

Chương 2

Dữ liệu hệ gen được tạo ra như thế nào?

Ở chương trước, chúng ta nói rằng “đọc” một hệ gen người giờ chỉ tốn vài trăm đô la. Câu nói đó đúng nhưng chưa đầy đủ, vì không có máy nào đọc DNA giống như ta đọc một cuốn sách. Mỗi công nghệ chỉ cho ta một góc nhìn, với độ phân giải nhất định, và luôn kèm theo lỗi. Một bộ dữ liệu hệ gen vì vậy không phải là “sự thật” về sinh vật, mà là một phép đo có giới hạn.

Chương này trả lời ba câu hỏi, đó là đọc cái gì, đọc bằng cách nào và đọc ở độ phân giải nào. Đặt hai câu hỏi đầu cạnh nhau, ta được một ma trận, còn câu hỏi thứ ba cho biết mỗi ô của ma trận có thể được đo trên cả một mẫu mô, trên từng tế bào hay ngay tại vị trí của tế bào trong mô. Hầu hết các công nghệ tạo dữ liệu hệ gen hiện nay đều nằm ở một ô trong ma trận đó. Cuối chương, chúng ta sẽ đọc lại ma trận theo một góc nhìn khác để xác định mỗi ô có những giới hạn gì, và những giới hạn đó đã sinh ra những bài toán tính toán nào.

2.1 Phân loại công nghệ theo tầng phân tử và nền tảng đọc

Câu hỏi thứ nhất, đọc cái gì, liên quan đến tầng phân tử. Quay lại central dogma, ta có thể đo trình tự DNA và các biến thể của nó, đo lượng RNA được phiên mã, đo các dấu ấn hóa học trên DNA như methylation, đo vùng chromatin nào đang mở và protein nào đang gắn vào DNA, hay đo cách hệ gen gấp cuộn trong không gian ba chiều. Mỗi tầng trả lời một câu hỏi sinh học khác nhau. Trình tự DNA cho biết một cá thể mang những biến thể nào, gần như không đổi trong suốt đời sống và giống nhau giữa các mô. Các tầng còn lại thì thay đổi theo loại tế bào, theo giai đoạn phát triển và theo điều kiện môi trường, nên cho biết phần nào của hệ gen đang thật sự được sử dụng. Vì vậy, cùng một cá thể có thể cho ra một bộ dữ liệu DNA duy nhất nhưng rất nhiều bộ dữ liệu RNA hay chromatin khác nhau, tùy theo mô được lấy mẫu.

Câu hỏi thứ hai, đọc bằng cách nào, liên quan đến nền tảng công nghệ. Có ba cách đọc chính, trong đó **array** (chip) dựa trên sự lai phân tử (*hybridization*) giữa DNA mẫu và các probe (đầu dò) đã thiết kế sẵn, **short-read sequencing** đọc song song hàng tỷ đoạn ngắn, còn **long-read sequencing** đọc từng phân tử DNA hoặc RNA dài. Ba cách đọc này khác nhau ở một điểm cơ bản. Array chỉ kiểm tra những gì người thiết kế đã biết trước, trong khi hai cách giải trình tự đọc trực tiếp trình tự base nên có thể phát hiện điều mới. Giữa hai cách giải trình tự, chiều dài của mỗi lần đọc quyết định ta giữ được bao nhiêu thông tin về những gì nằm cạnh nhau trên cùng một phân tử.

Câu hỏi thứ ba, đọc ở độ phân giải (*resolution*) nào, liên quan đến đơn vị mà mỗi phép đo mô tả. Cách đo truyền thống, gọi là **bulk**, đo chung cả một mẫu mô gồm hàng triệu tế

bào, nên kết quả là giá trị trung bình của mọi tế bào trong mẫu. Từ cuối thập niên 2000, các kỹ thuật **single-cell** cho phép đo riêng từng tế bào,¹ và gần đây các kỹ thuật **spatial** còn giữ được vị trí của từng tế bào hay từng vùng nhỏ trong mô. Giữa các mức này còn nhiều nấc trung gian, ví dụ như đo trên một nhóm tế bào đã được tách riêng theo loại, đo trên nhân thay vì cả tế bào, hay đo trên những điểm của một lát cắt mô có kích thước từ vài tế bào xuống đến nhỏ hơn một tế bào. Song song với việc tăng độ phân giải là một xu hướng khác, đó là đo đồng thời nhiều tầng phân tử trên cùng một tế bào hay cùng một lát cắt mô, thay vì đo từng tầng trên những mẫu khác nhau. Hai xu hướng này sẽ được trình bày kỹ ở cuối chương, sau khi người đọc đã nắm được từng tầng phân tử và từng nền tảng đọc.

Đặt hai trục đầu tiên vuông góc với nhau, ta được ma trận trong Hình 2.1. Ma trận này có hai cách đọc. Đọc theo hàng, ta thấy cùng một câu hỏi sinh học có thể được tiếp cận bằng nhiều công nghệ, với những đánh đổi khác nhau. Đọc theo cột, ta thấy lịch sử phát triển công nghệ, trong đó mỗi cột ra đời sau, một phần để giải quyết giới hạn của cột trước. Độ phân giải không được vẽ thành một trục riêng, vì nó cắt ngang gần như mọi ô của ma trận, ví dụ như RNA-seq hay ATAC-seq đều có phiên bản bulk, single-cell và spatial.

		Array	Short read	Long read
tăng phân tử (<i>central dogma</i>) ↓	Biến thể DNA	SNP array hàng trăm nghìn đến vài triệu vị trí định sẵn	WGS, WES, low-pass WGS	Long-read WGS: structural variant, phasing, vùng lặp, T2T
	Transcriptome	Expression microarray	RNA-seq (bulk, single-cell)	Iso-Seq, direct RNA: isoform đầy đủ, biến đổi trên RNA
	Methylation	Methylation array (450K, EPIC)	WGBS, RRBS, EM-seq	Đọc trực tiếp 5mC trên phân tử gốc, không cần bisulfite
	Chromatin, gắn protein- DNA	ChIP-chip	ChIP-seq, ATAC- seq, CUT&Tag	Fiber-seq: chromatin trên từng phân tử
	Cấu trúc 3D	(không phổ biến)	Hi-C	Pore-C: tiếp xúc đa điểm

thời gian: mỗi cột giải quyết một phần giới hạn của cột trước →

Hình 2.1: Ma trận công nghệ tạo dữ liệu hệ gen. Hàng biểu diễn tầng phân tử được đo, cột biểu diễn nền tảng đọc. Mỗi ô là một nhóm công nghệ tiêu biểu.

Phần còn lại của chương đi theo đúng cấu trúc này. Trước hết, chúng ta tìm hiểu nguyên lý của ba cách đọc, mỗi cách chỉ giải thích một lần. Sau đó, chúng ta đi qua từng hàng của ma trận. Cách sắp xếp này giúp tránh lặp lại, vì cùng một nền tảng như short read xuất hiện ở hầu hết các hàng, chỉ khác nhau ở bước chuẩn bị mẫu trước khi đọc. Khi đã nắm được nguyên lý của một nền tảng, người đọc có thể đoán trước phần lớn điểm mạnh và điểm yếu của nó ở bất kỳ hàng nào. Ví dụ như giới hạn về chiều dài read của short read sẽ lần lượt hiện ra khi ta bàn về structural variant, về isoform, tức các transcript khác nhau được tạo ra từ cùng một gen, và về phasing, tức việc xác định các biến thể có nằm trên cùng một nhiễm sắc thể hay không.

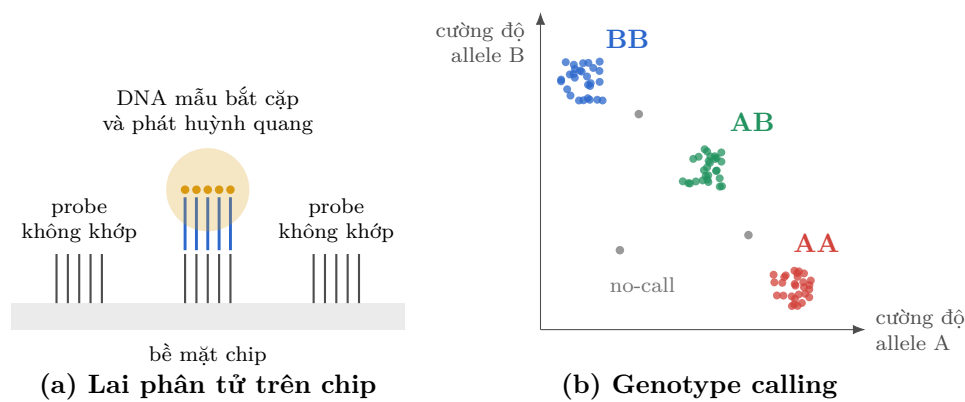
¹Tang F. và cộng sự (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* 6:377–382. <https://doi.org/10.1038/nmeth.1315>

2.2 Ba nền tảng công nghệ

2.2.1 Array

Nguyên lý của array là sự lai phân tử, tức sự bắt cặp giữa hai mạch DNA bổ sung. Trên một con chip nhỏ, người ta gắn sẵn hàng trăm nghìn đến vài triệu loại probe, mỗi loại là một đoạn oligonucleotide ngắn được thiết kế để bắt cặp với một vị trí cụ thể trong hệ gen. DNA mẫu được cắt nhỏ, đánh dấu huỳnh quang và đổ lên chip. Đoạn DNA nào bổ sung với probe nào sẽ bám vào đó, và cường độ huỳnh quang tại mỗi điểm cho biết lượng DNA tương ứng.

Với SNP array, mỗi vị trí SNP có tín hiệu cho hai allele. Vẽ cường độ của allele A và allele B cho nhiều mẫu, ta thấy các mẫu tự tách thành ba đám, tương ứng với ba kiểu gen AA, AB và BB (Hình 2.2). Việc gán kiểu gen cho từng mẫu, gọi là **genotype calling**, thực chất là một bài toán phân cụm (*clustering*). Những mẫu nằm lưng chừng giữa các đám sẽ không được gán kiểu gen và được ghi nhận là no-call.



Hình 2.2: Nguyên lý SNP array. (a) Chỉ DNA bổ sung với probe mới bám lại và phát tín hiệu. (b) Mỗi điểm là một mẫu; các mẫu tự tách thành ba đám kiểu gen. Điểm xám nằm giữa các đám không được gán kiểu gen.

Điểm mạnh của array là rẻ, nhanh và ổn định. Chi phí mỗi mẫu chỉ khoảng vài chục đô la, quy trình phân tích đã chuẩn hóa từ lâu, và dữ liệu giữa các phòng thí nghiệm dễ so sánh với nhau. Nhờ vậy, array là lựa chọn gần như mặc định cho các nghiên cứu cần hàng trăm nghìn mẫu, như UK Biobank với khoảng 500 nghìn người tham gia, hay các chương trình chọn giống vật nuôi.

Điểm yếu của array cũng nằm ngay trong nguyên lý của nó, đó là chỉ đo được những gì đã được thiết kế sẵn. Một biến thể không có probe tương ứng thì hoàn toàn vô hình. Hơn nữa, các vị trí trên array thường được chọn dựa trên những quần thể đã có nhiều dữ liệu, chủ yếu là người gốc châu Âu. Khi dùng cho quần thể khác, array có thể bỏ sót những biến thể quan trọng riêng của quần thể đó, và hiện tượng này được gọi là **ascertainment bias**, tức sai lệch sinh ra vì các vị trí được đo đã được chọn từ trước dựa trên một số quần thể nhất định.

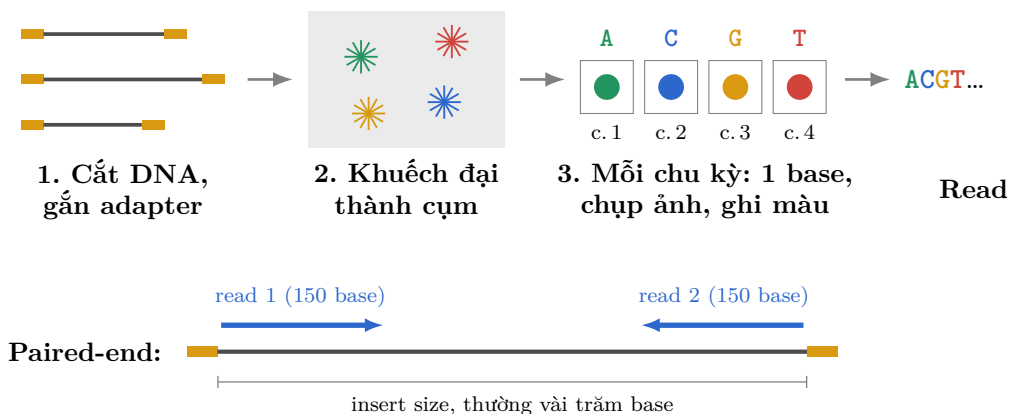
2.2.2 Short-read sequencing

Nền tảng short read phổ biến nhất hiện nay là Illumina, dựa trên nguyên lý **sequencing by synthesis** (giải trình tự bằng tổng hợp). Khác với array, giải trình tự không cần biết trước vị trí nào sẽ được đo, mà đọc trực tiếp thứ tự các base trên từng đoạn DNA. Nhờ vậy, mọi biến thể nằm trên đoạn được đọc đều có cơ hội được phát hiện, kể cả những biến thể chưa từng được ghi nhận. Ý tưởng cốt lõi là ghi lại từng nucleotide khi polymerase tổng hợp mạch

bổ sung, và làm điều đó cùng lúc trên một số lượng cực lớn các đoạn DNA. Quá trình gồm ba bước chính (Hình 2.3).

1. **Tạo thư viện** (*library preparation*). DNA được cắt thành các đoạn vài trăm base, hai đầu được gắn các đoạn **adapter**, tức những đoạn DNA ngắn có trình tự biết trước, giúp đoạn DNA bám vào flow cell và được máy nhận diện.
2. **Khuếch đại thành cụm** (*cluster generation*). Mỗi đoạn DNA bám vào bề mặt flow cell, tức tấm kính có các kênh nhỏ nơi phản ứng giải trình tự diễn ra, và được nhân lên tại chỗ thành một cụm gồm hàng nghìn bản sao giống hệt nhau. Mục đích là khuếch đại tín hiệu đủ mạnh để camera nhìn thấy.
3. **Đọc theo chu kỳ**. Ở mỗi chu kỳ, polymerase gắn thêm đúng một nucleotide vào mạch đang tổng hợp. Mỗi loại nucleotide mang một màu huỳnh quang khác nhau và có một nhóm chặn tạm thời (*reversible terminator*), nên quá trình dừng lại sau mỗi base. Camera chụp ảnh toàn bộ flow cell, ghi lại màu của từng cụm, rồi nhóm chặn được cắt bỏ để bắt đầu chu kỳ tiếp theo.

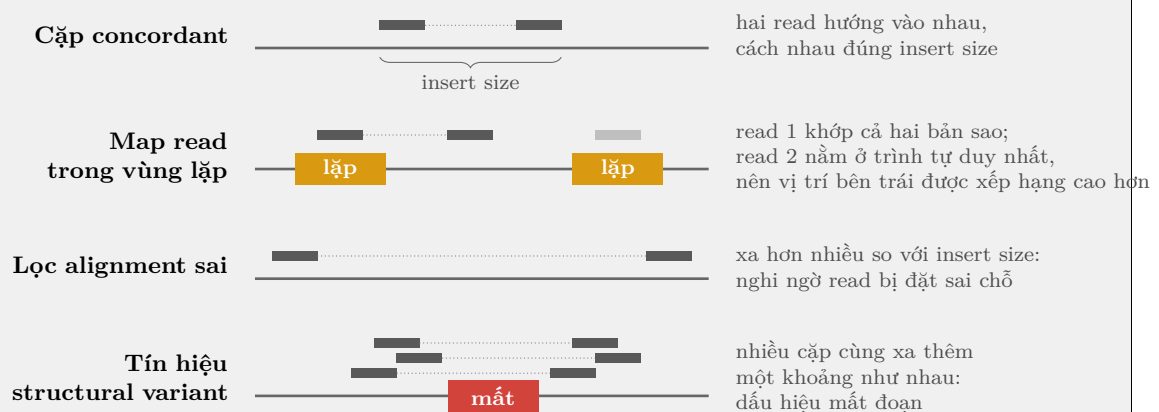
Sau 150 chu kỳ, mỗi cụm cho ra một read dài 150 base. Hàng tỷ cụm được đọc song song trên cùng một flow cell, và đó là lý do chi phí giảm mạnh. Phần lớn dữ liệu hiện nay là **paired-end**, tức mỗi đoạn DNA được đọc từ cả hai đầu, cho ra một cặp read cách nhau một khoảng xấp xỉ biết trước, gọi là insert size, tức chiều dài của đoạn DNA được đọc từ hai đầu. Thông tin về khoảng cách này rất hữu ích khi alignment và khi phát hiện structural variant.



Hình 2.3: Sequencing by synthesis (Illumina). Phía trên là ba bước từ DNA đến read; phía dưới là một đoạn DNA được đọc từ hai đầu, cho ra một cặp read.

Có thể bạn chưa biết. Vì sao lại đọc từ hai đầu?

Một đoạn DNA trong thư viện thường dài vài trăm base, trong khi mỗi read chỉ đọc được khoảng 150 base ở một đầu. Với paired-end, máy đọc thêm đầu còn lại của cùng đoạn DNA, nên mỗi đoạn cho ra hai read hướng vào nhau, nằm trên hai mạch ngược chiều và cách nhau một khoảng gần bằng insert size. Nhìn qua thì đây chỉ là gấp đôi lượng dữ liệu, nhưng điều quý nhất là hai read biết “tọa độ tương đối” của nhau, và aligner, tức chương trình thực hiện alignment, có thể dùng thông tin đó theo ba cách (xem hình minh họa bên dưới).



Ba cách dùng này dựa trên cùng một nguyên tắc. Một cặp read được gọi là **concordant** khi hai read nằm đúng chiều, tức hướng vào nhau, và cách nhau một khoảng nằm trong phân bố insert size của thư viện; ngược lại, cặp đó là **discordant**. Khi một read rơi vào vùng lặp, nó có thể map vào nhiều vị trí với điểm alignment gần như nhau, gọi là multiple hits, và nếu chỉ nhìn riêng read đó thì aligner không biết vị trí nào đúng. Lúc này read còn lại trong cặp trở thành thông tin bổ sung. Với mỗi vị trí ứng viên của read thứ nhất, aligner kiểm tra xem read thứ hai có map được ở đúng chiều và đúng khoảng cách insert size hay không. Những vị trí tạo thành một cặp concordant được xếp hạng cao hơn và được gán **mapping quality** cao hơn, tức mức độ tin cậy rằng read được đặt đúng chỗ, còn những vị trí còn lại bị xếp hạng thấp. Paired-end không xóa bỏ hoàn toàn sự mơ hồ, vì nếu cả hai read cùng nằm gọn trong vùng lặp thì các vị trí vẫn ngang nhau, nhưng trong rất nhiều trường hợp nó giúp aligner chọn ra alignment đáng tin cậy nhất. Tương tự, khi một cặp bị discordant mà không có lý do sinh học rõ ràng, aligner hạ mapping quality hoặc đánh dấu cặp đó để các bước sau thận trọng hơn. Các công cụ phát hiện structural variant lại làm điều ngược lại, tức đi tìm những nhóm cặp discordant cùng kiểu lặp lại ở một vùng, vì một cặp discordant đơn lẻ thường chỉ là lỗi, nhưng nhiều cặp cùng lệch một khoảng như nhau thì thường là dấu vết của một đoạn DNA bị mất, bị chèn hay bị đảo trong mẫu so với reference.

Phred quality score. Mỗi base trong read đi kèm một điểm chất lượng, gọi là **Phred quality score**, phản ánh xác suất máy đọc sai base đó theo công thức

$$Q = -10 \log_{10} p,$$

trong đó p là xác suất lỗi. Cách biểu diễn theo thang logarit giúp các con số dễ nhớ (Bảng 2.1). Q20 nghĩa là cứ 100 base thì có khoảng một base sai, Q30 là một trên một nghìn. Dữ liệu Illumina hiện nay thường có phần lớn base đạt Q30 trở lên, và chất lượng có xu hướng giảm dần về cuối read. Từ công thức có thể thấy mỗi khi Q tăng thêm 10 đơn vị, xác suất lỗi giảm đi mười lần, nên khoảng cách giữa Q20 và Q30 lớn hơn nhiều so với cảm giác ban đầu khi nhìn hai con số. Trong thực hành, Phred score được dùng ngay từ những bước xử lý đầu tiên, ví dụ như để cắt bỏ phần cuối read có chất lượng thấp hoặc để giảm trọng số của những base kém tin cậy khi xác định biến thể.

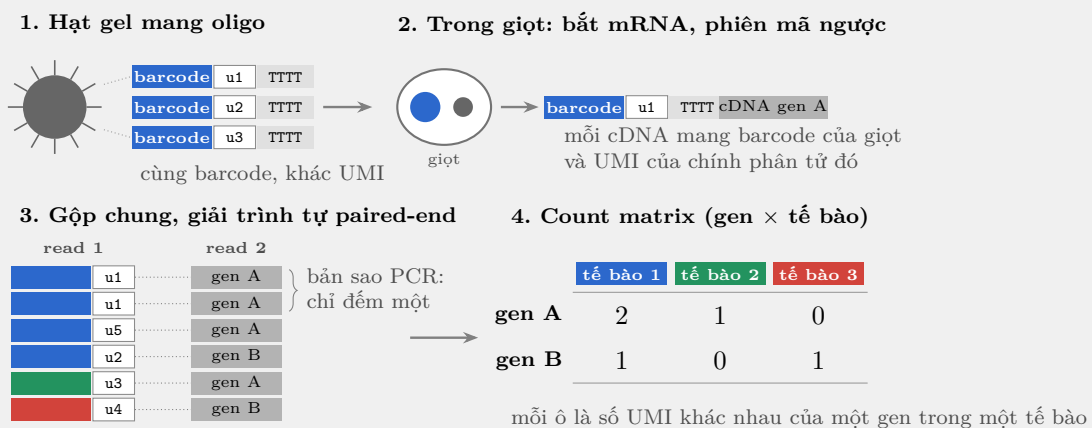
Phred score	Xác suất lỗi	Độ chính xác	Diễn giải
Q10	1/10	90%	chất lượng kém
Q20	1/100	99%	chấp nhận được
Q30	1/1.000	99,9%	ngưỡng “tốt” thường dùng
Q40	1/10.000	99,99%	rất tốt

Bảng 2.1: Phred score và xác suất lỗi tương ứng.

Cần đặt câu hỏi con số xác suất này đến từ đâu và có đáng tin cậy hay không. Máy giải trình tự không “biết” base nào sai, mà chỉ ước tính xác suất lỗi dựa trên đặc điểm tín hiệu, bằng một mô hình thống kê được hiệu chỉnh từ dữ liệu. Nói cách khác, ngay từ bước đầu tiên, dữ liệu hệ gen đã là một kết quả của suy luận thống kê. Chúng ta sẽ quay lại ý này ở phần sau của sách.

Có thể bạn chưa biết. Làm sao biết mỗi read đến từ tế bào nào?

Short read cũng là nền tảng của phần lớn các phương pháp single-cell hiện nay, và mấu chốt nằm ở chỗ mọi phân tử của cùng một tế bào được “đóng dấu” bằng cùng một mã trước khi được trộn chung với các tế bào khác (xem hình minh họa bên dưới). Trong các hệ thống dạng giọt (*droplet*), mỗi hạt gel mang rất nhiều oligo, tức những đoạn DNA ngắn được tổng hợp nhân tạo. Mọi oligo trên cùng một hạt có chung một **cell barcode**, nhưng mỗi oligo lại mang một **UMI** (*unique molecular identifier*) riêng, tức một đoạn trình tự ngẫu nhiên, và tận cùng bằng một chuỗi poly(dT) có thể bắt cặp với đuôi poly(A) của mRNA. Tế bào và hạt gel được đưa vào một thiết bị vi lỏng với nồng độ được điều chỉnh sao cho mỗi giọt thường chỉ chứa nhiều nhất một tế bào và một hạt. Trong giọt, tế bào bị phá vỡ, mRNA được oligo bắt lại và phiên mã ngược thành cDNA, nên mỗi phân tử cDNA mang theo barcode của giọt và UMI của chính nó. Sau đó, các giọt được phá vỡ, toàn bộ cDNA được gộp chung để tạo thư viện và giải trình tự paired-end, trong đó read 1 đọc barcode và UMI, còn read 2 đọc trình tự cDNA.^a Khi phân tích, read được nhóm theo barcode để biết đến từ tế bào nào và được map lên reference để biết thuộc gen nào. Với mỗi cặp tế bào và gen, ta đếm số UMI khác nhau thay vì số read, vì những read có cùng barcode và cùng UMI chỉ là bản sao PCR của một phân tử ban đầu. Kết quả là một count matrix với mỗi hàng là một gen và mỗi cột là một tế bào. Với cách làm này, một lần chạy có thể đo transcriptome của hàng chục nghìn tế bào.^b



^aChen X. và cộng sự. scg_lib_structs, 10x Genomics Single Cell 3' V3.

<https://scg-lib-structs.readthedocs.io/en/latest/ge/10xChromium3v3.html>

^bZheng G.X.Y. và cộng sự (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 8:14049. <https://doi.org/10.1038/ncomms14049>

2.2.3 Long-read sequencing

Short read có một giới hạn mang tính bản chất, đó là read quá ngắn so với nhiều cấu trúc trong hệ gen. Các đoạn lặp, structural variant hay một phân tử RNA hoàn chỉnh thường dài hơn read nhiều lần, nên khi DNA bị cắt thành những mảnh vài trăm base, thông tin về việc các mảnh nằm cạnh nhau ra sao trên phân tử ban đầu phần lớn bị mất. Không có cách tính toán nào khôi phục được trọn vẹn thông tin đã không được đo. Long read ra đời để vượt qua giới hạn đó bằng cách đọc từng phân tử đơn lẻ, dài, thay vì khuếch đại và đọc hàng tỷ mảnh ngắn. Hai nền tảng chính hiện nay dựa trên hai nguyên lý hoàn toàn khác nhau.

PacBio dùng công nghệ single-molecule real-time (SMRT, giải trình tự phân tử đơn theo thời gian thực), trong đó một phân tử polymerase được cố định ở đáy một giếng siêu nhỏ, và mỗi khi gắn một nucleotide phát huỳnh quang, nó tạo ra một tia sáng được ghi lại theo thời gian thực. Điểm đặc biệt nằm ở khuôn DNA, được nối thành vòng tròn để polymerase có thể đọc đi đọc lại cùng một phân tử nhiều lần. Gộp các lần đọc đó lại thành một trình tự đồng thuận (*consensus sequence*), ta được **HiFi read** (high-fidelity read, read có độ chính xác cao), có chiều dài khoảng 15–20 nghìn base, với độ chính xác trên 99,9%.

Oxford Nanopore (ONT) không dùng ánh sáng, mà kéo một mạch DNA qua một lỗ protein cỡ nanomet (*nanopore*) nằm trên màng, và mỗi nhóm vài nucleotide đi qua lỗ làm dòng điện ion thay đổi theo một kiểu đặc trưng. Một mạng neural (*neural network*) sau đó “dịch” tín hiệu dòng điện thành trình tự base, một bước gọi là **basecalling**. Vì không có giới hạn cứng về chiều dài, ONT có thể đọc các phân tử dài hàng trăm nghìn base, thậm chí hơn một triệu base. Thiết bị nhỏ nhất của ONT chỉ bằng chiếc điện thoại và cắm trực tiếp vào máy tính.

Có thể bạn chưa biết. *Lịch sử công nghệ nanopore*

Ý tưởng kéo DNA qua một lỗ nhỏ để đọc trình tự được David Deamer phác thảo trong sổ ghi chép ngày 25 tháng 6 năm 1989.^a Phải mất hơn hai mươi năm nghiên cứu, từ việc chọn loại protein làm lỗ đến kiểm soát tốc độ DNA đi qua, ý tưởng đó mới thành sản phẩm thương mại vào năm 2014. Những thiết bị đầu tiên có tỉ lệ lỗi rất cao, và nhiều người từng nghi ngờ công nghệ này. Ngày nay, cùng với PacBio, nó đã góp phần giải trình tự hoàn chỉnh hệ gen người.

^aDeamer D., Akeson M., Branton D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology* 34:518–524. <https://pmc.ncbi.nlm.nih.gov/articles/PMC6733523/>

Một ưu điểm ít được chú ý của cả hai nền tảng long read là khả năng phát hiện **biến đổi base** (*base modification*) ngay trên phân tử gốc. Với PacBio, một cytosine bị methyl hóa làm polymerase gắn nucleotide chậm hơn một chút. Với ONT, nó làm tín hiệu dòng điện khác đi. Nhờ vậy, ta có thể đọc đồng thời trình tự và methylation từ cùng một lần giải trình tự, mà không cần xử lý hóa học như các phương pháp short read. Ưu điểm này có được chính vì long read đọc phân tử gốc, không qua bước khuếch đại. Khi DNA được nhân bản bằng polymerase, các dấu ấn hóa học trên base không được sao chép sang bản sao, nên mọi phương pháp có bước khuếch đại đều làm mất thông tin này nếu không xử lý từ trước.

Bảng 2.2 tóm tắt đặc điểm của ba cách đọc, và khi đọc bảng, cần nhớ rằng các đặc điểm này mang tính tương đối và thay đổi nhanh theo từng thế hệ thiết bị. Điều ít thay đổi hơn là kiểu lỗi đặc trưng của mỗi nền tảng, vì nó gắn liền với nguyên lý đo. Kiểu lỗi đó quyết định phần mềm phân tích phía sau phải được thiết kế ra sao, nên một công cụ viết cho short read thường không dùng được ngay cho long read và ngược lại.

2.3 Biến thể DNA

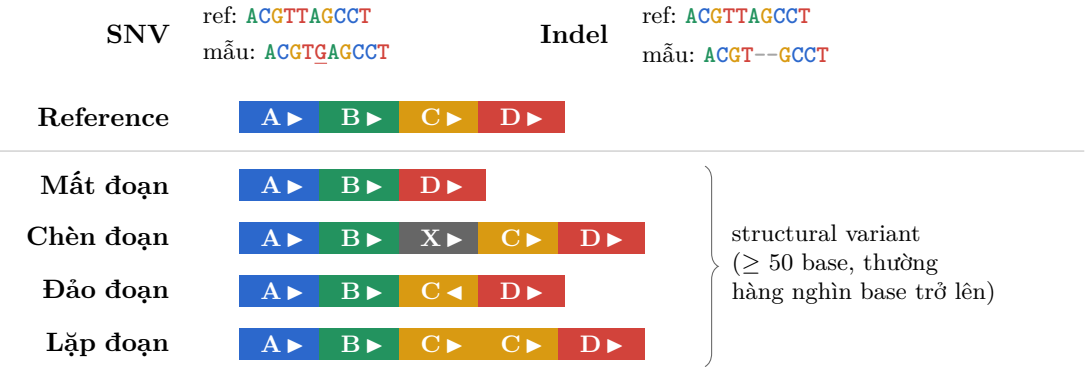
Hai hệ gen người bất kỳ giống nhau khoảng 99,9%. Phần khác biệt nhỏ đó quyết định rất nhiều điều, từ màu tóc đến nguy cơ mắc bệnh. So với reference genome, mỗi người mang khoảng 4–5 triệu biến thể nhỏ, và khoảng 20–30 nghìn structural variant. Các biến thể được chia thành ba nhóm theo kích thước (Hình 2.4). Cách phân nhóm này không chỉ mang tính quy ước. Kích thước của biến thể so với chiều dài read quyết định công nghệ nào nhìn thấy được nó, và vì vậy cũng quyết định phương pháp tính toán nào được dùng để phát hiện nó.

- **SNV** (*single nucleotide variant*) là thay đổi một base. Khi phổ biến trong quần thể, chúng

	Array	Short read	Long read
Nguyên lý	lai phân tử với probe	sequencing by synthesis	đọc phân tử đơn (SMRT, nanopore)
Đơn vị dữ liệu	cường độ tín hiệu tại vị trí định sẵn	read 50–300 base	read hàng chục nghìn base trở lên
Kiểu lỗi chính	gán sai kiểu gen, no-call	thay thế base, tăng dần cuối read	indel, đặc biệt ở đoạn lặp một base (đã giảm nhiều)
Khám phá biến thể mới	không	có	có
Chi phí mỗi mẫu	thấp nhất	trung bình	cao nhất
Điểm mạnh riêng	rẻ, chuẩn hóa, phù hợp cỡ mẫu rất lớn	thông lượng cao, chính xác, hệ sinh thái công cụ phong phú	vùng lặp, structural variant, phasing, isoform, methylation trực tiếp

Bảng 2.2: So sánh ba nền tảng công nghệ. Các đặc điểm mang tính tương đối và thay đổi nhanh theo từng thế hệ thiết bị.

- thường được gọi là SNP.
- **Indel** là thêm hoặc mất một đoạn ngắn, thường dưới 50 base.
 - **Structural variant** (SV) là thay đổi từ 50 base trở lên, gồm mất đoạn (*deletion*), chèn đoạn (*insertion*), đảo đoạn (*inversion*), lặp đoạn (*duplication*) và chuyển đoạn (*translocation*). Dù số lượng ít hơn nhiều so với SNV, tổng số base bị ảnh hưởng bởi SV trong một hệ gen lại lớn hơn.



Hình 2.4: Các nhóm biến thể theo kích thước. Phía trên là SNV và indel, nhìn thấy trực tiếp ở mức trình tự; phía dưới là structural variant, biểu diễn bằng các khối đoạn DNA; mũi tên chỉ chiều của đoạn.

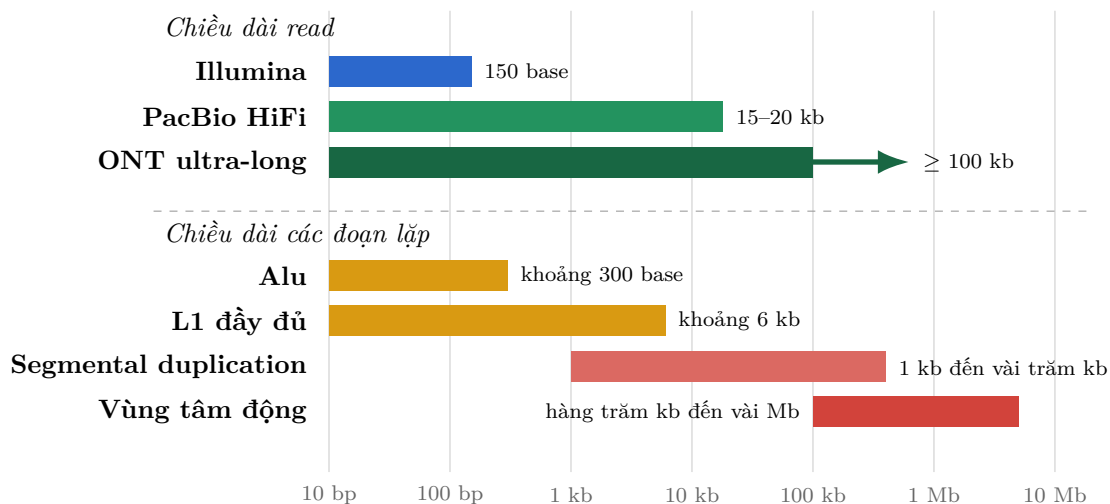
Cả ba cột của ma trận đều có công nghệ để đo biến thể DNA, nhưng mỗi cột nhìn thấy một phần khác nhau. SNP array đo trực tiếp hàng trăm nghìn đến vài triệu SNP đã chọn sẵn. Những biến thể còn lại được suy ra bằng **imputation**, tức dự đoán kiểu gen tại những biến thể không được genotype trực tiếp, trong đó, dựa vào một bảng tham chiếu (*reference panel*) gồm các hệ gen đã được giải trình tự đầy đủ, ta dự đoán kiểu gen tại những vị trí không có trên array. Imputation làm được điều này vì các biến thể nằm gần nhau trên nhiễm sắc thể thường được di truyền cùng nhau thành từng khối qua nhiều thế hệ. Biết allele tại vài vị trí đã đo trong một khối, ta có thể dò trong bảng tham chiếu xem những haplotype

nào phù hợp, tức những tổ hợp allele cùng nằm trên một nhiễm sắc thể, rồi suy ra allele tại các vị trí lân cận. Cần nhớ rằng kết quả imputation là một dự đoán kèm mức độ không chắc chắn, không phải một phép đo, và độ chính xác của nó phụ thuộc vào việc bảng tham chiếu có gần với quần thể đang nghiên cứu hay không.

Short read có nhiều dạng tùy mục đích, trong đó giải trình tự toàn bộ hệ gen (WGS) ở coverage khoảng $30\times$ là chuẩn phổ biến cho nghiên cứu cá thể. Giải trình tự exome (WES) chỉ đọc khoảng 1–2% hệ gen là vùng mã hóa protein, nên rẻ hơn nhiều, phù hợp cho chẩn đoán bệnh di truyền. **Low-pass WGS** đọc toàn bộ hệ gen nhưng ở coverage rất thấp, khoảng $0,5\text{--}4\times$, rồi dùng imputation để lấp khoảng trống; đây là một lựa chọn cạnh tranh trực tiếp với array.

Long read là công cụ mạnh nhất cho structural variant và các vùng lặp. Nó còn cho phép **phasing**, vì có thể xác định trực tiếp hai biến thể nằm trên cùng một nhiễm sắc thể (cùng từ cha hoặc cùng từ mẹ) hay nằm trên hai nhiễm sắc thể khác nhau. Với short read, thông tin này gần như mất hẳn, vì hai biến thể cách nhau vài nghìn base hiếm khi nằm trên cùng một read.

Hình 2.5 giải thích vì sao chiều dài read quan trọng đến vậy. Hệ gen người chứa rất nhiều đoạn lặp, bao gồm hàng triệu bản sao của transposable element Alu dài khoảng 300 base, các transposable element L1 dài khoảng 6 nghìn base, các segmental duplication (vùng lặp đoạn) dài từ vài nghìn đến hàng trăm nghìn base, và các vùng tâm động (*centromere*) dài hàng triệu base. Một read ngắn hơn đoạn lặp sẽ nằm gọn bên trong nó, và ta không thể biết read đó đến từ bản sao nào. Một read dài hơn đoạn lặp sẽ vắt qua cả hai đầu, bám vào trình tự duy nhất ở hai bên, và định vị được chính xác.



Hình 2.5: Chiều dài read so với chiều dài các loại đoạn lặp trong hệ gen người (thang logarit, số liệu xấp xỉ). Short read ngắn hơn gần như mọi loại đoạn lặp; long read vắt qua được phần lớn chúng.

2.4 Transcriptome

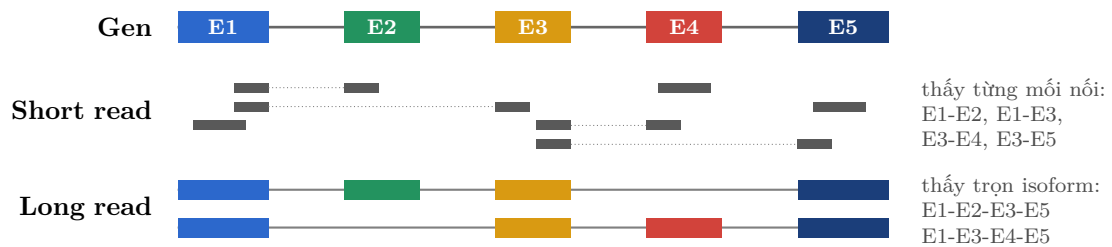
Nếu hệ gen là cuốn sách, thì transcriptome cho biết trang nào đang được đọc, trong mô nào, và nhiều đến mức nào. Đo transcriptome giúp trả lời câu hỏi gen nào đang hoạt động, và hoạt động đó thay đổi ra sao giữa các điều kiện. Khác với trình tự DNA gần như giống nhau ở mọi tế bào của một cơ thể, transcriptome của tế bào gan khác hẳn transcriptome của tế bào thần kinh, và có thể thay đổi nhanh khi tế bào phản ứng với môi trường. Vì vậy, một phép đo transcriptome luôn gắn với một mô, một thời điểm và một điều kiện cụ thể. Về mặt

kỹ thuật, ba cột của ma trận đo tăng này theo những cách khác nhau, và mỗi cách lại để lộ một giới hạn riêng.

Expression microarray, phổ biến từ cuối những năm 1990, đo lượng RNA của các gen đã biết bằng probe, giống như SNP array. Nó rẻ nhưng chỉ đo được gen đã có probe, tín hiệu dễ bão hòa khi gen biểu hiện quá mạnh và lẫn với nhiễu nền khi gen biểu hiện quá yếu. Hệ quả là microarray phù hợp để so sánh tương đối các gen đã biết giữa hai nhóm mẫu, nhưng khó đo chính xác ở hai đầu của dải biểu hiện (*dynamic range*). Đây cũng chính là giới hạn “chỉ đo được những gì đã thiết kế sẵn” của array mà ta đã gặp ở hàng biến thể DNA, nay xuất hiện lại ở tầng RNA.

RNA-seq, xuất hiện khoảng năm 2008, thay thế gần như hoàn toàn microarray. RNA được chuyển thành cDNA rồi giải trình tự bằng short read. Số read rơi vào mỗi gen phản ánh mức biểu hiện của gen đó. So với microarray, RNA-seq đo được trên một dải biểu hiện rộng hơn nhiều và phát hiện được các transcript chưa từng biết. Nhờ các junction read, tức read vắt qua ranh giới giữa hai exon (đoạn của gen được giữ lại trong RNA sau khi intron bị cắt bỏ), nó còn cho thấy cách gen được cắt nối (*splicing*).

Cắt nối khác nhau (*alternative splicing*) khiến một gen có thể tạo ra nhiều **isoform**. Đây là chỗ short read gặp khó (Hình 2.6), vì mỗi read chỉ thấy một hoặc hai exon, nên ta biết được từng mối nối riêng lẻ, nhưng không biết chắc các mối nối đó kết hợp với nhau ra sao trên cùng một phân tử RNA. Long read, với các công nghệ như **Iso-Seq** của PacBio hay **direct RNA sequencing** (giải trình tự RNA trực tiếp) của ONT, đọc trọn một phân tử từ đầu đến cuối, nên cho thấy trực tiếp isoform đầy đủ. Direct RNA sequencing còn phát hiện được các biến đổi hóa học trên RNA, như m6A, tức adenosine được methyl hóa tại vị trí N6.



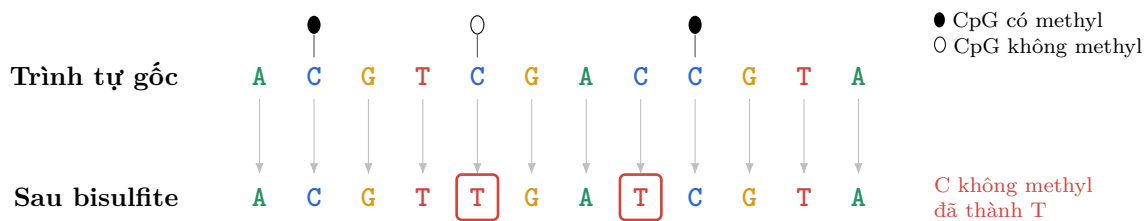
Hình 2.6: Short read cho biết từng mối nối riêng lẻ, nhưng các mối nối này phù hợp với ít nhất bốn tổ hợp isoform khác nhau. Long read đọc trọn phân tử nên xác định được trực tiếp hai isoform thật sự có mặt.

Ngoài mức biểu hiện và isoform, RNA-seq còn cho thấy nhiều dạng biến đổi khác ở mức RNA, và ở đây ta chỉ điểm qua vài dạng chính. **Circular RNA** (circRNA, RNA vòng) là những phân tử RNA có hai đầu nối lại với nhau thành vòng khép kín, được nhận biết qua các read vắt qua một back-splice junction, tức mối nối mà đầu sau của một exon nối ngược về đầu trước của một exon nằm phía trước nó. **Fusion transcript** (transcript dung hợp) là RNA được ghép từ hai gen khác nhau, thường là hệ quả của một structural variant như chuyển đoạn hay mất đoạn trên DNA, và hay gặp trong ung thư. **RNA editing** là hiện tượng một base trên RNA bị biến đổi sau phiên mã, ví dụ như adenosine bị chuyển thành inosine, khiến trình tự RNA khác với trình tự DNA tại vị trí đó. Điểm chung của các dạng này là chúng được phát hiện từ những read có cấu trúc hoặc trình tự khác thường so với reference, trong khi lỗi trong quá trình tạo thư viện hay lỗi alignment ở các vùng trình tự tương đồng cũng có thể tạo ra những read nhìn giống hệt như vậy. Vì vậy, danh sách ứng viên luôn lẫn nhiều kết quả dương tính giả (*false positive*), và cần được lọc bằng các phương pháp thống kê sẽ được đề cập ở phần sau của sách.

2.5 DNA methylation

DNA methylation là một dấu ấn hóa học gắn lên base, không làm thay đổi trình tự nhưng ảnh hưởng đến việc gen có được biểu hiện hay không. Ở động vật có vú, dạng phổ biến nhất là **5-methylcytosine** (5mC, cytosine mang nhóm methyl ở carbon số 5) tại các vị trí CpG, tức một C đứng ngay trước một G. Hệ gen người có khoảng 28 triệu vị trí CpG. Nhìn chung, CpG ở vùng promoter, tức đoạn DNA ở đầu gen nơi bộ máy phiên mã gắn vào để bắt đầu phiên mã, của gen đang hoạt động thường không bị methyl hóa, còn phần lớn CpG ở các vùng khác thì có.

Vấn đề là các máy giải trình tự thông thường không phân biệt được C có methyl và C không có methyl. Giải pháp kinh điển là **xử lý bisulfite** (*bisulfite conversion*) (Hình 2.7), trong đó hóa chất bisulfite biến C không methyl hóa thành U, và sau khi khuếch đại, U được đọc thành T. C đã methyl hóa thì được bảo vệ và vẫn là C. So sánh read với reference genome, vị trí nào vẫn là C thì đã được methyl hóa, vị trí nào thành T thì không.



Hình 2.7: Nguyên lý bisulfite. C được methyl hóa (chấm đen) giữ nguyên; mọi C không methyl hóa, kể cả C không nằm trong CpG, đều bị chuyển thành T.

Ba cột của ma trận tiếp cận methylation theo ba cách. **Methylation array**, như Illumina EPIC, đo khoảng 850 nghìn vị trí CpG chọn sẵn, tức chỉ khoảng 3% tổng số CpG, nhưng rất rẻ và được dùng rộng rãi trong các nghiên cứu dịch tễ quy mô lớn. Với short read, **WGBS** (whole-genome bisulfite sequencing, giải trình tự bisulfite toàn bộ hệ gen) đo gần như toàn bộ CpG nhưng tốn kém, và bisulfite làm DNA bị đứt gãy đáng kể; **RRBS** (reduced representation bisulfite sequencing) chỉ tập trung vào các vùng giàu CpG để giảm chi phí; **EM-seq** (enzymatic methyl-seq) thay bisulfite bằng enzyme để ít làm hư DNA hơn. Long read thì như đã nói ở trên, đọc methylation trực tiếp trên phân tử gốc, không cần chuyển đổi hóa học, và còn cho biết methylation trên từng haplotype.

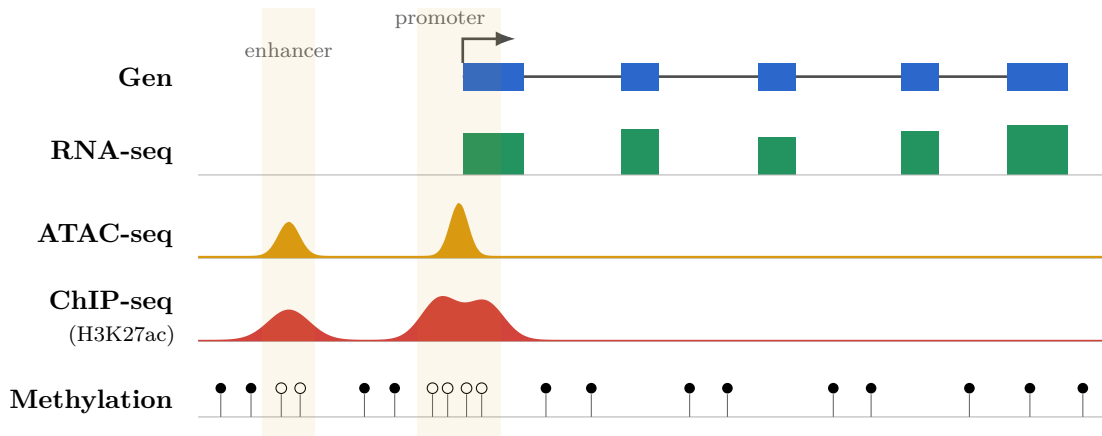
2.6 Chromatin và cấu trúc ba chiều của hệ gen

DNA trong nhân không nằm trần trụi mà quấn quanh các protein histone, tạo thành chromatin. Ở những vùng chromatin đóng chặt, bộ máy phiên mã khó tiếp cận DNA. Ở những vùng chromatin mở, các yếu tố phiên mã có thể gắn vào promoter và enhancer, tức những vùng điều hòa có thể nằm xa gen nhưng làm tăng mức phiên mã của gen đó, để điều hòa gen. Do trạng thái chromatin vì vậy là cách trực tiếp để tìm các vùng điều hòa trong hệ gen.

ChIP (chromatin immunoprecipitation, kết tủa miễn dịch chromatin) dùng kháng thể để kéo xuống một protein cụ thể cùng với đoạn DNA mà nó đang gắn vào, ví dụ như một yếu tố phiên mã hay một dạng histone biến đổi. Đoạn DNA thu được ban đầu được đo bằng array (ChIP-chip), rồi chuyển sang short read (ChIP-seq, từ khoảng 2007). **ATAC-seq** (2013) đi theo hướng khác, dựa trên enzyme transposase Tn5, loại enzyme chỉ chèn được vào những vùng chromatin mở, đồng thời gắn luôn adapter giải trình tự. Số read tập trung ở đâu thì vùng đó đang mở. ATAC-seq nhanh, cần ít mẫu, và trở thành phương pháp phổ biến nhất để

lập bản đồ vùng điều hòa. **CUT&Tag** là một phiên bản cải tiến của ChIP, cần ít tế bào hơn và cho tín hiệu sạch hơn. Ở cột long read, **Fiber-seq** đánh dấu các vùng DNA để tiếp cận bằng một enzyme methyl hóa, rồi đọc bằng long read, cho thấy trạng thái chromatin trên từng phân tử thay vì trung bình của hàng triệu tế bào.

Hình 2.8 minh họa cách các tầng dữ liệu này nhìn ra sao khi đặt cạnh nhau trên cùng một vùng hệ gen, theo kiểu hiển thị của các genome browser, tức phần mềm hiển thị nhiều loại dữ liệu dọc theo tọa độ hệ gen. Mỗi track, tức một hàng hiển thị một loại dữ liệu, được vẽ theo cùng một trục tọa độ hệ gen, nên người đọc có thể dóng thẳng từ trên xuống để so sánh các tầng tại cùng một vị trí. Mỗi tầng dữ liệu riêng lẻ chỉ cung cấp một phần bằng chứng, vì một đỉnh ATAC-seq cho biết chromatin mở nhưng chưa cho biết vùng đó làm gì. Khi nhiều tầng cùng chỉ về một vùng, như chromatin mở, histone marker hoạt động và CpG không bị methyl hóa, ta có cơ sở vững hơn nhiều để xem đó là một vùng điều hòa thật sự.



Hình 2.8: Các tầng dữ liệu trên cùng một vùng hệ gen (minh họa). Tại promoter và enhancer (vùng tô nhạt), chromatin ở trạng thái mở (ATAC-seq), mang histone marker của vùng hoạt động (H3K27ac), và CpG không bị methyl hóa (chấm trắng). Gen phía sau được phiên mã (RNA-seq).

Hàng cuối cùng của ma trận đo cấu trúc 3D của hệ gen. Một enhancer có thể nằm cách gen mà nó điều hòa hàng trăm nghìn base, nhưng trong không gian ba chiều, hai vùng này lại nằm sát nhau. **Hi-C** (2009) nối các đoạn DNA nằm gần nhau trong không gian rồi giải trình tự các cặp nối, tạo ra một ma trận tần suất tiếp xúc (*contact matrix*) giữa mọi cặp vùng trong hệ gen. Ngoài ứng dụng sinh học, dữ liệu Hi-C còn là công cụ quan trọng để sắp xếp các contig, tức các đoạn trình tự liên tục thu được từ assembly, thành nhiễm sắc thể hoàn chỉnh khi xây dựng reference genome. **Pore-C** kết hợp ý tưởng này với nanopore, cho phép thấy nhiều vùng cùng tiếp xúc với nhau trên một phân tử, thay vì chỉ từng cặp.

2.7 So sánh long read và short read

Đọc lại ma trận theo chiều ngang, ta thấy một mẫu hình lặp lại, trong đó cột long read thường giải quyết đúng những chỗ mà cột short read gặp khó. Bảng 2.3 tổng hợp các điểm này, và nếu nhìn kỹ, hầu hết các dòng trong bảng đều quy về cùng một nguyên nhân. Short read cắt phân tử thành những mảnh nhỏ, nên mọi thông tin đòi hỏi biết những gì nằm cạnh nhau trên một khoảng dài, từ vị trí của một đoạn lắp đến tổ hợp exon của một isoform, đều bị mất hoặc chỉ suy ra được gián tiếp. Long read giữ lại thông tin đó vì mỗi read tương ứng với một đoạn dài của chính phân tử ban đầu.

Tuy vậy, short read vẫn chưa bị thay thế, vì long read vẫn đắt hơn cho mỗi mẫu, đòi hỏi DNA có chất lượng cao và phân tử dài, và thông lượng thấp hơn. Với một nghiên cứu cần

Vấn đề	Vì sao short read gặp khó	Long read giải quyết ra sao
Vùng lặp	read nằm gọn trong đoạn lặp, không biết đến từ bản sao nào	read vắt qua đoạn lặp, bám vào trình tự duy nhất hai bên
Structural variant	SV thường dài hơn read, chỉ suy ra gián tiếp từ tín hiệu bất thường	một read chứa trọn SV cùng hai điểm đứt gãy
Assembly	đồ thị lắp ráp đứt gãy tại mọi đoạn lặp dài	contig dài hàng triệu base, dẫn đến T2T
Phasing	hai biến thể xa nhau hiếm khi cùng nằm trên một read	nhiều biến thể trên cùng một read, nối được thành haplotype dài
Isoform	chỉ thấy từng mối nối riêng lẻ	đọc trọn phân tử RNA
Methylation	cần xử lý bisulfite, làm hư DNA	đọc trực tiếp trên phân tử gốc

Bảng 2.3: Những vấn đề của short read mà long read giải quyết được.

hàng nghìn mẫu, short read hay thậm chí array vẫn là lựa chọn hợp lý hơn. Trên thực tế, nhiều nghiên cứu hiện nay kết hợp cả hai, dùng long read trên một số ít mẫu để xây dựng pangenome và danh mục structural variant chất lượng cao, rồi dùng thông tin đó để phân tích dữ liệu short read hoặc array của hàng nghìn mẫu khác.

2.8 Độ phân giải và đo đồng thời nhiều tầng

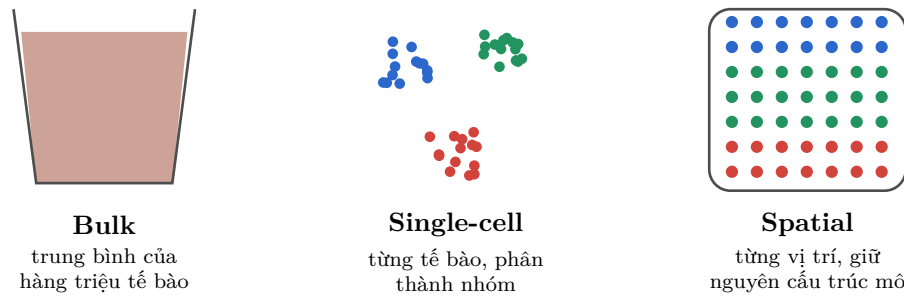
Như đã nói ở đầu chương, ma trận hai trục ở trên chưa thể hiện độ phân giải của phép đo. Hầu hết các phương pháp ở trên được gọi là bulk, vì chúng đo trên một mẫu mô gồm hàng triệu tế bào trộn lẫn, và kết quả là giá trị trung bình. Có thể so sánh dữ liệu bulk với một ly sinh tố trái cây, trong đó ta biết có những thành phần nào nhưng không biết từng miếng trái cây ban đầu ra sao (Hình 2.9).

Single-cell đo riêng từng tế bào, giống như một đĩa trái cây cắt sẵn, trong đó ta thấy được từng loại và đếm được mỗi loại có bao nhiêu. Phổ biến nhất là single-cell RNA-seq, nhưng các tầng khác như chromatin (scATAC-seq) cũng đã đo được ở mức đơn bào. Các công nghệ spatial đi thêm một bước, giữ nguyên vị trí của tế bào trong mô, giống như một chiếc bánh tart trái cây, khi ta không chỉ biết có những loại nào mà còn biết chúng được sắp xếp ra sao.

Mỗi mức độ phân giải lại có nhiều nấc. Ở giữa bulk và single-cell, các tế bào có thể được tách theo loại bằng kháng thể đánh dấu rồi đo chung từng nhóm, cho dữ liệu bulk của một quần thể tế bào tương đối đồng nhất. Ở mức single-cell, nhiều nghiên cứu đo nhân tế bào (*single-nucleus*) thay vì cả tế bào, vì nhân dễ tách hơn từ mô đông lạnh hay từ những mô khó tách thành tế bào rời như mô não, và vì các phép đo chromatin vốn chỉ cần đến nhân. Các công nghệ spatial cũng khác nhau nhiều về độ phân giải. Trên nền tảng Visium của 10x Genomics, mỗi điểm đo có đường kính 55 μm và thường chứa nhiều tế bào, trong khi phiên bản Visium HD dùng một lưới các ô vuông cạnh 2 μm , nhỏ hơn kích thước của một tế bào.² Vì vậy, độ phân giải không phải ba nấc rời rạc mà là một dãy liên tục, và mỗi nấc là một sự đánh đổi giữa mức chi tiết của phép đo, chi phí và lượng tín hiệu thu được trên mỗi đơn vị đo.

Điều quan trọng cần nhận ra là single-cell và spatial không phải một “loại dữ liệu” mới, mà là cách đo lại các hàng quen thuộc của ma trận ở độ phân giải cao hơn. Cái giá phải trả là dữ liệu của mỗi tế bào rất thưa (*sparse*), vì nhiều gen đang biểu hiện nhưng không được

²10x Genomics. Space Ranger glossary. <https://www.10xgenomics.com/support/software/space-ranger/latest/getting-started/space-ranger-glossary>



Hình 2.9: Ba mức độ phân giải. Cùng một mô, dữ liệu bulk cho giá trị trung bình, single-cell tách riêng từng loại tế bào, còn spatial cho biết thêm vị trí của chúng trong mô.

ghi nhận, đơn giản vì lượng RNA trong một tế bào quá nhỏ. Một lần nữa, giới hạn của phép đo lại đặt ra yêu cầu mới cho phân tích.

Xu hướng thứ hai là đo đồng thời nhiều tầng phân tử trên cùng một tế bào, thường được gọi là **multiome** hay single-cell multi-omics. Ví dụ phổ biến nhất là đo cùng lúc RNA và độ mở chromatin trong từng nhân tế bào, như phương pháp SHARE-seq,³ và sau đó là các bộ kit thương mại đo ATAC-seq và RNA-seq trên cùng một nhân. Các phương pháp ba tầng (*trimodal*) như TEA-seq và DOGMA-seq còn đo thêm lượng protein trên bề mặt tế bào bằng những kháng thể được gắn barcode, tức là đi qua cả ba tầng của central dogma, từ chromatin đến RNA và protein, trong cùng một tế bào.⁴

Lý do phải đo cùng lúc nằm ở chỗ dữ liệu từ những tế bào khác nhau không thể ghép nối trực tiếp với nhau. Nếu RNA được đo trên một nhóm tế bào và chromatin trên một nhóm khác, ta chỉ có thể ghép hai bộ dữ liệu bằng thuật toán, dựa trên giả định rằng cùng một loại tế bào thì có trạng thái tương ứng ở cả hai tầng. Giả định này thường hợp lý ở mức loại tế bào, nhưng không cho biết trong từng tế bào cụ thể, một vùng chromatin mở có thật sự đi kèm với việc gen lân cận được biểu hiện hay không. Khi hai tầng được đo trên cùng một tế bào, mối liên hệ đó được quan sát trực tiếp. Nhờ vậy, ta có thể dùng RNA, vốn dễ diễn giải hơn, để xác định loại tế bào, rồi gán loại tế bào đó cho dữ liệu chromatin vốn thừa hơn và khó phân loại hơn. Ta cũng có thể nối từng vùng điều hòa với gen mà nó điều khiển, hay quan sát thứ tự các sự kiện trong quá trình biệt hóa. Ví dụ như với SHARE-seq, các tác giả nhận thấy ở nhiều gen quan trọng cho sự biệt hóa, chromatin mở ra trước khi gen được biểu hiện, gợi ý rằng chromatin “chuẩn bị” cho tế bào trước khi nó thật sự chuyển sang trạng thái mới.

Ở mức mô, cùng một logic dẫn đến việc kết hợp dữ liệu spatial với hình ảnh mô học. Trong spatial transcriptomics, lát cắt mô thường được nhuộm H&E (*hematoxylin and eosin*), tức cách nhuộm mô chuẩn mà các bác sĩ giải phẫu bệnh dùng hằng ngày để quan sát hình thái tế bào và cấu trúc mô, và được chụp ảnh trước khi RNA từ chính lát cắt đó được bắt lên các điểm mang barcode vị trí.⁵ Vì hình ảnh và dữ liệu RNA đến từ cùng một lát cắt, mỗi điểm đo có thể được đặt đúng vào vùng mô tương ứng trên ảnh, ví dụ như vùng khối u, vùng mô đệm hay vùng có tế bào miễn dịch thâm nhiễm. Hình thái mô cho biết một vùng nhìn ra sao dưới kính hiển vi, còn dữ liệu RNA cho biết các tế bào ở đó đang làm gì, và chỉ

³Ma S. và cộng sự (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 183:1103–1116. <https://doi.org/10.1016/j.cell.2020.09.056>

⁴Swanson E. và cộng sự (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* 10:e63632, <https://doi.org/10.7554/eLife.63632>; Mimitou E.P. và cộng sự (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology* 39:1246–1258. <https://doi.org/10.1038/s41587-021-00927-2>

⁵Ståhl P.L. và cộng sự (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353:78–82. <https://doi.org/10.1126/science.aaf2403>

khi có cả hai, ta mới nói được những gì nhà giải phẫu bệnh quan sát với cơ chế phân tử phía sau. Gần đây, các phương pháp spatial còn đo được đồng thời độ mở chromatin hay histone marker cùng với RNA trên cùng một lát cắt, ở độ phân giải gần mức tế bào.⁶

Cái giá của các phép đo kết hợp này là dữ liệu phức tạp hơn nhiều. Mỗi tầng có kiểu nhiễu và độ thưa riêng, số đặc trưng đo được trên mỗi tế bào tăng lên nhiều lần, và việc phân tích đòi hỏi những phương pháp có thể xử lý nhiều tầng cùng lúc thay vì từng tầng riêng lẻ. Đây là một trong những hướng phát triển nhanh nhất của hệ gen học tính toán hiện nay, và sẽ được đề cập ở phần sau của sách.

2.9 Lựa chọn công nghệ theo câu hỏi nghiên cứu

Với một ngân sách cố định, người thiết kế nghiên cứu luôn phải đánh đổi giữa ba yếu tố là số mẫu, lượng thông tin trên mỗi mẫu, và độ chính xác. Không có lựa chọn nào tốt nhất cho mọi trường hợp; chỉ có lựa chọn phù hợp nhất với câu hỏi. Khi câu hỏi liên quan đến những hiệu ứng nhỏ phân bố trong quần thể, số mẫu thường quan trọng hơn mọi thứ khác, vì chỉ có cỡ mẫu lớn mới tách được tín hiệu yếu khỏi dao động ngẫu nhiên. Ngược lại, khi cần biết đầy đủ cấu trúc hệ gen của từng cá thể, lượng thông tin trên mỗi mẫu lại được ưu tiên, dù phải chấp nhận ít mẫu hơn. Bảng 2.4 đưa ra một vài ví dụ cho thấy cùng một logic đánh đổi này dẫn đến những thiết kế rất khác nhau.

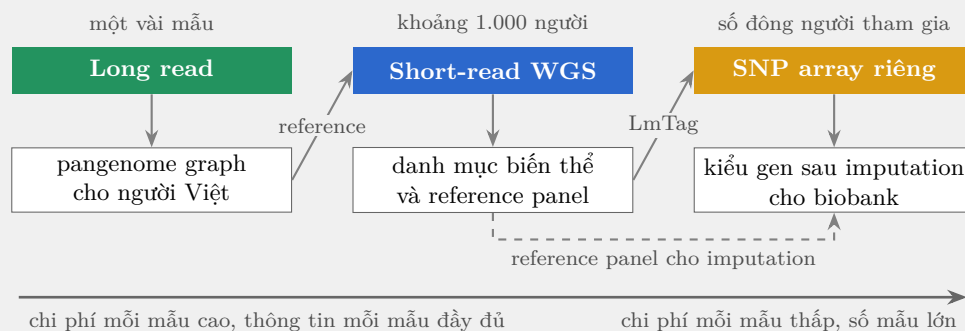
Câu hỏi	Thiết kế thường dùng	Lý do
GWAS cho một tính trạng phức tạp	SNP array và imputation trên hàng trăm nghìn người	hiệu ứng của từng biến thể rất nhỏ, cần cỡ mẫu cực lớn; chi phí mỗi mẫu phải thấp
Chọn giống vật nuôi (genomic selection)	SNP array mật độ trung bình	cần kiểu gen của rất nhiều cá thể mỗi thế hệ, với chi phí thấp
Danh mục structural variant, pangenome	long read trên vài chục đến vài trăm mẫu đa dạng	cần độ đầy đủ tối đa cho từng mẫu; số mẫu ít hơn là chấp nhận được
eQTL, vùng điều hòa	kiểu gen kết hợp RNA-seq, ATAC-seq trên cùng cá thể	cần nhiều tầng dữ liệu trên cùng một người hoặc con vật
Tìm nguyên nhân bệnh di truyền hiếm	WES hoặc WGS	biến thể gây bệnh thường hiếm, không có sẵn trên array
Tìm các loại tế bào trong một mô, kể cả loại hiếm	single-cell RNA-seq trên hàng nghìn đến hàng trăm nghìn tế bào	bulk RNA-seq chỉ cho trung bình của mọi tế bào, nên các loại tế bào hiếm bị che khuất
Cơ chế điều hòa trong từng loại tế bào	multiome (RNA và ATAC trên cùng một tế bào)	nổi vùng chromatin mở với gen mà nó điều hòa ngay trong cùng một tế bào

Bảng 2.4: Một vài câu hỏi nghiên cứu và thiết kế dữ liệu tương ứng.

⁶Zhang D. và cộng sự (2023). Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* 616:113–122. <https://doi.org/10.1038/s41586-023-05795-1>

Nghiên cứu thực tế. Dự án 1000 hệ gen người Việt (VN1K)

Dự án 1000 hệ gen người Việt (VN1K) là một ví dụ cho thấy bài toán đánh đổi ở trên có thể được giải bằng cách phối hợp cả ba cột của ma trận công nghệ thay vì chọn một (sơ đồ bên dưới).^a Long read (PacBio và ONT) chỉ được dùng cho một vài mẫu, đủ để lắp ráp các hệ gen chất lượng cao và xây dựng một pangenome graph, tức reference dạng đồ thị chứa sẵn các biến thể của quần thể, cho người Việt. Pangenome này sau đó được dùng làm reference để phân tích dữ liệu short-read WGS ở độ sâu cao của khoảng một nghìn người, giúp phát hiện structural variant đầy đủ hơn đáng kể so với khi chỉ dùng GRCh38, và tạo ra một danh mục biến thể cùng một reference panel cho imputation phù hợp với người Việt. Dữ liệu SNP từ short read lại là nguyên liệu để thiết kế một SNP array riêng cho người Việt, vì phần lớn array thương mại được thiết kế dựa trên dữ liệu người gốc châu Âu nên hiệu quả imputation có thể giảm khi dùng cho quần thể châu Á như Việt Nam.^b Vì số vị trí trên chip có hạn, bài toán là chọn những **tag SNP**, tức những SNP mà kiểu gen của chúng đủ để suy ra kiểu gen của nhiều SNP lân cận, sao cho độ chính xác của imputation trên toàn hệ gen là cao nhất, và đây chính là bài toán tối ưu mà thuật toán LmTag được phát triển để giải.^c



Tầm nhìn phía sau cách làm này là xây dựng hạ tầng cho một **biobank**, tức kho lưu trữ mẫu sinh học kèm dữ liệu di truyền và dữ liệu sức khỏe của một số lượng rất lớn người tham gia. Không công nghệ nào vừa đầy đủ vừa rẻ, vì vậy mỗi công nghệ được đặt vào đúng chỗ của nó. Long read đầy đủ nhất nhưng đắt nhất, nên chỉ dùng cho vài mẫu để dựng reference. Short read cân bằng giữa độ đầy đủ và chi phí, nên dùng cho khoảng một nghìn người để mô tả biến thể của quần thể. Array rẻ nhất, nên dùng cho số đông, còn phần thông tin array không đo được thì lấy lại bằng imputation từ reference panel. Mục tiêu cuối cùng là dữ liệu nghiên cứu tốt nhất với chi phí thấp nhất có thể, và bài học chung cho người làm hệ gen học là thiết kế một dự án lớn, xét cho cùng, là bài toán phân bổ ngân sách giữa các công nghệ dựa trên điểm mạnh và điểm yếu của từng loại.

^aTran T.T.H. và cộng sự (2026). VN1K is a pangenome-informed multi-omics and phenomics resource for the Vietnamese population. *Nature Communications* 17. <https://doi.org/10.1038/s41467-026-75375-0>

^bNguyen D.T. và cộng sự (2022). A comprehensive evaluation of polygenic score and genotype imputation performances of human SNP arrays in diverse populations. *Scientific Reports* 12:17556. <https://doi.org/10.1038/s41598-022-22215-y>

^cNguyen D.T., Nguyen Q.H., Duong N.T., Vo N.S. (2022). LmTag: functional-enrichment and imputation-aware tag SNP selection for population-specific genotyping arrays. *Briefings in Bioinformatics* 23:bbac252. <https://doi.org/10.1093/bib/bbac252>

2.10 Dữ liệu hệ gen công cộng

Để bắt đầu nghiên cứu, người học không nhất thiết phải có máy giải trình tự hay kinh phí lớn. Một lượng khổng lồ dữ liệu hệ gen đã được công bố và có thể tải về miễn phí, hoặc đăng ký sử dụng. Các nguồn này trải rộng trên nhiều hàng của ma trận, từ biến thể DNA đến biểu hiện gen và chromatin, và nhiều nguồn đo nhiều tầng trên cùng một nhóm cá thể. Với sinh viên, đây là cách thực tế nhất để làm quen với dữ liệu thật trước khi tự thiết kế một thí nghiệm, và Bảng 2.5 liệt kê một vài nguồn tiêu biểu.

Nguồn dữ liệu	Nội dung chính
Ensembl	reference genome, annotation gen (GTF, GFF3), trình tự cDNA và protein, dữ liệu biến thể cho hàng nghìn loài; phát hành theo từng release
1000 Genomes Project	hệ gen của hơn 2.500 người thuộc 26 quần thể, trong đó có người Kinh ở TP. Hồ Chí Minh; bảng tham chiếu kinh điển cho imputation
gnomAD	tần số biến thể trong exome và hệ gen của hàng trăm nghìn người
UK Biobank	kiểu gen, WGS và dữ liệu kiểu hình của khoảng 500 nghìn người (cần đăng ký)
GTEX	RNA-seq của hàng chục loại mô từ gần một nghìn người hiến tặng, kèm kiểu gen; nguồn dữ liệu eQTL quan trọng
ENCODE	dữ liệu chromatin, ChIP-seq, ATAC-seq, RNA-seq cho người và chuột
Human Pangenome Reference Consortium	hệ gen long read chất lượng cao của nhiều người, dùng xây dựng pangenome
FAANG	dữ liệu chú giải chức năng hệ gen cho vật nuôi và cá
SRA, ENA	kho lưu trữ dữ liệu giải trình tự thô từ hầu hết các bài báo đã công bố

Bảng 2.5: Một vài nguồn dữ liệu hệ gen công cộng tiêu biểu.

Bên cạnh dữ liệu của từng dự án, có một loại nguồn mà gần như dự án nào cũng cần đến, đó là reference genome và annotation đi kèm. Trong số các nguồn này, **Ensembl** là một trong những nguồn được dùng nhiều nhất.⁷ Ensembl tích hợp dữ liệu hệ gen của hàng nghìn loài sinh vật nhân thực, từ người, chuột đến vật nuôi, cá và cây trồng, và với mỗi loài, cung cấp gần như đủ những file mà một pipeline phân tích cần, gồm trình tự hệ gen dạng FASTA, trình tự cDNA, ncRNA và protein, file annotation GTF hoặc GFF3 cho biết mỗi gen, transcript và exon nằm ở đâu, cùng với dữ liệu biến thể và annotation vùng điều hòa cho một số loài. Đây chính là những thành phần của reference theo nghĩa rộng đã nói ở phần trước của sách. Muốn alignment read RNA-seq lên hệ gen, ta cần FASTA của hệ gen cùng file GTF, còn muốn định lượng biểu hiện trực tiếp trên từng transcript, ta cần trình tự cDNA. Ngoài trang web để tra cứu từng gen, Ensembl còn có BioMart để lọc và tải dữ liệu theo nhu cầu, và Variant Effect Predictor (VEP) để dự đoán tác động của từng biến thể lên gen và protein. Dữ liệu Ensembl được phát hành theo từng phiên bản đánh số (*release*), và một thói quen tốt là luôn ghi lại release đã dùng, đồng thời lấy FASTA và GTF từ cùng một release và cùng một bản lắp ráp hệ gen (*genome assembly*).

Rất nhiều bài báo có giá trị được xây dựng hoàn toàn từ dữ liệu công cộng. Kỹ năng tìm, đánh giá và kết hợp các bộ dữ liệu có sẵn là một trong những kỹ năng quan trọng nhất của người làm hệ gen học tính toán. Kỹ năng này đòi hỏi chính những hiểu biết đã trình bày

⁷Dyer S.C. và cộng sự (2025). Ensembl 2025. *Nucleic Acids Research* 53:D948–D957. <https://doi.org/10.1093/nar/gkae1071>

trong chương. Trước khi dùng một bộ dữ liệu, người phân tích cần biết nó được tạo ra bằng công nghệ nào, đo tầng phân tử nào, ở độ phân giải nào, và vì vậy mang những giới hạn gì. Hai bộ dữ liệu đo cùng một tầng nhưng bằng hai nền tảng khác nhau, hay được tạo ra ở hai phòng thí nghiệm khác nhau, thường không thể gộp lại một cách đơn giản mà cần được xử lý để có thể so sánh với nhau.

2.11 Giới hạn của dữ liệu và các bài toán tính toán

Đến đây, chúng ta đã đi hết ma trận công nghệ, và đã đến lúc đọc nó theo chiều ngược lại. Thay vì hỏi mỗi công nghệ làm được gì, phần này hỏi mỗi công nghệ không làm được gì. Câu hỏi nghe có vẻ bi quan, nhưng với người làm tính toán, đó lại là câu hỏi quan trọng nhất. Những gì công nghệ đo tốt thì chỉ cần xử lý theo quy trình chuẩn, và phần lớn quy trình đó đã được người đi trước viết sẵn thành công cụ. Còn những gì công nghệ đo không tới, hoặc đo không chắc, thì phải được suy luận, ước lượng hoặc mô hình hóa, và đó chính là phần việc chính của hệ gen học tính toán. Nói cách khác, giới hạn của dữ liệu thường quyết định công việc của người phân tích nhiều hơn là điểm mạnh của nó. Bảng 2.6 tổng hợp các giới hạn chính đã gặp trong chương, cùng với bài toán tính toán mà mỗi giới hạn sinh ra.

Giới hạn của dữ liệu	Bài toán tính toán sinh ra	Ví dụ phương pháp
Array chỉ đo vị trí định sẵn	suy ra kiểu gen tại vị trí không đo (imputation); chọn vị trí tối ưu để đo	Beagle, Minimac
Low-pass WGS có coverage thấp	imputation từ dữ liệu thừa và không chắc chắn	GLIMPSE, QUILT
Short read khó thấy structural variant	suy ra SV gián tiếp; imputation SV; ước tính SV đóng góp bao nhiêu vào tính trạng	imputation dựa trên reference panel có SV
Read ngắn hơn đoạn lặp; hệ gen nhân đôi	read map vào nhiều vị trí; alignment mơ hồ	mapping quality, mô hình multi-mapping
Mọi phép đo đều có lỗi và nhiễu	định lượng bất định; phân biệt tín hiệu thật với giả; kiểm soát false positive	Phred score, genotype likelihood, FDR
Kiểm định hàng triệu giả thuyết cùng lúc	kiểm soát FDR	Bonferroni, Benjamini-Hochberg
Số mẫu nhỏ	tăng power bằng cách tận dụng thêm thông tin	RASQUAL
Nhiều kỹ thuật giữa các đợt thí nghiệm	tách nhiễu kỹ thuật khỏi tín hiệu sinh học	hiệu chỉnh batch effect
Dữ liệu single-cell rất thưa	mô hình hóa dữ liệu đếm thưa và nhiễu nhiều	mô hình thống kê cho dữ liệu đếm
Không thể kiểm chứng thực nghiệm từng biến thể	dự đoán tác động của biến thể từ trình tự	mô hình sequence-to-function (ví dụ như Enformer)

Bảng 2.6: Giới hạn của dữ liệu và bài toán tính toán tương ứng. Cột cuối chỉ nêu một vài ví dụ tiêu biểu, không phải danh sách đầy đủ.

Nhìn vào bảng này, có thể thấy phần lớn các phương pháp trong hệ gen học tính toán không ra đời một cách tùy ý, mà mỗi phương pháp là câu trả lời cho một giới hạn cụ thể của dữ liệu. Lấy dòng đầu tiên làm ví dụ. Array chỉ đo được vài trăm nghìn vị trí, trong khi một nghiên cứu GWAS muốn biết kiểu gen ở hàng triệu vị trí. Khoảng trống đó sinh ra cả một

hướng nghiên cứu về imputation, và khi imputation đã đủ tốt, câu hỏi lại lật ngược thành nên đặt những vị trí nào lên array để imputation chính xác nhất, như ta đã thấy ở ví dụ về VN1K. Một giới hạn của công nghệ, nhìn từ phía người làm tính toán, vì vậy thường là điểm khởi đầu của cả một chuỗi bài toán nối tiếp nhau. Khi gặp một công cụ mới, câu hỏi đầu tiên nên đặt ra là “nó sinh ra để giải quyết giới hạn nào?”. Trả lời được câu hỏi đó, người đọc sẽ hiểu công cụ nhanh hơn, biết khi nào nó phù hợp, khi nào không, và quan trọng hơn cả, biết nó có thể sai ở đâu.

Bảng này cũng thay đổi theo thời gian. Khi long read ngày càng rẻ, những dòng như “short read khó thấy structural variant” sẽ dần bớt quan trọng, giống như cách short read đã làm cho nhiều bài toán của thời kỳ microarray trở nên ít quan trọng. Tuy vậy, mỗi công nghệ mới lại mang theo những giới hạn mới. Dữ liệu single-cell và spatial cho ta độ phân giải mà mười năm trước khó hình dung, nhưng đổi lại là dữ liệu thưa hơn, nhiễu hơn và đắt hơn trên mỗi mẫu, và chính những đặc điểm đó lại sinh ra một thể hệ phương pháp tính toán mới. Vì vậy, thứ đáng học nhất không phải là danh sách công cụ hiện tại, vốn sẽ thay đổi sau vài năm, mà là thói quen luôn hỏi dữ liệu trong tay mình không nhìn thấy gì, và có thể sai ở đâu. Trong thực tế, có những giới hạn không đến từ công nghệ mà đến từ điều kiện nghiên cứu, và cỡ mẫu nhỏ là ví dụ phổ biến nhất.

Nghiên cứu thực tế. *eQTL* khi cỡ mẫu nhỏ

Dòng “số mẫu nhỏ” trong Bảng 2.6 là một thực tế mà rất nhiều nghiên cứu phải đối mặt, nhất là khi câu hỏi đòi hỏi kết hợp nhiều hàng của ma trận trên cùng một nhóm cá thể. Một *eQTL* là một biến thể DNA có liên quan đến mức biểu hiện của một gen. Tìm *eQTL* đòi hỏi đo cả hàng thứ nhất (kiểu gen) lẫn hàng thứ hai (RNA-seq) trên cùng một nhóm cá thể. Đặt thêm dữ liệu ATAC-seq vào, ta có thể xem biến thể đó có nằm trong một vùng chromatin mở hay không, và từ đó đặt giả thuyết về cơ chế. Ở vật nuôi như cá hồi Đại Tây Dương, khó khăn lớn là số mẫu có đủ các tầng dữ liệu thường khá nhỏ, làm giảm power thống kê, tức khả năng phát hiện một hiệu ứng có thật. Các phương pháp như RASQUAL, được đóng gói thành pipeline nf-RASQUAL cho dữ liệu vật nuôi,^a tận dụng thêm thông tin từ biểu hiện đặc hiệu allele (*allele-specific expression*) để bù lại phần nào cỡ mẫu nhỏ. Các dự án như FAANG (Functional Annotation of Animal Genomes) đang tạo ra loại dữ liệu nhiều tầng này cho nhiều loài vật nuôi.

^aKumasaka N., Knights A.J., Gaffney D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics* 48:206–213. <https://doi.org/10.1038/ng.3467>

^bNguyen D.T. và cộng sự (2025). A nextflow pipeline for molecular quantitative trait loci mapping in small sample size datasets with an application in Atlantic salmon. *BMC Genomics* 26:1070. <https://doi.org/10.1186/s12864-025-12302-5>

Đọc lại cả bảng lẫn ví dụ *eQTL* ở trên, ta sẽ thấy một nguyên lý chung đúng với gần như mọi dòng, đó là dữ liệu hệ gen không bao giờ hoàn hảo, và ta cần một ngôn ngữ để nói về sự không hoàn hảo đó một cách chính xác. Lỗi đo lường, dữ liệu thiếu, nhiễu kỹ thuật, cỡ mẫu nhỏ hay hàng triệu phép kiểm định cùng lúc, nghe thì khác nhau, nhưng đều là những câu hỏi của thống kê. Điều này thật ra đã xuất hiện ngay từ đầu chương, khi bản thân mỗi base trong một read đã là một ước lượng kèm xác suất sai, được ghi lại bằng Phred score. Trong hệ gen học, một con số không kèm theo mức độ không chắc chắn của nó gần như là một con số chưa hoàn chỉnh. Hiểu dữ liệu được tạo ra như thế nào vì vậy mới là bước đầu tiên, còn biết cách lý luận một cách định lượng về sự không chắc chắn của nó là bước thứ hai, và đó cũng là nội dung của phần tiếp theo.

Tóm tắt chương

- Các công nghệ tạo dữ liệu hệ gen có thể xếp vào một ma trận hai trục gồm tầng phân tử được đo (biến thể DNA, transcriptome, methylation, chromatin, cấu trúc 3D) và nền tảng đọc (array, short read, long read).
- Array rẻ và chuẩn hóa nhưng chỉ đo được vị trí định sẵn. Short read có thông lượng cao và chính xác nhưng gặp khó ở vùng lặp, structural variant, phasing và isoform. Long read giải quyết phần lớn các vấn đề đó nhưng đắt hơn.
- Phred score biểu diễn xác suất lỗi của mỗi base trên thang logarit. Ngay từ bước đầu, dữ liệu giải trình tự đã là kết quả của suy luận thống kê.
- Độ phân giải là chiều thứ ba của phép đo, trải dài từ bulk, qua single-cell và single-nucleus, đến spatial với các điểm đo từ vài tế bào xuống nhỏ hơn một tế bào. Xu hướng hiện nay là đo đồng thời nhiều tầng trên cùng một tế bào hay cùng một lát cắt mô (multiome), vì chỉ khi đó các tầng mới được nối trực tiếp với nhau.
- Chọn công nghệ là đánh đổi giữa số mẫu, lượng thông tin trên mỗi mẫu và độ chính xác, tùy theo câu hỏi nghiên cứu.
- Mỗi giới hạn của dữ liệu sinh ra một bài toán tính toán. Hiểu giới hạn là cách nhanh nhất để hiểu phương pháp.

Thuật ngữ

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>5-methylcytosine (5mC)</i>	–	Cytosine mang nhóm methyl ở carbon số 5, dạng DNA methylation phổ biến nhất ở động vật có vú
<i>adapter</i>	–	Đoạn trình tự biết trước gắn vào hai đầu mỗi đoạn DNA khi tạo thư viện
<i>allele-specific expression</i>	biểu hiện đặc hiệu allele	Mức biểu hiện khác nhau giữa hai allele của cùng một gen trong một cá thể dị hợp
<i>alternative splicing</i>	cắt nối khác nhau	Cùng một gen được cắt nối theo nhiều cách, tạo ra nhiều isoform
<i>array</i>	chip	Nền tảng đo dựa trên lai phân tử giữa mẫu và các probe đã thiết kế sẵn, chỉ đo được những vị trí định trước
<i>ascertainment bias</i>	–	Sai lệch do các vị trí trên array được chọn dựa trên một số quần thể nhất định, nên bỏ sót biến thể riêng của quần thể khác
<i>back-splice junction</i>	–	Mối nối mà đầu sau của một exon nối ngược về đầu trước của một exon nằm phía trước nó; dấu hiệu nhận biết circRNA
<i>base modification</i>	biến đổi base	Dấu ấn hóa học trên base, ví dụ 5mC, không làm thay đổi trình tự
<i>basecalling</i>	–	Chuyển tín hiệu thô của máy giải trình tự thành trình tự base

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>biobank</i>	–	Kho lưu trữ mẫu sinh học kèm dữ liệu di truyền và dữ liệu sức khỏe của một số lượng rất lớn người tham gia, phục vụ nhiều nghiên cứu khác nhau
<i>bisulfite conversion</i>	xử lý bisulfite	Biến C không methyl hóa thành U (đọc thành T sau khuếch đại), trong khi C đã methyl hóa giữ nguyên
<i>bulk</i>	–	Đo trên một mẫu gồm hàng triệu tế bào trộn lẫn, cho giá trị trung bình
<i>cell barcode</i>	–	Đoạn trình tự ngắn gắn vào mọi phân tử của cùng một tế bào, giúp biết mỗi read đến từ tế bào nào
<i>ChIP (chromatin immunoprecipitation)</i>	kết tủa miễn dịch chromatin	Dùng kháng thể kéo xuống một protein cùng với đoạn DNA mà nó đang gắn vào
<i>circular RNA (circRNA)</i>	RNA vòng	Phân tử RNA có hai đầu nối lại thành vòng khép kín
<i>cluster generation</i>	khuếch đại thành cụm	Nhân mỗi đoạn DNA tại chỗ trên flow cell thành một cụm gồm hàng nghìn bản sao giống hệt nhau
<i>clustering</i>	phân cụm	Chia các điểm dữ liệu thành những nhóm gồm các điểm giống nhau
<i>concordant, discordant</i>	–	Cặp read nằm đúng chiều và cách nhau đúng khoảng insert size (concordant), hoặc không (discordant)
<i>consensus sequence</i>	trình tự đồng thuận	Trình tự gộp từ nhiều lần đọc của cùng một phân tử, giúp loại bỏ lỗi ngẫu nhiên
<i>contact matrix</i>	ma trận tần suất tiếp xúc	Ma trận cho biết mỗi cặp vùng trong hệ gen tiếp xúc với nhau trong không gian thường xuyên đến mức nào
<i>contig</i>	–	Đoạn trình tự liên tục thu được từ assembly
<i>deletion, insertion, inversion, duplication, translocation</i>	mất đoạn, chèn đoạn, đảo đoạn, lặp đoạn, chuyển đoạn	Các dạng structural variant
<i>direct RNA sequencing</i>	giải trình tự RNA trực tiếp	Đọc trực tiếp phân tử RNA bằng nanopore, cho thấy isoform đầy đủ và biến đổi hóa học trên RNA
<i>enhancer</i>	–	Vùng điều hòa có thể nằm xa gen, làm tăng phiên mã của gen mà nó điều hòa
<i>exon</i>	–	Đoạn của gen được giữ lại trong RNA sau khi intron bị cắt bỏ
<i>false positive</i>	dương tính giả	Kết quả được báo cáo là có nhưng thực tế không có
<i>flow cell</i>	–	Tấm kính có các kênh nhỏ nơi phản ứng giải trình tự short read diễn ra
<i>fusion transcript</i>	transcript dung hợp	RNA được ghép từ hai gen khác nhau, thường do một structural variant trên DNA
<i>genome browser</i>	–	Phần mềm hiển thị nhiều loại dữ liệu dọc theo tọa độ hệ gen
<i>genotype calling</i>	gán kiểu gen	Gán kiểu gen cho từng mẫu từ dữ liệu đo, ví dụ từ cường độ tín hiệu của array

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>H&E (hematoxylin and eosin)</i>	–	Cách nhuộm mô chuẩn trong giải phẫu bệnh, cho thấy hình thái tế bào và cấu trúc mô
<i>haplotype</i>	–	Tổ hợp các allele cùng nằm trên một nhiễm sắc thể
<i>Hi-C</i>	–	Nối các đoạn DNA nằm gần nhau trong không gian rồi giải trình tự các cặp nối để đo cấu trúc 3D của hệ gen
<i>HiFi read (high-fidelity read)</i>	read có độ chính xác cao	Long read PacBio có độ chính xác cao nhờ đọc lặp nhiều lần một phân tử vòng
<i>hybridization</i>	lai phân tử	Sự bắt cặp giữa hai mạch nucleic acid bổ sung
<i>imputation</i>	–	Dự đoán kiểu gen tại những biến thể không được genotype trực tiếp, dựa vào một reference panel
<i>insert size</i>	–	Chiều dài của đoạn DNA được đọc từ hai đầu trong dữ liệu paired-end
<i>isoform</i>	–	Các transcript khác nhau được tạo ra từ cùng một gen
<i>junction read</i>	–	Read vắt qua ranh giới giữa hai exon, cho thấy cách gen được cắt nối
<i>library preparation</i>	tạo thư viện	Cắt DNA và gắn adapter để chuẩn bị đưa vào máy giải trình tự
<i>low-pass WGS</i>	–	WGS ở coverage rất thấp, kết hợp imputation để lấp khoảng trống
<i>mapping quality</i>	–	Mức độ tin cậy rằng một read được đặt đúng vị trí trên reference; thấp khi read khớp nhiều vị trí (multiple hits)
<i>multiome</i>	–	Đo đồng thời nhiều tầng phân tử (ví dụ RNA và chromatin) trên cùng một tế bào hay cùng một lát cắt mô
<i>nanopore</i>	lỗ protein cỡ nanomet	Lỗ trên màng mà DNA hoặc RNA được kéo qua; thay đổi dòng điện ion cho biết trình tự
<i>no-call</i>	–	Mẫu không được gán kiểu gen vì tín hiệu nằm lưng chừng giữa các đám
<i>paired-end</i>	–	Đọc một đoạn DNA từ cả hai đầu, cho ra một cặp read
<i>phasing</i>	–	Xác định các biến thể có nằm trên cùng một nhiễm sắc thể (cùng từ cha hoặc cùng từ mẹ) hay không
<i>Phred quality score</i>	điểm chất lượng	$Q = -10 \log_{10} p$, với p là xác suất máy đọc sai một base
<i>power (statistical power)</i>	–	Khả năng phát hiện một hiệu ứng có thật
<i>probe</i>	đầu dò	Đoạn oligonucleotide gắn trên array để bắt cặp với một trình tự đích
<i>promoter</i>	–	Vùng DNA ở đầu gen, nơi bộ máy phiên mã gắn vào để bắt đầu phiên mã
<i>pangenome graph</i>	–	Reference dạng đồ thị chứa nhiều hệ gen và các biến thể của một quần thể hoặc một loài

Thuật ngữ	Tên tiếng Việt	Giải thích
<i>reference panel</i>	bảng tham chiếu	Tập hệ gen đã được giải trình tự đầy đủ, dùng làm cơ sở cho imputation
<i>resolution</i>	độ phân giải	Mức độ chi tiết của phép đo, ví dụ như bulk, single-cell hay spatial
<i>reversible terminator</i>	nhóm chặn tạm thời	Nhóm hóa học làm dừng tổng hợp sau mỗi base, được cắt bỏ trước chu kỳ tiếp theo
<i>RNA editing</i>	–	Biến đổi một base trên RNA sau phiên mã, ví dụ như adenosine thành inosine, khiến RNA khác với DNA tại vị trí đó
<i>segmental duplication</i>	vùng lặp đoạn	Đoạn DNA dài từ vài nghìn đến hàng trăm nghìn base, có các bản sao gần giống nhau trong hệ gen
<i>sequencing by synthesis</i>	giải trình tự bằng tổng hợp	Đọc trình tự bằng cách ghi nhận từng nucleotide khi polymerase tổng hợp mạch bổ sung
<i>single-molecule real-time (SMRT)</i>	giải trình tự phân tử đơn theo thời gian thực	Công nghệ của PacBio, ghi lại tín hiệu huỳnh quang khi polymerase gắn từng nucleotide
<i>single-nucleus</i>	–	Đo trên từng nhân tế bào thay vì cả tế bào, thuận tiện cho mô đông lạnh và các phép đo chromatin
<i>SNV (single nucleotide variant)</i>	–	Biến thể thay đổi một base
<i>spatial</i>	–	Đo dữ liệu omics mà vẫn giữ nguyên vị trí của tế bào trong mô
<i>splicing</i>	cắt nối	Cắt bỏ intron và nối các exon trong RNA
<i>tag SNP</i>	–	SNP được chọn để đo trên array sao cho, sau imputation, giữ lại nhiều thông tin nhất về các biến thể khác
<i>track</i>	–	Một lớp dữ liệu được vẽ theo trục tọa độ hệ gen trong genome browser
<i>UMI (unique molecular identifier)</i>	–	Đoạn trình tự ngẫu nhiên gắn vào từng phân tử ban đầu, giúp đếm số phân tử thay vì số read
<i>WGBS, RRBS, EM-seq</i>	–	Các phương pháp đo methylation bằng short read; lần lượt đo toàn bộ hệ gen, chỉ vùng giàu CpG, và dùng enzyme thay bisulfite

Đọc thêm

- Goodwin S., McPherson J.D., McCombie W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17:333–351. <https://doi.org/10.1038/nrg.2016.49>
- Logsdon G.A., Vollger M.R., Eichler E.E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics* 21:597–614. <https://doi.org/10.1038/s41576-020-0236-x>
- Marchini J., Howie B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 11:499–511. <https://doi.org/10.1038/nrg2796>
- Stark R., Grzelak M., Hadfield J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics* 20:631–656. <https://doi.org/10.1038/s41576-019-0150-2>
- Buenrostro J.D. và cộng sự (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10:1213–1218. <https://doi.org/10.1038/nmeth.2688>

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526:68–74. <https://doi.org/10.1038/nature15393>