

SEQUENCING AND ASSEMBLY HUMAN MITOCHONDRIAL GENOME USING LONG READ SEQUENCING

Le Thi Dung^{1,2}, Le Tung Lam¹, Nguyen Quang Huy², Truong Nam Hai^{1,3*}

¹Institute of Biology - Vietnam Academy of Science and Technology

²University of Science and Technology of Hanoi - Vietnam Academy of Science and Technology

³Graduate University of Science and Technology - Vietnam Academy of Science and Technology

ARTICLE INFO	ABSTRACT
Received: 05/5/2025	The human mitochondrial genome (mtDNA) plays a vital role in medical research, kinship identification, and population evolutionary analysis, especially when dealing with highly degraded samples. However, in Vietnam, research on mtDNA has primarily focused on the hypervariable regions within the D-loop, while the entire mitochondrial genome has not yet been fully explored. In this study, we employed high-accuracy PacBio HiFi sequencing technology to sequence and assemble the complete mitochondrial genome from a Vietnamese individual. The assembled genome is 16,809 base pairs in length, with a GC content of 44.4%. The findings provide more comprehensive data for mtDNA-related studies and highlight the advantages of long-read sequencing technology in mitochondrial genome assembly. Our result contributes to providing data for more in-depth applications in genetics and forensic science in Vietnam, especially expanding complete mtDNA analysis beyond the D-loop region to enhance the understanding of mitochondrial diversity and its implications for biomedical and anthropological research in the region.
Revised: 30/10/2025	
Published: 31/10/2025	
KEYWORDS	
mtDNA	
Sequencing	
Assembly	
HiFi read	
SMRTbell	

GIẢI TRÌNH TỰ VÀ LẮP RÁP GENOME TY THỂ NGƯỜI BẰNG CÔNG NGHỆ GIẢI TRÌNH TỰ ĐOẠN DÀI

Lê Thị Dung^{1,2}, Lê Tùng Lâm¹, Nguyễn Quang Huy², Trương Nam Hải^{1,3*}

¹Viện Sinh học – Viện Hàn lâm Khoa học và Công nghệ Việt Nam

²Trường Đại học Khoa học và Công nghệ Hà Nội – Viện Hàn lâm Khoa học và Công nghệ Việt Nam

³Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 05/5/2025	Hệ gen ty thể (mtDNA) người đóng vai trò quan trọng trong nghiên cứu y học, giám định huyết thống và phân tích tiến hóa quần thể, đặc biệt với các mẫu bị thoái hóa nặng. Tuy nhiên, tại Việt Nam, các nghiên cứu về mtDNA vẫn chủ yếu tập trung vào các vùng siêu biến thuộc vùng D-loop, trong khi toàn bộ hệ gen ty thể vẫn chưa được khai thác đầy đủ. Trong nghiên cứu này, chúng tôi sử dụng công nghệ giải trình tự PacBio HiFi với độ chính xác cao để giải trình tự và lắp ráp toàn bộ hệ gen ty thể từ một cá thể người Việt Nam. Hệ gen thu được có chiều dài 16.809 bp và hàm lượng GC là 44,4%. Kết quả nghiên cứu đóng góp vào việc cung cấp dữ liệu đầy đủ hơn cho các nghiên cứu liên quan đến mtDNA và cho thấy lợi ích của công nghệ giải trình tự đọc dài trong việc lắp ráp hệ gen ty thể. Nghiên cứu góp phần cung cấp dữ liệu cho các ứng dụng sâu hơn trong di truyền học và giám định tại Việt Nam, đặc biệt là các nghiên cứu phân tích mtDNA hoàn chỉnh vượt ra ngoài vùng D-loop để nâng cao hiểu biết về sự đa dạng ty thể và ý nghĩa của nó đối với nghiên cứu y sinh và nhân chủng học trong khu vực.
Ngày hoàn thiện: 30/10/2025	
Ngày đăng: 31/10/2025	
TỪ KHÓA	
mtDNA	
Giải trình tự	
Lắp ráp	
Đoạn đọc HiFi	
SMRTbell	

DOI: <https://doi.org/10.34238/tnu-jst.12719>

* Corresponding author. Email: tnhai@ibt.ac.vn

1. Giới thiệu

Ty thể là một bào quan thiết yếu trong hầu hết các tế bào nhân thực, mang theo nhiều bản sao của một bộ gen nhỏ dạng vòng có nguồn gốc từ một sinh vật cổ từng sống tự do. Bộ gen ty thể của người dài 16.569 cặp base, mã hóa tổng cộng 37 gen – bao gồm 13 gen cho các protein thuộc chuỗi hô hấp tế bào, 22 tRNA và 2 rRNA – đóng vai trò then chốt trong quá trình phosphoryl hóa oxy hóa [1], [2]. Sự hiện diện đồng thời của nhiều biến thể thuộc mtDNA trong cùng một tế bào – còn gọi là dị hợp tử ty thể (heteroplasmy) – là một hiện tượng được ghi nhận ở nhiều cá thể người, đặc biệt trong các mô có tốc độ phân chia cao hoặc trong bối cảnh lão hóa, bệnh lý hoặc di truyền mẹ - con. Mặc dù heteroplasmy không phổ biến trong tất cả mọi cá thể hay mô, nhưng khi xuất hiện, nó có thể gây ảnh hưởng đáng kể đến sức khỏe, phụ thuộc vào tỷ lệ đột biến, vị trí xảy ra và tương tác với các yếu tố nội sinh – ngoại sinh [3], [4]. Do vậy, việc phát hiện và định lượng chính xác các biến thể mtDNA là yếu tố cốt lõi trong chẩn đoán bệnh, tư vấn di truyền và xây dựng phác đồ điều trị [5].

Các phương pháp truyền thống như PCR kết hợp với giải trình tự đọc ngắn đã giúp xác định nhiều biến thể ty thể. Tuy nhiên, những kỹ thuật này thường bị hạn chế bởi chiều dài đọc ngắn, không cho phép phát hiện đầy đủ các tái tổ hợp lớn hay biến thể ở pha liên kết, đồng thời tiềm ẩn nguy cơ tạo lỗi polymerase trong quá trình khuếch đại. Gần đây, công nghệ giải trình tự đọc dài PacBio HiFi đã mở ra một hướng đi mới nhờ khả năng bao phủ toàn diện các vùng phức tạp của bộ gen, mặc dù vẫn còn những thách thức về độ chính xác trong việc nhận diện đột biến nhỏ và indel [6], [7].

Trước đây, với hạn chế về kinh phí và quy mô thực hiện, các nghiên cứu về mtDNA tại Việt Nam chủ yếu tập trung vào hai vùng siêu biến HVI và HVII nằm trong vùng D-loop – nơi có tần suất đột biến cao – phục vụ cho giám định huyết thống và nghiên cứu mối liên hệ giữa các nhóm dân cư. Đặc biệt, trong các mẫu DNA thoái hóa, khả năng truy xuất thông tin từ mtDNA vượt trội hơn DNA nhân nhờ số lượng bản sao cao, làm tăng độ thành công trong lập hồ sơ di truyền [8]. Tuy nhiên, trong những năm gần đây, một số nhiệm vụ cấp quốc gia đã tiến hành giải trình tự toàn bộ hệ gen ty thể của người Việt thuộc nhiều nhóm dân tộc khác nhau. Những kết quả này góp phần làm sáng tỏ sự đa dạng di truyền và nguồn gốc tiến hóa của các dân tộc sinh sống tại Việt Nam.

Kể từ khi bản đồ di truyền đầu tiên của mtDNA người được công bố [1] đã mở đường cho hàng trăm nghiên cứu chuyên sâu về mtDNA trong các quần thể khác nhau. Đến nay, hàng chục nghìn trình tự hệ gen ty thể hoàn chỉnh đã được lưu trữ tại Ngân hàng gen thuộc Trung tâm Thông tin Công nghệ Sinh học Quốc gia (NCBI) của Viện Y tế Quốc gia Hoa Kỳ, phản ánh sự đa dạng di truyền sâu sắc giữa các dân tộc trên thế giới.

Trong nghiên cứu này, chúng tôi tiến hành giải trình tự và lắp ráp một mtDNA người từ dữ liệu PacBio HiFi có độ chính xác cao. Nghiên cứu không chỉ đóng góp vào hiểu biết về biến thể mtDNA trong bệnh học, mà còn mở rộng tiềm năng ứng dụng trong nghiên cứu tiến hóa, dân số học và các loài sinh vật nhân chuẩn khác.

2. Phương pháp nghiên cứu

2.1. Thu thập mẫu

Nghiên cứu được thực hiện theo phê duyệt của Hội đồng Khoa học Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, vào ngày 01 tháng 6 năm 2021. Mẫu máu được thu thập từ tình nguyện viên nam đảm bảo các tiêu chí: 1/Có độ tuổi nằm trong khoảng từ 30-59 nhằm đảm bảo sự ổn định nhất về mặt sinh học và sự thay đổi tối thiểu về mặt di truyền và sức khỏe. 2/Tình nguyện viên tự kê khai tình trạng sức khỏe, không có bệnh di truyền, có tổ tiên là người Kinh, có gia đình sống tại Việt Nam tám thế hệ. 3/Tình nguyện viên cần có ít nhất một người con trai khỏe mạnh không có bất kỳ bệnh di truyền hoặc tiền sử bệnh tật nào.

2.2. Tách chiết

DNA tổng số được tách chiết từ mẫu máu mới bằng bộ kit Blood & Cell Culture DNA Midi Kit (QIAGEN) theo hướng dẫn của nhà sản xuất. Chất lượng DNA được kiểm tra bằng điện di trên

gel agarose 1%, đo nồng độ Qubit 3 Fluorometer (Invitrogen), và kiểm tra độ dài bằng hệ thống Bioanalyzer 2100 (Agilent).

2.3. Chuẩn bị thư viện và giải trình tự SMRTbell

Thư viện SMRTbell cho giải trình tự PacBio được thực hiện theo hướng dẫn của nhà sản xuất với một vài thay đổi nhỏ. Nghiên cứu đã bỏ qua bước phân đoạn DNA tổng số thành các đoạn 10–15 kb do kích thước DNA được kiểm tra đã nằm trong phạm vi này. Thư viện SMRTbell được chuẩn bị với bộ SMRTbell Express Template Prep Kit 2.0 (PacBio) và tinh sạch bằng AMPure PB beads (PacBio). Thư viện được kiểm tra chất lượng lần cuối với Bioanalyzer 2100 (Agilent) trước khi tải lên SMRT Cells (PacBio) và giải trình tự bằng hệ thống giải trình tự đoạn dài PacBio theo quy trình được tạo bởi SMRTLink portal phiên bản 9.0.

2.4. Trình tự DNA ty thể tham chiếu

Nghiên cứu sử dụng trình tự DNA ty thể Cambridge (rCRS) mã số NC012920 của Ngân hàng gen làm tham chiếu. Đây là trình tự chuẩn có chiều dài 16.569 bp và được sử dụng phổ biến nhất trong các nghiên cứu về mtDNA người.

2.5. Lắp ráp hệ gen ty thể từ các đoạn đọc SMRT sequencing

Trình tự DNA thu được có nguồn gốc từ hệ gen ty thể được xác định thông qua căn chỉnh với NC012920 bằng chương trình pbmm2. Sau đó, các đoạn đọc có chiều dài trên 18 kbp sẽ được loại bỏ bằng việc sử dụng option filter tùy chỉnh của awk, các đoạn đọc còn lại được giữ cho các phân tích tiếp theo. Kết quả căn chỉnh dạng BAM được chuyển đổi thành dạng FASTQ bằng cách sử dụng samtools fastq để tạo đầu vào cho bước lắp ráp bằng Hifiasm phiên bản v0.19.8. Kết quả lắp ráp là tệp tin GFA được chuyển thành dạng fasta bằng awk. Cuối cùng, nghiên cứu tiến hành lọc, loại bỏ các trình tự lắp ráp được có chiều dài dưới 15 kb và trên 20 kb, để thu trình tự có kích thước kỳ vọng của mtDNA người (~16,5 kb).

2.6. Biểu diễn dữ liệu

Trình tự mtDNA lắp ráp được biểu diễn bằng cách so sánh với rCRS bằng chương trình MITOMASTER.

2.7. Lưu trữ dữ liệu

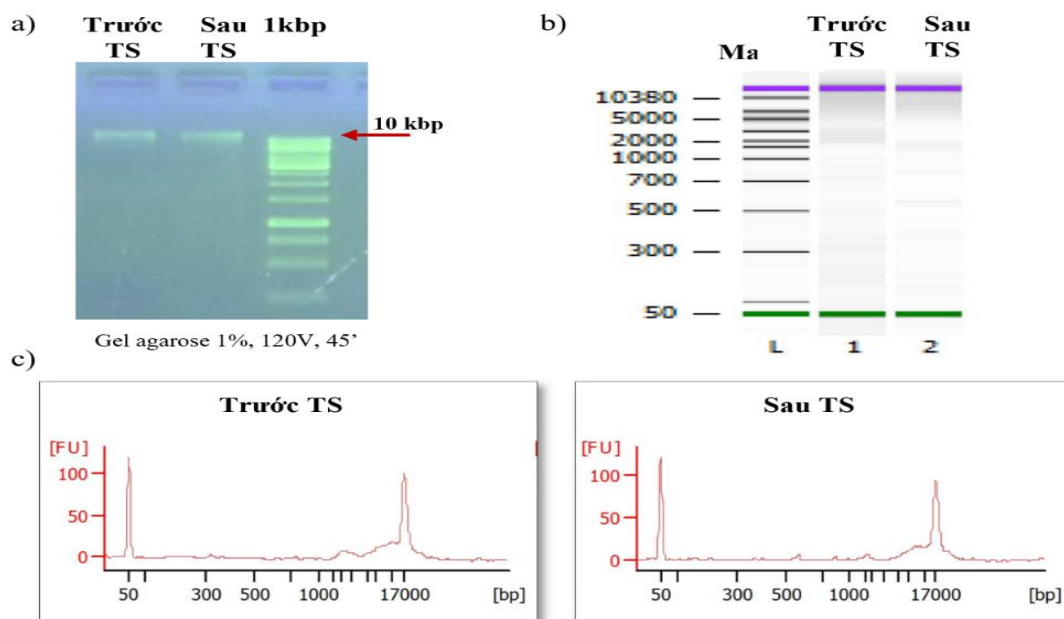
Trình tự mtDNA lắp ráp được và toàn bộ dữ liệu liên quan được lưu trữ tại hệ thống máy chủ thuộc quản lý của Viện Hàn lâm Khoa học và Công nghệ Việt Nam, với ký hiệu mtDNA do Dự án có mã số CT0000.09/21-22 quản lý và sử dụng.

3. Kết quả và thảo luận

3.1. Kết quả tách chiết DNA tổng số từ mẫu máu

Để thu được nồng độ DNA cao và có độ sạch tối đa, nghiên cứu đã tiến hành tách chiết DNA từ mẫu máu bằng bộ kit QIAamp DNA Blood Midi (QIAGEN) theo hướng dẫn của nhà sản xuất. Sau đó tiến hành tinh sạch (TS) bằng hạt từ PB (PacBio) để loại các đoạn ngắn và các thành phần ngoài DNA.

Dựa trên kết quả điện di trên gel agarose cho thấy mẫu DNA tách chiết được có kích thước nằm trong khoảng 10-15 kb (Hình 1a). Sau khi tinh sạch bằng PB bead, vạch DNA (Sau TS) thu hẹp và phần kích thước nhỏ đã giảm bớt do đã bị loại trong quá trình tinh sạch. Điều này hoàn toàn trùng khớp với kết quả điện di (Hình 1b) và điện tâm đồ trên Bioanalyzer (Hình 1c). Để biết chính xác nồng độ DNA thu được, nghiên cứu tiến hành đo nồng độ DNA sử dụng chất nhuộm huỳnh quang Qubit. Kết quả nồng độ DNA của mẫu được thể hiện ở Bảng 1.



Hình 1. Hình ảnh DNA tổng số trên a) Agarose gel 1%, maker 1kbp, b) Bản gel và c) Điện tâm đồ trên Bioanalyzer điện di sử dụng thang chuẩn ADN thuộc bộ ADN 12000 HS

(Ghi chú: Trước TS: Mẫu DNA tách chiết được. Sau TS: Mẫu DNA sau tinh sạch bằng PB bead)

Bảng 1. Kết quả đo nồng độ mẫu DNA tổng số bằng Qubit

Mẫu	Trước TS	Sau TS
Nồng độ (ng/ul)	150	62,35

Mẫu DNA sau tinh sạch có kích thước từ 10-15 kb, nồng độ 62,35 ng/ul (Bảng 1) là đủ điều kiện cho chuẩn bị thư viện cho giải trình tự PacBio.

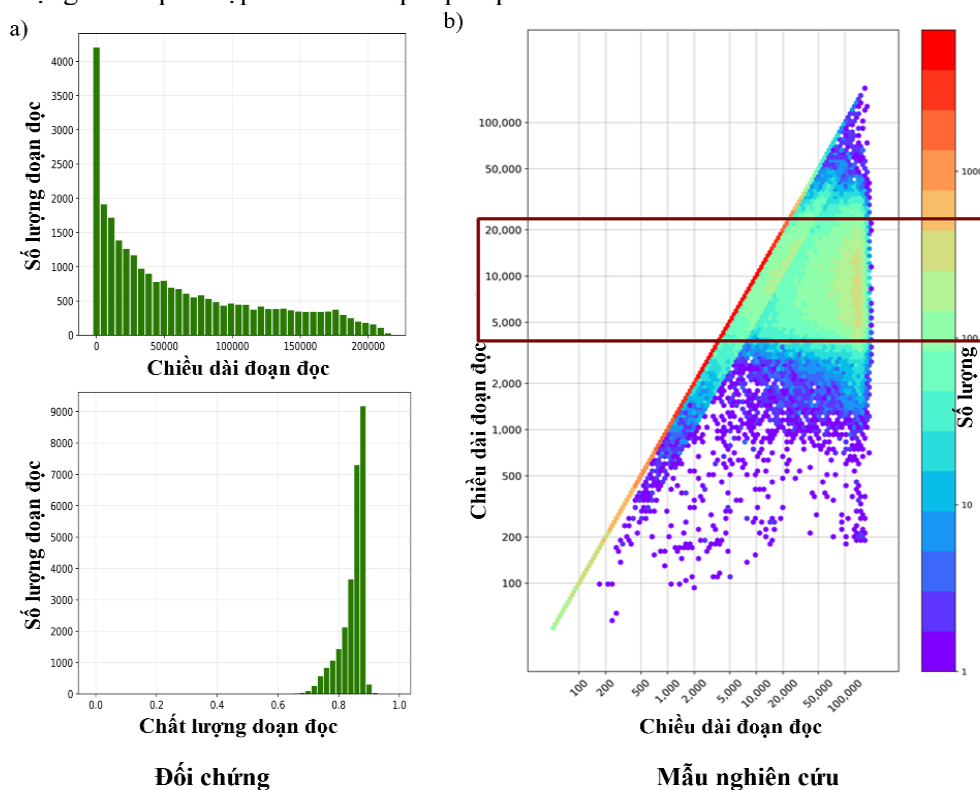
3.2. Kết quả chuẩn bị thư viện SMRTbell và giải trình tự PacBio

Bảng 2. Kết quả chuẩn bị thư viện và giải trình tự SMRT

Thư viện SMRTbell	Thông số
Nồng độ (ng/ul)	3,13
Nồng độ (nM/ul)	0,48
Kích thước trung bình	10.000 bp
Hiệu suất	82%
Thư viện tải SMRTcell	Thông số
Nồng độ	0,8 ng/uL
Chất lượng giải trình tự	Thông số
Chiều dài đoạn đọc trung bình	11.727 bp
Chất lượng giải trình tự	Q36

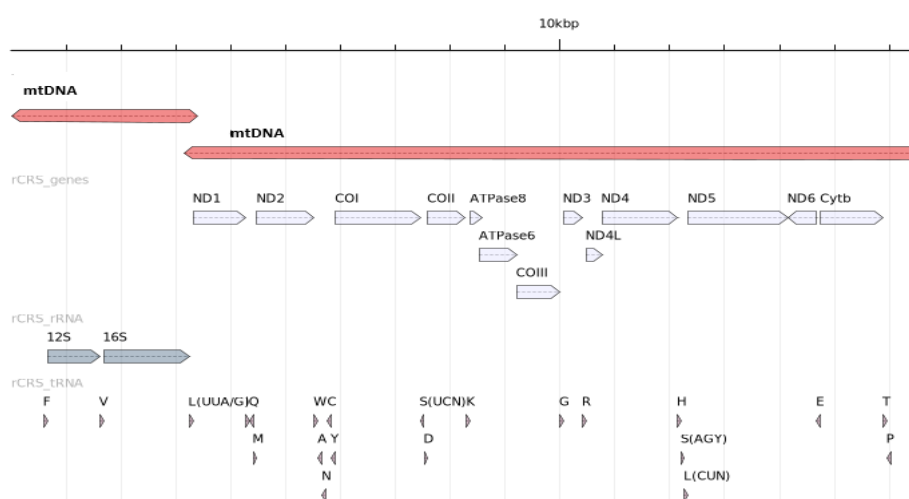
Mặc dù, một số hệ thống giải trình tự đoạn dài như Pacbio hay Oxford Nanopore vẫn còn hạn chế về độ chính xác khi so sánh với các hệ thống giải trình tự đoạn ngắn. Tuy nhiên, gần đây PacBio đã phát triển một phương pháp gọi là giải trình tự đồng thuận vòng tròn (CCS - circular consensus sequencing) [6], [7] nhằm đạt được tỷ lệ lỗi trình tự tương đương với Illumina. Phương pháp này tạo ra một mẫu “SMRTbell” bằng cách gắn các adapter ssDNA vào dsDNA đích, cho phép polymerase giải trình tự trên từng sợi của dsDNA đích nhiều lần. Quá trình này dẫn đến một đoạn đọc dài liên tục (CLR - continuous long read) bao gồm nhiều đoạn đọc con của trình tự mục tiêu [7]. Do đó, để cải thiện độ chính xác của trình tự lên tới 99%, nghiên cứu đã chuẩn bị thư viện

SMRTbell với hiệu suất 82% (Bảng 2) và tiến hành giải trình tự PacBio CCS giúp thu được đầy đủ độ dài trình tự tối đa của công nghệ PacBio CCS mà không ảnh hưởng đến chất lượng trình tự. Dữ liệu trình tự thô đã được tạo ra với độ dài đoạn đọc trung bình là 11,727 kb và chất lượng Q36 gần tương đương chất lượng đoạn đọc của giải trình tự Illumina (Bảng 2, Hình 2). Đây là dữ liệu có chất lượng tốt và phù hợp cho bước lắp ráp tiếp theo.



Hình 2. Chất lượng giải trình tự. a) Phân bố chiều dài đoạn đọc và chất lượng đoạn đọc của mẫu đối chứng nội kiểm (internal control). b) Phân bố chiều dài đoạn đọc và số lượng đoạn đọc của mẫu nghiên cứu

3.3. Kết quả lắp ráp hệ gen ty thể từ dữ liệu giải trình tự đoạn dài



Hình 3. So sánh trình tự mtDNA lắp ráp được của mẫu nghiên cứu với trình tự tham chiếu rCRS (NC012920)

Mặc dù hệ gen ty thể thường có cấu trúc dạng vòng, chứa nhiều nhánh phức tạp và các trình tự lặp lại, việc áp dụng chiến lược giải trình tự HiFi CCS kết hợp với lắp ráp từ các đoạn đọc dài (lên đến 18 kb) đã cho phép thu nhận hệ gen ty thể ở dạng thẳng (linear) (Hình 3), với tổng chiều dài 16.809 bp và hàm lượng GC đạt 44,4%. Đặc tính của hệ gen ty thể mà nghiên cứu xây dựng được là hoàn toàn phù hợp với các công bố trước đó về hệ gen ty thể ở người về chiều dài hoặc tính ổn định của thành phần GC [1], [2].

Kết quả so sánh trình tự DNA ty thể đã lắp ráp với trình tự tham chiếu cho thấy mức độ tương đồng gần như hoàn toàn, không ghi nhận sự thay đổi cấu trúc rõ rệt. Các kết quả giải trình tự và lắp ráp thu được cho thấy phương pháp này là một hướng tiếp cận hiệu quả và đơn giản trong việc tách chiết và xây dựng trình tự DNA ty thể ở người.

4. Kết luận

DNA tổng số được tách chiết bằng phương pháp đơn giản sử dụng bộ kit Blood & Cell Culture DNA Midi (QIAGEN) đạt chất lượng cao để tiến hành giải trình tự bằng công nghệ PacBio CCS (circular consensus sequencing) với đoạn đọc dài. Dữ liệu thu được từ hệ gen ty thể có độ dài đoạn đọc trung bình là 11.727 kb và chất lượng Q36, tương đương với chất lượng của các đoạn đọc từ công nghệ giải trình tự Illumina. Quá trình lắp ráp tạo ra hệ gen ty thể hoàn chỉnh với kích thước 16.809 bp và hàm lượng GC đạt 44,4%, tương đương với các hệ gen ty thể đã công bố trước đó. Nghiên cứu này cho thấy tiềm năng ứng dụng công nghệ giải trình tự đoạn dài trong việc xác định trình tự mtDNA, phục vụ cho các nghiên cứu về bệnh học, tiến hóa, dân số học và đa dạng sinh học ở các loài sinh vật nhân chuẩn.

Lời cảm ơn

Nghiên cứu được hỗ trợ bởi Dự án có mã số CT0000.09/21-22 của Viện Hàn lâm Khoa học và Công nghệ Việt Nam, do GS.TS. Trương Nam Hải chủ trì.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] P. F. Chinnery, "Mitochondrial DNA in Homo Sapiens," in *Human Mitochondrial DNA and the Evolution of Homo sapiens, Nucleic Acids and Molecular Biology*, H. J. Bandelt, V. Macaulay, M. Richards, Ed., vol. 18. Springer, Berlin, Heidelberg, 2006, pp. 3-15.
- [2] A. Piovesan, M. C. Pelleri, F. Antonaros *et al.*, "On the length, weight and GC content of the human genome," *BMC Res Notes*, vol. 12, no. 106, pp. 1-7, 2019.
- [3] K. Ye, J. Lu, F. Ma, A. Keinan, and Z. Gu, "Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals," *Proceedings of the National Academy of Sciences*, vol. 111, no. 29, pp. 10654-10659, 2014.
- [4] C. L. Hernández, "Mitochondrial DNA in Human Diversity and Health: From the Golden Age to the Omics Era," *Genes (Basel)*, vol. 14, no. 8, 2023, Art. no. 1534.
- [5] A. Sobenin, K. Y. Mitrofanov, A. V. Zhelankin, M. A. Sazonova, A. Y. Postnov, V. V. Revin, Y. V. Bobryshev, and A. N. Orekhov, "Quantitative assessment of heteroplasmy of mitochondrial genome: perspectives in diagnostics and methodological pitfalls," *BioMed Research International*, vol. 2014, pp. 1-9, 2014.
- [6] Y. Lin, J. Wang, R. Xu, Z. Xu, Y. Wang, S. Pan, Y. Zhang, Q. Tao, Y. Zhao, C. Yan, Z. Cao, and K. Ji, "HiFi long-read amplicon sequencing for full-spectrum variants of human mtDNA," *BMC Genomics*, vol. 25, no. 538, pp. 1-12, 2024.
- [7] A. M. Wenger, P. Peluso, W. J. Rowell *et al.*, "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome," *Nature Biotechnology*, vol. 37, pp. 1155–1162, 2019.
- [8] D. D. La, T. H. Dinh, and D. T. Nguyen, "Genetic variation of the D-loop region of the Kinh ethnic group from the northern, central, and southern Vietnam," *Academia Journal of Biology*, vol. 47, no. 1, pp. 75–85, 2025.