

XÂY DỰNG BỘ RNA DÀI KHÔNG MÃ HÓA CHẨN ĐOÁN UNG THƯ VÚ TỪ DỮ LIỆU GIẢI TRÌNH TỰ

Trần Thanh Bình¹, Nguyễn Thị Việt Hà² và Trần Thị Thu Huyền^{2,✉}

¹Bệnh viện Quân y 103

²Học viện Quân y

Nghiên cứu này nhằm mục tiêu xây dựng bộ chẩn đoán lncRNA có độ chính xác cao để chẩn đoán ung thư biểu mô tuyến vú dựa trên dữ liệu giải trình tự thế hệ mới (RNA-seq) được khai thác từ cơ sở dữ liệu TCGA. Các lncRNA biểu hiện khác biệt được sàng lọc bằng thuật toán DESeq2, sau đó lựa chọn đặc trưng bằng hồi quy Logistic LASSO với kiểm định chéo 5 lần (chọn ngưỡng λ_{1se}). Độ ổn định của mô hình được kiểm chứng thông qua phương pháp lấy mẫu lại phân tầng lặp lại 50 lần. Từ 3.293 DE-lncRNAs ban đầu, nghiên cứu đã thiết lập được bộ chẩn đoán gồm 06 lncRNA (LINC01614, LINC02202, LINC01537, AC134312.5, LINC02185 và AC106897.1). Mô hình đạt AUC = 0,9997 (KTC 95%: 0,9987 - 1,0000); độ nhạy 100% (KTC 95%: 98,02 - 100,0%) và độ đặc hiệu 93,75% (KTC 95%: 69,77 - 99,84%) tại điểm cắt tối ưu. Kết quả thẩm định cho thấy tính ổn định cao (AUC trung bình 0,9976). Bộ chẩn đoán gồm 06 lncRNA cho thấy khả năng phân biệt mô ung thư biểu mô tuyến vú xâm nhập và mô vú bình thường với hiệu năng cao trên tập dữ liệu TCGA-BRCA. Kết quả nghiên cứu gợi ý vai trò tiềm năng của lncRNA trong hỗ trợ chẩn đoán ung thư vú. Tuy nhiên, cần có thêm các nghiên cứu xác thực trên các quần thể lâm sàng độc lập.

Từ khóa: Ung thư vú, lncRNA, LASSO, TCGA, dấu ấn sinh học.

I. ĐẶT VẤN ĐỀ

Ung thư biểu mô tuyến vú xâm nhập (Breast Invasive Carcinoma - BRCA) hiện là ung thư phổ biến nhất và là nguyên nhân gây tử vong hàng đầu do ung thư ở phụ nữ trên toàn cầu.¹ Theo thống kê của GLOBOCAN, tỷ lệ mắc mới và tử vong do bệnh lý này vẫn tiếp tục gia tăng, đặt ra gánh nặng lớn cho hệ thống y tế.² Mặc dù các phương pháp sàng lọc truyền thống như chụp nhũ ảnh (mammography), siêu âm và chụp cộng hưởng từ đã góp phần cải thiện tỷ lệ sống sót, nhưng chúng vẫn tồn tại những hạn chế nhất định về độ nhạy, đặc biệt trên mô vú đặc và tỷ lệ dương tính giả cao dẫn đến các can thiệp xâm lấn không cần thiết.^{3,4} Được xem là

tiêu chuẩn vàng trong chẩn đoán xác định ung thư vú, sinh thiết mô bệnh học, một phương pháp có tính xâm lấn, lại không phù hợp cho sàng lọc diện rộng và có thể gặp sai lệch do lấy mẫu không đại diện trong bối cảnh khối u có tính dị hợp cao.⁵ Do đó, việc phát triển các dấu ấn sinh học phân tử có độ đặc hiệu vượt trội nhằm hỗ trợ nâng cao độ chính xác trong chẩn đoán là một nhu cầu cấp thiết.

Trong những năm gần đây, các đoạn RNA dài không mã hóa (long non-coding RNAs - lncRNAs) ngày càng được ghi nhận là những nhân tố điều hòa quan trọng trong cơ chế sinh bệnh học ung thư. Khác với các gen mã hóa protein, lncRNA thể hiện tính đặc hiệu mô và trạng thái bệnh lý cao hơn, đồng thời có thể phát hiện được trong các mẫu sinh thiết lỏng, khiến chúng trở thành biomarker lý tưởng cho mục tiêu chẩn đoán và tiên lượng.^{6,7,8} Sự phát triển bùng nổ của công nghệ giải trình tự thế hệ

Tác giả liên hệ: Trần Thị Thu Huyền

Học viện Quân y

Email: tranthuhuyen.cnsh@vmmu.edu.vn

Ngày nhận: 08/05/2026

Ngày được chấp nhận: 08/06/2026

mới (NGS) cùng các cơ sở dữ liệu lớn như The Cancer Genome Atlas (TCGA) đã cung cấp nguồn tài nguyên vô giá để khai thác toàn diện hệ phiên mã của ung thư vú. Tuy nhiên, thách thức lớn nhất khi làm việc với dữ liệu biểu hiện gen đa chiều là hiện tượng đa cộng tuyến và nhiễu, khiến việc xác định các dấu ấn sinh học thực sự trở nên khó khăn. Các phương pháp thống kê truyền thống thường không đủ mạnh để xử lý hàng nghìn biến số cùng lúc. Trong bối cảnh đó, các thuật toán học máy, đặc biệt là hồi quy LASSO (Least Absolute Shrinkage and Selection Operator), đã chứng minh hiệu quả trong việc chọn lọc biến số tối ưu thông qua cơ chế điều chuẩn hoá, từ đó giúp giảm nguy cơ quá khớp và xây dựng các mô hình đa gen có tính gọn nhẹ hơn.⁹

Mặc dù nhiều nghiên cứu đã đề xuất các dấu ấn phân tử dựa trên dữ liệu TCGA, tính ổn định và khả năng lặp lại của bộ chẩn đoán lncRNA vẫn còn là vấn đề cần được đánh giá thêm, đặc biệt trong bối cảnh số lượng biến lớn và nguy cơ sai lệch do mất cân bằng mẫu.^{10,11} Do đó, việc xây dựng một bộ tối giản dựa trên lncRNA và kiểm chứng độ ổn định bằng các phương pháp lấy mẫu lặp lại là cần thiết để tăng tính tin cậy của mô hình. Xuất phát từ lý do trên, chúng tôi tiến hành xây dựng và đánh giá hiệu năng của bộ RNA dài không mã hóa chẩn đoán ung thư vú từ dữ liệu giải trình tự.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Dữ liệu giải trình tự RNA (RNA-seq) và thông tin lâm sàng cơ bản đi kèm của bệnh nhân ung thư biểu mô tuyến vú xâm nhập (Breast Invasive Carcinoma - BRCA) được khai thác từ cơ sở dữ liệu The Cancer Genome Atlas (TCGA) thông qua cổng dữ liệu Genomic Data Commons (GDC) (<https://portal.gdc.cancer.gov/>).

Tiêu chuẩn chọn mẫu: Các mẫu được đưa vào nghiên cứu phải thỏa mãn đồng thời các điều kiện sau:

- (1) Được chẩn đoán xác định là ung thư biểu mô tuyến vú qua mô bệnh học;
- (2) Có sẵn dữ liệu giải trình tự RNA (RNA-seq) hoàn chỉnh;
- (3) Có đầy đủ thông tin lâm sàng cơ bản.

Tiêu chuẩn loại trừ: Nghiên cứu loại bỏ các trường hợp

- (1) mẫu thiếu dữ liệu biểu hiện;
- (2) Các mẫu có dữ liệu lâm sàng không rõ ràng hoặc thiếu hụt thông tin quan trọng ảnh hưởng đến quá trình phân loại.

2. Phương pháp

Thiết kế nghiên cứu

Nghiên cứu được thiết kế theo phương pháp hồi cứu, sử dụng dữ liệu thứ cấp.

Cỡ mẫu và phương pháp chọn mẫu

Các mẫu nghiên cứu được lựa chọn theo phương pháp chọn mẫu toàn bộ dựa trên các tiêu chí đưa vào và loại trừ. Cỡ mẫu phân tích gồm 200 mẫu bệnh phẩm từ 184 bệnh nhân ung thư biểu mô tuyến vú, trong đó có 184 mẫu mô khối ung thư nguyên phát và 16 mẫu mô lành cận u.

Thu thập và tiền xử lý

Dữ liệu giải trình tự RNA (RNA-seq) của các bệnh nhân ung thư biểu mô tuyến vú xâm nhập được thu thập từ cơ sở dữ liệu The Cancer Genome Atlas (TCGA) thông qua cổng GDC Data Portal. Nghiên cứu sử dụng dữ liệu *định lượng dạng thô* (STAR-Counts) được căn chỉnh theo hệ gen tham chiếu GRCh38. Sau khi tải xuống, dữ liệu được sử dụng thông tin chú giải gen đi kèm từ GDC để tách lọc các lncRNA. Các lncRNA có count ≥ 10 ở dưới 3 mẫu được loại bỏ như một bước lọc thô (pre-filtering) trước khi thực hiện phân tích DESeq2.

Đánh giá và sàng lọc các lncRNA biểu hiện khác biệt

Phân tích biểu hiện khác biệt được thực hiện bằng thuật toán DESeq2 trên môi trường R. Các lncRNA được xác định là biểu hiện khác biệt có ý nghĩa khi đồng thời thỏa mãn hai tiêu chí: $|\log_2\text{FoldChange}| > 1$ và $\text{padj} < 0,05$ (hiệu chỉnh theo phương pháp Benjamini-Hochberg).

Xây dựng và tối ưu hóa bộ chẩn đoán

Dữ liệu đếm của các lncRNA có mức ý nghĩa thống kê cao nhất được chuyển đổi chuẩn hóa theo phương pháp Ổn định phương sai (Variance Stabilizing Transformation - VST) trước khi đưa vào phân tích. Bộ chẩn đoán được xây dựng bằng thuật toán hồi quy Logistic LASSO, trong đó tham số phạt λ_{1se} được xác định thông qua quy trình kiểm định chéo 5 lần (5-fold cross-validation). Chỉ các lncRNA có hệ số hồi quy khác 0 tại ngưỡng λ_{1se} mới được giữ lại để thiết lập công thức tính điểm chẩn đoán (Diagnostic Score).

Đánh giá độ ổn định và hiệu năng của Bộ chẩn đoán

Để kiểm chứng độ ổn định và giảm ảnh hưởng của tình trạng mất cân bằng mẫu giữa nhóm bệnh và nhóm chứng, mô hình được đánh giá bằng phương pháp lấy mẫu lại phân tầng lặp lại 50 lần độc lập với tỷ lệ phân chia tập huấn luyện và tập kiểm tra lần lượt là 80% và 20%, đồng thời vẫn bảo toàn tỷ lệ giữa hai nhóm trong mỗi lần lặp. Hiệu năng tổng thể của mô hình được đánh giá dựa trên giá trị trung bình diện tích dưới đường cong ROC (Mean AUC) qua 50 lần lặp. Bên cạnh đó, các chỉ số độ nhạy, độ đặc hiệu và độ chính xác được tính toán từ ma trận nhầm lẫn tại ngưỡng cắt xác

suất mặc định là 0,5. Đồng thời, chỉ số Youden được sử dụng để xác định điểm cắt tối ưu nhằm tối đa hóa tổng độ nhạy và độ đặc hiệu trên toàn bộ tập dữ liệu.

Phân tích số liệu

Toàn bộ số liệu được xử lý và phân tích bằng R phiên bản 4.2.1 thông qua môi trường RStudio với các gói TCGAbiolinks, SummarizedExperiment, DESeq2, glmnet, pROC, dplyr và pheatmap.

3. Đạo đức nghiên cứu

Nghiên cứu sử dụng nguồn dữ liệu thứ cấp đã được mã hóa và ẩn danh hoàn toàn từ cơ sở dữ liệu công khai The Cancer Genome Atlas (TCGA). Do không can thiệp trực tiếp lên đối tượng con người, nghiên cứu được miễn trừ các yêu cầu về sự chấp thuận của Hội đồng Đạo đức và văn bản đồng ý tham gia. Nhóm nghiên cứu cam kết tuân thủ nghiêm ngặt các quy định về quyền truy cập, trích dẫn dữ liệu của mạng lưới TCGA và các nguyên tắc trong Tuyên ngôn Helsinki.

III. KẾT QUẢ

1. Đặc điểm chung đối tượng nghiên cứu

Đối tượng tham gia nghiên cứu bao gồm 200 mẫu bệnh phẩm được thu thập từ 184 bệnh nhân ung thư biểu mô tuyến vú xâm nhập trên cơ sở dữ liệu TCGA, bao gồm 184 mẫu mô ung thư và 16 mẫu mô lành cận u. Kết quả thống kê cho thấy độ tuổi trung vị là 62 tuổi, trong đó có 1 trường hợp (0,55%) là nam giới, và phần lớn các khối u được phát hiện ở giai đoạn sớm. Chi tiết các đặc điểm lâm sàng và mô bệnh học của đối tượng nghiên cứu được trình bày tại Bảng 1.

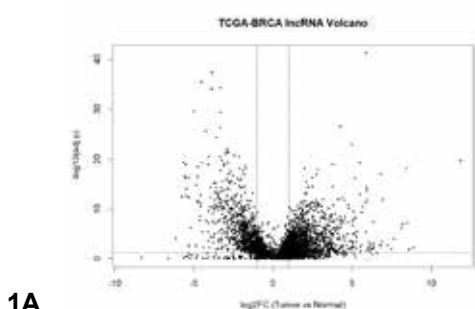
Bảng 1. Đặc điểm chung của đối tượng tham gia nghiên cứu (n = 184)

Đặc điểm	Giá trị (n = 184)
Tuổi tại thời điểm chẩn đoán (năm)	
Trung vị (Min-Max)	62 (29-90)
Giới tính, n (%)	
Nữ	183 (99,45)
Nam	01 (0,55)
Giai đoạn bệnh (AJCC Stage), n (%)	
Giai đoạn 0-I	39 (21,2)
Giai đoạn II	84 (45,65)
Giai đoạn III	37 (20,11)
Giai đoạn IV	02 (1,09)
Không rõ (Unknown)	22 (11,96)
Phân nhóm phân tử (PAM50), n (%)	
Luminal A	92 (50,0)
Luminal B	38 (20,7)
HER2-enriched	11 (6,0)
Basal-like	30 (16,3)
Normal-like	3 (1,6)
Không xác định	10 (5,4)

2. Kết quả phân tích biểu hiện khác biệt lncRNA giữa mô u và mô lành

Phân tích biểu hiện khác biệt lncRNA giữa nhóm nhóm mô ung thư (n = 184) và nhóm mô

lành (n = 16) bằng DESeq2 cho thấy sự khác biệt rõ rệt. Tổng cộng ghi nhận 3.293 lncRNA khác biệt có ý nghĩa thống kê với ($|\log_2FC| > 1$ và $\text{padj} < 0,05$) (Biểu đồ 1A).

**1A****1B**

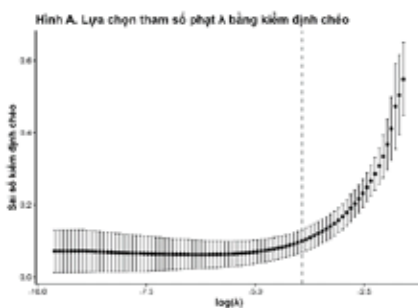
Biểu đồ 1. Sàng lọc các lncRNA biểu hiện khác biệt trong ung thư biểu mô tuyến vú xâm nhập trên tập dữ liệu TCGA-BRCA

Hình 1A. Biểu đồ Volcano. Biểu diễn sự phân bố mức độ thay đổi biểu hiện của toàn bộ lncRNA được phân tích. Trục hoành biểu thị logarit cơ số 2 của hệ số thay đổi ($\log_2\text{FoldChange}$), trục tung biểu thị giá trị đối nghịch logarit cơ số 10 của giá trị p điều chỉnh ($-\log_{10}p_{\text{adj}}$). Các điểm chấm nằm ngoài hai đường kẻ dọc và phía trên đường kẻ ngang đại diện cho các lncRNA có sự khác biệt có ý nghĩa thống kê ($|\log_2\text{FC}| > 1$ và $p_{\text{adj}} < 0,05$).

Hình 1B. Biểu đồ nhiệt (Heatmap): Phân cụm phân cấp (Hierarchical clustering) dựa trên mức độ biểu hiện của 20 lncRNA ứng viên hàng đầu (top 20 DE-lncRNAs) có giá trị p_{adj} . Các cột tương ứng với các mẫu bệnh phẩm (nhóm Mô lành - Normal và nhóm mô u - Tumor), các hàng tương ứng với tên các lncRNA. Màu sắc thể hiện giá trị Z-score của dữ liệu đếm đã chuẩn hóa VST: màu xanh lam biểu thị mức biểu hiện thấp (down-regulation) và màu vàng/đỏ biểu thị mức biểu hiện cao (up-regulation).

Để lựa chọn các dấu ấn sinh học tiềm năng nhất cho xây dựng mô hình chẩn đoán, nghiên cứu đã xếp hạng và lựa chọn 20 lncRNA có giá trị p_{adj} nhỏ nhất làm tập ứng viên. Sự phân tách rõ rệt giữa nhóm bệnh và nhóm chứng dựa trên mức độ biểu hiện của 20 ứng viên hàng đầu được minh họa cụ thể qua biểu đồ heatmap (Biểu đồ 1B).

3. Xây dựng bộ chẩn đoán bằng thuật toán LASSO



Để sàng lọc các biến số quan trọng nhất và hạn chế hiện tượng quá khớp, mô hình hồi quy logistic kết hợp chọn biến LASSO đã được áp dụng cho 20 lncRNA ứng viên được lựa chọn từ phân tích biểu hiện khác biệt.

Tham số phạt tối ưu của mô hình được xác định thông qua giá trị kiểm định chéo 5 lần. Kết quả cho thấy giá trị λ_{1se} là 0,01968834 (Biểu đồ 2A).

Tại ngưỡng λ_{1se} , thuật toán đã loại bỏ các biến ít đóng góp và giữ lại 06 lncRNA có hệ số hồi quy khác 0 để thiết lập Bộ chẩn đoán bao gồm: *LINC01614*, *LINC02202*, *LINC01537*, *AC134312.5*, *LINC02185* và *AC106897.1*. Mức độ đóng góp là chiều tác động của từng lncRNA trong mô hình đa biến được thể hiện thông qua hệ số hồi quy (Biểu đồ 2B).

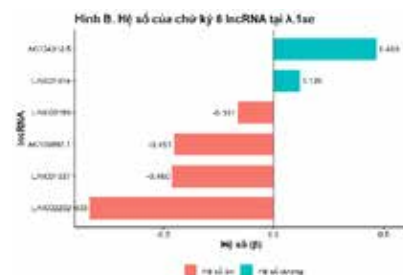
Công thức tính điểm dự báo (Diagnostic Score) của 6 lncRNA được thiết lập như sau:

$$\text{Diagnostic Score} = 10317 + 0,120 \times \text{LINC01614} + 0,469 \times \text{AC134312.5} - 0,161 \times \text{LINC02185} - 0,451 \times \text{AC106897.1} - 0,4 \times \text{LINC01537} - 0,5 \times \text{LINC02202}$$

(Trong đó, giá trị của từng lncRNA là mức biểu hiện đã được chuẩn hóa bằng phương pháp variance stabilizing transformation (VST).)

Xác suất mẫu thuộc nhóm mô ung thư vú được suy ra từ điểm dự báo theo hàm logistic:

$$P(\text{mô ung thư vú}) = \frac{1}{1 + e^{\text{diagnosticscore}}}$$



Biểu đồ 2. Quy trình chọn biến tối ưu và xác định hệ số hồi quy cho bộ chẩn đoán lncRNA chẩn đoán bằng thuật toán LASSO

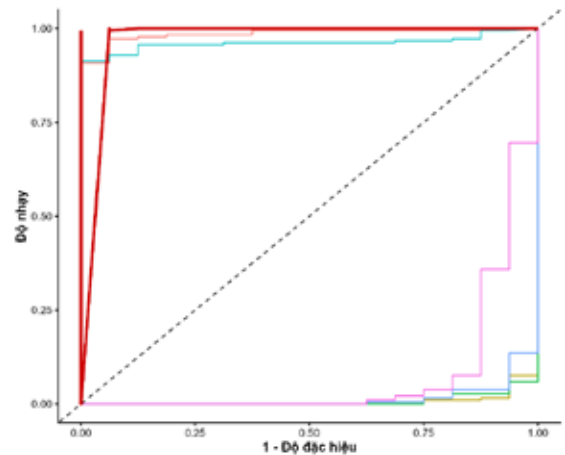
Biểu đồ 2A. Lựa chọn tham số phạt λ : Biểu đồ hiển thị sai số kiểm định chéo theo logarit của tham số λ . Đường cong biểu diễn sai số trung bình; các thanh dọc thể hiện độ lệch chuẩn. Đường kẻ dọc nét đứt chỉ giá trị λ_{1se} (giá trị λ lớn nhất nằm trong khoảng 1 độ lệch chuẩn của sai số tối thiểu), được chọn để xây dựng mô hình tinh gọn và tránh quá khớp.

Biểu đồ 2B. Hệ số hồi quy của 06 lncRNA trong bộ chẩn đoán. Biểu đồ cột ngang thể hiện độ lớn và chiều hướng đóng góp của 6 lncRNA được mô hình LASSO lựa chọn tại λ_{1se} . Các thanh màu xanh (hệ số dương) cho thấy mối tương quan thuận với nguy cơ ung thư (*LINC01614*, *AC134312.5*), trong khi các thanh màu đỏ (hệ số âm) cho thấy mối tương quan nghịch (*LINC02185*, *AC106897.1*, *LINC01537*, *LINC02202*).

4. Đánh giá hiệu năng và độ ổn định của bộ chẩn đoán

Kết quả kiểm chứng độ ổn định bằng phương pháp lấy mẫu lại phân tầng lặp lại (50 lần) cho thấy mô hình hoạt động ổn định với giá trị AUC trung bình trên các tập kiểm tra đạt 0,9976 (KTC 95%: 0,9865 - 1,0000). Khi đánh giá trên toàn bộ đoàn hệ nghiên cứu, bộ chẩn đoán được thiết lập từ 06 lncRNA thể hiện hiệu năng chẩn đoán vượt trội với diện tích dưới đường cong ROC (AUC) đạt 0,9997 (KTC 95%: 0,9987 - 1,0000) (**Biểu đồ 3**).

Tại điểm cắt tối ưu được xác định theo chỉ số Youden, mô hình đạt độ nhạy 100,0% (KTC 95%: 98,02 - 100,0%) và độ đặc hiệu 93,75% (KTC 95%: 69,77 - 99,84%) (tương ứng với giá trị chỉ số Youden = 0,9375). Đặc biệt, khả năng phân loại của bộ chẩn đoán cao hơn đáng kể so với từng lncRNA đơn lẻ (AUC dao động từ 0,9249 đến 0,9922), khẳng định giá trị hiệp đồng của mô hình đa gen trong chẩn đoán ung thư biểu mô tuyến vú.



Biểu đồ 3. So sánh hiệu năng chẩn đoán giữa bộ 06 lncRNA và các lncRNA đơn lẻ thông qua phân tích đường cong ROC

Biểu đồ thể hiện các đường cong đặc trưng hoạt động của bộ thu (ROC) dùng để đánh giá khả năng phân loại giữa mô ung thư và mô lành.

Đường màu đỏ đậm: Đại diện cho bộ chẩn đoán phối hợp 06 lncRNA. Đường này bao trùm lên tất cả các đường khác và tiệm cận sát góc trên bên trái, tương ứng với giá trị AUC cao nhất (0,9997), khẳng định độ chính xác vượt trội của mô hình đa biến.

Các đường màu mảnh: Đại diện cho đường cong ROC của từng lncRNA đơn lẻ trong bộ chẩn đoán. Các đường này đều nằm dưới đường màu đỏ, cho thấy hiệu năng chẩn đoán độc lập của từng gen thấp hơn so với khi phối hợp.

Trục tọa độ: Trục tung biểu thị Độ nhạy (Sensitivity) và trục hoành biểu thị tỷ lệ dương tính giả (1 - Độ đặc hiệu).

IV. BÀN LUẬN

Nghiên cứu của chúng tôi đã xây dựng được một bộ chẩn đoán ung thư biểu mô tuyến vú dựa trên biểu hiện của 06 lncRNA (*LINC01614*, *LINC02202*, *LINC01537*, *AC134312.5*, *LINC02185* và *AC106897.1*) sử dụng thuật

toán LASSO. Kết quả cho thấy bộ chẩn đoán này đạt diện tích dưới đường cong ROC (AUC) là 0,9997 trên tập dữ liệu TCGA, cho thấy khả năng phân biệt rõ rệt giữa mô ung thư và mô lành. Mặc dù mô hình cho thấy hiệu năng phân loại rất cao trên dữ liệu TCGA-BRCA, các kết quả này vẫn cần được diễn giải trong bối cảnh nghiên cứu mang tính chất khám phá ban đầu dựa trên dữ liệu tin sinh học. Giá trị AUC rất cao có thể chịu ảnh hưởng bởi đặc điểm của bộ dữ liệu nghiên cứu và việc mô hình chưa được kiểm định trên tập dữ liệu độc lập bên ngoài. Do đó, cần có thêm các nghiên cứu xác thực trên các cohort lâm sàng độc lập nhằm đánh giá thêm tính ổn định và khả năng khái quát hóa của mô hình.

Đặt trong bối cảnh nghiên cứu dấu ấn sinh học ung thư vú, các phương pháp tiếp cận dựa trên biểu hiện gen đang được quan tâm nhờ khả năng phản ánh những thay đổi phân tử ở giai đoạn sớm của quá trình sinh ung thư. Trong nghiên cứu này, mô hình 06 lncRNA cho thấy hiệu năng phân loại cao trên dữ liệu RNA-seq mô của TCGA-BRCA, gợi ý giá trị tiềm năng của lncRNA trong hỗ trợ phân biệt mô ung thư và mô bình thường. Tuy nhiên, các kết quả hiện tại vẫn cần được tiếp tục xác thực trên các quần thể lâm sàng độc lập và các loại mẫu sinh học khác trước khi xem xét khả năng ứng dụng thực tế. Về mặt cơ chế sinh học, các gen được thuật toán lựa chọn trong bộ chẩn đoán có sự tương đồng với các công bố trước đây, củng cố tính hợp lý của mô hình. Cụ thể, lncRNA *LINC01614* được xác định có hệ số hồi quy dương (+0,120), đóng vai trò là yếu tố nguy cơ. Kết quả này phù hợp với các báo cáo gần đây về vai trò sinh học của lncRNA trong ung thư. Đặc biệt, Tian và cộng sự (2024) đã chỉ ra rằng lncRNA *LINC01614* được biểu hiện tăng cao trong nhiều loại ung thư và có liên quan chặt chẽ đến tiến triển khối u cũng như tiên lượng sống còn. Ngoài ra, *LINC01614* còn

tương quan với xâm nhập miễn dịch và trạng thái ức chế miễn dịch trong vi môi trường khối u, gợi ý vai trò điều hòa miễn dịch của lncRNA này. Tương tự, Wang và cs (2019) đã xác định lncRNA *LINC01614* là một dấu ấn sinh học tiên lượng tiềm năng trong ung thư vú dựa trên dữ liệu biểu hiện từ cơ sở TANCIC. Nghiên cứu cho thấy *LINC01614* được biểu hiện tăng cao rõ rệt trong mô ung thư so với mô bình thường lân cận. Đặc biệt, phân tích hồi quy Cox đơn biến và đa biến xác nhận rằng *LINC01614* là một yếu tố tiên lượng độc lập, trong đó mức biểu hiện cao liên quan đến nguy cơ tử vong tăng lên (hệ số hồi quy dương). Ngoài ra, *LINC01614* tăng biểu hiện trong các nhóm phân tử như ER+, PR+ và HER2+, đồng thời tương quan với các pathway liên quan đến TGF- β và kết dính tế bào. Những kết quả này gợi ý rằng *LINC01614* có thể đóng vai trò thúc đẩy tiến triển và di căn trong ung thư vú.⁸ Một nghiên cứu gần đây của Chen và cs (2024) cho thấy *LINC01614* tăng biểu hiện trong ung thư vú thể bộ ba âm tính và thúc đẩy khả năng di cư, xâm lấn cũng như quá trình EMT thông qua hoạt hóa tín hiệu Wnt/ β -catenin.¹⁴ Đối với các lncRNA khác như *AC134312.5* và *AC106897.1*, hiện chưa có nhiều dữ liệu trong y văn. Đây có thể là những dấu ấn sinh học mới đặc hiệu cho quần thể bệnh nhân trong dữ liệu TCGA mà nghiên cứu này đã phát hiện ra.

Trong nghiên cứu này, chúng tôi xây dựng bộ chẩn đoán dựa trên hồi quy LASSO nhằm lựa chọn các biến dự báo quan trọng và giảm nguy cơ overfitting. Phương pháp này áp dụng ràng buộc chuẩn L1 giúp shrinkage các hệ số và đưa một số biến về 0, tạo mô hình gọn và dễ diễn giải. Friedman và cs (2010) đã phát triển thuật toán coordinate descent cho các mô hình tuyến tính tổng quát có regularization, bao gồm logistic LASSO, giúp nâng cao hiệu quả lựa chọn biến trong các bộ dữ liệu có số chiều lớn như dữ liệu biểu hiện gen.⁹ Trong nghiên cứu này, từ

hơn 3.000 lncRNA biểu hiện khác biệt ban đầu, LASSO đã tinh gọn xuống còn 06 lncRNA. Số lượng gen này tương đương hơn so với một số bộ chẩn đoán đa gen đã công bố trước đây giúp tăng tính khả thi và giảm chi phí khi triển khai ứng dụng trên lâm sàng thông qua kỹ thuật RT-qPCR, phù hợp với xu hướng phát triển các công cụ chẩn đoán tối thiểu xâm lấn.¹¹⁻¹³

Tuy nhiên, nghiên cứu của chúng tôi tồn tại một hạn chế quan trọng cần được ghi nhận. Hiện tại, các kết quả xây dựng bộ chẩn đoán mới chỉ dừng lại ở bước phân tích tin sinh học dựa trên cơ sở dữ liệu giải trình tự công khai. Nhóm nghiên cứu chưa tiến hành các thử nghiệm kiểm chứng thực tế bằng kỹ thuật định lượng gen (như RT-qPCR) trên các mẫu bệnh phẩm lâm sàng thu thập trực tiếp từ bệnh nhân. Do sự khác biệt tiềm năng về đặc điểm di truyền giữa các quần thể, việc ngoại suy kết quả từ dữ liệu TCGA cần được thực hiện thận trọng trước khi áp dụng cho bệnh nhân Việt Nam. Đây chính là nội dung trọng tâm mà chúng tôi sẽ giải quyết trong các nghiên cứu xác thực tiếp theo.

Tóm lại, bộ chẩn đoán 06 lncRNA được xây dựng thể hiện hiệu năng chẩn đoán cao và phù hợp với các cơ chế sinh học đã biết. Kết quả này cung cấp thêm bằng chứng khoa học cho việc sử dụng lncRNA như một công cụ hỗ trợ chẩn đoán ung thư biểu mô tuyến vú xâm nhập.

Hạn chế của nghiên cứu

Nghiên cứu này mang tính chất sàng lọc và khám phá ban đầu dựa trên dữ liệu RNA-seq từ cơ sở dữ liệu TCGA-BRCA. Bộ dữ liệu nghiên cứu có sự khác biệt về số lượng giữa mẫu mô u và mô lành, đồng thời mô hình hiện mới được đánh giá thông qua phương pháp lấy mẫu lại phân tầng trong cùng một cohort mà chưa được thẩm định trên tập dữ liệu độc lập bên ngoài. Ngoài ra, dữ liệu nghiên cứu chủ yếu được khai thác từ quần thể TCGA của phương

Tây. Trong thời gian tới, chúng tôi sẽ tiếp tục tiến hành các nghiên cứu xác thực trên quần thể lâm sàng độc lập và đánh giá bằng các kỹ thuật thực nghiệm nhằm kiểm chứng thêm tính ổn định và khả năng ứng dụng của mô hình.

V. KẾT LUẬN

Nghiên cứu đã xây dựng được bộ chẩn đoán gồm 06 lncRNA (*LINC01614*, *LINC02202*, *LINC01537*, *AC134312.5*, *LINC02185*, *AC106897.1*) có khả năng phân biệt mô ung thư biểu mô tuyến vú xâm nhập và mô vú bình thường thông qua thuật toán chọn biến LASSO. Mô hình đạt AUC = 0,9997 (KTC 95%: 0,9987 - 1,0000); độ nhạy 100% (KTC 95%: 98,02 - 100,0%) và độ đặc hiệu 93,75% (KTC 95%: 69,77 - 99,84%) tại điểm cắt tối ưu. Kết quả nghiên cứu cung cấp thêm cơ sở cho việc tiếp tục đánh giá vai trò của lncRNA trong hỗ trợ chẩn đoán ung thư vú ở các nghiên cứu xác thực tiếp theo. Tuy nhiên, các kết quả hiện tại vẫn cần được kiểm chứng trên các quần thể lâm sàng độc lập trước khi xem xét khả năng ứng dụng thực tế.

TÀI LIỆU THAM KHẢO

1. Sung H, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021; 71(3): 209-249.
2. Kim J, et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nat Med*. 2025; 31(4): 1154-1162.
3. Boyd NF, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*. 2007; 356(3): 227-236.
4. Wang S, et al. Over-detection and over-surveillance in breast screening: current status and the potential for artificial intelligence optimisation. *Insights Imaging*. 2025; 16(1): 276.
5. Polyak K. Heterogeneity in breast cancer.

J Clin Invest. 2011; 121(10): 3786-3788.

6. Xu S, et al. Oncogenic long noncoding RNA landscape in breast cancer. *Mol Cancer.* 2017; 16(1): 129.

7. El-Helkan B, et al. Long non-coding RNAs as novel prognostic biomarkers for breast cancer in Egyptian women. *Sci Rep.* 2022; 12(1): 19498.

8. Wang Y, Song B, Zhu L, et al. Long non-coding RNA LINC01614 as a potential biomarker for prognostic prediction in breast cancer. *PeerJ.* 2019; 7:e7976.

9. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1-22. Shaath H, Elango R, Alajez NM. Molecular classification of breast cancer utilizing long non-coding RNA transcriptomes identifies novel diagnostic lncRNA panel for triple-negative breast cancer. *Cancers (Basel).*

2021; 13(21).

10. Fan CN, Ma L, Liu N. Comprehensive analysis of novel three-long noncoding RNA signatures as diagnostic and prognostic biomarkers of human triple-negative breast cancer. *J Cell Biochem.* 2019; 120(3): 3185-3196.

11. Wu J, Cai Y, Zhao G, et al. A ten N6-methyladenosine-related long non-coding RNAs signature predicts prognosis of triple-negative breast cancer. *J Clin Lab Anal.* 2021; 35(6): e23779.

12. Xu M, et al. A seven-lncRNA signature for predicting prognosis in breast carcinoma. *Transl Cancer Res.* 2021;10(9):4033-4046.

13. Chen T, Shi K, Sun S. LINC01614 activated by SP1 promoted malignant behavior of triple-negative breast cancer cells via the WNT/ β -catenin signaling pathway. *Rev Invest Clin.* 2024; 76(4): 185-198.

Summary

CONSTRUCTION OF A LONG NON-CODING RNA PANEL FOR BREAST CANCER DIAGNOSIS USING RNA SEQUENCING DATA

This study aims to construct a robust lncRNA-based diagnostic panel for Breast Invasive Carcinoma (BRCA) using next-generation sequencing data. RNA-seq data of 184 tumor and 16 normal tissue samples were retrieved from The Cancer Genome Atlas (TCGA). Differentially expressed lncRNAs (DE-lncRNAs) were identified using DESeq2. To mitigate overfitting and select optimal features, the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression model was applied with 5-fold cross-validation (λ_{1se}). Model stability was validated via 50-times repeated stratified subsampling. From 3,293 significant DE-lncRNAs, a diagnostic panel comprising six lncRNAs (*LINC01614*, *LINC02202*, *LINC01537*, *AC134312.5*, *LINC02185*, and *AC106897.1*) was established. The panel demonstrated superior diagnostic performance with an Area Under the Curve (AUC) of 0.9997, significantly outperforming individual genes. At the optimal Youden index cut-off ($J = 0.9375$), the model achieved 100% sensitivity and 93.75% specificity. Validation via repeated subsampling confirmed high stability with a mean AUC of 0.9976. We successfully constructed a highly accurate and stable 6-lncRNA diagnostic panel for BRCA using LASSO. This panel may serve as a potential candidate for further evaluation in breast cancer diagnostic research.

Keywords: Breast Cancer, lncRNA, LASSO, Diagnosis, TCGA, molecular biomarker.