KHẢO SÁT CÁC BIẾN THỂ GEN LIÊN QUAN ĐẾN UNG THƯ VÚ BẰNG DỮ LIỆU GIẢI TRÌNH TỰ ARN

Phan Huy Giang[1], Hoàng Hồng Thắm[2], Võ Sỹ Nam[2], Nguyễn Hoàng Quân[3]
Trịnh Lê Huy[1], Vũ Minh Giang[2], Hoàng Yến[1] và Trần Huy Thịnh[1,✉]

[1]Trường Đại học Y Hà Nội
[2]Công ty GeneStory
[3]Đại học Queensland - Australia

Nghiên cứu này nhằm khảo sát các biến thể gen liên quan tới ung thư vú bằng dữ liệu giải trình tự ARN. Chúng tôi thực hiện nghiên cứu với 5 người bệnh ung thư vú và 8 đối chứng lấy từ dữ liệu VN1K. Trên 5 phụ nữ ung thư vú và 8 người khoẻ mạnh đối chứng có độ tuổi tương đồng nhau. Chúng tôi áp dụng phương pháp mô tả cắt ngang để tìm hiểu các biến thể dòng mầm có mặt ở bệnh nhân ung thư vú thông qua giải trình tự ARN mẫu máu. Phân tích kết quả giải trình tự ARN chúng tôi đã xác định được 143 gen có sự khác biệt biểu hiện đáng kể giữa nhóm ung thư vú và nhóm khỏe mạnh. Tiếp đó, chúng tôi thực hiện gọi biến thể trong 143 gen này được 3515 biến thể.Trong đó, 8 biến thể nguy cơ cao được phát hiện, có 2 biến thể đã được biết trong cơ sở dữ liệu dbSNP là rs35400274 và rs34406374. Một biến thể đáng chú ý là c.456G>A (rs35400274) tạo thành bộ ba kết thúc sớm nằm trên gen C17orf107, biến thể này làm giảm mức độ biểu hiện gen ở nhóm ung thư vú so với nhóm chứng Log2FoldChange = -2,49 và p-value =0,002, biến thể này cũng đã được báo cáo có mặt ở bệnh nhân ung thư đại trực tràng trong cơ sở dữ liệu COSMIC. Biến thể rs7937 được báo cáo liên quan tới tình trạng ung thư vú trong cơ sở dữ liệu GWAS-Catalog do ảnh hưởng tới nồng độ thuốc Letrozole. Kết quả từ nghiên cứu này cung cấp thêm những hiểu biết về con đường bệnh sinh và sự tác động của các biến thể gen với nguy cơ ung thư vú ở phụ nữ.

Từ khóa: Biến thể gen, ung thư vú, giải trình tự ARN.

I. ĐẶT VẤN ĐỀ

Ung thư vú (UTV) là bệnh lý ác tính phổ biến hàng đầu ở phụ nữ ở cả các nước phát triển và đang phát triển. Bệnh chiếm 25% tỉ lệ chết do ung thư ở các nước phát triển. Theo GLOBOCAL 2020, UTV ở nữ đã vượt qua ung thư phổi, trở thành bệnh ung thư hang đầu trong số ca mắc mới năm 2020, với gần 2.300.000 ca, chiếm 11,7% tổng số ca mới mắc. Tại Việt Nam, năm 2020 nữ giới UTV đứng hàng thứ 3 trong số các ca mới mắc với tỷ lệ mắc chuẩn theo tuổi 34,2/100.000 dân, độ tuổi hay gặp 40 - 49.[1] Căn nguyên bệnh sinh ung thư vú rất phức tạp, việc phòng ngừa, phát hiện sớm và điều trị còn gặp nhiều khó khăn. Nhiều nghiên cứu ở Mỹ và châu Âu cho rằng khoảng 10 - 15% ung thư vú có yếu tố gia đình, nghĩa là người bệnh mang các biến thể gen di truyền từ mẹ.[2] Dựa trên các kĩ thuật sinh học phân tử hiện đại ngày nay như giải trình tự thế hệ mới cho phép xác định chính xác các biến thể gen làm tang nguy cơ gây ung thư, giúp cho việc chẩn đoán và điều trị lâm sàng hiệu quả hơn.

Các nghiên cứu trước đây đã chỉ ra đa hình đơn nucleotide (SNP) có thể làm thay đổi biểu hiện của gen, do đó có thể ảnh hưởng tới chức năng và liên quan đến tăng hoặc giảm nguy cơ ung thư. Đối với ung thư vú, một số nghiên cứu trên thế giới chứng minh các SNP của gen *AKT1* và *PTEN* thuộc con đường truyền tín hiệu nội bào liên quan tới sự hình thành ung thư.[3]

Tác giả liên hệ: Trần Huy Thịnh
Trường Đại học Y Hà Nội
Email: tranhuythinh@hmu.edu.vn
Ngày nhận: 08/11/2023
Ngày được chấp nhận: 22/11/2023

Giải trình tự ARN (RNA-Seq) là một phương pháp được sử dụng rộng rãi để lập hồ sơ biểu hiện gen. Ngoài ra, tính hữu ích của RNA-Seq trong việc phát hiện tính đa hình nucleotide đơn (SNP) hiệu quả ở các gen mã hóa đã được chứng minh ở các mô và loài khác nhau.[4] So với giải trình tự DNA, gọi biến thể từ RNA-seq hiệu quả gần như tương tự. Hơn nữa, hầu hết SNP được xác định thông qua phân tích này đều nằm ở các vùng được phiên mã của gen, do đó các biến thể có nhiều khả năng dẫn đến thay đổi kiểu hình. Do đó, phương pháp RNA-Seq đã được sử dụng để xác định SNP dựa trên gen và điều tra xem liệu SNP đó có liên quan đến sự khác biệt về biểu hiện gen giữa bệnh nhân ung thư vú và người bình thường.

Tại Việt Nam, các biến thể gen gây bệnh ở bệnh nhân ung thư vú chưa được nghiên cứu rộng rãi. Các nghiên cứu trước đây tập trung vào các nhóm gen đã biết có mối liên quan với ung thư vú như *BRCA1* và *BRCA2* và dựa trên dữ liệu giải trình tự ADN.[5] Mặt khác, dữ liệu biến thể gen của người Việt đã có từ cơ sở dữ liệu VN1K, việc khảo sát các biến thể gen ở các bệnh nhân ung thư vú là rất giá trị nhằm xác định và phân loại các biến thể gen liên quan tới ung thư vú, phục vụ cho nghiên cứu và ứng dụng sàng lọc nguy cơ bệnh tật trong quần thể người Việt. Vì các lý do như trên, chúng tôi tiến hành khảo sát các biến thể gen liên quan đến ung thư vú bằng dữ liệu giải trình tự ARN.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Nghiên cứu được thực hiện trên 5 người bệnh ung thư vú khám và điều trị tại bệnh viện Đại Học Y Hà Nội và 8 đối chứng.

Nhóm bệnh: Những bệnh nhân được chẩn đoán xác định ung thư vú bằng kết quả xét nghiệm giải phẫu bệnh, chẩn đoán giai đoạn bệnh từ I - II theo phân loại TNM, không mắc các ung thư khác, và đồng ý tham gia nghiên cứu.

Nhóm chứng: Những người không có tiền sử mắc ung thư vú hay các bệnh ung thư khác đến khám sức khỏe tại Bệnh viện Đại Học Y Hà Nội.

2. Phương pháp

Địa điểm và thời gian nghiên cứu

Bệnh viện Đại học Y Hà Nội.

Thời gian nghiên cứu

Từ 6/2022 tời 10/2023.

Thiết kế nghiên cứu

Nghiên cứu mô tả cắt ngang.

Cỡ mẫu và cách chọn mẫu

Lấy mẫu thuận tiện.

Thí nghiệm

- Thực hiện tại: Công ty Genestory, Đại học Queensland (Australia), công ty Novogene.

- Tách chiết và giải trình tự

Tất cả các bệnh nhân được lấy 4mL máu tĩnh mạch lưu trữ trong ống chống đông EDTA. Các mẫu khỏe được lấy từ dũ liệu VN1K (https://genome.vinbigdata.org), kí hiệu bắt đầu bằng VN_ma_doi_tuong. Các mẫu bệnh ung thư vú được bắt đầu bằng kí tự S. Cả 2 đều được tách chiết bằng bộ kit tách chiết DNA/RNA Qiagen blood mini kit. Chất lượng của các mẫu RNA sequencing được đánh giá bằng hệ thống Thermo Scientific NanoDrop 2000 và điện di trên gel agarose 1%. Sau đó, các mẫu chất lượng cao (28S/18S > 1,8 và tỷ lệ OD 260/280 > 1,9) được gửi đến Đại học Queensland(Australia) và công ty Novogene để xây dựng thư viện cDNA và giải trình tự RNA. Các mẫu được giải trình tự RNA nếu tỷ số RIN > 7 dựa trên hệ thống Agilent 2.100. Giải trình tự được thực hiện trên nền tảng Illumina HiSeq 2000 theo phương thức pair-end với độ dài đọc 150 bp.

Xử lý dữ liệu giải trình tự RNA

Các lần đọc RNA-seq được xử lý và kiểm

soát chất lượng bằng cách sử dụng các công cụ FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) và Trimmomatic để loại bỏ các trình tự adaptor và đoạn đọc chất lượng thấp. Ngoài ra, điểm phred tối thiểu là 30 và độ dài tối thiểu là 150bp.

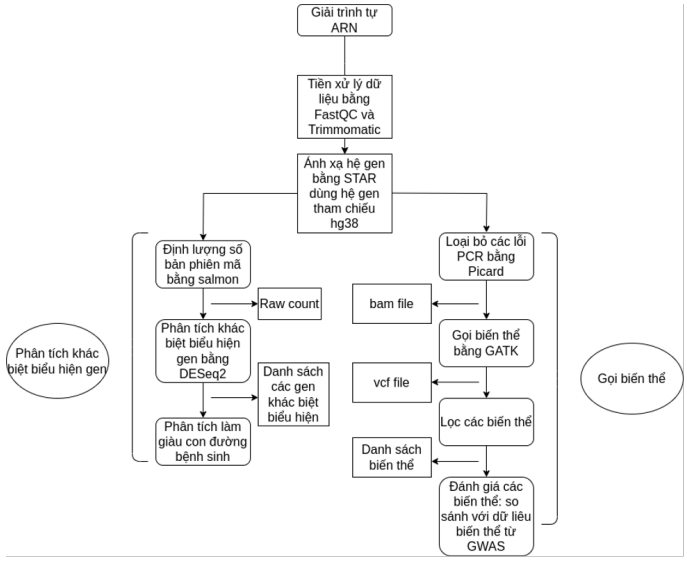
Phân tích dữ liệu giải trình tự RNA

Với quy trình gọi biến thể, các đoạn đọc đạt chất lượng được sử dụng để gọi biến thể trong đó tập trung vào SNP và các đoạn chèn và xóa ngắn (indels), các loại biến thể khác không được phân tích do giới hạn về độ dài đọc. Các lần đọc chất lượng tốt được căn chỉnh theo bộ gen tham chiếu của người Ensembl GRCh38.p14 (GCA_000001405.29) bằng phần mềm Hisat2. Các lần đọc giống hệt nhau (hoặc các bản sao PCR) được căn chỉnh vào cùng một vị trí, được đánh dấu bằng công cụ MarkDuplicates từ Picard (v1.104) (https://picard.sourceforge.net/) và bị bỏ qua trong phân tích downstream. Tiếp đó công cụ GATK (https://gatk.broadinstitute.org) được sử dụng để gọi biến thể. Dữ liệu biến thể sẽ được tập trung vào các gen khác biệt biểu hiện từ quá trình phân tích khác biệt biểu hiện gen. Cuối cùng các biến thể này sẽ được chú thích chức năng bằng công cụ SnpEff (https://pcingola.github.io/SnpEff/) và so sánh với bộ dữ liệu biến thể liên quan ung thư vú của GWAS Catalog (https://www.ebi.ac.uk/gwas/).

Các đoạn đọc được căn chỉnh dựa trên bộ gen tham chiếu hg38 thực hiện bằng công cụ STAR với các tham số mặc định. Phân tích biểu hiện gen khác biệt được thực hiện bằng package R DESeq2. Kết quả sau tính toán là giá trị Log2FoldChange thể hiện cho chênh lệch mức độ biểu hiện gen giữa nhóm bệnh và nhóm chứng. Công thức tính:

Log2 Fold change = Log2(giá trị biểu hiện gen nhóm bệnh) - Log2(giá trị biểu hiện gen nhóm chứng).

Các gen được xác định là biểu hiện khác biệt giữa nhóm bệnh và nhóm chứng nếu có mức chênh lệch mức độ biểu hiện |Log2FoldChange| > 0,5 và có giá trị p-value hiệu chỉnh < 0,05. Các gen có mức chênh lệch biểu hiện |Log2FoldChange| > 2x và p-value hiệu chỉnh < 0,05 được coi là khác biệt đáng kể. Các gen khác biệt đáng kể sẽ được đưa vào để phân tích làm giàu bộ gen để xác định quá trình bệnh sinh tiềm năng.



Hình 1. Quy trình phân tích dữ liệu giải trình tự ARN

3. Đạo đức nghiên cứu

Nghiên cứu đã thông qua Hội đồng đạo đức Trường Đại học Y Hà Nội.

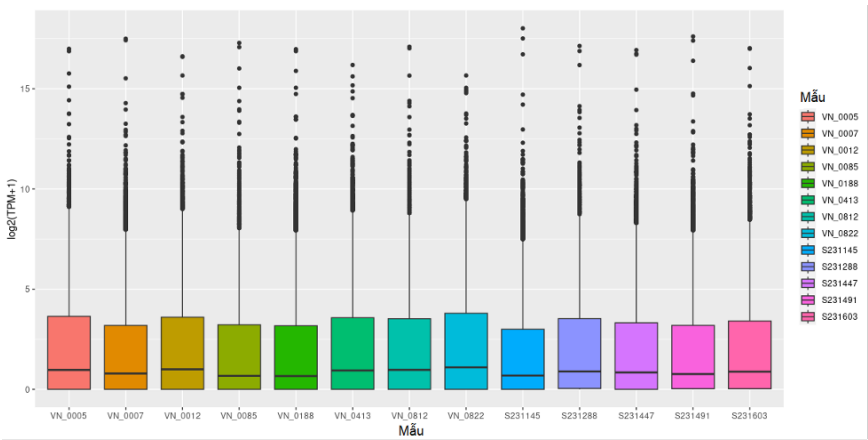
III. KẾT QUẢ

1. Đặc điểm chung của đối tượng nghiên cứu

Những đặc điểm chung như tuổi trung bình, phân bố các nhóm tuổi được phân tích. Độ tuổi giữa hai nhóm bệnh và nhóm chứng là tương đồng. Tuổi trung bình nhóm bệnh là 57 và tuổi trung bình nhóm đối chứng là 54,71. Ở những bệnh nhân ung thư vú, nhóm tuổi mắc bệnh nhiều nhất là nhóm tuổi từ 40 đến 60 tuổi chiếm 60% và không có bệnh nhân nào dưới 40 tuổi. Phân loại theo giải phẫu bệnh, các bệnh nhân đều thuộc thể ung thư biểu mô tuyến. Trong 5 bệnh nhân đưa vào nghiên cứu, có 4 bệnh nhân thuộc giai đoạn II, 1 bệnh nhân thuộc giai đoạn III theo TNM. Không có bệnh nhân nào có xâm nhập mạch thần kinh. Nhuộm hóa mô miễn dịch cho thấy: 1 mẫu dương tính với ER, 2 mẫu vơi PR và 2 mẫu với HER2. Chỉ số Ki67 cho thấy 4 trên 5 bệnh nhân có mức độ phân chia tế bào mạnh.

2. Giải trình tự RNA và ánh xạ hệ gen

Khoảng 1063 triệu lượt đọc ghép cặp được tạo ra từ giải trình tự RNA của 5 mẫu máu của bệnh nhân ung thư vú và 8 mẫu máu người khỏe mạnh, trung bình xấp xỉ 82 triệu lần đọc ghép cặp trên mỗi mẫu. Tỷ lệ phần trăm trung bình của số lần đọc duy nhất là 47,72%.

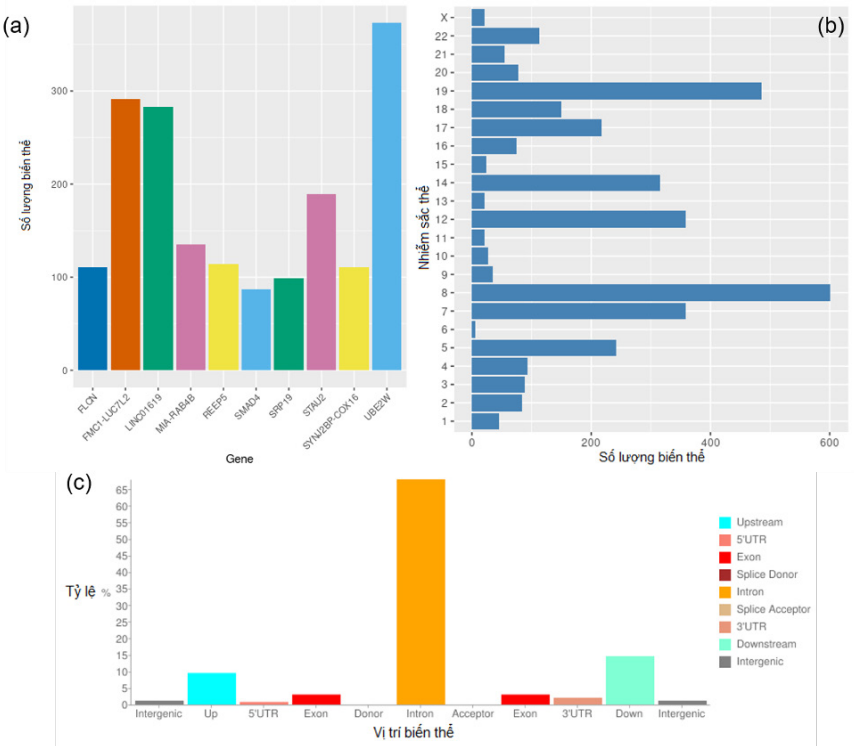


Hình 2. Dữ liệu biểu hiện gen sau khi được chuẩn hóa.

Dữ liệu biểu hiện gen của các mẫu bệnh và chứng sau chuẩn hóa bằng phương pháp tính TPM (transcript per million), sau đó cân bằng bằng log2 (TPM+1)

Chuẩn hóa dữ liệu RNA-seq cần thiết để loại bỏ các khác biệt về điều kiện thí nghiệm như kích thước thư viện và độ sâu giải trình tự, từ đó cho phép phát hiện chính xác sự khác biệt sinh học giữa các mẫu. **Hình 2** thể hiện dữ liệu biểu hiện gen của các mẫu bệnh và chứng sau chuẩn hóa bằng phương pháp tính TPM (transcript per million), sau đó cân bằng bằng log2(TPM+1). Đối với mỗi mẫu ta dựng một biểu đồ hộp chứa các giá trị biểu hiện gen của mẫu đó, các chấm biểu thị cho mức độ biểu hiện của một gen (được đo bằng cách tính log2(TPM+1), như ta thấy thì trung vị và IQR (Interquartile range) của các mẫu gần giống nhau. Do vậy, dữ liệu biểu hiện gen sau khi chuẩn hóa phân phối khá tương đồng, điều đó có nghĩa chất lượng các mẫu tương đối đồng đều và phương pháp chuẩn hóa lựa chọn là phù hợp, dữ liệu đầu ra sau bước này đảm bảo cho các phân tích biểu hiện gen tiếp theo.

3. Gọi biến thể và chú thích chức năng



Hình 3. Phân bố của các biến thể trong hệ gen.

Các gen có số lượng biến thể nhiều nhất được thể hiện trong hình (a); hình (b) liệt kê biến thể theo từng nhiễm sắc thể (NST); biến thể cũng được chú thích theo từng vùng gen như hình (c)

Việc khai thác và phân tích toàn bộ các biến thể từ hệ gen của người bệnh là tương đối khó khăn và kém hiệu quả do hạn chế về cỡ mẫu khiến sự so sánh giữa mẫu bệnh và mẫu người khoẻ mạnh ít có ý nghĩa thống kê. Do vậy, chúng tôi tập trung vào tìm hiểu các biến thể gen nằm trên các gen có sự khác biệt biểu hiện giữa người bệnh và người khoẻ mạnh. Quá trình phân tích biểu hiện gen cho thấy 143 gen khác biệt biểu hiện nhất, là những gen tiềm năng mang những biến thể liên quan tới ung thư vú. Từ đó, chúng tôi thực hiện quá trình gọi biến thể trên những gen này ở các mẫu bệnh. Tổng cộng 3515 biến thể trong 143 gen được xác định, gồm 1568 biến thể mới phát hiện và 1947 biến thể đã được ghi nhân trong cơ sở dữ liệu dbSNP. Trong 3515 biến thể có 2837 SNP, 327 thêm đoạn và 349 mất đoạn. Số lượng biến thể theo từng gen riêng biệt, các gen có số lượng biến thể nhiều nhất được thể hiện trong hình (a). Kết quả quá trình gọi biến thể được liệt kê trong bảng theo từng NST, như ta thấy trong hình (b) các NST chứa nhiều biến thể nhất là NST số 8 và 19, các NST có ít biến thể nhất là NST số 6, 11, và NST giới tính X. Các biến thể cũng được chú thích theo từng vùng gen, các vùng intron chứa nhiều biến thể nhất chiếm tới 68% số biến thể phát hiện, điều này liên quan tới độ dài lớn của vùng này, các biến thể ở vùng này sẽ ít ảnh hưởng tới chức năng của gen hơn nên ít bị chọn lọc. Các biến thể quan trọng như exon và vùng cắt nối xuất hiện rất ít chỉ chiếm 3% và 0,3% quan sát trong hình (c).

Các biến thể có tần số xuất hiện cao rất đáng quan tâm, vì đây là các biến thể tiềm năng có liên quan với kiểu hình ung thư vú. Thống kê trong 5 mẫu người bệnh có 921 biến thể xuất hiện ở ít nhất 2 mẫu (AF > 0,2). Các biến thể xuất hiện với tần số (AF) cao nhất ở nhóm bệnh được liệt kê ở bảng 1.

Bảng 1. Các biến thể gen xuất hiện nhiều nhất ở nhóm ung thư vú

NST	Vị trí	rs ID*	Trình tự	Loại biến thể	Mức tác động	Gen	Tần số nhóm bệnh	Tần số nhóm chứng
10	63465371	*rs2393979*	A → G	Đột biến im lặng	Trung tính	*JMJD1C*	1	0,375
10	63465485	*rs10761770*	A → G	Đột biến im lặng	Trung tính	*JMJD1C*	1	0,1875
11	16738697	*rs1846936*	T → G	Đột biến im lặng	Trung tính	*C11orf58*	1	0,9375
11	16739562	*rs3802963*	C → G	Đột biến im lặng	Trung tính	*C11orf58*	1	0,1875
14	23310425	*rs1535094*	C → G	Đột biến im lặng	Trung tính	*BCL2L2*	1	0,25

Tiếp đó, chúng tôi cũng thực hiện dự đoán chức năng của biến thể. Các biến thể có mức tác động cao được liệt kê ở bảng 2.

Bảng 2. Các biến thể gen có nguy cơ cao được phát hiện

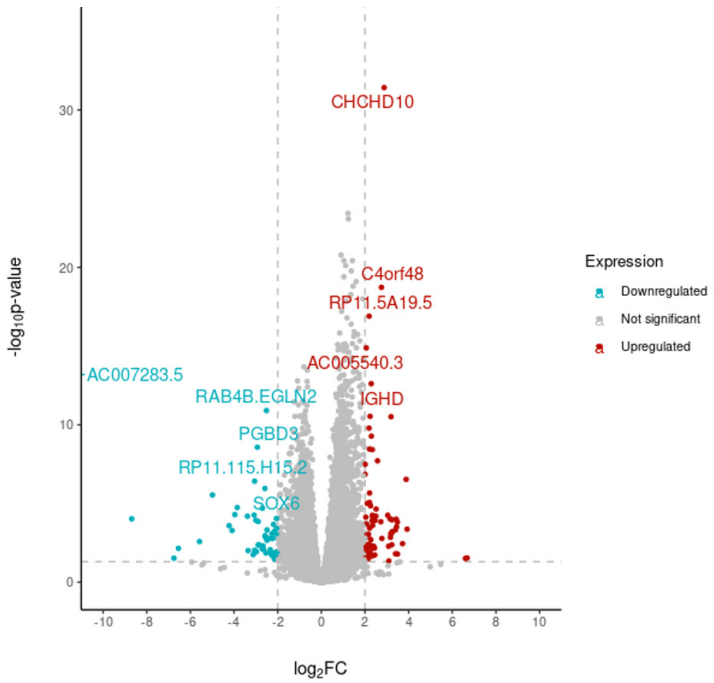
NST	Vị trí	rs ID*	Trình tự	Loại Biến thể	Mức độ tác động	Gen	Tần số nhóm bệnh	Tần số nhóm chứng
12	51822274	.	GC → G	Dịch khung	CAO	*FIGNL2*	0,1	0,4375
17	4900416	*rs35400274*	G→ A	Đột biến vô nghĩa	CAO	*C17orf107*	0,3	0,25
17	29593481	.	C → CT	Dịch khung	CAO	*ANKRD13B*	0,1	0,0625
19	7632999	*rs34406374; rs78473084*	TGA→ T	Dịch khung	CAO	*PCP2*	0,2	0,25
19	45768497	.	CG →	Dịch khung	CAO	SIX5	0,1	0,375
2	86976205	.	T→ A	Đột biến vô nghĩa	CAO	*RGPD1*	0,1	0,0625

NST	Vị trí	rs ID*	Trình tự	Loại Biến thể	Mức độ tác động	Gen	Tần số nhóm bệnh	Tần số nhóm chứng
20	34999055	.	G → GA	Dịch khung	CAO	*MYH7B*	0,1	0,5625
22	20241951	.	CG→ C	Dịch khung	CAO	*RTN4R*	0,1	0,0625

** các biến thể chưa được báo cáo trong cơ sở dbSNP đánh dấu "."*

Cuối cùng, chúng tôi sử dụng cơ sở dữ liệu các biến thể gen liên quan tới ung thư vú của GWAS-Catalog. So sánh các biến thể đã được báo cáo trên GWAS với kết quả gọi biến thể, chúng tôi xác định biến thể *rs7937-G* nằm trong vùng 3'UTR gen RAB4B có liên quan tới ung thư vú và tần số xuất hiện rất cao AF = 0,7.

4. Phân tích khác biệt biểu hiện gen

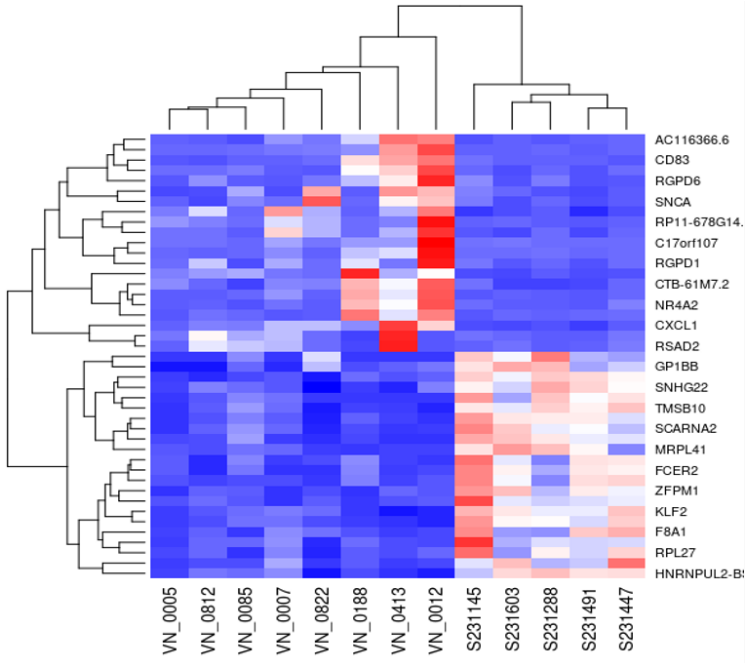


Hình 4. Biểu đồ núi lửa thể hiện các gen biểu hiện khác biệt giữa 2 nhóm.

Các đốm đỏ biểu thị các gen được điều chỉnh tăng Log2FoldChange > 2; các đốm xanh tượng trưng cho các gen được điều hòa giảm; Log2FoldChange < -2; đốm xám thể hiện gen có biểu hiện |Log 2 (Fold Change)| < 2 và p value > 0,05. Các gen được gắn nhãn là các gen có giá trị p value nhỏ nhất (có ý nghĩa nhất) trong từng nhóm và |Log 2 (Fold Change)| > 2

Chúng tôi thu được trung bình 82 triệu lần đọc với độ dài đọc là 150 (bp), trung bình 89,7% các lần đọc được ánh xạ tới bộ gen tham chiếu (hg38). Phần lớn của các gen biểu hiện khác nhau được xác định là protein gen mã hóa. Trong số 18.168 gen, có tổng số lần đọc lớn hơn 50, có 4817 gen biểu hiện khác biệt giữa bệnh nhân ung thư vú và người bình thường

với giá trị p-value hiệu chỉnh < 0,05, trong đó có 2314 gen biểu hiện tăng và 2503 gen biểu hiện giảm. Chúng tôi thu được 143 gen khác biệt biểu hiện đáng kể (Log2FoldChange > 2x) bao gồm 78 gen tăng biểu hiện và 65 gen giảm biểu hiện. Các gen được điều chỉnh tăng nhiều nhất ở bệnh nhân ung thư vú bao gồm: *CHCHD10* (Log2FoldChange:2,88, p-value: 3,8e-18), *C4orf48* (Log2FoldChange:2,75, p-value: 1.,85e-5), *RP11.5A19.5* (Log2FoldChange: 2,19, p-value: 1,26e-3). Các gen bị giảm biểu hiện nhất ở nhóm ung thư vú là: *AC007283.5* (Log2FoldChange: -23,12, p-value: 9,46e-13), *RAB4B-EGLN2* (Log2FoldChange: -2,51, p-value: 1,01e-3), *PGBD3* (Log2FoldChange: -2,93, p-value: 2e-5).



Hình 5. Biểu đồ nhiệt thể hiện mức độ biểu hiện gen theo từng mẫu.

Hàng ngang thể hiện các gen thay đổi biểu hiện, hàng dọc biểu thị cho từng mẫu (các mẫu VN là các mẫu khỏe mạnh, các mẫu S là mẫu ung thư vú)

Để hiển thị và so sánh trực quan hơn sự khác biệt biểu hiện gen giữa nhóm bệnh nhân ung thư vú và người khỏe mạnh, chúng tôi vẽ biểu đồ nhiệt bằng phần mềm R và chọn các gen có khác biệt biểu hiện nhất giữa 2 nhóm để thực hiện phân cụm.

5. Phân tích chức năng gen và làm giàu con đường bệnh sinh

Các gen khác biệt biểu hiện tiếp tục được lọc với các điều kiện là LFC > 2, cuối cùng chúng tôi thu được 143 gen khác biệt biểu hiện nhất bao gồm 78 gen tăng biểu hiện và 65 gen giảm biểu hiện, các gen này tiếp tục được đưa vào chú thích chức năng Gene ontology và phân tích làm giàu con đường bệnh sinh GSEA (gene set enrichment analysis).

Các quá trình sinh học Gene Ontology và nghiên cứu làm giàu con đường bệnh sinh cho các gen khác biệt biểu hiện giữa nhóm bệnh và nhóm chứng được thực hiện.

HALLMARK: HALLMARK_WNT_BETA_CATENIN_SIGNALING được làm giàu với các gen tăng biểu hiện và HALLMARK_HEME_METABOLISM làm giàu với các gen giảm biểu hiện.

IV. BÀN LUẬN

Trong nghiên cứu này, chúng tôi đã nghiên cứu sự khác biệt biểu hiện gen giữa nhóm ung thư vú và nhóm bệnh nhân bình thường từ dữ liệu RNA seq, kết quả phân tích cho thấy có 4817 gen khác biệt biểu hiện giữa hai nhóm, sau đó chọn ra 143 gen có Log2FoldChange > 2 là các gen có sự khác biệt đáng kể. Chúng tôi phát hiện được 3515 biến thể nằm trong 143 gen khác biệt biểu hiện đáng kể. Trong đó có biến thể gen rs7937 nằm trên gen *RAB4B* đã được báo cáo liên quan tới ung thư vú trong cơ sở dữ liệu GWAS Catalog.[6]

Trong các nghiên cứu trước đây, *ABHD14A. ACY1* có liên quan đến chuyển hóa lipid. Chuyển hoá lipid bị thay đổi thường được quan sát ở bệnh nhân ung thư vú và gen này có thể góp phần vào những thay đổi trao đổi chất. LYPD2 biểu hiện cao ở bệnh nhân ung thư vú với LFC là 3,88 và p-value 9,73e-06. *LYPD2* được biết đến liên quan tới sự bám dính của tế bào, điều này rất quan trọng trong quá trình phát triển và di căn của ung thư. Sự biểu hiện quá mức của gen này có thể đóng một vai trò trong việc tăng cường di căn và xâm lấn của các tế bào ung thư vú. Một gen được tăng biểu hiện khác là *PAK6*, với Log2FoldChange là 3,48 và p-value là 0,0015, *PAK6* là một kinase đóng vai trò trong các con đường truyền tín hiệu tế bào, có thể góp phần vào sự tăng sinh và xâm lấn tế bào không kiểm soát được, những đặc điểm nổi bật của bệnh ung thư.

Trong khi một số gen biểu hiện sự biểu hiện gia tăng, một số gen khác lại giảm bộc lộ đáng kể ở bệnh nhân ung thư vú. *RP11.343C2.12* giảm biểu hiện đáng kể, đây là một gen nằm trên NST số 16, chức năng của gen này chưa được nghiên cứu rõ, nhưng có những báo cáo cho thấy gen này liên quan đến chuyển hoá lipid và các tính trạng liên quan.

Sự phân bố các SNP tập trung ở các NST số 8 và số 19, điều này gợi ý mối liên quan giữa ung thư vú và các gen nằm trên NST này. Thể đơn bội NST số 8 đã được báo cáo thấy ở 2/3 số tế bào ung thư biểu mô tuyến vú thể ống xâm nhập.[7] Giả thuyết đặt ra các gen trên NST số 8 đóng vai trò ức chế khối u, các biến thể gen làm giảm chức năng của chúng tạo thuận lợi cho sự tăng sinh không kiểm soát của tế bào.

Có một biến thể đáng chú ý là c.456G>A (*rs35400274*) tạo thành bộ ba kết thúc sớm nằm trên gen *C17orf107*, biến thể này làm giảm mức độ biểu hiện gen ở nhóm ung thư vú so với nhóm chứng Log2FoldChange = -2,49 và p-value =0,002, biến thể này cũng đã được báo cáo có mặt ở bệnh nhân ung thư đại trực tràng trong cơ sở dữ liệu COSMIC.[8] Biến thể có mức tác động cao khác là *rs78473084* mã hoá cho gen Purkinje cell protein 2, các nghiên cứu trước đây cũng chưa xác định chức năng của SNP này. *rs7937* là biến thể duy nhất được báo cáo trong cơ sở dữ liệu GWAS-catalog là liên quan tới chuyển hoá letrozole ở các bệnh nhân ung thư vú với tương quan trên toàn hệ gen (p value = 5,26e-10). *rs7937* có liên quan đến nồng độ letrozole ngay cả sau khi điều chỉnh theo độ tuổi và BMI do ảnh hưởng tới hoạt tính chuyển hóa CYP2A6. Biến thể *rs7937* cũng đã được nghiên cứu chỉ ra có liên quan tới một số bệnh lý khác như COPD, một nghiên cứu trên tạp chí Oxford Academic năm 2018 đã chỉ ra *rs7937* ảnh hưởng tới trình methyl hóa DNA hệ gen của các tế bào máu.[9]

Ở người bệnh ung thư vú, nguyên nhân chính gây ung thư là do đột biến mới phát sinh trong tế bào sinh dưỡng (soma), vì vậy sẽ là lý tưởng nếu thực hiện giải trình tự ARN mẫu mô của người bệnh ung thư vú, ngoài ra việc thiết kế thí nghiệm cần mở rộng về số lượng mẫu bệnh phẩm được phân tích, với các mẫu lặp về sinh học và mẫu lặp về kĩ thuật, điều đó

loại bỏ các sai lệch và độ nhiễu do quá trình thí nghiệm và chọn mẫu. Việc hạn chế về cỡ mẫu cũng khiến nghiên cứu các biến thể hiếm MAF < 0,01 trở nên khó khăn vì không thể so sánh phù hợp. Do đó, trong nghiên cứu này, phần lớn SNP hiếm không thể được nghiên cứu đầy đủ. Các nghiên cứu độc lập sâu hơn với cỡ mẫu lớn hơn sẽ giúp xác nhận những phát hiện từ nghiên cứu này.

V. KẾT LUẬN

Tóm lại, nghiên cứu của chúng tôi đã xác định được một số biến thể gen tiềm năng liên quan tới ung thư vú ở nhóm đối tượng nghiên cứu. Các biến thể gen xuất hiện nhiều nhất ở nhóm bệnh là *rs2393979, rs10761770, rs1846936, rs3802963, rs1535094*. 8 biến thể nguy cơ cao được phát hiện, trong đó có 2 biến thể đã được biết trong cơ sở dữ liệu dbSNP là *rs35400274* và *rs34406374*. Một biến thể đáng chú ý là c.456G>A (*rs34406374*) tạo thành bộ ba kết thúc sớm nằm trên gen C17orf107, biến thể này làm giảm mức độ biểu hiện gen ở nhóm ung thư vú so với nhóm chứng Log2FoldChange = -2,49 và p-value =0,002, biến thể này cũng đã được báo cáo có mặt ở bệnh nhân ung thư đại trực tràng trong cơ sở dữ liệu COSMIC. Có 1 biến thể đã được chứng minh liên quan tới nồng độ letrozole (một thuốc điều trị nội tiết) trong dữ liệu GWAS Catalog là *rs7937*. Ngoài ra chúng tôi còn phát hiện được một số gen khác biệt biểu hiện ở nhóm bệnh nhân ung thư vú bao gồm: *CHCHD10, C4orf48, RP11.5A19.5, AC007283.5, RAB4B-EGLN2, PGBD3*, điều này rất hữu ích cho các nghiên cứu sau này nhằm phát hiện và phát triển các dấu ấn phục vụ chẩn đoán và điều trị ung thư vú. Kết quả từ nghiên cứu này cung cấp thêm những hiểu biết sâu sắc về con đường bệnh sinh và sự tác động của các gen mang biến thể với nguy cơ ung thư vú ở phụ nữ.

LỜI CẢM ƠN

Bài báo được thực hiện trong khuôn khổ đề tài nghiên cứu Giải pháp đánh giá nguy cơ gây bệnh dựa trên dữ liệu bộ gen người Việt, mã số VINIF.2020.DA.02 do Quỹ VINIF tài trợ.

TÀI LIỆU THAM KHẢO

1. Jenkins C, Ha DT, Lan VT, et al. Breast Cancer messaging in Vietnam: an online media content analysis. *BMC Public Health*. 2020; 20:966. doi:10.1186/s12889-020-09092-8.

2. Ellisen LW, Haber DA. Hereditary breast cancer. *Annu Rev Med*. 1998; 49: 425-436. doi:10.1146/annurev.med.49.1.425.

3. de Nóbrega M, Cilião HL, de Souza MF, et al. Association of polymorphisms of PTEN, AKT1, PI3K, AR, and AMACR genes in patients with prostate cancer. *Genet Mol Biol*. 2020; 43(3): e20180329. doi:10.1590/1678-4685-GMB-2018-0329.

4. Bakhtiarizadeh MR, Alamouti AA. RNA-Seq based genetic variant discovery provides new insights into controlling fat deposition in the tail of sheep. *Sci Rep*. 2020; 10(1): 13525. doi:10.1038/s41598-020-70527-8.

5. Le TNN, Tran VK, Nguyen TT, et al. BRCA1/2 Mutations in Vietnamese Patients with Hereditary Breast and Ovarian Cancer Syndrome. *Genes*. 2022; 13(2): 268. doi:10.3390/genes13020268.

6. Hertz DL, Douglas JA, Kidwell KM, et al. Genome-wide association study of letrozole plasma concentrations identifies non-exonic variants that may affect CYP2A6 metabolic activity. *Pharmacogenet Genomics*. 2021; 31(5): 116-123. doi:10.1097/FPC.0000000000000429.

7. GARCÍA PARRA-PÉREZ FA, ZAVALA-POMPA A, PACHECO-CALLEROS J, et al. Monosomy of chromosome 8 could be

considered as a primary preneoplastic event in breast cancer: A preliminary study. *Oncol Lett*. 2012; 3(2): 445-449. doi:10.3892/ol.2011.484.

8. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019; 47(D1):D941-D947. doi:10.1093/nar/gky1015.

9. Nedeljkovic I, Lahousse L, Carnero-Montoro E, et al. COPD GWAS variant at 19q13.2 in relation with DNA methylation and gene expression. *Hum Mol Genet*. 2018; 27(2): 396-405. doi:10.1093/hmg/ddx390.

Summary

INVESTING GENETIC VARIATIONS ASSOCIATED WITH BREAST CANCER USING RNA SEQUENCING DATA

This study aims to investigate gene genetic variants associated with breast cancer using RNA sequencing (RNA-seq) data. We conducted a study with 5 breast cancer (BC) patients and 8 controls from 1000 Vietnamese genomes project (VN1K) data. Over 5 women with breast cancer and 8 healthy controls of similar age. We apply a cross-sectional method to understand germline variants present in breast cancer patients through RNA sequencing of blood samples. By analyzing the RNA-seq data, we identified 143 genes with significant expression differences between the BC group and the healthy controls. Next, we observed a total of 3515 variants located on such genes. Among them, 8 variants have been reported to probably increase BC risk, 2 variants were included in the dbSNP, *rs35400274* and *rs34406374*. *rs35400274* forms a premature termination triplet located on the C17orf107 gene, reducing gene expression levels in the BC group, Log2FoldChange = -2,49 and p-value = 0,002, this variant has also been reported to be present in colorectal cancer patients in the COSMIC database. Meanwhile, *rs7937* is associated with breast cancer traits in the GWAS-Catalog database due to its influence on Letrozole drug concentrations. Results from this study provide additional insights into the pathogenesis and impact of genetic variants on breast cancer risk in women.

Keywords: Gene variation, breast cancer, RNA sequencing.