NGHIÊN CỨU ĐA DẠNG DI TRUYỀN CỦA *Pseudomonas aeruginosa* TẠI VIỆT NAM THÔNG QUA PHÂN TÍCH PAN GENOME, CORE GENOME MLST VÀ CORE GENOME SNP

Vương Thị Hương[1], Nguyễn Tiến Đạt[1], Trịnh Thị Xuân[2], Nguyễn Cường[3*]

[1]*Công ty TNHH LOBI Việt Nam*

[2]*Khoa Công nghệ thông tin, Đại học Mở Hà Nội*

[3]*Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

TÓM TẮT

Pseudomonas aeruginosa, một vi khuẩn có khả năng thích nghi cao, là một trong các nguyên nhân chính gây nhiễm trùng bệnh viện ở Việt Nam với sự phổ biến của các chủng kháng carbapenem, đặc biệt là ST235. Kiểm soát nhiễm trùng hiệu quả đòi hỏi phải hiểu rõ chi tiết về sự lây truyền và mô hình kháng thuốc của nó. Nghiên cứu này nhằm phân tích pan genome, core genome MLST (cgMLST) và core genome SNP (cgSNP) của các chủng *P. aeruginosa* được phân lập ở Việt Nam, so sánh và tích hợp các phương pháp này để nâng cao độ phân giải dịch tễ học. Chúng tôi đã kiểm tra 196 bộ gene của *P. aeruginosa* được phân lập ở Việt Nam và 956 hệ gene hoàn chỉnh từ cơ sở dữ liệu NCBI. Định dạng chủng ban đầu được thực hiện bằng cách sử dụng MLST. Phân tích pan genome được thực hiện bằng Roary, trong khi các sơ đồ cgMLST được thiết lập với ChewBBACA và phân tích SNP của core genome được thực hiện bằng Parsnp. Sau đó, cây phát sinh loài được xây dựng và hiển thị bằng GrapeTree và iTOL. Phân tích xác định 334 sequence type (ST) đã biết từ cơ sở dữ liệu NCBI, bao gồm 127 ST mới, và 25 ST đã biết từ bộ dữ liệu Việt Nam, với 8 ST mới. ST235 là phổ biến nhất, chiếm 77,04% mẫu Việt Nam. Phân tích bộ gene toàn phần tiết lộ 68.963 gene, với 4,16% là các gen lõi. Sơ đồ cgMLST bao gồm 3.289 locus, cung cấp độ phân giải cao hơn so với MLST truyền thống. Phân tích phát sinh loài trên các chủng ST235 cho thấy các mô hình phân cụm rõ ràng, thể hiện khả năng phân biệt vượt trội của cgMLST và phân tích cgSNP. Nghiên cứu này làm nổi bật hiệu quả của các công cụ genome tiên tiến trong phân loại và so sánh các chủng *P. aeruginosa*, góp phần nâng cao hiểu biết về đa dạng di truyền và hỗ trợ phát triển các biện pháp kiểm soát nhiễm trùng và kháng thuốc của *P. aeruginosa* hiệu quả tại Việt Nam.

Từ khóa: Core genome MLST, core genome SNP, đa dạng di truyền, pan genome, *Pseudomonas aeruginosa*, Việt Nam.

MỞ ĐẦU

Pseudomonas aeruginosa là một vi khuẩn có khả năng biến đổi di truyền và chuyển hóa vượt trội, cho phép nó sống sót và phát triển trong nhiều môi trường khác nhau. Khả năng thích nghi linh hoạt của *P. aeruginosa* khiến nó trở thành một trong những tác nhân gây bệnh cơ hội nguy hiểm nhất, được xếp vào nhóm các tác nhân gây bệnh quan trọng lâm sàng ESKAPE (Rice, 2008) (bao gồm *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *P. aeruginosa* và *Enterobacter* spp.). Ở người, *P. aeruginosa* là nguyên nhân gây ra nhiều loại nhiễm trùng trong cộng đồng và bệnh viện, từ nhiễm trùng da và mô mềm đến viêm phổi và nhiễm trùng máu phức tạp (Driscoll *et al*., 2007). Đặc biệt, *P. aeruginosa* là tác nhân gây nhiễm trùng chủ yếu ở bệnh nhân xơ nang và là nguồn gốc chính của nhiễm trùng ở vết bỏng, dẫn đến tỷ lệ mắc bệnh và tử vong đáng kể (Kerr and Snelling, 2009).

Khu vực Đông Nam Á được coi là một "điểm nóng" về kháng kháng sinh, và *P. aeruginosa* đã được xác định là nguyên nhân phổ biến gây nhiễm trùng bệnh viện ở Việt Nam. Theo báo cáo giám sát kháng kháng sinh của Bộ Y tế Việt Nam năm 2020 (công bố tháng 11/2023), dữ liệu từ 16 bệnh viện tại 10 tỉnh cho thấy trong tổng số 69.715 chủng vi khuẩn được thu thập, *P. aeruginosa* đứng thứ năm với tỷ lệ 9,4% (6.564 chủng), là nguyên nhân hàng thứ ba khi phân lập bệnh phẩm từ đường hô hấp dưới (12,9% - 3.087 chủng), bệnh phẩm nước tiểu (8,5% - 826 chủng) và bệnh phẩm dịch ổ bụng (7,7% - 341 chủng). Theo dữ liệu từ Trung tâm Động lực học Bệnh tật, Kinh tế & Chính sách (CDDEP) năm 2016, 36% các chủng *P. aeruginosa* ở Việt Nam kháng carbapenems, đứng thứ hai toàn cầu chỉ sau Ấn Độ về tỷ lệ kháng thuốc. Một nghiên cứu từ năm 2016 tại một bệnh viện ở Hà Nội đã ghi nhận sự xuất hiện của các chủng *P. aeruginosa* thuộc kiểu ST235 mang các gen kháng carbapenemase như blaIMP-15, blaIMP-26, và blaIMP-51 (Tada *et al*., 2016).

Để nhanh chóng và hiệu quả triển khai các biện pháp kiểm soát trong quản lý đợt bùng phát, việc hiểu rõ sự lây truyền của các tác nhân kháng thuốc là điều cần thiết. Việc xác định các cụm và phân loại các tác nhân gây

bệnh, bao gồm các đặc điểm kháng thuốc và gen độc lực, đóng vai trò quan trọng trong việc kiểm soát nhiễm trùng một cách hiệu quả (Leopold *et al*., 2014). Điện di gel với trường điện từ (Pulsed-field gel electrophoresis-PFGE) và phân loại trình tự nhiều locus (multilocus sequence typing-MLST) đã được sử dụng như các tiêu chuẩn vàng trong nhiều năm để xác nhận các đợt bùng phát nghi ngờ. Tuy nhiên, PFGE thường gặp vấn đề về tính chủ quan, khó khăn trong việc giải thích kết quả và tốn thời gian. Ngược lại, MLST, được phát triển vào năm 2004 cho *P. aeruginosa*, dựa trên việc giải trình tự và đánh giá sự biến đổi alen của bảy gen bảo tồn, tạo ra các kiểu trình tự (sequence type - ST) để đặc trưng hóa các chủng (Curran *et al*., 2004). Tuy nhiên, khả năng phân biệt của MLST không phải lúc nào cũng đủ để giải quyết các đợt bùng phát.

Gần đây, các phương pháp dựa trên giải trình tự toàn bộ hệ gen (WGS) đã trở nên khả thi nhờ chi phí giải trình tự giảm; cùng với các công cụ tin sinh học cho phép phân biệt cao giữa các chủng có cùng ST trong các nghiên cứu dịch tễ học (Tang *et al*., 2017) như pan genome, core genome MLST, core genome SNP, ... Pan genome là tập hợp toàn bộ các gen trong một loài vi khuẩn, bao gồm cả gen lõi và gen phụ thuộc vào từng chủng. Phân tích Pan genome xác định gen lõi (có mặt trong tất cả các chủng) và gen phụ (chỉ có trong một số chủng), cung cấp cái nhìn về sự đa dạng di truyền và các yếu tố ảnh hưởng đến khả năng gây bệnh và kháng thuốc. Core Genome MLST (cgMLST) mở rộng MLST bằng cách sử dụng dữ liệu giải trình tự toàn bộ genome (WGS) để kiểm tra nhiều gen đích, tạo ra hệ thống đánh số alen có hệ thống, giúp phân giải cao hơn trong phát hiện và phân loại các đợt bùng phát. Phân tích cgSNP tập trung vào các SNPs trong các vùng bảo tồn của bộ gen lõi, xác định mối quan hệ tiến hóa và theo dõi sự lây lan của các chủng vi khuẩn, phát hiện khác biệt di truyền nhỏ và cung cấp cái nhìn chi tiết về sự phát tán của tác nhân gây bệnh.

Nghiên cứu này nhằm phân tích pan genome, core genome MLST và core genome SNP cho các chủng *P. aeruginosa* phân lập được tại Việt Nam. So sánh và kết hợp các phương pháp này có thể đạt được độ phân giải tối đa cho các phân tích dịch tễ học và giám sát *P. aeruginosa* (Zhou *et al*., 2017), hỗ trợ xác định và theo dõi sự lây truyền, tiến hóa và kháng thuốc của các chủng vi khuẩn *P. aeruginosa* tại Việt Nam.

DỮ LIỆU VÀ PHƯƠNG PHÁP

Dữ liệu

Dữ liệu giải trình tự của các chủng phân lập tại Việt Nam

Các mã Sequence Read Archive (SRA) của các mẫu *P. aeruginosa* phân lập tại Việt Nam được lưu trữ trên cơ sở dữ liệu National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). Tương ứng với các mã SRA này, 307 trình tự đọc đã được truy xuất. Chúng đều được phân lập từ bệnh phẩm của người trong khoảng thời gian từ năm 2013 đến năm 2019, thuộc bốn dự án PRJEB29424, PRJEB28400, PRJDB4025, và PRJDB2736. Các trình tự này đã được tinh sạch và lắp ráp theo các bước sau: 1. Đánh giá chất lượng dữ liệu ban đầu bằng FastQC với thông số mặc định; 2. Tinh sạch dữ liệu bằng Trimmomatic với các thông số: SLIDING WINDOW=25:4, MINLEN=100, ILLUMINACLIP=ON; 3. Đánh giá chất lượng dữ liệu sau tinh sạch bằng FastQC với thông số mặc định; 4. Lắp ráp hệ gen bằng SPAdes với thông số mặc định; 5. Đánh giá chất lượng lắp ráp bằng QUAST với thông số mặc định; 6. Đánh giá độ toàn vẹn hệ gen bằng BUSCO với thông số auto-lineage-prok; 7. Định danh taxonomy cho hệ gen lắp ráp bằng Kraken2 với database Minikraken2; 8. Lọc bỏ các contigs có độ dài nhỏ hơn 500bp. Các hệ gen sau lắp ráp thỏa mãn các điều kiện sau sẽ được đưa vào các phân tích phylogeny tiếp theo: kết quả định danh Loài (Species -S) gần nhất với *P. aeruginosa* và %S >=90%, độ hoàn thiện hệ gen >=90%, kích thước hệ gen sau lọc <=7,5M bp và số lượng contigs sau lọc <=400 (gọi tắt là tập dữ liệu VN).

Dữ liệu các hệ gen hoàn chỉnh của Pseudomonas aeruginosa có sẵn trên NCBI

Để phục vụ cho phân tích cgMLST và tìm hiểu sự liên hệ của các loài phân lập ở Việt Nam với các loài trên thế giới, tất cả 956 trình tự bộ gen hoàn chỉnh (tệp fasta) của *P. aeruginosa* có sẵn công khai tại cơ sở dữ liệu NCBI Reference Sequence (RefSeq) tính đến ngày 20 tháng 6 năm 2024 (gọi tắt là tập dữ liệu NCBI) đã được đưa vào các phân tích xây dựng cây phát sinh loài, cùng với các hệ gen đã được lắp ráp của *P. aeruginosa*. Thống kê cho thấy phần lớn các hệ gen này có từ 0 (738, 77,2%) hoặc 1 plasmid (156, 16,0%). Các hệ gen này được phân lập và lắp ráp trong khoảng thời gian từ năm 2006 đến năm 2024, với phần lớn được thực hiện từ năm 2018 đến 2024 (848, 88,7%). Các hệ gen *P. aeruginosa* này chủ yếu đến từ Trung Quốc (239, 25,0%), Mỹ (151, 15,8%), và Đan Mạch (87, 9,1%), và đa số được phân lập từ mẫu bệnh phẩm của người (767, 80,2%). Có duy nhất 1 hệ gen được phân lập tại Việt Nam nhưng từ mẫu đất.

Phương pháp

Phân tích MLST

956 hệ gen hoàn chỉnh được tải về từ cơ sở dữ liệu genome NCBI và 196 hệ gen được lắp ráp từ các trình tự của các mẫu phân lập tại Việt Nam đã được quét đối chiếu với các sơ đồ phân loại truyền thống của PubMLST (https://pubmlst.org/) bằng công cụ mlst. Các ST mới được gán khi không tìm thấy ST tương ứng trong cơ sở dữ liệu PubMLST trong khuôn khổ của nghiên cứu này.

Phân tích Pan Genome

956 hệ gen hoàn chỉnh và 196 hệ gen lắp ráp được chú giải bằng công cụ Prokka. Các tập tin GFF3 được tạo ra bởi Prokka được sử dụng để thực hiện phân tích pan genome bằng công cụ Roary (phiên bản 3.11.2). Theo đó, các gen được phân loại thành bốn nhóm khác nhau: nhóm "cốt lõi" (99% ≤ số chủng ≤ 100%), nhóm "cốt lõi mềm" (95% ≤ số chủng < 99%), nhóm "vỏ" (15% ≤ số chủng < 95%) và nhóm "mây" (0% ≤ số chủng < 15%). Cây pan genome phản ánh mối quan hệ tiến hóa giữa các chủng *P. aeruginosa* dựa trên sự tương đồng và khác biệt trong toàn bộ gen của chúng, được xây dựng từ dữ liệu sự hiện diện và vắng mặt của các gen và trực quan hóa bằng phần mềm Phandango (Phandango.net).

Phân tích core genome MLST

Sơ đồ cgMLST của *P. aeruginosa* (scheme) được thiết lập với công cụ Comprehensive and High Efficient Workflow for a Blast Score Ratio Based Allele Calling Algorithm (ChewBBACA) (Silva *et al*., 2018). Trong đó, trình tự genome của *P. aeruginosa* PAO1 (RefSeq assembly accession: GCF_000006765.1) chỉ được sử dụng làm genome tham chiếu để dự đoán các locus whole-genome MLST (wgMLST). Chúng tôi đã chọn các locus ứng viên cho sơ đồ cgMLST có mặt trong 99% các hệ gen hoàn chỉnh hiện có (956 hệ gen). Sau đó, 196 hệ gen lắp ráp từ các mẫu được phân lập tại Việt Nam đã được gọi alen theo sơ đồ cgMLST đã xác định. Cây phân loài cgMLST được xây dựng bằng phần mềm GrapeTree (phiên bản 1.5.0) (Zhou *et al*., 2018) với các thông số được triển khai trong MSTree v2, và được trực quan hóa bằng phần mềm Interactive Tree Of Life (iTOL, phiên bản 4.2.3) (Letunic and Bork, 2016). Qua đó, khả năng phân loại các mẫu trình tự của *P. aeruginosa* bằng cgMLST và MLST được nhận xét và so sánh.

So sánh phân tích core genome MLST với core genome SNP cho các mẫu Pseudomonas aeruginosa ST235

Tất cả các hệ gen *P. aeruginosa* ST235 của Việt Nam và các quốc gia khác trong nghiên cứu này sẽ được chọn. Một cây phát sinh loài dựa trên các cgMLST đã được xây dựng bằng phần mềm GrapeTree với các thông số được triển khai trong NJ. Đối với phân tích SNP dựa trên bộ gen lõi, bộ gen lõi được căn chỉnh bằng Parsnp, là một phần của gói phần mềm Harvest, sử dụng NCGM2.S1 (RefSeq assembly accession: GCF_000284555.1) làm bộ gen tham chiếu. Cả hai cây cùng được trực quan hóa bằng iTOL. Sự tương đồng giữa hai phương pháp, cgMLST và phân tích SNP dựa trên bộ gen lõi, sẽ được thảo luận dựa trên những điểm tương đồng và khác biệt trong việc phân cụm của hai cây phát sinh loài.

KẾT QUẢ VÀ THẢO LUẬN

Lắp ráp các chủng phân lập tại Việt Nam

Sau khi áp dụng các tiêu chuẩn loại trừ, 196 hệ gen lắp ráp (PRJEB29424: 110; PRJEB28400: 37; PRJDB2736: 26; PRJDB4025: 23) được chọn để tiến hành các phân tích tiếp theo. Các hệ gen này có kích thước từ 6,39 Mbp đến 7,45 Mbp và đạt mức độ hoàn thiện rất cao, với giá trị trung vị đạt 100% và các giá trị dao động từ 96,8% đến 100%. Tỷ lệ phần trăm tương đồng với *P. aeruginosa* (%S_*P. aeruginosa*) có giá trị trung bình khoảng 94,36%. Các hệ gen có chất lượng lắp ráp tốt, với số lượng contigs dao động từ 35 đến 390, đảm bảo chất lượng cao cho các phân tích cây phát sinh loài tiếp theo.

Xác định MLST

Trong tập dữ liệu từ NCBI, tổng số 334 ST được xác định, trong đó 207 ST đã được ghi nhận trong cơ sở dữ liệu pubMLST và 127 ST là các ST mới. Các ST đã biết xuất hiện nhiều nhất trong tập dữ liệu này bao gồm ST235 (54 mẫu), ST111 (43 mẫu), ST463 (29 mẫu), ST309 (18 mẫu), ST357 (17 mẫu), ST253 (17 mẫu), ST277 (15 mẫu), ST233 (13 mẫu), ST395 (13 mẫu)... Đối với tập dữ liệu VN (**Bảng 1**), có 25 ST được xác định, trong đó 17 ST đã được ghi nhận và 8 ST là các ST mới. Tương tự như tập dữ liệu NCBI, ST235 là ST phổ biến nhất với 151 mẫu được định danh, chiếm 77.04% tổng số mẫu. Tiếp theo là ST357 với 10 mẫu, chiếm 19.60%. Các ST khác như ST644, ST360, ST310, ST1650, ST2332,... có số lượng mẫu dao động từ 1-4. Các ST mới chỉ xuất hiện ở mức thấp, với số lượng mẫu dao động từ 1-2 cho mỗi ST mới.

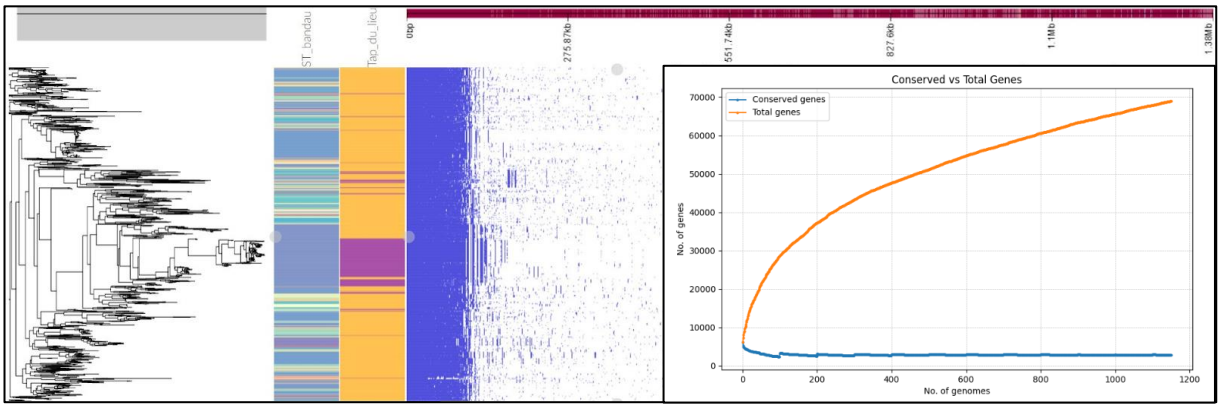
Bảng 1. Kết quả định danh MLST của 1.152 hệ gen *Pseudomonas aeruginosa*

ST	Sheme	acsA	aroE	guaA	mutL	nuoD	ppsA	trpE	NCBI	VN
235	paeruginosa	38	11	3	13	1	2	4	54	151
357	paeruginosa	2	4	5	3	1	6	11	17	10
644	paeruginosa	28	3	94	13	1	4	10	3	4
360	paeruginosa	15	5	36	11	27	4	2	2	4
310	paeruginosa	5	59	60	3	1	6	4	0	4

1650	paeruginosa	16	5	26	3	4	15	8	0	2
2332	paeruginosa	15	5	91	11	3	4	2	0	2
new_1	paeruginosa	38	~11	3	13	1	2	4	0	2
277	paeruginosa	39	5	9	11	27	5	2	15	1
773	paeruginosa	5	4	5	5	5	7	8	11	1
1971	paeruginosa	32	190	3	62	8	7	26	8	1
1212	paeruginosa	11	10	11	72	3	10	3	4	1
664	paeruginosa	9	5	11	3	4	40	18	2	1
2165	paeruginosa	87	42	114	37	53	97	147	1	1
830	paeruginosa	5	13	109	5	1	1	47	0	1
1410	paeruginosa	13	158	5	5	12	7	15	0	1
1649	paeruginosa	16	5	12	77	11	4	18	0	1
3359	paeruginosa	47	50	65	31	1	6	8	0	1
new_2	paeruginosa	38	~11	3	~13	1	2	4	0	1
new_3	paeruginosa	9	~5	~11	~21	4	40	18	0	1
new_4	paeruginosa	39	5	9	11	~27	5	2	0	1
new_5	paeruginosa	27	~14	25	~23	1	16	~46	0	1
new_6	paeruginosa	~5	4	~5	~5	5	7	8	0	1
new_7	paeruginosa	~28	3	~94	~13	1	4	~10	0	1
new_8	paeruginosa	11	~6	11	13	2	7	53	0	1

Phân tích Pan genome từ Roary

Phân tích Roary của pan genome của tổng số 1.152 chủng *P. aeruginosa* đã xác định được 2.868 gen lõi (4,16%), 1.619 gen lõi mềm (2,45%), 2.330 gen vỏ (3,38%) và 62.146 gen đám mây (90,01%), trong tổng số 68.963 gen. **Hình 1** (bên trái) cho thấy các hệ gen từ các mẫu phân lập tại Việt Nam (màu tím) có số lượng các gen phụ nhiều hơn so với các hệ gen hoàn thiện đã được công bố trên NCBI. Đặc biệt, có thể dễ dàng thấy tồn tại một nhánh nhỏ các hệ gen trong tập dữ liệu VN nằm trong số 1% không chứa đầy đủ một số lượng khá lớn các gen lõi nhưng lại cùng có một số lượng gen phụ đặc trưng ở vị trí khoảng 551,7kb. **Hình 1** (bên phải) cho thấy số lượng gen lõi được cân bằng khi số lượng hệ gen tăng cho thấy rằng các gen lõi là cần thiết và bảo tồn cao giữa các chủng khác nhau, phản ánh tính ổn định và thiết yếu của chúng trong bộ gen. Số lượng lớn các gen đám mây cho thấy có sự dị biệt lớn giữa tổng số 1.152 chủng *P. aeruginosa* được xem xét, nhấn mạnh tính chất 'mở' của pan genome *P. aeruginosa.*

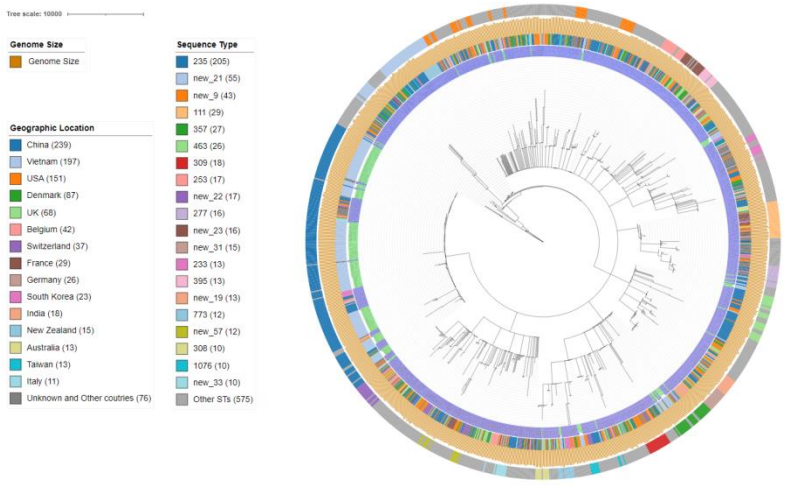


Hình 1. Pan genome của *Pseudomonas aeruginosa*

Bên trái: cây phân loài và biểu đồ ma trận (bên trái) dựa trên sự hiện diện hay vắng mặt của các gen lõi và gen phụ cùng với thông tin về sequence type (ST) và tập dữ liệu (NCBI: màu cam, VN: màu tím); Bên phải: đường màu đỏ thể hiện tổng số gen và đường màu xanh thể hiện số gen bảo tồn (gen lõi) khi số lượng hệ gen Pseudomonas aeruginosa tăng lên.

Phân tích core genome MLST

Một sơ đồ cgMLST bao gồm 3.289 gen đích đã được xác định, bao phủ 73,58% trong số 5.570 ORF được dự đoán cho chủng tham chiếu PAO1. Kết quả này khá đồng nhất với kết quả tổng có 3.168 gen đích trong nghiên cứu phân tích cgMLST của *P. aeruginosa* của Romário Oliveira de Sales và cộng sự. Theo đó, 196 chủng phân lập tại Việt Nam được xác định sơ đồ cgMLST theo bộ gen đích trên. Kết quả của bộ sơ đồ cgMLST của 1.152 chủng *P. aeruginosa* sau đó được sử dụng để xây dựng cây phân loài (**Hình 2**). Tính từ trong ra ngoài, trong cùng là cây phân loài được xây dựng từ kết quả cgMLST. Vòng tròn đầu tiên phân biệt giữa các chủng được phân lập tại Việt Nam (màu xanh lá) và trên thế giới (màu tím). Vòng tròn thứ hai phân loại cụ thể nguồn gốc quốc gia của các chủng *P. aeruginosa.* Vòng tròn thứ ba thể hiện kích thước hệ gen của từng chủng. **Hình 2** đã cho thấy được sự vượt trội về khả năng phân loại các chủng *P. aeruginosa* và thể hiện chi tiết hơn về mối quan hệ di truyền giữa các chủng này của phân tích cgMLST so với phương pháp truyền thống MLST.



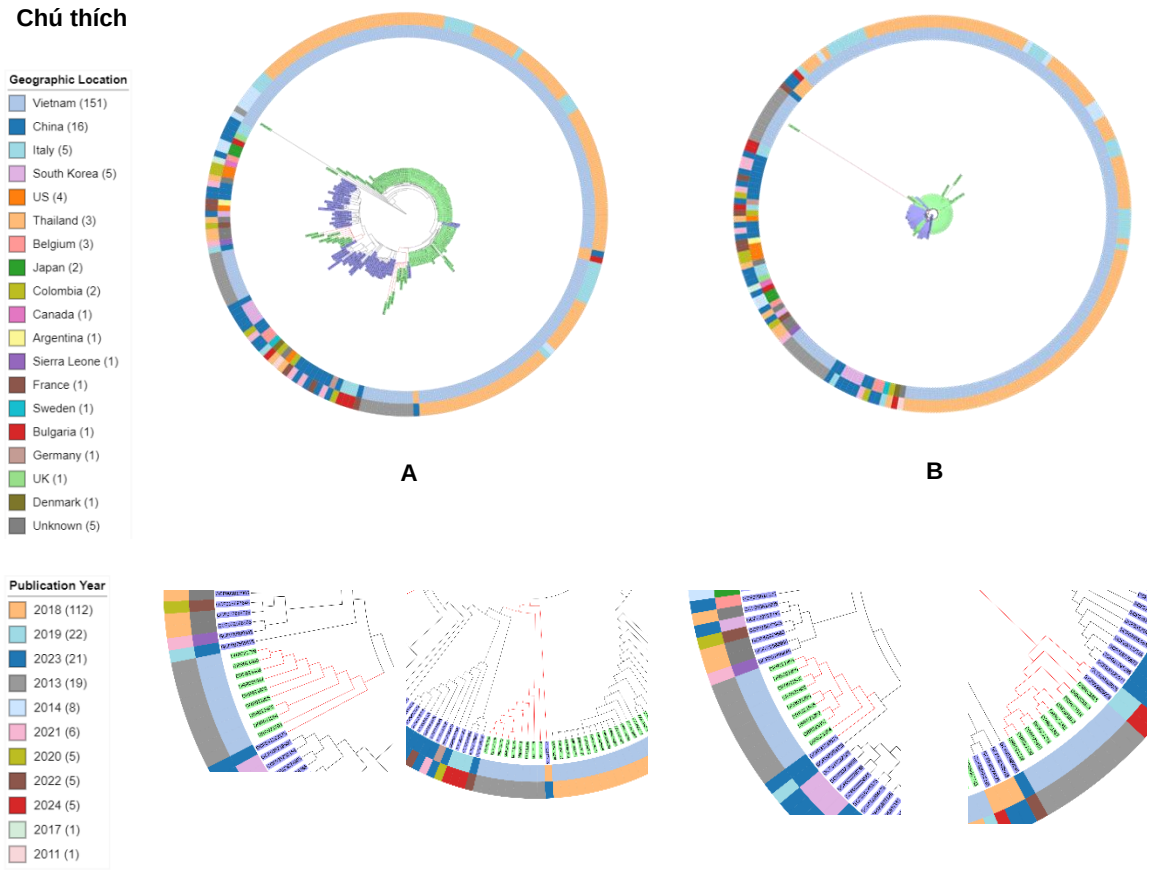
Hình 2. Cây phát sinh loài dựa vào phân tích cgMLST

Tính từ trong ra ngoài: Tập dữ liệu (xanh lá: VN, tím: NCBI); Quốc gia phân lập; Kích thước hệ gen; Sequence Type

So sánh phân tích cgMLST và cgSNP cho các chủng Pseudomonas aeruginosa ST235

Nghiên cứu sự đa dạng của *P. aeruginosa* ST235 ở Việt Nam là quan trọng vì đây là nhóm vi khuẩn đa kháng phổ biến tại Việt Nam (chiếm tới 77,14% tổng số mẫu trong nghiên cứu này), đòi hỏi sự quan tâm đặc biệt để hỗ trợ dịch tễ học và phát triển chiến lược kiểm soát. ST235 thể hiện sự thay đổi không đáng kể trong kiểu gen qua một thập kỷ, như được ghi nhận trong nghiên cứu của Wei Feng và cộng sự, cho phép khảo sát các phương pháp phân loại có độ phân giải cao hơn cấp độ gen như cgMLST và cgSNP, giúp đánh giá chính xác biến đổi di truyền và hiệu quả của các công cụ phân tích này.

Hình 3A và **Hình 3B** lần lượt là cây phân loài được vẽ từ phân tích cgMLST và cgSNP cho 205 chủng *P. aeruginosa* ST235 được xác định trong nghiên cứu của chúng tôi. Ở cả hai hình, chúng ta có thể thấy được sự tương đồng của đa số các chủng ở Việt Nam, và chúng có sự khác biệt với các chủng trên thế giới. Đặc biệt, chủng có ID DRR021817 (Việt Nam, 2013) có khoảng cách xa hẳn các chủng còn lại; các chủng còn lại này đều được phân lập năm 2013 và chúng tách thành hai nhóm rõ rệt ở cả hai hình. Tuy nhiên, vẫn có sự khác biệt giữa hai phương pháp phân loại này. Cụ thể, ở cây phân loài cgMLST, chúng tôi phát hiện hai nhóm trên lần lượt có mối quan hệ gần nhất với hai chủng được phân lập năm 2023 ở châu Á (Thái Lan - 2023, Trung Quốc - 2023); trong khi đó, ở cây phân loài cgSNP, hai nhóm này lần lượt có độ tương đồng cao nhất lần lượt với một chủng ở Trung Quốc năm 2023 và một chủng ở Ý năm 2024.



Hình 3. Cây phát sinh loài cho các chủng *Pseudomonas aeruginosa* ST235

A. Cây từ phân tích cgMLST; B. Cây từ cgSNP. Tính từ trong ra ngoài, trong cùng là cây phân loài với nhãn màu tím thuộc dữ liệu NCBI và nhãn màu xanh thuộc tập dữ liệu VN; vòng ở trong thể hiện vị trí địa lý; vòng ngoài cùng thể hiện năm phân lập của các chủng Pseudomonas aeruginosa ST235

Nghiên cứu này có một vài hạn chế. Thứ nhất, chỉ các hệ gen hoàn chỉnh có sẵn trên cơ sở dữ liệu NCBI được đưa vào đồng phân tích cây phát sinh loài, do đó thiếu các dữ liệu để kết luận về mối quan hệ của các hệ gen chưa hoàn thiện thuộc các ST được quan tâm và công bố ở thời điểm khác. Thứ hai, dựa trên kết quả tìm kiếm trên NCBI, chỉ có dữ liệu của các chủng *P. aeruginosa* Việt Nam được phân lập vào các năm 2013, 2014, 2018, và 2019, dẫn đến thiếu hụt dữ liệu hệ gen của vi khuẩn này trong những năm gần đây để đánh giá và nhận xét về mức độ phát triển của chúng. Bên cạnh ST235, ST357 cũng nằm trong top 10 sequence type đáng quan tâm về vấn đề kháng thuốc nhưng chưa được tìm hiểu sâu hơn về mức độ đa dạng. Trong các nghiên cứu tương lai, chúng tôi sẽ cố gắng thu thập dữ liệu và giải trình tự các chủng *P. aeruginosa* mới nhất, cũng như khám phá mức độ phân loại của các chủng sequence type khác để bổ sung vào hiểu biết hiện có.

KẾT LUẬN

Nghiên cứu này đã chứng minh được mức độ phân giải cao của các công cụ được khảo sát trong việc phân loại các chủng *P. aeruginosa* ở Việt Nam và so sánh chúng với các chủng trên thế giới. Các kết quả của nghiên cứu này sẽ góp phần vào việc nâng cao hiểu biết về sự đa dạng di truyền và sự lan truyền của *P. aeruginosa* nói riêng và các chủng vi khuẩn đa kháng nói chung, từ đó hỗ trợ phát triển các biện pháp kiểm soát và phòng ngừa hiệu quả hơn, từ đó bảo vệ sức khỏe cộng đồng trong tình hình kháng thuốc dữ dội tại Việt Nam.

TÀI LIỆU THAM KHẢO

Curran B, Jonas D, Grundmann H, Pitt T, Dowson CG, 2004. Development of a Multilocus Sequence Typing Scheme for the Opportunistic Pathogen *Pseudomonas aeruginosa*. *J Clin Microbiol* 42, 5644-5649. https://doi.org/10.1128/JCM.42.12.5644-5649.2004

Driscoll JA, Brody SL, Kollef MH, 2007. The Epidemiology, Pathogenesis and Treatment of Pseudomonas aeruginosa Infections. *Drugs* 67, 351-368. https://doi.org/10.2165/00003495-200767030-00003

Kerr KG, Snelling AM, 2009. *Pseudomonas aeruginosa*: a formidable and ever-present adversary. *J Hospital Infect*, Proceedings of The Lancet Conference on Healthcare-Associated Infections 73, 338-344. https://doi.org/10.1016/j.jhin.2009.04.020

Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A, 2014. Bacterial Whole-Genome Sequencing Revisited: Portable, Scalable, and Standardized Analysis for Typing and Detection of Virulence and Antibiotic Resistance Genes. *J Clin Microbiol* 52, 2365–2370. https://doi.org/10.1128/JCM.00262-14

Letunic I, Bork P, 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44, W242-W245. https://doi.org/10.1093/nar/gkw290

Rice LB, 2008. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. *J Infect Dis* 197, 1079-1081. https://doi.org/10.1086/533452

Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA, 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 4, e000166. https://doi.org/10.1099/mgen.0.000166

Tada T, Nhung PH, Miyoshi-Akiyama T, Shimada K, Tsuchiya M, Phuong DM, Anh NQ, Ohmagari N, Kirikae T, 2016. Multidrug-Resistant Sequence Type 235 Pseudomonas aeruginosa Clinical Isolates Producing IMP-26 with Increased Carbapenem-Hydrolyzing Activities in Vietnam. *Antimicrob Agents Chemother* 60, 6853-6858. https://doi.org/10.1128/AAC.01177-16

Tang P, Croxen MA, Hasan MR, Hsiao WWL, Hoang LM, 2017. Infection control in the new age of genomic epidemiology. *Amer J Infect Cont* 45, 170-179. https://doi.org/10.1016/j.ajic.2016.05.015

Zhou H, Liu W, Qin T, Liu C, Ren H, 2017. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Klebsiella pneumoniae*. *Front Microbiol* 8, 371. https://doi.org/10.3389/fmicb.2017.00371

Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M, 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 28, 1395–1404. https://doi.org/10.1101/gr.232397.117

STUDY ON THE GENETIC DIVERSITY OF *Pseudomonas aeruginosa* IN VIETNAM THROUGH PAN-GENOME, CORE GENOME MLST, AND CORE GENOME SNP ANALYSES

Vuong Thi Huong[1], Nguyen Tien Dat[1], Trinh Thi Xuan[2], Nguyen Cuong[3*]

[1]LOBI Vietnam Ltd

[2]Faculity of Information Technology, Hanoi Open University

[3]Institute of Information Technology, Vietnam Academy of Science and Technology

SUMMARY

Pseudomonas aeruginosa, a highly adaptable pathogen, is one of the major causes of hospital infections in Vietnam, particularly with the prevalence of carbapenem-resistant strains, notably ST235. Effective infection control requires a detailed understanding of its transmission and resistance patterns. This study aims to analyze the pan-genome, core genome MLST (cgMLST), and core genome SNP (cgSNP) of *P. aeruginosa* strains isolated in Vietnam, comparing and integrating these methods to enhance epidemiological resolution. We examined 196 *P. aeruginosa* gene profiles from Vietnam and 956 complete genome sequences from the NCBI database. Initial strain typing was performed using MLST. Pan-genome analysis was conducted with Roary, while cgMLST schemas were developed with ChewBBACA, and core genome SNP analysis was performed using Parsnp. Phylogenetic trees were constructed and visualized using GrapeTree and iTOL. The analysis identified 334 known sequence types (ST) from the NCBI database, including 127 new STs, and 25 known STs from the Vietnamese dataset, with 8 new STs. ST235 was the most prevalent, accounting for 77.04% of the Vietnamese samples. The pan-genome analysis revealed 68,963 genes, with 4.16% being core genes. The cgMLST schema included 3,289 loci, providing higher resolution compared to traditional MLST. Phylogenetic analysis of ST235 strains demonstrated clear clustering patterns, showcasing the superior discriminative ability of cgMLST and cgSNP analysis. This study highlights the effectiveness of advanced genomic tools in classifying and comparing *P. aeruginosa* strains, enhancing understanding of genetic diversity, and supporting the development of effective infection control and resistance management strategies in Vietnam.

Keywords: Core genome MLST, core genome SNP, genetic diversity, pan-genome, *Pseudomonas aeruginosa*, Vietnam.

* Author for correspondence: Tel: 0916110333; Email: ncuong@ioit.ac.vn