REVIEW AND COMPARISON OF HLASCAN AND HLAMINER FOR HLA ANALYSIS OF NGS DATA

Tran Thi Bich Ngoc[1], Vu Phuong Nhung[1], Ma Thi Huyen Thuong[1], Nguyen Hai Ha[1,2], Nguyen Dang Ton[1,2*]
[1]Institute of Genome Research – VAST, [2]Graduate University of Science and Technology - VAST

ARTICLE INFO

Received: 14/10/2022
Revised: 04/11/2022
Published: 24/11/2022

KEYWORDS

HLA
HLAminer
HLAscan
WES
IMGT/HLA

ABSTRACT

Human leukocyte antigen (HLA), an essential factor for a successful organ transplant, is related to many human diseases. The recent advances in next-generation sequencing technologies (NGS) have led to increased bandwidth and throughput as well as reduced cost for genome sequencing. Several HLA typing algorithms and assays have been developed to take advantage of this new development. However, due to the variety and polymorphisms of HLA loci, HLA typing is still a challenge, even with NGS data. For practical usage of this technology in both research and clinical environments, it is necessary to investigate which software tool could provide better HLA typing results from available sequencing data. We have therefore conducted experiments to assess two HLA typing tools, HLAscan and HLAminer, on whole exome sequencing data (WES). HLAscan predicted 21 HLA genotypes with better accuracy and performance than HLAminer. This result can provide directions for future HLA typing studies on WES data.

GIỚI THIỆU VÀ SO SÁNH CÔNG CỤ HLASCAN VÀ HLAMINER CHO PHÂN TÍCH HLA TỪ DỮ LIỆU GIẢI TRÌNH TỰ GEN THẾ HỆ MỚI

Trần Thị Bích Ngọc[1], Vũ Phương Nhung[1], Ma Thị Huyền Thương[1] Nguyễn Hải Hà[1,2], Nguyễn Đăng Tôn[1,2*]
[1]Viện Nghiên cứu hệ gen - Viện Hàn lâm Khoa học và Công nghệ Việt Nam
[2]Học Viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

THÔNG TIN BÀI BÁO

Ngày nhận bài: 14/10/2022
Ngày hoàn thiện: 04/11/2022
Ngày đăng: 24/11/2022

TỪ KHÓA

HLA
HLAminer
HLAscan
WES
IMGT/HLA

TÓM TẮT

Hệ thống phức hợp kháng nguyên bạch cầu người (HLA) đóng vai trò quan trọng quyết định kết quả của việc cấy ghép nội tạng và có liên quan đến nhiều bệnh ở người. Với các tiến bộ gần đây trong công nghệ giải trình tự thế hệ mới (NGS) thì độ chính xác và thông lượng của việc giải trình tự đã tăng lên và chi phí cũng giảm đi. Nhiều thuật toán và assay để định kiểu HLA đã được phát triển để tận dụng tiến bộ công nghệ mới này. Tuy nhiên, do tính đa dạng và đa hình của các locus HLA, việc định kiểu HLA là một thách thức ngay cả với dữ liệu NGS. Để sử dụng công nghệ này trong thực tế, ở môi trường nghiên cứu lẫn lâm sàng cần phải khảo sát để xác định công cụ phần mềm nào cho kết quả định kiểu HLA tốt hơn trên dữ liệu giải trình tự đã có. Chúng tôi thực hiện thử nghiệm và đánh giá 02 phần mềm định kiểu HLA là HLAscan và HLAminer trên dữ liệu toàn bộ hệ gen biểu hiện (WES). HLAscan dự đoán được 21 kiểu gen HLA khác nhau, cho độ chính xác và hiệu quả hơn HLAminer. Kết quả này giúp định hướng cho các nghiên cứu sử dụng phân loại HLA trên dữ liệu WES.

DOI: https://doi.org/10.34238/tnu-jst.6670

* Corresponding author. *Email: dtnguyen@igr.ac.vn*

1. Giới thiệu

Hệ thống phức hợp kháng nguyên bạch cầu người (HLA- Human Leukocyte Antigen) do một nhóm gen mã hóa các protein trình diện kháng nguyên trên bề mặt hầu hết các tế bào trong cơ thể, tạo ra các peptide kháng nguyên đối với thụ thể tế bào T trên tế bào lympho T. Các gen này chịu trách nhiệm chính trong cảm ứng, đáp ứng và điều chỉnh các phản ứng miễn dịch, giúp hệ thống miễn dịch của cơ thể phát hiện các kháng nguyên ngoại lai, để từ đó có thể tấn công các tế bào lạ xâm nhập [1]-[3]. HLA là vùng gen đa hình nhất trong bộ gen người, gồm các gen nằm trên cánh ngắn nhiễm sắc thể số 6 (NST6), chia làm 3 lớp chính: lớp I, II, III. HLA lớp I bao gồm các phân nhóm A, B, C, E, F, G có trên bề mặt của tất cả các tế bào có nhân, có chức năng trình diện kháng nguyên nội sinh, hoạt hóa tế bào lympho T độc – TCD8+, trực tiếp tham gia phản ứng tiêu diệt tế bào đích có kháng nguyên lạ đặc hiệu. HLA lớp II hay D gồm DP, DQ, DR, DO, DM, chỉ có trên bề mặt một số tế bào như tế bào B, tế bào đuôi gai, đại thực bào và các tế bào trình diện kháng nguyên khác (APCs). HLA lớp II chủ yếu trình diện kháng nguyên ngoại bào cho TCD4+, sau đó TCD4+ sẽ hoạt hóa TCD8+ để tác động trực tiếp tiêu diệt tế bào lạ, đồng thời hoạt hóa các tế bào B để sản xuất các kháng thể đặc hiệu chống lại virus, vi khuẩn, trong đó có SARS-CoV-2. HLA lớp III nằm xen giữa lớp I và lớp II, chứa các gen sản xuất các bổ thể C4, C2, yếu tố B và TNF∝.

Cấu trúc tên gọi của HLA có trình tự sau:

*Tên gen * Mã số alen : Mã số protein riêng : Mã số DNA tương ứng trong vùng mã hóa : Mã số DNA ngoài vùng mã hóa (có thể có thêm hậu tố).*

Ví dụ: *HLA-A*24:02:01:02L*

Có rất nhiều phương pháp định danh HLA. Trước đây người ta sử dụng phương pháp huyết thanh học cổ điển; nhưng từ khi phương pháp PCR ra đời, có nhiều kỹ thuật phát triển dựa trên PCR và gần đây sử dụng giải trình tự Sanger để định danh HLA. Các phương pháp trên vẫn đang được sử dụng với các mức độ khác nhau, tuy nhiên với số lượng alen HLA lớn khoảng hơn 34000 alen đã biết (số liệu tính đến 6/2022) [4] cùng với việc các alen HLA có trình tự giống nhau cao trên các locus khác nhau khiến việc xác định kiểu gen HLA trở nên khó khăn hơn. Vấn đề này được khắc phục bằng công nghệ giải trình tự thế hệ mới (NGS – Next Generation Sequencing) [5].

Với các bệnh nhân đã có dữ liệu hệ gen/hệ gen biểu hiện (WGS/WES), việc xác định kiểu gen HLA bằng cách tái phân tích dữ liệu NGS sẽ tiết kiệm được thời gian và chi phí thiết kế mẫu thí nghiệm. Hiện tại, NGS đã đứng vững và là phương pháp thường được chọn cho xác định kiểu gen HLA độ phân giải cao, áp dụng trong các phòng thí nghiệm hệ gen miễn dịch cũng như trong nghiên cứu cơ bản và đang mở rộng ra các phòng thí nghiệm y khoa và chẩn đoán [6]. Nhưng do bản chất đa hình cao của hệ thống HLA và việc thiếu trình tự hoàn thiện đã biết của vùng NST6 nên việc xác định kiểu gen cho HLA vẫn còn nhiều thách thức. Để giải quyết vấn đề này, nhiều phần mềm phân tích tin sinh định loại HLA đã ra đời. Chúng tôi thực hiện thử nghiệm và đánh giá 02 phần mềm HLAscan và HLAminer trên dữ liệu WES và đối chiếu trong cơ sở dữ liệu (CSDL) IMGT/HLA để tìm ra phương pháp phân tích HLA tối ưu, làm cơ sở để sử dụng tiếp cho các vấn đề nghiên cứu sau này.

2. Phương pháp

2.1. Đặc điểm chung của các công cụ phân tích HLA từ dữ liệu NGS

Số lượng lớn đoạn đọc từ WGS/WES được đối chiếu vào CSDL IMGT/HLA để tìm kiếm các alen khớp tốt nhất dựa trên thống kê đối chiếu, số lượng đoạn đọc phủ exon và mức độ phủ exon. Để xác định alen HLA, dữ liệu đọc trình tự từ WGS/WES được đối chiếu với toàn bộ CSDL IMGT/HLA, chỉ lưu lại các đoạn đọc có thể bắt cặp tương đồng với alen trong CSDL với số lượng khớp cao mới được sử dụng để phân tích thống kê [5].

Hiện nay, có rất nhiều phần mềm phân tích HLA khác nhau sử dụng dữ liệu từ các nguồn giải trình tự khác nhau. Trong đó, hai phần mềm phân tích HLAscan và HLAminer được áp dụng phổ

biến với dữ liệu giải trình tự DNA từ máy giải trình tự gen thế hệ mới. HLAscan được phát triển vào năm 2017, sử dụng ngôn ngữ Python, đến nay đã được trích dẫn 56 lần ở các tạp chí khác nhau, tốc độ tính toán nhanh hơn HLAminer. HLAminer là phần mềm được xây dựng từ sớm (từ năm 2012), đến nay đã có rất nhiều nghiên cứu đã sử dụng phần mềm này (có tới 149 trích dẫn) [7]. HLAminer sử dụng ngôn ngữ Perl. HLAscan phát hiện được 21 loại HLA (cả lớp I, II và một phần lớp III), còn HLAminer chỉ phát hiện được 16 loại HLA khác nhau ở lớp I và II [7].

2.2. Thực nghiệm

Đối với nghiên cứu này, chúng tôi tiến hành sử dụng dữ liệu giải trình tự gen thế hệ mới của 30 mẫu COVID-19 để làm đầu vào cho hai phần mềm phân tích HLAscan và HLAminer, sử dụng CSDL IGMT/HLA [8]. Kết quả 2 phần mềm này giao nhau ở 11 loại HLA (xem trong bảng 1). Đây là căn cứ để so sánh độ hiệu quả hai phần mềm nêu trên.

Bảng 1. *Bảng so sánh 02 phần mềm HLAscan và HLAminer*

Tiêu chí so sánh		**Phần mềm HLAminer**	**Phần mềm HLAscan**
Loại dữ liệu phân tích	*DNA*		x
	RNA	x	
Dạng file đầu vào	*BAM*		x
	FASTQ		x
Phiên bản được sử dụng ở nghiên cứu này		v1.4	v2.1
Loại HLA	*A*		x
	B		
	C		
	F		
	G		
	H	x	
	E		x
	DPA1		x
	DPA2	x	
	DPB1		x
	DPB2	x	
	DQA1		x
	DQB1		
	DRA		
	DRB1		
	DRB3	x	
	DRB4		
	DRB5		x
	MICA		
	MICB		
	DMA		
	DMB		
	DOB		
	DOA		
	TAP1		
	TAP2		

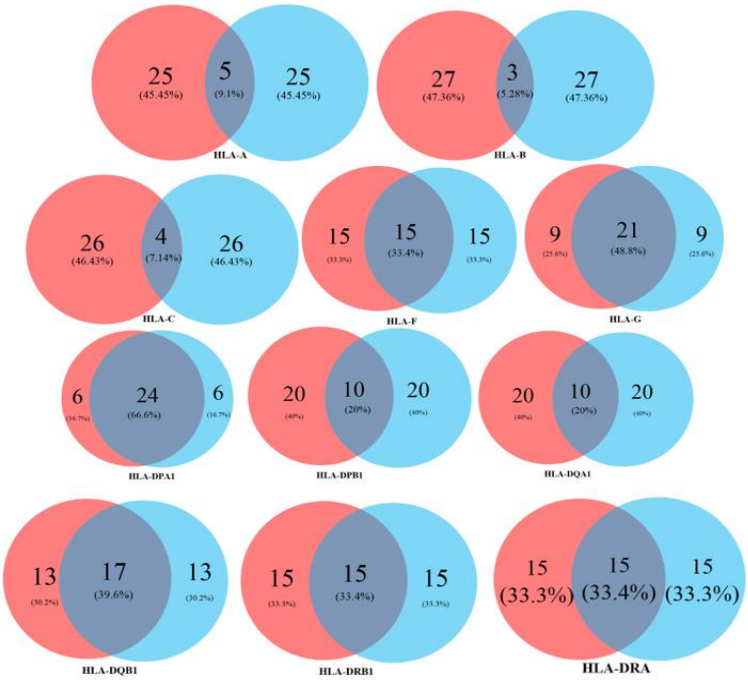
Chúng tôi sử dụng bộ số liệu 30 mẫu WES của bệnh nhân COVID-19 (ở dạng file *FASTQ), kích thước dữ liệu thô của mỗi mẫu dao động từ 16Gb-20Gb/pair-end. Các bước phân tích được thực hiện trên máy tính hiệu năng cao tại Viện Nghiên cứu hệ gen, trên hệ điều hành Linux với 20 core/node, cấu hình Intel(R) Xeon(R) CPU E5-2690 v3, tốc độ 2.60GHz và 128GB Ram.

3. Kết quả và bàn luận

Hai phần mềm HLAscan và HLAminner có ưu điểm sử dụng trực tiếp dữ liệu thô (file *FASTQ) sau khi giải trình tự gen thế hệ mới, tiết kiệm chi phí thiết kế thí nghiệm để định kiểu

HLA hơn các phương pháp thực nghiệm đang được sử dụng hiện nay. HLAscan phát hiện được 21 loại HLA (cả lớp I, II và một phần lớp III), còn HLAminer chỉ phát hiện được 16 loại HLA khác nhau ở lớp I và II.

Kết quả là 2 phần mềm này đưa ra dự đoán chung 11 loại HLA (chi tiết trong bảng 1). Phần mềm HLAminer cần trung bình 5h để phân tích xong một mẫu. Phần mềm HLAscan có thời gian phân tích ngắn hơn (~3h/mẫu). Kết quả dự đoán của HLAminer là hình tròn đỏ bên phải, còn HLAscan là hình tròn xanh bên trái. Từ kết quả ở hình 1 cho thấy, 30 mẫu bệnh nhân COVID-19 khi phân tích bằng HLAminer và HLAscan có độ overlap khoảng từ 5,28 - 66,6%, tập trung cao vào loại HLA-DPA1.



Hình 1. *Kết quả phân tích HLA từ 2 phần mềm HLAminer và HLAscan*

3.1. Phần mềm HLAscan

HLAscan là một phần mềm dựa trên căn chỉnh (*align*) để xác định các dạng haplotype có xét đến phân phối đoạn đọc (*read*). Phần mềm thực hiện mapping các read với trình tự HLA từ CSDL IMGT/HLA. Sự phân bố của các read đã căn chỉnh được sử dụng để tính toán giá trị *hàm cho điểm* nhằm xác định các alen có phân loại chính xác bằng cách loại bỏ dần các alen dương tính giả. HLAscan có thể được tin cậy, ứng dụng xác định loại HLA trên WGS, WES và target sequencing. Ka và cộng sự (năm 2017) nhận thấy rằng, độ sâu (depth) là một yếu tố quan trọng liên quan tới độ chính xác khi định kiểu HLA. Với độ sâu > 90x thì HLAscan cho độ chính xác 100% cho việc sử dụng lâm sàng [9].

Đầu tiên, HLAscan bắt đầu với các read ở định dạng FASTQ để mapping với trình tự tham chiếu từ CSDL IMGT/HLA. Trong suốt quá trình này, các read tương ứng với vùng exon của gen HLA được chọn dựa trên sự liên kết ban đầu tạo ra bằng cách sử dụng công cụ GATK với trình tự tham chiếu hệ gen người (hg19). Bước lọc này không phân loại các read thành các gen HLA.

Quá trình phân tích gồm 5 bước chính:

- Bước 1: Thu thập dữ liệu thô từ giải trình tự gen thế hệ mới.
- Bước 2: Mapping trình tự gen HLA (ví dụ HLA-A) vào trình tự hệ gen tham chiếu hg19.
- Bước 3: Các trình tự HLA (ví dụ HLA-A) được sắp hàng vào các kiểu alen cụ thể. Từ các alen ứng viên, kiểu alen thật sự được xác định bằng *hàm cho điểm* và giải quyết các vấn đề

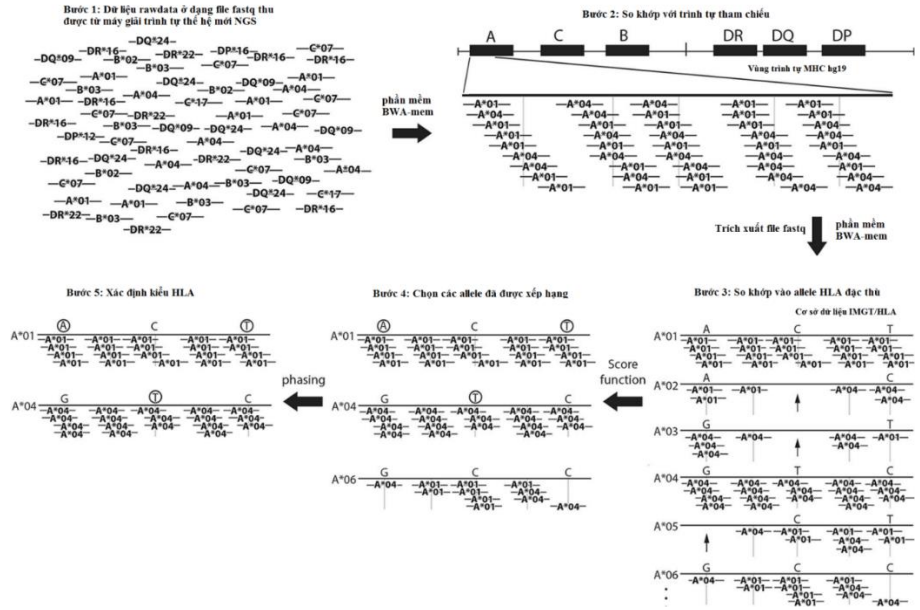
phasing (bước 4 và 5). Các đường thẳng đứng màu xám dưới trình tự tham chiếu biểu diễn các vị trí với biến thiên trình tự. Các mũi tên đen ở các alen A*02, A*03, và A*05 của bước 3 chỉ các vị trí gen không có trình tự nào được sắp hàng.

- Bước 4: Chọn ra các base được xếp hạng.

Các base xếp hạng được khoanh tròn ở hình 2: nucleotide A và T ở A*01 cùng T ở A*04 biểu diễn các trình tự duy nhất không thừa với trình tự base ở bất cứ alen đã được xếp hạng khác.

- Bước 5: Xác định kiểu HLA.

Sau khi chọn ra các base được xếp hạng, tiến hành định kiểu HLA cho mẫu nghiên cứu.



Hình 2. *Quy trình phân tích HLAscan*

Do tính đa hình cao và mỗi gen có nhiều loại alen làm giảm hiệu suất của HLAscan. HLAscan cần giảm thiểu số lượng alen bằng cách loại bỏ alen lỗi qua mỗi bước. Để lọc các alen sai ra khỏi nhóm alen ứng viên ban đầu, HLAscan sử dụng một hàm điểm đánh giá sự phân bố của các đoạn đọc được căn chỉnh trên vùng mục tiêu. *'Read$_i$'* được định nghĩa là tọa độ trên chuỗi đích khớp với tâm của đoạn đọc thứ *i* khi có *n* đoạn đọc ($1 \le i \le n$). Số vị trí liên tiếp trong chuỗi đích không có số đọc là khoảng cách giữa tâm của đoạn đọc liền kề, được xác định là D_j *($1 \le j \le m$)*.

Sau đó, *hàm cho điểm* được tính như sau:

$$\text{Score} = \sum_{j=1}^{m} \left(\frac{D_j}{c}\right)^3$$ trong đó *c* là hằng số.

Hằng số *c* có thể được xác định dựa trên độ sâu trình tự và độ dài của đoạn đọc. Với dữ liệu NGS có độ sâu 60x (với dữ liệu giải trình tự đích) hoặc 30x (với giải trình tự toàn bộ hệ gen) thì *c* thường được đặt thành 30. Giả sử độ dài trung bình của các đoạn đọc là khoảng 150 bp và hằng số *c* được đặt thành 30. Khi độ dài của một đoạn đọc là 150 bp thì khoảng cách giữa tâm của hai đoạn đọc liền kề D_j sẽ là 150 và *hàm cho điểm* là 125. HLAscan sẽ phân loại các alen có điểm trên 125.

Các alen vượt qua *hàm cho điểm* được coi là alen ứng viên. Nhiều alen sai sẽ bị loại bỏ bởi *hàm cho điểm*, việc loại bỏ các alen lặp thường để lại một số hoặc ít hơn các alen ứng viên. Số đoạn đọc trình tự duy nhất trên mỗi alen ứng viên sẽ được đếm lại, vì số lượng trình tự duy nhất trong các alen ứng viên có thể bị đếm sai do sự hiện diện của các alen sai đã bị loại bỏ ở bước trước. Sau đó, các alen ứng viên thứ nhất và thứ hai được xác định dựa trên có số đoạn đọc duy nhất cao hơn. Cuối cùng, hệ thống tạo ra lệnh gọi dị hợp tử nếu hai alen ứng cử viên cuối cùng sở hữu các đoạn đọc được căn chỉnh duy nhất hoặc lệnh gọi đồng hợp tử nếu chỉ một alen sở hữu các đoạn đọc căn chỉnh duy nhất. Một ví dụ được cung cấp trong bước 4 của hình 2. Các alen

A*01, A*04 và A*06 thể hiện sự liên kết với độ bao phủ độ sâu tốt và phân bố đọc tương đối đồng đều. Mặc dù alen A*06 có cách đọc phổ biến với alen A*01 hoặc A*04, nhưng alen A*01 và A*04 đều có cách đọc độc đáo của riêng chúng. Trong trường hợp này, HLAscan sẽ chọn các alen A*01 và A*04 làm loại HLA cuối cùng.

Ví dụ một mẫu đầu ra dự đoán của phần mềm HLAscan được trình bày trong bảng 2.

Bảng 2. *Bảng dữ liệu đầu ra của phần mềm HLAscan*

HLAscan v2.1
Report created
2022. 2. 7. 16:27:52
==========
HLA gene : HLA-A
of considered types : 3182
---------- HLA-Types ----------
[Type 1] 24:07:01 EX3_9.24638_100 EX2_2.7963_0 EX4_19.4094_100 EX5_7.64103_100
[Type 2] 24:07:01 EX3_9.24638_100 EX2_2.7963_0 EX4_19.4094_100 EX5_7.64103_100

3.2. Phần mềm HLAminer

Khác với HLAscan, HLAminner phân tích với nhiều dạng dữ liệu khác nhau WGS/WES, RNA-Seq hay lắp ráp de novo, lắp ráp mục tiêu, đoạn đọc ngắn. Tùy theo cách giải trình tự, cả alignment và short read đều dùng TASR [10]. TASR là công cụ lắp ráp chỉ sử dụng các read khớp với một đoạn trình tự đầu vào để cho ra trình tự thực tế của trình tự đó và các vị trí xung quanh. Khi đầu vào là các biến thể của cùng một locus, nó có thể cho biết biến thể nào tồn tại trong dữ liệu. HLA CDS hoặc trình tự bộ gen từ CSDL IMGT/HLA được đọc bởi TASR, tạo ra một bảng băm của mỗi nhóm 15 nucleotide (k-mers) có thể gặp phải và sử dụng chúng để nhận diện các đoạn đọc cho bước lắp ráp tiếp theo. Phương pháp này áp dụng cho cả dữ liệu DNA và RNA. Target sequencing chỉ lắp ráp cho một đoạn nhất định để tìm biến thể trên đoạn đấy chứ không lắp ráp toàn bộ. Sau khi lắp ráp trình tự mục tiêu thì tạo ra các Contig. Các contig này tiếp tục được BLAST lên CSDL IMGT/HLA và kiểu gen được xác định dựa trên *hàm cho điểm* và xác suất quan sát thấy trong vùng contig. Tuy nhiên, công thức tính *hàm cho điểm* khác với HLAscan. Với sắp hàng trình tự các đoạn ngắn thì dùng trực tiếp các đoạn đọc thô hoặc đưa qua công cụ TASR. Việc sắp hàng vào CSDL IMGT/HLA cho dữ liệu WES/WGS có thể bỏ sót một số gen nhưng tiết kiệm thời gian và tính toán mà không giảm độ chính xác. Đầu ra của phần mềm HLAminner là 16 loại HLA gồm HLA lớp I (A, B, C) và HLA lớp II (DP, DQ, DR). Các HLA tốt nhất của mỗi contig được nhận diện. Mỗi HLA được cho điểm theo các contig chứa nó bao gồm độ phủ và độ dài contig cùng mức độ phù hợp với HLA.

Các contig được xây dựng bằng cách tăng số lượng trình tự HLA, theo dõi các chuỗi HLA tương ứng, từ đó xác định contig tốt nhất. Đối với mỗi HLA giả định, điểm S_{HLA} là tổng số điểm được tính toán cho từng contig được ghép nối. Điểm số mỗi contig phụ thuộc độ sâu, chiều dài và phần trăm trình tự phân lập được (%sequence_identity), sao cho điểm số phản ánh số lượng base được align với một alen HLA cụ thể. Điểm số cho trình tự HLA đã nhận diện được nhân đôi nếu HLA đã cho khớp tốt nhất với contig.

$S_{HLA}=\sum_{contig=1}^{n} Score_{Contig}$= *size*depth*%sequence_identity*

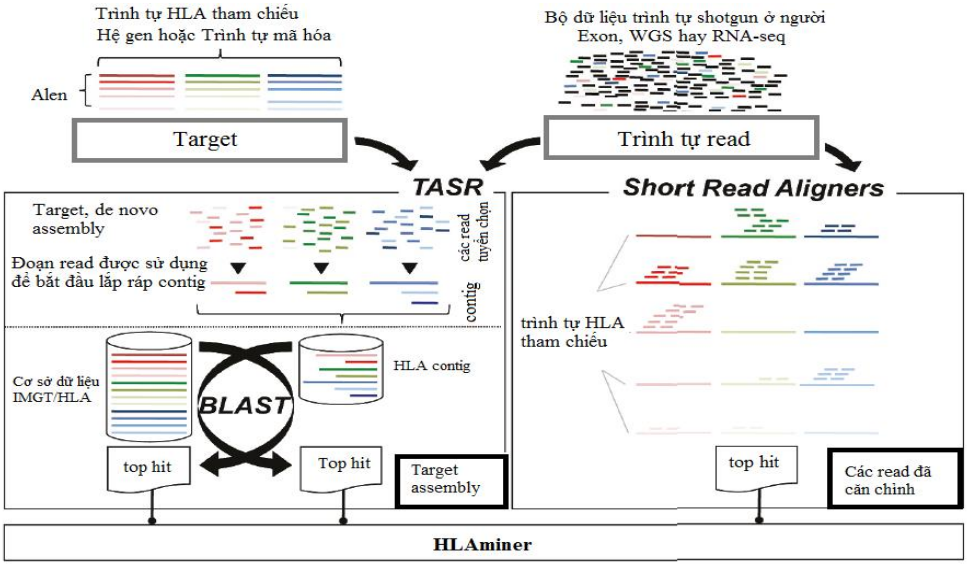
Đối với bất kỳ contig nào, xác suất mô tả một alen HLA duy nhất bằng nhau bằng tỷ lệ nghịch đảo của các chuỗi HLA trong CSDL trình tự.

Vì các contig ngắn hơn có thể không đủ các căn cứ để mô tả bất kỳ type nào một cách rõ ràng, thì xác suất mà một contig đặc trưng cho các HLA loại này hay loại khác:

$P_{Contig1\_is\_HLAx}= \sum P_{HLA}$

Giá trị kỳ vọng (eval) của mỗi HLA, được tính là:

$Eval_{HLA_x}$ =($P_{contig_1\_is\_HLA_x} * P_{HLA_x\_is\_Contig_1}$)*($P_{contig_2\_is\_HLA_x} * P_{HLA_x\_is\_Contig_2}$)*…*$P_{contig_x\_is\_HLA_x} * P_{HLA_x\_is\_Contig_x}$)



Hình 3. *Quy trình phân tích HLAminer*

Dự đoán tính toán cho HLA lớp I từ dữ liệu shotgun sử dụng lắp ráp mục tiêu (hình 3, bên trái) hoặc sắp hàng đoạn đọc ngắn (hình 3, bên phải). Cho lắp ráp mục tiêu, các đoạn đọc NGS có 15 base 5' khớp với một trình tự CDS HLA (RNA-Seq) hoặc trình tự hệ gen (WGS/WES) sẽ được chọn và lắp ráp denovo sử dụng TASR. Các trình tự của contig được sắp hàng vào trình tự CSDL bao gồm tất cả các CDS HLA (RNA-Seq) hoặc trình tự WGS/exon đã được dự đoán, cho ra các HLA khớp tốt nhất. Việc chỉ ra các alen tiềm năng từ dữ liệu shotgun (HLAminer) sử dụng độ dài contig, độ sâu và mức độ tương đồng với trình tự tham chiếu. Xác suất của mỗi dự đoán được ước lượng bằng cách xác định xác suất của dự đoán đó được quan sát ngẫu nhiên.

Độ chính xác của HLAminer dựa trên độ sâu, độ bao phủ, độ dài đọc trình tự và lỗi trình tự. Đầu ra của HLAminer theo thứ tự: Alen, Điểm (*Score*), Giá trị kỳ vọng (*Eval*) và Độ tin cậy *Confidence* = (-10 * log10(Eval)). Ví dụ: *A*24:387,6037.01,1.00e-61,610.0*. Dưới đây là một số lưu ý về kết quả dự đoán của HLAminer ở một số trường hợp đặc biệt (xem trong bảng 3).

Bảng 3. *Mô tả dữ liệu đầu ra của phần mềm HLAminner và một số trường hợp đặc biệt.*

Alen[a]	**Điểm**[b]	**Giá trị kỳ vọng (Eval)**	**Độ tin cậy (-10*log10(Eval))**
HLA-A[c]			
Dự đoán 1 – A*02:01P	64038.03	1.63E-06	57.9
Dự đoán 2 – A*11:01P	5463.99	5.30E-09	82.8
HLA-B			
Dự đoán 1 – B*27:05P	64579.61	2.67E-18	175.7
Dự đoán 2 – B*07:02P	56662.08	6.63E-12	111.8
HLA-C			
Dự đoán 1 – C*07:02P	49419.33	5.23E-08	72.8
Dự đoán 2 – C*02			
C*02:02P[e]	20466.00	6.64E-16	151.8
C*02:21[e]	20466.00	6.64E-16	151.8

a. Loại HLA đã được kiểm chứng bằng PCR.
b. Giá trị điểm các dự đoán alen mã hóa protein được sắp xếp theo điểm giảm dần từ nhiều nhất đến ít khả năng hơn.
c. HLA lớp I và protein alen mã hóa (Confidence (-10 × log10(Eval)) ≥ 20 Score ≥ 1,000) cho mỗi gen.
d. Thứ hạng dự đoán trong điểm tối đa cho mỗi alen HLA được dự đoán.
e. Trường hợp xảy ra 2 hoặc nhiều dự đoán HLA có cùng điểm số (C*02:02P và C*02:21).

Tuy nhiên, khi phân tích HLA sẽ gặp trở ngại khi trong quần thể xảy ra mất cân bằng liên kết (linkage disequilibrium). Ví dụ, sự kết hợp của HLA-A* 01; C*07 và B*08 là phổ biến ở một số quần thể Tây Âu [10]. Ngoài ra, HLAminer chưa dự đoán tốt và đa dạng các kiểu HLA như các phần mềm phổ biến hiện nay.

4. Kết luận

Công nghệ giải trình tự gen thế hệ mới, đặc biệt là WES ngày càng trở nên phổ biến trong nghiên cứu hệ gen người liên quan đến các vấn đề sức khỏe. Việc phát triển và ứng dụng các công cụ phân tích HLA từ dữ liệu WES giúp cho các nghiên cứu đạt hiệu quả cao hơn. Nghiên cứu này đã giới thiệu và thực hiện so sánh ưu nhược điểm của 2 phần mềm phân tích HLA khác nhau và chỉ ra công cụ HLAscan có độ chính xác cao và thời gian phân tích ngắn hơn so với HLAminer. Do đó, chúng tôi sẽ sử dụng HLAscan để phân tích tiếp với những bài toán nhận dạng HLA trên mẫu WES.

Lời cảm ơn

Công trình này được hỗ trợ bởi đề tài "Nghiên cứu đặc điểm hệ gen người Việt Nam liên quan đến lây nhiễm và diễn biến của bệnh nhân Covid-19" - Mã số: ĐTĐLCN.49/20 do Bộ Khoa học và Công nghệ tài trợ kinh phí và được thực hiện tại Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] S. Y. Choo, "The HLA system: genetics, immunology, clinical testing, and clinical implications," *Yonsei medical Journal*, vol. 48, no. 1, pp. 11-23, 2007.

[2] M. Jeanmougin, J. Noirel, C. Coulonges, and J.-F. Zagury, "HLA-check: evaluating HLA data from SNP information," *BMC bioinformatics*, vol. 18, no. 1, pp. 1-8, 2017.

[3] J. Holoshitz, "The quest for better understanding of HLA-disease association: scenes from a road less travelled by," *Discovery medicine*, vol. 16, no. 87, p. 93, 2013.

[4] IPD-IMGT/HLA Database, "About the IPD-IMGT/HLA Database," 2022 [Online]. Available: https://www.ebi.ac.uk/ipd/imgt/hla/about/. [Accessed Jun. 12, 2022].

[5] K. Hosomichi, T. Shiina, A. Tajima, and I. Inoue, "The impact of next-generation sequencing technologies on HLA research," *Journal of human genetics*, vol. 60, no. 11, pp. 665-673, 2015.

[6] K. J. Ingram, H. Merkens, E. F. O'Shields, D. Kiger, and M. D. Gautreaux, "New HLA alleles discovered by next generation sequencing in routine histocompatibility lab work in a medium-volume laboratory," *Human immunology*, vol. 80, no. 7, pp. 465-467, 2019.

[7] P. Liu, M. Yao, Y. Gong, Y. Song, Y. Chen, Y. Ye, X. Liu, F. Li, H. Dong, and R. Meng, "Benchmarking the Human Leukocyte Antigen Typing Performance of Three Assays and Seven Next-Generation Sequencing-Based Algorithms," *Frontiers in Immunology*, vol. 12, p. 652258, 2021.

[8] J. Robinson, D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek, and S. G. Marsh, "Ipd-imgt/hla database," *Nucleic acids research*, vol. 48, no. D1, pp. D948-D955, 2020.

[9] S. Ka, S. Lee, J. Hong, Y. Cho, J. Sung, H.-N. Kim, H.-L. Kim, and J. Jung, "HLAscan: genotyping of the HLA region using next-generation sequencing data," *BMC bioinformatics*, vol. 18, no. 1, pp. 1-11, 2017.

[10] R. L. Warren, G. Choe, D. J. Freeman, M. Castellarin, S. Munro, R. Moore, and R. A. Holt, "Derivation of HLA types from shotgun sequence datasets," *Genome medicine*, vol. 4, no. 12, pp. 1-8, 2012.