

Chương 1

Lược sử sinh học phân tử và sự ra đời của hệ gen học tính toán

Tháng 4 năm 1953, tạp chí *Nature* đăng một bài báo dài chưa đến hai trang của hai nhà khoa học trẻ ở Cambridge. Bài báo không có số liệu thí nghiệm nào do chính họ đo được, chỉ có một mô hình, một hình vẽ đơn giản và vài lập luận ngắn gọn. Tuy vậy, bài báo này đã đặt nền tảng cho sinh học phân tử hiện đại. Bẩy mươi năm sau, chúng ta có thể đọc toàn bộ hệ gen (*genome*) của một người trong vài ngày, với chi phí chỉ vài trăm đô la. Khó khăn giờ đây không còn nằm ở việc *đọc* DNA, mà ở việc *hiểu* những gì vừa đọc được. Mỗi người tạo ra hàng trăm gigabyte dữ liệu với hàng triệu vị trí khác biệt so với hệ gen tham chiếu (*reference genome*), trong khi chỉ một số ít trong đó có ý nghĩa sinh học. Hệ gen học tính toán (*computational genomics*) ra đời để giải quyết vấn đề này.

Chương này điểm lại lược sử sinh học phân tử (*molecular biology*) theo hai chặng. Chặng thứ nhất, từ khi *nuclein* được phát hiện năm 1869 đến khi mã di truyền (*genetic code*) được giải xong vào giữa thập niên 1960, trả lời câu hỏi vật chất di truyền (*genetic material*) là gì và hoạt động ra sao. Chặng thứ hai, từ phương pháp Sanger năm 1977 đến pangenome người năm 2023 (bộ hệ gen tham chiếu gồm hệ gen của nhiều người thay vì của một người), là lịch sử của việc đọc DNA ở quy mô ngày càng lớn. Trên nền đó, phần cuối chương giải thích vì sao hệ gen học tính toán ra đời, và những bài toán nào đang đặt ra cho lĩnh vực này hiện nay. Mục đích của chương không nằm ở các mốc lịch sử, mà ở việc cho thấy mỗi bước tiến trong sinh học lại kéo theo nhu cầu tính toán mới.

1.1 Phát hiện vật chất di truyền

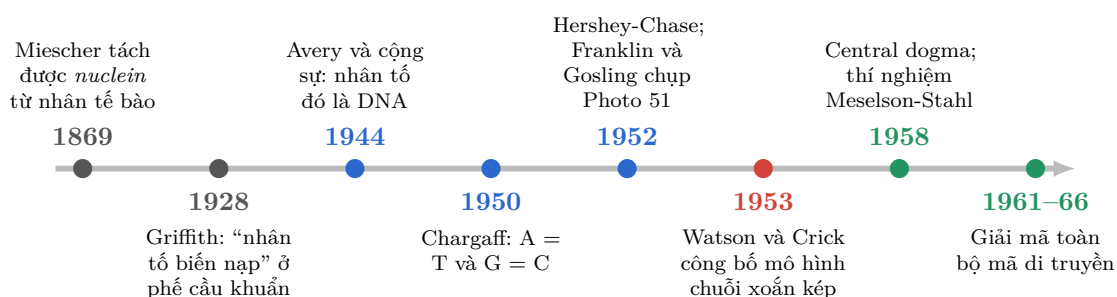
Năm 1866, Gregor Mendel đã chứng minh rằng tính trạng (*trait*) được truyền qua các thế hệ bằng những “nhân tố” riêng rẽ. Tuy nhiên, bản chất vật chất của các nhân tố này vẫn chưa được xác định trong gần một thế kỷ sau đó. Chỉ ba năm sau, vào năm 1869, Friedrich Miescher, một bác sĩ trẻ người Thụy Sĩ làm việc tại Tübingen, tách được từ nhân tế bào bạch cầu (lấy từ mủ trên băng gạc phẫu thuật) một chất giàu phosphor mà ông gọi là *nuclein*. Đó chính là DNA, dù phải mất thêm tám mươi năm nữa người ta mới hiểu vai trò của nó.¹ Như vậy, ngay từ cuối thế kỷ 19 đã có hai hướng nghiên cứu song song, một bên mô tả quy luật di truyền ở mức tính trạng quan sát được, một bên phân tích thành phần hóa học của tế bào. Câu hỏi nối hai hướng này lại, tức chất nào trong tế bào đóng vai trò “nhân tố” của Mendel, sẽ chi phối sinh học trong nhiều thập kỷ tiếp theo.

¹Dahm R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology* 278:274–288.
<https://doi.org/10.1016/j.ydbio.2004.11.028>

Trong nửa đầu thế kỷ 20, đa số các nhà khoa học tin rằng protein mới là vật chất di truyền. Quan điểm này có cơ sở, vì protein được xây dựng từ hai mươi loại amino acid (axit amin), còn DNA chỉ có bốn loại nucleotide, tức đơn vị cấu tạo của axit nucleic (*nucleic acid*), gồm một phân tử đường, một nhóm phosphate và một base. Một “bảng chữ cái” gồm bốn ký tự được cho là quá đơn điệu để mang toàn bộ thông tin của sự sống. Thêm vào đó, giả thuyết tetranucleotide của Phoebus Levene cho rằng bốn nucleotide lặp lại theo một khuôn mẫu cố định, khiến DNA nhìn giống một phân tử cấu trúc hơn là một vật mang thông tin. Để thay đổi một quan niệm đã phổ biến như vậy, cần những thí nghiệm tách riêng được vai trò của DNA và của protein, thay vì chỉ quan sát chúng cùng tồn tại trong tế bào. Quan niệm này dần bị bác bỏ qua ba thí nghiệm sau.

- **Griffith (1928)** phát hiện rằng vi khuẩn phế cầu *Streptococcus pneumoniae* chủng S (gây bệnh) đã bị giết bằng nhiệt vẫn có thể “biến nạp” (*transformation*) chủng R (không gây bệnh) thành chủng gây chết chuột. Kết quả này cho thấy có một “nhân tố biến nạp” được chuyển từ chủng S sang chủng R.²
- **Avery, MacLeod và McCarty (1944)** lần lượt loại bỏ từng thành phần của dịch chiết tế bào và chỉ ra rằng nhân tố biến nạp đó là DNA, không phải protein.³
- **Hershey và Chase (1952)** đánh dấu phóng xạ (*radioactive labeling*) riêng protein và DNA của thực khuẩn thể (*bacteriophage*), tức virus ký sinh trên vi khuẩn, và thấy rằng chỉ DNA đi vào tế bào vi khuẩn để tạo ra thế hệ virus mới.⁴

Song song đó, Erwin Chargaff (khoảng 1950) đo thành phần base của DNA (bốn loại base nitơ (*nitrogenous base*) A, T, G và C) từ nhiều loài và nhận thấy một quy luật, theo đó lượng A luôn gần bằng lượng T, và lượng G gần bằng lượng C. Tỷ lệ này thay đổi giữa các loài, qua đó bác bỏ giả thuyết tetranucleotide. Thật vậy, nếu bốn nucleotide lặp lại theo một khuôn mẫu cố định, tỷ lệ của chúng phải gần như giống nhau ở mọi sinh vật. Việc tỷ lệ này khác nhau giữa các loài cho thấy trình tự DNA có thể thay đổi, và vì vậy có khả năng mang thông tin riêng cho từng loài. Ngược lại, sự cân bằng giữa A với T và giữa G với C gợi ý rằng các base có quan hệ từng cặp với nhau, nhưng không ai biết quan hệ đó đến từ đâu. Quy luật này chưa được giải thích cho đến năm 1953.⁵ Hình 1.1 tóm tắt các mốc chính của giai đoạn này, từ khi phát hiện *nuclein* đến khi giải mã xong mã di truyền.



Hình 1.1: Các mốc chính trong việc xác định vật chất di truyền và cơ chế hoạt động của nó, từ *nuclein* (1869) đến mã di truyền (1966).

²Griffith F. (1928). *Journal of Hygiene* 27:113–159. <https://doi.org/10.1017/S0022172400031879>

³Avery O.T., MacLeod C.M., McCarty M. (1944). *Journal of Experimental Medicine* 79:137–158. <https://doi.org/10.1084/jem.79.2.137>

⁴Hershey A.D., Chase M. (1952). *Journal of General Physiology* 36:39–56. <https://doi.org/10.1085/jgp.36.1.39>

⁵Chargaff E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6:201–209. Thông tin trích dẫn: <https://historyofmedicine.com/id/8695>

1.2 Chuỗi xoắn kép

Đầu những năm 1950, cuộc đua giải cấu trúc DNA diễn ra giữa vài nhóm nghiên cứu. Tại King's College London, Rosalind Franklin, một chuyên gia về nhiễu xạ tia X (*X-ray diffraction*), cùng nghiên cứu sinh Raymond Gosling đã thu được những ảnh nhiễu xạ sắc nét nhất về DNA. Bức ảnh nổi tiếng nhất, được gọi là *Photo 51*, chụp tháng 5 năm 1952, cho thấy một hình chữ X đặc trưng của cấu trúc xoắn. Từ số liệu của mình, Franklin còn suy ra được rằng khung đường-phosphate (*sugar-phosphate backbone*) nằm ở phía ngoài phân tử. Cùng lúc đó ở Cambridge, James Watson và Francis Crick theo đuổi một cách tiếp cận khác, đó là dựng mô hình bằng que và tấm kim loại, thử các cách sắp xếp nguyên tử sao cho phù hợp với hóa học đã biết. Mô hình đầu tiên của họ năm 1951, gồm ba mạch với base quay ra ngoài, đã sai; chính Franklin là người chỉ ra điểm sai đó. Hai cách tiếp cận này bổ sung cho nhau, vì ảnh nhiễu xạ cho số liệu đo đạc trực tiếp nhưng khó diễn giải thành một cấu trúc cụ thể, còn mô hình que cho một cấu trúc cụ thể nhưng chỉ có giá trị khi khớp được với số liệu đo đạc.

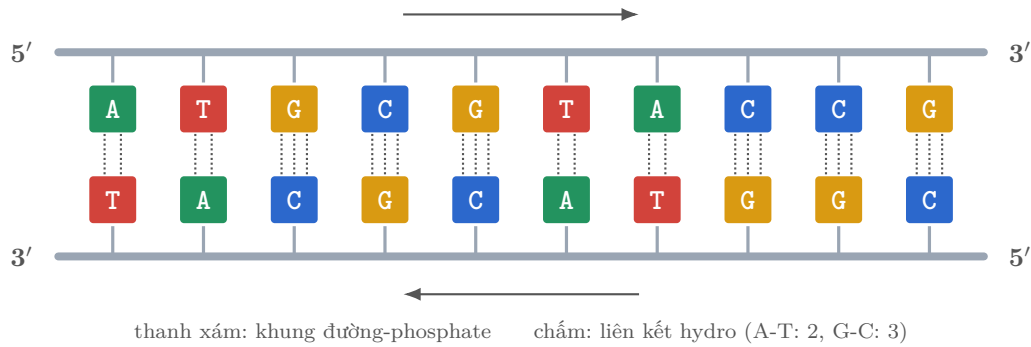
Tháng 1 năm 1953, Maurice Wilkins, đồng nghiệp của Franklin tại King's College, cho Watson xem *Photo 51* mà Franklin không hề hay biết.⁶ Watson và Crick cũng tiếp cận được một báo cáo nội bộ tóm tắt số liệu của nhóm King's. Hình chữ X trong *Photo 51* xác nhận DNA có dạng xoắn, còn việc khung đường-phosphate nằm ở phía ngoài buộc các base phải quay vào bên trong phân tử. Câu hỏi còn lại là các base bên trong được sắp xếp ra sao, và quy luật Chargaff cung cấp manh mối cho câu hỏi này. Kết hợp những số liệu này với quy luật Chargaff, họ nhanh chóng đi đến mô hình cuối cùng (Hình 1.2) với các đặc điểm chính sau.

- DNA gồm **hai mạch** xoắn quanh một trục chung, tạo thành **chuỗi xoắn kép** (*double helix*), chạy ngược chiều nhau (đối song song, **antiparallel**), một mạch theo hướng $5' \rightarrow 3'$, mạch kia theo hướng $3' \rightarrow 5'$.
- Khung đường-phosphate nằm bên ngoài; các base nằm bên trong và **bắt cặp bổ sung** (*complementary base pairing*), trong đó A bắt cặp với T bằng hai liên kết hydro (*hydrogen bond*) và G bắt cặp với C bằng ba liên kết hydro.
- Vì A luôn đi với T và G luôn đi với C, quy luật Chargaff được giải thích một cách tự nhiên.

Ngày 25 tháng 4 năm 1953, *Nature* đăng liên tiếp ba bài báo, gồm bài mô hình của Watson và Crick, và hai bài số liệu thực nghiệm của nhóm Wilkins và nhóm Franklin-Gosling. Bài của Watson và Crick kết thúc bằng một câu nhận xét nổi tiếng, trong đó các tác giả lưu ý rằng cách bắt cặp đặc hiệu này gợi ý ngay một cơ chế sao chép vật chất di truyền. Nói cách khác, cấu trúc phân tử đã gợi ý trực tiếp về chức năng của nó.⁷ Ý tưởng này khá trực quan, vì trình tự của một mạch xác định hoàn toàn trình tự của mạch kia, nên khi hai mạch tách rời nhau, mỗi mạch có thể làm khuôn để tổng hợp một mạch bổ sung (*complementary strand*) mới, và kết quả là hai phân tử DNA giống hệt phân tử ban đầu. Dự đoán này được kiểm chứng bằng thực nghiệm chỉ vài năm sau, như trình bày ở mục tiếp theo.

⁶King's College London. The story behind Photograph 51.
<https://www.kcl.ac.uk/the-story-behind-photograph-51>

⁷Watson J.D., Crick F.H.C. (1953). *Nature* 171:737–738. <https://doi.org/10.1038/171737a0>



Hình 1.2: Hai mạch DNA chạy ngược chiều và bắt cặp bổ sung (chuỗi xoắn được biểu diễn ở dạng duỗi thẳng). Trình tự của một mạch xác định hoàn toàn trình tự của mạch còn lại; đây cũng là nguyên lý của sao chép DNA và của hầu hết các kỹ thuật giải trình tự.

Có thể bạn chưa biết. *Rosalind Franklin và Photo 51*

Rosalind Franklin qua đời năm 1958 vì ung thư buồng trứng, khi mới 37 tuổi. Bốn năm sau, Watson, Crick và Wilkins nhận giải Nobel Sinh lý học hoặc Y học 1962.^a Giải Nobel không trao cho người đã mất, vì vậy không thể biết ủy ban sẽ quyết định ra sao nếu bà còn sống. Trong nhiều thập kỷ, vai trò của Franklin bị đánh giá thấp; ngày nay bà được ghi nhận là người đóng góp số liệu then chốt cho một trong những khám phá lớn nhất thế kỷ 20. Câu chuyện này cũng là lời nhắc về đạo đức trong chia sẻ và ghi nhận dữ liệu, một chủ đề sẽ được đề cập ở phần sau của sách.

^aArizona State University, Embryo Project Encyclopedia. Photograph 51, by Rosalind Franklin (1952). <https://embryo.asu.edu/pages/photograph-51-rosalind-franklin-1952>

1.3 Central dogma của sinh học phân tử

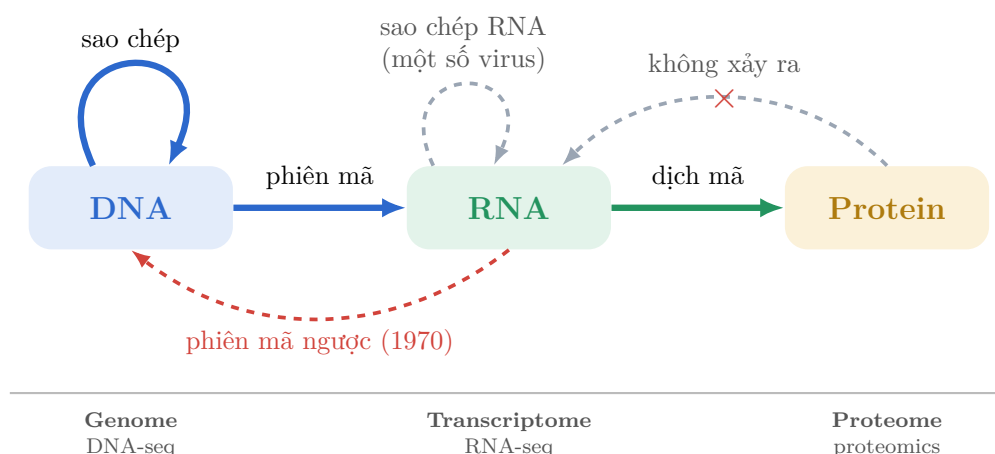
Cấu trúc DNA mới chỉ trả lời một nửa câu hỏi. Nửa còn lại là cơ chế nào giúp một chuỗi gồm bốn loại nucleotide quyết định hình dạng của tế bào, màu mắt hay nguy cơ mắc bệnh. Để trả lời, cần biết thông tin chứa trong trình tự DNA được chuyển sang những phân tử nào, theo chiều nào, và bằng cơ chế gì. Năm 1958, Francis Crick đề xuất một khung lý thuyết mà ông gọi là **central dogma**, trong sách giáo khoa tiếng Việt thường được dịch là học thuyết trung tâm. Theo đó, thông tin di truyền được sao chép (*replication*) từ DNA sang DNA, được **phiên mã** (*transcription*) từ DNA sang RNA, rồi được **dịch mã** (*translation*) từ RNA sang protein. Điểm cốt lõi trong phát biểu của Crick là một khi thông tin đã đi vào protein, nó không thể quay ngược trở lại axit nucleic, tức DNA và RNA. Như vậy, DNA là nơi thông tin được lưu trữ và truyền lại cho thế hệ sau, RNA là bản sao trung gian, còn protein là nơi phần lớn thông tin đó được thể hiện thành chức năng trong tế bào.

Những năm sau đó, các bằng chứng thực nghiệm lần lượt củng cố khung lý thuyết này. Cũng năm 1958, thí nghiệm của Meselson và Stahl chứng minh DNA sao chép theo kiểu bán bảo toàn (*semi-conservative*), nghĩa là mỗi phân tử DNA con mang một mạch cũ và một mạch mới, đúng như mô hình chuỗi xoắn kép dự đoán.⁸ Năm 1961, Marshall Nirenberg và Heinrich Matthaei cho một đoạn RNA nhân tạo chỉ gồm U vào hệ dịch mã ngoài tế bào (*cell-free translation system*) và thu được một chuỗi polypeptide chỉ gồm phenylalanine, chứng tỏ bộ ba (*codon*) UUU mã hóa phenylalanine. Mỗi bộ ba là một nhóm ba nucleotide liên tiếp

⁸Meselson M., Stahl F.W. (1958). *PNAS* 44:671–682. <https://doi.org/10.1073/pnas.44.7.671>

trên RNA, tương ứng với một amino acid hay một tín hiệu dừng dịch mã.⁹ Đến năm 1966, toàn bộ 64 bộ ba của mã di truyền đã được giải. Con số 64 đến từ việc mỗi bộ ba gồm ba vị trí, mỗi vị trí có thể là một trong bốn nucleotide, nên có $4^3 = 64$ tổ hợp. Số tổ hợp này nhiều hơn hẳn hai mươi loại amino acid, vì vậy phần lớn amino acid được mã hóa bởi nhiều hơn một bộ ba, và một số bộ ba là bộ ba kết thúc (*stop codon*), báo hiệu dừng dịch mã.

Năm 1970, Howard Temin và David Baltimore độc lập phát hiện enzyme phiên mã ngược (*reverse transcriptase*) ở retrovirus, nhóm virus có vật chất di truyền là RNA và nhân lên bằng cách chuyển RNA đó thành DNA rồi chèn vào hệ gen của tế bào chủ, cho phép thông tin đi từ RNA ngược về DNA.¹⁰ Nhiều người coi đây là một “ngoại lệ” của central dogma. Crick sau đó viết một bài trên *Nature* để làm rõ rằng phát biểu ban đầu của ông chưa bao giờ cấm dòng RNA→DNA; điều bị loại trừ là dòng thông tin từ protein ngược về axit nucleic, và cho đến nay điều đó vẫn đúng.¹¹ Phát hiện này cũng có ý nghĩa thực tiễn lớn đối với hệ gen học. Ngày nay, enzyme phiên mã ngược là công cụ quen thuộc trong phòng thí nghiệm, vì trong hầu hết các quy trình giải trình tự RNA (*RNA sequencing*, *RNA-seq*), RNA được chuyển thành cDNA (*complementary DNA*), tức DNA bổ sung được tổng hợp từ khuôn RNA, trước khi đưa vào máy giải trình tự.



Hình 1.3: Central dogma. Mũi tên liền biểu diễn các dòng thông tin phổ biến trong mọi tế bào, mũi tên đứt biểu diễn các dòng đặc biệt. Hàng dưới cho biết loại dữ liệu omics tương ứng với mỗi phân tử.

Có thể bạn chưa biết. Nguồn gốc tên gọi *central dogma*

Nhiều năm sau, Crick kể lại rằng nhà sinh học Jacques Monod từng nhận xét ông không dùng đúng nghĩa của chữ “dogma” (một niềm tin không được phép nghi ngờ). Crick giải thích rằng ông chỉ muốn nói đến một giả thuyết lớn, chưa có nhiều bằng chứng trực tiếp, và lẽ ra có thể gọi đó là “central hypothesis”.^a Dù cách dùng từ không chính xác, tên gọi này vẫn được giữ đến ngày nay.

^aCobb M. Francis Crick was misunderstood. *Asimov Press*. <https://press.asimov.com/articles/crick>. Câu chuyện cũng được Crick kể trong hồi ký *What Mad Pursuit* (1988).

Với người làm hệ gen học tính toán, central dogma không chỉ là kiến thức sinh học nền tảng. Nó còn là một *bản đồ dữ liệu*, trong đó mỗi phân tử trong Hình 1.3 là một tầng thông

⁹Nirenberg M.W., Matthaei J.H. (1961). *PNAS* 47:1588–1602. <https://doi.org/10.1073/pnas.47.10.1588>

¹⁰Temin H.M., Mizutani S. (1970). *Nature* 226:1211–1213, <https://doi.org/10.1038/2261211a0>; Baltimore D. (1970). *Nature* 226:1209–1211; thông tin trích dẫn: <https://www.historyofmedicine.com/id/3429>

¹¹Crick F. (1970). Central dogma of molecular biology. *Nature* 227:561–563. <https://doi.org/10.1038/227561a0>

tin có thể đo đạc. Mỗi tầng tương ứng với một loại dữ liệu **omics**, tức dữ liệu đo đồng thời toàn bộ một loại phân tử trong mẫu thay vì từng phân tử riêng lẻ. Giải trình tự DNA cho ta dữ liệu **hệ gen**, giải trình tự RNA cho ta **transcriptome** (hệ phiên mã), tức toàn bộ các phân tử RNA được phiên mã trong một tế bào hay một mô tại một thời điểm. Tầng protein cũng đo được, nhưng bằng những kỹ thuật khác, vì protein không có cơ chế bắt cặp bổ sung như axit nucleic để có thể được “đọc” theo nguyên lý của các phương pháp giải trình tự. Các mũi tên trong hình cũng cho thấy các tầng dữ liệu không độc lập với nhau. Một **biến thể** (*variant*), tức một vị trí mà trình tự DNA khác nhau giữa các cá thể hay khác với trình tự tham chiếu, có thể làm thay đổi lượng RNA được phiên mã, và từ đó thay đổi lượng protein, vì vậy nhiều phân tích trong hệ gen học tìm cách kết nối dữ liệu từ nhiều tầng khác nhau. Phần lớn cuốn sách xoay quanh hai tầng đầu tiên này.

1.4 Lịch sử giải trình tự và các dự án hệ gen

Sau khi mã di truyền được giải, câu hỏi tiếp theo là làm sao để đọc trình tự của một đoạn DNA cụ thể, tức **giải trình tự** (*sequencing*). Biết mã di truyền mà không đọc được trình tự thì cũng giống như biết bảng chữ cái nhưng không có cách nào đọc được văn bản. Năm 1977, Frederick Sanger công bố phương pháp giải trình tự dựa trên dideoxynucleotide, tức các nucleotide làm dừng quá trình tổng hợp DNA tại đúng vị trí của chúng. Khi tổng hợp mạch bổ sung với một lượng nhỏ dideoxynucleotide, ta thu được một tập hợp các đoạn DNA có chiều dài khác nhau, mỗi đoạn kết thúc tại một vị trí mang một loại base nhất định. Sắp xếp các đoạn này theo chiều dài cho phép đọc lần lượt từng base của trình tự. Cùng năm đó, nhóm của ông đọc xong hệ gen hoàn chỉnh đầu tiên, của thực khuẩn thể phiX174, dài khoảng 5.400 nucleotide.¹² Phương pháp Sanger thống trị gần ba mươi năm và mang về cho ông giải Nobel Hóa học thứ hai (giải thứ nhất năm 1958, cho việc xác định trình tự insulin).

Ngay từ giai đoạn này, máy tính đã trở nên cần thiết, vì so sánh một trình tự mới với hàng trăm, hàng nghìn trình tự đã biết, từng ký tự một, là việc không thể làm bằng tay một cách đáng tin cậy. Margaret Dayhoff tập hợp các trình tự protein đã biết vào cuốn *Atlas of Protein Sequence and Structure* (1965) và xây dựng những substitution matrix (ma trận thay thế) đầu tiên, tức bảng cho biết khả năng một amino acid bị thay bằng một amino acid khác trong quá trình tiến hóa. Năm 1970, Needleman và Wunsch công bố thuật toán sequence alignment (sắp xếp trình tự) bằng quy hoạch động (*dynamic programming*), và cũng năm đó Paulien Hogeweg cùng Ben Hesper đặt ra thuật ngữ *bioinformatics* (tin sinh học).¹³ Ngân hàng trình tự GenBank ra đời năm 1982, và BLAST, công cụ tìm trong cơ sở dữ liệu những trình tự tương đồng với một trình tự cho trước, ra đời năm 1990 và nhanh chóng trở thành một trong những công cụ được trích dẫn nhiều nhất trong sinh học. Các thuật toán này sẽ được đề cập ở phần sau của sách.

¹²Sanger F. và cộng sự (1977). *Nature* 265:687–695. <https://doi.org/10.1038/265687a0>

¹³Hogeweg P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology* 7:e1002021. <https://doi.org/10.1371/journal.pcbi.1002021>

Có thể bạn chưa biết. Nguồn gốc của chữ *bioinformatics*

Paulien Hogeweg và Ben Hesper, hai nhà sinh học lý thuyết ở Đại học Utrecht (Hà Lan), dùng chữ *bioinformatica* từ năm 1970 trong một bài viết bằng tiếng Hà Lan, dù nhiều nguồn sau này chỉ truy về các công bố năm 1978 của họ. Định nghĩa ban đầu của họ là “nghiên cứu các quá trình thông tin trong hệ thống sống” (*the study of informatic processes in biotic systems*). Điều họ quan tâm là sự sống xử lý thông tin ra sao ở nhiều cấp độ, từ sự tích lũy thông tin qua tiến hóa đến sự truyền thông tin từ DNA tới các quá trình bên trong và giữa các tế bào. Theo nghĩa đó, *bioinformatics* là một hướng nghiên cứu sinh học, không phải một bộ công cụ phân tích dữ liệu. Đến cuối thập niên 1980, khi dữ liệu trình tự tăng theo cấp số nhân, chữ này mới trở nên phổ biến, nhưng với nghĩa hẹp hơn nhiều, là phát triển và sử dụng các phương pháp tính toán để quản lý và phân tích dữ liệu trình tự.^a Hogeweg cho rằng những hiểu biết thu được từ các dự án giải trình tự lớn là “thân cây” chứ không phải “gốc rễ” của *bioinformatics*. Vì chữ này phổ biến cùng lúc với sự gia tăng nhanh chóng của dữ liệu trình tự, đến nay nhiều người vẫn đồng nhất *bioinformatics* với phân tích dữ liệu hệ gen, dù tin sinh học hiện đại đã rộng hơn nhiều, bao gồm cả cấu trúc protein, sinh học hệ thống và mô hình hóa các quá trình sống.

^aHogeweg P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology* 7(3):e1002021. <https://doi.org/10.1371/journal.pcbi.1002021>

1.4.1 Dự án hệ gen người

Năm 1990, dự án hệ gen người (*Human Genome Project*) chính thức khởi động với ngân sách khoảng 3 tỷ đô la và sự tham gia của nhiều quốc gia. Mục tiêu của dự án là đọc toàn bộ khoảng 3 tỷ cặp base (*base pair, bp*) của hệ gen người. Dự án đi theo chiến lược thận trọng, lập bản đồ vật lý (*physical mapping*) trước, tức xác định vị trí tương đối của các đoạn DNA lớn trên nhiễm sắc thể (*chromosome*), rồi mới giải trình tự từng đoạn lớn đã biết vị trí. Năm 1995, nhóm của Craig Venter tại TIGR chứng minh một cách làm táo bạo hơn, gọi là **whole-genome shotgun**, trong đó toàn bộ DNA được cắt ngẫu nhiên thành các mảnh nhỏ, đọc hết, rồi ghép lại bằng máy tính. Họ công bố hệ gen hoàn chỉnh đầu tiên của một sinh vật sống độc lập, vi khuẩn *Haemophilus influenzae* (khoảng 1,8 triệu cặp base).¹⁴ Sau đó, công ty Celera của Venter áp dụng cách làm này cho hệ gen người, dẫn đến sự cạnh tranh giữa hai phía. Cách làm shotgun nhanh hơn vì bỏ qua bước lập bản đồ, nhưng dôn toàn bộ khó khăn sang khâu ghép trình tự (*assembly*) bằng máy tính, nhất là ở những vùng có trình tự lặp lại (*repetitive sequence*).

Tháng 2 năm 2001, cả hai phía cùng công bố bản nháp hệ gen người. Năm 2003, dự án được tuyên bố hoàn thành, dù thực tế vẫn còn khoảng 8% hệ gen, chủ yếu là các vùng lặp lại dày đặc, chưa giải được. Nguyên nhân nằm ở chính cách đọc DNA theo từng mảnh. Khi một đoạn lặp lại dài hơn chiều dài của mỗi mảnh được đọc, các mảnh rơi vào đoạn đó nhìn gần như giống hệt nhau, và máy tính không thể xác định chúng thuộc về bản sao nào hay ghép chúng theo thứ tự nào. Vì vậy, “hoàn thành” ở đây cần được hiểu là hoàn thành phần hệ gen có thể giải được bằng công nghệ lúc bấy giờ. Những vùng còn lại chỉ dần được lấp đầy khi có công nghệ đọc được những đoạn DNA dài hơn.

¹⁴Fleischmann R.D. và cộng sự (1995). *Science* 269:496–512. <https://doi.org/10.1126/science.7542800>

Có thể bạn chưa biết. Cuộc cá cược GeneSweep

Năm 2000, khi dự án sắp hoàn thành, Ewan Birney (EMBL-EBI) tổ chức một cuộc cá cược mang tên *GeneSweep* để đoán hệ gen người có bao nhiêu gen. Hơn 460 lượt đặt cược đã được ghi nhận, với các dự đoán trải từ dưới 26 nghìn đến hơn 300 nghìn gen. Năm 2003, người thắng cuộc là Lee Rowen, với con số *thấp nhất* là 25.947.^a Ước tính hiện nay còn thấp hơn, khoảng 20 nghìn gen mã hóa protein (*protein-coding gene*), tương đương số gen của giun tròn *Caenorhabditis elegans*.

^aEMBL-EBI. Five things you probably didn't know about the Human Genome Project. <https://www.cambridgenetwork.co.uk/news/five-things-you-probably-didnt-know-about-human-genome-project>; xem thêm <https://en.wikipedia.org/wiki/GeneSweep>.

1.4.2 Ba thế hệ công nghệ giải trình tự

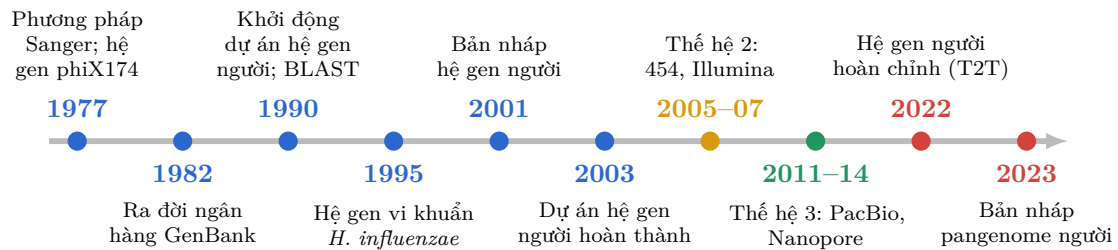
Sau dự án hệ gen người, công nghệ giải trình tự thay đổi nhanh đến mức người ta chia nó thành các “thế hệ” (Hình 1.4). Ba tiêu chí thường dùng để so sánh các thế hệ là độ dài **read**, tức độ dài của mỗi đoạn trình tự mà máy đọc được từ một mảnh DNA, độ chính xác và chi phí cho mỗi base được đọc. Mỗi thế hệ cải thiện một vài tiêu chí nhưng phải đánh đổi ở tiêu chí khác, vì vậy thế hệ sau không thay thế hoàn toàn thế hệ trước, mà mở ra những loại dữ liệu và bài toán phân tích mới.

- **Thế hệ thứ nhất** là phương pháp Sanger, đọc từng đoạn dài khoảng 800–1000 base, chính xác nhưng chậm và đắt.
- **Thế hệ thứ hai** (từ 2005–2007, với 454 và Illumina) đọc song song hàng triệu đến hàng tỷ đoạn ngắn 50–300 base (**short read**). Đây là bước ngoặt làm chi phí giảm hàng nghìn lần.
- **Thế hệ thứ ba** (từ 2011–2014, với PacBio và Oxford Nanopore) đọc các phân tử DNA đơn lẻ (*single-molecule sequencing*), không cần nhân bản trước, cho ra đoạn dài hàng chục nghìn đến hàng triệu base (**long read**).

Long read đã giúp hoàn thiện những vùng cuối cùng của hệ gen người. Lý do là một read đủ dài có thể phủ trọn một đoạn lặp lại cùng với phần trình tự không lặp lại ở hai đầu, nhờ đó xác định được đoạn lặp lại nằm ở đâu. Năm 2022, consortium Telomere-to-Telomere (T2T), tên gọi lấy từ telomere là đầu mút của nhiễm sắc thể, công bố hệ gen người đầu tiên thực sự hoàn chỉnh, từ đầu mút này đến đầu mút kia của mọi nhiễm sắc thể trừ nhiễm sắc thể Y. Năm 2023, bản nháp **pangenome** người đầu tiên ra đời, gồm hệ gen của 47 người thuộc nhiều quần thể khác nhau, thay cho một hệ gen tham chiếu duy nhất. Một hệ gen tham chiếu duy nhất không thể đại diện cho sự đa dạng của cả loài, và những đoạn DNA có ở người này nhưng vắng mặt trong hệ gen tham chiếu rất dễ bị bỏ sót khi phân tích. Hiện tượng này gọi là reference bias (thiên lệch tham chiếu), và pangenome ra đời để giảm bớt thiên lệch đó. Các chủ đề này sẽ được đề cập ở phần sau của sách.

1.5 Quy mô và thành phần của hệ gen

Các dự án giải trình tự lần đầu tiên cho phép mô tả hệ gen một cách định lượng. **Hệ gen** là toàn bộ vật chất di truyền của một sinh vật. Ở người, phần lớn hệ gen nằm trong nhân tế bào, được đóng gói thành 23 cặp nhiễm sắc thể; một phần rất nhỏ nằm trong ty thể (*mitochondria*). Hình 1.5 cho thấy các cấp độ tổ chức, từ tế bào xuống đến từng nucleotide. Một điểm quan trọng là DNA trong nhân không nằm duỗi thẳng mà bị nén rất chặt; muốn một gen được



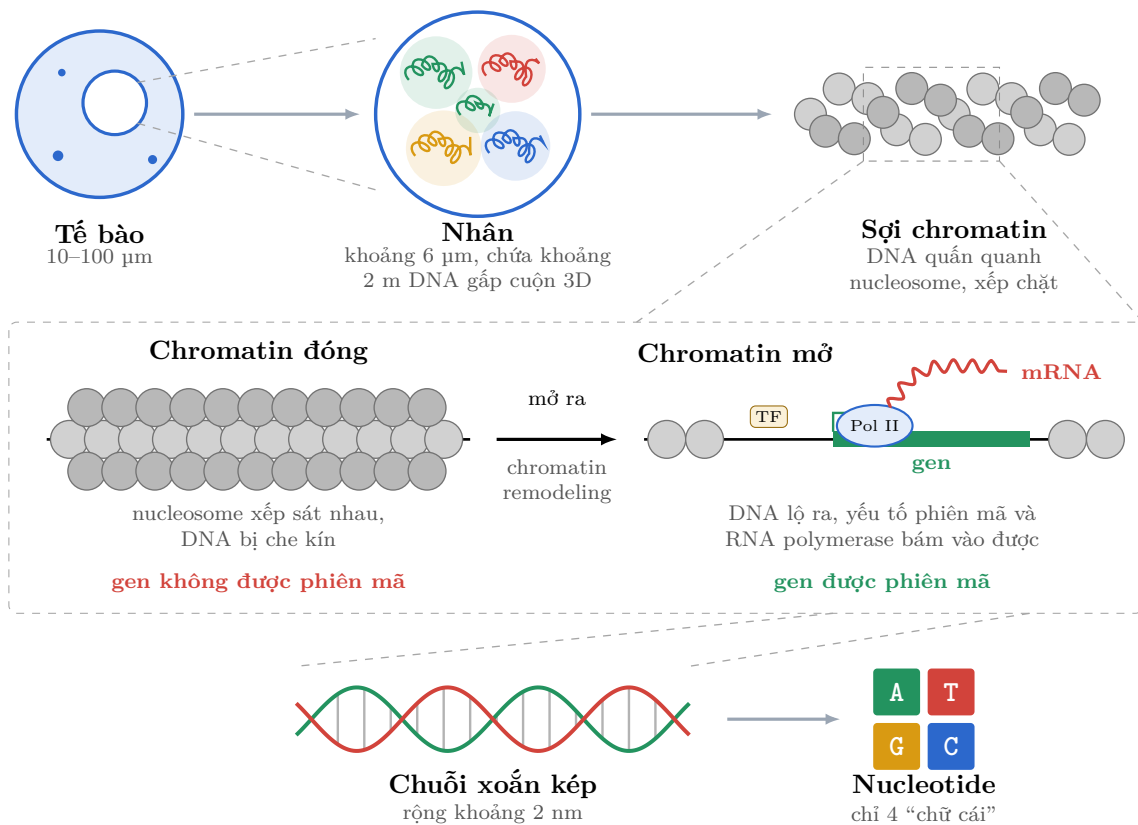
Hình 1.4: Các mốc chính trong lịch sử giải trình tự, từ hệ gen của thực khuẩn thể phiX174 (khoảng 5.400 nucleotide) đến pangenome người.

biểu hiện, tức được phiên mã thành RNA trong quá trình biểu hiện gen (*gene expression*), vùng **chromatin** (chất nhiễm sắc) chứa gen đó phải được mở ra trước. Chromatin là phức hợp gồm DNA và protein, chủ yếu là histone, tạo nên nhiễm sắc thể. Đây cũng là cơ sở của các phương pháp đo độ mở chromatin (*chromatin accessibility*) sẽ được trình bày ở phần sau của sách.

Một vài con số giúp hình dung quy mô của hệ gen người được tóm tắt dưới đây. Những con số này quyết định quy mô của hầu hết các bài toán tính toán sẽ gặp trong sách.

3,1 tỷ cặp base trong một bộ hệ gen đơn bội	20.000 gen mã hóa protein (ước tính)	1,5% hệ gen là trình tự mã hóa protein	50% hệ gen là trình tự lặp lại, phần lớn là transposable element
---	--	--	--

Nếu in hệ gen của một người ra giấy A4 với cỡ chữ thông thường, ta sẽ cần khoảng **một triệu trang**. Mỗi tế bào lưỡng bội (*diploid*) mang hai bản như vậy, một từ cha và một từ mẹ. Các con số trên cũng gợi ý những khó khăn đặc trưng của việc phân tích dữ liệu hệ gen. Vì trình tự mã hóa protein chỉ chiếm khoảng 1,5% hệ gen, phần lớn các biến thể tìm thấy ở một người nằm ngoài gen mã hóa protein, và việc diễn giải ý nghĩa của chúng khó hơn nhiều so với một biến thể làm thay đổi amino acid. Việc khoảng một nửa hệ gen là trình tự lặp lại khiến nhiều mảnh DNA được đọc không thể được đặt chính xác vào một vị trí duy nhất. Cả hai khó khăn này sẽ còn xuất hiện nhiều lần trong các phần sau của sách.

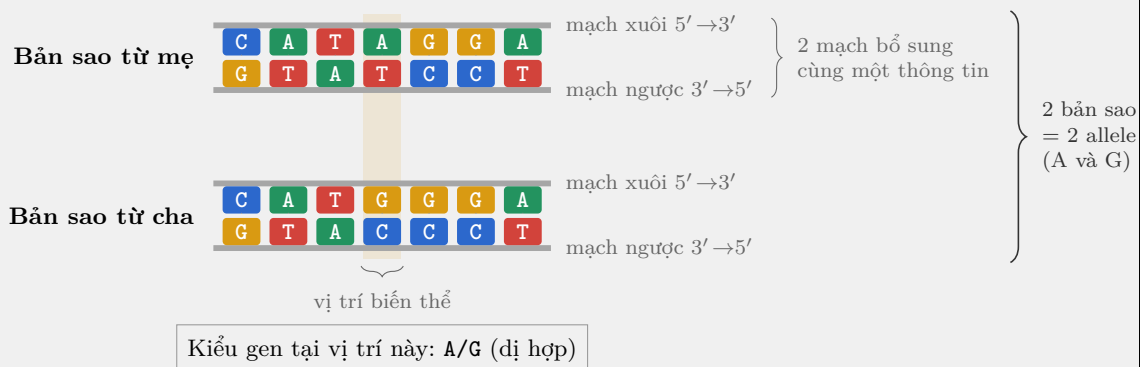


Hình 1.5: Các cấp độ tổ chức của hệ gen trong một tế bào người. Khoảng 2 mét DNA được nén vào một nhân có đường kính vài micromet nhờ quấn quanh nucleosome (đơn vị gồm một đoạn DNA quấn quanh một lõi protein histone) và gấp cuộn trong không gian ba chiều, mỗi nhiễm sắc thể chiếm một vùng riêng trong nhân. Để một gen được phiên mã, vùng chromatin chứa gen đó phải mở ra, để yếu tố phiên mã (TF) và RNA polymerase II (Pol II) tiếp cận được DNA.

Có thể bạn chưa biết. Hai allele hay hai mạch DNA?

Một lầm lẫn rất phổ biến ở người mới làm quen với dữ liệu hệ gen là đồng nhất hai khái niệm “hai” khác nhau. Thứ nhất, tế bào lưỡng bội có hai **bản sao** của mỗi nhiễm sắc thể, một từ cha và một từ mẹ. Tại mỗi vị trí, hai bản sao này mang hai **allele**, tức hai phiên bản trình tự tại cùng một vị trí, có thể giống hoặc khác nhau. Thứ hai, mỗi bản sao lại là một phân tử DNA xoắn kép gồm hai **mạch bổ sung**. Hai mạch bổ sung không phải hai allele, vì chúng mang cùng một thông tin: biết mạch này là suy ra được mạch kia.

Như vậy, tại một vị trí bất kỳ trong tế bào lưỡng bội có bốn mạch DNA nhưng chỉ có hai allele (xem hình minh họa bên dưới). Một người có kiểu gen (*genotype*) dị hợp (*heterozygous*) A/G, tức mang hai allele khác nhau, nghĩa là bản sao từ mẹ mang cặp base A-T và bản sao từ cha mang cặp base G-C, chứ không phải “mạch này là A, mạch kia là G”. Khi gọi kiểu gen (*genotype calling*) từ dữ liệu giải trình tự, read có thể đến từ cả hai mạch của cả hai bản sao, nhưng tất cả đều được quy về chiều của mạch tham chiếu trước khi đếm. Vì vậy, tỉ lệ read mang A và mang G phản ánh hai allele, không phải hai mạch. Cũng cần lưu ý rằng theo quy ước, kiểu gen và biến thể được ghi theo mạch xuôi (*forward strand*) của hệ gen tham chiếu, theo chiều 5' → 3'. Vì vậy, kiểu gen trong hình được ghi là A/G, chứ không phải T/C như khi đọc trên mạch ngược (*reverse strand*).



Phần lớn hệ gen không mã hóa protein, nhưng cũng không phải là “DNA rác” (*junk DNA*) như từng được gọi. Phần này bao gồm intron, tức những đoạn nằm trong gen nhưng bị cắt bỏ khỏi RNA trước khi dịch mã, các vùng điều hòa (*regulatory region*) quyết định gen được biểu hiện ở đâu, khi nào và nhiều hay ít, các gen RNA không mã hóa (*non-coding RNA gene*), và một lượng lớn **transposable element** (TE, yếu tố di động), tức những đoạn DNA có khả năng di chuyển hoặc tự sao chép rồi chèn vào vị trí mới trong hệ gen. Chức năng của nhiều vùng trong số này đến nay vẫn còn đang được tranh luận. Tuy vậy, vai trò của các vùng điều hòa thì đã rõ. Mọi tế bào trong cơ thể mang cùng một hệ gen, nhưng tế bào gan và tế bào thần kinh biểu hiện những tập gen rất khác nhau, và sự khác biệt đó được quyết định phần lớn bởi việc vùng điều hòa nào được kích hoạt trong loại tế bào nào. Vì vậy, một biến thể nằm ngoài gen vẫn có thể ảnh hưởng đến tính trạng nếu nó làm thay đổi hoạt động của một vùng điều hòa.

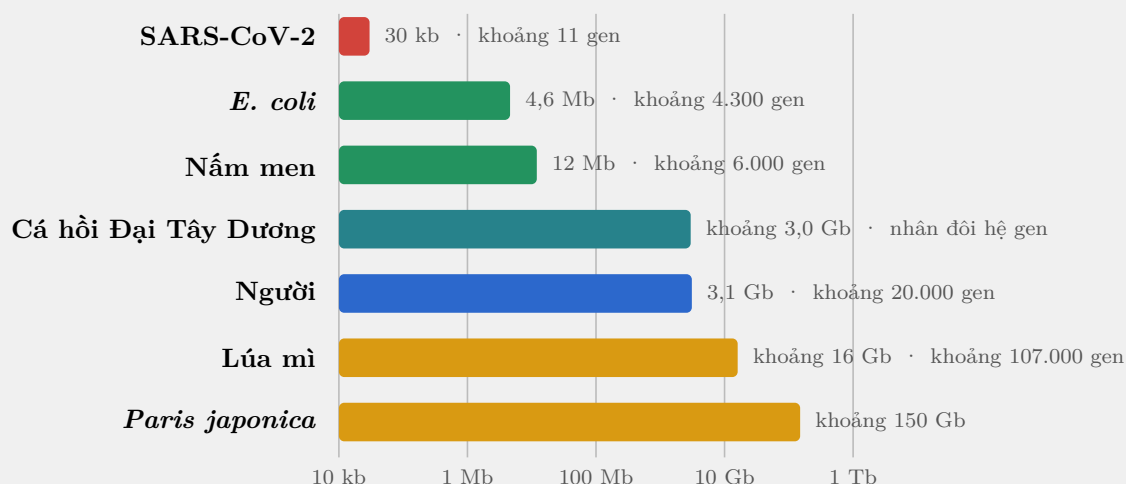
1.5.1 Kích thước hệ gen

Kích thước hệ gen khác nhau rất lớn giữa các loài (xem hình minh họa trong box bên dưới). Hệ gen lúa mì lớn gấp khoảng năm lần hệ gen người, và hệ gen của một loài thực vật có hoa

ở Nhật Bản, *Paris japonica*, lớn gấp gần năm mươi lần. Như vậy, kích thước hệ gen không phản ánh trực tiếp độ phức tạp của một sinh vật. Một hệ gen lớn chưa chắc chứa nhiều gen hơn, vì phần chênh lệch thường nằm ở các trình tự không mã hóa. Đối với người phân tích dữ liệu, kích thước hệ gen lại là một thông số rất thực tế. Hệ gen càng lớn thì càng cần nhiều dữ liệu giải trình tự để mỗi vị trí được đọc đủ số lần, và càng cần nhiều bộ nhớ cũng như thời gian tính toán cho các bước xử lý về sau.

Có thể bạn chưa biết. Vì sao kích thước hệ gen chênh lệch lớn giữa các loài?

Nếu đo độ “phức tạp” của một sinh vật bằng kích thước hệ gen, con người sẽ thua xa một loài hoa nhỏ ở Nhật Bản. Hệ gen của *Paris japonica* dài khoảng 150 tỷ cặp base, gần gấp năm mươi lần hệ gen người, còn hệ gen lúa mì cũng lớn gấp khoảng năm lần. Hình minh họa bên dưới so sánh kích thước hệ gen và số gen ước tính của một số sinh vật, trên thang logarit (mỗi vạch lớn gấp 100 lần vạch trước; số liệu đã làm tròn).



Phần lớn sự chênh lệch này đến từ hai nguồn. Nguồn thứ nhất là **transposable element**. Ở nhiều loài thực vật, retrotransposon, loại transposable element tự nhân lên thông qua một bản sao trung gian bằng RNA, đã tích lũy qua hàng triệu năm và chiếm tới 80% hệ gen, khiến hệ gen phình to mà số gen gần như không tăng. Nguồn thứ hai là **đa bội hóa** (*polyploidization*), tức quá trình tạo ra những loài đa bội (*polyploid*) mang nhiều hơn hai bộ nhiễm sắc thể. Ở bánh mì hằng ngày được làm từ lúa mì *Triticum aestivum*, một loài lục bội (*hexaploid*) mang ba bộ hệ gen tổ tiên khác nhau, nên cả kích thước lẫn số gen đều lớn. Ở động vật, tổ tiên của họ cá hồi đã trải qua một lần nhân đôi toàn bộ hệ gen (*whole-genome duplication*) khoảng 80 triệu năm trước, và đến nay nhiều cặp gen bản sao ở cá hồi Đại Tây Dương (*Salmo salar*) vẫn giống nhau đến mức khó phân biệt. Những đặc điểm này cũng gây khó khăn đáng kể cho người phân tích dữ liệu, vì trình tự lặp lại làm cho assembly (ghép các read thành trình tự dài) và alignment (đặt read lên hệ gen tham chiếu) trở nên mơ hồ, còn đa bội và nhân đôi hệ gen khiến ta khó phân biệt các paralog (bản sao gen tương đồng).

1.6 Sự ra đời của hệ gen học tính toán

Trước khi có các hệ gen hoàn chỉnh, sinh học phân tử chủ yếu làm việc với từng gen, từng protein, và dữ liệu của một thí nghiệm thường gói gọn trong vài bảng số hay vài tấm ảnh điện di (*electrophoresis*). Từ giữa thập niên 1990, khi các hệ gen hoàn chỉnh đầu tiên lần lượt được công bố, sinh học bắt đầu có một loại dữ liệu hầu như chưa từng xuất hiện trước đó. Mỗi sinh vật cho ra hàng triệu đến hàng tỷ ký tự trình tự, một thí nghiệm có thể đo biểu hiện của hàng chục nghìn gen cùng lúc, và về sau, các dự án lớn giải trình tự hệ gen của hàng trăm nghìn người. Loại dữ liệu này không thể đọc bằng mắt hay xử lý bằng bảng tính, và nó đòi hỏi một ngành khoa học mới.

Sự chuyển mình đó in dấu khá rõ lên tên của các tạp chí khoa học trong ngành. Năm 1987, Victor McKusick và Frank Ruddle sáng lập tạp chí *Genomics*, và chính tên tạp chí, do Thomas Roderick đề xuất, đã đưa thuật ngữ *genomics* (hệ gen học) vào từ vựng khoa học.¹⁵ Năm 1995, tạp chí *PCR Methods and Applications* đổi tên thành *Genome Research*; năm 1998, tạp chí *Computer Applications in the Biosciences* (CABIOS) đổi tên thành *Bioinformatics*.¹⁶ Tiếp theo là hàng loạt tạp chí mới, như *Genome Biology* (2000), *PLoS Computational Biology* (2005) hay *NAR Genomics and Bioinformatics* (2019).¹⁷ Những thay đổi về tên gọi này cho thấy cộng đồng nghiên cứu dần xem dữ liệu hệ gen và các phương pháp tính toán đi kèm là một lĩnh vực riêng, chứ không còn là công cụ phụ trợ cho các thí nghiệm sinh học.

Tóm lại, **hệ gen học tính toán** (*computational genomics*) ra đời từ dữ liệu, và tồn tại để khai thác dữ liệu đó. Lĩnh vực này dùng thuật toán, phương pháp thống kê và hạ tầng tính toán để trả lời những câu hỏi sinh học mà trước đây không thể trả lời nếu không có tính toán, ví dụ như so sánh hệ gen của hàng nghìn người để tìm biến thể liên quan đến bệnh, hay đo cùng lúc hoạt động của hàng chục nghìn gen trong một mẫu mô. Điểm khác biệt của hệ gen học tính toán so với việc đơn thuần dùng máy tính để xử lý số liệu nằm ở chỗ câu hỏi sinh học và phương pháp tính toán phải được thiết kế cùng nhau. Một thuật toán tốt nhưng không tính đến đặc điểm của dữ liệu, ví dụ như kiểu sai số riêng của từng công nghệ giải trình tự, vẫn có thể cho ra kết quả sai một cách có hệ thống.

Lĩnh vực này nằm ở giao điểm của ba ngành (Hình 1.6). Sinh học phân tử và di truyền học đặt ra câu hỏi và giúp diễn giải kết quả. Khoa học máy tính cung cấp thuật toán và cấu trúc dữ liệu (*data structure*) để xử lý hàng tỷ ký tự một cách hiệu quả. Thống kê giúp phân biệt tín hiệu thật với nhiễu, một việc tưởng đơn giản nhưng lại là nguồn gốc của rất nhiều kết luận sai trong các nghiên cứu hệ gen. Các vùng giao nhau từng đôi trong hình cũng là những lĩnh vực đã hình thành từ trước, như thuật toán trình tự giữa sinh học và khoa học máy tính, di truyền quần thể (*population genetics*) giữa sinh học và thống kê, hay học máy (*machine learning*) giữa khoa học máy tính và thống kê. Trong thực tế, ít người thành thạo cả ba ngành, vì vậy các dự án hệ gen thường do những nhóm liên ngành thực hiện. Với một người làm nghiên cứu sinh học, khả năng trao đổi được với người làm tính toán, và hiểu được giới hạn của các công cụ mình dùng, cũng quan trọng không kém việc tự mình viết được chương trình. Ngược lại, người làm tính toán cũng cần trao đổi thường xuyên với đồng nghiệp sinh học để hiểu dữ liệu được tạo ra từ thí nghiệm ra sao, câu hỏi sinh học thật sự là gì, và kết quả nào là hợp lý về mặt sinh học. Một phân tích đúng về mặt kỹ thuật nhưng bỏ qua bối cảnh sinh học vẫn có thể dẫn đến kết luận sai.

¹⁵Yadav S.P. (2007). The wholeness in suffix -omics, -omes, and the word om. *Journal of Biomolecular Techniques* 18:277. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2392988>

¹⁶Genome Research: https://en.wikipedia.org/wiki/Genome_Research; Bioinformatics, mục lục NLM của CABIOS: <https://www.ncbi.nlm.nih.gov/nlmcatalog/8511758>

¹⁷Genome Biology: <https://genomebiology.biomedcentral.com/20years>; PLoS Computational Biology: <https://doi.org/10.1371/journal.pcbi.0010004>; NAR Genomics and Bioinformatics: <https://doi.org/10.1093/nargab/lqz001>



Hình 1.6: Hệ gen học tính toán nằm ở giao điểm của sinh học, khoa học máy tính và thống kê.

Ngoài ra, thuật ngữ *computational genomics* thường bị dùng lẫn với hai khái niệm gần gũi. **Tin sinh học** (*bioinformatics*) có phạm vi rộng hơn, vì ngoài dữ liệu hệ gen, nó còn bao gồm phân tích cấu trúc protein, cơ sở dữ liệu sinh học nói chung và nhiều chủ đề khác. **Hệ gen học** (*genomics*) thì có thể hoàn toàn mang tính thực nghiệm, như tối ưu quy trình tách chiết hay tạo thư viện giải trình tự (*sequencing library*), tức tập hợp các mảnh DNA đã được xử lý để có thể đưa vào máy giải trình tự. Cuốn sách này chọn một phạm vi hẹp hơn tin sinh học nhưng đi sâu hơn, tập trung vào dữ liệu DNA và RNA, và cách xử lý chúng ở quy mô lớn. Trong thực tế, ranh giới giữa các thuật ngữ này không cứng nhắc, và cùng một nhà nghiên cứu có thể tự giới thiệu mình bằng bất kỳ tên gọi nào trong số đó tùy theo bối cảnh.

1.7 Quy trình phân tích của một dự án hệ gen

Một hiểu lầm phổ biến là máy giải trình tự “đọc” hệ gen từ đầu đến cuối như ta đọc một cuốn sách. Thực tế, không có công nghệ nào hiện nay làm được như vậy với hệ gen người. Thay vào đó, DNA được cắt thành hàng triệu mảnh, và máy đọc từng mảnh một. Mỗi mảnh được đọc như vậy chính là một read. Như vậy, đầu ra trực tiếp của máy giải trình tự không phải là một hệ gen, mà là một tập hợp rất lớn các read ngắn hơn hệ gen rất nhiều lần và nằm lẫn lộn, không theo thứ tự nào. Quá trình này có thể được hình dung qua một phép so sánh. Giả sử có nhiều bản sao của cùng một cuốn sách, và tất cả đều bị cắt vụn. Nội dung gốc có thể được ghép lại vì các mảnh từ những bản sao khác nhau bị cắt ở những vị trí khác nhau, nên chúng *chồng lấp* lên nhau (Hình 1.7). Dựa vào phần chồng lấp, ta có thể nối mảnh này với mảnh kia, và có càng nhiều bản sao thì càng dễ tìm được những mảnh nối liền nhau.

Có hai cách để “ghép” các read, tùy theo loài đã có hệ gen tham chiếu hay chưa. Nếu đã có sẵn một hệ gen tham chiếu (**reference genome**) của loài, ta chỉ cần tìm vị trí của từng read trên reference đó; việc này gọi là **alignment**. Nếu chưa có reference, ta phải ghép các read với nhau từ đầu; việc này gọi là **assembly**. Alignment nhanh và đơn giản hơn, nhưng chỉ nhìn thấy những gì reference cho phép nhìn thấy, vì một đoạn DNA không có trong reference sẽ không có chỗ để đặt read vào. Assembly không phụ thuộc vào reference, nhưng đòi hỏi

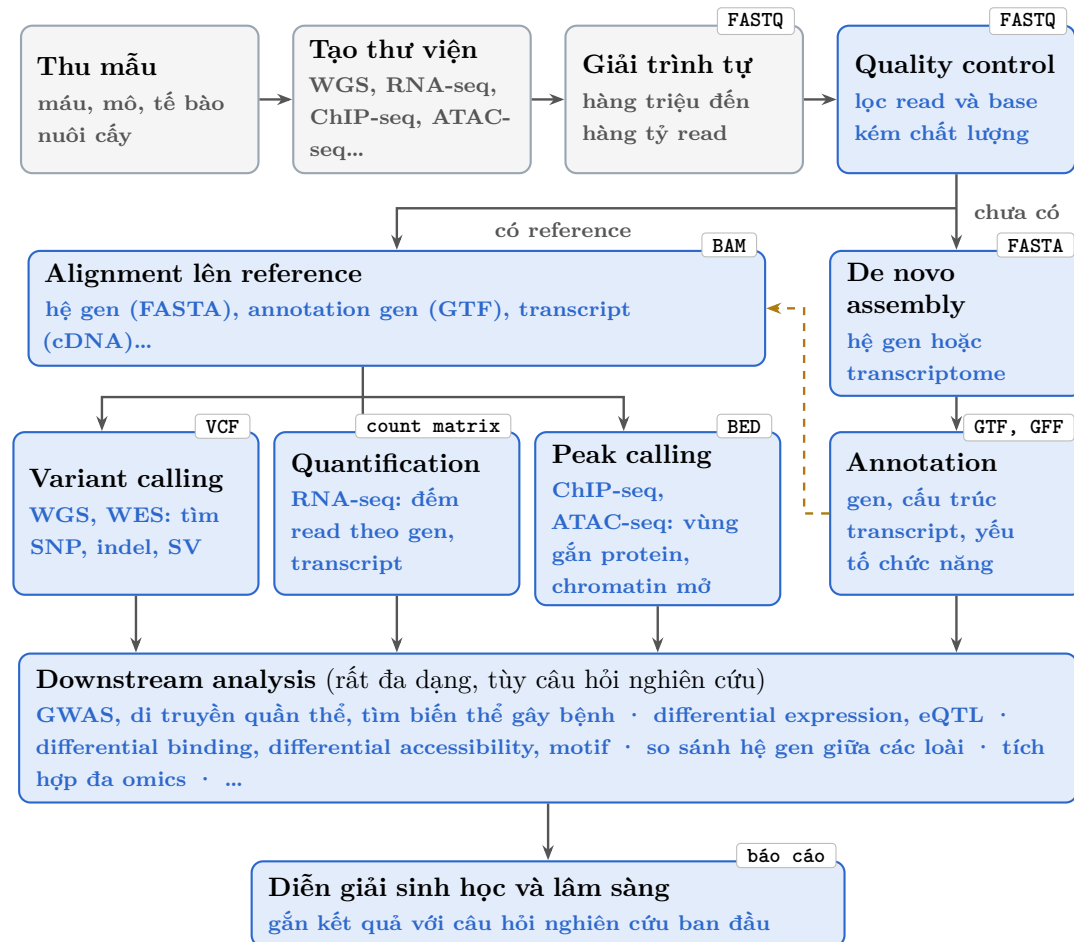


Hình 1.7: Nguyên lý ghép trình tự. Các read chồng lấp nhau cho phép suy ra trình tự dài ban đầu. Một hệ gen người cần từ hàng trăm triệu đến vài tỷ read như vậy.

nhiều dữ liệu hơn, tốn nhiều tài nguyên tính toán hơn, và gặp khó khăn lớn ở các vùng lặp lại như đã thấy ở trên. Cả hai đều là những bài toán thuật toán quan trọng và sẽ được trình bày chi tiết ở phần sau của sách.

Như vậy, điểm chung của mọi dữ liệu giải trình tự là máy không trả về một trình tự hoàn chỉnh, mà trả về rất nhiều read, mỗi read là một mảnh DNA ngắn được lấy ra từ mẫu. Phần lớn công việc tính toán trong hệ gen học, xét cho cùng, là tìm cách sử dụng những mảnh này. Hình 1.8 là một cái nhìn tổng quan về các cách sử dụng đó, dưới dạng một **pipeline**, tức quy trình phân tích gồm nhiều bước nối tiếp nhau, trong đó đầu ra của bước trước là đầu vào của bước sau. Hình chỉ mô tả các dự án dùng dữ liệu giải trình tự, còn những loại dữ liệu khác như SNP array (chip đo kiểu gen tại các vị trí biến thể đã biết) hay microarray (chip đo mức biểu hiện gen bằng phương pháp lai phân tử) có quy trình riêng và sẽ được đề cập ở phần sau của sách. Ba bước đầu diễn ra trong phòng thí nghiệm, và bước tạo thư viện quyết định các mảnh DNA mang loại thông tin gì. Thư viện làm từ DNA toàn hệ gen cho các mảnh phủ đều lên hệ gen, gọi là whole-genome sequencing (WGS, giải trình tự toàn bộ hệ gen). Thư viện làm từ RNA cho các mảnh phản ánh những gen đang được phiên mã, tức dữ liệu RNA-seq. Thư viện ChIP-seq (*chromatin immunoprecipitation sequencing*) dùng kháng thể để chỉ giữ lại các mảnh DNA gắn với một protein nhất định, còn thư viện ATAC-seq (*assay for transposase-accessible chromatin using sequencing*) chỉ giữ lại các mảnh nằm trong vùng chromatin mở. Dù thư viện được tạo theo cách nào, đầu ra của máy giải trình tự vẫn là các read trong file FASTQ, định dạng lưu trình tự của từng read kèm quality score (điểm chất lượng) của từng base, và sự khác biệt giữa các loại dữ liệu chỉ thật sự hiện ra ở cách ta xử lý chúng về sau.

Sau bước quality control (kiểm soát chất lượng), trong đó các read và base kém chất lượng bị loại bỏ, pipeline tách thành hai hướng, tùy theo loài nghiên cứu đã có dữ liệu reference hay chưa. Cần hiểu chữ reference (dữ liệu tham chiếu) ở đây theo nghĩa rộng. Nó không chỉ là trình tự hệ gen tham chiếu, thường được lưu trong file FASTA (định dạng văn bản đơn giản dùng để lưu trình tự), mà còn gồm annotation (chú giải hệ gen) cho biết vị trí và cấu trúc của các gen (file GTF hay GFF), trình tự của các transcript (trong các cơ sở dữ liệu thường được gọi là trình tự cDNA, cũng ở dạng FASTA), và nhiều loại dữ liệu đi kèm khác như danh sách các biến thể đã biết. Mỗi loại dữ liệu cần một phần khác nhau của bộ reference này. Hướng thứ nhất, cũng là hướng phổ biến nhất trong các nghiên cứu trên người và các loài mô hình (*model species*) như chuột, ruồi giấm hay *Arabidopsis thaliana*, bắt đầu bằng việc align read lên reference để biết mỗi mảnh đến từ đâu. Từ đó, mỗi loại dữ liệu sử dụng các mảnh theo một cách riêng. Với dữ liệu DNA như WGS hay WES (*whole-exome sequencing*), tức



Ở quy mô lớn và tái lập được: workflow, HPC, quản trị dữ liệu

● bước thực nghiệm ● bước tính toán - - - kết quả assembly và annotation trở thành reference mới

Hình 1.8: Tổng quan pipeline của một dự án hệ gen dùng dữ liệu giải trình tự. Sau bước quality control, pipeline tách thành hai hướng. Khi loài đã có reference (bên trái), gồm trình tự hệ gen, annotation gen, trình tự transcript và các dữ liệu đi kèm khác, read được align lên reference rồi được xử lý tiếp tùy loại dữ liệu, với variant calling cho DNA, quantification cho RNA và peak calling cho dữ liệu epigenome. Khi chưa có reference (bên phải), read được ghép de novo thành hệ gen hay transcriptome mới rồi được annotation, và kết quả trở thành reference cho các nghiên cứu sau. Downstream analysis rất đa dạng, và hình chỉ nêu một vài ví dụ. Nhân ở góc mỗi bước là định dạng dữ liệu đầu ra.

chỉ giải trình tự exome, tập hợp các vùng mã hóa protein, ta *so sánh* các mảnh với reference để tìm những vị trí mà mẫu khác với reference, tức variant calling (gọi biến thể). Các biến thể này gồm SNP (*single nucleotide polymorphism*), tức sai khác tại một nucleotide, indel, tức những đoạn ngắn bị chèn thêm hay mất đi, và structural variant (SV, biến thể cấu trúc), tức những thay đổi lớn như mất, lặp, đảo hay chuyển chỗ cả một đoạn DNA dài, và kết quả được lưu trong file VCF (*Variant Call Format*). Với RNA-seq, ta *đếm* số mảnh rơi vào mỗi gen hay mỗi transcript để ước lượng mức biểu hiện, tức quantification (định lượng), và kết quả là một count matrix (bảng đếm), với mỗi hàng là một gen và mỗi cột là một mẫu. Bước này cần đến annotation để biết mỗi read rơi vào gen nào, và cũng có những phương pháp so read trực tiếp với trình tự transcript mà không cần align lên hệ gen. Với ChIP-seq hay ATAC-seq, ta tìm những vùng trên hệ gen nơi các mảnh *chồng chất* lên nhau nhiều hơn hẳn so với nền, tức peak calling, vì đó là dấu vết của nơi protein gắn vào hay nơi chromatin mở, và kết quả thường được lưu dưới dạng file BED, định dạng lưu tọa độ các vùng trên hệ gen. Hướng thứ hai là de novo assembly, tức ghép trình tự từ đầu mà không dựa vào reference, trong đó các mảnh được *nối* với nhau dựa trên phần chồng lấp như trong Hình 1.7. Tương tự như reference, kết quả của hướng này cũng không chỉ là một hệ gen. Từ dữ liệu DNA, ta có thể assembly hệ gen của loài, còn từ dữ liệu RNA-seq, ta có thể assembly transcriptome, tức tập hợp trình tự các transcript (bản phiên mã, tức các phân tử RNA được phiên mã từ gen; một gen có thể cho nhiều transcript khác nhau), một cách làm phổ biến khi loài chưa có hệ gen tham chiếu. Sau đó, bước annotation xác định vị trí các gen, cấu trúc transcript và các yếu tố chức năng khác. Như vậy, toàn bộ hướng thứ hai thực chất là xây dựng bộ reference cho một loài, và bộ reference này sẽ được các nghiên cứu sau trên cùng loài sử dụng theo hướng thứ nhất.

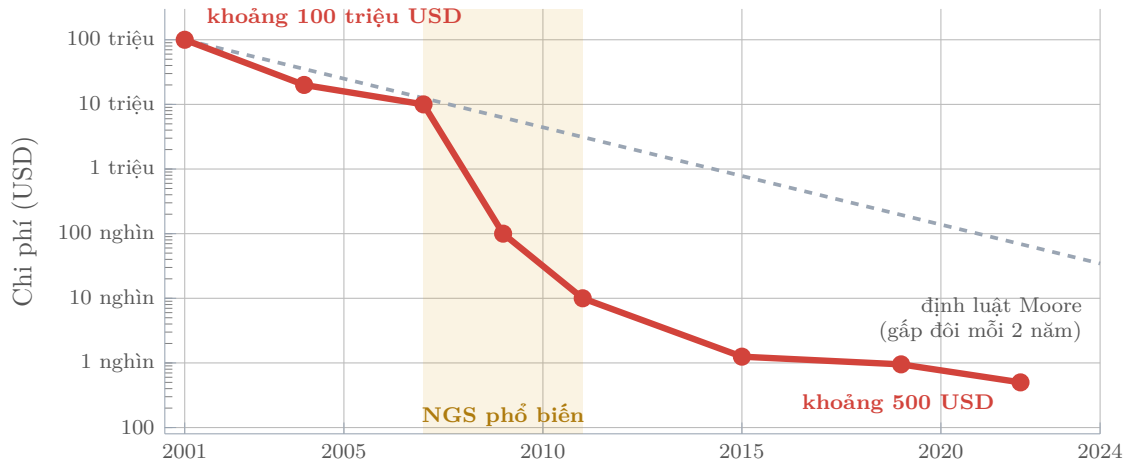
Các file VCF, count matrix hay file peak chưa phải là câu trả lời cho câu hỏi nghiên cứu. Chúng là đầu vào cho downstream analysis, tức các phân tích dùng những kết quả này để trả lời câu hỏi sinh học, phần đa dạng nhất của pipeline vì nó phụ thuộc trực tiếp vào câu hỏi được đặt ra. Cùng một file VCF có thể dùng cho GWAS (*genome-wide association study*), tức nghiên cứu liên kết toàn hệ gen, phân tích population structure (cấu trúc quần thể) hay tìm biến thể gây bệnh. Cùng một count matrix có thể dùng để phân tích differential expression (biểu hiện khác biệt), tức tìm những gen có mức biểu hiện khác nhau giữa hai nhóm mẫu, hay kết hợp với dữ liệu kiểu gen để tìm eQTL (*expression quantitative trait locus*), tức những vị trí trên hệ gen mà biến thể di truyền tại đó ảnh hưởng đến mức biểu hiện của gen. Peak từ ChIP-seq và ATAC-seq có thể được so sánh giữa các điều kiện, hay được quét để tìm motif, tức những đoạn trình tự ngắn đặc trưng mà các yếu tố phiên mã (*transcription factor*) nhận biết và gắn vào. Hình chỉ nêu một vài ví dụ, và danh sách này còn dài hơn rất nhiều. Cuối cùng, kết quả được diễn giải trong bối cảnh sinh học hay lâm sàng để trả lời câu hỏi ban đầu.

Vì các bước nối tiếp nhau, chất lượng của bước sau phụ thuộc vào bước trước. Một mẫu bị nhiễm hay một thư viện kém chất lượng thường không thể được sửa hoàn toàn bằng tính toán, vì vậy người phân tích dữ liệu cũng cần hiểu các bước thực nghiệm phía trước. Dải nền ở dưới của hình nhắc rằng khi số mẫu tăng lên, việc tổ chức các bước thành quy trình tự động và tái lập được (*reproducible*) trở thành một yêu cầu bắt buộc. Cần nhấn mạnh rằng Hình 1.8 chỉ là bức tranh tổng quan. Mỗi loại dữ liệu có những đặc điểm, nguồn lỗi và công cụ riêng, và pipeline cụ thể cho từng loại sẽ được trình bày chi tiết ở phần sau của sách.

1.8 Dữ liệu hệ gen và bài toán dữ liệu lớn

Trong hai thập kỷ qua, chi phí giải trình tự một hệ gen người đã giảm từ khoảng 100 triệu đô la xuống còn khoảng 500 đô la, tức rẻ hơn khoảng 200 nghìn lần (Hình 1.9). Để so sánh, sức mạnh máy tính theo định luật Moore (*Moore's law*) chỉ tăng gấp đôi sau mỗi hai năm.

Từ khoảng năm 2008, khi công nghệ thế hệ thứ hai (*next-generation sequencing*, *NGS*) trở nên phổ biến, đường chi phí giải trình tự tách hẳn khỏi đường Moore. Hệ quả là dữ liệu được sinh ra nhanh hơn nhiều so với tốc độ phần cứng máy tính phát triển để xử lý chúng. Vì vậy, khó khăn chính của nhiều dự án hệ gen hiện nay không còn nằm ở chi phí giải trình tự, mà ở việc lưu trữ, tính toán và phân tích lượng dữ liệu được tạo ra. Đây chính là đặc trưng của một bài toán dữ liệu lớn (*big data*).



Hình 1.9: Chi phí giải trình tự một hệ gen người theo thời gian (số liệu xấp xỉ, theo thống kê của NHGRI). Trục tung là thang logarit.

Để ước lượng quy mô dữ liệu, xét một hệ gen người được giải trình tự ở coverage (độ phủ) $30\times$, tức mỗi vị trí trong hệ gen được đọc trung bình 30 lần (Hình 1.10). Dữ liệu thô ban đầu vào khoảng 100 GB, và qua mỗi bước xử lý, dữ liệu được cô đọng dần, cho đến khi chỉ còn một danh sách vài triệu biến thể, và cuối cùng là một vài biến thể thực sự có ý nghĩa. Kết quả alignment thường được lưu trong file BAM, còn file CRAM nhỏ hơn BAM nhiều vì nó được nén dựa trên reference, tức chủ yếu chỉ lưu những chỗ read khác với reference thay vì toàn bộ trình tự, nhưng về nguyên tắc vẫn giữ được thông tin của từng read. Bước tạo file VCF thì khác, vì nó bỏ đi từng read và chỉ giữ lại các vị trí mà người đó khác với reference. Thông tin đã bị lược bỏ ở một bước như vậy không thể khôi phục lại từ các bước sau, vì vậy dữ liệu ở các tầng đầu thường vẫn phải được lưu giữ để có thể phân tích lại khi có phương pháp tốt hơn.



Hình 1.10: “Phễu dữ liệu” của một hệ gen người ở coverage $30\times$. Chiều dài thanh tỉ lệ với dung lượng lưu trữ.

Dữ liệu thô của một người vào khoảng 100 GB. Một dự án nghiên cứu một nghìn người là khoảng 100 TB, vượt xa khả năng của bất kỳ máy tính cá nhân nào.

Con số này vẫn còn đánh giá thấp quy mô thật sự của dữ liệu, vì giải trình tự không chỉ đơn giản là đọc hệ gen của mỗi cá thể một lần. Hệ gen gần như giống nhau ở mọi tế bào trong cơ thể, nhưng transcriptome thì khác nhau giữa các mô, các giai đoạn phát triển và các điều kiện môi trường, vì vậy một nghiên cứu có thể cần giải trình tự RNA ở nhiều mô của cùng một cá thể, tại nhiều thời điểm khác nhau. Tương tự, mỗi loại tế bào có một trạng thái epigenome (hệ biểu sinh) riêng, tức tập hợp những biến đổi hóa học và cấu trúc trên DNA và chromatin không làm thay đổi trình tự nhưng ảnh hưởng đến việc gen nào được biểu hiện. Trạng thái này gồm vùng chromatin mở, các histone marker (biến đổi hóa học trên protein histone đánh dấu trạng thái hoạt động của một vùng chromatin) và mức DNA methylation (methyl hóa DNA), và mỗi lớp thông tin này lại cần một thí nghiệm riêng. Gần đây hơn, các công nghệ single-cell (đơn tế bào) cho phép đo transcriptome hay epigenome của từng tế bào riêng lẻ thay vì của cả một mẫu mô, nên một thí nghiệm có thể tạo ra dữ liệu cho hàng nghìn tế bào hoặc hơn. Khi nhân số cá thể với số mô, số thời điểm, số lớp dữ liệu và số tế bào, khối lượng dữ liệu tăng lên rất nhanh. Quan trọng hơn, dữ liệu không chỉ lớn mà còn phức tạp, vì các lớp thông tin có cấu trúc khác nhau, nguồn nhiều khác nhau, và phải được kết nối với nhau mới trả lời được một câu hỏi sinh học.

Các kho dữ liệu công cộng quốc tế như Sequence Read Archive (SRA) của NCBI (National Center for Biotechnology Information, Hoa Kỳ) hiện lưu trữ hàng chục petabyte dữ liệu trình tự. Ở quy mô này, câu hỏi không chỉ là “phân tích ra sao”, mà còn là “lưu ở đâu, chạy trên máy nào, và làm sao để người khác chạy lại được đúng kết quả của mình”. Những vấn đề này sẽ được đề cập ở phần sau của sách.

Từ Miescher với chiếc băng gạc năm 1869 đến những kho dữ liệu hàng petabyte hôm nay, câu hỏi cơ bản của sinh học vẫn không thay đổi, đó là thông tin di truyền là gì, và nó tạo nên sự sống như thế nào, còn điều thay đổi là công cụ. Nếu thế hệ của Franklin, Watson và Crick trả lời câu hỏi đó bằng tia X và mô hình que, thì các nhà nghiên cứu hiện nay trả lời bằng thuật toán, thống kê và dòng lệnh. Tuy vậy, công cụ mới không thay thế cho tư duy sinh học. Một phân tích chỉ có giá trị khi nó xuất phát từ một câu hỏi sinh học rõ ràng và kết thúc bằng một diễn giải sinh học hợp lý.

Tóm tắt chương

- Nhờ chuỗi thí nghiệm của Griffith, Avery, Hershey và Chase, cùng quy luật Chargaff, DNA được xác định là vật chất di truyền. Số liệu nhiễu xạ tia X của Franklin và Gosling đóng vai trò quyết định giúp Watson và Crick xây dựng mô hình chuỗi xoắn kép năm 1953.
- Central dogma mô tả dòng thông tin DNA→RNA→protein. Mỗi phân tử tương ứng với một tầng dữ liệu omics.
- Công nghệ giải trình tự đã trải qua ba thế hệ, từ Sanger đến short read song song và long read phân tử đơn, dẫn đến hệ gen người hoàn chỉnh (2022) và pangenome (2023).
- Hệ gen người có khoảng 3,1 tỷ cặp base và 20 nghìn gen mã hóa protein, nhưng chỉ khoảng 1,5% là trình tự mã hóa. DNA trong nhân bị nén chặt, và một gen chỉ được phiên mã khi vùng chromatin chứa nó được mở ra.
- Khi khó khăn dần dịch chuyển từ phòng thí nghiệm sang máy tính, hệ gen học tính toán ra đời ở giao điểm của sinh học, khoa học máy tính và thống kê. Chi phí giải trình tự giảm nhanh hơn định luật Moore, biến dữ liệu hệ gen thành một bài toán dữ liệu lớn.

Thuật ngữ

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>alignment</i>	sắp xếp trình tự	Đặt các trình tự cạnh nhau để tìm phần tương đồng; với dữ liệu giải trình tự là xác định vị trí của từng read trên reference
<i>annotation</i>	chú giải hệ gen	Xác định vị trí và cấu trúc của gen, transcript và các yếu tố chức năng trên hệ gen
<i>antiparallel</i>	đối song song	Hai mạch DNA chạy theo hai hướng ngược nhau
<i>assembly (de novo)</i>	ghép trình tự	Ghép các read thành trình tự dài mà không cần reference
<i>ATAC-seq</i>	–	Giải trình tự các mảnh DNA nằm trong vùng chromatin mở
<i>BAM / CRAM</i>	–	Định dạng file lưu kết quả alignment; CRAM là dạng nén dựa trên reference
<i>base, base pair (bp)</i>	base, cặp base	Base nitơ A, T, G, C của nucleotide; cặp base là đơn vị đo chiều dài DNA sợi đôi
<i>BED</i>	–	Định dạng file lưu tọa độ các vùng trên hệ gen, ví dụ các peak
<i>big data</i>	dữ liệu lớn	Dữ liệu có quy mô và độ phức tạp vượt quá khả năng lưu trữ và xử lý của các công cụ thông thường
<i>bioinformatics</i>	tin sinh học	Ngành dùng phương pháp tính toán để quản lý và phân tích dữ liệu sinh học
<i>cDNA (complementary DNA)</i>	DNA bổ sung	DNA được tổng hợp từ khuôn RNA nhờ enzyme phiên mã ngược
<i>central dogma</i>	học thuyết trung tâm	Nguyên lý về dòng thông tin DNA→RNA→protein
<i>ChIP-seq</i>	–	Giải trình tự các mảnh DNA gắn với một protein nhất định, được tách bằng kháng thể
<i>chromatin</i>	chất nhiễm sắc	Phức hợp DNA và protein, chủ yếu là histone, tạo nên nhiễm sắc thể
<i>codon, stop codon</i>	bộ ba, bộ ba kết thúc	Nhóm ba nucleotide liên tiếp trên RNA, mã hóa một amino acid; bộ ba kết thúc báo hiệu dừng dịch mã
<i>complementary base pairing</i>	bắt cặp bổ sung	A bắt cặp với T, G bắt cặp với C giữa hai mạch DNA
<i>computational genomics</i>	hệ gen học tính toán	Lĩnh vực dùng thuật toán, thống kê và hạ tầng tính toán để khai thác dữ liệu hệ gen
<i>count matrix</i>	bảng đếm	Bảng số read của mỗi gen (hàng) trong mỗi mẫu (cột)
<i>coverage, depth</i>	độ phủ	Số read trung bình phủ lên mỗi vị trí trong hệ gen
<i>differential expression</i>	biểu hiện khác biệt	Gen có mức biểu hiện khác nhau giữa các nhóm mẫu
<i>diploid / haploid</i>	lượng bội / đơn bội	Mang hai bộ / một bộ nhiễm sắc thể
<i>DNA methylation</i>	methy hóa DNA	Gắn nhóm methyl lên base của DNA, thường là C trong CpG, ảnh hưởng đến biểu hiện gen mà không đổi trình tự
<i>double helix</i>	chuỗi xoắn kép	Cấu trúc hai mạch DNA xoắn quanh một trục chung

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>downstream analysis</i>	–	Các phân tích dùng VCF, count matrix hay peak để trả lời câu hỏi nghiên cứu
<i>epigenome</i>	hệ biểu sinh	Các biến đổi hóa học và cấu trúc trên DNA và chromatin không làm thay đổi trình tự nhưng ảnh hưởng đến biểu hiện gen
<i>eQTL</i>	–	Vị trí trên hệ gen mà biến thể tại đó ảnh hưởng đến mức biểu hiện của gen
<i>FASTA / FASTQ</i>	–	Định dạng lưu trình tự / lưu read kèm điểm chất lượng của từng base
<i>gene expression</i>	biểu hiện gen	Quá trình gen được phiên mã (và dịch mã) thành sản phẩm
<i>genome</i>	hệ gen	Toàn bộ vật chất di truyền của một sinh vật
<i>genotype</i>	kiểu gen	Tổ hợp allele của một cá thể tại một vị trí; dị hợp (<i>heterozygous</i>) khi hai allele khác nhau
<i>GTF / GFF</i>	–	Định dạng file annotation, lưu vị trí và cấu trúc của gen và transcript
<i>GWAS</i>	nghiên cứu liên kết toàn hệ gen	Tìm các biến thể liên quan đến một tính trạng hay bệnh trên toàn hệ gen
<i>histone marker</i>	–	Biến đổi hóa học trên protein histone, đánh dấu trạng thái hoạt động của một vùng chromatin
<i>indel</i>	–	Đoạn trình tự ngắn bị chèn thêm (insertion) hay mất đi (deletion)
<i>intron</i>	–	Đoạn nằm trong gen nhưng bị cắt bỏ khỏi RNA trước khi dịch mã
<i>model species</i>	loài mô hình	Loài được nghiên cứu rộng rãi, có sẵn reference và công cụ, ví dụ chuột, ruồi giấm, <i>Arabidopsis thaliana</i>
<i>motif</i>	–	Đoạn trình tự ngắn đặc trưng mà một yếu tố phiên mã nhận biết và gắn vào
<i>NGS (next-generation sequencing)</i>	giải trình tự thế hệ mới (thế hệ thứ hai)	Các công nghệ giải trình tự song song hàng loạt từ thế hệ thứ hai trở đi
<i>nucleotide</i>	–	Đơn vị cấu tạo của axit nucleic, gồm một phân tử đường, một nhóm phosphate và một base
<i>omics</i>	–	Dữ liệu đo đồng thời toàn bộ một loại phân tử, ví dụ genome, transcriptome
<i>pangenome</i>	–	Tập hợp hệ gen của nhiều cá thể trong cùng một loài, dùng làm reference thay cho một hệ gen duy nhất
<i>paralog</i>	bản sao gen tương đồng	Các bản sao gen trong cùng một hệ gen, sinh ra từ sự kiện nhân đôi
<i>peak calling</i>	–	Tìm những vùng có nhiều read chồng chất hơn hẳn so với nền trong ChIP-seq, ATAC-seq
<i>physical mapping</i>	lập bản đồ vật lý	Xác định vị trí tương đối của các đoạn DNA lớn trên nhiễm sắc thể
<i>pipeline</i>	quy trình phân tích	Chuỗi các bước xử lý nối tiếp, đầu ra của bước trước là đầu vào của bước sau

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>polyploidy, polyploidization</i>	đa bội, đa bội hóa	Mang nhiều hơn hai bộ nhiễm sắc thể / quá trình tạo ra loài đa bội
<i>quality control (QC)</i>	kiểm soát chất lượng	Đánh giá dữ liệu và loại bỏ read, base kém chất lượng
<i>quantification</i>	định lượng	Đếm read theo gen hay transcript để ước lượng mức biểu hiện
<i>read</i>	–	Một đoạn trình tự do máy giải trình tự đọc được từ một mảnh DNA
<i>reference</i>	dữ liệu tham chiếu	Bộ dữ liệu chuẩn của một loài, gồm hệ gen (FASTA), annotation (GTF/GFF), transcript (cDNA) và dữ liệu đi kèm
<i>reference bias</i>	thiên lệch tham chiếu	Xu hướng bỏ sót hay đánh giá sai những trình tự khác biệt nhiều so với reference
<i>reference genome</i>	hệ gen tham chiếu	Hệ gen chuẩn của một loài, dùng để so sánh
<i>regulatory region</i>	vùng điều hòa	Vùng DNA quyết định gen được biểu hiện ở đâu, khi nào và nhiều hay ít
<i>replication</i>	sao chép DNA	Tổng hợp phân tử DNA mới từ khuôn DNA, theo kiểu bán bảo toàn
<i>reverse transcription, reverse transcriptase</i>	phiên mã ngược, enzyme phiên mã ngược	Tổng hợp DNA từ khuôn RNA / enzyme thực hiện quá trình đó
<i>RNA-seq</i>	giải trình tự RNA	Giải trình tự RNA (sau khi chuyển thành cDNA) để đo biểu hiện gen
<i>sequencing</i>	giải trình tự	Xác định thứ tự các nucleotide trong một phân tử DNA hay RNA
<i>sequencing library</i>	thư viện giải trình tự	Tập hợp các mảnh DNA đã được xử lý để đưa vào máy giải trình tự
<i>short read / long read</i>	–	Read ngắn (50–300 base) / read dài (hàng chục nghìn base trở lên)
<i>single-cell</i>	đơn tế bào	Đo dữ liệu omics ở từng tế bào riêng lẻ thay vì cả một mẫu mô
<i>SNP</i>	–	Sai khác tại một nucleotide giữa các cá thể (<i>single nucleotide polymorphism</i>)
<i>structural variant (SV)</i>	biến thể cấu trúc	Thay đổi lớn như mất, lặp, đảo hay chuyển chỗ một đoạn DNA dài
<i>transcript</i>	bản phiên mã	Một phân tử RNA được phiên mã từ gen; một gen có thể cho nhiều transcript
<i>transcription / translation</i>	phiên mã / dịch mã	Tổng hợp RNA từ DNA / tổng hợp protein từ RNA
<i>transcription factor (TF)</i>	yếu tố phiên mã	Protein gắn vào DNA tại các motif đặc trưng để điều hòa phiên mã
<i>transcriptome</i>	hệ phiên mã	Toàn bộ các phân tử RNA trong một tế bào hay mô tại một thời điểm
<i>transposable element (TE)</i>	yếu tố di động	Đoạn DNA có thể di chuyển hoặc tự sao chép trong hệ gen
<i>variant</i>	biến thể	Vị trí mà trình tự DNA khác nhau giữa các cá thể hay khác với reference

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>variant calling</i>	gọi biến thể	Tìm các biến thể từ dữ liệu alignment
<i>VCF</i>	–	Định dạng file lưu các biến thể và kiểu gen
<i>WGS / WES</i>	–	Giải trình tự toàn bộ hệ gen / chỉ exome, tức các vùng mã hóa protein
<i>whole-genome duplication</i>	nhân đôi toàn bộ hệ gen	Sự kiện tiến hóa trong đó toàn bộ hệ gen được nhân đôi
<i>whole-genome shotgun</i>	–	Cắt ngẫu nhiên toàn bộ DNA thành mảnh nhỏ, đọc hết rồi ghép lại bằng máy tính

Đọc thêm

Watson J.D., Crick F.H.C. (1953). Molecular structure of nucleic acids. *Nature* 171:737–738. <https://doi.org/10.1038/171737a0>

Franklin R.E., Gosling R.G. (1953). Molecular configuration in sodium thymonucleate. *Nature* 171:740–741. <https://doi.org/10.1038/171740a0>

Crick F. (1970). Central dogma of molecular biology. *Nature* 227:561–563. <https://doi.org/10.1038/227561a0>

Sanger F., Nicklen S., Coulson A.R. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS* 74:5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860–921. <https://doi.org/10.1038/35057062>

Nurk S. và cộng sự (2022). The complete sequence of a human genome. *Science* 376:44–53. <https://doi.org/10.1126/science.abj6987>

Liao W.-W. và cộng sự (2023). A draft human pangenome reference. *Nature* 617:312–324. <https://doi.org/10.1038/s41586-023-05896-x>