Tạp chí Công nghệ Sinh học **19**(1): 1-14, 2021

BÀI TỔNG QUAN

PROTEOGENOMICS, CÁC ỨNG DỤNG TRONG SINH HỌC VÀ Y HỌC CHÍNH XÁC

Phan Văn Chi✉, Lê Thị Bích Thảo

Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

✉Người chịu trách nhiệm liên lạc. E-mail: pvchi@yahoo.com

Ngày nhận bài: 17.1.2020
Ngày nhận đăng: 20.4.2020

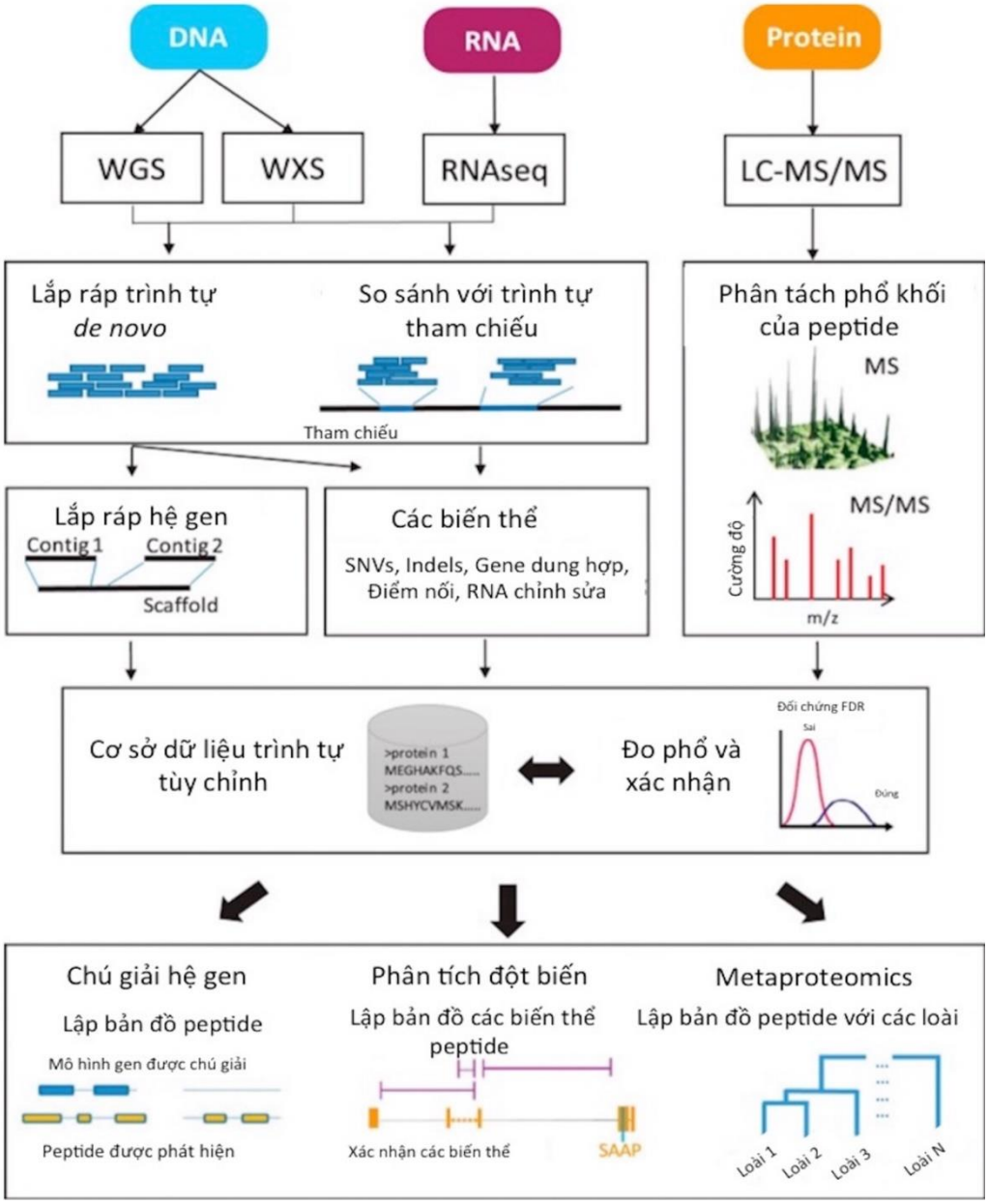
TÓM TẮT

Trong tổng quan này, chúng tôi thảo luận ngắn gọn về proteogenomics, tích hợp của proteomics với genomics và transcriptomics, theo đó các công nghệ nền tảng là giải trình tự thế hệ tiếp theo (NGS) và phép đo phổ khối (MS) với xử lý các dữ liệu thu được, một lĩnh vực mới nổi hứa hẹn thúc đẩy nhanh những nghiên cứu cơ bản liên quan đến quá trình phiên mã, dịch mã, cũng như các khả năng ứng dung. Bằng cách kết hợp các thông tin của hệ gen và hệ protein, các nhà khoa học đang đạt được những kết quả mới do sự hiểu biết đầy đủ và thống nhất hơn về các quá trình sinh học phân tử phức tạp. Một phần của tổng quan này giới thiệu một số kết quả sử dung proteogenomics trong giải quyết các vấn đề như chú giải, chú giải lại gen/hệ gen, bao gồm cả chỉnh sửa các khung đọc mở (ORF), hoặc cải thiện quá trình phát hiện gen mới ở một số cơ thể sinh vật khác nhau, kể cả con người. Đặc biệt, bài báo cũng thảo luận về tiềm năng của proteogenomics thông qua các thành tựu nghiên cứu về bộ gen/hệ protein người trong y học chính xác, đặc biệt là trong các dự án về nghiên cứu quá trình phát sinh, chẩn đoán và điều trị ung thư. Những thách thức và tương lai của proteogenomics cũng được thảo luận và ghi nhận.

Từ khóa: Proteogenomics, Genomics, Transcriptomics, Proteomics, Next-generation sequencing (NGS), Mass spectrometry (MS)

MỞ ĐẦU

Sự hợp nhất của proteomics với genomics với tên gọi proteogenomics là một một lĩnh vực mới nổi được thiết lập tốt nhất trong nghiên cứu multi-omics, theo đó các công nghệ nền tảng chính là giải trình tự thế hệ tiếp theo (NGS) và phép đo phổ khối (MS). Trong phương pháp tiếp cận phân tích proteogenomics, dữ liệu genome bao gồm trình tự DNA, ESTs (expressed sequence tags) và dữ liệu transcriptome bao gồm RNA-Seq, RIBO-Seq (ribosome profiling) được sử dụng để tạo cơ sở dữ liệu trình tự protein tùy chỉnh để giúp diễn giải dữ liệu proteomics (LC-MS/MS). Ngược lại, dữ liệu proteomics cung cấp xác nhận ở mức độ protein dữ liệu về biểu hiện gen, cũng như giúp tinh chỉnh mô hình gen. Các mô hình gen nâng cao có thể giúp cải thiện cơ sở dữ liệu trình tự protein để phân tích proteomics truyền thống (Nesvizhskii, 2014; Ruggles *et al*, 2017; Low *et al*, 2019). Với proteogenomics, các nhà sinh học đã tạo ra nghiên cứu sâu sắc mà không thể đạt được chỉ bằng genomics hoặc proteomics. Proteogenomics có thể kết hợp các kỹ thuật MS với NGS để nghiên cứu vai trò của các biến thể protein trong cơ chế sinh học và bệnh lý. Trong một thí nghiệm proteomics điển hình, phổ MS/MS của peptide thường được giải thích bằng thuật toán tìm kiếm cơ sở dữ liệu khi được cho là có sự khớp và tương đồng của từng phổ khối thí nghiệm so với phổ khối mô hình được xây dựng từ trình tự peptide có trong cơ sở dữ liệu trình tự protein do người dùng cung cấp (Aebersold, Mann, 2016). Trong nghiên cứu về protein, dữ liệu MS thường được so khớp với các peptide hiện có trong cơ sở dữ liệu protein tham chiếu.



Hình 1. Mô phỏng quy trình xác định trình tự tập trung theo công nghệ proteogenomics (Ruggles *et al*, 2017), theo đó việc giải trình tự DNA (toàn bộ trình tự bộ gen, WGS; toàn bộ trình tự exome, WXS) và RNA (RNA-seq) tạo ra hàng triệu lần đọc trình tự ngắn được tập hợp thành bộ genome, bộ exome hoặc bộ transcriptome bằng cách tiếp cận *de novo* hoặc dựa trên mẫu chuẩn được so sánh với trình tự tham chiếu. Quang sai về trình tự đặc hiệu mẫu được xác định và trình tự nucleotide được chuyển thành cơ sở dữ liệu trình tự amino acid đã cá nhân hóa. Phổ khối peptide thu được từ phân tích LC-MS/MS từ một mẫu tương ứng sau đó được ghi và xác nhận dựa trên cơ sở dữ liệu được cá nhân hóa cho phép phát hiện các trình tự peptide đặc trưng cho mẫu. Tùy thuộc vào phạm vi của dự án proteogenomics, những peptide này sau đó có thể được sử dụng để: (i) hỗ trợ chú giải bộ gen bằng cách phát hiện các peptide ở vùng genome chưa được chú giải; (ii) xác định các đột biến đặc hiệu của khối u được dịch mã trong proteome cũng như các biến thể protein mới; và (iii) phát hiện các peptide đặc trưng loài trong các cộng đồng vi sinh vật.

Một số vấn đề có thể chính phát sinh ở đây là protein trong câu hỏi có thể là mới và do đó không thể tham chiếu trong cơ sở dữ liệu và peptide có thể chứa đột biến hoặc đại diện cho một dạng mới thay thế. Bằng cách kết hợp proteomics và genomics, proteogenomics tích hợp bộ dữ liệu bộ/hệ gen (genome), bộ/hệ phiên mã (transcriptome) và hệ protein (proteome) để khắc phục những vấn đề nêu trên. Proteogenomics cho phép phân tích mối tương quan giữa: (i) các cặp mRNA và protein; (ii) các đột biến và biến đổi sau dịch mã và đường dẫn tín hiệu; (iii) các tác động điều hòa đối với mức độ biểu hiện RNA và protein do biến thể di truyền (eQTL), microRNAs (miRNAs). Như vậy, các kỹ thuật phổ biến sử dụng trong nghiên cứu proteogenomics sẽ phải bao gồm không chỉ giải trình tự DNA, RNA, phân tích dữ liệu, mà còn phải là các phép đo khối phổ LC-MS/MS và MALDI (Datta *et al*, 2016a; Ruggles *et al*, 2017). Proteogenomics cho thấy một quan hệ đối tác bình đẳng về đóng góp và lợi ích của mỗi thành phần. NGS cho phép các nhà nghiên cứu mô tả các biến thể trong bộ gen, chẳng hạn như đa hình nucleotide đơn (SNPs) và dịch mã. Sử dụng các phương pháp *in silico*, các biến thể này sau đó có thể được dịch thành các proteoforms thêm vào cơ sở dữ liệu protein hiện có và được sử dụng để giải thích dữ liệu MS, làm cho cơ sở dữ liệu toàn diện hơn. Hình 1 dưới đây là mô hình nguyên lý cơ bản của Ruggles và đồng tác giả về quy trình công nghệ proteogenomics dựa trên xác định trình tự tập trung (sequence-centric proteogenomics) (Ruggles *et al*, 2017). Quy trình bao gồm phân tích tích hợp cơ sở dữ liệu genome và proteome để chú giải exome dưới dạng khám phá gen và sàng lọc mô hình gen (PAGA, proteomics aiding genome annotation); phát hiện ở mức độ protein các biến thể amino acid đơn lẻ (SAAV), chèn, xóa, nối ghép thay thế và sự dung hợp gen mới liên quan đến trình tự bộ gen tham chiếu); ứng dụng proteomics để nghiên cứu đặc tính của kháng thể; nghiên cứu ảnh hưởng của nhiễm virut và transposons đối với biểu hiện gen ở sinh vật nhân chuẩn; và các ứng dụng của proteogenomics để điều tra/phát hiện metaproteomics (metaproteogenomics).

PROTEOGENOMICS VÀ CÁC VẤN ĐỀ TRONG CHÚ GIẢI/CHÚ GIẢI LẠI GEN/HỆ GEN

Theo truyền thống, chú giải bộ/hệ gen được thực hiện bằng thuật toán dự đoán gen với hướng dẫn từ cơ sở dữ liệu của bộ/hệ phiên mã. Rất thường xuyên, các chú giải không chính xác dẫn đến các vùng mã hóa bị bỏ lỡ hoặc bị chú giải sai điểm bắt đầu phiên mã (TSSs, Transcriptions Start Sites) và điểm đầu dịch mã (TISs, Translational Initiation Sites). Ngoài việc đa hình di truyền có thể làm thay đổi trình tự amino acid, người ta cũng đã chứng minh rằng các RNA không mã hóa, khung đọc mở (ORF, Open Reading Frame) ngắn và pseudogene trên thực tế có thể mã hóa các protein mới và các trình tự này thường không thấy ở đường dẫn chú giải gen/hệ gen (Datta *et al*, 2016a). Việc chú giải các TIS có thể phức tạp bởi có nhiều codon khởi đầu AUG trong một bản phiên mã và vấn đề này có thể được giải quyết bằng cách kết hợp định dạng ribosome (RIBO-Seq) và xác định trình tự đầu N bằng proteomics. Giess và đồng tác giả đã xây dựng một mô hình để dự đoán các TIS, sau đó được xác nhận bằng proteomics đầu N. Theo đó, họ cũng đã chú giải lại một số TIS và xác nhận cắt ngắn đầu cuối N và mở rộng các trình tự mã hóa được chú giải trước đó ở các prokaryote đã chọn (Giess *et al*, 2017). Tương tự, proteogenomics có thể được áp dụng để phát hiện sự kết thúc bất ngờ của quá trình dịch mã. Armengaud và đồng tác giả đã ~~được~~ chứng minh ở loài *Blastocystis* một cơ chế dừng/chấm dứt mới, theo đó các đoạn giàu GU nằm ở phía sau các vị trí polyadenylation của mRNA tạo ra nhiều các codon kết thúc bổ sung trước codon kết thúc thực tế, dẫn đến protein bị cắt ngắn (Armengaud *et al*, 2017). Những ví dụ về ứng dụng proteogenomics trong chú giải lại gen/hệ gen đã được mô tả khá rõ đối với các sinh vật không phải là mô hình, đặc biệt là các vi khuẩn, mà thiếu bộ gen tham chiếu. Bằng cách hợp nhất các trình tự mã hóa đã được chú giải (CDSs, annotated coding sequences), dự đoán gen *ab initio* và *in silico* các khung đọc mở vào một cơ sở dữ liệu tích hợp duy nhất (iPtgxDB, intergrated proteogenomics database,

https://iptgxdb.expasy.org/database/), và sử dụng nó để tìm kiếm cơ sở dữ liệu proteomics, Omasits và đồng tác giả đã tìm thấy các sai sót trong chú giải pseudogen, ORF và TIS ở *Bartonellahenselae*, *Bradyrhizobium diazoefficiens* và *Escherichia coli* (Omasits *et al*, 2017).

Genome của tằm đã được giải trình tự và lắp ráp khá chắc chắn, nhưng chú giải chính xác về bộ gen đối với các vấn đề sinh học hiện đại vẫn chưa hoàn chỉnh. Để cải thiện phần chú giải nay, nhóm Ye X và đồng tác giả đã thực hiện phân tích proteogenomics, sử dụng 9,8 triệu phổ khối thu thập từ các mô khác nhau ở các giai đoạn phát triển của tằm (Ye *et al*, 2019). Kết quả đã xác nhận các sản phẩm dịch mã của 4.307 mô hình gen hiện có và xác định được 1.701 peptit mới đặc hiệu cho tìm kiếm bộ gen (GSSP, genome search-specific peptides). Sử dụng các GSSP này, 74 trình tự mã hóa gen mới đã được xác định và 121 mô hình gen hiện có đã được chỉnh sửa. Nhóm tác giả cũng đã xác định được 1.182 peptit tiếp giáp mới dựa trên cơ sở dữ liệu bỏ qua exon dẫn đến việc xác định 973 vị trí nối thay thế. Hơn nữa, họ thực hiện phân tích RNA-seq để cải thiện chú giải bộ gen của tằm ở cấp độ phiên mã. Tổng cộng có 1704 phiên mã mới và 1136 exon mới đã được xác định, 2581 vùng chưa được dịch mã (UTR, untranslated regions) đã được chỉnh sửa và 1301 gen nối thay thế (AS, alternative splicing) đã được xác định. Kết quả transcriptomics được tích hợp với dữ liệu proteomics để bổ sung và xác minh thêm các chú giải mới. Ngoài ra, 14 gen không chính xác và 10 exon bị bỏ qua đã được xác minh bằng hai phương pháp phân tích. Như vậy, họ đã xác định được 1838 bản sao mã mới và 1593 gen AS, chỉnh sửa 5074 gen hiện có bằng cách sử dụng phân tích proteogenomics and transcriptomics. Các dữ liệu có thể tra cứu qua ProteomeXchange (http://www.proteomexchange.org/) với số nhận dạng PXD009672 .

Dựa trên các dữ liệu về proteogenomics, Mao và đồng tác giả chú giải lại bộ gen của chủng *Yersinia pestis* 91001, loại bỏ 137 CDS không đáng tin cậy, tái định vị TIS cho 41 gen tương đồng và sửa đổi chức năng của 7 pseudogenes và 392 gen giả định (Mao *et al*, 2016). Một nghiên cứu proteogenomics của Rang và đồng tác giả cho thấy 39 trình tự mã hóa trong toàn bộ bộ gen của *Bacillus thuringiensis* có liên quan đến khả năng gây bệnh của côn trùng, bao gồm 5 gen *cry*. Tuy nhiên, các protein kháng sâu như Cry2Ab, Cry1Ia, Cytotoxin K, Bacteriocin, Exoenzyme C3 và Alveolysin đã không thể được xác định trong dữ liệu proteomics thu được (Rang *et al*, 2015). Ở nấm, các kỹ thuật proteogenomics cũng đã hỗ trợ chú giải lại bộ gen cho (i) *Coccidiodes posadasii* gây ra bệnh cầu trùng (Valley fever) (Mitchell *et al*, 2018); (ii) *Malassezia sympodialis*, một loại nấm men da (Zhu *et al*, 2017); (iii) *Candida tropicalis*, một mầm bệnh cơ hội gây ra bệnh nấm candida ở những người bị suy giảm miễn dịch (Datta *et al*, 2016b) và (iv) *Parastagonospora nodorum*, một loại mầm bệnh của lúa mì gây ra bệnh đốm đỏ Septoria (SNB) (Syme *et al*, 2016). Một giống nho *Cabernet Sauvignon* cũng đã được giải trình tự theo cách tiếp cận proteogenomics và dữ liệu RNA-seq đã cho thấy có tới 341 chú giải mới (Chapman, Bellgard, 2017).

Việc nhân dòng phân tử ở lúa (*Oryza sativa*) đã thu hút được sự chú ý đáng kể trong những năm gần đây, nhưng việc chú giải bộ gen không chính xác đã cản trở tiến trình nay, cũng như và các nghiên cứu chức năng của bộ gen lúa. Rất gần đây, Chen EX và cs khi áp dụng phương pháp giải mã trình tự RNA đơn phân tử đọc dài (lrRNA_seq) dựa trên proteogenomics để tiết lộ sự phức tạp của bộ phiên mã ở lúa và khả năng mã hóa của nó (Chen *et al*, 2020). Đáng ngạc nhiên là khoảng 60% các locus được xác định bởi lrRNA_seq có liên quan đến các bản sao antisense tự nhiên (NAT, natural antisense transcripts). Sự sắp xếp bộ gen mật độ cao của các gen NAT cho thấy vai trò tiềm năng của chúng trong việc kiểm soát nhiều mặt sự biểu hiện của gen. Ngoài ra, một số lượng lớn các bản sao dung hợp và giữa các gen đã được quan sát thấy. Có đến 906.456 đồng dạng phiên mã (transcript isoforms) đã được xác định, và 72,9% gen có thể tạo ra đồng dạng nối ghép. Có tổng cộng 706.075 biến đổi sau phiên mã sau đó đã

được phân loại thành 10 loại phụ, chứng tỏ sự phụ thuộc lẫn nhau của các cơ chế sau phiên mã góp phần vào sự đa dạng của nhóm phiên mã. Việc giải trình tự RNA đọc ngắn song song chỉ ra rằng lrRNA_seq có khả năng vượt trội hơn trong việc xác định các bản sao dài hơn. Ngoài ra, việc xác định được 190.000 peptide duy nhất thuộc 9.706 proteoforms/nhóm protein cho thấy sự đa dạng hơn của hệ protein lúa. Phát hiện của nhóm tác giả chỉ ra rằng tổ chức bộ gen, sự đa dạng của bộ phiên mã và tiềm năng mã hóa của bộ gen phiên mã ở lúa phức tạp hơn nhiều so với dự đoán trước đây.

Proteogenomics cũng cho phép phát hiện ra các peptide mới (từ các locus mã hóa protein không được chú giải) và các peptide có các biến thể amino acid đơn (có nguồn gốc từ các đột biến và đa hình nucleotit đơn). Ví dụ, do các trình tự peptide của ong mật (*Apis mellifera*) có tỷ lệ nhận dạng protein thấp, McAfee và đồng tác giả đã thực hiện phân tích proteogenomic với ~ 1500 tệp MS thô và tìm thấy đến hơn 2000 vùng mã hóa/exon mới bị bỏ lỡ, cũng như các chú giải bị bỏ lỡ trước đó (McAfee *et al*, 2017). Bằng một chiến lược phân tích và chú giải proteogenomics mới, peptide conorfamide-Vc1 (CNF-Vc1), một họ gen mới cũng đã được xác định từ nọc độc của loài ốc biển săn mồi *Conus victoriae* (Robinson *et al*, 2015). Mahadevan và đồng tác giả khi thử gây bệnh thối rễ hạt tiêu đen (*Piper nigrum* L.) với *Phytophthora capsici* và kết hợp phân tích transcriptomics và proteomics, đã phát hiện cái nhìn mới lạ về quan hệ tương tác này. Họ đã nhận dạng được tổng số 532 protein lá mới từ hạt tiêu đen, trong đó 518 protein được chú giải về mặt chức năng bằng công cụ BLAST2GO, trong đó có 22 protein điều hòa tăng và 134 protein điều hòa giảm (Mahadevan *et al*, 2016).

Một vấn đề nổi tiếng khác trong proteomics là vấn đề "protein bị thiếu", khi xem xét chú giải gen liên quan đến các protein chưa được phát hiện, đặc biệt cả trong Dự án HPP (Human Proteome Project, https://www.hupo.org/human-proteome-project) nói chung và C-HPP (Chromosome-Centric Human Proteome Project, https://www.hupo.org/C-HPP) nói riêng (Omenn *et al*, 2017, González-Gomariz *et al*, 2019). Nhóm nghiên cứu Chromosome-11 trong dự án C-HPP đã ghép nối cơ sở dữ liệu NeXtProt (https://www.nextprot.org/) và GENCODE (https://www.gencodegenes.org/) để phân tích các bộ dữ liệu MS từ các mô não, vỏ não, tủy sống, não thai nhi, tinh hoàn và tinh trùng. Họ đã xác định các protein bị thiếu và các biến thể ghép nối thay thế (ASV, alternative spliced variants) của GENCODE ở các phần chèn exon mới, các bản dịch mã thay thế ở vùng 5' chưa được dịch mã hoặc trình tự mã hóa protein mới (Hwang *et al*, 2017). Dự án HPP hàng năm báo cáo về những tiến bộ đạt được trong việc xác định và mô tả một cách đáng tin cậy danh sách các bộ phận protein hoàn chỉnh của cơ thể người và biến proteomics trở thành một phần không thể thiếu của các nghiên cứu multiomics trong khoa học sự sống và y học. Bản phát hành NeXtProt 2019-01-11 chứa 17.694 protein với bằng chứng cấp độ protein mạnh mẽ (PE1), tuân thủ hướng dẫn của HPP về diễn giải dữ liệu MS v2.1; những gen này đại diện cho 89% trong tổng số 19.823 gen mã hóa dự đoán neXtProt (tất cả các protein PE1,2,3,4), tăng từ 17.470 một năm trước đó. Ngược lại, số lượng protein neXtProt PE2,3,4, được gọi là "protein bị thiếu" (MPs), đã giảm từ 2.949 xuống 2.129 kể từ năm 2016 thông qua những nỗ lực trong toàn cộng đồng, bao gồm cả C-HPP. PeptideAtlas (http://www.peptideatlas.org/) là nguồn dữ liệu khối phổ thô được phân tích lại đồng nhất cho NeXtProt; PeptideAtlas đã bổ xung thêm 495 protein chuẩn từ năm 2018 đến năm 2019, đặc biệt là từ các nghiên cứu được thiết kế để phát hiện các protein khó nhận dạng. Trong khi đó, Bản đồ Protein Người (The Human Protein Atlas, https://www.proteinatlas.org/) đã phát hành phiên bản 19.3 với bằng chứng hóa mô miễn dịch về sự biểu hiện của 17.058 protein trên cơ sở phân tích với 26.371 kháng thể. Nhiều nhà nghiên cứu áp dụng các proteomics theo chiến lược SRM (selected reaction monitoring) để định lượng các protein phổ biến đặc hiệu cho các bào quan trong các nghiên cứu về các bệnh khác nhau ở người. 19 nhóm nghiên cứu theo hướng Sinh học và Dịch bệnh của dự án (B/D-

HPP, Biology and Disease-driven B/D-HPP) đã xuất bản tổng cộng 160 ấn phẩm vào năm 2018, đưa proteomics và proteogenomics thành cách tiếp cận/kỹ thuật không thể thiếu trong nghiên cứu của y-sinh học (Omenn *et al*, 2019).

Một vấn đề có thể gọi là "hóc búa" trong proteomics là phát hiện biến thể amino acid đơn (SAAV) trong các peptide đa hình, một sản phẩm của các biến thể gen không đồng nghĩa. Nhiều peptide biến thể tương ứng với các biến thể nucleotide đơn (SNV, single nucleotide variants) có liên quan đến các bệnh cụ thể. Trình tự các biến thể được mã hóa trong bộ gen như vậy, tất nhiên, có thể nhận được từ giải trình tự exome (Lobas *et al*, 2016), RNA-Seq (Cesnik *et al*, 2016) hoặc từ cơ sở dữ liệu SNP/SNV hiện có. Ngoài ra, các trình tự amino acid cũng có thể được thay đổi bằng cách chỉnh sửa RNA, theo đó, adenosine deaminase (ADAR) đặc hiệu với RNA chuyển đổi adenosine thành inosine. Trong quá trình dịch mã, inosine có thể được nhận diện như là guanine, dẫn đến việc thay thế amino acid. Bằng cách tìm kiếm các bộ dữ liệu proteome sâu dựa trên cơ sở dữ liệu trình tự protein tùy chỉnh được tạo ra từ các nghiên cứu RNA adenosine-to-inosine trên toàn bộ gen ở *D. melanogaster*, Kuznetsova và đồng tác giả đã xác định 56 protein được chỉnh sửa, theo đó 7 protein được chia sẻ giữa các hệ protein tổng số, đầu và não của côn trùng (Kuznetsova *et al*, 2018). Trong khi đó, Wingo và đồng tác giả khi xác định các protein đặc hiệu alen và định lượng chúng trong hai mẫu não sau khi chết của người và phát hiện ra hơn 400 cặp peptide tham chiếu và SAV (Wingo *et al*, 2017). Dimitrakopoulos và đồng tác giả cũng xác định được các đột biến p53 ở mức protein trong các mẫu khối u ung thư vú mà trước đây đã được giải trình tự sử dụng phương pháp phân tích SRM (selected reaction monitoring) (Dimitrakopoulos *et al*, 2017). Khi dịch mã *in silico* các trình tự phiên mã được nối ghép xen kẽ (AST, alternatively spliced transcript) thu được thông qua RNA-Seq, sẽ có thể nhận được cả các protein dạng nối ghép (spliceforms) theo FASTA. Ví dụ minh họa là kết quả thí nghiệm của Nassa và đồng tác giả khi kích hoạt các tiểu cầu với thụ thể collagen và thrombin hoạt hóa peptide và phân tích tổng hợp protein *de novo* trong tiểu cầu bằng các kỹ thuật proteogenomics. Các tác giả đã xác nhận một quần thể các RNA chứa intron cư trú trong các tiểu cầu đang ở trạng thái nghỉ ngơi và sau đó chúng được tách ra để tạo ra protein trưởng thành khi kích hoạt (Nassa *et al*, 2018).

PROTEOGENOMICS VÀ Y HỌC CHÍNH XÁC

Từ quan điểm về triển vọng của "omics", khá nhiều các lĩnh vực của y học chính xác, bao gồm cả ung thư lâm sàng đã bị chi phối trong trong thời gian vừa qua chỉ bởi những nghiên cứu về genomics. Tuy nhiên, xem cách mà Văn phòng Nghiên cứu Proteomics Ung thư lâm sàng của NCI (NCI's Office of Cancer Clinical Proteomics Research (OCCPR)) đã thay đổi động lực của lĩnh vực này thông qua Hiệp hội Phân tích Khối u Proteomic lâm sàng (CPTAC, Clinical Proteomic Tumor Analysis Consortium, https://proteomics.cancer.gov/programs/ cptac) và Hiệp hội Proteogenome Ung thư Quốc tế (ICPC, International Cancer Proteogenome Consortium, https://proteomics.cancer.gov/programs/international-cancer-proteogenome-consortium), cho thấy vai trò quan trọng của công nghệ proteomics và proteogenomics như thế nào trong y học chính xác, và đặc biệt đối với ung thư. Thêm nữa, CPTAC trong hợp tác với DREAM Challenges (http://dreamchallenges.org/) đã công bố NCI-CPTAC DREAM Proteogenomics Challenge với mục đích là việc tạo ra các phương pháp tính toán để trích xuất thông tin từ proteome ung thư, đặc biệt là phosphoproteome và để liên kết những dữ liệu đó với các thông tin của genome và transcriptome. Kết quả của phương pháp sẽ được đánh giá bằng cách sử dụng tập dữ liệu xác thực chưa từng thấy trước đây do CPTAC tạo ra. Dựa trên mô hình thành công cao của CPTAC, ICPC khuyến khích hợp tác quốc tế và đầu tư vào nghiên cứu proreogenomics ung thư. Thông qua những nỗ lực của CPTAC, ngày càng rõ ràng rằng để hiểu rõ genome, người ta cũng cần có một sự hiểu biết vững chắc về proteome, bao gồm cả các

sửa đổi sau dịch mã (PTMs). Rõ ràng, cùng với các kết quả nghiên cứu được phân tích ở phần trên đã cho thấy, việc tích hợp dữ liệu genome và proteome thông qua các phương pháp/cách tiếp cận proteogenomics có thể làm sáng tỏ các cơ sở sinh học mà khó có thể có được hoặc không thể thông qua chỉ bằng genomics. Những dự án như thế này, cùng với Dự án HPP nói chung và C-HPP nói riêng đều sẽ là những bước đột phá quan trọng trong hiểu biết về các cơ chế phân tử của ung thư, thúc đẩy phát triển khoa học và ứng dụng công nghệ proteogenomics vào y học chính xác trong tương lai.

Bản đồ bộ gen ung thư (TCGA, The Cancer Genome Atlas, https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas) đã mô tả đặc điểm bộ gen của nhiều loại ung thư ở người, bao gồm CRC và CPTAC cũng đã thực hiện các phân tích protein tích hợp CRC (Zhang *et al*, 2014). Tuy nhiên, cơ sở di truyền chính của CLM (colorectal cancer liver metastasis) vẫn chưa được làm sáng tỏ đầy đủ. Phân tích đặc điểm proteogenomics, rồi tích hợp và so sánh bộ gen có thể cung cấp các thông tin liên quan đến chức năng để chú giải các bất thường về bộ gen với giá trị tiên lượng. Ma Y và đồng tác giả đã tiến hành phân tích các proteome, giải trình tự toàn bộ exome và transcriptome và xác định đa hình nucleotide đơn cho 2 bộ mẫu bao gồm mô đại trực tràng bình thường, mô CRC nguyên phát và mô di căn gan CLM khớp đồng bộ (Ma *et al*, 2018). Họ đã xác định được 112 phân tử tương quan CNV-mRNA-protein, bao gồm COL1A2 và BGN được điều chỉnh tăng liên quan đến tiên lượng và bốn điểm nóng mạnh nhất (nhiễm sắc thể X, 7, 16 và 1) thúc đẩy sự biến đổi phong phú của mRNA trong di căn gan CRC. Hai vị trí ($DMRTB1^{R202H}$ và $PARP4^{V458I}$) được phát hiện là những đột biến thường xuyên chỉ ở nhóm di căn gan và hiển thị lượng protein bị rối loạn điều hòa. Hơn nữa, nhóm nghiên cứu cũng xác nhận rằng số lượng peptide bị đột biến có giá trị tiên lượng tiềm năng và các biến thể soma cho thấy lượng protein tăng lên, bao gồm mức biểu hiện cao của MYH9 và CCT6A, là có ý nghĩa lâm sàng. Phân tích tổng hợp toàn diện về 44 khối u đại trực tràng và các cặp bình thường cung cấp một số hiểu biết sâu sắc về sinh học của CLM và xác định các mục tiêu điều trị tiềm năng. Hơn nữa, kết quả nghiên cứu/phân tích đặc tính của CRC di căn bằng cách tiếp cận proteogenomics cho thấy rõ sức mạnh của việc tích hợp genomics và proteomics. Cách tiếp cận này cung cấp cái nhìn mới về vai trò của những thay đổi protein này trong CLM, có thể được mở rộng để hiểu vai trò của đột biến protein trong các bệnh ung thư khác. Đặc biệt, việc khám phá về dấu ấn (marker) sinh học mới cho phép thiết lập các xét nghiệm không xâm lấn có thể làm tăng khả năng tuân thủ điều trị của bệnh nhân. Chúng cũng cho phép chẩn đoán sớm hiệu quả về chi phí, cũng như các phương pháp điều trị phù hợp với bệnh nhân, cải thiện sự sống sót. Các dấu ấn sinh học CRC cũng có thể có giá trị tiên lượng và thông thường, chúng được đưa vào các chương trình theo dõi. Tuy nhiên, bất chấp sự tiến bộ liên tục của các công nghệ mới, việc xác nhận lâm sàng của chúng vẫn còn đang tranh cãi. Trong bối cảnh như vậy, các nghiên cứu lâm sàng bổ sung vẫn cần thiết để xác định các dấu ấn hiệu quả nhất trong số các dấu ấn tiềm năng (Binetti *et al*, 2020).

Sự thay đổi về mã hóa protein được thể hiện ở mức protein vật lý, những thông tin không thể được suy luận chỉ từ một loại dữ liệu về genome hoặc MS. Vasaikar và đồng tác giả đã thực hiện nghiên cứu phân tích proteogenomics đầu tiên trên một nhóm ung thư ruột kết được thu thập tiền cứu (Vasaikar et al, 2019). Kết quả phân tích proteomics và phosphoproteomics khối u và các mô lân cận bình thường đã tạo được một danh mục các protein và các điểm phosphoryl hóa liên quan đến ung thư ruột kết, bao gồm các dấu ấn sinh học mới đã biết và giả định, điểm tác động/mục tiêu của thuốc và kháng nguyên ung thư. Tính tích hợp trong proteogenomics không chỉ ưu tiên các mục tiêu được suy luận theo hệ gen, chẳng hạn như số bản sao trình điều khiển và các kháng nguyên mới có nguồn gốc đột biến (mutation-derived neoantigens), mà còn mang lại những phát hiện mới. Dữ liệu về phosphoproteomics liên quan đến sự phosphoryl hóa Rb với sự tăng sinh và giảm quá trình

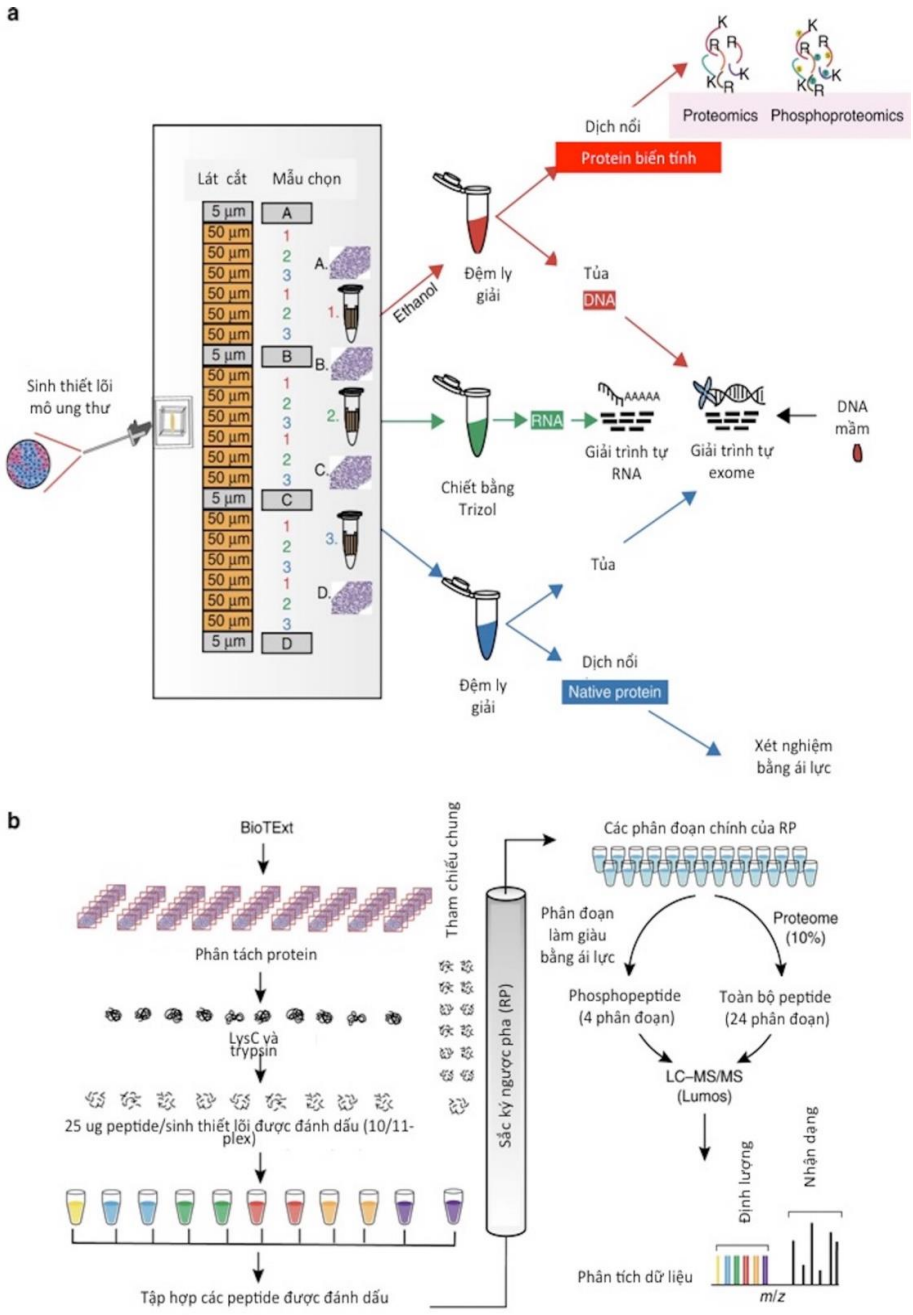
apoptosis trong ung thư ruột kết. Điều này giải thích tại sao chất ức chế khối u cổ điển này được khuếch đại trong các khối u ruột kết và gợi ý lý do để nhắm mục tiêu phosphoryl hóa Rb trong ung thư ruột kết. Proteomics đã xác định được mối liên quan giữa việc giảm thâm nhập tế bào T CD8 và tăng đường phân trong các khối u có độ bất ổn định cao (MSI-H, microsatellite instability-high) của tế bào vi mô, cho thấy quá trình đường phân là một mục tiêu tiềm năng để vượt qua sức đề kháng của khối u MSI-H đối với sự phong tỏa điểm kiểm tra miễn dịch. Proteogenomics đưa ra những con đường mới cho những khám phá sinh học và phát triển liệu pháp điều trị.

Để làm sáng tỏ các mô-đun chức năng đã bị thay đổi dẫn đến ung thư biểu mô tế bào thận (tế bào sáng) (ccRCC, clear cell Renal Cell Carcinoma), Nhóm của Clark và đồng tác giả đã nghiên cứu đặc tính toàn diện về genome, epigenome, transcriptome, proteome và phosphoproteome của ccRCC được điều trị và so sánh các mẫu mô lân cận bình thường (Clark *et al*, 2020). Các phân tích đã xác định được một phân nhóm phân tử riêng biệt có liên quan đến sự bất ổn định của bộ gen. Tích hợp các phép đo proteogenomics đã xác định sự rối loạn điều hòa protein duy nhất của các cơ chế tế bào bị ảnh hưởng bởi sự thay đổi bộ gen, bao gồm chuyển hóa liên quan đến các quá trình phosphoryl hóa-oxy hóa, dịch mã protein và mô-đun tín hiệu. Để đánh giá mức độ xâm nhập miễn dịch ở từng khối u, họ đã xác định các dấu hiệu trong môi trường tế bào mô tả bốn phân nhóm ccRCC dựa trên miễn dịch được đặc trưng bởi các con đường chuyển hóa riêng biệt. Nghiên cứu này cho thấy một phân tích proteogenomics quy mô lớn về ccRCC để phân biệt tác động chức năng của sự thay đổi bộ gen và cung cấp bằng chứng cho việc lựa chọn phương pháp điều trị hợp lý bắt nguồn từ bệnh học của ccRCC.

Sự thật là bệnh nhân và bác sĩ trong lĩnh vực ung thư phải đối mặt với một vấn đề ngày càng tăng, đó là khả năng chống lại các phương pháp điều trị ung thư. 90% thất bại của hóa trị liệu trong quá trình xâm lấn và di căn của bệnh ung thư liên quan đến vấn đề kháng thuốc. Proteogenomics có thể giúp hiểu được sự đề kháng này, bằng cách khám phá tầm quan trọng của một số biến thể gen và protein nhất định trong việc quyết định thành công của quá trình điều trị. Ví dụ, trong CRC, bệnh nhân thường được điều trị bằng kháng thể đơn dòng cetuximab và panitumumab (anti-EGFR drugs). Woo và đồng tác giả mô tả phân tích proteogenomics tích hợp mở rộng giới hạn tìm kiếm phân tích protein sử dụng dữ liệu trình tự RNA tùy chỉnh từ cơ sở dữ liệu của ICGC (International Cancer Genome Consortium, https://icgc.org/) và TCGA để nghiên cứu vai trò của các peptide biến thể bên cạnh các biến thể gen immunoglobulin trong liệu pháp anti-GFR (Woo *et al*, 2015). Các tác giả đã phát hiện sự hiện diện của gen KRAS kiểu hoang dã là bắt buộc để thuốc anti-EGFR có hiệu quả đối với dạng ung thư này. Do vậy, sự thay đổi trong gen này có thể dẫn đến phản ứng trị liệu kém. Đây cũng là minh chứng đầu tiên về đặc tính mở rộng đáp ứng miễn dịch khối u và chứng minh tiềm năng của proteogenomics trong cải thiện đặc tính phân tử của các phân nhóm khối u. Trong một nghiên cứu khác, khi phân tích proteogenomics dựa trên MS để khám phá các đột biến của gene "gác cổng" trong ung thư phổi các tác giả cho thấy hiệu quả của thuốc ức chế tyrosine kinase (một loại thuốc chống ung thư) có thể khác nhau giữa các nhóm chủng tộc. Họ cũng cho rằng giá trị của các phương pháp proteogenomics dựa trên MS là ở chỗ nó cho phép phân tích trực tiếp các protein bị đột biến trong một mẫu lâm sàng, cung cấp cho các nhà khoa học khả năng phát hiện và phát triển thuốc heo phân tầng bệnh nhân (Nishimura, Nakamura, 2016).

Một vấn đề đang rất được lưu tâm là yêu cầu về lượng mô đối với các phân tích proteogenomics, một điểm hạn chế các cơ hội nghiên cứu về các quá trình dịch mã, cũng như khả năng ứng dụng để chẩn đoán ung thư. Nếu so với yêu cầu của CPTAC là phải có ít nhất 100 mg mô khối u, đủ để cung cấp thông tin định lượng với >10.000 protein và >30.000 điểm phosphoryl hóa trên mỗi mẫu (Mertins *et al*, 2018). Còn đối với chẩn đoán lâm sàng,

một sinh thiết kim lõi giàu khối u đông lạnh (<20 mg) cũng phải cung cấp đủ DNA, RNA và protein để định dạng proteogenomics quy mô sâu.



Hình 2. Quy trình của công nghệ MiProt (Microscaled Proteogenomics) dựa trên phân tách sinh thiết lõi sử dụng Trifecta EXTraction (Satpathy *et al*, 2020). **a** Trong quy trình của Trifecta EXTraction (BioTEXT), sinh thiết lõi từ bệnh nhân được cắt lát, tiếp theo là chiết xuất DNA, RNA, protein và các bước phân tích đặc tính bằng proteogenomics ở quy mô sâu và hình ảnh dựa trên hóa mô miễn dịch. **b.** Công nghệ MiProt cho phép xác định đặc tính hệ protein và hệ phosphoprotein ở quy mô sâu chỉ với 25 µg peptide trên mỗi sinh thiết lõi. MiProt sử dụng một tham chiếu chung để so sánh trên tất cả các mẫu trong một plex TMT10/11 và trên một số plex TMT10/11 trải qua một số sinh thiết lõi.

Để giảm bớt những yêu cầu về mô này, nhóm của Satpathy và đồng tác giả đã phát triển các phương pháp tạo ra DNA, RNA và protein chất lượng cao để giải trình tự DNA và RNA quy mô sâu, đồng thời phân tích proteome và phosphoproteome từ một sinh thiết kim lõi 14 G (Sinh thiết Trifecta Extraction, (BioTExt)) và một quy trình phân tích proteome dựa trên microLC-MS/MS và phosphoproteome dựa trên MiProt (Microscaled Proteogenomics) với yêu cầu chỉ là 25 µg peptide cho mỗi mẫu (Satpathy *et al*, 2020). Hình 2 dưới đây mô tả khá rõ quy trình quy trình về công nghệ MiProt dựa trên phân tách sinh thiết lõi từ bệnh nhân được cắt lát, tiếp theo là chiết xuất DNA, RNA, protein và các bước phân tích đặc tính bằng proteogenomics ở quy mô sâu chỉ với 25 µg peptide. Nhóm tác giả đã phân tích sinh thiết kim lõi từ ung thư vú dương tính với ERBB2 trước và 48-72 giờ sau khi bắt đầu hóa trị liệu có trastuzumab bổ trợ và cho thấy sự ức chế mạnh hơn với ERBB2 và cả mức độ phosphoryl hóa của ERBB2 và mTOR trong các trường hợp liên quan đến đáp ứng bệnh lý. Các tác giả cũng cho rằng, nguyên nhân tiềm tàng của kháng thuốc bao gồm một số yếu tố sau: (i) không có khuếch đại ERBB2; (ii) hoạt động ERBB2 không đủ cho độ nhạy điều trị mặc dù có khuếch đại ERBB2; (iii) và các cơ chế kháng thuốc bao gồm truyền tín hiệu thụ thể androgen, biểu hiện quá mức của mucin và vi môi trường miễn dịch không hoạt động. Rõ ràng, các kết nghiên cứu được trình bay ở trên có thể được coi là những minh chứng đảm bảo tiện ích lâm sàng và tiềm năng khám phá của proteogenomics ở quy mô sinh thiết cho những nghiên cứu tiếp theo.

THÁCH THỨC VÀ TƯƠNG LAI

Các lĩnh vực nghiên cứu mới luôn gặp phải những thách thức trong việc thành lập và hoàn thiện các kỹ thuật được sử dụng, và proteogenomics cũng không phải là ngoại lệ. Thực tế, proteogenomics là một lĩnh vực sử tích hợp các dữ liệu genome, transcriptome và proteome để rút ra mối tương quan giữa gen và protein. Tuy nhiên, việc kết hợp ba chuyên ngành mà mỗi chuyên ngành tạo ra các tập dữ liệu lớn đáng kể, đưa ra những thách thức đáng kể liên quan đến phân tích. Trong công trình về khái niệm, ứng dụng của proteogenomics, Nesvizhskii A (Nesvizhskii, 2014) đã chỉ ra nguồn sai sót có thể có trong phân tích, bao gồm việc ứng dụng các ngưỡng lọc giống nhau cho cả peptide đã biết và mới, xác định không chính xác các peptide mới tương đồng với các trình tự đã biết và đưa ra kết luận không được hỗ trợ dựa trên các peptide đã chia sẻ. Tác giả cũng đã khuyến khích tập trung vào việc thiết lập các hướng dẫn phân tích dữ liệu kỹ lưỡng để khắc phục các vấn đề đã nêu. Hơn nữa, do lượng dữ liệu khổng lồ được tạo ra trong các thí nghiệm dựa trên phép đo phổ khối, những cải tiến trong thuật toán tin sinh học là một chiến lược thiết yếu cho tương lai của proteomics nói chung và proteogenomic lâm sàng nói riêng. Đặc biệt, nhiều phương pháp tiếp cận proteomics để nghiên cứu các mẫu ung thư đã được đề xuất, nhưng vẫn còn những thách thức nghiêm trọng về phương pháp luận, đặc biệt là trong việc xác định các biến thể đột biến hoặc các biến thể cấu trúc như các trường hợp gen dung hợp. Trong khi các công nghệ giải trình tự thông lượng cao đã khá phổ biến trong việc tạo ra dữ liệu gen và phiên mã, thì proteomics vẫn bị tụt hậu về cả phạm vi và chi phí do những hạn chế về công nghệ. Các công nghệ dựa trên khối phổ (MS) cũng đã trở nên phổ biến trong nghiên cứu protein/hệ protein, mặc dù vẫn gặp phải những hạn chế về khả năng lặp lại trong nhận dạng và tính nhất quán của vấn đề định lượng (Schubert *et al*, 2017). Hạn chế đối với việc sử dụng proteogenomics để dự đoán ở mức độ protein là sự phức tạp của các proteome nói chung, và đặc biệt là của người nói riêng. Một đánh giá của Kendrick và đồng tác giả cho thấy có mối tương quan khá hạn chế giữa phiên mã mRNA và mức độ biểu hiện protein (Kendrick, 2016). Nhiều yếu tố có thể góp phần vào sự tương quan thấp, bao gồm các kiểu biểu hiện cụ thể của tế bào, các sửa đổi sau dịch mã và môi trường vi mô phức tạp của tế bào, trong đó nhiều tương tác mRNA-mRNA, mRNA-protein và protein-protein thường xuyên xảy ra. Tuy nhiên, các mối tương quan yếu tồn tại giữa các bản sao liên quan và protein mở ra khả năng dự đoán thuần túy theo hướng dữ liệu về mức protein từ các mức phiên mã.

Mặc dù kỹ thuật MS đã tiến bộ và cải thiện rất nhiều trong những năm gần đây, các vấn đề liên quan đến độ nhạy, kích thước của protein, độ hòa tan mẫu, sự phân tách và phân tích dữ liệu vẫn còn. Các phương pháp tiếp cận proteome dựa trên MS vẫn còn cần tối ưu hóa ở nhiều điểm. Tuy nhiên, tính linh hoạt và tiềm năng của phép đo phổ khối vẫn được khai thác triệt để, cho phép tiến hành những nghiên cứu proteomics chính xác quy mô lớn (Poulos *et al*, 2020). Trong những thời gian tới, nó sẽ cung cấp cái nhìn sâu sắc hơn về các góc không thể tiếp cận trước đây của sinh học tế bào. Sẽ có thể thấy được những tiến bộ trong proteogenomics về cả hình ảnh quang phổ khối của các tế bào đơn lẻ với độ phân giải rõ ràng của các tiểu phần tế bào chất và nhân. Hy vọng, khả năng LC-MS/MS sẽ cho phép nhận được dữ liệu chính xác, có thể tái tạo từ một tế bào duy nhất. Chúng ta đang học được rất nhiều về nguồn gốc phân tử của bệnh ung thư từ những tiến bộ nhanh chóng trong công nghệ đo lường phân tử ... kiến thức được chuyển thành những tiến bộ hữu hình trong hiểu biết về sinh học ung thư, dẫn đến nhiều lý do hơn bao giờ hết để hy vọng. Một số nghiên cứu đã chứng minh sự liên quan của proteogenomics trong nghiên cứu ung thư. Xem xét những tiến bộ đạt được trong lĩnh vực này trong những năm gần đây, các nhà nghiên cứu cho rằng chính proteogenomics là sự tích hợp có hệ thống và toàn diện của proteomics với genomics và transcriptomics. Nghiên cứu proteogenomic có khả năng tiết lộ những hiểu biết có thể mở ra những bí ẩn của các quá trình sinh học phức tạp. Trong tương lai, các nhà khoa học đang hướng tới việc tích hợp thêm dữ liệu về chuyển hóa (metabolomics) để tạo ra bức tranh hoàn chỉnh hơn về một sinh vật và trạng thái sinh học của nó. Việc kết hợp các lĩnh vực này lại với nhau sẽ đòi hỏi sự hợp tác của các nhà khoa học với nhiều chuyên môn đa dạng, cùng với những tiến bộ hơn nữa của các công cụ tin sinh học có thể tích hợp lượng lớn các dữ liệu định lượng với các quá trình chuyển hóa sinh học đã biết. Proteogenomics đang mở ra những dấu ấn mới trong nghiên cứu y sinh và sẽ làm cho y học chính xác trong tương lai không xa.

Lời cảm ơn: *Công trình được hoàn thành với sự hỗ trợ của đề tài "Xây dựng Bản đồ Công nghệ Protein & Enzyme" (2019-2020, Mã số ĐM.43.DA/19, Chương trình Đổi mới Công nghệ Quốc gia đến năm 2020).*

REFERENCES

Aebersold R, Mann M (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* 537(7620):347-355. doi:10.1038/nature19949.

Armengaud J, Pible O, Gaillard JC, Cian A, Gantois N, Tan KSW, Chabe M, Viscogliosi E (2017). Proteogenomic Insights into the Intestinal Parasite Blastocystis sp. Subtype 4 Isolate WR1. *Proteomics* 17(21):10.1002/pmic.201700211. doi:10.1002/pmic.201700211.

Binetti M, Lauro A, Vaccari S, Cervellera M, Tonini V (2020). Proteogenomic biomarkers in colorectal cancers: clinical applications. *Expert Rev Proteomics* 17(5):355-363. doi:10.1080/ 14789450.2020.1782202

Cesnik AJ, Shortreed MR, Sheynkman GM, Frey BL, Smith LM (2016). Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. *J Proteome Res* 15(3):800-808. doi:10.1021/acs.jproteome.5b00817.

Chapman B, Bellgard M (2017). Plant Proteogenomics: Improvements to the Grapevine Genome Annotation. *Proteomics* 17(21):10, doi 10.1002/pmic.201700197.

Chen MX, Zhu FY, Gao B, Ma KL, Zhang Y, Fernie AR, Chen X, Dai L, Ye NH, Zhang X, Tian Y, Zhang D, Xiao S, Zhang J, Liu YG (2020). Full-Length Transcript-Based Proteogenomics of Rice Improves Its Genome and Proteome Annotation. *Plant Physiol* 182(3):1510-1526. doi:10.1104/ pp.19.00430.

Clark DJ, Dhanasekaran SM, Petralia F, *et al* (2020). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma [published correction appears in *Cell*. 179(4):964-983.e31. doi:10.1016/j.cell.2019.10.007.

Datta KK, Madugundu AK, Gowda H (2016a). Proteogenomic Methods to Improve Genome Annotation. *Methods Mol Biol* 1410:77-89. doi:10.1007/978-1-4939-3524-6_5.

Datta KK, Patil AH, Patel K, G. Dey, Madugundu AK, Renuse S, Kaviyil JE, Sekhar R, Arunima A, Daswani B, Kaur I, Mohanty J, Sinha R, Jaiswal S, Sivapriya S, Sonnathi Y, Chattoo BB, Gowda H, Ravikumar R, Prasad TSK (2016b). Proteogenomics of *Candida tropicalis*-An Opportunistic Pathogen with Importance for Global Health. *OMICS* 20, 239.

Dimitrakopoulos L, Prassas I, Berns EMJJ, Foekens JA, Diamandis EP, Charames GS (2017). Variant peptide detection utilizing mass spectrometry: laying the foundations for proteogenomic identification and validation. *Clin Chem Lab Med* 55(9):1291-1304. doi:10.1515/cclm-2016-0947.

Giess A, Jonckheere V, Ndah E, Chyżyńska K, Van Damme P, Valen E (2017). Ribosome signatures aid bacterial translation initiation site identification. *BMC Biol* 15(1):76. doi:10.1186/s12915-017-0416-0.

González-Gomariz J, Guruceaga E, López-Sánchez M, Segura V (2019). Proteogenomics in the context of the Human Proteome Project (HPP). *Expert Rev Proteomics*. 16(3):267-275. doi:10.1080/ 14789450.2019.1571916

Hwang H, Park GW, Park JY, Lee HK, Lee JY, Jeong JE, Park SKR, Yates 3rd JR, Kwon KH, Park YM, Lee HJ, Paik YK, Kim JY, Yoo JS (2017). Next Generation Proteomic Pipeline for Chromosome-Based Proteomic Research Using NeXtProt and GENCODE Databases. *J Proteome Res* 16(12):4425-4434. doi:10.1021/acs.jproteome.7b00223.

Kendrick N (2016). A gene's mRNA level does not usually predict its protein level. https://kendricklabs.com/wp-content/uploads/2016/08/WP1_mRNAvsProtein_KendrickLabs.pdf

Kuznetsova KG, Kliuchnikova AA, Ilina IU, Chernobrovkin AL, Novikova SE, Farafonova TE, Karpov DS, Ivanov MV, Goncharov AO, Ilgisonis EV, Voronko OE, Nasaev SS, Zgoda VG, Zubarev RA, Gorshkov MV, and Moshkovskii SA (2018). Proteogenomics of adenosine-to-inosine RNA editing in the fruit fly. *J Proteome Res* 17: 3889-3903. doi: https://doi.org/10.1021/ acs.jproteome.8b00553.

Lobas AA, Karpov DS, Kopylov AT, Solovyeva EM, Ivanov MV, IlinaIY, Lazarev N, Kuznetsova KG, Ilgisonis EV, Zgoda VG, Góhkov MV, Moshkovskii SA (2016). Exome-based proteogenomics of HEK-293 human cell line: Coding genomic variants identified at the level of shotgun proteome. *Proteomics* 16(14):1980-1991. doi:10.1002/pmic.201500349

Low TY, Mohtar MA, Ang MY, Jamal R (2019). Connecting Proteomics to Next-Generation Sequencing: Proteogenomics and Its Current Applications in Biology. *Proteomics*. 19(10): e1800235. doi:10.1002/pmic.201800235.

Ma Y, Huang T, Zhong X, Zhong XM, Zhang HW, Cong XL, Xu H, Lu GX, Yu F, Xue SB & Fu D (2018). Proteogenomic characterization and comprehensive integrative genomic analysis of human colorectal cancer liver metastasis. *Mol Cancer* 17, 139. https://doi.org/10.1186/s12943-018-0890-1

McAfee A, Harpur BA, Michaud S, Beavis RC, Kent CF, Zayed A, Foster LJ (2016). Toward an Upgraded Honey Bee (*Apis mellifera* L.) Genome Annotation Using Proteogenomics. *Proteome Res* 15: 411.

Mahadevan C, Krishnan A, Saraswathy GG, Surendran A, Jaleel A, Sakuntala M (2016). Transcriptome-Assisted Label-Free Quantitative Proteomics Analysis Reveals Novel Insights into Piper nigrum-Phytophthora capsici Phytopathosystem. *Front. Plant Sci* 7:785.

Mao Y, Yang Y, Liu Y, Yan Y, Du Z, Han Y, Song Y, Zhou L, Cui Y, Yang R (2016). Reannotation of Yersinia pestis Strain 91001 Based on Omics Data. *Am J Trop Med Hyg* 95:562.

Mitchell NM, Sherrard AL, Dasari S, Magee DM, Grys TE, Lake DF (2018). Proteogenomic Re-Annotation of *Coccidioides posadasii* Strain Silveira. *Proteomics* 18(1):10, doi 10.1002/pmic. 201700173.

Mertins P, Tang LC, Krug K, Clark DJ, Gritsenko MA, Chen L, Clauser KR, Clauss TR, Shah P, Gillette MA, Petyuk VA, Thomas SN, Mani DR, Mundt F, Mồe RJ, Hu Y, Zhao R, Schnaubelt M, Keshishian H, Monroe ME, Zhang Z, Udeshi ND, Mani D, Davies SR, Townsend RR, Chan DW, Smith RD, Zhang H, Liu T, Carr SA (2018). Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nat Protoc* 13(7):1632-1661. doi:10.1038/s41596-018-0006-9.

Mun D, Bhin J, Kim S, Kim H, Jung J, Jung Y, Jang Y, Park J, Kim H, Jung Y, Lee H, Bae J, Back S, Kim S, Kim J, Park H, Li H, Hwang K, Park Y, Yook J, Kim B, Kwon S, Ryu S, Park D, Jeon T, Kim D, Lee J, Han S, Song K, Park D, Park J, Rodriguez H, Kim

J, Lee H, Kim K, Yang E, Kim H, Paek E, Lee S, Lee S and Hwang D (2019). Proteogenomic Characterization of Human Early-Onset Gastric Cancer. *Cancer Cell* 35(1): 111-124.e10.

Nassa G, Giurato G, Cimmino G, Rizzo F, Ravo M, Salvati A, Nyman TA, Zhu Y, Vesterlund M, Lehtio J, Golino P, Weisz A, Tarallo R (2018). Splicing of platelet resident pre-mRNAs upon activation by physiological stimuli results in functionally relevant proteome modifications. *Sci Rep* 8(1): 498. doi:10.1038/s41598-017-18985-5.

Nesvizhskii A (2014). Proteogenomics: concepts, applications and computational strategies. *Nature Methods*. 11(11):1114-1125.

Nishimura T and Nakamura H (2016). Developments for Personalized Medicine of Lung Cancer Subtypes: Mass Spectrometry-Based Clinical Proteogenomic Analysis of Oncogenic Mutations *Adv Exp Med Biol* 926:115-137. doi: 10.1007/978-3-319-42316-6_8.

Omasits U, Varadarajan AR, Schmid M, Goetze S, Melidis D, Bourqui M, Nikolayeva O, Quebatte M, Patrignani A, Dehio C, Frey JE, Robinson MD, Wollscheid B, Ahrens CH (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res* 27: 2083. 27(12):2083-2095. doi:10.1101/gr.218255.116.

Omenn GS, Lane L, Lundberg EK, Overall CM, Deutsch EW (2017). Progress on the HUPO Draft Human Proteome: Metrics of the Human Proteome Project. *J Proteome Res* 16(12):4281-4287. doi:10.1021/acs.jproteome.7b00375.

Omenn GS, Lane L, Overall CM, Corrales FJ, Schwenk JM, Paik YK, Van Eyk JE, Liu S, Pennington S, Snyder M, Baker MS, Deutsch EW (2019). Progress on Identifying and Characterizing the Human Proteome: 2019 Metrics from the HUPO Human Proteome Project. *J Proteome Res* 18(12):4098-4107. doi:10.1021/acs.jproteome.9b00434

Poulos RC, Hains PG, Shah R, Jucas N, Xavier D, Manda SS, Anees A, Koh JMS, Mahboob S, Wittman M, William SG, Sykes EK, Hecker M, Dausmann M, Wouters MA, Ashman K, Yang J, Wild PJ, deFazio A, Balleine RL, Tully B, Aebersold R, Speed TP, Liu Y, Reddel RR, Rbinson PJ & Zhong Q (2020). Strategies to enable large-scale proteomics for reproducible research. *Nat Commun* 11**,** 3793. https://doi.org/10.1038/s41467-020-17641-3.

Rang J, He H, Wang T, Ding X, Zuo M, Quan M, Sun Y, Yu Z, Hu S, Xia L (2015). Comparative analysis of genomics and proteomics in *Bacillus thuringiensis* 4.0718. *PLoS One* 10(3):e0119065. doi:10.1371/journal.pone.0119065.

Robinson SD, Safavi-Hemami H, Raghuraman S, Imperial JS, Papenfuss AT, Teichert RW, Purcell AW, Olivera BM, Norton RS (2015). Discovery by proteogenomics and characterization of an RF-amide neuropeptide from cone snail venom. *J Proteomics* 114:38-47. doi:10.1016/j.jprot.2014. 11.003.

Ruggles KV, Krug K, Wang X, Clauser K R, Wang J, Payne SH, Fenyö D, Zhang B, & Mani DR (2017). Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteom* 16(6), 959-981. https://doi.org/10.1074/mcp.MR117.000024.

Satpathy S, Jaehnig EJ, Krug K, Kim BJ, Saltzman AB, Chan DW, Holloway KR, Anurag M, Huang C, Singh P, Gao A, Namai N, Dou Y, Wen B, Vasaikar SV, Mutch D, Watson MA, Ma C, Ademuyiwa FO, Rimawi MF, Schiff R, Hoog J, Jacobs S, Malovannaya A, Hyslop T, Clauser KR, Mani DR, Perou CM, Miles G, Zhang B, Gillette MA, Carr SA, Ellis MJ (2020). Microscaled proteogenomic methods for precision oncology. *Nat Commun* 11(1):532. doi:10.1038/s41467-020-14381-2.

Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R (2017). Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat Protoc* 12(7):1289-1294. doi:10.1038/nprot.2017.040

Syme RA, Tan KC, Hane JK, Dodhia K, Stoll T, Hastie M, Furuki E, Ellwood RS, Williams AH, Tan YF, Testa AC, Gorman JJ, Oliver RP (2016). Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLoS One* 11(2):e0147221. https// doi:10.1371/journal.pone.0147221.

Timp W và Tim G (2020). Beyond mass spectrometry, the next step in proteomics. *Sci Adv* 6: eaax8978. doi: 10.1126/sciadv.aax8978.

Vasaikar S, Huang C, Wang X, et al (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* 177(4):1035-1049.e19. doi:10.1016/j.cell.2019. 03.030

Wingo TS, Duong DM, Zhou M, Dammer EB, Wu H, Cutler DJ, Lah JJ, Levey AI, Seyfried NT (2017). Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. *J Proteome Res* 16(9):3336-3347. doi:10.1021/acs.jproteome.7b00324.

Woo S, Cha S, Bonissone S, Na S, Tabb D, Pevzner P and Bafna V (2015). Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.* 14(9): 3555-3567.

Ye X, Tang X, Wang X, Che J, Wu M, Liang J, Qian Q, Li J, You Z, Zhang Y, Wang S, Zhong B (2019). Improving Silkworm Genome Annotation Using a Proteogenomics Approach. *J Proteome Res* 2019;18(8):3009-3019. doi:10.1021/acs.jproteome.8b00965.

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers M, Zimmerman L, Shaddox K, Kim S, Davies S, Wang S, Wang P, Kinsinger C, Rivers R, Rodriguez H, Townsend R, Ellis, M, Carr S, Tabb D, Coffey R, Slebos R and Liebler D (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature* 513(7518): 382-387.

Zhu Y, Engstrom PG, Tellgren-Roth C, Baudo CD, Kennell JC, Sun S, Billmyre RB, Schroder MS, Andersson A, Holm T, Sigurgeirsson B, Wu G, Sankaranarayanan SR, Siddharthan R, Sanyal K, Lundeberg J, Nystedt B, Boekhout T, Dawson TLJ, Heitman J, Scheynius A, Lehtio J (2017). Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of *Malassezia sympodialis*. *Nucleic Acids Res* 45(5):2629-2643. doi: 10.1093/nar/gkx006.

PROTEOGENOMICS AND ITS APPLICATIONS IN BIOLOGY AND PRECISION MEDICINE

Phan Van Chi, Le Thi Bich Thao

Institute of Biotechnology, Vietnam Academy of Science and Technology

SUMMARY

In this review, we briefly discuss proteogenomics, the integration of proteomics with genomics and transcriptomics, whereby the underlying technologies are next-generation sequencing (NGS) and mass spectrometry (MS) with processing the resulting data, an emerging field that promises to accelerate fundamental research related to transcription and translation, as well as its applicability. By combining genomic and proteomic information, scientists are achieving new results due to a more complete and unified understanding of complex molecular biological processes. Part of this review introduces some of the results of using proteogenomics in solving problems such as annotation, gene/genome re-annotation, including editing of open reading frames (ORFs), or improving a process to detect new genes in a number of different organisms, including humans. In particular, the paper also discusses the potential of proteogenomics through research achievements on human genome/proteome in precision medicine, especially in projects on phylogenetic and diagnostic research. and cancer treatment. The challenges and future of proteogenomics are also discussed and documented.

Keywords: Proteogenomics, Genomics, Transcriptomics, Proteomics, Next-generation sequencing (NGS), Mass spectrometry (MS)