ gen. Một số quốc gia đang xây dựng cơ sở hạ tầng như các tiêu chuẩn chung cùng các nền tảng và chính sách chia sẻ dữ liệu, một số quốc gia mới khởi xướng chương trình hệ gen quốc gia, trong khi một số quốc gia khác đã triển khai nhiều chương trình và thu được những kết quả giá trị. Ví dụ, Vương quốc Anh đã hoàn thành Dự án 100.000 hệ gen và đưa vào khai thác dữ liệu phục vụ chăm sóc sức khỏe hàng ngày cho người dân. Các dự án tương tự có thể sẽ trở nên rất nhỏ so với dự án Y học chính xác của Trung Quốc, dự kiến thực hiện trong 15 năm, với hạn mức đầu tư 9,2 tỷ USD và đặt mục tiêu hoàn thành nhiệm vụ giải trình tự 100 triệu hệ gen vào năm 2030. Vương quốc Anh, Hoa Kỳ, Pháp, Australia, Trung Quốc, Nhật Bản là những quốc gia điển hình, từ rất sớm đã triển khai những dự án quy mô, xây dựng được các hệ thống dữ liệu hệ gen quốc gia, quốc tế và các công cụ khai thác dữ liệu hệ gen hoạt động hiệu quả (Stark *et al*., 2019).

Vương quốc Anh

Được xem là quốc gia đi tiên phong trong lĩnh vực nghiên cứu hệ gen, năm 2013 chính phủ đã thành lập Genomics England (GEL) với mức đầu tư 415 triệu USD và năm 2018, GEL đã hoàn thành việc giải trình tự 100.000 hệ gen từ các bệnh nhân, với trên 100 bệnh hiếm và 7 loại bệnh ung thư phổ biến cùng các thành viên của gia đình họ. GEL đã xây dựng cơ sở hạ tầng để thực hiện các dịch vụ giải trình tự hệ gen WGS bao gồm từ máy móc, đến các công cụ phân tích tin sinh học tiêu chuẩn, các trung tâm lưu trữ mẫu sinh học và quản lý dữ liệu. Mạng lưới các phòng thí nghiệm hệ gen quốc gia mới được thành lập và liên kết với GEL để nhận và chia sẻ cơ sở hạ tầng về tin sinh học và dữ liệu hệ gen WGS. Gần đây, ngành khoa học sự sống đã nhận được 92,5 triệu USD đầu tư từ Viện Nghiên cứu dữ liệu sức khỏe Vương quốc Anh để thực hiện Dự án giải trình tự 5.000.000 hệ gen trong vòng 5 năm tới (Stark *et al*., 2019).

Hoa Kỳ

Trung tâm Thông tin Công nghệ sinh học Quốc gia NCBI thúc đẩy sự phát triển khoa học và quản lý sức khỏe thông qua chia sẻ các thông tin di truyền và y sinh học. NCBI đã xây dựng các hệ thống cơ sở dữ liệu lớn và phức tạp cho phép lưu trữ số lượng khổng lồ các trình tự gen, protein của mọi loài sinh vật được cung cấp bởi các nhà khoa học trên toàn cầu và các công cụ tin sinh học hỗ trợ phân tích, khai thác thông tin nhằm tăng cường hiểu biết về vật chất di truyền của sinh vật và vai trò hay sự liên quan của chúng đối với sức khỏe và bệnh tật (www.ncbi.nlm.nih.gov). Năm 2016, Dự án nghiên cứu thuộc Chương trình Y học chính xác đã được khởi động nhằm thu thập dữ liệu từ tối thiểu 1 triệu người sinh sống ở Hoa Kỳ, hướng tới ứng dụng trong y học chính xác, chẩn đoán và điều trị các loại bệnh.

Pháp

Năm 2015, Thủ tướng đã thông qua Kế hoạch quốc gia về y học hệ gen đến 2025, trong đó đặt mục tiêu tích hợp y học hệ gen vào chăm sóc sức khỏe và xây dựng ngành công nghiệp y học - hệ gen quốc gia nhằm thúc đẩy đổi mới sáng tạo và phát triển kinh tế. Trung tâm Phân tích dữ liệu quốc gia đảm nhận việc lưu trữ và phân tích dữ liệu cũng như tương tác với các cơ sở dữ liệu quốc gia và quốc tế khác (Stark *et al*., 2019).

Liên minh châu Âu

Viện Nghiên cứu tin sinh học châu Âu (European Bioinformatics Institute - EBI) xây dựng cơ sở dữ liệu trình tự gen, protein và các công cụ tin sinh học cho phép các nhà khoa học trên toàn cầu truy cập và khai thác miễn phí (www.ebi.ac.uk).

Australia

Hiệp hội Sức khỏe - Hệ gen Australia (The Australian Genomics Health Alliance) đã kết nối hơn 80 tổ chức trong nước nhằm tích hợp dữ liệu y học hệ gen vào chăm sóc sức khỏe, trong đó tập trung vào bệnh hiếm và ung thư. Trung tâm Hệ gen học so sánh (Centre for Comparative Genomics – CCG) đã triển khai các nghiên cứu tin sinh học và hệ gen học so sánh giữa động vật và các tác nhân gây bệnh cho người, trên lúa mạch và các cây họ đậu…nhằm ứng dụng trong y học và nông nghiệp (http://ccg.murdoch.edu.au).

Trung Quốc

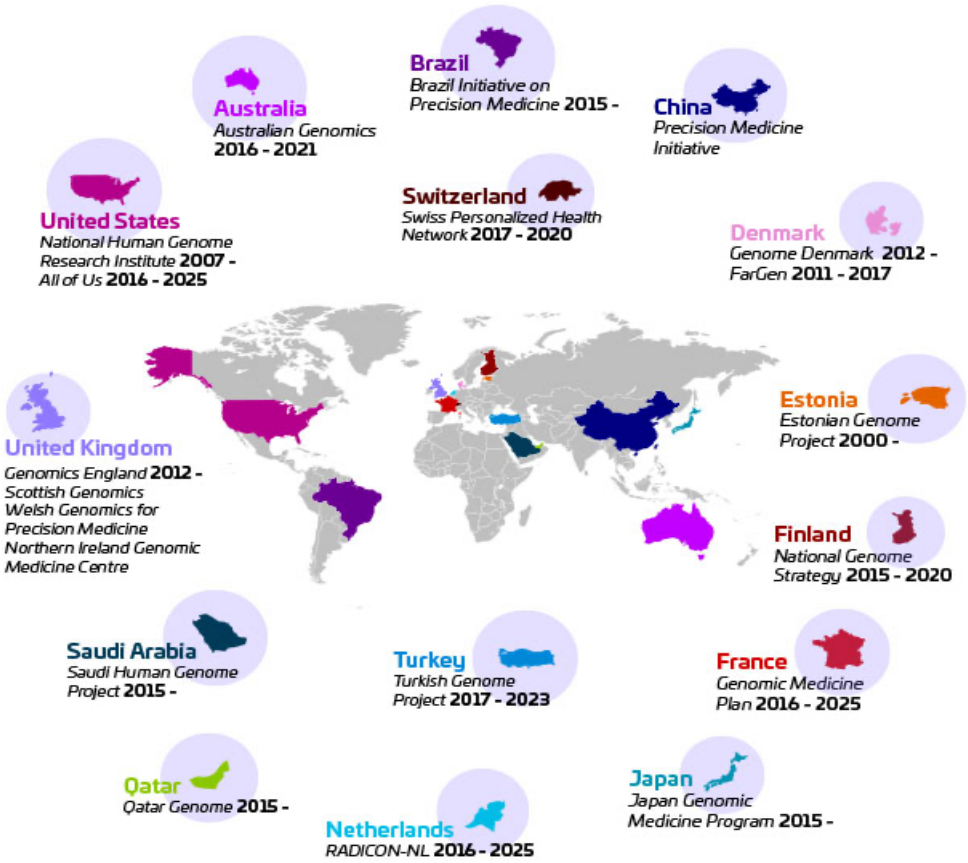
Là quốc gia sớm khởi động các hoạt động liên quan đến xây dựng, quản trị và khai thác dữ liệu lớn về hệ gen. Viện Nghiên cứu hệ gen học Bắc Kinh (Beijing Institute of Genomics - BIG) (www.big.cas.cn), Viện Hệ gen Bắc Kinh (www.genomics.cn), Trung tâm Hệ gen người quốc gia tại Thượng Hải (Chinese National Human Genome Center (http://chgc.sh.cn/)…là các đơn vị đã và đang thực hiện nhiều dự án nghiên cứu quan trọng của quốc gia và quốc tế liên quan đến xây dựng và khai thác dữ liệu gen, hệ gen sinh vật vào các lĩnh vực y dược học, nông nghiệp, môi trường, ví dụ: Dự án Hệ gen người đầu tiên, Dự án HapMap quốc tế, Dự án Hệ gen siêu lúa lai, Hệ gen tằm, Hệ gen virus SARS và phát triển các bộ KIT chẩn đoán, Hệ gen người châu Á đầu tiên, 100 hệ gen người Trung Quốc, 1.000 hệ gen người quốc tế, 1.000 hệ gen thực vật, 1.000 hệ gen động vật… Năm 2017, Trung Quốc xây dựng Dự án xác định hệ gen của 100.000 người. Với tài trợ từ Bộ Khoa học và Công nghệ, các nhà khoa học đã thiết lập dữ liệu hệ gen của dân tộc Hán và 9 dân tộc thiểu số khác để tìm hiểu thông tin di truyền trong gen và thu thập dữ liệu hệ gen của bệnh nhân nhằm làm rõ mối liên quan giữa gen và bệnh, ví dụ, tiểu đường. Viện Hàn lâm Khoa học Trung Quốc (Chinese Academy of Sciences - CAS) đã triển khai Dự án Y học chính xác quốc gia với mục tiêu hướng tới giải trình tự 100 triệu hệ gen vào năm 2030 (Stark *et al*., 2019).

Nhật Bản

Nhiều trung tâm thuộc các viện nghiên cứu/trường đại học như Trung tâm Y học hệ gen thuộc Viện Nghiên cứu lý hóa RIKEN (http://www.src.riken.jp/english/), Trung tâm Hệ gen người thuộc Đại học Tokyo (www.hgc.jp) đã tham gia các dự án giải trình tự hệ gen quốc gia và quốc tế như Dự án Hệ gen người đầu tiên, Dự án HapMap, Hệ gen đầu tiên người Nhật Bản, Nghiên cứu hệ gen học một số bệnh ung thư nhằm xác định các chỉ thị phân tử để chẩn đoán và điều trị. Nhật Bản cũng là một trong 3 quốc gia có cơ sở dữ liệu quốc tế về gen và protein lớn nhất thế giới (www.ddbj.nig.ac.jp). Năm 2015, Chương trình Y học hệ gen Nhật Bản được khởi xướng bởi Tổ chức Nghiên cứu phát triển và y học Nhật Bản (Japan Medical and Research Development Agency - AMED) nhằm chia sẻ thông tin về tần suất xuất hiện các allele và các đa hình liên kết với bệnh trong quần thể người Nhật Bản.

Hàn Quốc

Viện Dữ liệu lớn (Big Data Institute) của Hàn Quốc thuộc Đại học Quốc gia Seoul đã được thành lập vào năm 2014, liên kết khoảng 220 giáo sư người Hàn Quốc hoạt động trong lĩnh vực liên ngành này. Kể từ sau năm 2008, khi Hàn Quốc công bố hệ gen tham chiếu người Hàn đầu tiên, đến nay có nhiều hệ gen người đã được xác định trình tự và cơ sở dữ liệu đa hình hệ gen đã được xây dựng. Trong khuôn khổ Dự án Hệ gen người Hàn, đến 2020, 1094 hệ gen cá thể của người Hàn với các thông tin lâm sàng đã được công bố (Jeon *et al*., 2020).



Hình 1. Một số chương trình hệ gen quốc gia trên thế giới (https://www.bio-itworld.com/).

Như vậy, các quốc gia trên đều nhận thức rõ sự cần thiết và ưu tiên đầu tư cho dự án nghiên cứu và ứng dụng về hệ gen, đều có các trung tâm khoa học công nghệ chịu trách nhiệm xây dựng và quản lý dữ liệu hệ gen sinh vật. Một số quốc gia thành lập mạng lưới các trung tâm và thiết lập các cơ chế phối hợp hoạt động của các cơ quan này. Các dự án quốc gia được chính phủ tài trợ đóng vai trò quan trọng trong các nỗ lực toàn cầu nhằm phát triển, chia sẻ và khai thác dữ liệu, thông tin, kiến thức có được về hệ gen. Hiện nay, các thách thức trong xây dựng chiến lược, lộ trình chia sẻ công cụ, dữ liệu và các khung, tiêu chuẩn kỹ thuật quốc tế thống nhất cho các chương trình hệ gen đang được các quốc gia phối hợp giải quyết, hướng tới mục tiêu khai thác ứng dụng hiệu quả nguồn dữ liệu hệ gen khổng lồ trên quy mô toàn cầu.

TÌNH HÌNH NGHIÊN CỨU VÀ ỨNG DỤNG DỮ LIỆU LỚN VỀ HỆ GEN SINH VẬT Ở VIỆT NAM

Việt Nam đã rất chú trọng tới các chính sách tạo điều kiện cho sự phát triển của khoa học và công nghệ, trong đó công nghệ sinh học đã sớm được xác định là một trong bốn hướng công nghệ cần ưu tiên phát triển phục vụ công cuộc công nghiệp hóa, hiện đại hóa đất nước (Nghị quyết số 26/BCT). Các chương trình, đề án phát triển công nghệ sinh học các ngành y dược, nông nghiệp, thủy sản, công nghiệp sinh học ngành nông nghiệp, chế biến…được chính phủ phê duyệt trong những năm gần đây như đã góp phần thúc đẩy công nghệ sinh học phát triển, tăng cường ứng dụng các nghiên cứu về công nghệ sinh học vào nhiều lĩnh vực của đời sống xã hội, tăng cường vai trò của công nghệ sinh học đối với sự

phát triển của nền kinh tế. Trong khuôn khổ các dự án hợp tác quốc tế, các đề tài khoa học công nghệ thuộc các chương trình do Bộ Nông nghiệp và Phát triển Nông thôn (Bộ NN&PTNT), Bộ Khoa học và Công nghệ (Bộ KH&CN), Viện Hàn lâm Khoa học và Công nghệ Việt Nam (Viện HLKH&CNVN) quản lý như Chương trình Công nghệ sinh học Nông nghiệp-Thủy sản (CNSHNN-TS); Chương trình Bảo tồn và sử dụng bền vững nguồn gen đến năm 2015, định hướng đến năm 2030…, một số hệ gen người, động vật, thực vật, vi sinh vật đã và đang được các nhóm nghiên cứu chuyên sâu về gen, hệ gen tại các viện nghiên cứu, trường đại học và các đối tác quốc tế tiến hành giải trình tự thành công sử dụng các công nghệ NGS. Tuy nhiên, đối tượng nghiên cứu và số lượng hệ gen được xác định chưa nhiều cũng như cơ sở dữ liệu và việc khai thác dữ liệu còn rất hạn chế do hướng nghiên cứu khá mới, nguồn nhân lực, kinh phí nghiên cứu, cơ sở hạ tầng và trang thiết bị còn thiếu và chưa đồng bộ (Lê Thị Thu Hiền *et al*., 2016; Nông Văn Hải, 2019; Tran *et al*., 2021).

Trong lĩnh vực y dược

Ở quy mô hệ gen, ngay từ những năm 2008, 10 hệ gen ty thể hoàn chỉnh của người Việt Nam thuộc dân tộc Kinh, Tày, Mường đã được nhóm nghiên cứu tại Viện Công nghệ sinh học công bố (Trần Thị Minh Nguyệt *et al*., 2008). Năm 2015, hệ gen của 1 cá thể và của 1 gia đình (bộ ba) người Việt đã được nhóm nghiên cứu tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội giải trình tự và phân tích (Dang *et al*., 2014; 2015). Tại Viện Nghiên cứu hệ gen, năm 2018, hệ gen của 11 gia đình (bao gồm bố, mẹ và con), trong đó bố là nạn nhân phơi nhiễm dioxin đã được giải trình tự hoàn chỉnh sử dụng công nghệ NGS (Nguyen *et al*., 2018a). Cũng trong năm 2018, hệ gen của 10 cá thể người Việt Nam khỏe mạnh thuộc 3 gia đình và toàn bộ hệ gen ty thể cùng vùng không trao đổi chéo trên nhiễm sắc thể Y của hơn 600 cá thể thuộc 17 dân tộc với 5 nhóm ngôn ngữ đã được giải trình tự (Nguyen *et al*., 2018c). Đặc biệt, hướng khai thác dữ liệu giải trình tự gen, hệ gen và xác định các biến thể di truyền của nhiều bệnh, trong đó có các bệnh hiếm được nhiều nhóm tại các viện nghiên cứu, trường đại học triển khai thực hiện. Tại Viện Nghiên cứu hệ gen, toàn bộ exome ở một bệnh nhân Việt Nam mắc chứng rối loạn phổ tự kỷ đã được giải trình tự và phân tích, xác định được hai đột biến nhầm nghĩa mới (p.L111P và p.R3048C) trên gen *RYR3* (Nguyen *et al*., 2017). Các đột biến trên gen RB1 ở các bệnh nhân mắc u nguyên bào võng mạc đã được sàng lọc (Nguyen *et al*., 2018b). Tần số các allele *CYP2C19 *2*, *CYP2C19 *3* và *CYP2C19 *17* có liên quan đến hiệu quả sử dụng thuốc Clopidogrel trên 96 bệnh nhân mắc bệnh động mạch vành ở Việt Nam đã được khảo sát. Tỷ lệ bệnh nhân có kiểu hình chuyển hóa trung bình là 41,67%, số bệnh nhân có kiểu hình chuyển hóa kém chiếm 10,42%. Đặc biệt, nghiên cứu phát hiện 2 bệnh nhân (chiếm 2,08%) có kiểu gen dị hợp *CYP2C19 *1/*17* có khả năng chuyển hóa thuốc cực nhanh. Kết quả của nghiên cứu là tiền đề cho việc đưa ra liệu pháp chống ngưng tập tiểu cầu cá thể hóa ở Việt Nam dựa vào xét nghiệm di truyền (Nguyễn Hải Hà *et al*., 2020). Tại Viện Công nghệ gen và tế bào gốc Vinmec, hệ gen của 105 cá thể người Kinh không quan hệ họ hàng và hệ gen biểu hiện WES của 200 cá thể là bố mẹ của 100 trẻ tự kỷ đã được công bố trong năm 2019 (Le *et al*., 2019). Cơ sở dữ liệu các biến thể gen người Việt Nam cũng được xây dựng và phát triển (https://genomes.vn).

Viện Dữ liệu lớn VinBigdata thuộc Tập đoàn Vingroup đã công bố xây dựng và phát triển hệ thống quản lý và phân tích dữ liệu y sinh lớn nhất Việt Nam, phối hợp với 21 tổ chức nghiên cứu uy tín trên thế giới và trong nước thuộc lĩnh vực y học chính xác, đầu tư triển khai các dự án xây dựng hệ gen tham chiếu cho người Việt, giải trình tự hệ gen của hơn 1.000 người Việt nhằm nghiên cứu về các đặc điểm di truyền quần thể người Việt, phục vụ chăm sóc sức khỏe người Việt thông qua các giải pháp dự đoán nguy cơ bệnh và đáp ứng thuốc dựa trên hệ gen, tìm ra các phác đồ điều trị chuẩn xác (https://genome.vinbigdata.org).

Trong lĩnh vực nông - lâm nghiệp

Trong khuôn khổ các chương trình khoa học và công nghệ các cấp, một số đề tài thực hiện giải

trình tự một phần hoặc toàn bộ hệ gen ở các loài thực vật, động vật, vi sinh vật đã được thực hiện, với sự phối hợp của các nhóm nghiên cứu chủ yếu đến từ các đơn vị thuộc Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Bộ Nông nghiệp và Phát triển Nông thôn, như: (1) Nghiên cứu giải mã hệ gen một số giống lúa địa phương của Việt Nam (Viện Di truyền nông nghiệp chủ trì): Lần đầu tiên tại Việt Nam, hệ gen của 36 giống lúa đã được giải trình tự hoàn chỉnh, mở ra hướng nghiên cứu về hệ gen học và ứng dụng tin sinh học để khai thác dữ liệu hệ gen phục vụ công tác nghiên cứu và chọn tạo giống lúa. Thông tin có được từ dự án là nguồn vật liệu có giá trị để tầm soát các gen chức năng như kháng rầy nâu, đạo ôn, bạc lá, chịu hạn, chịu mặn, gen chất lượng, gen thơm; định vị chính xác các gen đích trên bản đồ, thiết kế các chỉ thị chức năng là những chỉ thị liên kết chặt với các gen đích giúp chọn lọc cá thể mang gen đích một cách chính xác phục vụ công tác lai tạo giống (https://most.gov); (2) Xác định các QTL mới liên quan đến tính trạng thiếu nước trong giai đoạn sinh dưỡng ở các giống lúa Việt Nam sử dụng công nghệ GWAS (Hoang *et al.,* 2019); (3) Giải trình tự hệ gen lục lạp của sâm Ngọc Linh và các loài thuộc chi Nhân sâm (Viện Nghiên cứu hệ gen): Nhằm phân tích và khai thác cơ sở dữ liệu của toàn bộ hệ gen lục lạp trong nghiên cứu phát sinh chủng loại, quá trình thích nghi, nhận dạng loài phục vụ giám sát thương mại cũng như góp phần định hướng ứng dụng trong giám định chất lượng sâm Ngọc Linh và các loài thuộc chi Nhân sâm ở Việt Nam, hệ gen lục lạp của sâm Ngọc Linh và các loài khác thuộc chi Nhân sâm như sâm Vũ diệp (*Panax bipinnatifidus*), Tam thất hoang (*Panax stipuleanatus*), sâm Nghệ An (*Panax* sp. puxailaileng) đã được giải trình tự, phân tích và chú giải thành công sử dụng công nghệ giải trình tự gen thế hệ mới. Trên cơ sở phân tích và so sánh, 04 chỉ thị có tiềm năng làm mã vạch phân tử cho phân loại sâm Ngọc Linh và các loài khác thuộc chi Nhân sâm đã được phát hiện (Manzanilla *et al*., 2018); (4) Giải trình tự và phân tích hệ gen phiên mã của sâm Ngọc Linh (Viện Nghiên cứu hệ gen): Sử dụng các công nghệ giải trình tự gen NGS, đề tài tập trung đánh giá đa dạng di truyền các quần thể sâm Ngọc Linh ở các khu vực phân bố thuộc tỉnh Quảng Nam và Kon Tum; giải trình tự, phân tích và xây dựng dữ liệu hệ gen phiên mã đặc thù mô và các giai đoạn phát triển khác nhau của sâm Ngọc Linh; xác định các gen tham gia chuỗi sinh tổng hợp ginsenoside và ginsenoside đặc thù Sâm Ngọc Linh phục vụ bảo tồn và phát triển bền vững nguồn gen sâm Ngọc Linh quý hiếm; (5) Giải trình tự hệ gen loài vi tảo biển dị dưỡng của Việt Nam *Schizochytrium mangrovei* PQ6 (Viện Công nghệ sinh học): Trình tự toàn bộ hệ gen 59,97 Mb và hệ gen phiên mã 20,7 Mb của loài *S. mangrovei* PQ6 cùng dữ liệu về các gen tham gia vào con đường sinh tổng hợp các chất quan trọng đã được xác định (Nguyễn Văn Lâm *et al*., 2015); (6) Nghiên cứu giải trình tự một phần bộ gen và xây dựng cơ sở dữ liệu genome tôm sú (Viện Nghiên cứu hệ gen): Đây là đề tài nghiên cứu về hệ gen loài thủy sản đầu tiên được khởi động ở Việt Nam, trong đó đã phát hiện mới một số cDNA mã hóa cho các protein quan trọng liên quan sinh trưởng và miễn dịch, giải trình tự hệ gen ty thể phục vụ nghiên cứu đa dạng di truyền. Dữ liệu giải trình tự được lưu trữ trên cơ sở dữ liệu hệ gen tôm sú và GenBank, là nguồn thông tin di truyền hữu ích phục vụ nghiên cứu các gen chức năng và chọn giống tôm sú; (7) Lập bản đồ bộ gen tôm sú (Viện Công nghệ sinh học); (8) Ứng dụng công nghệ sinh học trong chọn tạo giống tôm sú tăng trưởng nhanh (Viện Nghiên cứu nuôi trồng thủy sản II); (9) Nghiên cứu phát triển và ứng dụng chỉ thị phân tử để chọn tạo tôm chân trắng bố mẹ tăng trưởng nhanh (Viện Nghiên cứu nuôi trồng thủy sản III); (10) Phân tích hệ gen biểu hiện (exome + transcriptome) của cá tra nhằm phát triển chỉ thị phân tử phục vụ chọn giống cá tra theo hướng tăng trưởng (Viện Nghiên cứu hệ gen): Toàn bộ hệ gen của một cá thể cá tra đực đã được xác định và lắp ráp thành công. Hệ gen biểu hiện được chú giải. Các SNP ứng viên tiềm năng có sự khác biệt giữa nhóm cá tra sinh trưởng nhanh và sinh trưởng chậm đã được sàng lọc…(Kim *et al*., 2018).

Trong công tác quản lý dịch bệnh

Đối với công tác quản lý dịch bệnh trên người, cây trồng, vật nuôi, hệ gen của nhiều chủng vi sinh vật gây bệnh cũng được xác định

trình tự sử dụng công nghệ NGS phục vụ chẩn đoán và giám sát dịch bệnh. Ví dụ, hệ gen SARS-CoV-2 ở 44 bệnh nhân dương tính với virus tại Bệnh viện Nhiệt đới Trung ương đã được xác định và phân tích nhằm tìm kiếm các đa hình di truyền và quan hệ phát sinh chủng loại với các chủng trên thế giới (Nguyen *et al*., 2020). Toàn bộ hệ gen của chủng vi khuẩn *Neisseria meningitidis* B phân lập từ một đơn vị quân đội ở Việt Nam đã được xác định trình tự sử dụng công nghệ WGS và được phân tích đặc tính kháng kháng sinh, hỗ trợ nghiên cứu dịch tễ học và kháng kháng sinh cũng như giám sát bệnh viêm màng não ở Việt Nam (Tran *et al*., 2019). Các phân tích metagenomic được sử dụng để phát hiện virus gây bệnh trên tu hài (*Lutraria rhynchaena*) ở Việt Nam (Kim *et al*., 2020). Ngoài ra, các công nghệ NGS cũng được khai thác để tìm kiếm các chủng vi sinh vật hữu ích phục vụ công tác phòng trị bệnh, ví dụ, hệ gen WGS của chủng vi khuẩn *Bacillus thuringiensis* bản địa đã được xác định và dữ liệu hệ gen được phân tích nhằm sàng lọc các gen có hoạt lực diệt sâu đục quả đậu tương (Pham *et al*., 2021).

Trong lĩnh vực môi trường

Đánh giá đa dạng di truyền nhiều nhóm loài vi sinh vật trong các môi trường sinh thái khác nhau thông qua nuôi cấy hoặc không qua nuôi cấy đã được thực hiện dựa trên phân tích vùng gen 16S rDNA sử dụng công nghệ NGS (Tang *et al*., 2018).

Với nhiều ý nghĩa khoa học và thực tiễn, ở nước ta, hướng nghiên cứu hệ gen và xây dựng cơ sở dữ liệu hệ gen của các loài đã và đang được các nhà khoa học và các nhà quản lý quan tâm. Thông tin về hệ gen là nền tảng cho các nghiên cứu cơ bản và ứng dụng. Gần đây, các viện nghiên cứu và trường đại học cùng khối tư nhân đã hợp tác thực hiện các dự án nghiên cứu hệ gen và xây dựng cơ sở dữ liệu về hệ gen quy mô lớn. Đây là cơ sở để phát triển tiềm lực và tiếp cận được các thành tựu khoa học và công nghệ của thế giới, phục vụ phát triển kinh tế - xã hội của đất nước.

KẾT LUẬN

Hiện nay, nghiên cứu và khai thác dữ liệu lớn về hệ gen của sinh vật là một hướng khoa học và công nghệ chuyên sâu, có tác động đến nhiều lĩnh vực của đời sống xã hội. Các công nghệ giải trình tự gen thế hệ mới NGS cho phép tạo ra các nguồn dữ liệu khổng lồ về hệ gen của sinh vật. Nhờ các công cụ toán - tin - sinh, việc quản trị và phân tích dữ liệu hệ gen đã có thể thực hiện và được khai thác ứng dụng vào cuộc sống. Nhiều quốc gia trên thế giới đã ưu tiên đầu tư xây dựng và phát triển các cơ sở dữ liệu lớn về hệ gen. Việc thiết lập các mạng lưới gồm các viện nghiên cứu, trường đại học, công ty ở mỗi quốc gia và liên minh quốc tế nhằm phối hợp thực hiện các dự án nghiên cứu quy mô lớn, quản trị và khai thác hiệu quả nguồn dữ liệu khổng lồ về hệ gen, giải quyết các khó khăn cũng như xây dựng các tiêu chuẩn kỹ thuật quốc tế chung sẽ góp phần thúc đẩy sự phát triển ổn định và bền vững của lĩnh vực khoa học công nghệ hiện đại và hữu ích này.

Lời cảm ơn: *Công trình được thực hiện trong khuôn khổ đề tài: "Giải trình tự và phân tích hệ gen phiên mã (transcriptome) ở sâm Ngọc Linh (Panax vietnamensis* Ha et Grushv.*)", mã số: 16/17-HĐ-NVQG.*

TÀI LIỆU THAM KHẢO

Bah SY, Morang'a CM, Kengne-Ouafo JA, Amenga–Etego L and Awandare GA (2018) Highlights on the application of genomics and bioinformatics in the fight against infectious diseases: Challenges and opportunities in Africa. *Front Genet* 9: 575. doi: 10.3389/fgene.2018.00575.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: 745-755.

Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, *et al*. (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 44: 685-689.

Becker N, Jöst H, Ziegler U, Eiden M, Höper D, Emmerich P, Fichet-Calvet E, Ehichioya DU, Czajka C, Gabriel M, Hoffmann B, Beer M, Tenner-Racz K, Racz P, Günther S, Wink M, Bosch S, Konrad A, Pfeffer M, Groschup MH, Schmidt-Chanasit J (2012)

Epizootic emergence of Usutu virus in wild and captive birds in Germany. *PLoS One* 7(2): e32604. doi:10.1371/journal.pone.0032604.

Bellmunt J, Orsola A, Sonpavde G (2015) Precision and predictive medicine in urothelial cancer: Are we making progress? *Eur Urol* 68: 547-549.

Berry IM, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, Morton L, Jarman RG (2020) Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: Approaches, applications, and considerations for development of laboratory capacity. *J Infect Dis* 221(Suppl 3): S292-S307. doi: 10.1093/infdis/jiz286.

Biek R, O'Hare A, Wright D, Mallon T, McCormick C, Orton RJ, McDowell S, Trewby H, Skuce RA, Kao RR (2012) Whole genome sequencing reveals local transmission patterns of *Mycobacterium bovis* in sympatric cattle and badger populations. *PLoS Pathog* 8: e1003008.

Biesecker LG (2013) Hypothesis-generating research and predictive medicine. *Genome Res* 23: 1051-1053.

Birney E, Soranzo N (2015) The end of the start for population sequencing. *Nature* 526: 52-53.

Birney E, Vamathevan J, Goodhand P (2017) Genomics in healthcare: GA4GH looks to 2022. bioRxiv. https://doi.org/10.1101/203554.

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: pastsuccesses for Mendelian disease, future approaches for complex disease. *Nat Genet* 33: 228-237.

Calduch-Giner JA, Bermejo-Nogales A, Benedito-Palos L, Estensoro I, Ballester-Lozano G, Sitjà-Bobadilla, Pérez-Sánchez A*,* Pérez-Sánchez J (2013) Deep sequencing for *de novo* construction of a marine fish *(Sparus aurata)* transcriptome database with a large coverage of protein-coding transcripts. *BMC Genomics* 14: 178. https://doi.org/10.1186/1471-2164-14-178.

Chen X, Kang Y, Luo J, Pang K, Xu X, Wu J, Li X, Jin S (2021) Next-generation sequencing reveals the progression of COVID-19. *Front Cell Infect Microbiol* 11: 142. doi: 10.3389/fcimb.2021.632490.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM (2014) SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat Methods* 12: 966-968.

Chitty LS, Mason S, Barrett AN, McKay F, Lench N, Daley R, Jenkins LA (2015) Non-invasive prenatal diagnosis of achondroplasia and thanatophoric dysplasia: next-generation sequencing allows for a safer, more accurate, and comprehensive approach. *Prenat Diagn* 35(7): 656-662. doi:10.1002/pd.4583.

Dang TH, Nguyen DT, Pham TMT, Dang CC, Hoang KP, Pham PS, Le SV, Le SQ, Phan TTH, Do DD, Nguyen HD (2014) Prelimitary results on the whole genome analysis of a Vietnamese individual. *VNU Journal of Science: Comp Science & Com Eng* 30(3): 31-35.

Dang TH, Nguyen DT, Pham TMT, Le SQ, Phan TTH, Dang CC, Hoang KP, Nguyen HD, Do DD, Bui QM, Pham BS, Le SV (2015) Whole genome analysis of a Vietnamese trio. *J Biosci* 40(1): 113-124.

Della BM, Byrne S, Mullins E (2020) Characterization of potato virus Y isolates and assessment of nanopore sequencing to detect and genotype potato viruses. *Viruses* 12: 478. doi: 10.3390/v12040478.

Fellers JP, Webb C, Fellers MC, Shoup RJ, De Wolf E (2019) Wheat virus identification within infected tissue using nanopore sequencing technology. *Plant Dis* 103: 2199-2203. doi: 10.1094/pdis-09-18-1700-re.

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.

Fitak RR, Antonides JD, Baitchman EJ, Bonaccorso E, Braun J, Kubiski S, Chiu E, Fagre AC, Gagne RB, Lee JS, Malmberg JL, Stenglein MD, Dusek RJ, Forgacs D, Fountain-Jones NM, Gilbertson MLJ, Worsley-Tonks KEL, Funk WC, Trumbo DR, Ghersi BM, Grimaldi W, Heisel SE, Jardine CM, Kamath PL, Karmacharya D, Kozakiewicz CP, Kraberger S, Loisel DA, McDonald C, Miller S, O'Rourke D, Ott-Conn CN, Páez-Vacas M, Peel AJ, Turner WC, VanAcker MC, VandeWoude S, Pecon-Slattery J (2019) The expectations and challenges of wildlife disease research in the era of genomics: Forecasting with a horizon scan-like exercise. *J Hered* 110(3): 261-274. doi: 10.1093/jhered/esz001.

Garnica DP, Upadhyaya NM, Dodds PN, Rathjen JP **(**2013) Strategies for wheat stripe rust pathogenicity

identified by transcriptome sequencing. *PLoS One* 8: e67150.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6): 333-351.

Hadidi A, Flores R, Candresse T, Barba M (2016) Next-generation sequencing and genome editing in plant virology. *Front Microbiol* 7:1325. doi:10.3389/fmicb.2016.01325.

IHGSC (International Human Genome Sequencing Consortium) (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921. http://dx.doi.org/10.1038/35057062

Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, Jang J, Blazyte A, Kim C, Kim Y, Shim J, Kim N, Kim YJ, Park SG, Kim J, Cho YS, Park Y, Kim HM, Kim BC, Park NH, Shin ES, Kim BC, Bolser D, Manica A, Edwards JS, Church G, Lee S, Bhak J (2020) Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci Adv* 27: EAAZ7835. doi: 10.1126/sciadv.aaz7835.

Joly D, Faure D (2015) Next-generation sequencing propels environmental genomics to the front line of research. *Heredity* 114: 429-430. https://doi.org/10.1038/hdy.2015.23.

Kim KD, Kang Y, Kim C (2020) Application of genomic big data in plant breeding: Past, present, and future. *Plants (Basel)*. 9(11): 1454. doi:10.3390/plants9111454.

Kim TPO, Kagaya Y, Tran SH, Minei R, Tran THT, Duong TTH, Le TNB, Dang TL, Kinoshita K, Ogura A, Yura K (2020) A novel circular ssDNA virus of phylum Cressdnaviricota discovered in metagenomic data of otter clam (*Lutraria rhynchaena*). *Arch Virol* 165(12): 2921-2926. https://doi.org/10.1007/s00705-020-04819-9.

Kim TPO, Nguyen TP, Shoguchi E, Hisata K, Vo TBT, Inoue J, Shinzato C, Le TNB, Nishitsuji K, Knada M, Nguyen HV, Nong NV, Satoh N (2018) A draft genome of the striped catfish, *Pangasianodon hypophthalmus*, for comparative analysis of genes relevant to development and a resource for aquaculture improvement. *BMC Genomics* 19: 733. https://doi.org/10.1186/s12864-018-5079-x.

Ku CS, Naidoo N, Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. *Hum Genet* 129: 351-370.

Lappalainen T, Sammeth M, Friedländer MR, Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, *et al*. (2013) Transcriptome and genome sequencing uncovers functional variationin humans. *Nature* 501: 506-511.

Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K, Tran LS (2011) Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. *DNA Res* 18: 263-276.

Lê Thị Thu Hiền, Hugo De Boer, Vincent Manzanilla, Hà Văn Huân, Nông Văn Hải (2016) Giải mã hệ gen ở thực vật và các loài thuộc chi Nhân sâm (*Panax* L.). *Tạp chí Công nghệ Sinh học* 14(1): 1-13.

Le VS, Tran KT, Bui HTP, Le HTT, Nguyen CD, Do DH, Ly HTT, Pham LTD, Dao LTM, Nguyen LT (2019) A Vietnamese human genetic variation database. *Hum Mutat.* doi: 10.1002/humu.23835.

Lefebure T, Bitar PD, Suzuki H, Stanhope MJ (2010) Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol Evol* 2: 646-655.

Lefterova MI, Suarez CJ, Banaei N, Pinsky BA (2015) Next-generation sequencing for infectious disease diagnosis and management: A report of the association for molecular pathology. *J Mol Diagn* 17(6): 623-634. https://doi.org/10.1016/j.jmoldx.2015.07.004.

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.

Manzanilla V, Kool A, Nguyen NL, Nong VH, Le TTH, de Boer HJ (2018) Phylogenomics and barcoding of *Panax*: toward the identification of ginseng species. *BMC Evol Biol.* https://doi.org/10.1186/s12862-018-1160-y.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6(7): e22751. doi:

10.1371/journal.pone.0022751.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46.

Nguyen DT, Nakagawa H, Nguyen HH, Nguyen TD, Vu PN, Le TTH, Huynh TTH, Nguyen HH, Wong JH, Nakano K, Maejima, Sasaki-Oku A, Tsunoda T, Fujimoto A, Nong VH (2018a) Whole genome sequencing and mutation rate analysis of trios with paternal dioxin exposure. *Hum Mutat.* doi: 10.1002/humu.23585.

Nguyễn Hải Hà, Lê Thị Bích Thảo, Nguyễn Thị Thanh Hoa, Lê Thị Thu Hiền (2020) Nghiên cứu đa hình kiểu gen cyp2C19*2, *3 và *17 trên người Việt Nam mắc bệnh động mạch vành. *Tạp chí Công nghệ Sinh học* 18(1): 41-48.

Nguyen HH, Nguyen TTH, Vu PN, Le TQ, Pham MC, Ma THT, Do MH, Pham LBH, Nguyen DT, Le TTH, Nong VH (2018b) Mutational screening of germline *RB1* gene in Vietnamese patients with retinoblastoma reveals three novel mutations. *Mol Vis* 24: 231-238. http://www.molvis.org/molvis/v24/231.

Nguyen TD, Macholdt E, Nguyen DT, Arias L, Schröder R, Nguyen VP, Vo TBT, Nguyen HH, Huynh TTH, Nguyen TX, Kim TPO, Le TTH, Nguyen HH, Pakendorf B, Stoneking M, Nong VH (2018c) Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. *Sci Rep* 8: 11651. doi :10.1038/s41598-018-29989-0.

Nguyen TH, Nguyen TTN, Le BV, Thanh NM, Nguyen TKL, Nong VH, Nguyen HH (2017) Whole-exome sequencing identifies two novel missense mutations (p.L111P and p.R3048C) of RYR3 in a Vietnamese patient with autism spectrum disorders. Genes Genom 39: 301-306. https://doi.org/10.1007/s13258-016-0495-2.

Nguyen TT, Pham TN, Van TD, Nguyen TT, Nguyen DTN, Le HNM, Eden JS, Rockett RJ, Nguyen TTH, Vu BTN, Tran GV, Le TV, Dwyer DE, van Doorn HR; OUCRU COVID-19 Research Group (2020) Genetic diversity of SARS-CoV-2 and clinical, epidemiological characteristics of COVID-19 patients in Hanoi, Vietnam. *PLoS One* 15(11): e0242537. doi: 10.1371/journal.pone.0242537.

Nguyễn Văn Lâm, Phạm Quang Huy, Nguyễn Quốc Đại, Hoàng Minh Hiền, Đặng Diễm Hồng, Lê Văn Sơn, Chu Hoàng Hà, Trương Nam Hải, Nguyễn Cường (2015) Lắp ráp và chú giải hệ gen vi tảo biển dị dưỡng *Schizochytrium mangrovei* PQ6 của Việt Nam. *Bản B của Tạp chí Khoa học và Công nghệ Việt Nam 2*(6). https://b.vjst.vn/index.php/ban_b/article/view/742.

Nông Văn Hải (2019) Một số kết quả nghiên cứu gen và hệ gen người Việt Nam. Nhà xuất bản Khoa học Tự nhiên và Công nghệ.

One Thousand Plant Transcriptomes Initiative (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679-685. https://doi.org/10.1038/s41586-019-1693-2.

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. *Genomics* 93 (2): 105-111.

Pham LBH, Nguyen NL, Nguyen HH, Nguyen VD, Le TTH (2020) Genome sequence of a Vietnamese *Bacillus thuringiensis* strain TH19 reveals two potential insecticidal crystal proteins against *Etiella zinckenella* larvae. *Biol Control* 152, 104473.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and IlluminaMiSeq sequencers. *BMC Genomics* 13(1): 341.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328: 636-639.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74: 5463-5467.

Schadt EE, Turner S, Kasarskis A (2010) Window into third-generation sequencing. *Hum Mol Genet* 19 (R2): R227-240.

Shahid MS, Sattar MN, Iqbal Z, Raza A, Al-Sadi AM (2021) Next-generation sequencing and the CRISPR-Cas nexus: A molecular plant virology perspective. *Front Microbiol* 11: 609376. doi: 10.3389/fmicb.2020.609376.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135-1145.

Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, Levy Y, Glazer D, Wilson J, Lawler M, Boughtwood T, Braithwaite J, Goodhand P,

Birney E, North KN (2019) Integrating genomics into healthcare: A global responsibility. *Am J Hum Genet* 104(1): 13-20. doi: 10.1016/j.ajhg.2018.11.014.

Sudhagar A, Kumar G, El-Matbouli M (2018) Transcriptome analysis based on RNA-seq in understanding pathogenic mechanisms of diseases and the immune system of fish: A comprehensive review. *Int J Mol Sci* 19(1): 245. doi:10.3390/ijms19010245

Tan MP, Wong LL, Razali SA, Afiqah-Aleng N, Mohd Nor SA, Sung YY, Van de Peer Y, Sorgeloos P, Danish-Daniel M (2019) Applications of next-generation sequencing technologies and computational tools in molecular evolution and aquatic animal conservation studies: A short review. *Evol Bioinform Online* 15: 1176934319892284. doi: 10.1177/1176934319892284.

Tang S, Liang H, Yan D, Zhao Y, Han X, Carlson JE, Xia X, Yin W (2013) *Populus euphratica*: the transcriptomic response to drought stress. *Plant Mol Biol* 83: 539-557.

Tang TC, Phung DH, Bui VC, Nguyen NL, Nguyen NL, Nguyen SN, Nguyen QH, Le TTH (2018) Sequencing batch reactor and bacterial community in aerobic granular sludge for wastewater treatment of noodle-manufacturing sector. *Appl Sci*. https://doi.org/10.3390/app8040509.

Tran DK, Vu XD, Phi CN, Tran DX, Nguyen TT, Khuat HT, Dong HG, Nguyen HH, Tran HD, Do MT, Bui TMH (2021) Rice breeding in Vietnam: Retrospects, challenges and prospects. *Agriculture* 11(5): 397. https://doi.org/10.3390/agriculture11050397.

Trần Thị Minh Nguyệt, Lê Thị Bích Thảo, Bùi Thị Huyền, Phạm Đình Minh, Trần Thế Thành, Nguyễn Thị Tỵ, Nguyễn Bích Nhi, Đặng Diễm Hồng, Lê Quang Huấn, Quyền Đình Thi, Nguyễn Đăng Tôn, Nông Văn Hải, Phan Văn Chi (2008) Trình tự toàn bộ genome ty thể từ 9 cá thể người Việt Nam. *Tạp chí Công nghệ Sinh học* 6(4A): 569-578.

Tran TX, Le TT, Trieu LP, Austin CM, Dong VQ, Nguyen HM (2019) Whole-genome sequencing and characterization of an antibiotic resistant Neisseria meningitidis B isolate from a military unit in Vietnam. *Ann Clin Microbiol Antimicrob* 18(1):16. doi: 10.1186/s12941-019-0315-z.

Van Borm S, Belák S, Freimanis G, Fusaro A, Granberg F, Höper D, King DP, Monne I, Orton R, Rosseel T (2015) Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? *Methods Mol Biol* 1247: 415-436. doi:10.1007/978-1-4939-2004-4_30.

Venter JC, Adam MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA *et al*. (2001) The sequence of the human genome. *Science* 16: 1304-1351. http://www.sciencemag.org/content/291/5507/1304.full.

Vlk D, Řepková J (2017) Application of next-generation sequencing in plant breeding. *Czech J Genet Plant Breed*. doi: 10.17221/192/2016-CJGPB.

Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, Chan TL, Kan Z, Chan ASY, Tsui WY, Lee SP, Ho SL, Chan AKW, Cheng GHW, Roberts PC, Rejto PA, Gibson NW, Pocalyko DJ, Mao M, Xu J, Leung SY (2011) Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* 43: 1219-1223.

Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 46: 1160-1165.

Wilkins KE, Booher NJ, Wang L, Bogdanove AJ (2015) TAL effectors and activation of predicted host targets distinguish Asian from African strains of the rice pathogen *Xanthomonas oryzae* pv. oryzicola while strict conservation suggests universal importance of five TAL effectors. *Front Plant Sci* 6: 536.

Wu J, Wu M, Chen T, Jiang R (2016) Whole genome sequencing and its applications in medical genetics. *Quant Biol* 2016, 4(2): 115-128. doi: 10.1007/s40484-016-0067-0.

Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A (2020) Enhancing genetic gain through genomic selection: From livestock to plants. *Plant Commun*. https://doi.org/10.1016/j.xplc.2019.100005.

You X, Shan X, Shi Q (2020) Research advances in the genomics and applications for molecular breeding of aquaculture animals. Aquaculture 526. https://doi.org/10.1016/j.aquaculture.2020.735357.

GENOMICS AND BIG DATA: RESEARCH, DEVELOPMENT AND APPLICATIONS

Le Thi Thu Hien[1,2], Nguyen Tuong Van[3], Kim Thi Phuong Oanh[1,2], Nguyen Dang Ton[1,2], Huynh Thi Thu Hue[1,2], Nguyen Thuy Duong[1,2], Pham Le Bich Hang[1], Nguyen Hai Ha[1,2]

[1]*Institute of Genome Research, Vietnam Academy of Science and Technology*
[2]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology*
[3]*Institute of Biotechnology, Vietnam Academy of Science and Technology*

SUMMARY

Recent years, genomics and big data analytics have been widely applied and have significant impacts in various important areas of social life worldwide. The development of the next-generation sequencing (NGS) technologies, such as whole-genome sequencing (WGS), whole-exome sequencing (WES), transcriptome, and/or targeted sequencing, has enabled quickly generating the genomes of interested living organisms. Around the world many nations have invested in and promoted the development of genomics and big data analytics. A number of well-established projects on sequencing of human, animal, plant, and microorganism genomes to generate vast amounts of genomic data have been conducted independently or as collaborative efforts by national or international research networks of scientists specializing in different technical fields of genomics, bioinformatics, computational and statistical biology, automation, artificial intelligence, etc. Complicated and large genomic datasets have been effectively established, storage, managed, and used. Vietnam supports this new field of study through setting up governmental authorized institutions and conducting genomic research projects of human and other endemic organisms. In this paper, the research, development, and applications of genomic big data are reviewed with focusing on: (i) Available sequencing technologies for generating genomic datasets; (ii) Genomics and big data initiatives worldwide; (iii) Genomics and big data analytics in selected countries and Vietnam; (iv) Genomic data applications in key areas including medicine for human health care, agriculture - forestry, food safety, and environment.

Keywords: exome, genome, genomic big data, next generation sequencing, transcriptome