BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Lã Đức Duy

PHÂN TÍCH TOÀN BỘ HỆ GEN TY THỂ VÀ ĐA HÌNH NUCLEOTIDE ĐƠN VÙNG KHÔNG TRAO ĐỔI CHÉO CỦA NHIỄM SẮC THỂ Y NGƯỜI VIỆT NAM THUỘC NĂM DÂN TỘC PA KÔ, CƠ-TU, RƠ-MĂM, KINH MIỀN TRUNG VÀ KINH MIỀN NAM

LUẬN VĂN THẠC SĨ SINH HỌC

Hà Nội – 2024

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Lã Đức Duy

PHÂN TÍCH TOÀN BỘ HỆ GEN TY THỂ VÀ ĐA HÌNH NUCLEOTIDE ĐƠN VÙNG KHÔNG TRAO ĐỔI CHÉO CỦA NHIỄM SẮC THỂ Y NGƯỜI VIỆT NAM THUỘC NĂM DÂN TỘC PA KÔ, CƠ-TU, RƠ-MĂM, KINH MIỀN TRUNG VÀ KINH MIỀN NAM

LUẬN VĂN THẠC SĨ SINH HỌC

Ngành: Sinh học thực nghiệm

Mã số: 8420114

NGƯỜI HƯỚNG DẪN KHOA HỌC:
PGS. TS. Nguyễn Thùy Dương

Hà Nội – 2024

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận văn này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu dưới sự hướng dẫn của PGS. TS. Nguyễn Thuỳ Dương. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, các số liệu, kết quả nêu trong luận văn là mới, trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác.

Tác giả luận văn ký và ghi rõ họ tên

Lã Đức Duy

LỜI CẢM ƠN

Trước hết, tôi xin được bày tỏ sự kính trọng và biết ơn sâu sắc nhất đối với PGS. TS. Nguyễn Thùy Dương, Trưởng phòng phòng Hệ gen học người, Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Trong quá trình học tập và làm việc tại phòng, cô đã hướng dẫn, tạo điều kiện cũng như giúp đỡ tôi rất nhiều.

Tiếp theo, tôi xin cảm ơn các đồng nghiệp của tôi tại phòng Hệ gen học người, Viện Nghiên cứu hệ gen vì đã tận tình giúp đỡ trong quá trình học tập và làm việc.

Tôi xin cảm ơn Ban Lãnh Đạo, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ, cũng như các thầy cô giáo đã tận tình giảng dạy và tạo điều kiện thuận lợi để tôi có thể hoàn thành được luận văn này.

Cuối cùng, tôi xin được bày tỏ sự biết ơn đối với bố mẹ tôi, đặc biệt là mẹ tôi vì họ đã cho tôi những lời khuyên chân thành, đồng thời cũng động viên tôi khi tôi gặp khó khăn trong quá trình học tập và làm việc. Tôi cũng rất biết ơn sự kiên trì của họ vì không phải lúc nào tôi cũng phải là đứa con dễ bảo.

Ngoài ra, luận văn được thực hiện dưới sự hỗ trợ kinh phí của đề tài nghiên cứu độc lập cấp Quốc gia "Xây dựng cơ sở dữ liệu hệ gen biến thể ty thể và nhiễm sắc thể Y của một số dân tộc người Việt Nam" chủ nhiệm bởi PGS. TS. Nguyễn Thuỳ Dương trong khoảng thời gian 2019 – 2024.

Tác giả luận văn ký và ghi rõ họ tên

Lã Đức Duy

Mục lục

Danh mục các ký hiệu, các chữ cái viết tắt

Chữ viết tắt	Tiếng Anh	Tiếng Việt
AMOVA	Analysis of Molecular Variance	Phân tích phương sai phân tử
BTB		Bắc Trung Bộ
CA	Correspondence analysis	Phân tích tương quan
ĐBB		Đông Bắc Bộ
ĐBSH		Đồng Bằng sông Hồng
ĐNA		Đông Nam Á
ĐNB		Đông Nam Bộ
H	Haplotype diversity	Đa dạng kiểu đơn bội
Indel	Indel	Biến thể thêm mất đoạn
KMB		Kinh miền Bắc
KMN		Kinh miền Nam
KMT		Kinh miền Trung
Kya	Kilo years ago	Nghìn năm trước
MDS	Non-metric multidimensional scaling	
MPD	Mean pairwise difference	Trung bình số lượng nucleotide khác nhau theo từng cặp
MSEA	Mainland Southeast Asia	Đông Nam Á lục địa
mtDNA	Mitochondrial DNA	Hệ gen ty thể

NGS	Next generation sequencing	Giải trình tự hệ gen thế hệ mới
NHHM		Ngữ hệ Hmông-Miền
NHHT		Ngữ hệ Hán Tạng
NHNA		Ngữ hệ Nam Á
NHNĐ		Ngữ hệ Nam Đảo
NHTK		Ngữ hệ Tai-Kadai
NRY	Non-recombining Y chromosome	Vùng không trao đổi chéo của nhiễm sắc thể Y
NST		Nhiễm sắc thể
NTB		Nam Trung Bộ
rARN		ARN ribosome
RSRS	Reconstructed Sapiens Reference Sequence	
SNP	Single nucleotide polymorphism	Đa hình nucleotide đơn
SPRI	Solid Phase Reversible Immobilization	
STR	Short tandem repeat	Đoạn lặp ngắn
tARN		ARN vận chuyển
TBB		Tây Bắc Bộ
TN		Tây Nguyên
π	Nucleotide diversity	Đa dạng nucleotide

Danh mục bảng

Danh mục hình

MỞ ĐẦU

Hệ gen ty thể và vùng không trao đổi chéo trên nhiễm sắc thể (NST) Y là hai chỉ thị ADN được sử dụng rộng rãi để nghiên cứu cấu trúc di truyền giữa các quần thể khác nhau, nghiên cứu di truyền quần thể, nghiên cứu về đa hình cá thể phục vụ cho khoa học giám định hình sự và pháp y. Hệ gen ty thể được lựa chọn vì có tốc độ tiến hóa nhanh hơn so với ADN nhân, gen trong hệ gen ty thể có nhiều bản sao hơn so với chỉ hai bản sao của gen trong ADN nhân và chỉ di truyền theo dòng mẹ. Giống với hệ gen ty thể, nhiễm sắc thể Y là một locus đơn bội, di truyền theo dòng bố (từ cha sang con trai), phù hợp với các nghiên cứu về lịch sử dòng bố của các quần thể.

Việt Nam rất đa dạng về nhóm dân tộc với 54 dân tộc anh em, được phân chia theo 5 ngữ hệ khác nhau ngữ hệ Nam Á (NHNA), Nam Đảo (NHNĐ), Tai-Kadai (NHTK), Hán-Tạng (NHHT) và Hmông-Miền (NHHM), trong đó ngữ hệ Nam Á chiếm khoảng 90% dân số và phân bố khắp cả nước. Phần lớn người nói ngữ hệ Nam Á thuộc dân tộc Kinh (85,3%) và 4,7% còn lại thuộc 24 dân tộc thiểu số phân bố trải dài khắp cả nước với nhiều dân tộc bị thu hẹp quy mô hoặc sống ở những vùng ẩn dật. Do đó, những dân tộc sống tách biệt này có thể mang một số thông tin di truyền đặc trưng cho mỗi dân tộc, mà không có ở dân tộc Kinh. Hầu hết các nghiên cứu về hệ gen ty thể hoàn chỉnh và nhiễm sắc thể Y của quần thể người ở Việt Nam đều tập trung vào dân tộc Kinh. Ngoài ra, nghiên cứu đa dạng di truyền sử dụng hệ gen ty thể hoàn chỉnh và nhiễm sắc thể Y dòng bố và dòng mẹ trên ngữ hệ Nam Á được thực hiện ở sáu dân tộc thiểu số (Mảng, Churu, Eđê, Gia-rai, Raglay, Chăm), trong đó có năm dân tộc (Churu, Eđê, Gia-rai, Raglay, Chăm) ở khu vực Tây Nguyên và một dân tộc (Mảng) ở khu vực Đông Bắc Bộ của Việt Nam. Như vậy, nghiên cứu hệ gen ty thể và NST Y trên các dân tộc thuộc ngữ hệ Nam Á còn rất hạn chế. Xuất phát từ lý do trên, chúng tôi thực hiện nghiên cứu "Phân tích toàn bộ hệ gen ty thể và đa hình nucleotide đơn vùng không trao đổi chéo của nhiễm sắc thể Y người Việt Nam thuộc năm dân tộc Pa Kô, Cơ-tu, Rơ-măm, Kinh miền Trung và Kinh miền Nam ".

Mục tiêu đề tài

Xác định và phân tích được toàn bộ hệ gen ty thể của người Việt Nam thuộc năm dân tộc Pa Kô, Cơ-tu, Rơ-măm, Kinh miền Trung và Kinh miền Nam

Phân tích được các đa hình nucleotide đơn từ vùng không trao đổi chéo của nhiễm sắc thể Y của người Việt Nam thuộc năm dân tộc nêu trên.

Nội dung nghiên cứu

1. Giải mã trình tự toàn bộ hệ gen ty thể của nam giới thuộc năm dân tộc Pa Kô, Cơ-tu, Rơ-măm, Kinh miền Trung và Kinh miền Nam

2. Phân tích các biến thể trên hệ gen ty thể của năm dân tộc trong nghiên cứu và so sánh các biến thể giữa các dân tộc thuộc các nhóm ngữ hệ khác nhau ở Việt Nam

3. Phân tích đa dạng di truyền quần thể dựa vào hệ gen ty thể của các dân tộc Việt Nam và các dân tộc ở châu Á (đa dạng kiểu đơn bội, đa dạng nucleotide, tỉ lệ tương đồng kiểu đơn bội giữa các quần thể, khoảng cách di truyền quần thể, xác định các yếu tố ảnh hưởng đến cấu trúc di truyền quần thể)

4. Phân tích các đa hình nucleotide đơn từ nhiễm sắc thể Y của một số nam giới từ mỗi dân tộc thuộc năm dân tộc người Việt Nam nêu trên và so sánh các đa hình nucleotide đơn giữa các dân tộc thuộc các nhóm ngữ hệ khác nhau ở Việt Nam và giữa dân tộc ở các châu lục khác nhau trên thế giới

Chương 1. TỔNG QUAN NGHIÊN CỨU

1.1. Hệ gen ty thể và nhiễm sắc thể Y

1.1.1. Hệ gen ty thể

mtDNA là một phân tử dạng vòng (khác với nhiễm sắc thể dạng thẳng trong nhân) và khá nhỏ gọn. Nó chỉ có khoảng 16.500 base (so với khoảng 3,2 tỷ base trong bộ gen nhân đơn bội) và chỉ chứa 37 gen: 13 gen mã hóa protein, 2 gen ARN ribosome (rARN) và 22 gen ARN vận chuyển (tARN). 13 gen mã hóa protein bao gồm ba tiểu đơn vị của cytochrome oxidase, hai tiểu đơn vị của F1-ATPase, bảy tiểu đơn vị của NADH–dehydrogenase và gen của cytochrome B. Tất cả những gen này đều liên quan đến chức năng chính của ty thể, đó là thực hiện hô hấp tế bào rồi chính quá trình này lại liên quan đến việc sản xuất năng lượng từ các chất chuyển hóa [1]. Tất cả các polypeptide được mã hóa bởi các gen mtDNA kết hợp với các polypeptide khác được mã hóa bởi ADN nhân để tạo thành các phức hợp protein liên quan đến quá trình sản xuất năng lượng. Hơn nữa, tất cả các protein cần thiết để sao chép mtDNA và thực hiện phiên mã, xử lý và dịch mã ARN thông tin từ mtDNA cũng được mã hóa bởi ADN nhân, vì vậy tất cả các protein ty thể được mã hóa bởi nhân phải được vận chuyển vào ty thể.

Một trong những nguyên nhân khiến mtDNA được lựa chọn cho các nghiên cứu nhân chủng học phân tử là tốc độ tiến hóa nhanh gấp khoảng 10 lần so với ADN nhân [2] mặc dù tính "gọn nhẹ" mtDNA (các gen sẽ nằm ngay cạnh nhau mà không có khoảng cách giữa các gen do mtDNA không có intron và vùng điều khiển hầu hết nằm ở vùng điều khiển hay còn gọi là vùng D-loop) trên lý thuyết sẽ giúp nó chống chịu được trước áp lực chọn lọc tiến hóa. Nguyên nhân của hiện tượng này hiện vẫn chưa được giải thích rõ ràng [3], nhưng đặc điểm này của mtDNA giúp các nhà nghiên cứu có thể thu được nhiều biến thể ADN hơn so với việc phân tích ADN nhân (do tỉ lệ đột biến ở mtDNA cao hơn nên nếu phân tích cùng lượng base thì phân tích mtDNA sẽ có nhiều biến thể hơn).

Một đặc điểm hữu ích khác của mtDNA là nó có nhiều bản sao trên mỗi tế bào - trung bình ty thể có 5 - 10 bộ gen mtDNA, và trung bình tế bào có vài trăm đến vài nghìn ty thể - so với chỉ hai bản sao của bất kỳ gen nào trong ADN nhân [4]. Điều này khiến mtDNA trở thành bộ gen phù hợp để phân tích ADN

từ các mẫu vật cổ xưa, cũng như từ một số loại mẫu vật pháp y (xương cũ, tóc, xác bị cháy, v.v.) bởi vì đối với mẫu vật có khả năng có rất ít (nếu có) ADN còn sót lại, việc mtDNA tồn tại ở nhiều bản sao hơn ADN nhân làm tăng đáng kể cơ hội thu được ADN.

Đặc điểm cuối cùng làm mtDNA được lựa chọn trong các nghiên cứu nhân chủng học phân tử là mtDNA chỉ di truyền theo dòng mẹ. Hiện tượng này được phát hiện đầu tiên bởi hai nghiên cứu độc lập trong thập kỉ 70 [5,6] và được giải thích là do sự loại trừ mtDNA từ bố trong trứng trong quá trình thụ tinh. Sau đấy, một giả thuyết khác về hiện tượng này cho rằng bởi vì số lượng bản sao của mtDNA từ mẹ lớn hơn rất nhiều so với bố nên trong một vài chu kì đầu của quá trình phân chia tế bào sau khi thụ tinh, khoảng thời gian không có sự tái bản của mtDNA, sự chênh lệch về số lượng bản sao dẫn đến sự chênh lệch về tỉ lệ mtDNA được phân chia vào các tế bào giữa hai giới. Để kiểm chứng giả thuyết này, nhiều nghiên cứu đã được thực hiện [7,8] bằng việc lai con cái từ một loài với con đực từ loài thứ hai, lấy con cái ở đời con và lai ngược chúng với con đực từ loài thứ hai và lặp lại quá trình này trong nhiều thế hệ để tạo nên một lượng mtDNA từ bố đủ để phát hiện và sử dụng loài khác nhau để dễ dàng phân biệt. Mặc dù những thí nghiệm ở trên có tìm thấy mtDNA của con đực trên thế hệ sau, khi lặp lại thí nghiệm trên cùng một loài, kết quả lại không tìm thấy mtDNA đực [9]. Hiện tượng này được giải thích là do khi lai khác loài, cơ chế loại bỏ mtDNA từ con đực đã bị làm hỏng [9], vì vậy thí nghiệm lai khác loài ở trên đã chứng minh rằng cơ chế đằng sau di truyền theo dòng mẹ của mtDNA chịu sự ảnh hưởng của yếu tố di truyền [3]. Cho đến nay, chỉ có duy nhất một trường hợp di truyền mtDNA theo dòng bố ở người được tìm thấy [10]. Vì vậy, mtDNA không có sự xuất hiện của hiện tượng tái tổ hợp nên sự khác duy nhất giữa các biệt trình tự mtDNA khác nhau là các đột biến và số lượng đột biến chênh lệch giữa hai trình tự mtDNA bất kì phản ánh khoảng cách giữa chúng và tổ tiên gần nhất.

Tuy nhiên, việc mtDNA là đơn bội khiến nó dễ bị ảnh hưởng bởi các yếu tố ngẫu nhiên và chọn lọc tự nhiên. Do đó, lịch sự của mtDNA chưa chắc đã phản ánh chính xác lịch sử của một quần thể hoặc một loài bởi vì có xác suất xảy ra hiện tượng trôi dạt di truyền hoặc các yếu tố ngẫu nhiên khác, hoặc mtDNA đã chịu ảnh hưởng của chọn lọc tự nhiên. Để đưa ra kết luận chính xác

về lịch sử của quần thể, việc nghiên cứu các locus di truyền độc lập với nhau là cần thiết [3].

1.1.2. Nhiễm sắc thể Y

NST Y chỉ tồn tại ở các cá thể nam và được di truyền từ cha sang con, cho nên việc phân tích các biến thể trên NST này sẽ giúp tìm hiểu về lịch sử dòng bố của các quần thể. Một số vùng của NST Y có thể bắt cặp và tái tổ hợp với NST X trong quá trình giảm phân, và chúng được gọi là vùng giả NST thường. Vùng giả NST thường 1 và vùng giả NST thường 2 của NST Y là các vùng tương đồng ngắn giữa NST X và Y ở động vật có vú, trong đó vùng giả NST thường 1 nằm ở đầu cánh tay p và vùng giả NST thường 2 nằm ở đầu cánh tay q. Do sự khác biệt trong trình tự ADN của nhiễm sắc thể X và Y, chúng không trải qua quá trình bắt cặp trong quá trình giảm phân, ngoại trừ các vùng trong vùng giả NST sẽ bắt cặp và tái tổ hợp với các vùng giả NST trong NST X trong quá trình giảm phân. Tuy nhiên, sự tái tổ hợp và bắt cặp của các vùng giả NST có sự khác biệt về mặt thời gian và di truyền so với các phần khác của bộ gen [11]. Sự hình thành và bắt cặp của mạch kép gãy của vùng giả NST xảy ra muộn hơn so với các NST thường bởi vì các vùng giả NST chỉ bắt đầu quá trình làm gãy mạch kép sau khi tất cả các mạch kép gãy của nhiễm sắc thể thường đã được sửa chữa. Mặc dù vậy, tỉ lệ trao đổi chéo ở vùng giả NST 1 vẫn xảy ra cao hơn so với tỉ lệ trao đổi chéo ở NST thường [12]. Sự suy giảm về tỉ lệ này ở vùng giả NST 1 đã được phát hiện có liên quan đến việc gia tăng tần suất của hiện tượng bất thường về số lượng NST giới tính trong tinh trùng, dẫn đến hội chứng Turner hoặc Kleinfelter ở đời con [13,14]. Phần còn lại của NST Y không trải qua quá trình tái tổ hợp nên thường được gọi là vùng không trao đổi chéo của NST Y (non-recombining Y chromosone - NRY). Về mặt di truyền tế bào học, vùng này được chia thành hai vùng bao gồm vùng dị sắc và vùng euchromatin. Vùng dị sắc của NST Y nằm ở vùng gần đầu của cánh tay Yq chứa hai họ trình tự lặp lại cao, DYZ1 và DYZ2 [15]. Sự thay đổi về kích thước của vùng dị sắc ở cánh tay dài NST Y trong một cá thể đã được báo cáo [16] tuy nhiên ý nghĩa lâm sàng của nó hiện vẫn chưa được giải đáp. Vùng euchromatin bao gồm vùng cận tâm động ở cánh tay ngắn và dài của Y. Trước đây, vùng này được cho là một vùng không có nhiều ý nghĩa chức năng, tuy nhiên sự phát triển của khoa học trong hai thập kỉ gần đây đã phát hiện được

vai trò của vùng euchromatin đối với sự phát triển của các mô và trong người trưởng thành [17]. NRY có chứa một số vùng trình tự đặc trưng, có thể dễ dàng phân tích bằng nhiều phương pháp khác nhau, cũng như nhiều vùng có nhiều bản sao của các đoạn lặp lại dài. Ở thời kì đầu, NRY được cho là không có nhiều biến đổi ở trên con người [18], tuy nhiên với sự phát triển của các kĩ thuật phát hiện các điểm đa hình nucleotide đơn (single nucleotide polymorphism - SNP), hàng loạt các SNP marker đã được tạo ra [19]. Công nghệ giải trình tự hệ gen thế hệ mới còn cho phép đọc trình tự của một đoạn dài của NRY để có thể phân tích chuyên sâu hơn [20]. Giống như mtDNA, NRY là một locus đơn bội, có chung điểm mạnh cũng như điểm yếu của chỉ thị di truyền theo dòng mẹ.

1.2. Tổng quan về ngữ hệ Nam Á

1.2.1. Nguồn gốc của ngữ hệ Nam Á

Đông Nam Á (ĐNA) là nơi giao thoa của sự đa dạng dân tộc ngôn ngữ được hình thành bởi nhiều sự kiện nhân khẩu học, bắt đầu với sự xuất hiện đầu tiên của con người hiện đại về mặt giải phẫu ít nhất là 65 nghìn năm trước (kilo years ago - kya) [21,22]. Ở Đông Á, dấu vết đầu tiên của loài người tại đây cũng xuất hiện từ rất sớm, ít nhất 45 kya [23]. Các bằng chứng khảo cổ học đã chỉ ra hai luồng di chuyển chính về phía nam của nền văn hóa nông nghiệp, với một hướng đi vào ĐNA còn hướng còn lại đi về phía Đài Loan và tiếp tục mở rộng tạo thành ngữ hệ Nam Đảo (NHNĐ) ngày nay [24].

Trong năm ngữ hệ chính hiện tại ở ĐNA, ngữ hệ Nam Á (NHNA) là ngữ hệ xuất hiện đầu tiên, tuy nhiên nguồn gốc của ngữ hệ này hiện vẫn là vấn đề đang được bàn cãi. Các nhà khảo cổ học đã đưa ra hai giả thuyết về nguồn gốc của nền văn hóa nông nghiệp ở ĐNA, trong đó một giả thuyết đề cập đến những người nông dân sau khi di cư từ Đông Á đã tương tác với người săn bắt hái lượm bản địa còn giả thuyết còn lại cho rằng nền văn hóa nông nghiệp được phát tán do người bản địa tự phát triển hoặc học từ người nông dân ở Đông Á. Hiện tại, ngày càng có nhiều bằng chứng ủng hộ giả thuyết đầu tiên [25,26]. Với sự phát triển của công nghệ giải trình tự gen thế hệ mới (next generation sequencing - NGS), các nghiên cứu về ADN từ mẫu vật cổ đại đã phát hiện rằng người săn bắt hái lượm Hòabìnhian từ Lào và Malaysia (mẫu xuất hiện lần lượt từ ~7,8 - 8,0 kya và ~4,2 - 4,4 kya) có quan hệ gần gũi nhất với người

bản địa hiện tại từ ĐNA và Nam Á, trong khi các cá thể từ thời đồ đá mới từ vùng lục địa Đông Nam Á (mainland Southeast Asia - MSEA) xuất hiện từ ~4 kya có thể là kết quả của sự pha trộn giữa những người săn bắn hái lượm Hòabìnhian và những người nông dân đầu tiên từ Trung Quốc [27,28]. Việc những cá thể từ thời kì đồ đá mới này có chung tổ tiên với các dân tộc nói ngôn ngữ thuộc NHNA từ MSEA [27,28] chỉ ra rằng sự lan truyền của NHNA có thể liên quan đến sự mở rộng của nông nghiệp vào MSEA.

ADN cổ đại từ các cá thể MSEA xuất hiện từ ~2 kya có mối quan hệ với văn hóa thời đại đồ đồng và từ các mốc thời gian tiếp đó như thời kì đồ sắt và các mốc lịch sử gần hơn, có thêm sự pha trộn tổ tiên mới từ Đông Á so với các cá thể từ thời kì đồ đá mới [27–29]. Phần lớn cấu trúc của các quần thể hiện nay ở MSEA được hình thành từ kết quả những làn sóng di cư đề cập ở trên, vì những cá thể cổ đại từ thời điểm này bắt đầu giống hơn về mặt di truyền với quần thể MSEA hiện tại từ cùng khu vực.

1.2.2. Sự phân bố của ngữ hệ Nam Á

Ở thời điểm hiện tại, các dân tộc nói ngôn ngữ thuộc NHNA ở ĐNA sống phân bố rải rác ở các quốc gia trong khu vực ngoại trừ Việt Nam và Campuchia, nơi NHNA trở thành quốc ngữ. Hiện tượng này có thể là kết quả của sự xuất hiện và phát triển của các ngữ hệ khác [24]. Theo Ethnologue [30], NHNA, một trong những ngữ hệ chính trên thế giới, bao gồm 167 ngôn ngữ chia thành hai nhánh lớn Mon-Khmer và Munda, và phân bố ở Nam Á, Đông Á và ĐNA. Ở Nam Á, nơi các ngôn ngữ thuộc NHNA được phân tầng thành nhánh phụ Munic, ngữ hệ này là một trong bốn ngôn ngữ chính của miền nam và miền trung Ấn Độ và là ngôn ngữ thiểu số ở một số vùng của Bangladesh và Nepal [30]. Ở Đông Á, nó phân bố ở khu vực phía nam Trung Quốc, được gọi là tiếng Bolyu và tiếng Bagan [31]. Khu vực ĐNA được chia thành MSEA và vùng ĐNA hải đảo (island Southeast Asia - ISEA), trong đó NHNA chỉ được nói bởi một vài dân tộc thiểu số ở ISEA nơi NHNĐ được nói chủ yếu, còn ở MSEA ngoài Việt Nam và Campuchia, ngữ hệ này cũng được sử dụng bởi một số dân tộc thiểu số ở Thái Lan, Lào, Myanmar và vùng bán đảo của Malaysia. Với xấp xỉ 126 triệu người sử dụng NHNA [30], ngữ hệ này là ngữ hệ đứng thứ 8 về số lượng người nói trên thế giới và thứ 3 tại ĐNA, với 3/4 số lượng người nói là người Việt Nam (Tổng điều tra dân số và nhà ở năm 2019, www.gso.gov.vn),

tương đương với khoảng 90% dân số. Trong 25 dân tộc sử dụng NHNA tại Việt Nam, dân tộc Kinh chiếm khoảng 85% số lượng người sử dụng ngôn ngữ và 24 dân tộc thiểu số còn lại chiếm khoảng 5%. Ở vị trí thứ hai về số lượng người dùng trong danh sách các ngôn ngữ sử dụng NHNA, tiếng Khơ-me, quốc ngữ của Campuchia, được sử dụng bởi 15.900.000 người (~95 % dân số) [30].

Dân tộc Kinh thuộc nhánh con Viet-Muong của nhánh Mon-Khmer của NHNA [30] là dân tộc bản địa, đã sinh sống lâu đời tại Việt Nam và hiện có khoảng 82 triệu người (Tổng điều tra dân số và nhà ở năm 2019, www.gso.gov.vn), chiếm khoảng 85% tổng dân số cả nước. Người Kinh ban đầu sinh sống tại vùng Bắc Bộ và Bắc Trung Bộ, sau đó mở rộng khu vực cư trú đến các vùng khác, trở thành dân tộc có số lượng người lớn nhất và có mặt trên mọi địa bàn của Việt Nam. Tuy nhiên, tộc người này sống tập trung ở khu vực đồng bằng, chiếm tỉ lệ lớn thành phần cư dân của các đô thị phát triển trong cả nước.

Dân tộc Cơ-tu thuộc nhánh con Katuic của nhánh Mon-Khmer của NHNA [30] có khoảng 74.173 người (Tổng điều tra dân số và nhà ở năm 2019, www.gso.gov.vn), cư trú chủ yếu ở tỉnh Quảng Nam và Thừa Thiên Huế. Hiện nay, người Cơ-tu chưa hình thành các nhóm dân tộc theo địa phương mà phân thành các nhóm dựa vào khu vực cư trú như người vùng cao, người vùng trung và người vùng thấp [32].

Dân tộc Pa Kô thuộc nhánh con Katuic của nhánh Mon-Khmer của NHNA có khoảng 19.000 người [30], cư trú chủ yếu ở tỉnh Quảng Trị và Thừa Thiên Huế, gần biên giới với Lào. Ngoài Việt Nam, người Pa Kô còn sống ở Lào tại các tỉnh Salavan, Samouay, Savannahkhet và Xekong với dân số khoảng 23.000 người. Người Pa Kô ở Việt Nam chưa được coi là một dân tộc riêng mà mới đang được xếp vào dân tộc Tà Ôi, còn người Pa Kô ở Lào đã được coi là một dân tộc mà không xếp chung với người Tà Ôi ở đó.

Dân tộc Rơ-măm thuộc nhánh con Bahnaric của nhánh Mon-Khmer của NHNA [30] có khoảng 600 người, cư trú chủ yếu ở huyện Sa Thầy, tỉnh Kon Tum, Việt Nam. Người Rơ-măm là dân tộc có dân số ít thứ ba ở Việt Nam, chỉ cao hơn dân tộc Brâu và Ơ Đu [32]. Nguồn gốc xuất hiện của người Rơ-măm vẫn chưa được xác định rõ, tuy nhiên dân tộc này đã có mặt ở Việt Nam từ lâu.

1.3. Tình hình nghiên cứu hệ gen ty thể và nhiễm sắc thể Y về đa dạng di truyền quần thể của ngữ hệ Nam Á

1.3.1. Tình hình nghiên cứu quốc tế

1.3.1.1. Hệ gen ty thể

Việc các hai nhánh con lớn của NHNA là Mon-Khmer và Munda chỉ được sử dụng lần lượt ở ĐNA (và Trung Quốc) và Ấn Độ dẫn đến phần lớn các nghiên cứu mtDNA ở các dân tộc NHNA được thực hiện trên các cá thể ở MSEA [33–39] và Ấn Độ [40–42]. Dữ liệu mtDNA từ nhóm nói tiếng Munda thuộc NHNA từ Ấn Độ và nhóm nói tiếng Khasi–Aslian thuộc NHNA từ ĐNA đã thể hiện sự khác biệt về nguồn gốc tổ tiên dòng mẹ giữa hai nhóm này bởi vì hai nhóm đều có nguồn gốc tổ tiên gần hơn với các dân tộc lân cận [42]. Các dân tộc nói tiếng Munda có thành phần nhóm đơn bội phần lớn được tìm thấy ở các nhóm thuộc ngữ hệ Dravidia và Ấn-Âu [43,44], trong khi đó các nhóm nói ngôn ngữ thuộc nhánh Mon-Khmer có nhiều nhóm đơn bội đặc trưng cho quần thể Đông Á hơn [42] cho nên có thể phân biệt người NHNA ở Ấn Độ và ĐNA dựa vào nhóm đơn bội của họ [45]. Ở các quần thể MSEA, các bằng chứng về nguồn gốc từ Nam Á đã được tìm thấy thông qua việc phân tích các nhánh đơn bội cụ thể, tần suất của các nhóm đơn bội và khoảng cách di truyền giữa các dân tộc [36,46]. Ngoài ra, các nhóm người cổ từ thời đồ đá mới ở MSEA chịu ảnh hưởng các các làn sóng di cư từ Trung Quốc [27,28], dẫn tới nhóm đơn bội của các dân tộc NHNA ở khu vực MSEA có nhiều điểm tương đồng với các dân tộc ở Trung Quốc [47,48]. Ngoài các nghiên cứu quy mô lớn đã đề cập, một số nghiên cứu ở quy mô nhỏ hơn, tập trung vào sự đa dạng di truyền theo dòng mẹ ở các nhóm người nhất định cũng đã được thực hiện [38,49].

1.3.1.2. Nhiễm sắc thể Y

Các nghiên cứu sử dụng NST Y thường được thực hiện cùng với các nghiên cứu về dòng mẹ để phản ánh sự đa dạng về văn hóa và phong tục tập quán của các dân tộc [48]. Sự khác biệt về đặc điểm di truyền giữa hai chỉ thị giới tính cũng được quan sát ở nhiều nghiên cứu khác nhau, có thể dẫn tới những kết quả ủng hộ các sự kiện di cư thiên về một giới tính nhất định [36,42,48]. Kết quả quan sát thấy ở các bộ tộc NHNA sống ở đồi núi ở Thái Lan đã thể hiện rõ xu hướng này, trong đó các bộ tộc sống theo phong tục ở rể

sẽ có thông tin di truyền đa dạng hơn đối với NST Y so với các bộ tộc cư trú theo phụ hệ và xu hướng ngược lại được quan sát trên mtDNA [48]. Tuy nhiên, đặc điểm khác biệt giữa hai chỉ thị giới tính quan sát được trong nghiên cứu của Kutanan và cộng sự không được thể hiện rõ ràng, điển hình là dân tộc Htin thuộc NHNA đều có mức độ đa dạng di truyền thấp ở cả hai chỉ thị. Hiện tượng này thể hiện rằng các yếu tố khác cũng đóng vai trò đối với sự đa dạng di truyền sau hôn nhân [48]. Kutanan và cộng sự đã lí giải rằng hiện tượng này có thể là kết quả của sự cô lập về mặt địa lý dẫn tới hiện tượng trôi dạt di truyền và giao phối trong cùng bộ tộc, làm giảm sự đa dạng di truyền trong quần thể và làm tăng sự khác biệt giữa các quần thể với nhau. Điều này cũng nhất quán với các kết luận về sự ảnh hưởng của các yếu tố văn hóa đối với sự khác biệt về đa dạng di truyền giữa dòng bố và dòng mẹ của các nghiên cứu trước đây [50,51]. Ngoài ra, bằng việc sử dụng chỉ thị NST Y, nhóm tác giả trên cũng phát hiện được mối quan hệ di truyền gần hơn giữa dân tộc Mon thuộc NHNA với các dân tộc NHTK ở vùng trung tâm Thái Lan và với các dân tộc Ấn Độ, nhất quán với kết quả quan sát được khi sử dụng chỉ thị mtDNA [33,48]. Kết quả từ đa dạng di truyền trên NST Y cũng chỉ ra rằng nhóm người Mon cổ- hiện nay gọi là nhóm Nyahkur, không có chung kiểu đơn bội này với các nhóm Mon khác, củng cố giả thuyết là người Mon hiện tại ở Thái Lan là người tị nạn chính trị từ Myanmar vào khoảng thế kỉ 16 đến 19. Ngoài ra, tần suất xuất hiện cao của nhóm đơn bội NST Y O2a* và C* và mối quan hệ di truyền gần gũi của người Nyahkur với các dân tộc NHTK và NHHT cũng thể hiện sự khác biệt và dòng bố giữa họ và người Mon.

1.3.2. Tình hình nghiên cứu trong nước

1.3.2.1. Hệ gen ty thể

Ở thời kì đầu, các nghiên cứu về đa dạng di truyền quần thể người có sử dụng mẫu người Việt hoặc nghiên cứu về quần thể người Việt đều không ghi rõ tên dân tộc nghiên cứu mà chỉ đề cập là người Việt Nam [52–55]. Ở các nghiên cứu có xác định nhóm dân tộc, chỉ có dân tộc Kinh và Mường sử dụng ngôn ngữ thuộc NHNA được nghiên cứu [56–59]. Vào năm 2017, Pischedda và cộng sự đã thực hiện nghiên cứu phân tích vùng điều khiển của 622 cá thể người Việt thuộc 7 dân tộc khác nhau, trong đó chỉ có dân tộc Kinh (399 cá thể) thuộc NHNA [60]. Kết quả nghiên cứu cho thấy, các cá thể Việt Nam mang

các nhóm đơn bội Đông Nam Á, thể hiện sự phân tầng địa lý và dân tộc ở mức trung bình, trong khi đó dân tộc Mông khác biệt nhất so với các dân tộc còn lại. Hai nhánh mtDNA mới (M7b1a1f1 và F1f1) đã chỉ ra dòng gen giữa Việt Nam và các nước lân cận. Sự giảm dân số mạnh mẽ của quần thể người Chăm 700 năm trước phù hợp với quá trình Nam tiến từ nơi cư trú chính ở đồng bằng sông Hồng của người Việt. Không lâu sau đó, Macholdt và công sự đã giải trình tự toàn bộ mtDNA và một phần trình tự NRY khoảng 2,3 Mb của 600 cá thể nam người Việt từ 17 nhóm dân tộc thuộc 5 ngữ hệ chính tại Việt Nam [49]. Kết quả nghiên cứu tìm thấy sự đa dang di truyền độc lập cao giữa các nhóm mà không liên quan đến ngữ hệ hay phân bố địa lý của chúng. Cụ thể hơn, dân tộc Mảng và dân tộc Sila đã chịu ảnh hưởng của hiện tượng thắt cổ chai gần đây, trong khi đó, dân tộc Kinh có sự tương đồng lớn với các dân tộc khác. Sự đa dạng di truyền của hai dân tộc thuộc NHNĐ là Giarai và Êđê có xu hướng ngược lại đối với các quần thể khác, ám chỉ sự đa dạng di truyền của hai dân tộc này có khả năng đã chịu tác động bởi phong tục ở rể (đặc trưng của NHNĐ). Nhìn chung, nghiên cứu của Macholdt và cộng sự đã chỉ ra tình trạng sống cô lập, ít có sự tương tác của các nhóm dân tộc thiểu số là yếu tố chính ảnh hưởng đến cấu trúc di truyền của các quần thể người Việt, một phần lớn đa dạng di truyền của người Việt không được tìm thấy trên người Kinh mà ở các dân tộc thiểu số. Tương tự với nghiên cứu của Macholdt, Dương và cộng sự đã phân tích các nhóm đơn bội mtDNA của cùng các cá thể trong nghiên cứu của nêu trên để hiểu rõ hơn về lịch sử phân bố của mtDNA ở khu vực MSEA [61]. Nhóm nghiên cứu của Dương đã tìm thấy 399 kiểu đơn bội thuộc 135 nhóm đơn bội, trong đó có 111 dòng mẹ lần đầu được tìm thấy ở người Việt Nam. Tuy nhiên, ngoài dân tộc Kinh, dân tộc Mảng là dân tộc duy nhất thuộc NHNA được giải trình tự toàn bộ mtDNA. Trong cùng năm, nghiên cứu đầu tiên về trình tự vùng D-loop của dân tộc Mảng được thực hiện bởi Ngọc và cộng sự vào năm 2018 [62]. Không lâu sau, Hằng và cộng sự, một nhóm nghiên cứu độc lập so với nhóm của Dương, đã giải trình tự và phân tích vùng siêu biến 1 và 2 của 517 cá thể người Việt thuộc 4 dân tộc, trong đó có 3 dân tộc thuộc NHNA là Kinh, Mường và Khơ-me [63]. Trong 50 nhóm đơn bội được tìm thấy, tần suất nhóm đơn bội F1a cao nhất ở mức 15,7%, tiếp theo là B5a (10,8%), M (8,9%) và M7b1 (7,7%). Các SNP xuất hiện nhiều nhất trong nghiên cứu này là A263G (100%), A73G (99,6%), 315insC (96%), 309insC

(56%), C16223T (41%) và T16189C (39%). Sự đa dạng di truyền của 4 dân tộc ở mức 99,83% và xác suất trùng khớp ngẫu nhiên của hai cá thể có cùng trình tự mtDNA là 0,37%. Gần đây, Thảo và cộng sự đã thực hiện một nghiên cứu về mối quan hệ di truyền giữa năm dân tộc NHNĐ ở Việt Nam với các nhóm người Việt khác và với các dân tộc NHNĐ ở nước ngoài thông qua chỉ thị mtDNA và NST Y [64]. Kết quả của nghiên cứu trên chỉ ra rằng các dân tộc NHNĐ ở Việt Nam gần gũi về mặt di truyền đối với các dân tộc NHNA trong nước hơn khi so với mối quan hệ giữa các dân tộc này với các dân tộc thuộc cùng ngữ hệ ở khu vực khác, phản ánh sự khuếch tán văn hóa đóng vai trò chính trong sự lan truyền của các ngôn ngữ NHNĐ ở Việt Nam. Tuy nhiên, các phân tích dựa trên kiểu đơn bội cũng chỉ ra vai trò của sự khuếch tán vốn gen trong việc du nhập các ngôn ngữ NHNĐ vào Việt Nam. Ngoài dữ liệu mới về các dân tộc NHNĐ trong nước, dữ liệu mới về mtDNA và SNP ở trên NST Y của năm dân tộc NHNA cũng được công bố, mặc dù vậy nghiên cứu chỉ lấy mẫu các cá thể NHNA ở vùng lân cận với các dân tộc NHNĐ cho nên chưa thể hiện hết được sự đa dạng của các nhóm NHNA tại Việt Nam.

1.3.2.2. Nhiễm sắc thể Y

Giống với mtDNA, di truyền theo dòng bố ở quần thể người Việt Nam cũng được các nhà khoa học trên thế giới nghiên cứu từ sớm [65]. Vào năm 2009, nghiên cứu đầu tiên về sự phân bố các SNP của nhóm đơn bội C và O bởi nhóm tác giả người Việt đã được thực hiện bởi Tôn và cộng sự [66,67], tuy nhiên cho đến nay số lượng nghiên cứu đến từ nhóm tác giả trong nước về vấn đề này vẫn còn hạn chế. Các kết quả nghiên cứu bước đầu cho thấy nhóm đơn bội C3 của nhiễm sắc thể Y có tần số phân bố tương đối cao, trong khi đó nhóm đơn bội C và C2 có tần số phân bố trung bình thấp ở người Việt Nam [66]. Trong khi đó, 6 trên 10 SNP thuộc nhóm đơn bội O (M175, P186, P191, P196, M119, M50, M103, M110, SRY465 và JST022454) được tìm thấy trong tổng số 248 mẫu nghiên cứu. Nhóm đơn bội O1a*-M119 có tần số phân bố cao nhất là 62,9%. Nhóm đơn bội O1a2 (M50, M103 và M110) và O2b-SRY465 không phát hiện thấy đa hình trong các mẫu nghiên cứu [67]. Vào năm 2014, để xác định mối quan hệ di truyền giữa các quần thể người Đài Loan và các tộc người trong vùng đảo Đông Nam Á, Trejaut và cộng sự tiến hành thu thập 1658 mẫu máu từ các cá thể từ Việt Nam (người Kinh), Thái Lan, Fujian, Đài Loan (những

người Hán thuộc bộ lạc cư trú ở đồng bằng và 14 nhóm bản địa), Philippines và Indonesia. Kết quả nghiên cứu phát hiện 81 chỉ thị (chủ yếu là các SNP) và 17 đoạn lặp ngắn trên NST Y (short tandem repeats – STR) được tìm thấy. Trong khi sự khác biệt giữa các nhóm đơn bội O1a*-M119, O1a1*-P203, O1a2-M50 và O3a2-P201 giảm dần từ Đài Loan về phía Tây Indonesia, ngược lại các khác biệt trong nhóm đơn bội O2a1-M95/M88, O3a*-M324, O3a1c-IMS-JST002611 và O3a2c1a-M133 giảm theo hướng ngược lại từ Tây Indonesia về phía Đài Loan. So với các nhóm thiểu số bộ lạc đồng bằng Đài Loan, các nhóm sử dụng ngôn ngữ thuộc NHNĐ ở Đài Loan chỉ có một ít thông tin di truyền từ dòng bố người Hán. Những quần thể này đặc trưng bởi sự đa dạng NST Y thấp, do đó củng cố cho bằng chứng về sự trôi dạt di truyền nhanh chóng ở các quần thể này. Tuy nhiên, trái ngược với dữ liệu được cung cấp từ các vùng khác của bộ gen, sự đa dạng gen NST Y ở các bộ lạc miền núi Đài Loan tăng đáng kể từ Bắc đến Nam. Nghiên cứu tiếp theo được thực hiện bởi Macholdt và cộng sự [49] đã được trình bày ở mục 1.3.2.1.

Ngoài các nghiên cứu dựa trên SNP, một nghiên cứu khác dựa trên 23 STR trên NST Y cũng được thực hiện [68]. Trong 200 cá thể Kinh, 200 kiểu đơn bội khác nhau đã được tìm thấy, với 196 kiểu đơn bội là đặc thù. Tuy nhiên, có một điều cần lưu ý khi sử dụng STR là các locus STR pháp y thường được sử dụng được chọn cụ thể cho mục đích pháp y vì chúng có sự khác biệt về mặt di truyền thấp hơn giá trị thường thấy giữa các quần thể [69]. Sự khác biệt về mặt di truyền giữa các quần thể là mối quan tâm khi cần tính xác suất một hồ sơ STR cụ thể sẽ được quan sát ở một cá thể khác từ cùng một quần thể. Để ước tính xác suất này, người sử dụng phải tìm ra quần thể nào là phù hợp và nếu hồ sơ STR khác nhau nhiều giữa các quần thể, thì việc sử dụng quần thể tham chiếu không chính xác có thể cung cấp giá trị xác suất sai. Việc sử dụng các locus có sự khác biệt nhỏ về mặt di truyền giữa các quần thể làm cho điều này không còn là mối lo ngại lớn vì các giá trị xác suất là tương tự nhau bất kể quần thể tham chiếu được sử dụng, và do đó các locus STR như vậy được ưu tiên cho pháp y. Nhưng điều này cũng có nghĩa là việc sử dụng các locus STR đó trong các nghiên cứu nhân chủng học phân tử sẽ không thu được ước tính chính xác về sự khác biệt về mặt di truyền giữa các quần thể [69].

Chương 2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. ĐỐI TƯỢNG NGHIÊN CỨU

Mẫu máu ngoại vi đã được thu thập từ 129 nam giới người Việt, bao gồm 22 cá thể thuộc dân tộc Pa Kô, 19 cá thể thuộc dân tộc Cơ-tu, 24 cá thể thuộc dân tộc Rơ-măm, 31 cá thể thuộc quần thể Kinh sống ở miền Trung (KMT) và 33 cá thể thuộc quần thể Kinh sống ở miền Nam (KMN). Các mẫu thuộc dân tộc Pa Kô, Cơ-tu, Rơ-măm, KMT và KMN lần lượt được thu từ Quảng Trị, Thừa Thiên Huế, Kon Tum, Gialai và thành phố Hồ Chí Minh. Mẫu được chọn là mẫu không có quan hệ huyết thống và có ít nhất ba thế hệ trong gia đình đều thuộc một dân tộc. Các cá thể tham gia nghiên cứu đều ký vào giấy đồng ý tự nguyện cho máu cho nghiên cứu. Nghiên cứu này được thông qua bởi Hội đồng Đạo đức trong nghiên cứu y sinh của Viện Nghiên cứu hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam (No: 9-2019/NCHG-HĐĐĐ).

Đối với các kết quả thống kê về các điểm biến đổi mtDNA ở các mẫu mới cũng như so sánh với các quần thể trong nước, nghiên cứu đã tích hợp 129 trình tự mtDNA mới này với 600 trình tự mtDNA từ nghiên cứu của Dương và cộng sự [61], và 329 trình tự mtDNA từ nghiên cứu của Thảo và cộng sự [64]. Đối với các kết quả nghiên cứu đa dạng di truyền theo chỉ thị mtDNA, ngoài dữ liệu từ các cá thể trong nước, nghiên cứu đã tích hợp thêm 2207 trình tự vùng điều khiển của mtDNA từ các dân tộc ở Thái Lan [33–36], 49 trình tự vùng điều khiển của mtDNA từ Lào [34], 272 trình tự vùng điều khiển của mtDNA từ Campuchia [38,70], 108 trình tự vùng điều khiển của mtDNA từ Myanmar [39,71], 237 trình tự vùng điều khiển của mtDNA từ Ấn Độ [40], 142 trình tự vùng điều khiển của mtDNA từ Trung Quốc [72], 86 trình tự vùng điều khiển của mtDNA từ Malaysia [73], 276 trình tự vùng điều khiển của từ Philippines [74,75], 72 trình tự vùng điều khiển của mtDNA từ Indonesia [76] và 550 trình tự vùng điều khiển của mtDNA từ Đài Loan [77]. Để đảm bảo tính nhất quán trong nghiên cứu, nhóm nghiên cứu chỉ sử dụng trình tự vùng điều khiển mtDNA để phân tích đa dạng di truyền.

Từ 129 cá thể trong nghiên cứu, 16 cá thể KMT, 20 cá thể KMN, 12 cá thể Cơ-tu, 19 cá thể Pa Kô và 15 cá thể Rơ-măm sẽ được chọn ngẫu nhiên từ mỗi dân tộc tương ứng để gửi đi đọc dữ liệu SNP toàn bộ hệ gen sử dụng chip Axiom Genome-Wide Human Origins. Sau đấy, dữ liệu SNP trên NST Y của

82 cá thể này sẽ được tích hợp với dữ liệu SNP NST Y từ hai nghiên cứu trên quần thể người Việt trước đây bao gồm 170 cá thể từ nghiên cứu của Thảo [64] và cộng sự và 598 từ nghiên cứu của Macholdt [49], và dữ liệu SNP NST Y từ dự án 1000 bộ gen [72]. Hai mẫu bị loại đi từ nghiên cứu của Macholdt do có độ bao phủ thấp.

2.2. PHƯƠNG PHÁP NGHIÊN CỨU

2.2.1. Tách chiết và tinh sạch ADN tổng số từ mẫu máu

Quy trình tách chiết ADN tổng số của 129 mẫu máu được thực hiện trên bộ kit tách chiết và tinh sạch thương mại Exgene™ Blood SV Mini (GeneAll, Hàn Quốc), từng bước cụ thể như sau:

- Bước 1: Máu cất giữ ở -80^{o}C được lấy ra và để vào tủ ấm 37^{o}C trong thời gian 30 phút.

- Bước 2: Thêm 20 µl Proteinase K vào 200 µl máu, mix bằng vortex.

Thêm 200 µl dung dịch đệm BL, mix hoàn toàn bằng vortex hoặc pipet.

- Bước 3: Ủ mẫu ở 56^{o}C trong 10 phút và thỉnh thoảng vortex hoặc sử dụng bể lắc đến khi tế bào tan hoàn toàn.

- Bước 4: Thêm 200 µl ethanol (96-100%) và mix bằng pipet.

- Bước 5: Chuyển hỗn hợp vào cột ly tâm. Ly tâm ở 10000 vòng/phút trong vòng 1 phút. Loại bỏ ống thu có chứa dịch. Loại bỏ phần dịch thu được và đặt cột vào một ống thu 2ml mới.

- Bước 6: Thêm 500 µl đệm BW (đã thêm ethanol). Ly tâm ở tốc độ 10.000 vòng/phút trong 1 phút.

- Bước 7: Thêm 700 µl đệm TW (đã thêm ethanol). Ly tâm ở tốc độ 12.000 vòng/phút trong 3 phút. Loại bỏ phần dịch thu được và đặt cột vào ống thu mới.

- Bước 8: Ly tâm cột ở tốc độ 12.000 vòng/phút trong vòng 2 phút để loại bỏ tất cả dịch còn lại.

- Bước 9: Đặt cột vào ống eppendorf 1,5 mL mới và thêm 200 µl đệm TE vào giữa màng cột để thu ADN. Ủ ở nhiệt độ phòng trong 10 phút và ly tâm ở 10000 vòng/phút trong 1 phút.

- Bước 10: Bỏ cột. Sử dụng ADN thu được hoặc lưu trữ ở -20^{o}C.

2.2.2. Điện di kiểm tra ADN trên gel agarose

- Hòa 0,8 g agarose trong 100 ml TAE 1X, đun sôi trong lò vi sóng cho tan hoàn toàn. Để cho nhiệt độ hạ xuống khoảng 500^{o}C, đổ vào khuôn có đặt sẵn răng lược thích hợp. Đợi cho thạch đông và ổn định trong vòng 1 giờ. Sau đó bỏ răng lược ra và đặt gel và hộp điện di. Đổ đệm TAE 1X sao cho mức đệm cao hơn mặt gel từ 1 - 2 mm.

- Trộn một lượng mẫu thích hợp với đệm tra mẫu (glycerol 20%; Tris-Cl 0,1M pH 8; EDTA 0,01M pH 8; bromophenol blue 0,25%). Điện di với dòng điện một chiều có hiệu điện thế 100V, dòng 60 - 80 mA, trong thời gian khoảng 30 phút.

- Nhuộm gel bằng Ethidium bromide 10 µg/ml trong khoảng 10 phút trên máy lắc, sau đó rửa sạch mẫu.

- Quan sát và chụp ảnh trên máy GEL - DOC.

2.2.3. Xác định nồng độ ADN bằng quang phổ kế

Các bước tiến hành:

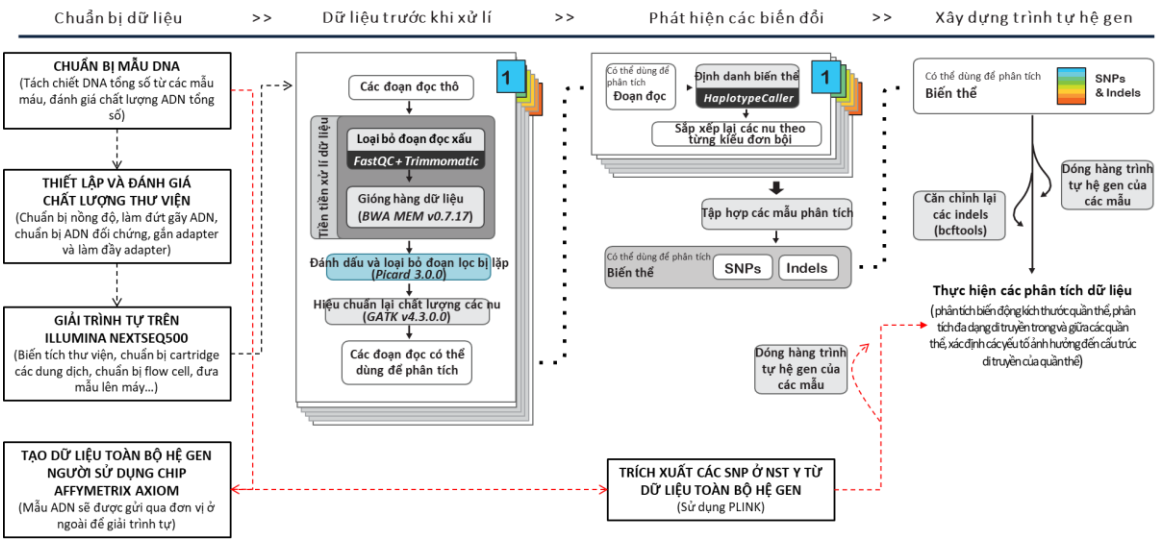
- Lấy 3 µl dung môi (TE pH 8,0 hoặc nước khử ion vô trùng) để làm Blank.

- Sử dụng 3 µl mỗi mẫu để xác định nồng độ ADN ở bước sóng 260 nm và 280 nm. Nồng độ ADN được thể hiện trên máy quang phổ (NanoDrop) được tính bằng công thức sau: Nồng độ ADN (ng/µl) = Độ hấp thụ ở bước sóng 260 nm x 50.

- Ghi lại kết quả nồng độ ADN cho mỗi mẫu nghiên cứu sau mỗi lần đo.

2.2.4. Giải trình tự hệ gen ty thể trên hệ thống máy giải trình tự thế hệ mới

Quy trình giải trình tự mtDNA, xử lí và phân tích dữ liệu của các cá thể trong nghiên cứu được mô tả chi tiết theo sơ đồ sau:



Hình 2.1. Sơ đồ mô tả quy trình nghiên cứu

2.2.4.1. Thiết lập thư viện ADN

Quá trình thiết lập thư viện ADN được dựa trên công bố của Maricic và cộng sự với một số cải biến gồm những bước sau [78]:

- Chuẩn bị ADN: Mẫu ADN tổng số được pha loãng về nồng độ 10 ng/µl sử dụng dung dịch đệm TE. Tiếp đến, 80 µl ADN tổng số đã pha loãng được chuyển vào ống 0,5ml để tiến hành cắt phân đoạn ADN tổng số. Các mẫu ADN đối chứng được chuẩn bị và sử dụng từ bước tạo đầu bằng cho tới bước làm đầy adapter. Đối chứng âm (-ve) là dung dịch TE và đối chứng dương (+ve) là đoạn ADN hoặc sản phẩm PCR có kích thước 300 bp.

- Phản ứng tạo đầu bằng cho các đoạn ADN được thực hiện trên đĩa PCR với tổng thể tích của một phản ứng là 70 µl gồm 50 µl thể tích dung dịch ADN được làm đứt gãy nhờ siêu âm và 20 µl thể tích dung dịch mastermix gồm 7,12 µl H_2O, 7,0 µl buffer Tango (10X), 0,28 µl dNTPs (25mM), 0,7 µl ATP (100mM), 3,5 µl T4 polynucleotide kinase (10U/µl), 1,4 µl T4 ADN polymerase (5U/µl). Thí nghiệm cũng sử dụng một đối chứng âm (-ve) và một đối chứng dương (+ve) trên cùng một đĩa. Đĩa mẫu sẽ được ủ ở 25^{o}C trong thời gian 15 phút. Sau đó giữ phản ứng ở 12^{o}C trong thời gian 5 phút.

- Tinh sạch sản phẩm bằng Solid Phase Reversible Immobilization (SPRI)

- Gắn adapter: Quá trình gắn adapter sử dụng ba loại adapter khác nhau, bao gồm Sol_MP_P5, Sol_MP_P7 và Sol_MP_M5P7_comp (Bảng 2.1). Để

tạo ra loại adapter trộn giữa adapter P5 và P7, cần phải chuẩn adapter P5 và P7 với thành phần lần lượt như sau: 20 µl Sol_MP_P5 (500µM), 20 µl Sol_MP_P5P7_comp (500µM), 5 µl HybBuffer (10X), 5 µl dung dịch TE (1X); 20 µl Sol_MP_P7 (500µM), 20 µl Sol_MP_P5P7_comp (500 µM), 5 µl HybBuffer (10X), 5 µl dung dịch TE (1X). Sau đó, ủ adapter P5 và P7 ở 95^{o}C trong 1 phút. Sau đó giảm nhiệt độ xuống 14^{o}C với tốc độ giảm nhiệt 0,1^{o}C/ s. Adapter trộn được chuẩn bị bằng việc trộn adapter P5 và adapter P7 theo tỉ lệ thể tích (1:1). Phản ứng gắn adapter được thực hiện với phản ứng có tổng thể tích 40 µl gồm 20 µl master mix (10 µL H_2O, 4 µL T4 ADN ligase buffer (10X), 4 µL PEG-4000 (50%), 1 µL adapter mix (100 µM), 1 µL T4 ADN ligase (5U/ µL)) và 20 µl sản phẩm đã được tạo đầu bằng trước đó. Đậy các ống của đĩa mẫu bằng một tấm thermo film và trộn nhẹ bằng máy vortex. Ly tâm nhẹ đĩa mẫu trong máy ly tâm đĩa. Ủ đĩa mẫu ở 60^{o}C trong 60 phút. Tinh sạch sản phẩm gắn adapter bằng SPRI.

Bảng 2.1. Trình tự nucleotide của adapter sử dụng trong quá trình gắn adapter

Tên adapter	Trình tự nucleotide	Tên adapter
Sol_MP_P5	A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T	Sol_MP_P5
Sol_MP_P7	Biotine-G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*C*T*T	Sol_MP_P7
Sol_MP_P5P7_comp	A*G*A*T*CGGAA*G*A*G*C	Sol_MP_P5P7_comp

Ghi chú: * thể hiện sự sửa đổi phosphorothioate.- Làm đầy adapter: thành phần phản ứng gồm 20 µl master mix (14,1 µL H_2O, 4,0 µL Thermopol buffer (10X), 0,4 µL dNTPs (25 mM), 1,5 µL Bst polymerase (8U/ µL)) và 20 µl sản phẩm đã được gắp adapter. Đậy các ống trong đĩa mẫu bằng một tấm thermo film và trộn nhẹ bằng máy vortex. Ly tâm nhẹ đĩa mẫu trong máy ly tâm đĩa. Ủ mẫu ở 37^{o}C trong 20 phút. Tinh sạch sản phẩm fill-in adapter bằng SPRI và kiểm tra sản phẩm làm đầy adapter trên gel agrose 2%.

- Định lượng sản phẩm làm đầy adapter bằng phương pháp PCR định lượng: Các ống trong đĩa mẫu sẽ được bổ sung 19 µl của dung dịch master mix (7,0 µL H_2O, 1,0 µL mồi xuôi (Sol_iPCR_P5), 1,0 µL mồi ngược (Sol_iPCR_P7), 10,0 µL Maxima Mastermix) và 1 µl mẫu vào các ống tương ứng trong đĩa mẫu. Đậy đĩa mẫu bằng một tấm thermo film và trộn nhẹ bằng máy vortex. Ly tâm nhẹ đĩa. Cụ thể, 1 µL của sản phẩm làm đầy adapter từ 8 mẫu ngẫu nhiên sẽ được sử dụng. Sau đó chuẩn bị các mẫu pha loãng với hai tỷ lệ (1:10) và (1:100); đối chứng âm và qPCR standard với các nồng độ giảm dần 10^9, 10^8, 10^7, 10^6, 10^5, 10^4, 10^3, 10^2.

- Quá trình index PCR: mỗi một ống trong đĩa mẫu gồm 37 µl của dung dịch master mix (26,0 µL H_2O; 10,0 µL Herculase buffer (5X); 0,5 µL dNTPs (25 mM) và 0,5 µL Herculase polymerase) và 10 µl sản phẩm làm đầy adapter. Bổ sung lần lượt 1,5 µl mồi index P5 và 1,5 µl mồi index P7 vào các ống tương ứng trong đĩa mẫu. Đậy đĩa mẫu bằng một tấm thermo film và vortex. Ly tâm đĩa mẫu trong máy ly tâm đĩa. Điều kiện của phản ứng Indexing PCR như sau: 95°C/ 2 phút, n chu kỳ của (95°C/ 20s, 60°C/ 20s, 72°C/30s), 72°C/ 3 phút. Trong đó, số chu kỳ n được xác định bằng phản ứng định lượng sản phẩm làm đầy adapter ở phía trên. Sản phẩm Indexing PCR sẽ được kiểm tra trên gel agarose 2% và tinh sạch bằng SPRI.

- Xác định nồng độ các sản phẩm index PCR: các sản phẩm Indexing PCR chưa pha loãng sẽ được định lượng bằng phương pháp đo quang phổ bằng máy Nanodrop. Dựa trên kết quả đo, thực hiện tính toán thể tích cần sử dụng để mỗi mẫu có được khoảng 200 ng sản phẩm Indexing PCR. Sau đó, chuyển các mẫu với thể tích đã được tính toán vào chung một ống eppendorf 0,5 mL.

- Phản ứng lai làm giàu trình tự đích:

+ Chuẩn bị các mẫu cho phản ứng lai-bắt giữ bao gồm 1 ống chứa ADN bait gồm 3 µl ADN bait, 7 µl H_2O, và 10 µl BWT buffer (1X) được ủ ở 95°C trong 1 phút và ngay lập tức đưa lên đá sau ủ; và 1 ống chứa bead gồm 7 µl bead để lên giá từ, sau đó rửa bead với lần lượt 200 µl buffer BWT (1X) và 200 µl TET buffer (1X) và cuối cùng elute bead với 50 µl TET trước khi đặt trực tiếp lên đá. Chuyển toàn bộ 20 µl thể tích ADN bait vào ống chứa bead ở trên và ủ ở nhiệt độ phòng trong vòng 20 phút. Rửa hai lần với dung dịch BWT buffer (1X) và để trên giá từ trong 2 phút và loại bỏ dịch huyền phù. Tiếp tục

rửa hai lần với BWT buffer (1X) với sau mỗi lần đều ủ ở 40^{o}C trong 2 phút. Rừa giải hỗn hợp ADN bait và bead trong 50 µl TET buffer.

+ ADN bait sẽ được gắn lên hạt M-270 Streptavidin trước khi sử dụng cùng với dung dịch lai để thực hiện phản ứng lai. 52 µl dung dịch lai (17,8 µl ADN (2 µg), 0,75 µl BO4_CSH (200 µM), 0,75 µl BO6_CSH (200 µM), 0,75 µl BO8.P5.part1.R, 0,75 µl BO10.P5.part2.R, 5,2 µl agilent blocking agent (2X), 26,0 µl agilent hybridization buffer (2X)) sẽ được ủ trong 3 phút ở 95^{o}C và tiếp tục ủ lắc ở 37^{o}C trong 30 phút trước khi sử dụng trong phản ứng lai. Đặt hỗn hợp bait, bead và TET lên trên giá từ và loại bỏ dịch huyền phù. Bổ sung dung dịch lai vào hỗn hợp trên và ủ ở lò lai nhiệt độ 65^{o}C trong khoảng 2 ngày. Sản phẩm lai sau khi ủ sẽ được đặt trên giá từ và rửa 4 lần với 200 µl buffer BWT (1X). Bổ sung 200 µl HWB buffer đã được làm nóng trước ở 60^{o}C vào ống và ủ ở 60^{o}C trong vòng 2 phút. Đặt lên giá từ và loại bỏ dịch huyền phù. Rửa với 200 µl BWT buffer (1X), loại bỏ dịch huyền phù. Bổ sung 100 µl TET và chuyển hỗn hợp dung dịch sang ống 0,5 ml mới. Chuyển lên giá từ và loại bỏ dịch huyền phù. Thêm 30 µl EBT buffer và ủ ở 95^{o}C trong 3 phút và chuyển lên đá. Chuyển toàn bộ phần dịch huyền phù vào ống 0,5 ml mới.

+ Thư viện sau khi làm giàu sẽ được định lượng bằng phản ứng PCR định lượng gồm 7,0 µl H_2O, 1,0 µl mồi Sol_bridge_P5 (10 µM), 1,0 µl mồi Sol_bridge_P7 (10 µM), 10,0 µl DyNAmo mastermix (2X) và 1 µl thư viện đã được làm giàu. Điều kiện phản ứng PCR định lượng được thực hiện như sau: 95^{o}C/ 15 phút, 40 chu kỳ của (94^{o}C/ 30s, 58^{o}C/30s, 72^{o}C/ 1 phút), 72^{o}C/ 10 phút.

- Khuếch đại thư viện ADN sau khi được làm giàu: Phản ứng PCR khuếch đại thư viện ADN có tổng thể tích 40 µl bao gồm các thành phần bao gồm 26,0 µl H_2O, 10,0 µl herculase buffer (5X), 0,5 µl dNTPs (25 mM), 1,5 µl P5 Bridge primer (10 µM), 1,5 P7 µl Bridge primer (10 µM), 0,5 µl herculase polymerase và 10 µl thư viện ADN đã được làm giàu. Điền kiện của phản ứng khuếch đại như sau: 95^{o}C/ 2 phút, n chu kỳ của (98^{o}C/ 20s, 60^{o}C/ 20s, 72^{o}C/30s) và 72^{o}C/3 phút. Trong đó, n chu kỳ được tính toán dựa trên kết quả định lượng thư viện sau khi làm giàu ở trên. Thư viện sau khi khuếch đại sẽ được tinh sạch bằng SPRI. Thư viện sau khi tinh sạch sẽ được mang đi kiểm tra chất lượng về

nồng độ và kích thước trên hệ thống Bioanalyzer trước khi được mang đi giải trình tự

2.2.4.2. Giải trình tự hệ gen ty thể trên hệ thống máy giải trình tự thế hệ mới

Các bước giải trình tự trên máy Illumina được thực hiện theo hướng dẫn của nhà sản xuất gồm các bước như sau:

- Làm tan các hộp hoá chất;
- Kiểm tra hộp hoá chất;
- Bổ sung dung dịch NaOCl mới pha loãng;
- Chuẩn bị flow cell;
- Kiểm tra flow cell;
- Tra mẫu vào khay;
- Cài đặt chương trình chạy trên máy;
- Đặt flow cell vào khay;
- Loại bỏ khay chất thải;
- Loại bỏ buffer cartridge đã sử dụng khỏi khoang trên.

2.2.5. Tạo dữ liệu đa hình nucleotide đơn nhiễm sắc thể Y sử dụng chip Axiom Genome-Wide Human Origins

- Mẫu ADN sẽ được gửi qua đơn vị ở ngoài để giải trình tự.

- Sau khi có dữ liệu SNP toàn bộ hệ gen, trích xuất các SNP ở NST Y bằng PLINK.

2.2.6. Tiền xử lý dữ liệu

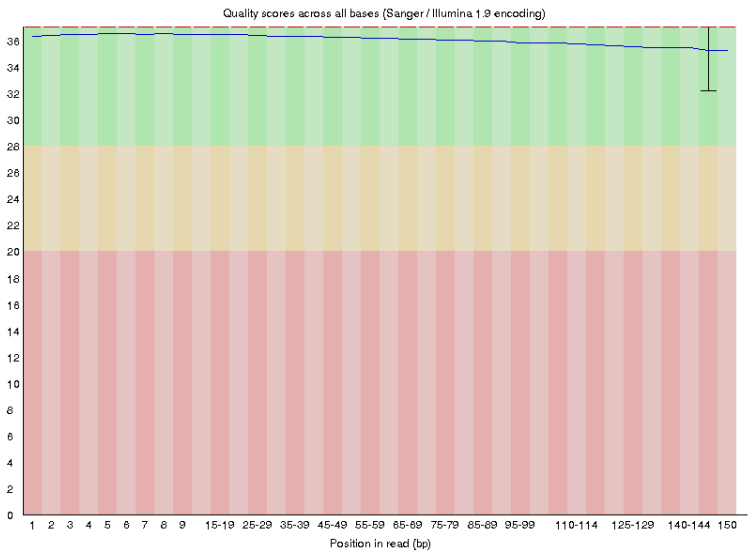
2.2.6.1. Đánh giá chất lượng và loại bỏ dữ liệu chất lượng thấp

Dữ liệu sẽ được đánh giá chất lượng và loại bỏ những dữ liệu chất lượng thấp hoặc không phù hợp với nhu cầu phân tích sử dụng FastQC v0.12.1 [79] và Trimmomatic v0.32 [80]. Việc kiểm tra chất lượng ở giai đoạn này bao gồm:

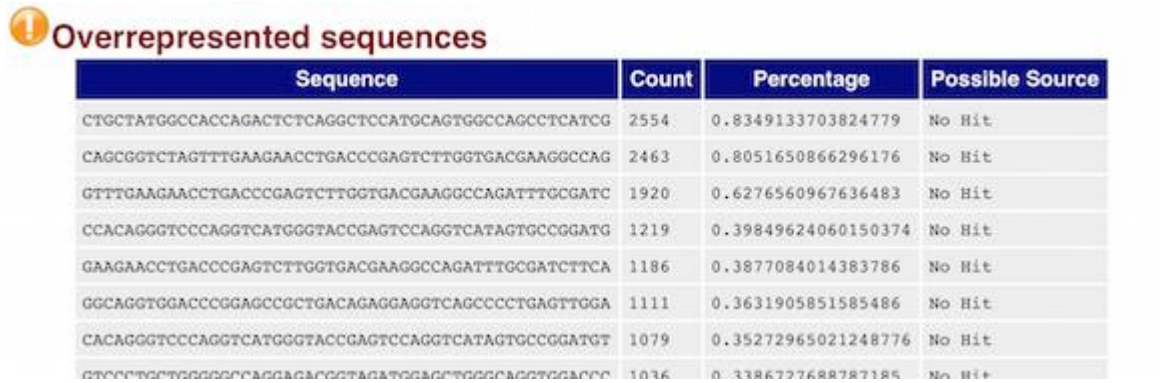
- Kiểm tra chất lượng của các nucleotide đọc được để đảm bảo không có vấn đề gì trong quá trình sắp xếp thứ tự

- Kiểm tra các đoạn đọc để đảm bảo số liệu chất lượng của chúng tuân thủ mong đợi
- Kiểm tra độ sạch của các đoạn đọc (khả năng xuất hiện của trình tự không mong muốn)

FastQC được sử dụng để xác định các vấn đề về dữ liệu thu được sau khi giải trình tự như xác định chất lượng tất cả các nucleotide tại mỗi vị trí trên các đoạn đọc (Hình 2.2), xác định trình tự không mong muốn và trình tự lặp lại quá nhiều lần (Hình 2.3), hiểu rõ được độ phức tạp của thư viện, đảm bảo các sinh vật được thể hiện chính xác bằng hàm lượng %GC (mặc dù đôi khi số liệu này có thể bị sai lệch nếu có nhiều gen được biểu hiện quá mức). Sau đấy, Trimmomatic được sử dụng để loại bỏ các base chất lượng thấp và các đoạn adapter. Sau khi cắt ngắn các đoạn đọc, chúng tôi sử dụng FastQC trên mẫu đã được cắt bớt để xem liệu có sự cải thiện nào ở các mẫu chất lượng thấp hay không.



Hình 2.2. Mô tả chất lượng dữ liệu theo toàn bộ các nucleotide trên trình tự bằng phần mềm FastQC



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit

Hình 2.3. Các trình tự xuất hiện nhiều lần ở một mẫu trong nghiên cứu

2.2.6.2. Dóng các đoạn đọc lên trình tự tham chiếu

Để xác định vị trí các đoạn đọc trên hệ gen người, các đoạn đọc sẽ được căn chỉnh với bộ gen tham chiếu bằng cách sử dụng BWA MEM [81]. Trước khi sử dụng BWA MEM, chúng tôi cần tải xuống hệ gen tham chiếu (tệp fasta) từ phylotree (http://www.phylotree.org/resources/RSRS.fasta). Sau đó, bộ gen sẽ được đánh dấu vì việc đánh dấu cho phép bộ căn chỉnh thu hẹp nguồn gốc tiềm năng của chuỗi truy vấn trong bộ gen, tiết kiệm cả thời gian và bộ nhớ. BWA index được sử dụng để đánh dấu bộ gen tham chiếu. Đầu ra ở bước này là tệp dữ liệu dưới định dạng .sam hoặc .bam.

2.2.6.3. Xác định, đánh dấu và loại bỏ các trình tự lặp lại

Sau khi lắp ráp xong, bước tiếp theo là xác định, đánh dấu và loại bỏ các trình tự lặp lại ở file BAM hoặc SAM sử dụng Picard 3.0.0/Samtools [82]. Các đoạn này sinh ra do quá trình khuếch đại PCR và có thể gây ra các kết quả phát hiện biến thể dương tính giả. Nếu sử dụng Picard thì sẽ dùng công cụ MarkDuplicates, còn sử dụng samtools thì sẽ dùng công cụ markdup.

2.2.6.4. Hiệu chỉnh chất lượng các nucleotide

Sau khi đã loại bỏ các đoạn đọc trùng lặp, các base được đánh giá lại chất lượng bằng việc sử dụng công nghệ học máy để phát hiện và sửa lỗi hệ thống trong việc đánh giá chất lượng các base dựa vào cách thức cho điểm từng base của máy giải trình tự. Điểm chất lượng của các nucleotide đóng vai trò quan trọng trong việc cân nhắc bằng chứng ủng hộ hoặc chống lại các alen biến thể có thể có trong quá trình phát hiện biến thể, do đó điều quan trọng là phải sửa các sai lệch hệ thống được quan sát thấy trong dữ liệu. Những sai lệch có thể bắt nguồn từ các quá trình sinh hóa trong quá trình chuẩn bị và giải trình tự

thư viện, từ các lỗi sản xuất trong chip hoặc các lỗi thiết bị của máy giải trình tự. Quy trình hiệu chỉnh lại bao gồm việc thu thập số liệu thống kê đồng biến từ tất cả các lệnh gọi nucleotide trong tập dữ liệu, xây dựng mô hình từ các số liệu thống kê đó và áp dụng các thuật toán điều chỉnh chất lượng nucleotide cho tập dữ liệu dựa trên mô hình kết quả. Việc thu thập số liệu thống kê ban đầu có thể được thực hiện song song bằng thu thập cùng lúc trên toàn bộ hệ gen, thường là theo nhiễm sắc thể hoặc các lô nhiễm sắc thể nhưng điều này có thể được chia nhỏ hơn nữa để tăng thông lượng nếu cần. Sau đó, số liệu thống kê theo vùng phải được tập hợp thành một mô hình cộng biến trên toàn bộ hệ gen; bước này không thể thực hiện song song nhưng nó không đòi hỏi nhiều về mặt tính toán. Cuối cùng, các quy tắc hiệu chỉnh xuất phát từ mô hình ở trên được áp dụng cho tập dữ liệu gốc để tạo ra tập dữ liệu được hiệu chỉnh lại. Quá trình này được thực hiện song song theo cách tương tự việc thu thập số liệu thống kê ban đầu, trên các vùng gen, sau đó là thao tác hợp nhất tệp cuối cùng để tạo ra một tệp sẵn sàng phân tích cho mỗi mẫu.

Công cụ BaseRecalibrator/ApplyBQSR của GATK v4.3.0.0 [83] được sử dụng cho bước này. Bước đầu tiên là hiệu chỉnh lại điểm chất lượng các nucleotide bằng cách sử dụng BaseRecalibrator (GATK v4.3.0.0). BaseRecalibrator tạo bảng dựa trên các hiệp biến cụ thể. Nó di chuyển qua các locus ở những vị trí có trong tệp gọi các biến thể đã biết. Các cơ sở dữ liệu ExAc, gnomAD hoặc dbSNP có thể được sử dụng để lấy vị trí các biến thể đã biết. Do đó, chúng tôi giả định rằng tất cả các tham chiếu không khớp đều là lỗi và cho thấy chất lượng base kém. Vì có một lượng lớn dữ liệu nên ta có thể tính toán xác suất xảy ra lỗi thực nghiệm dựa trên các hiệp biến cụ thể tại điểm này, trong đó p(error) = số lượng sai lệch / số lượng quan sát. Tệp đầu ra là một bảng có chứa một số giá trị đồng biến, số lượng giá trị quan sát được, số lượng sai lệch, điểm chất lượng thực nghiệm. Để chuẩn bị các cơ sở dữu liệu cần thiết cho bước này, chúng tôi cần tải xuống các tệp có biến thể đã biết được liệt kê ở định dạng vcf từ ensembl (https://ftp.ensembl.org/pub/release-109/variation/vcf/homo_sapiens/) và cơ sở dữ liệu NCBI (https://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/). Sau đó, chúng tôi áp dụng hiệu chỉnh lại điểm chất lượng các nucleotide bằng cách sử dụng ApplyBQSR (GATK v4.3.0.0). Công cụ này thực hiện bước thứ hai trong quy trình hai giai đoạn được gọi là hiệu chỉnh lại điểm chất lượng các nucleotide.

Cụ thể hơn, nó hiệu chỉnh lại chất lượng các nucleotide của các đoạn đọc dựa trên bảng hiệu chỉnh lại do công cụ BaseRecalibrator tạo ra và xuất ra tệp BAM được hiệu chỉnh lại.

2.2.6.5. Phát hiện các biến thể

Có nhiều phần mềm phát hiện biến thể khác nhau; mỗi phần mềm có thể phát hiện một hoặc nhiều dạng biến thể: biến thể nucleotide đơn, biến thể số lượng bản sao và biến thể cấu trúc. Trong nghiên cứu này chúng tôi sẽ sử dụng công cụ HaplotypeCaller của GATK v4.3.0.0 để phát hiện các biến thể ở trên hệ gen ty thể của người. HaplotypeCaller có khả năng định dạng đồng thời SNP và biến thể thêm mất đoạn (indel) thông qua việc tập hợp các kiểu đơn bội cục bộ trong một khu vực hoạt động. Nói cách khác, bất cứ khi nào chương trình gặp một vùng có dấu hiệu bị biến đổi, nó sẽ loại bỏ thông tin trình tự hiện có và tập hợp lại hoàn toàn các lần đọc trong vùng đó. Điều này cho phép HaplotypeCaller trở nên chính xác hơn khi định danh các vùng trước đây khó định danh, chẳng hạn như các vùng chứa các loại biến thể khác nhau gần nhau. Đầu ra của bước này là tệp dữ liệu dưới định dạng .vcf (Hình 2.4).

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	PAKON01_	PAKON01_	PAKON02_	PAKON02_	PAKON03_	PAKON03_
RSRS	125	.	T	C	2896.74	.	AC=2;AF=(	GT:AD:DP:	0:39,0:39:9	0:36,0:36:9	0:64,0:64:9	0:59,0:59:9	0:26,0:26:9	0:22,0:22:9
RSRS	126	.	A	G	2941.74	.	AC=2;AF=(	GT:AD:DP:	0:39,0:39:9	0:36,0:36:9	0:64,0:64:9	0:59,0:59:9	0:26,0:26:9	0:22,0:22:9
RSRS	127	.	T	C	2941.74	.	AC=2;AF=(	GT:AD:DP:	0:39,0:39:9	0:36,0:36:9	0:64,0:64:9	0:59,0:59:9	0:26,0:26:9	0:22,0:22:9
RSRS	143	.	G	A	2716.74	.	AC=2;AF=(	GT:AD:DP:	0:39,0:39:9	0:36,0:36:9	0:64,0:64:9	0:59,0:59:9	0:26,0:26:9	0:22,0:22:9
RSRS	146	.	C	T	96929.55	.	AC=42;AF=	GT:AD:DP:	1:0,40:40:9	1:0,50:50:9	1:0,84:84:9	1:0,82:82:9	1:0,40:40:9	1:0,23:23:9
RSRS	150	.	C	T	30128.7	.	AC=10;AF=	GT:AD:DP:	0:40,0:40:9	0:50,0:50:9	0:82,0:82:9	0:81,0:81:9	0:37,0:37:9	0:21,0:21:9
RSRS	152	.	C	T	98731.55	.	AC=42;AF=	GT:AD:DP:	1:0,39:39:9	1:0,50:50:9	1:0,79:79:9	1:0,81:81:9	1:0,37:37:9	1:0,21:21:9
RSRS	185	.	G	A	9489.17	.	AC=4;AF=(	GT:AD:DP:	0:38,0:38:9	0:46,0:46:9	1:1,67:68:9	1:0,75:75:9	0:33,0:33:9	0:17,0:17:9
RSRS	195	.	C	T	74626.8	.	AC=38;AF=	GT:AD:DP:	1:0,41:41:9	1:0,46:46:9	1:0,65:65:9	1:0,70:70:9	1:0,32:32:9	1:0,22:22:9
RSRS	198	.	C	T	6785.74	.	AC=2;AF=(	GT:AD:DP:	0:42,0:42:9	0:45,0:45:9	0:66,0:66:9	0:67,0:67:9	0:28,0:28:9	0:22,0:22:9
RSRS	199	.	T	C	15960.86	.	AC=6;AF=(	GT:AD:DP:	0:42,0:42:9	0:45,0:45:9	0:66,0:66:9	0:67,0:67:9	0:28,0:28:9	0:22,0:22:9
RSRS	210	.	A	G	25497.21	.	AC=14;AF=	GT:AD:DP:	0:42,0:42:9	0:45,0:45:9	0:66,0:66:9	0:67,0:67:9	1:0,30:30:9	1:0,22:22:9
RSRS	246	.	TA	T	9490.13	.	AC=4;AF=(	GT:AD:DP:	0:42,0:42:9	0:45,0:45:9	0:66,0:66:9	0:67,0:67:9	0:26,0:26:9	0:23,0:23:9
RSRS	247	.	A	G,*	78108.76	.	AC=40,4;A	GT:AD:DP:	1:0,45,0:45	1:0,44,0:44	1:1,75,0:76	1:0,73,0:73	1:0,28,0:28	1:0,26,0:26
RSRS	248	.	A	G	9490.17	.	AC=4;AF=(	GT:AD:DP:	0:31,0:31:9	0:44,0:44:9	0:75,0:75:9	0:68,0:68:9	0:22,0:22:9	0:27,0:27:9
RSRS	302	.	A	AC,ACC	19253.44	.	AC=22,6;A	GT:AD:DP:	0:31,0,0:3	0:24,0,0:24	1:2,31,7:40	1:0,24,5:32	1:0,6,4:10:	1:1,11,2:14
RSRS	310	.	T	TC	43206.12	.	AC=44;AF=	GT:AD:DP:	1:0,27:27:9	1:0,20:20:9	1:0,44:44:9	1:0,35:39:9	1:0,12:12:9	1:0,17:17:9
RSRS	332	.	C	T	6656.17	.	AC=4;AF=(	GT:AD:DP:	0:31,0:31:9	0:23,0:23:9	0:50,0:50:9	0:47,0:47:9	0:14,0:14:9	0:19,0:19:9
RSRS	482	.	T	C	5805.74	.	AC=2;AF=(	GT:AD:DP:	0:31,0:31:9	0:23,0:23:9	0:50,0:50:9	0:47,0:47:9	0:14,0:14:9	0:19,0:19:9
RSRS	489	.	T	C	21002.76	.	AC=12;AF=	GT:AD:DP:	0:31,0:31:9	0:23,0:23:9	0:50,0:50:9	0:47,0:47:9	0:14,0:14:9	0:19,0:19:9
RSRS	709	.	G	A	36152.94	.	AC=16;AF=	GT:AD:DP:	0:46,0:46:9	0:45,0:45:9	0:64,0:64:9	0:67,0:67:9	1:0,38:38:9	1:1,44:45:9
RSRS	737	.	C	T	11629.17	.	AC=4;AF=(	GT:AD:DP:	0:46,0:46:9	0:45,0:45:9	0:64,0:64:9	0:67,0:67:9	0:38,0:38:9	0:40,0:40:9
RSRS	769	.	A	G	115231.2	.	AC=44;AF=	GT:AD:DP:	1:0,58:58:9	1:0,70:70:9	1:1,106:10	1:0,99:99:9	1:0,42:42:9	1:0,43:43:9
RSRS	825	.	A	T	118348.2	.	AC=44;AF=	GT:AD:DP:	1:1,59:60:9	1:0,76:76:9	1:0,120:12	1:2,105:10	1:2,30:32:9	1:3,39:42:9

Hình 2.4. Định dạng của một file vcf

2.2.6.6. Tạo tệp dữ liệu dưới dạng .fasta

Sau khi xác định được các biến thể trong các mẫu, bước tiếp theo là tạo tệp dữ liệu trình tự mtDNA của các mẫu dưới dạng .fasta. Công cụ FastaAlternateReferenceMaker của GATK v4.3.0.0 với hệ gen tham chiếu là

RSRS (Reconstructed Sapiens Reference Sequence) được sử dụng để thực hiện bước này.

2.2.7. Phân tích đa dạng di truyền quần thể

2.2.7.1. So sánh nhiều trình tự bằng MAFFT v7.520

Trước hết, chúng tôi sử dụng công cụ MAFFT v7.520 để gióng hàng tất cả các trình tự với trình tự tham chiếu RSRS [84]. Sau đó, để đảm bảo độ chính xác của quá trình phân tích, chúng tôi tiến hành loại bỏ tất cả các vị trí không có thông tin và 8 vùng trình tự có thể gây nhiễu trên hệ gen biến thể ty thể bao gồm: (1) vùng poly-C tại vùng siêu biến II (303–317); (2) vùng lặp lại CA (514–523); (3) chuỗi C-1 (568–573); (4) 12S rARN (np 956–965); (5) vị trí có liên quan đến quá trình lịch sử (historical site) (3107); (6) chuỗi C-2 (5895–5899); (7) vùng mất đoạn/ thêm đoạn 9 bp (8272–8289); và (8) vùng poly-C tại vùng siêu biến I (16180–16195). Các đoạn trình tự sau khi được xử lý sẽ được dùng cho tất cả các nghiên cứu chuyên sâu tiếp theo.

2.2.7.2. Xác định các giá trị đa dạng quần thể sử dụng Arlequin v3.5.2.2

Arlequin [85] là một phần mềm di truyền quần thể có khả năng xử lý các mẫu dữ liệu phân tử lớn (RFLP, trình tự ADN, microsatellite), đồng thời vẫn duy trì khả năng phân tích dữ liệu di truyền thông thường (dữ liệu đa vị trí chuẩn hoặc dữ liệu tần số alen đơn thuần). Để sử dụng các công cụ tính toán, Arlequin yêu cầu người dùng chuẩn bị một tệp tin đầu vào có định dạng ARP từ dữ liệu tệp .fasta đã có (Hình 2.5).

```
NBSamples=3

DataType=STANDARD
GenotypicData=0
LocusSeparator=WHITESPACE

[Data]
[[Samples]]

SampleName="Cotu"
SampleSize=19
SampleData={

Cotu01 1 G A T C A C A G G T C T A T C A C C C T A T T A A C C A C T C A C G G G A G C T C T C C
C T C G C C C A T C C T A C C C A G ? ? ? ? ? ? ? ? ? A C C G C T G C T A A C C C C C C A A C C A
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
```

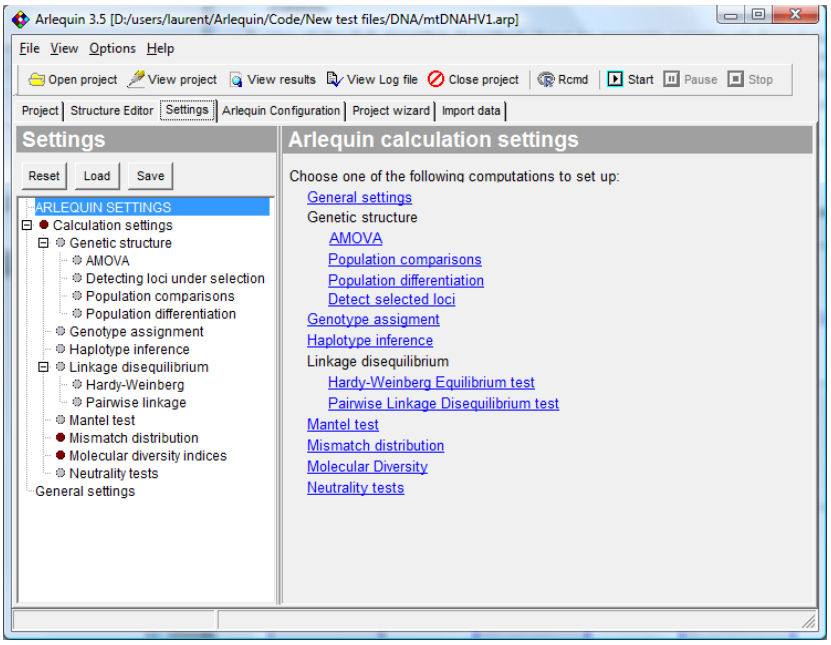
Hình 2.5. Cấu trúc tệp tin .arp sử dụng cho Arlequin

Trong hình 2.5, những thông tin cần phải khai báo bao gồm:

- Số lượng quần thể: NBSamples = 3.
- Dạng xử lí dữ liêu: DataType = STANDARD. Ở đây nghiên cứu sẽ dùng dạng "STANDARD" do mẫu đầu vào là các trình tự hệ gen biến thể ty thể.
- Dạng dữ liệu đầu vào: GenotypicData = 0. Sử dụng giá trị 0 (đơn bội) do dữ liệu đầu vào là ty thể nên sẽ ở dạng đơn bội.
- Kí tự sử dụng để tách riêng các alen ở locus khác nhau: LocusSeparator = WHITESPACE.
- Tên quần thể dân tộc: SampleName = "Cotu".
- Kích thước quần thể: SampleSize = 19.
- Tên, tần số và trình tự của từng mẫu trong quần thể: SampleData = {}. Tất cả các thông tin trên sẽ được khai báo trong dấu ngoặc {}.

Để thực hiện tính toán các giá trị đa dạng của các dân tộc trong nghiên cứu, bao gồm số lượng kiểu đơn bội (hai trình tự ADN giống hệt nhau được coi là thuộc cùng một kiểu đơn bội) của toàn bộ các dân tộc và từng dân tộc, chỉ số đa dạng kiểu đơn bội (*H*), đa dạng nucleotide (π), và trung bình số lượng

nucleotide khác nhau theo từng cặp trong tập mẫu (MPD - mean pairwise differences), chúng tôi lựa chọn mục "Haplotype inferences" và "Molecular diversity" trong mục "Settings" của Arlequin (Hình 2.6). Để quan sát mức độ đa dạng *H* và π, chúng tôi tính toán độ chênh lệch (%) so với giá trị trung bình của tất cả dân tộc đối với từng dân tộc. Để nghiên cứu sâu hơn về mức độ giống kiểu đơn bội giữa các dân tộc chúng tôi sử dụng package pegas [86] và ape [87] trong R để so sánh các trình tự kiểu đơn bội của từng mẫu trong các dân tộc.



Hình 2.6. Giao diện settings của Arlequin

Để tính khoảng cách di truyền theo cặp (khoảng cách Φ_{ST}) giữa các dân tộc trong nghiên cứu bằng phần mềm Arlequin v3.5.2.2, chúng tôi lựa chọn mục "Population comparisons" (Hình 2.6), với số lần hoán vị 10,000 với độ tin cậy 0,05 và lựa chọn tính ma trận khoảng cách di truyền dựa trên số lượng alen khác nhau. Ma trận khoảng cách di truyền và giá trị p tương ứng được ghi vào tệp kết quả XML và được lưu trữ ở hai mục lần lượt là "//PairFstMat" và "//PairFstPvalMat". Ma trận này sẽ tiếp tục được sử dụng để tính toán các giá trị nonmetric multidimensional scaling (MDS) bằng function isoMDS của package MASS [88] trong R. Kết quả MDS sẽ được biểu diễn dưới dạng biểu đồ hai chiều hoặc biểu đồ heatplot từ giá trị MDS năm chiều với các giá trị của mỗi chiều được chuẩn hóa về khoảng 0 đến 1.

Phân tích phương sai phân tử (Analysis of Molecular Variance - AMOVA) được tính toán trong Arlequin bằng cách lựa chọn mục "Standard

AMOVA computations (haplotypic format)" trong mục "AMOVA" của "Settings" (Hình 2.6).

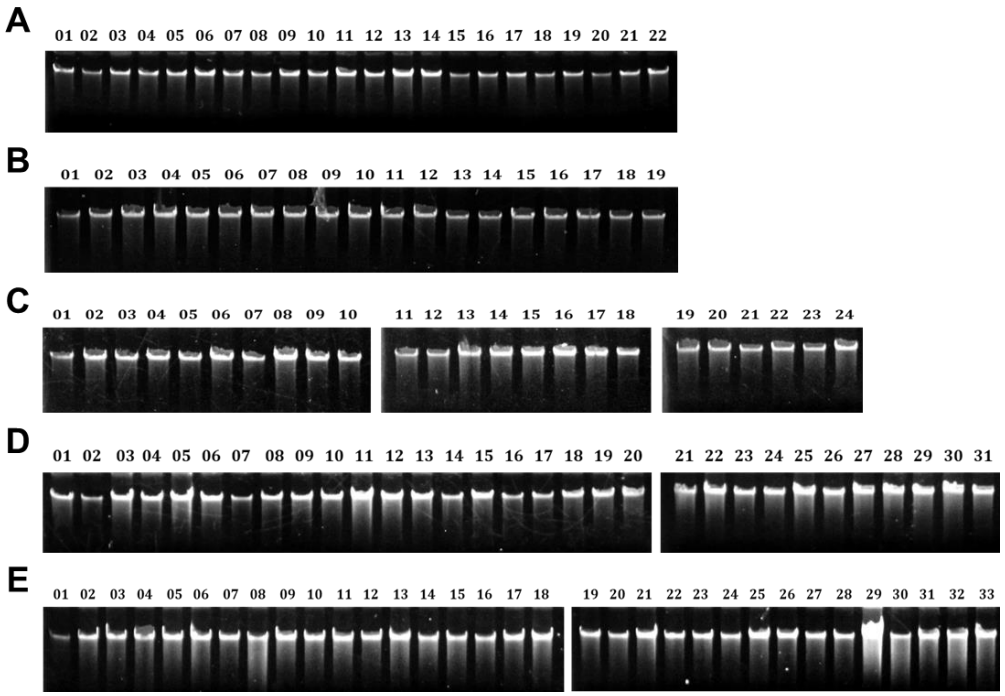
2.2.8. Kiểm định xác suất thống kê

Kiểm định Mann-Whitney U (hai phía) được sử dụng để xác định mối quan hệ giữa các nhóm ngữ hệ ở Việt Nam qua giá trị *H*, π và MPD của mỗi dân tộc, đồng thời kiểm định này cũng được sử dụng để kiểm tra sự khác biệt giữa số lượng biến thể trên mtDNA và NST Y tìm thấy trên các cá thể thuộc hai nhóm khác nhau. Kiểm định G được sử dụng để so sánh tần suất xuất hiện của một số biến thể nổi bật trong tất cả các nhóm và giữa các nhóm. Tất cả giá trị p nhỏ hơn 0,05 đều được coi là có ý nghĩa.

Chương 3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả tách chiết và tinh sạch ADN tổng số từ mẫu máu

Quy trình tách chiết ADN tổng số được thực hiện như đã trình bày trong mục 2.2.1. Nồng độ và độ tinh sạch ADN của các mẫu nghiên cứu được đánh giá bằng phương pháp đo quang phổ và phương pháp điện di trên gel agarose 0,8%. Kết quả điện di của ADN tổng số tách chiết được từ mẫu máu cho thấy các băng ADN sắc nét và sáng (Hình 3.1).



Hình 3.1. Kết quả điện di ADN tổng số các mẫu trên gel agarose 0,8%

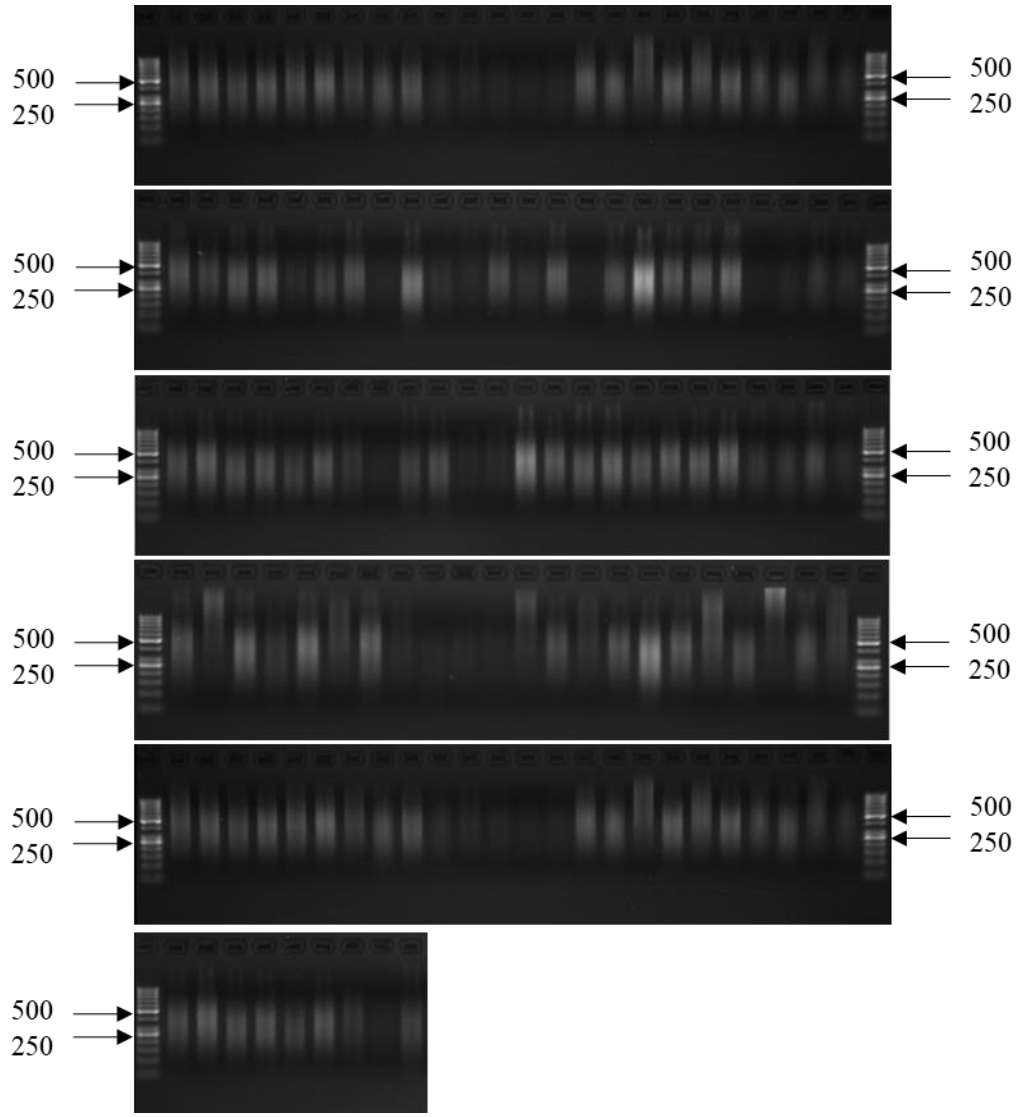
A: dân tộc Pa Kô, B: dân tộc Cơ-tu, C: dân tộc Rơ-măm, D: dân tộc KMT, E: dân tộc KMN.

Toàn bộ 129 mẫu ADN gồm 22 mẫu Pa Kô, 19 mẫu Cơ-tu, 24 mẫu Rơ-măm, 31 mẫu KMT và 33 mẫu KMN cũng được định lượng bằng máy quang phổ NanoDrop để xác định nồng độ và độ tinh sạch của mẫu. Kết quả đo nồng độ ADN của các mẫu trong nghiên cứu (Phụ lục 1) cho thấy nồng độ ADN của các mẫu đều trên 40 ng/µL và tỷ lệ giữa độ hấp thụ ở bước sóng 260 và 280nm (A260/A280) trong khoảng 1,7 - 1,9.

3.2. Kết quả thiết lập, làm giàu và giải trình tự thư viện ADN

3.2.1. Kết quả cắt phân đoạn ADN tổng số

Sau khi pha loãng, 129 mẫu ADN tổng số sẽ được tiến hành cắt phân đoạn thành các đoạn ADN ngắn. Kết quả điện di kiểm tra trên gel agarose 2% sau khi làm đứt gãy của tất cả các mẫu được thể hiện ở hình 3.2. Kết quả điện di cho thấy ADN tổng số được làm đứt gãy thành công, có kích thước băng nằm trong khoảng 200 bp đến 400 bp.

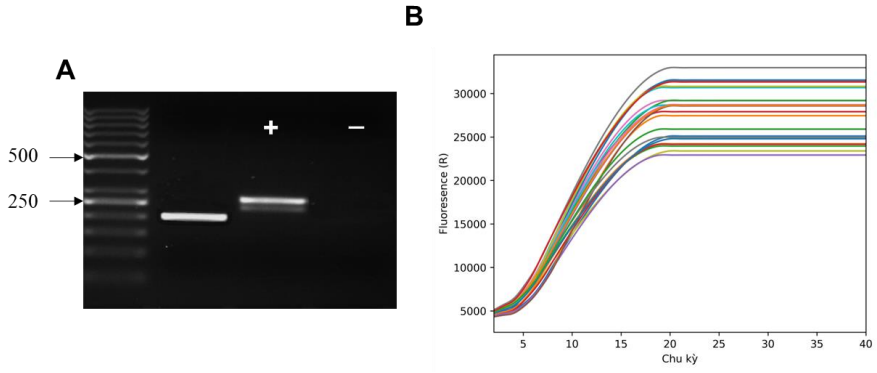


Hình 3.2. Kết quả cắt phân đoạn ADN tổng số trên gel agarose 2%

3.2.2. Kết quả làm đầy adapter

Tiếp theo, chúng tôi làm đầy adapter trên 129 mẫu ADN được làm đứt gãy ở trên cùng với 1 đối chứng dương và 1 đối chứng âm. Kết quả điện di kiểm tra ở hình 3.3A cho thấy mẫu đối chứng dương đã được gắn và làm đầy adapter thành công với kích thước lớn hơn kích thước mẫu đối chứng chưa được gắn adapter cũng như mẫu đối chứng âm không xuất hiện băng. Sau đó,

24 mẫu trong 129 mẫu được chọn ngẫu nhiên để chạy qPCR để xác định được số chu kỳ cần phải sử dụng trong phản ứng indexing PCR ở bước tiếp theo (Hình 3.3B). Kết quả xác định số chu kỳ cần sử dụng là 20 chu kỳ.

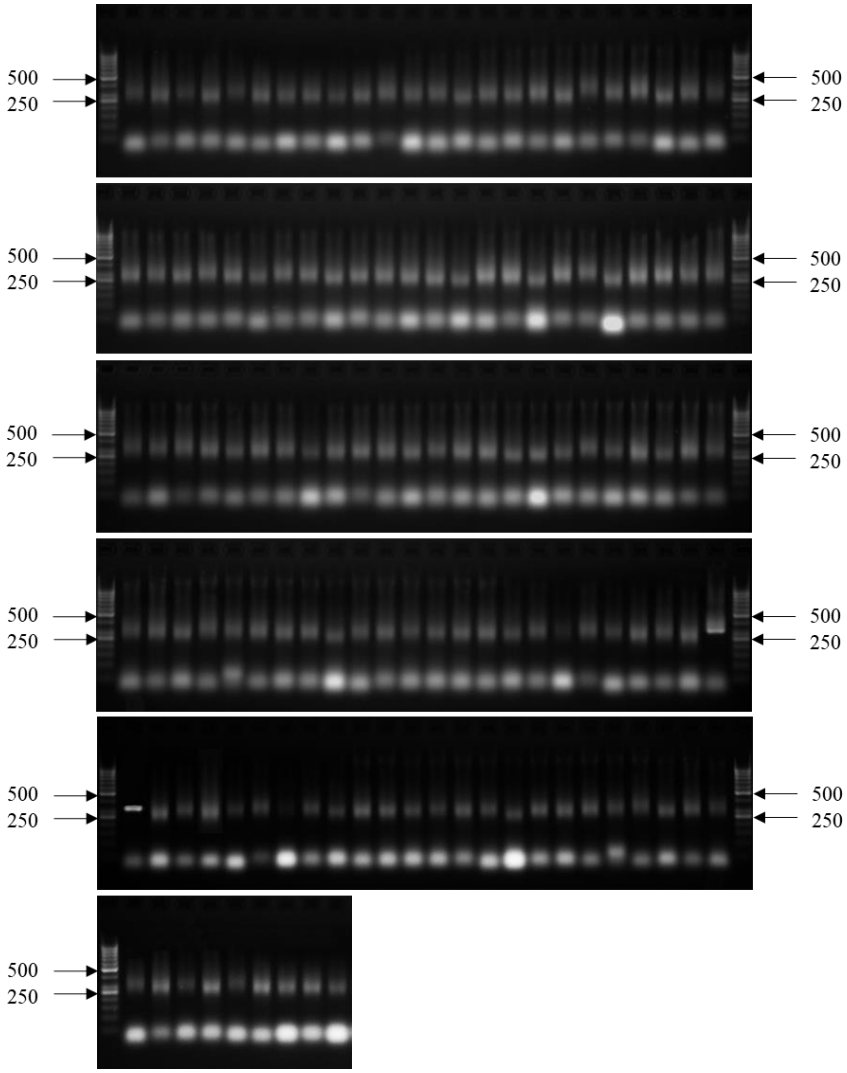


Hình 3.3. Kết quả làm đầy adapter

A) Kết quả làm đầy adapter được điện di trên gel agarose 2%. B) Kết quả qPCR định lượng sản phẩm làm đầy adapter

3.2.3. Kết quả indexing PCR

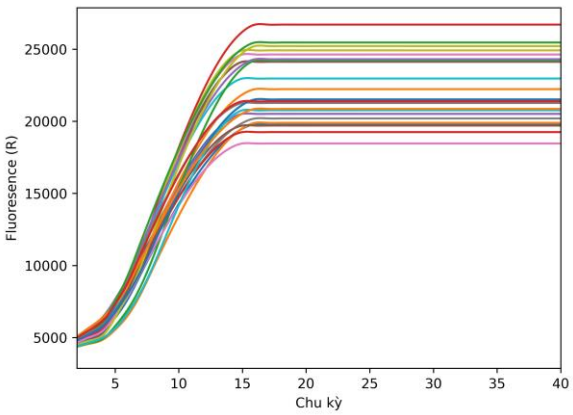
Phản ứng indexing PCR được tiến hành với số chu kỳ đã được xác định ở trên. Kết quả điện di kiểm tra chất lượng indexing PCR trên gel agarose 2% (Hình 3.4) cho thấy sản phẩm indexing PCR xuất hiện một dải băng tập trung ở vùng 200 - 400 bp giống với lý thuyết. Sau đó, chúng tôi sẽ kiểm tra chất lượng các sản phẩm indexing PCR bằng máy đo quang phổ Nanodrop. Tất cả các sản phẩm indexing đều có nồng độ đạt đủ tiêu chuẩn để sử dụng cho các bước tiếp theo.



Hình 3.4. Kết quả điện di sản phẩm indexing PCR trên gel agarose 2%

3.2.4. Kết quả định lượng thư viện sau khi làm giàu

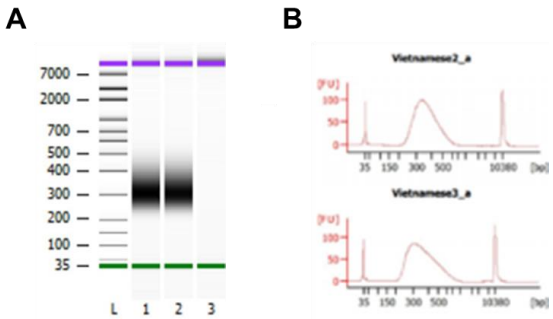
Sau khi kiểm tra nồng độ, các sản phẩm indexing PCR sẽ được góp chung thành một sản phẩm duy nhất sao cho lượng ADN của mỗi mẫu vào khoảng 200 ng để làm giàu thư viện ADN thông qua phản ứng lai – bắt giữ với ADN bait và bead. Thư viện sau khi làm giàu sẽ được định lượng bằng phản ứng qPCR (Hình 3.5). Dựa vào kết quả qPCR này, thực hiện tính toán cho ra cần thực hiện 15 chu kỳ cho phản ứng khuếch đại thư viện ADN.



Hình 3.5. Kết quả qPCR định lượng thư viện sau khi làm giàu

3.2.5. Kết quả đánh giá chất lượng thư viện ADN sau khi tinh sạch bằng Bioanalyzer

Sau khi khuếch đại, các thư viện ứng với số lượng đĩa mẫu sử dụng sẽ được mang đi tinh sạch và được đánh giá chất lượng, đo chính xác nồng độ của thư viện ADN, cũng như xác định kích thước (bp) trung bình của các đoạn ADN thông qua hệ thống máy Bioanalyzer 2100 (Hình 3.6). Kết quả phân tích cho thấy kích thước của 2 thư viện nằm trong khoảng 290 - 340 bp và nồng độ phân tử trong khoảng 16 - 27,0 nmol/L. Do đó, kết quả đã phản ảnh việc thư viện ADN đã được thiết lập và làm giàu tốt, đủ chất lượng để tiến hành giải trình tự.



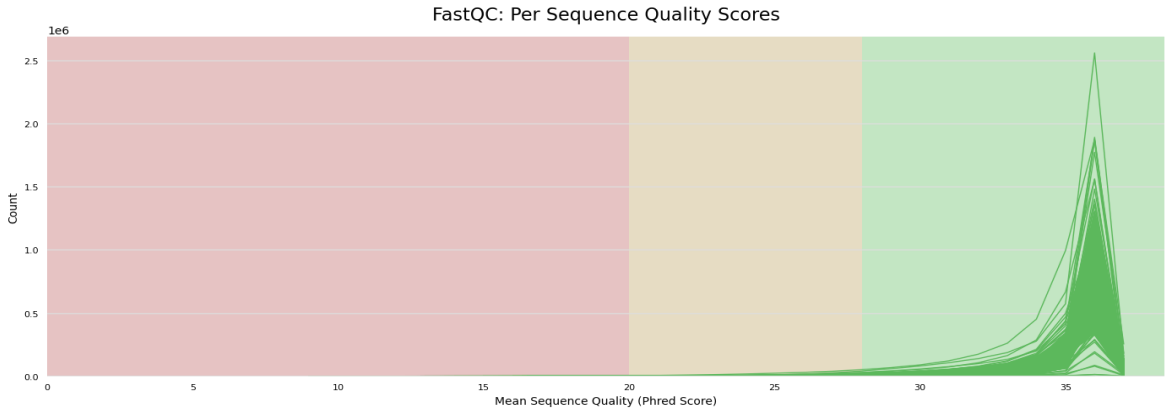
Hình 3.6. Kết quả đánh giá chất lượng thư viện ADN bằng Bioanalyzer (A). Kết quả điện di các mẫu thư viện ADN. (B) Biểu đồ kết quả của Bioanalyzer

3.3. Tiền xử lý dữ liệu

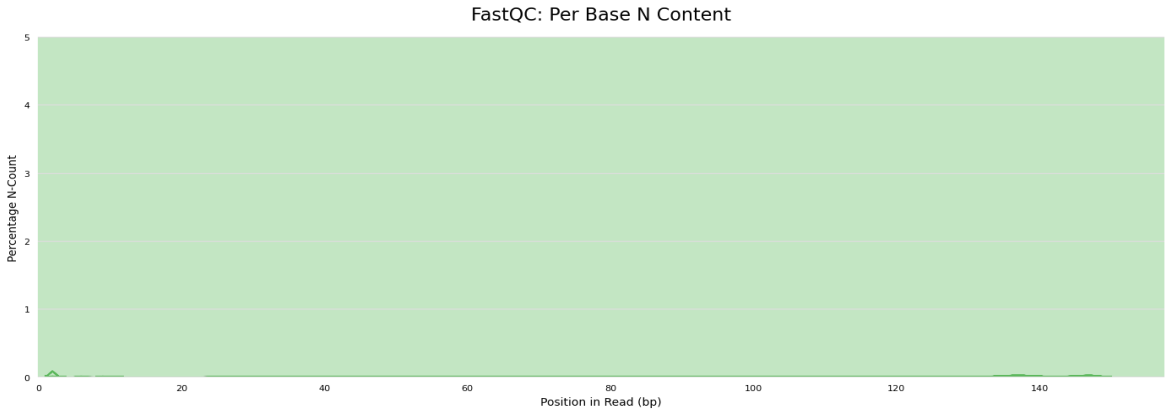
Kết quả kiểm tra chất lượng của dữ liệu sau khi giải trình tự cho thấy chất lượng của từng nucleotide đều ở ngưỡng tốt (khoảng màu xanh) (Hình 3.7), hầu hết các đoạn đọc đều ở ngưỡng tốt (khoảng màu xanh) (Hình 3.8) và

tất cả các nucleotide cả 129 mẫu đều có tần suất bị mất dữ liệu thấp (< 0,2%) (Hình 3.9). Sau khi dóng hàng với trình tự tham chiếu RSRS, 129 trình tự mtDNA có độ bao phủ dao động từ 23X - 1049X với giá trị trung bình là 203X.

Hình 3.7. Chất lượng trung bình trên từng vị trí của đoạn đọc của 129 mẫu



Hình 3.8. Chất lượng trung bình của từng đoạn đọc của 129 mẫu



Hình 3.9. Hàm lượng nucleotide N ở từng vị trí trên đoạn đọc của 129 mẫu

3.4. Phân tích dữ liệu hệ gen ty thể

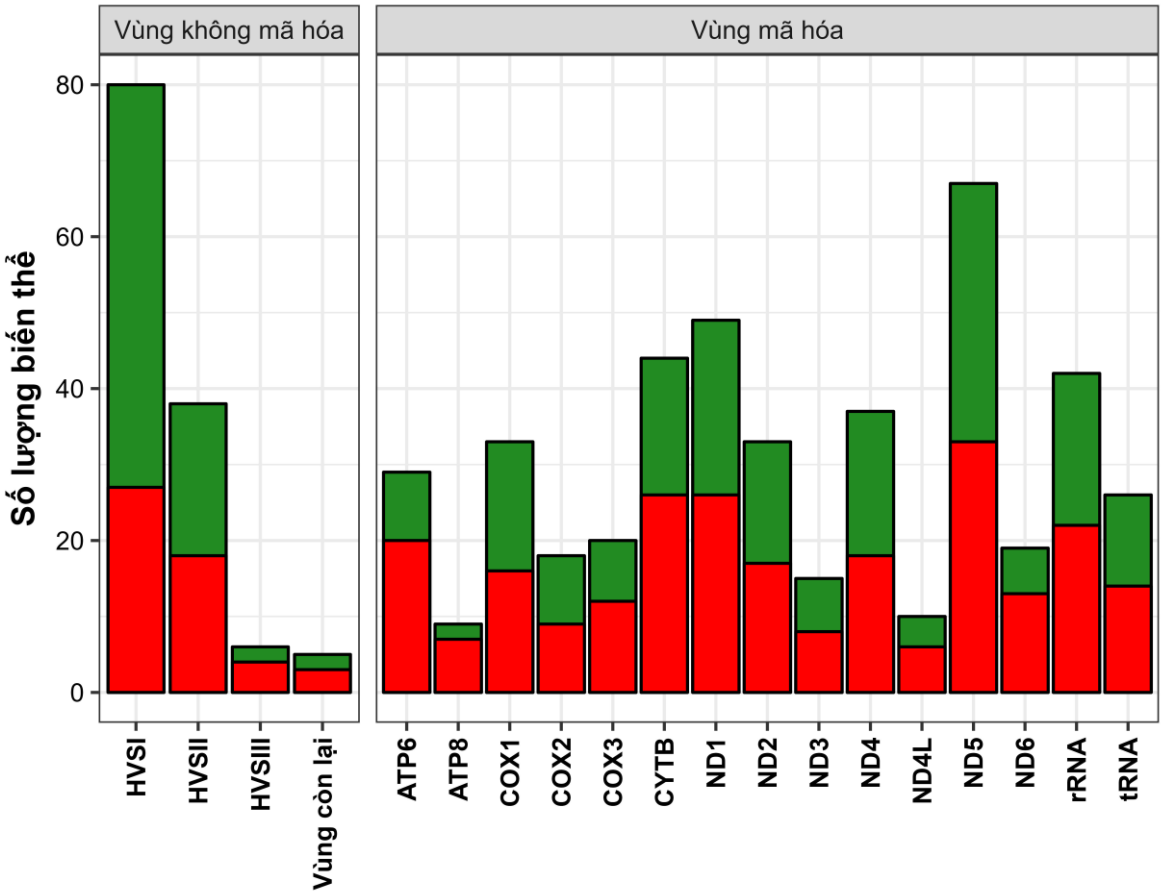
3.4.1. So sánh sự phân bố của các biến thể ở hệ gen ty thể trên các dân tộc trong nghiên cứu và các dân tộc Việt Nam

3.4.1.1. So sánh số lượng biến thể trong hệ gen ty thể

Kết quả xác định các biến thể tìm thấy 587 biến thể trong 129 trình tự mtDNA, trong đó 458 điểm nằm ở vùng mã hóa và 129 điểm nằm ở vùng điều khiển. Số lượng biến thể được tìm thấy nhiều nhất trên tất cả các dân tộc được nghiên cứu nằm ở vùng HVSI của vùng không mã hóa, còn số lượng ít nhất nằm ở vùng HVSIII và vùng còn lại của vùng không mã hóa (Hình 3.10). Số lượng biến thể chỉ xuất hiện ở một cá thể cũng được tìm thấy ở số lượng trong tất cả các vùng, chiếm từ ~50% đến ~70% số lượng biến thể ở mỗi vùng tương ứng. Tuy nhiên, vùng HVSI chỉ có khoảng 30% số lượng biến thể là biến thể chỉ xuất hiện trên một cá thể. Ở vùng mã hóa, gen *ATP8* có số lượng biến thể tìm thấy thấp nhất (9 biến thể) và gen *ND5* có số lượng biến thể tìm thấy cao nhất (67 biến thể). Khi so sánh số lượng biến thể ở hệ gen ty thể giữa các cặp dân tộc trong nghiên cứu khác nhau bằng kiểm định Mann-Whitney U (Bảng 3.1), nghiên cứu không tìm thấy sự khác biệt đáng kể giữa KMT và KMN và giữa KMT với Rơ-măm. Sự khác biệt đáng kể giữa KMN và người Rơ-măm chỉ xuất hiện ở gen *ND2*. Người Pa Kô và Cơ-tu có số lượng biến thể ở các vùng tương đối đồng đều, chỉ có sự khác biệt đáng kể ở gen *ND2*.

Kiểm định Mann-Whitney U được sử dụng để so sánh số lượng biến thể tìm thấy giữa từng cặp dân tộc trong các dân tộc Việt Nam được nghiên cứu mới, và giữa từng cặp ngữ hệ trong năm ngữ hệ trong nước (Hình 3.11, 3.12). Đối với năm nhóm NHNA trong nghiên cứu, không có sự khác biệt đáng kể giữa số lượng biến thể tìm thấy ở vùng điều khiển (Hình 3.11A) còn ở vùng mã hóa chỉ có dân tộc Rơ-măm có số lượng điểm nhiều hơn đáng kể so với dân tộc Pa Kô và Cơ-tu (Hình 3.11B). Các cá thể người Cơ-tu đều có khoảng phân bố số lượng biến thể tìm thấy trên từng cá thể hẹp nhất trên cả hai vùng mã hóa và điều khiển, trong khi các cá thể dân tộc Rơ-măm có khoảng phân bố lớn, đặc biệt là ở vùng điều khiển. Kết quả này có thể cho thấy sự đa dạng di truyền dòng mẹ của người Rơ-măm và sự tương đồng theo dòng mẹ của người Cơ-tu-kết quả của hiện tượng sống cô lập trong thời gian dài. Khi so sánh số lượng biến thể ở các cặp ngữ hệ khác nhau, sự khác biệt đáng kể duy nhất tìm được ở vùng mã hóa là giữa cặp của NHHT với các ngữ hệ khác (Hình 3.12B) còn ở

vùng điều khiển xuất hiện nhiều sự khác biệt đáng kể giữa các cặp ngữ hệ hơn (Hình 3.12A). Ở vùng điều khiển, các cá thể NHNĐ và NHHT có sự khác biệt về số lượng biến thể trong nhóm lớn nhất còn sự khác biệt này ở các cá thể NHHM là nhỏ nhất. Outlier lớn nhất và nhỏ nhất ở vùng không mã hóa được tìm thấy lần lượt trong nhóm NHNA và NHNĐ còn ở vùng mã hóa kết quả ngược lại được tìm thấy (Hình 3.12A, 3.12B).



Hình 3.10. Số lượng biến thể tìm thấy trên hệ gen ty thể của năm dân tộc trong nghiên cứu

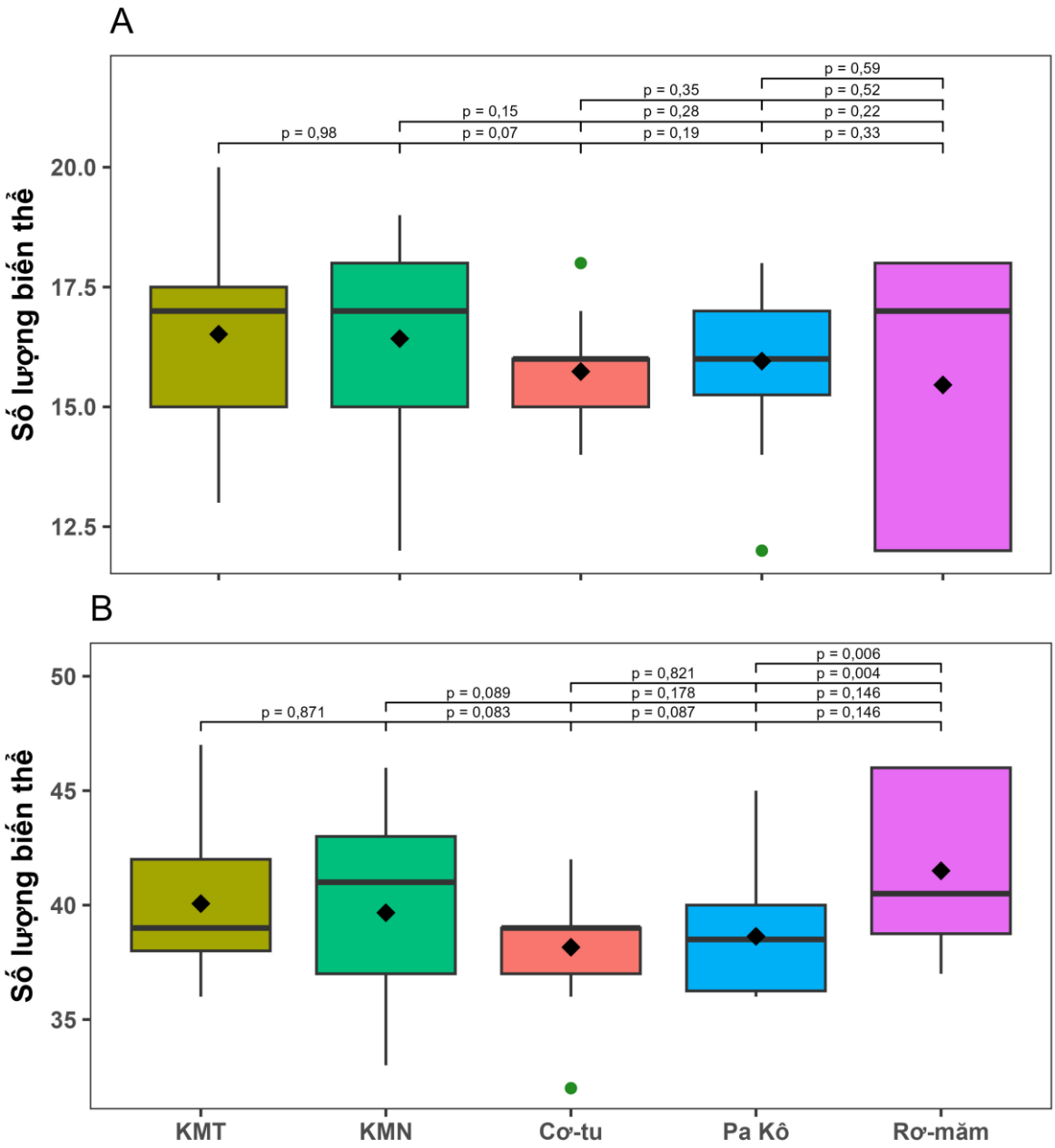
Phần màu đỏ của thanh thể hiện số lượng biến thể chỉ tìm thấy trên một cá thể, phần màu đỏ và xanh thể hiện tất cả biến thể tìm thấy trên các vùng tương ứng.

Bảng 3.1. Giá trị p khi so sánh số lượng biến thể trong các vùng khác nhau giữa các cặp dân tộc được nghiên cứu

Vùng	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**
HVSI	0,769	0,552	0,561	0,269	0,510	0,448	0,300	0,867	0,248	0,133

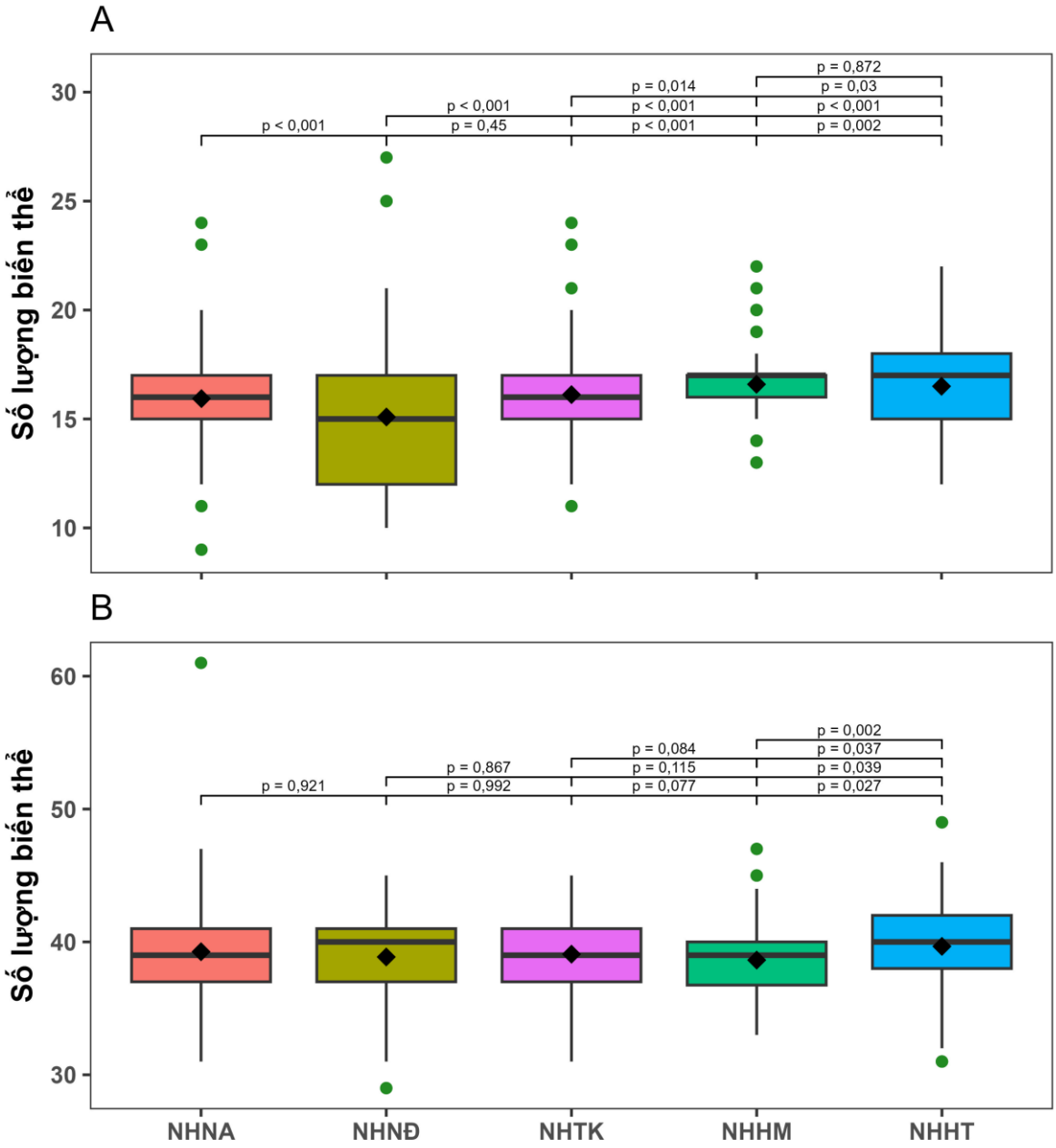
HVSII	0,957	0,163	0,809	0,109	0,173	0,727	0,173	0,318	0,003	0,122
HVSIII	0,603	0,552	0,192	0,284	0,910	0,418	0,636	0,543	0,762	0,687
Vùng không mã hóa còn lại	0,093	0,163	0,218	0,966	0,003	0,004	0,095	0,846	0,193	0,251
rRNA	0,255	0,089	0,021	0,358	0,821	0,458	0,602	0,632	0,837	0,564
tRNA	0,689	0,060	0,635	0,102	0,047	0,448	0,307	0,227	0,001	0,057
ND1	0,557	0,015	0,045	0,095	0,192	0,190	0,051	0,651	0,006	0,010
ND2	0,446	0,006	0,375	0,143	0,034	0,898	0,027	0,030	0,000	0,014
COX1	0,291	0,210	0,040	0,463	0,043	0,005	0,108	0,305	1,000	0,492
COX2	0,371	0,890	0,405	0,886	0,234	0,062	0,381	0,414	0,744	0,251
COX3	0,176	0,331	0,405	0,073	0,036	0,727	0,307	0,108	0,080	0,320
ATP8	0,883	0,463	0,436	0,423	0,375	0,352	0,339	1,000	1,000	1,000
ATP6	0,199	0,539	0,490	0,550	0,057	0,615	0,475	0,189	0,104	0,991
ND3	0,371	0,015	0,043	0,584	0,001	0,003	0,101	0,400	0,018	0,077
ND4	0,841	0,463	0,907	0,528	0,299	0,714	0,648	0,476	0,127	0,452
ND4L	0,709	0,692	0,239	0,584	0,940	0,105	0,339	0,127	0,350	0,520
ND5	0,689	0,552	1,000	0,414	0,865	0,753	0,659	0,460	0,913	0,534
ND6	0,471	0,372	0,573	0,730	0,120	0,898	0,292	0,164	0,586	0,377
CYTB	0,593	0,663	0,169	0,046	0,880	0,512	0,136	0,578	0,134	0,342

1: KMT, 2: KMN, 3: Cơ-tu, 4: Pa Kô, 5: Rơ-măm.



Hình 3.11. Biểu đồ số lượng biến thể ở mtDNA của năm dân tộc trong nghiên cứu

Điểm hình thoi thể hiện giá trị trung bình của số lượng biến thể tìm được trong một cá thể của các dân tộc tương ứng. A: vùng điều khiển. B: vùng mã hóa.



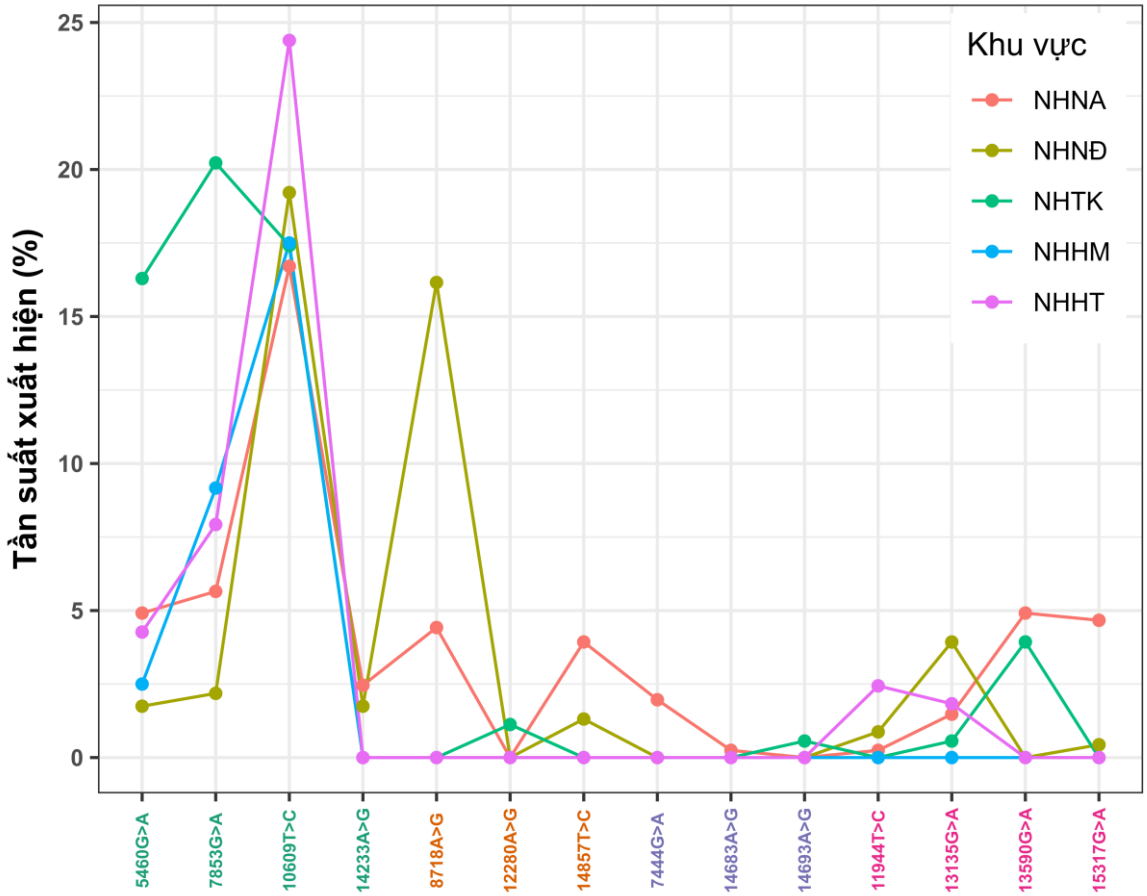
Hình 3.12. Biểu đồ số lượng biến thể ở mtDNA của năm ngữ hệ tại Việt Nam

Điểm hình thoi thể hiện giá trị trung bình của số lượng biến thể tìm được trong một cá thể của ngữ hệ tương ứng. A: vùng điều khiển. B: vùng mã hóa.

3.4.1.2. So sánh tần suất xuất hiện của một số biến thể nổi bật ở hệ gen ty thể

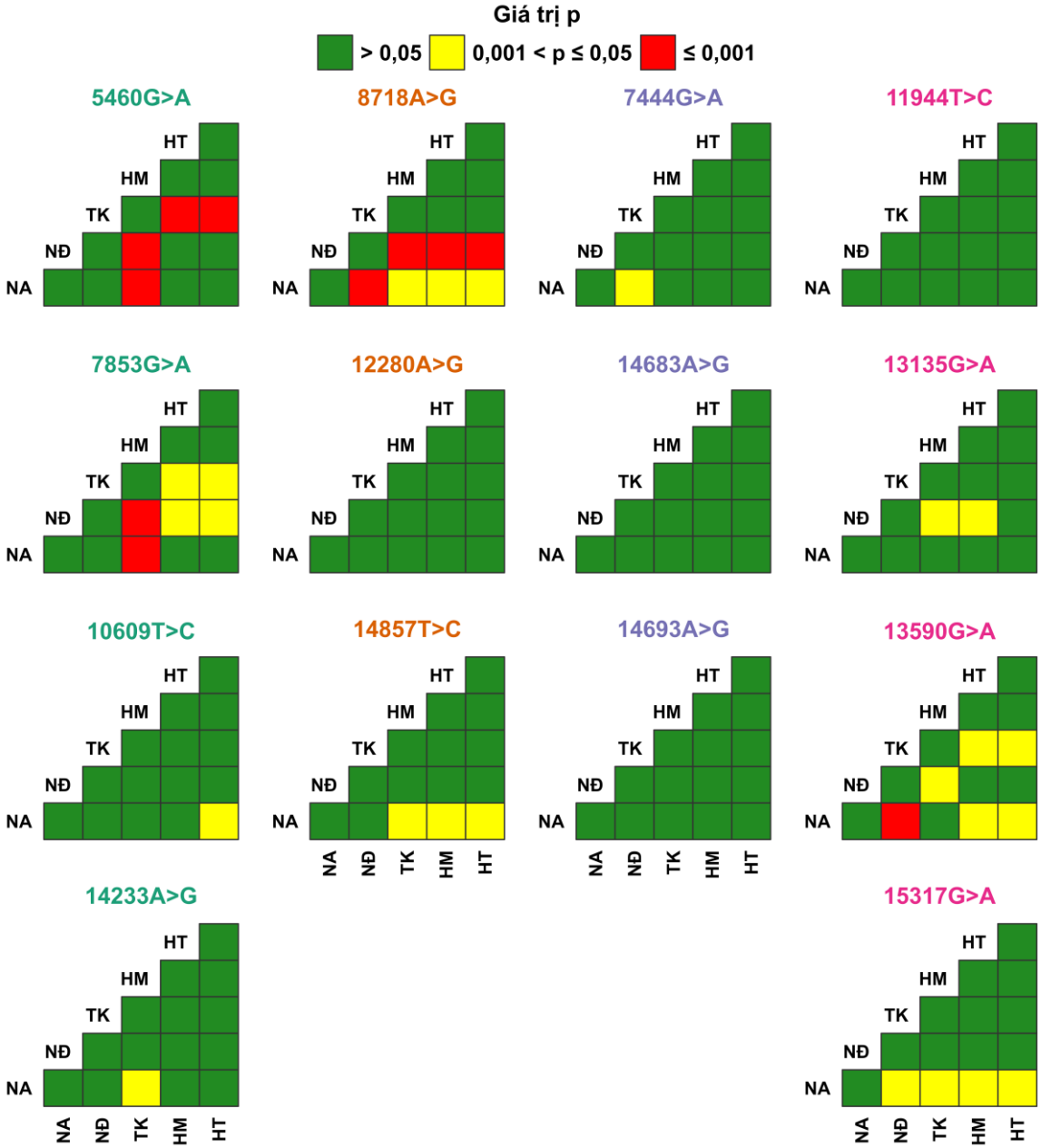
Các biến thể sẽ được lọc qua cơ sở dữ liệu MITOMAP [89] và các nghiên cứu mối quan hệ giữa các biến thể trên mtDNA với bệnh để chọn các biến thể nổi bật. Sau khi lọc, nhóm nghiên cứu không phát hiện được biến thể nào liên quan tới bệnh ở vùng điều khiển và phát hiện tổng cộng 14 biến thể bao gồm 3 điểm liên quan tới bệnh mất thính giác không hội chứng (7444G>A,

14683A>G, 14693A>G) [90], 4 điểm liên quan đến bệnh béo phì (5460G>A, 7853G>A, 10609T>C, 14233A>G) [91], 4 điểm liên quan đến bệnh đau nửa đầu (13590G>A, 11944T>C, 15317G>A, 13135G>A) [92] và 3 điểm liên quan đến bệnh cận thị nặng (8718A>G, 12280A>G, 14857T>C) [93] xuất hiện trong vùng mã hóa của các cá thể người Việt (Hình 3.13). Trong đó, 5 điểm (14683A>G, 14693A>G, 12280A>G, 11944T>C, 7444G>A) có tần suất xuất hiện nhỏ hơn 1% trong tập mẫu, đặc biệt là 2 điểm (14683A>G, 14693A>G) có tần suất nhỏ hơn 0,1%.



Hình 3.13. Tần suất xuất hiện của các biến thể trên mtDNA liên quan đến bệnh trên năm ngữ hệ ở Việt Nam

Các biến thể được đổ màu tên dựa vào bệnh liên quan với màu xanh lá cây là bệnh béo phì, màu da cam là bệnh cận thị nặng, màu xanh tím là bệnh mất thính giác không hội chứng và màu tím hồng là bệnh đau nửa đầu.



Hình 3.14. Giá trị p khi so sánh tần suất xuất hiện của các biến thể giữa các cặp ngữ hệ

Các biến thể được đổ màu tên dựa vào bệnh liên quan với màu xanh lá cây là bệnh béo phì, màu da cam là bệnh cận thị nặng, màu xanh tím là bệnh mất thính giác không hội chứng và màu tím hồng là bệnh đau nửa đầu.

Các nghiên cứu trước đây đã xác định được mối liên hệ giữa các nhóm đơn bội ty thể cũng như các điểm biến đổi trên mtDNA đối với bệnh béo phì, bao gồm cả mối liên hệ tích cực (có ít lượng mỡ cơ thể hơn) [94,95] và liên hệ tiêu cực (tăng nguy cơ bị béo phì) [95–97]. Trong bốn biến thể xuất hiện ở các cá thể Việt Nam (Hình 3.13), ba điểm có sự khác biệt đáng kể giữa các ngữ hệ

khác nhau (5460G>A, 7853G>A, 14233A>G) (Hình 3.14), trong đó 7853G>A có liên hệ tiêu cực với khả năng bị bệnh béo phì [91]. Trong đó, điểm 14233A>G là điểm đột biến được dùng để nhận biến nhóm đơn bội T- nhóm đơn bội đã được nhiều nghiên cứu chỉ ra là làm tăng nguy cơ bị béo phì [96,97], và cũng được tìm thấy có mối liên hệ tích cực với nguy cơ bị bệnh béo phì ở quần thể người Kuwait [91], còn điểm 5460G>A nằm trên gen *ND2* của mtDNA- gen đã được phát hiện có các biến thể liên quan tới khối lượng mã trong cơ thể và nguy cơ có BMI cao [94,98]. Biến thể không có sự khác biệt đáng kể về tần suất xuất hiện giữa các ngữ hệ (10609T>C) là chỉ thị cho một nhánh con của nhóm đơn bội F (F1) và có liên hệ tiêu cực với khả năng bị bệnh béo phì [91] đồng thời cũng có mối liên hệ tích cực tới khả năng chạy trong quần thể người Hàn Quốc [99]. Tần suất cao của điểm này trong quần thể người Việt có thể liên quan đến việc nhóm đơn bội F1 là một trong những nhóm chính ở MSEA [61]. Ngoại trừ tần suất thấp của 7853G>A ở NHNĐ, sự khác biệt giữa tần suất xuất hiện của ba biến thể liên quan đến bệnh béo phì ở các cá thể thuộc các ngữ hệ khác nhau trong Việt Nam hầu hết đều xảy ra giữa NHTK và các ngữ hệ khác (Hình 3.14). Mặc dù các cá thể NHTK có tần suất của điểm 5460G>A cao, tuy nhiên nhiều cá thể NHTK cũng mang các biến thể có lợi như 7853G>A và 10609T>C.

Bệnh cận thị nặng có nguy cơ toàn cầu ở mức 4%, đặc biệt là ở các dân tộc Đông Á [100], đồng thời nó cũng thường đi kèm với việc tăng nguy cơ bị các biến chứng viễn thị nghiêm trọng, có thể dẫn tới việc suy giảm thị lực nặng nề hoặc mù lòa [101]. Stress oxy hóa ở ty thể được chỉ ra là có liên quan đến cận thị và sự hình thành của cận thị nặng [102,103]. Nhìn chung, cả ba biến thể (8718A>G, 12280A>G, 14857T>C) đều xuất hiện ở tần suất thấp trên quần thể người Việt ngoại trừ tần suất của 8718A>G ở nhóm NHNĐ, trong đó điểm 12280A>G chỉ xuất hiện trên hai cá thể NHTK (Hình 3.13, 3.14). Việc hai điểm 8718A>G và 14857T>C phần lớn chỉ được tìm thấy trên các nhóm đơn bội cụ thể [93] có thể giải thích cho kết quả quan sát được, với 8718A>G (liên kết với nhóm đơn bội M71- nhóm đơn bội xuất hiện ở nhiều cá thể NHNĐ ở Việt Nam [64]) có tần suất cao ở nhóm NHNĐ và 14857T>C (liên kết với nhóm D4e3) hầu hết chỉ được tìm thấy trên người NHNA. Điều này cũng có thể chỉ ra rằng việc mang một số nhóm đơn bội cụ thể có thể làm tăng nguy cơ bị bệnh cận thị nặng [93].

Mất thính giác (điếc) là một trong những vấn đề sức khỏe thường gặp trong cuộc sống với tỉ lệ mắc bệnh là 1 trên 1000 trẻ sơ sinh [104]. Cho đến nay, nguyên nhân dẫn đến hiện tượng này vẫn chưa được hiểu rõ nhưng ngày càng có nhiều bằng chứng chỉ ra rằng các yếu tố di truyền và môi trường có thể gây ra bệnh mất thính giác [105]. Các biến thể trên mtDNA đã được phát hiện là nguyên nhân lớn gây ra mất thính giác do aminoglycoside và không hội chứng ở nhiều trường hợp trên thế giới [106–108]. Cũng giống với các biến thể liên quan đến bệnh cận thị nặng, các biến thể liên quan đến bệnh mất thính giác không hội chứng xuất hiện ở tần suất thấp trong các cá thể người Việt Nam với hai trên ba điểm tìm thấy chỉ xuất hiện ở một cá thể (14683A>G, 14693A>G) (Hình 3.13, 3.14). Sự khác biệt đáng kể duy nhất về tần số xuất hiện của biến thể giữa các ngữ hệ được tìm thấy ở điểm 7444G>A ($p = 0,0031$), cụ thể hơn là giữa NHNA và NHNĐ ($p = 0,0452$) (Hình 3.14). 7444G>A là điểm biến đổi nằm ở codon kết thúc khiến đầu C của chuỗi polypeptit này có thêm ba axit amin và có khả năng làm giảm mức độ biểu hiện của $tARN^{Ser(UCN)}$, và khi có thêm sự hiện diện của 14692A>G, bệnh mất thính giác không hội chứng có thể biểu hiện ở mức độ cao hơn [90]. Tuy nhiên, không có cá thể nào mang cả hai điểm này được phát hiện trong bộ mẫu hiện tại.

Đau nửa đầu được đề xuất là một trong những bệnh thần kinh có thể được giải thích bằng đột biến mtDNA [92]. Điều này được ủng hộ bởi việc rối loạn chức năng ty thể có liên quan đến tình trạng tăng kích thích thần kinh, sự thất bại của quá trình tổng hợp ATP và các bất thường về mạch máu não, và hậu quả của những biểu hiện trên đối với bệnh sinh lý của chứng đau nửa đầu [109]. Thêm vào đó, độ phổ biến của di truyền từ mẹ của chứng đau nửa đầu cũng ủng hộ ảnh hưởng của các đột biến mtDNA đối với bệnh này [110]. Trong bốn biến thể phát hiện ở nhóm người Việt Nam, sự khác biệt đáng kể lớn nhất giữa các ngữ hệ được quan sát thấy ở điểm 13590G>A và 15317G>A (Hình 3.13, hình 3.14). Hai điểm này chủ yếu được tìm thấy trên quần thể người NHNA với tần suất ~5%, trong đó, 13590G>A và 15317G>A lần lượt nằm trên hai gen *ND5*- thuộc nhóm gen *ND* của ty thể có vai trò mã hóa các thành phần cần thiết của phức hợp I [111] và *CYTB*- có vai trò mã hóa một phần của phức hợp III (cytochrome b) [112]. Đột biến trên các gen này cũng đã được phát hiện có liên quan đến nhiều bệnh về thần kinh và cơ như bệnh cơ ty thể [113–115], và hiện tượng này có thể được giải thích bởi sự phụ thuộc của các mô như mô thần kinh

và cơ vào quá trình trao đổi chất hiếu khí và hệ thống phosphoryl oxy hóa của ty thể [92]. Vì vậy bất kỳ sự rối loạn nào trong cơ chế sản xuất năng lượng, đặc biệt là trong chuỗi vận chuyển điện tử hô hấp đều có thể dẫn đến các triệu chứng của bệnh thần kinh và cơ [92].

Nhìn chung, hầu hết các biến thể liên quan đến bệnh xuất hiện trong quần thể người Việt đều được tìm thấy trên người NHNA, đồng thời tần suất xuất hiện của các điểm này giữa nhóm NHNA và NHNĐ không có sự khác biệt rõ rệt (ngoại trừ điểm biến đổi 8718A>G, 13590G>A và 15317G>A). Ngoài các điểm liên quan đến bệnh béo phì, các cá thể NHTK, NHHM và NHHT hầu như không có sự khác biệt về tần suất xuất hiện của 14 biến thể.

3.4.2. Kết quả phân tích đa dạng nucleotide và kiểu đơn bội ty thể của 32 dân tộc Việt Nam

Trình tự mtDNA hoàn chỉnh của 32 dân tộc Việt Nam (17 dân tộc ở nghiên cứu của Macholdt và cộng sự [49], 10 dân tộc ở nghiên cứu của Thảo và cộng sự [64] và 5 dân tộc ở nghiên cứu hiện tại) được sử dụng để tính các giá trị đa dạng di truyền như *H*, π và MPD (Bảng 3.2, hình 3.15). Hai quần thể Kinh mới (KMT và KMN) cũng có các giá trị *H* và π tương ứng với quần thể KMB đã được nghiên cứu [49], trong khi dân tộc Rơ-măm lại có cả hai giá trị *H* và π thấp hơn so với các dân tộc NHNA ở khu vực Tây Nguyên. Mặc dù dân tộc Pa Kô và Cơ-tu thuộc cùng một nhánh con Katuic của NHNA [30] và có khoảng cách địa lý thấp, độ đa dạng của hai dân tộc này lại tương đối khác biệt do dân tộc Cơ-tu có giá trị π thấp nhất (0,0013) trong 32 quần thể người Việt Nam. Sự chênh lệch giữa hai giá trị đa dạng di truyền của dân tộc Cơ-tu thể hiện rằng mặc dù dân tộc Cơ-tu có độ đa dạng về kiểu đơn bội ngang với giá trị trung bình của các dân tộc Việt Nam, tuy nhiên sự khác biệt giữa hai trình tự mtDNA bất kì trong dân tộc này lại thấp hơn nhiều. Còn đối với người Rơ-măm, sự đối lập về giá trị *H* và π so với dân tộc Cơ-tu có thể chỉ ra rằng dân tộc Rơ-măm sống tập trung thành các cộng đồng nhỏ hơn và giữa các cộng đồng có sự khác biệt về vốn gen ty thể. Nhìn chung, hầu hết các dân tộc NHNA đều có độ đa dạng di truyền cao hơn giá trị trung bình cho tất cả các nhóm, đồng thời một số dân tộc NHNA cũng có đặc điểm riêng biệt.

Bảng 3.2. Giá trị đa dạng di truyền của 32 dân tộc Việt Nam

Dân tộc	Số lượng kiểu đơn bội	Số lượng cá thể trong quần thể	H^a	π^b	MPD[c]
Mảng	13	37	0,8739	0,0017	28,6186
KMB	49	50	0,9992	0,0022	39,5184
Pa Kô	16	22	0,9697	0,0018	30,5974
Cơ-tu	17	19	0,9825	0,0013	21,8480
KMT	31	31	10,000	0,0022	37,1828
Hrê	24	30	0,9724	0,0022	36,5977
Ba na	29	36	0,9889	0,0022	37,6587
Rơ-măm	11	24	0,9094	0,0021	35,5036
Mnông	38	53	0,9811	0,0023	38,3077
Mạ	23	26	0,9908	0,0023	38,8800
Cơ ho	24	46	0,9478	0,0022	38,5981
KMN	33	33	10,000	0,0022	37,3106
Gia-rai-I	15	27	0,8832	0,0018	30,8604
Gia-rai-II	21	37	0,9625	0,0022	37,0991
Eđê-I	15	24	0,9312	0,0017	29,7464
Eđê-II	25	43	0,9313	0,0019	32,2924
Churu	31	44	0,9767	0,0023	41,3605
Raglay	23	37	0,9580	0,0020	34,0495
Chăm	10	17	0,9265	0,0022	36,2059

Tày	40	47	0,9880	0,0021	35,7280
Thái	21	24	0,9891	0,0023	37,9420
Nùng	27	37	0,9805	0,0021	34,6982
Cờ Lao	25	34	0,9733	0,0022	36,9109
La Chí	13	36	0,8921	0,0019	32,1698
Pà Thẻn	14	36	0,9000	0,0020	33,5921
Hmông	19	41	0,9354	0,0019	32,9415
Dao	21	43	0,9491	0,0020	33,7231
La Hủ	16	31	0,9247	0,0017	29,0108
Hà Nhì	23	33	0,9754	0,0022	38,2197
Phù Lá	23	35	0,9664	0,0022	37,2168
Lô Lô	10	36	0,8651	0,0017	29,0254
Si La	10	29	0,8325	0,0016	27,3645

a: Độ đa dạng kiểu đơn bội được tính bằng công thức

$$H = \left(1 - \sum x_i^2\right) \times \frac{n}{n-1}$$

x là tần suất của một kiểu đơn bội trong quần thể

n là số lượng cá thể trong quần thể

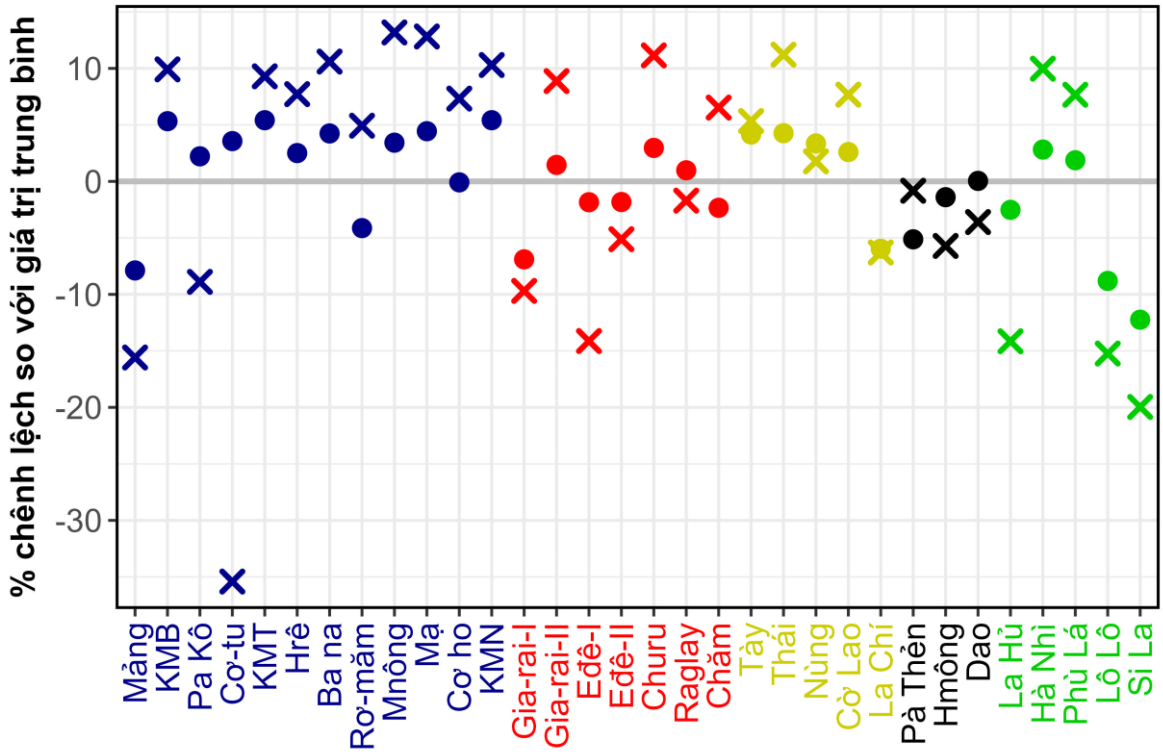
b: Độ đa dạng nucleotide được tính bằng lệnh nuc.div của gói hỗ trợ "pegas" trong R

c: Giá trị trung bình sự khác biệt nucleotide giữa hai trình tự được tính bằng công thức

$$MPD = \frac{\sum x_i}{\frac{n \times (n-1)}{2}}$$

x là số lượng nucleotide khác nhau giữa hai cá thể trong quần thể

n là số lượng cá thể trong quần thể



Hình 3.15. Thống kê đa dạng di truyền mtDNA biểu thị dưới dạng phần trăm khác biệt so với giá trị trung bình 32 dân tộc Việt Nam

Hình tròn: giá trị *H* của dân tộc tương ứng. Dấu X: giá trị đa dạng nucleotide của dân tộc tương ứng. Tên các dân tộc được đổ màu dựa theo ngữ hệ tương ứng của các dân tộc đó với màu xanh nước biển là NHNA, màu đỏ là NHNĐ, màu vàng là NHTK, màu đen là NHHM và màu xanh lá cây là NHHT.

Sau đấy, kiểm định Mann-Whitney U được sử đụng để xác định mối liên hệ giữa các cặp ngữ hệ (Bảng 3.3). Nhóm nghiên cứu xác định được sự khác biệt đáng kể giữa NHNA và NHNĐ, và NHNA và NHHT khi sử dụng *H*, tuy nhiên không có sự khác biệt đáng kể nào được tìm thấy đối với giá trị đa dạng MPD và π.

Bảng 3.3. Kiểm định Mann-Whitney U khảo sát mối quan hệ giữa từng cặp ngữ hệ sử dụng giá trị đa dạng di truyền *H*, π và MPD

Cặp ngữ hệ	*H* (giá trị p)	MPD (giá trị p)	π (giá trị p)
NHNA - NHNĐ	0,047	0,432	0,340
NHNA - NHTK	0,712	0,506	0,646
NHNA - NHHM	0,097	0,233	0,233
NHNA - NHHT	0,040	0,279	0,234
NHNĐ - NHTK	0,106	0,639	0,530
NHNĐ - NHHM	0,833	0,833	0,833
NHNĐ - NHHT	0,530	0,432	0,343
NHTK - NHHM	0,250	0,250	0,250
NHTK - NHHT	0,095	0,548	0,310
NHHM - NHHT	1,000	0,786	0,786

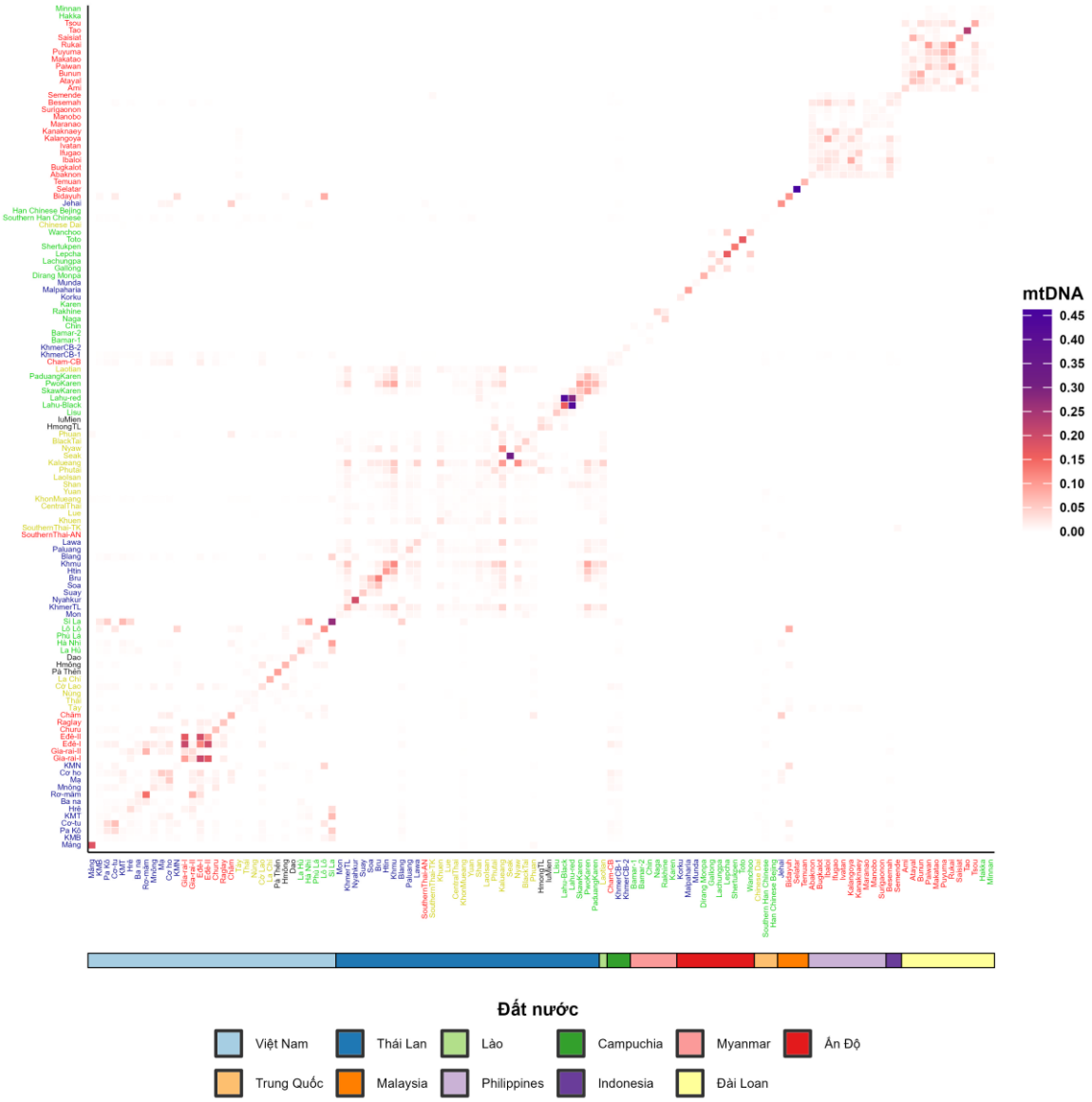
3.4.3. Mối quan hệ di truyền giữa các dân tộc

Tỉ lệ một kiểu đơn bội được tìm thấy ở trong hai cá thể ngẫu nhiên từ hai quần thể khác nhau được sử dụng để giúp đưa ra các kết luận về tổ tiên gần hay các sự kiện trao đổi thông tin di truyền gần đây (Hình 3.16). Nhìn chung, các dân tộc NHNA có tỉ lệ tương đồng kiểu đơn bội với dân tộc khác trong ngữ hệ thấp hơn NHNĐ và tương đồng đối với các cá thể cùng dân tộc thấp hơn dân tộc NHHT, có thể phản ánh sự thiếu tương tác với nhau và sự đa dạng của các dân tộc NHNA. Tương tự với kết quả tìm được ở nghiên cứu của Macholdt và công sự [49], dân tộc Mảng không có sự tương đồng về kiểu đơn bội với bất kì dân tộc nào. Kết quả tỉ lệ tương đồng kiểu đơn bội của hai quần thể KMT và KMM cũng nhất quán với Kinh miền Bắc (KMB) và kết quả đa dạng di truyền ở trên (Hình 3.15), với tỉ lệ tương đồng thấp với các dân tộc khác và với các cá thể cùng dân tộc. Dân tộc Pa Kô và Cơ-tu đều có tỉ lệ tương đồng kiểu đơn bội thấp với các dân tộc khác, ngoại trừ Si La (đối với Pa Kô) và Lô Lô (Cơ-tu). Tỉ lệ tương đồng kiểu đơn bội trong dân tộc cao ở người Rơ-măm cũng phản ánh độ đa dạng kiểu đơn bội và nucleotide thấp (Hình 3.15). Với sự bổ sung của dân tộc Rơ-măm, một sự phân nhóm mới được quan sát thấy ở trong nhóm các dân tộc NHNA ở Tây Nguyên, bao gồm nhóm Hrê và Ba na, Rơ-măm và Bana, và Cơ-ho, Mạ, Mnông so với nghiên cứu trước [64]. Khi xác định tỉ lệ này giữa các dân tộc trong nước với các dân tộc nước ngoài cũng như các dân tộc nước ngoài với nhau, kết quả chỉ ra rằng các dân tộc từ các nước khác nhau có sự

giao thoa rất hạn chế, hầu hết chỉ xuất hiện giữa dân tộc Việt Nam và một vài dân tộc ở các nước lân cận như Blang từ Thái Lan, Cham-CB và Khmer-CB-1 từ Campuchia, Jerai và Bidayuh từ Malaysia (Hình 3.16). Mặc dù đã sử dụng một tập mẫu lớn hơn, chúng tôi chỉ phát hiện được một tín hiệu tương đồng kiểu đơn bội thấp giữa dân tộc Mảng và một dân tộc khác (Phuan từ Thái Lan), điều này có thể chỉ ra rằng dân tộc Mảng đã sống cô lập trong một thời gian dài. Sự tương đồng về kiểu đơn bội giữa các dân tộc nước ngoài với các dân tộc Việt Nam chỉ được tìm thấy ở các dân tộc thuộc NHNA (Blang, Khmer-CB-1, Jehai), NHNĐ (Cham-CB, Bidayuh) và NHTK (Phuan, KhonMueang), trong đó tỉ lệ tương đồng kiểu đơn bội cao nhất đều xảy ra giữa các dân tộc thuộc các nhóm ngữ hệ khác nhau.

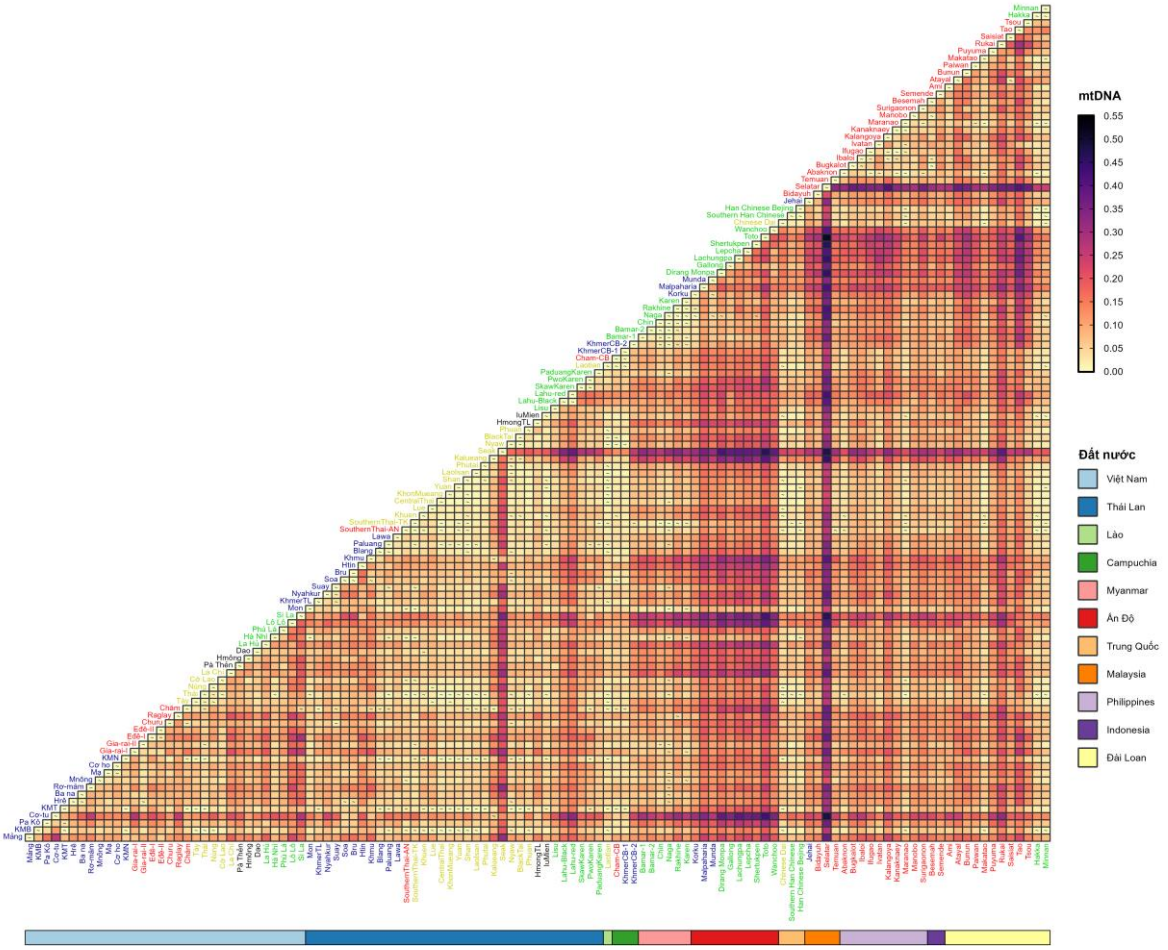
Để đánh giá độ "gần gũi" của các dân tộc, nhóm nghiên cứu sử dụng biểu đồ MDS để thể hiện ma trận khoảng cách di truyền lập bởi phần mềm Arlequin (Hình 3.17, hình 3.18). Bởi vì giá trị stress của biểu đồ MDS vẫn còn cao (> 15) [116] nên MDS 5 chiều biểu thị dưới dạng biểu đồ heatplot được sử dụng để phát hiện các yếu tố có thể chưa được biểu hiện ở biểu đồ 2 chiều (Hình 3.19). Trong đồ thị này, vị trí của các dân tộc NHNA, NHNĐ và NHHT của Việt Nam có thể chia thành ba nhóm: nhóm ở giữa biểu đồ (có khoảng cách di truyền thấp đối với các nhóm còn lại) (gồm ba quần thể người Kinh thuộc NHNA; dân tộc Churu và Chăm thuộc NHNĐ; dân tộc La Hủ, Hà Nhì và Phù Lá thuộc NHHT), nhóm NHNĐ (gồm các dân tộc NHNĐ trừ Raglay và Chăm; dân tộc Mảng, Ba na, Mnông thuộc NHNA) và nhóm Cơ-tu (dân tộc Pa Kô, dân tộc Cơ-tu thuộc NHNA). Giá trị ở chiều bốn và năm của dân tộc Rơ-măm có xu hướng khác so với các dân tộc NHNA ở khu vực Tây Nguyên và giống với dân tộc Raglay ở chiều bốn, khiến vị trí của hai dân tộc này tách ra khỏi cụm NHNĐ. Các dân tộc thuộc NHTK và NHHM đều nằm ở giữa biểu đồ ở hầu hết các chiều. Vị trí tương đối của các dân tộc NHTK, NHHM và NHHT giống với biểu đồ MDS hai chiều của Macholdt [49] còn vị trí của các nhóm NHNĐ có thể do tần suất cao của các nhóm đơn bội M (đặc trưng cho các dân tộc NHNĐ) và sự đa dạng về thành phần nhóm đơn bội giữa các dân tộc NHNĐ Việt Nam [64]. Nhất quán với kết quả tỉ lệ tương đồng kiểu đơn bội (Hình 3.16), các dân tộc nước ngoài có kiểu đơn bội tương đồng với các dân tộc trong nước tương ứng đều có khoảng cách di truyền thấp đối với các dân tộc đó. Ngoài ra, nhóm nghiên cứu cũng phát hiện được khoảng cách di truyền thấp giữa ba quần

thể Kinh và dân tộc NHNĐ và phần lớn các dân tộc NHTK từ Thái Lan, Laotian từ Lào, Cham-CB và Khmer-CB-1 từ Campuchia, các dân tộc từ Trung Quốc và hai dân tộc NHHT (Hakka và Minnan) từ Đài Loan. Kết quả này có thể có liên quan tới sự liên kết giữa nguồn gốc của các dân tộc NHNA ở Việt Nam với sự mở rộng của nền văn hóa nông nghiệp từ sông Hoàng Hà của Trung Quốc. Cụ thể hơn, sông Hoàng Hà là một trong những nơi khởi nguồn của người Hán cổ đại [117] và có liên quan đến sự lan tỏa của NHHT [118] cho nên có khả năng những người nông dân ở đây đã trao đổi thông tin di truyền với tổ tiên của người NHTK của các dân tộc NHTK ở Thái Lan và Lào và người NHHT ở Đài Loan trước khi họ di cư [34,77]. Khoảng cách di truyền không nhỏ giữa ba quần thể người Kinh với các dân tộc NHNA ở Thái Lan (ngoại trừ dân tộc Paluang) có thể được giải thích bởi một số dân tộc NHNA ở Thái Lan, đặc biệt là dân tộc Mon, có cả nguồn thông tin di truyền từ Nam Á [33,34,36]. Trong 85 dân tộc nước ngoài được sử dụng, dân tộc có khoảng cách di truyền thấp đối với dân tộc Pa Kô và Cơ-tu là hai dân tộc NHNA (Soa và Bru) và hai dân tộc NHTK (Kalueang và Seak) từ Thái Lan, tuy nhiên khoảng cách giữa dân tộc Seak và Pa Kô vẫn tương đối lớn. Việc hai dân tộc NHNA nói trên không có sự tương đồng về kiểu đơn bội với người Pa Kô và Cơ-tu có thể chỉ ra rằng cá thể của bốn dân tộc này đã tiếp xúc với nhau cách đây một khoảng thời gian dài trước khi tách hẳn ra thành các dân tộc riêng biệt. Hơn thế nữa, ngôn ngữ của người Pa Kô và Cơ-tu cũng thuộc cùng nhánh con Katuic của NHNA với người Soa và Bru [30].



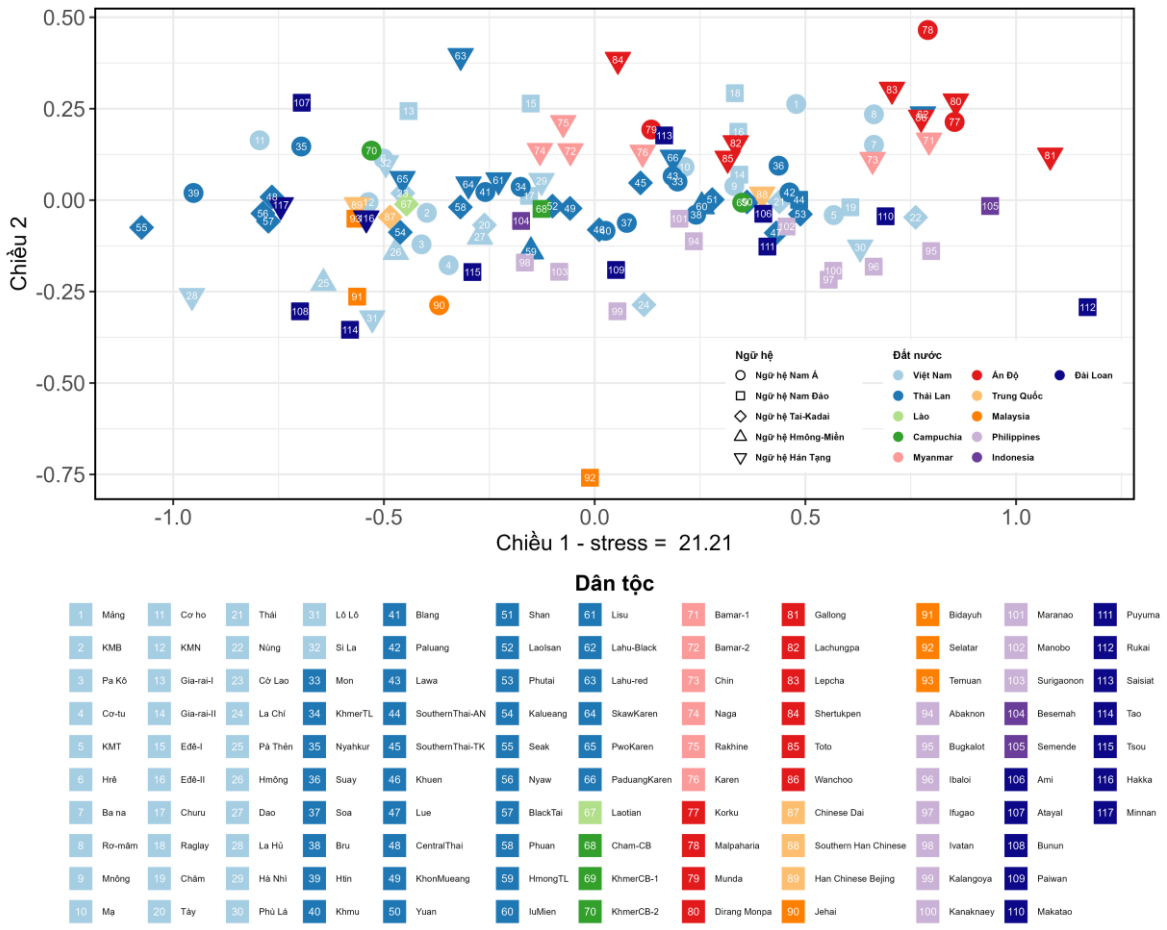
Hình 3.16. Tần suất của các kiểu đơn bội tương đồng trong cùng dân tộc/quần thể và giữa các dân tộc Việt Nam với quần thể các nước ở châu Á

Tên các dân tộc được đổ màu dựa theo ngữ hệ tương ứng của các dân tộc đó với màu xanh nước biển là NHNA, màu đỏ là NHNĐ, màu vàng là NHTK, màu đen là NHHM và màu xanh lá cây là NHHT.

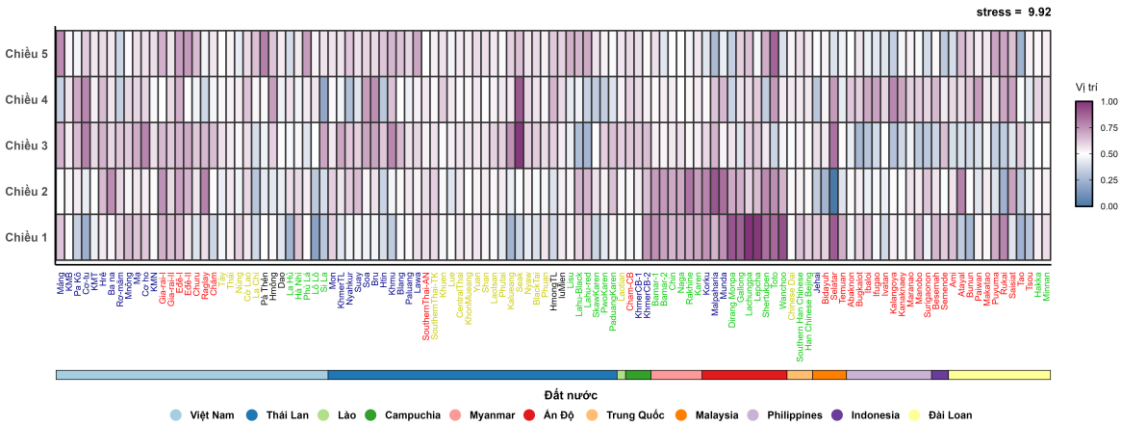


Hình 3.17. Khoảng cách di truyền Φ_{ST} theo cặp giữa các dân tộc Việt Nam và các quần thể ở châu Á tính theo hệ gen ty thể

Tên các dân tộc được đổ màu dựa theo ngữ hệ tương ứng của các dân tộc đó với màu xanh nước biển là NHNA, màu đỏ là NHNĐ, màu vàng là NHTK, màu đen là NHHM và màu xanh lá cây là NHHT. Kí hiệu "~" thể hiện khoảng cách di truyền không đáng kể ($p > 0,05$).



Hình 3.18. Biểu đồ MDS hai chiều dựa trên khoảng cách di truyền Φ_{ST} hệ gen ty thể



Hình 3.19. Biểu đồ heatplot dựa trên giá trị MDS năm chiều

Tên các dân tộc được đổ màu dựa theo ngữ hệ tương ứng của các dân tộc đó với màu xanh nước biển là NHNA, màu đỏ là NHNĐ, màu vàng là NHTK, màu đen là NHHM và màu xanh lá cây là NHHT.

3.4.4. Các yếu tố ảnh hưởng đến cấu trúc di truyền quần thể của các dân tộc Việt Nam

Để kiểm định mối quan hệ giữa các ngữ hệ hoặc vị trí địa lý với cấu trúc di truyền quần thể của các dân tộc Việt Nam, các dân tộc đã được phân loại vào các nhóm dựa trên tiêu chí ngữ hệ hoặc tiêu chí vị trí địa lý (cấp độ tỉnh và cấp độ vùng miền), rồi được phân tích bằng AMOVA (Bảng 3.4). Cả ba cách phân loại đều có giá trị tương đối giống nhau, tuy nhiên chỉ có cách phân loại theo ngữ hệ và phân loại theo vùng miền quan sát được sự khác biệt đáng kể giữa các nhóm trong một kiểu phân loại ($p < 0,01$). Vì vậy, từng cặp trong kiểu phân loại theo ngữ hệ và vùng miền đã được kiểm định thêm để xem sự khác biệt quan sát được đến từ nhóm nào. Đối với cách phân loại dựa vào ngữ hệ, NHTK có tỉ lệ sự khác biệt giữa các trình tự vùng điều khiển được giải thích bởi sự khác biệt giữa các cá thể trong một dân tộc cao nhất (95,27%) trong khi giá trị này được tìm thấy thấp nhất ở nhóm NHHT (86,34%). Sự khác biệt có ý nghĩa ở cách phân loại này ($p < 0,01$) là kết quả của sự khác biệt giữa NHNA và NHHT ($p < 0,05$), giữa NHNĐ và NHTK ($p < 0,05$), giữa NHNĐ và NHHM ($p < 0,05$) và giữa NHNĐ và NHHT ($p < 0,05$). Đối với cách phân loại dựa vào vùng miền, tất cả các nhóm đều có sự khác biệt đáng kể giữa các dân tộc trong cùng một nhóm ngoại trừ khu vực Đông Nam Bộ (ĐNB). Hiện tượng này có thể là kết quả của số lượng dân tộc ít (KMN và Chăm) trong khu vực và trình tự vùng điều khiển của quần thể KMN không có sự khác biệt đáng kể so với dân tộc Chăm, nhất quán với khoảng cách di truyền thấp giữa hai dân tộc nêu trên (Hình 3.19). Sự khác biệt đáng kể giữa hai vùng miền chủ yếu đến từ hai vùng Bắc Bộ (Tây Bắc Bộ - TBB và Đông Bắc Bộ - ĐBB) với vùng Bắc Trung Bộ (BTB) và Tây Nguyên (TN), và giữa BTB và TN. Kết quả này phù hợp với sự tách biệt giữa hai dân tộc ở vùng BTB (Pa Kô và Cơ-tu) với các dân tộc còn lại (Hình 3.19).

Bảng 3.4. Kết quả phân tích AMOVA

Phân loại dân tộc	Số nhóm	Số lượng dân tộc	Nguồn gốc sự khác biệt giữa các trình tự vùng điều khiển (%)

	phân loại	trong một nhóm	Giữa các nhóm	Giữa các dân tộc trong một nhóm	Giữa các cá thể trong một dân tộc
Tất cả dân tộc		32		8,94**	91,06
Tất cả các ngữ hệ	5	32	0,93**	8,19**	90,88**
NHNA		12		7,82**	92,18
NHNĐ		7		8,48**	91,52
NHTK		5		4,73**	95,27
NHHM		3		6,87**	93,13
NHHT		5		13,66**	86,34
NHNA - NHNĐ	2	19	0,06	8,05**	91,90**
NHNA - NHTK	2	17	0,36	6,99**	92,65**
NHNA - NHHM	2	15	1,05	7,59**	91,36**
NHNA - NHHT	2	17	1,30*	9,23**	89,47**
NHNĐ - NHTK	2	12	2,45*	6,72**	90,83**
NHNĐ - NHHM	2	10	2,89*	7,74**	89,37**
NHNĐ - NHHT	2	12	2,63*	10,35**	87,02**
NHTK - NHHM	2	8	0,37	5,49**	94,14**
NHTK - NHHT	2	10	-0,20	9,14**	91,06**
NHHM - NHHT	2	8	-1,09	11,19**	89,90**
Tất cả các tỉnh	15	32	1,11	7,92**	90,97**

Tất cả các khu vực	7	32	1,41**	7,80**	90,79**
TBB		7		9,59**	90,41
BTB		2		5,36*	94,64
NTB		2		9,94**	90,06
TN		11		6,77**	93,23
ĐNB		2		1,30	98,70
ĐBB		7		8,80**	91,20
TBB - ĐBSH	2	8	-4,56	9,89**	94,67**
TBB - BTB	2	9	6,80*	8,69**	84,51**
TBB - NTB	2	9	-0,50	9,68**	90,82**
TBB - TN	2	18	1,84**	7,65**	90,51**
TBB - ĐNB	2	9	-2,51	8,94**	93,57**
TBB - ĐBB	2	14	0,13	9,17**	90,70**
ĐBSH - BTB	2	3	3,92	4,16*	91,93**
ĐBSH - NTB	2	3	-4,53	10,10**	94,42**
ĐBSH - TN	2	12	-1,63	6,89**	94,74**
ĐBSH - ĐNB	2	3	0,66	1,35	97,99*
ĐBSH - ĐBB	2	8	-4,20	9,06**	95,14**
BTB - NTB	2	4	7,97	7,69**	84,34**
BTB - TN	2	13	7,11*	6,28**	86,61**
BTB - ĐNB	2	4	8,08	2,75*	89,17**

BTB - ĐBB	2	9	6,04*	8,08**	85,89**
NTB - TN	2	13	0,16	7,06**	92,78**
NTB - ĐNB	2	4	-1,39	6,62**	94,77**
NTB - ĐBB	2	9	1,74	8,79**	89,47**
TN - ĐNB	2	13	1,18	6,38**	92,44**
TN - ĐBB	2	18	2,95**	7,32**	89,74**
ĐNB - ĐBB	2	9	-0,62	8,11**	92,51**

*: 0,01 <= p < 0,05, **: p < 0,01, TBB: Tây Bắc Bộ, ĐBB: Đông Bắc Bộ, ĐBSH: Đồng bằng Sông Hồng, BTB: Bắc Trung Bộ, TN: Tây Nguyên, NTB: Nam Trung Bộ, ĐNB: Đông Nam Bộ. Các dân tộc được xếp vào các nhóm như sau:

NHNA: Mảng, KMB, Pa Kô, Cơ-tu, KMT, Hrê, Ba na, Rơ-măm, Mnông, Mạ, Cơ ho, KMN; NHNĐ: Gia-rai-I, Gia-rai-II, Eđê-I, Eđê-II, Churu, Raglay, Chăm; NHTK: Tày, Thái, Nùng, Cờ Lao, La chí; NHHM: Pà Thẻn, Hmông, Dao; NHHT: La Hủ, Hà Nhì, Phù Lá, Lô Lô, Si La.

Tỉnh Lai Châu: Mảng, La Hủ, Hà Nhì, Si La; tỉnh Điện Biên: Hmông, Thái; tỉnh Lào Cai: Tày; thành phố Hà Nội: KMB; tỉnh Quảng Trị: Pa Kô; tỉnh Thừa Thiên Huế: Cơ-tu; tỉnh Bình Định: KMT; tỉnh Kon Tum: Hrê, Ba na, Rơ-măm, Gia-rai-II; tỉnh Đăk Lăk: Mnông, Eđê-I, Eđê-II; tỉnh Gia Lai: Gia-rai-I; tỉnh Lâm Đồng: Mạ, Cơ ho, Churu; tỉnh Hà Giang: Nùng, Cờ Lao, La Chí, Pà Thẻn, Dao, Phù Lá, Lô Lô; tỉnh Khánh Hòa: Raglay; tỉnh Bình Phước: Chăm; thành phố Hồ Chí Minh: KMN.

ĐBB: Mảng, Hmông, Thái, Tày, La Hủ, Hà Nhì, Si La; TBB: Nùng, Cờ Lao, La Chí, Pà Thẻn, Dao, Phù Lá, Lô Lô; ĐBSH: KMB; BTB: Pa Kô, Cơ-tu; NTB: KMT, Raglay; TN: Hrê, Ba na, Rơ-măm, Mnông, Mạ, Cơ ho, Gia-rai-I, Gia-rai-II, Eđê-I, Eđê-II, Churu; ĐNB: KMN, Chăm.

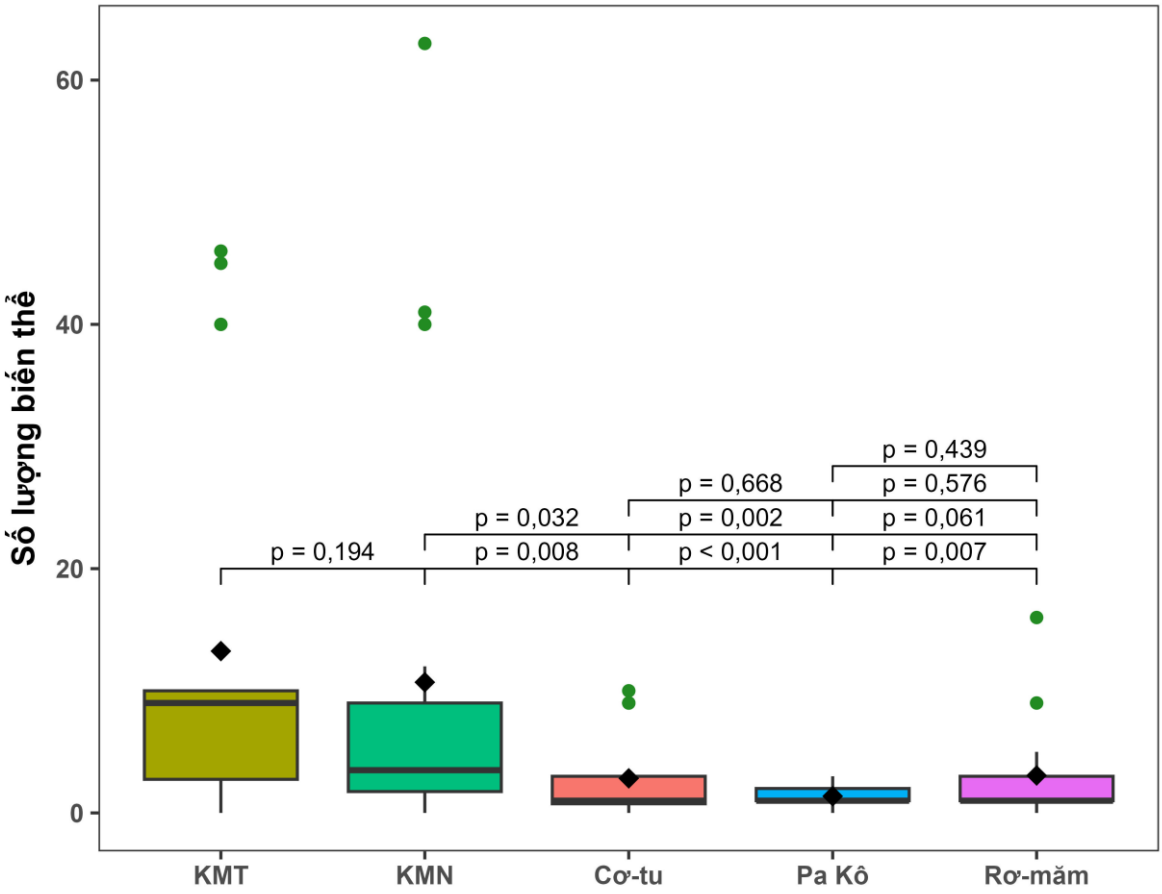
3.5. Phân tích đa hình nucleotide đơn từ nhiễm sắc thể Y

Trong 82 cá thể trong nghiên cứu, nhóm tác giả tìm thấy 189 SNP khác nhau và sau khi tích hợp với bộ dữ liệu SNP trên NST Y người Việt [49], số

lượng SNP có thể giữ lại để phân tích đa dạng di truyền quần thể là không đủ nên trong khuôn khổ nghiên cứu hiện tại chỉ có kết quả của phần so sánh số lượng và tần suất xuất hiện của các SNP được thực hiện.

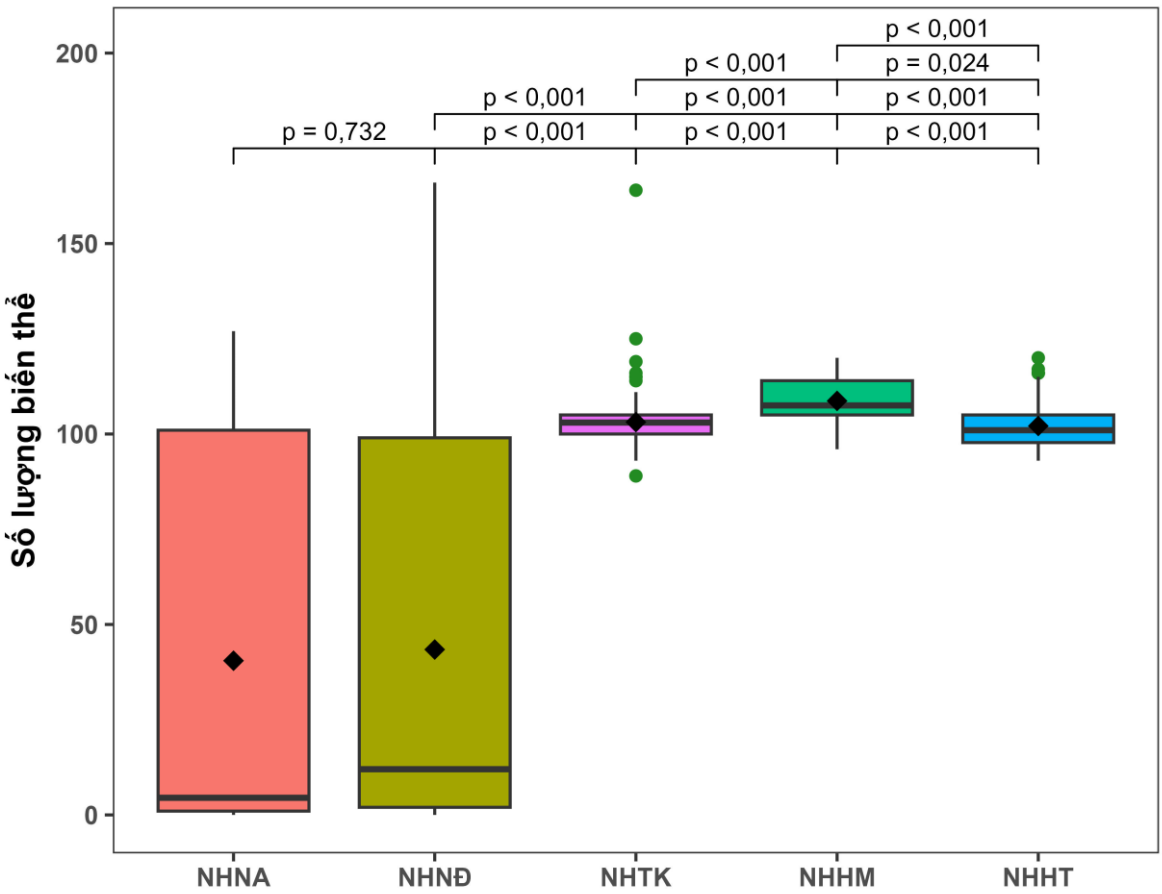
3.5.1. Kết quả phân tích số lượng đa hình nucleotide đơn trên nhiễm sắc thể Y

Nhìn chung, khác với mtDNA, số lượng SNP ở NST Y tìm được trong năm dân tộc được nghiên cứu cho thấy sự khác biệt rõ rệt giữa hai quần thể người Kinh với ba dân tộc còn lại (Hình 3.20). Các cá thể người Kinh (KMT, KMN) có sự đa dạng về số lượng biến thể trên mỗi cá thể lớn với việc có những cá thể không mang điểm SNP nào và có những cá thể outlier lớn hơn hẳn so với khoảng phân bố trong quần thể, trong khi các cá thể ở ba dân tộc còn lại có khoảng phân bố hẹp hơn, đồng nghĩa với sự khác biệt về số lượng SNP tìm được ở các cá thể này sẽ bé hơn. Khi chia các cá thể người Việt trong bộ mẫu thành các nhóm ngữ hệ khác nhau, kết quả thể hiện rõ sự khác biệt rõ rệt giữa hai ngữ hệ Nam Á và Nam Đảo với ba ngữ hệ còn lại (Hình 3.21). So với khoảng phân bố số lượng SNP trên mỗi cá thể lớn ở NHNA và NHNĐ, các cá thể NHTK, NHHM và NHHT có số lượng SNP đồng đều hơn nhiều, trong đó số lượng SNP tìm được trên nhóm NHHM lớn hơn đáng kể ($p < 0,001$) so với hai ngữ hệ còn lại. Hiện tượng này có thể phản ánh sự đa dạng về vốn gen của các cá thể nam thuộc NHNA và NHNĐ, còn các cá thể nam thuộc ba ngữ hệ còn lại có mối quan hệ gần gũi với nhau hơn.



Hình 3.20. Biểu đồ số lượng SNP trên NST Y của năm dân tộc trong nghiên cứu

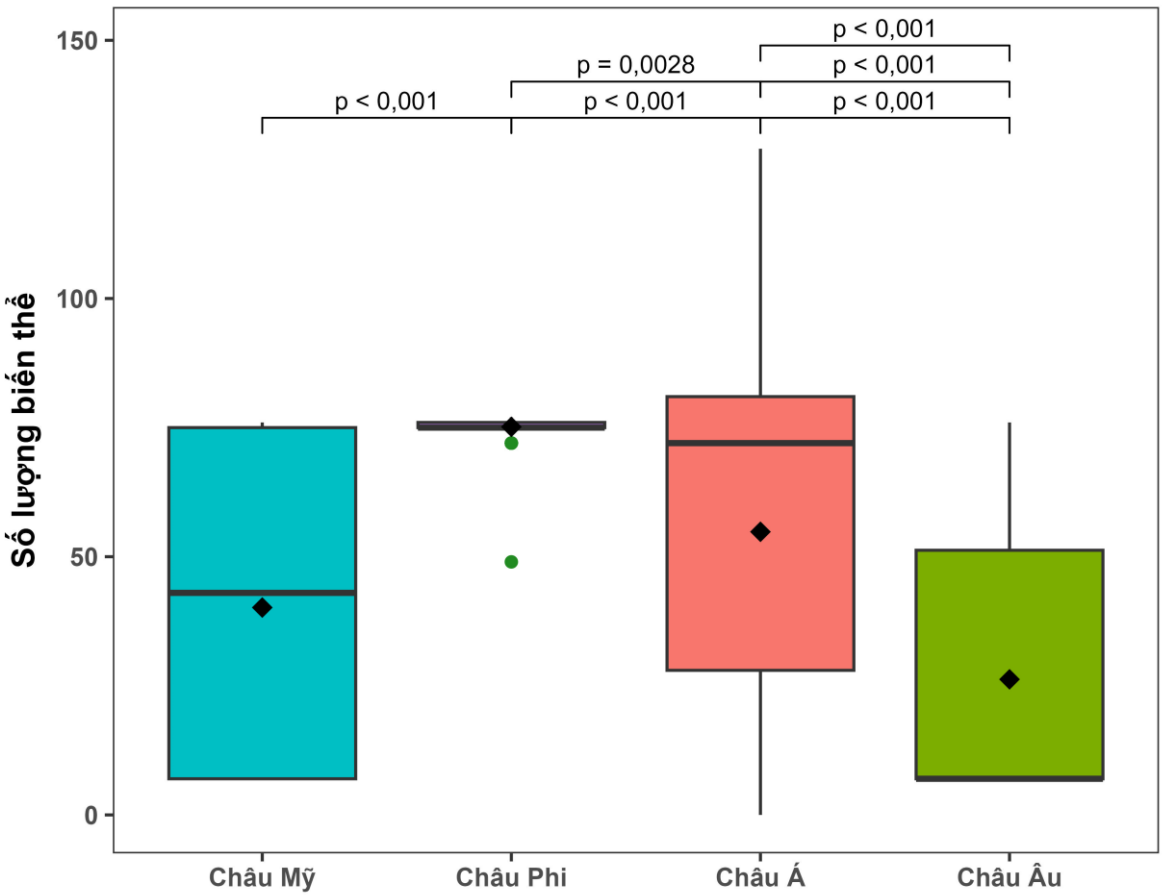
Điểm hình thoi thể hiện giá trị trung bình của số lượng biến thể tìm được trong một cá thể của các dân tộc tương ứng.



Hình 3.21. Biểu đồ số lượng SNP trên NST Y của năm ngữ hệ tại Việt Nam

Điểm hình thoi thể hiện giá trị trung bình của số lượng biến thể tìm được trong một cá thể của các ngữ hệ tương ứng.

Hiện tượng quan sát được trên quần thể người Việt cũng được quan sát giữa các khu vực khác nhau trên thế giới (Hình 3.22). Khoảng phân bố số lượng SNP rất hẹp ở quần thể người châu Phi phản ánh sự "thuần" của các cá thể ở khu vực này, cũng như số lượng hạn chế của các cuộc di cư ngược lại châu Phi. Sự khác biệt giữa ba châu lục còn lại cũng thể hiện sự đa dạng giữa các cá thể nam ở các châu lục khác nhau, trong đó người châu Âu có số lượng SNP ít hơn đáng kể so với người châu Âu, châu Phi và châu Á.



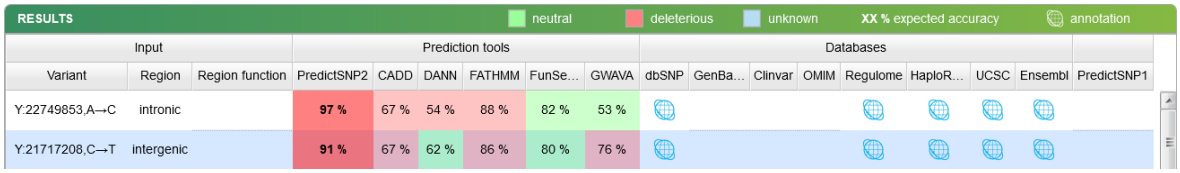
Hình 3.22. Biểu đồ số lượng SNP trên NST Y của bốn khu vực trên thế giới

Điểm hình thoi thể hiện giá trị trung bình của số lượng biến thể tìm được trong một cá thể của các châu lục tương ứng.

3.5.2. Kết quả phân tích tần suất xuất hiện của một số đa hình nucleotide đơn nổi bật

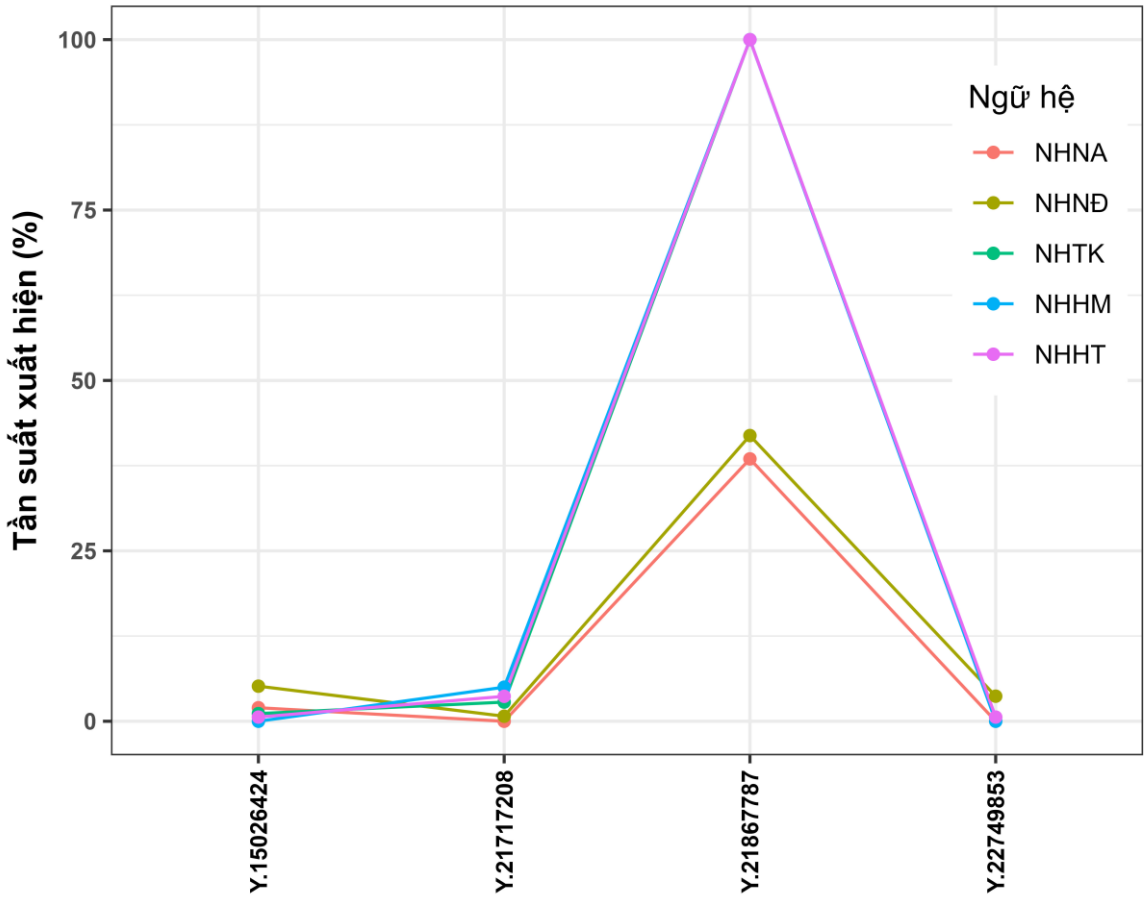
Đầu tiên, 189 SNP được kiểm tra khả năng gây bệnh bằng phần mềm online predictSNP2 [119], sau đó nhóm nghiên cứu sẽ sàng lọc thêm các SNP trên NST Y đã được nghiên cứu mối quan hệ với bệnh. Ở bước đầu, chúng tôi chọn được hai SNP (Hình 3.23) và chọn được thêm hai SNP ở bước thứ hai [120–122]. Ba trong bốn SNP (Y.15026424, Y.21867787, Y.22749853) nằm trong kiểu đơn bội gồm sáu SNP liên quan đến nguy cơ bị bệnh tự kỉ ở bé trai, trong đó Y.22749853 sự liên quan rõ rệt nhất [120]. Hơn nữa, các gen chứa những điểm này (Y.22749853-*EIF1AY*, Y.21867787-*KDM5D*, Y.15026424-*DDX3Y*) cũng có mối liên quan mạnh với các bệnh như bệnh vô tinh trùng vô căn (*KDM5D*, *DDX3Y*), mất một phần NST Y (*DDX3Y*), các bệnh hệ miễn dịch (*DDX3Y*, *EIF1AY*) và không có tinh trùng (*DDX3Y*, *KDM5D*, *EIF1AY*) [120].

Ngoài ra, mặc dù điểm Y.21717208 được dự đoán có khả năng gây bệnh cao nhưng chưa có nghiên cứu nào được thực hiện với điểm biến đổi này.

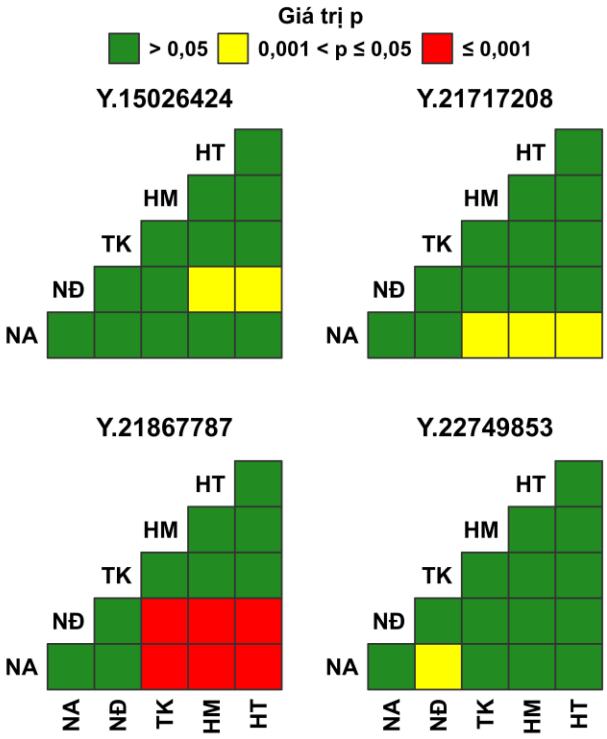


Hình 3.23. Kết quả sàng lọc khả năng gây bệnh của 189 SNP qua predictSNP2

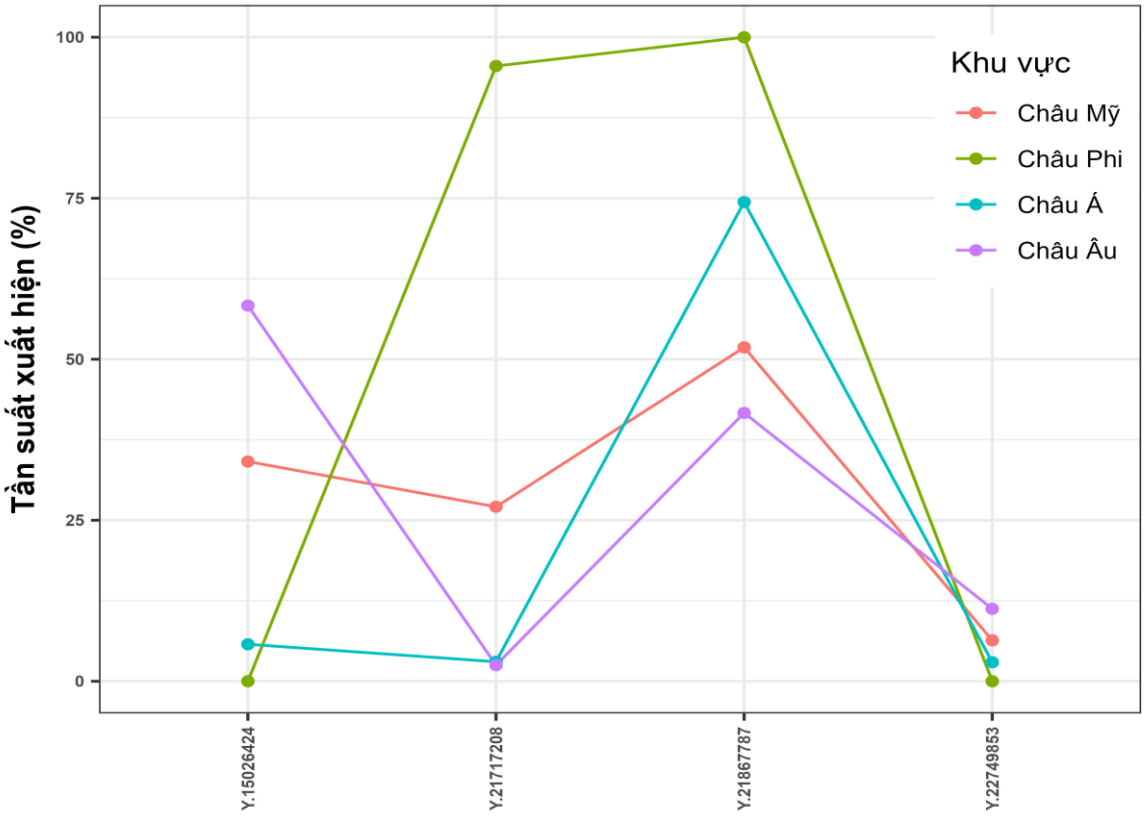
Đối với các ngữ hệ của Việt Nam, xu hướng của kết quả trông thấy ở hình 3.24 có thể chia thành hai nhóm: nhóm gồm NHHT, NHTK và NHHM, và nhóm gồm NHNĐ và NHHA. Sự khác biệt rõ rệt nhất giữa hai nhóm là tại điểm Y.21867787 với tần suất lên đến 100% ở NHTK, NHHM và NHHT (Hình 3.25). Xu hướng gộp nhóm này cũng nhất quán với kết quả so sánh số lượng biến thể. Khi so sánh giữa các khu vực khác nhau trên thế giới, kết quả cho thấy sự khác biệt đáng kể ở cả bốn SNP (Hình 3.26, hình 3.27). Cụ thể hơn, ở người châu Phi, hai điểm không tìm thấy trên nhóm người này lại xuất hiện với tần suất cao ở người châu Mỹ và châu Âu (Y.15026424) và một điểm xuất hiện ở tần suất trung bình ở các nhóm người còn lại (Y.22749853), còn một trong hai điểm xuất hiện ở tần suất cao ở người châu Phi (Y.21717208) lại gần như không được tìm thấy ở người châu Á và châu Âu và tần suất trung bình ở người châu Mỹ. Điểm Y.21867787 được tìm thấy ở tần suất cao ở tất cả nhóm người trong bộ mẫu, trong đó nhóm người châu Phi có tần suất xuất hiện ~100%.



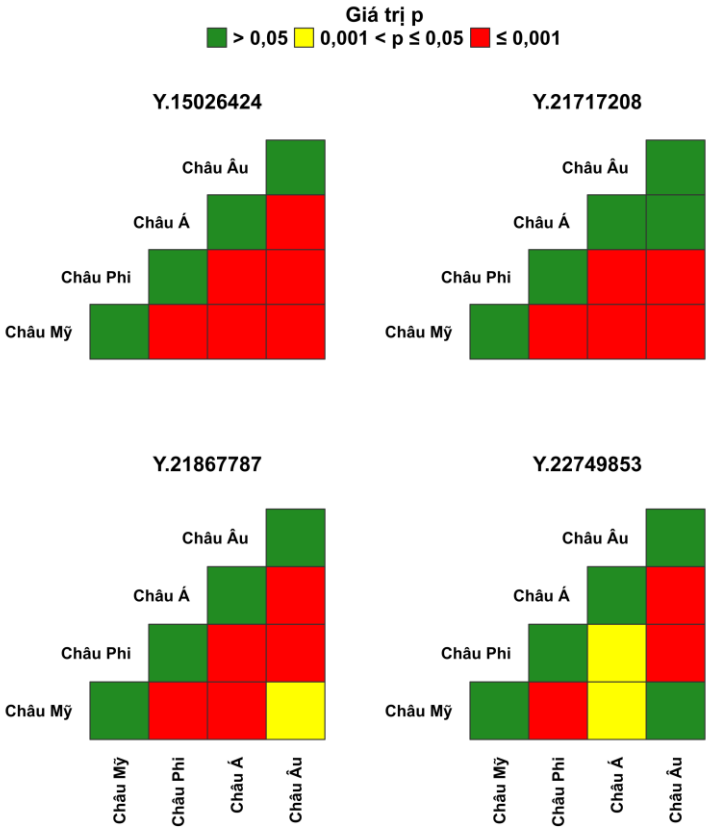
Hình 3.24. Tần suất xuất hiện của bốn SNP trên NST Y ở năm ngữ hệ tại Việt Nam



Hình 3.25. Giá trị p khi so sánh tần suất xuất hiện của bốn SNP trên NST Y giữa các cặp ngữ hệ



Hình 3.26. Tần suất xuất hiện của bốn SNP trên NST Y ở bốn khu vực trên thế giới



Hình 3.27. Giá trị p khi so sánh tần suất xuất hiện của bốn SNP trên NST Y giữa các cặp khu vực

KẾT LUẬN VÀ KIẾN NGHỊ

KẾT LUẬN

Thông qua các kết quả đã trình bày ở trên, luận văn "Phân tích toàn bộ hệ gen ty thể và đa hình nucleotide đơn vùng không trao đổi chéo của nhiễm sắc thể Y người Việt Nam thuộc năm dân tộc Pa Kô, Cơ-Tu, Rơ-Măm, Kinh miền Trung và Kinh Miền Nam" đã thu được các kết luận như sau:

1. Thu thập, tách chiết ADN tổng số và giải trình tự thành công hệ gen ty thể người Việt Nam bao gồm 22 cá thể người Pa Kô, 19 cá thể người Cơ-tu, 24 cá thể người Rơ-măm, 31 cá thể người Kinh miền Trung và 33 cá thể người Kinh miền Nam;

2. Xác định được 587 biến thể ở mtDNA của 129 cá thể thuộc 5 dân tộc nêu trên, trong đó vùng có nhiều và ít biến thể nhất lần lượt là HVSI và HVSIII. Xác định được sự khác biệt về số lượng điểm biến đổi trong vùng mã hóa hệ gen ty thể giữa người Rơ-măm với người Pa Kô và Cơ-tu, và giữa năm nhóm ngữ hệ tại Việt Nam;

3. Xác định được khoảng cách di truyền thấp giữa dân tộc Kinh với các dân tộc khác, sự khác biệt giữa dân tộc Pa Kô và Cơ-tu với các dân tộc còn lại, và đặc điểm di truyền đặc trưng của dân tộc Rơ-măm với các dân tộc chung ngữ hệ ở cùng khu vực. Sau khi tích hợp thêm trình tự vùng điều khiển của 85 dân tộc khác ở châu Á, xác định được khoảng cách di truyền thấp giữa dân tộc Kinh với một số dân tộc ở các quốc gia lân cận và giữa dân tộc Pa Kô và Cơ-tu với dân tộc Soa và Bru ở Thái Lan;

4. Xác định được 189 SNP trên 82 cá thể người Việt thuộc 5 dân tộc nêu trên. Xác định được sự khác biệt về số lượng và sự phân bố SNP trong NST Y giữa hai quần thể người Kinh với ba dân tộc còn lại, giữa năm nhóm ngữ hệ tại Việt Nam và giữa bốn nhóm châu lục trên thế giới. Xác định được sự khác biệt về tần suất xuất hiện của ba SNP liên quan tới bệnh **(**Y.15026424, Y.21867787, Y.22749853**)** và một điểm được dự đoán có khả năng gây bệnh cao **(**Y.21717208**)** trên NST Y giữa năm nhóm ngữ hệ tại Việt Nam và giữa bốn nhóm châu lục trên thế giới.

KIẾN NGHỊ

1. Tiếp tục nghiên cứu đa dạng di truyền sử dụng hai chỉ thị giới tính trên các dân tộc thuộc ngữ hệ Nam Á tại Việt Nam khác để quan sát rõ hơn sự di chuyển của dòng gen và sự đa dạng di truyền của ngữ hệ này, cũng như các sự kiện trao đổi thông tin di truyền với các quần thể nước ngoài;

2. Sử dụng chỉ thị khác để nghiên cứu đa dạng di truyền ở năm dân tộc trong nghiên cứu hiện tại như nghiên cứu toàn bộ hệ gen.

DANH MỤC CÔNG TRÌNH CÔNG BỐ CỦA TÁC GIẢ

Lã Đức Duy, Nông Văn Hải, Nguyễn Thùy Dương. Đa dạng di truyền vùng D-loop của hai dân tộc Cơ-Tu và Rơ-Măm. Tạp chí Khoa học và Công nghệ Việt Nam - B.*

*: Đã được chấp nhận đăng

DANH MỤC TÀI LIỆU THAM KHẢO

1. Taanman J.W., 1999, The mitochondrial genome: structure, transcription, translation and replication, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1410(2), pp. 103–123.
2. Brown W.M., George M., Wilson A.C., 1979, Rapid evolution of animal mitochondrial DNA, *Proceedings of the National Academy of Sciences of the United States of America*, 76(4), pp. 1967–1971.
3. Stoneking M., **2017**, *Sampling DNA Regions*, John Wiley & Sons, Ltd, pp.111–124.
4. Chial. H., Craig J., 2008, mtDNA and mitochondrial diseases, *Nature Education*, 1(1), pp. 217.
5. Hutchison C.A., Newbold J.E., Potter S.S., Edgell M.H., 1974, Maternal inheritance of mammalian mitochondrial DNA, *Nature*, 251(5475), pp. 536–538.
6. E. Giles R., Hugues B., M.Cann H., Wallace D.C., 1980, Maternal inheritance of human mitochondrial DNA, *Proceedings of the National Academy of Sciences of the United States of America*, 77(11), pp. 6715–6719.
7. Kondo R., Satta Y., Matsuura E.T., Ishiwa H., Takahata N., Chigusa S.I., 1990, Incomplete maternal transmission of mitochondrial DNA in Drosophila., *Genetics*, 126(3), pp. 657–663.
8. Gyllensten U., Wharton D., Josefsson A., Wilson A.C., 1991, Paternal inheritance of mitochondrial DNA in mice, *Nature 1991 352:6332*, 352(6332), pp. 255–257.
9. Kaneda H., Hayashi J.I., Takahama S., Taya C., Lindahl K.F., Yonekawa H., 1995, Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis, *Proceedings of the National Academy of Sciences of the United States of America*, 92(10), pp. 4542–4546.
10. Schwartz M., Vissing J., 2002, Paternal inheritance of mitochondrial DNA, *The New England Journal of Medicine*, 347(8), pp. 576–580.
11. Kauppi L., Barchi M., Baudat F., Romanienko P.J., Keeney S., Jasin M., 2011, Distinct properties of the XY pseudoautosomal region crucial for male meiosis, *Science (New York, N.Y.)*, 331(6019), pp. 916.
12. Page D.C., Mosher R., Simpson E.M., Fisher E.M.C., Mardon G., Pollack J., McGillivray B., de la Chapelle A., Brown L.G., 1987, The sex-determining region of the human Y chromosome encodes a finger protein, *Cell*, 51(6), pp. 1091–1104.
13. Hassold T.J., Sherman S.L., Pettay D., Page D.C., Jacobs P.A., 1991, XY

chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region., *American Journal of Human Genetics*, 49(2), pp. 253.

14. Shi Q., Martin R.H., 2001, Aneuploidy in human spermatozoa: FISH analysis in men with constitutional chromosomal abnormalities, and in infertile men, *Reproduction (Cambridge, England)*, 121(5), pp. 655–666.
15. Manz E., Alkan M., Bühler E., Schmidtke J., 1992, Arrangement of DYZ1 and DYZ2 repeats on the human Y-chromosome: a case with presence of DYZ1 and absence of DYZ2, *Molecular and Cellular Probes*, 6(3), pp. 257–259.
16. Cotter P.D., Norton M.E., 2005, Y chromosome heterochromatin variation detected at prenatal diagnosis, *Prenatal Diagnosis*, 25(11), pp. 1062–1063.
17. Colaco S., Modi D., 2018, Genetics of the human Y chromosome and its association with male infertility, *Reproductive Biology and Endocrinology : RB&E*, 16(1),.
18. Dorit R.L., Akashi H., Gilbert W., 1995, Absence of polymorphism at the ZFY locus on the human Y chromosome, *Science (New York, N.Y.)*, 268(5214), pp. 1183–1185.
19. Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonné-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L., Oefner P.J., 2000, Y chromosome sequence variation and the history of human populations, *Nature Genetics*, 26(3), pp. 358–361.
20. Lippold S., Xu H., Ko A., Li M., Renaud G., Butthof A., Schröder R., Stoneking M., 2014, Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences, *Investigative Genetics*, 5(1), pp. 13.
21. Westaway K.E., Louys J., Awe R.D., Morwood M.J., Price G.J., Zhao J.X., Aubert M., Joannes-Boyau R., Smith T.M., Skinner M.M., Compton T., Bailey R.M., Van Den Bergh G.D., De Vos J., Pike A.W.G., Stringer C., Saptomo E.W., Rizal Y., Zaim J., Santoso W.D., Trihascaryo A., Kinsley L., Sulistyanto B., 2017, An early modern human presence in Sumatra 73,000-63,000 years ago, *Nature*, 548(7667), pp. 322–325.
22. Demeter F., Shackelford L., Westaway K., Barnes L., Duringer P., Ponche J.L., Dumoncel J., Sénégas F., Sayavongkhamdy T., Zhao J.X., Sichanthongtip P., Patole-Edoumba E., Dunn T., Zachwieja A., Coppens Y., Willerslev E., Bacon A.M., 2017, Early modern humans from tam pà ling, laos fossil review and perspectives, *Current Anthropology*, 58, pp. S527–S538.

23. Yang M.A., 2022, A genetic history of migration, diversification, and admixture in Asia, *Http://Pivotscipub.Com*, 2(1), pp. 1–32.

24. Stoneking M., Arias L., Liu D., Oliveira S., Pugach I., Rodriguez J.J.R.B., 2023, The Past 12,000 Years of Behavior, Adaptation, and Evolution Shaped Who We Are Today: Genomic perspectives on human dispersals during the Holocene, *Proceedings of the National Academy of Sciences of the United States of America*, 120(4), pp. 2209475119.

25. Bellwood P., 2018, The search for ancient DNA heads east, *Science (New York, N.Y.)*, 361(6397), pp. 31–32.

26. Sidwell P., Jenny M., **2021**, *The neolithic occupation of Southeast Asia*, De Gruyter, pp.21–32.

27. Lipson M., Cheronet O., Mallick S., Rohland N., Oxenham M., Pietrusewsky M., Pryce T.O., Willis A., Matsumura H., Buckley H., Domett K., Nguyen G.H., Trinh H.H., Kyaw A.A., Win T.T., Pradier B., Broomandkhoshbacht N., Candilio F., Changmai P., Fernandes D., Ferry M., Gamarra B., Harney E., Kampuansai J., Kutanan W., Michel M., Novak M., Oppenheimer J., Sirak K., Stewardson K., Zhang Z., Flegontov P., Pinhasi R., Reich D., 2018, Ancient genomes document multiple waves of migration in Southeast Asian prehistory, *Science (New York, N.Y.)*, 361(6397), pp. 92.

28. McColl H., Racimo F., Vinner L., Demeter F., Gakuhari T., Víctor Moreno-Mayar J., Van Driem G., Wilken U.G., Seguin-Orlando A., De la Fuente Castro C., Wasef S., Shoocongdej R., Souksavatdy V., Sayavongkhamdy T., Saidin M.M., Allentoft M.E., Sato T., Malaspinas A.S., Aghakhanian F.A., Korneliussen T., Prohaska A., Margaryan A., De Barros Damgaard P., Kaewsutthi S., Lertrit P., Nguyen T.M.H., Hung H. chun, Tran T.M., Truong H.N., Nguyen G.H., Shahidan S., Wiradnyana K., Matsumae H., Shigehara N., Yoneda M., Ishida H., Masuyama T., Yamada Y., Tajima A., Shibata H., Toyoda A., Hanihara T., Nakagome S., Deviese T., Bacon A.M., Duringer P., Ponche J.L., Shackelford L., Patole-Edoumba E., Nguyen A.T., Bellina-Pryce B., Galipaud J.C., Kinaston R., Buckley H., Pottier C., Rasmussen S., Higham T., Foley R.A., Lahr M.M., Orlando L., Sikora M., Phipps M.E., Oota H., Higham C., Lambert D.M., Willerslev E., 2018, The prehistoric peopling of Southeast Asia, *Science*, 361(6397), pp. 88–92.

29. Wang C.C., Yeh H.Y., Popov A.N., Zhang H.Q., Matsumura H., Sirak K., Cheronet O., Kovalev A., Rohland N., Kim A.M., Mallick S., Bernardos R., Tumen D., Zhao J., Liu Y.C., Liu J.Y., Mah M., Wang K., Zhang Z., Adamski N., Broomandkhoshbacht N., Callan K., Candilio F., Carlson K.S.D., Culleton B.J., Eccles L., Freilich S., Keating D., Lawson A.M., Mandl K., Michel M., Oppenheimer J., Özdoğan K.T., Stewardson K., Wen S., Yan S., Zalzala F., Chuang R., Huang C.J., Looh H., Shiung

C.C., Nikitin Y.G., Tabarev A. V., Tishkin A.A., Lin S., Sun Z.Y., Wu X.M., Yang T.L., Hu X., Chen L., Du H., Bayarsaikhan J., Mijiddorj E., Erdenebaatar D., Iderkhangai T.O., Myagmar E., Kanzawa-Kiriyama H., Nishino M., Shinoda K. ichi, Shubina O.A., Guo J., Cai W., Deng Q., Kang L., Li D., Li D., Lin R., Nini, Shrestha R., Wang L.X., Wei L., Xie G., Yao H., Zhang M., He G., Yang X., Hu R., Robbeets M., Schiffels S., Kennett D.J., Jin L., Li H., Krause J., Pinhasi R., Reich D., 2021, Genomic insights into the formation of human populations in East Asia, *Nature*, 591(7850), pp. 413–419.

30. Eberhard, David M., Gary F. Simons and C.D.F., ed., **2023**, *Ethnologue: Languages of the World*, Twenty-six, SIL International.

31. Sidwell P., Jenny M., 2021, The languages and linguistics of Mainland Southeast Asia: A comprehensive guide, *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, pp. 1–968.

32. Dang N. Van, Chu T.S., Luu H., **2014**, *Ethnic Minorities in Vietnam*, The gioi.

33. Kutanan W., Kampuansai J., Brunelli A., Ghirotto S., Pittayaporn P., Ruangchai S., Schröder R., MacHoldt E., Srikummool M., Kangwanpong D., Hübner A., Arias L., Stoneking M., 2018, New insights from Thailand into the maternal genetic history of Mainland Southeast Asia, *European Journal of Human Genetics*, 26(6), pp. 898.

34. Kutanan W., Kampuansai J., Srikummool M., Kangwanpong D., Ghirotto S., Brunelli A., Stoneking M., 2017, Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages, *Human Genetics*, 136(1), pp. 85–98.

35. Kutanan W., Shoocongdej R., Srikummool M., Hübner A., Suttipai T., Srithawong S., Kampuansai J., Stoneking M., 2020, Cultural variation impacts paternal and maternal genetic lineages of the Hmong-Mien and Sino-Tibetan groups from Thailand, *European Journal of Human Genetics*, 28(11), pp. 1563–1579.

36. Woravatin W., Stoneking M., Srikummool M., Kampuansai J., Arias L., Kutanan W., 2023, South Asian maternal and paternal lineages in southern Thailand and the role of sex-biased admixture, *PLOS ONE*, 18(9), pp. e0291547.

37. Hill C., Soares P., Mormina M., Macaulay V., Meehan W., Blackburn J., Clarke D., Raja J.M., Ismail P., Bulbeck D., Oppenheimer S., Richards M., 2006, Phylogeography and ethnogenesis of aboriginal Southeast Asians, *Molecular Biology and Evolution*, 23(12), pp. 2480–2491.

38. Kloss-Brandstätter A., Summerer M., Horst D., Horst B., Streiter G., Raschenberger J., Kronenberg F., Sanguansermsri T., Horst J.,

Weissensteiner H., 2021, An in-depth analysis of the mitochondrial phylogenetic landscape of Cambodia, *Scientific Reports*, 11(1), pp. 10816.

39. Summerer M., Horst J., Erhart G., Weißensteiner H., Schönherr S., Pacher D., Forer L., Horst D., Manhart A., Horst B., Sanguansermsri T., Kloss-Brandstätter A., 2014, Large-scale mitochondrial DNA analysis in Southeast Asia reveals evolutionary effects of cultural isolation in the multi-ethnic population of Myanmar, *BMC Evolutionary Biology*, 14(1), pp. 1–12.
40. Chandrasekar A., Kumar S., Sreenath J., Sarkar B.N., Urade B.P., Mallick S., Bandopadhyay S.S., Barua P., Barik S.S., Basu D., Kiran U., Gangopadhyay P., Sahani R., Prasad B.V.R., Gangopadhyay S., Lakshmi G.R., Ravuri R.R., Padmaja K., Venugopal P.N., Sharma M.B., Rao V.R., 2009, Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor, *PloS One*, 4(10), pp. e7447.
41. Rajkumar R., Banerjee J., Gunturi H.B., Trivedi R., Kashyap V.K., 2005, Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages, *BMC Evolutionary Biology*, 5, pp. 26.
42. Chaubey G., Metspalu M., Choi Y., Mägi R., Romero I.G., Soares P., Van Oven M., Behar D.M., Rootsi S., Hudjashov G., Mallick C.B., Karmin M., Nelis M., Parik J., Reddy A.G., Metspalu E., Van Driem G., Xue Y., Tyler-Smith C., Thangaraj K., Singh L., Remm M., Richards M.B., Lahr M.M., Kayser M., Villems R., Kivisild T., 2011, Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture, *Molecular Biology and Evolution*, 28(2), pp. 1013.
43. Chaubey G., Metspalu M., Kivisild T., Villems R., 2007, Peopling of South Asia: investigating the caste-tribe continuum in India, *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 29(1), pp. 91–100.
44. Chaubey G., Karmin M., Metspalu E., Metspalu M., Selvi-Rani D., Singh V.K., Parik J., Solnik A., Naidu B.P., Kumar A., Adarsh N., Mallick C.B., Trivedi B., Prakash S., Reddy R., Shukla P., Bhagat S., Verma S., Vasnik S., Khan I., Barwa A., Sahoo D., Sharma A., Rashid M., Chandra V., Reddy A.G., Torroni A., Foley R.A., Thangaraj K., Singh L., Kivisild T., Villems R., 2008, Phylogeography of mtDNA haplogroup R7 in the Indian peninsula, *BMC Evolutionary Biology*, 8(1), pp. 227.
45. Ahlawat B., Dewangan H., Pasupuleti N., Dwivedi A., Rajpal R., Pandey S., Kumar L., Thangaraj K., Rai N., 2024, Investigating linguistic and

genetic shifts in East Indian tribal groups, *Heliyon*, 10(14), pp. e34354.

46. Jaisamut K., Pitiwararom R., Sukawutthiya P., Sathirapatya T., Noh H., Worrapitirungsi W., Vongpaisarnsin K., 2023, Unraveling the mitochondrial phylogenetic landscape of Thailand reveals complex admixture and demographic dynamics, *Scientific Reports*, 13(1), pp. 20396.

47. Hoh B.P., Deng L., Xu S., 2022, The Peopling and Migration History of the Natives in Peninsular Malaysia and Borneo: A Glimpse on the Studies Over the Past 100 years, *Frontiers in Genetics*, 13, pp. 767018.

48. Kutanan W., Kampuansai J., Srikummool M., Brunelli A., Ghirotto S., Arias L., Macholdt E., Hübner A., Schröder R., Stoneking M., 2019, Contrasting Paternal and Maternal Genetic Histories of Thai and Lao Populations, *Molecular Biology and Evolution*, 36(7), pp. 1490.

49. Macholdt E., Arias L., Duong N.T., Ton N.D., Van Phong N., Schröder R., Pakendorf B., Van Hai N., Stoneking M., 2020, The paternal and maternal genetic history of Vietnamese populations, *European Journal of Human Genetics*, 28(5), pp. 636–645.

50. Kumar V., Langstieh B.T., Madhavi K. V., Naidu V.M., Singh H.P., Biswas S., Thangaraj K., Singh L., Reddy B.M., 2006, Global Patterns in Human Mitochondrial DNA and Y-Chromosome Variation Caused by Spatial Instability of the Local Cultural Processes, *PLoS Genetics*, 2(4), pp. 420–424.

51. Arias L., Schröder R., Hübner A., Barreto G., Stoneking M., Pakendorf B., 2018, Cultural Innovations Influence Patterns of Genetic Diversity in Northwestern Amazonia, *Molecular Biology and Evolution*, 35(11), pp. 2719.

52. Ballinger S.W., Schurr T.G., Torroni A., Gan Y.Y., Hodge J.A., Hassan K., Chen K.H., Wallace D.C., 1992, Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations, *Genetics*, 130(1), pp. 139–152.

53. Ivanova R., Astrinidis A., Lepage V., Djoulah S., Wijnen E., Vu-Trieu A.N., Hors J., Charron D., 1999, Mitochondrial DNA polymorphism in the Vietnamese population, *European Journal of Immunogenetics : Official Journal of the British Society for Histocompatibility and Immunogenetics*, 26(6), pp. 417–422.

54. Irwin J.A., Saunier J.L., Strouss K.M., Diegoli T.M., Sturk K.A., O'Callaghan J.E., Paintner C.D., Hohoff C., Brinkmann B., Parsons T.J., 2008, Mitochondrial control region sequences from a Vietnamese population sample, *International Journal of Legal Medicine*, 122(3), pp. 257–259.

55. Oota H., Kitano T., Jin F., Yuasa I., Wang L., Ueda S., Saitou N., Stoneking M., 2002, Extreme mtDNA homogeneity in continental Asian populations, *American Journal of Physical Anthropology*, 118(2), pp. 146–153.

56. Huỳnh T.T.H., Hoàng T.T.Y., Nguyễn Đ.T., Lê T.T.H., Nguyễn Đ.C., Phan V.C., Nông V.H., 2005, Phân tích trình tự vùng điều khiển (D-loop) trên genome ty thể của 5 cá thể người Việt Nam, *Tạp Chí Công Nghệ Sinh Học*, 3(1), pp. 15–22.

57. Nguyễn Đ.T., Nguyễn T.T.L., Vũ H.C., Trần T.N.D., Địch T.K.H., Bùi T.T., Nguyễn H.H., Huỳnh T.T.H., Lê T.T.H., Trần T.P.L., Phan V.C., Nông V.H., 2008, Đa hình đơn bội DNA ty thể của các cá thể người Việt Nam, *Tạp Chí Công Nghệ Sinh Học*, 6(4), pp. 579–590.

58. Trần T.T.H., Trần H.T., Trần V.K., 2017, Đa hình thái đơn nucleotid vùng gen ty thể HV1 và HV2 trên người dân tộc Kinh và dân tộc Mường của Việt Nam, *Tạp Chí Nghiên Cứu Y Học*, 106(1), pp. 33–40.

59. Hung D.M., Ha N.H., Khoi P.N., Nhung V.P., Phong N. Van, Duong N.T., Hai N. Van, Ton N.D., 2016, Genetic variation of mitochondrial sequence-hv2 in Vietnamese populations, *Academia Journal of Biology*, 38(2), pp. 243–249.

60. Pischedda S., Barral-Arca R., Gómez-Carballa A., Pardo-Seco J., Catelli M.L., Álvarez-Iglesias V., Cárdenas J.M., Nguyen N.D., Ha H.H., Le A.T., Martinón-Torres F., Vullo C., Salas A., 2017, Phylogeographic and genome-wide investigations of Vietnam ethnic groups reveal signatures of complex historical demographic movements, *Scientific Reports*, 7(1), pp. 1–15.

61. Duong N.T., Macholdt E., Ton N.D., Arias L., Schröder R., Van Phong N., Thi Bich Thuy V., Ha N.H., Thi Thu Hue H., Thi Xuan N., Thi Phuong Oanh K., Hien L.T.T., Hoang N.H., Pakendorf B., Stoneking M., Van Hai N., 2018, Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia, *Scientific Reports*, 8(1), pp. 1–13.

62. Thy Ngọc N., Bảo Trang N., Quang Huy N., Đăng Tôn N., Thùy Dương N., 2018, Đa hình vùng D-loop hệ gen ty thể của các cá thể dân tộc Kinh và Mảng cùng trong nhóm ngữ hệ Nam Á, *Tạp Chí Công Nghệ Sinh Học*, 16(2), pp. 231–240.

63. Tran T.T.H., Nguyen D.H., Tran V.K., Nguyen Q.L., Trinh H.A., Luong L.H., Tran V.A., Pham L.A.T., Nguyen T.T., Nguyen V.B., Tran T.H., Van Ta T., 2020, Variation of Mitochondrial DNA HV1 AND HV2 of the Vietnamese Population, *Advances in Experimental Medicine and Biology*, 1292, pp. 37–63.

64. Thao D.H., Dinh T.H., Mitsunaga S., Duy L.D., Phuong N.T., Anh N.P.,

Anh N.T., Duc B.M., Hue H.T.T., Ha N.H., Ton N.D., Hübner A., Pakendorf B., Stoneking M., Inoue I., Duong N.T., Hai N. Van, 2024, Investigating demic versus cultural diffusion and sex bias in the spread of Austronesian languages in Vietnam, *PloS One*, 19(6), pp. e0304964.

65. Li H., Wen B., Chen S.J., Su B., Pramoonjago P., Liu Y., Pan S., Qin Z., Liu W., Cheng X., Yang N., Li X., Tran D., Lu D., Hsu M.T., Deka R., Marzuki S., Tan C.C., 2008, Paternal genetic affinity between western Austronesians and Daic populations, *BMC Evolutionary Biology*, 8(1), pp. 1–12.

66. Nguyễn Đ.T., Nguyễn T.D., Nông V.H., 2009, Sự phân bố các đa hình Nucleotide đơn của nhóm đớn bội C trên nhiễm sắc thể Y ở người Việt Nam, *Tạp Chí Công Nghệ Sinh Học*, 1, pp. 11–18.

67. Nguyễn Đ.T., Nguyễn T.D., Nông V.H., 2009, Sự phân bố các đa hình nucleotide đơn của nhóm đơn bội O trên nhiễm sắc thể Y ở người Việt Nam, *Tạp Chí Công Nghệ Sinh Học*, 7(3), pp. 285–294.

68. Ha H.H., Nguyen T.H., Tran L.H., Nguyen H.T.H., Hoang H., Chu H.H., 2019, Genetic characteristics of 23 Y-chromosomal STRs in the Kinh population in Northern Vietnam, *International Journal of Legal Medicine*, 133(5), pp. 1403–1404.

69. Stoneking M., **2017**, *Genetic Markers*, John Wiley & Sons, Ltd, pp.94–102.

70. Zhang X., Qi X., Yang Z., Serey B., Sovannary T., Bunnath L., Seang Aun H., Samnom H., Zhang H., Lin Q., Van Oven M., Shi H., Su B., 2013, Analysis of mitochondrial genome diversity identifies new and ancient maternal lineages in Cambodian aborigines, *Nature Communications 2013 4:1*, 4(1), pp. 1–11.

71. Li Y.C., Wang H.W., Tian J.Y., Liu L.N., Yang L.Q., Zhu C.L., Wu S.F., Kong Q.P., Zhang Y.P., 2015, Ancient inland human dispersals from Myanmar into interior East Asia since the Late Pleistocene, *Scientific Reports*, 5(9473),.

72. Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Bentley D.R., Chakravarti A., Clark A.G., Donnelly P., Eichler E.E., Flicek P., Gabriel S.B., Gibbs R.A., Green E.D., Hurles M.E., Knoppers B.M., Korbel J.O., Lander E.S., Lee C., Lehrach H., Mardis E.R., Marth G.T., McVean G.A., Nickerson D.A., Schmidt J.P., Sherry S.T., Wang J., Wilson R.K., Boerwinkle E., Doddapaneni H., Han Y., Korchina V., Kovar C., Lee S., Muzny D., Reid J.G., Zhu Y., Chang Y., Feng Q., Fang X., Guo X., Jian M., Jiang H., Jin X., Lan T., Li G., Li J., Li Y., Liu S., Liu X., Lu Y., Ma X., Tang M., Wang B., Wang G., Wu H., Wu R., Xu X., Yin Y., Zhang D., Zhang W., Zhao J., Zhao M., Zheng X., Gupta N., Gharani N., Toji L.H., Gerry N.P., Resch A.M., Barker J., Clarke L., Gil L., Hunt S.E.,

Kelman G., Kulesha E., Leinonen R., McLaren W.M., Radhakrishnan R., Roa A., Smirnov D., Smith R.E., Streeter I., Thormann A., Toneva I., Vaughan B., Zheng-Bradley X., Grocock R., Humphray S., James T., Kingsbury Z., Sudbrak R., Albrecht M.W., Amstislavskiy V.S., Borodina T.A., Lienhard M., Mertes F., Sultan M., Timmermann B., Yaspo M.L., Fulton L., Ananiev V., Belaia Z., Beloslyudtsev D., Bouk N., Chen C., Church D., Cohen R., Cook C., Garner J., Hefferon T., Kimelman M., Liu C., Lopez J., Meric P., O'Sullivan C., Ostapchuk Y., Phan L., Ponomarov S., Schneider V., Shekhtman E., Sirotkin K., Slotta D., Zhang H., Balasubramaniam S., Burton J., Danecek P., Keane T.M., Kolb-Kokocinski A., McCarthy S., Stalker J., Quail M., Davies C.J., Gollub J., Webster T., Wong B., Zhan Y., Campbell C.L., Kong Y., Marcketta A., Yu F., Antunes L., Bainbridge M., Sabo A., Huang Z., Coin L.J.M., Fang L., Li Q., Li Z., Lin H., Liu B., Luo R., Shao H., Xie Y., Ye C., Yu C., Zhang F., Zheng H., Zhu H., Alkan C., Dal E., Kahveci F., Garrison E.P., Kural D., Lee W.P., Leong W.F., Stromberg M., Ward A.N., Wu J., Zhang M., Daly M.J., DePristo M.A., Handsaker R.E., Banks E., Bhatia G., Del Angel G., Genovese G., Li H., Kashin S., McCarroll S.A., Nemesh J.C., Poplin R.E., Yoon S.C., Lihm J., Makarov V., Gottipati S., Keinan A., Rodriguez-Flores J.L., Rausch T., Fritz M.H., Stütz A.M., Beal K., Datta A., Herrero J., Ritchie G.R.S., Zerbino D., Sabeti P.C., Shlyakhter I., Schaffner S.F., Vitti J., Cooper D.N., Ball E. V., Stenson P.D., Barnes B., Bauer M., Cheetham R.K., Cox A., Eberle M., Kahn S., Murray L., Peden J., Shaw R., Kenny E.E., Batzer M.A., Konkel M.K., Walker J.A., MacArthur D.G., Lek M., Herwig R., Ding L., Koboldt D.C., Larson D., Ye K., Gravel S., Swaroop A., Chew E., Lappalainen T., Erlich Y., Gymrek M., Willems T.F., Simpson J.T., Shriver M.D., Rosenfeld J.A., Bustamante C.D., Montgomery S.B., De La Vega F.M., Byrnes J.K., Carroll A.W., DeGorter M.K., Lacroute P., Maples B.K., Martin A.R., Moreno-Estrada A., Shringarpure S.S., Zakharia F., Halperin E., Baran Y., Cerveira E., Hwang J., Malhotra A., Plewczynski D., Radew K., Romanovitch M., Zhang C., Hyland F.C.L., Craig D.W., Christoforides A., Homer N., Izatt T., Kurdoglu A.A., Sinari S.A., Squire K., Xiao C., Sebat J., Antaki D., Gujral M., Noor A., Ye K., Burchard E.G., Hernandez R.D., Gignoux C.R., Haussler D., Katzman S.J., Kent W.J., Howie B., Ruiz-Linares A., Dermitzakis E.T., Devine S.E., Kang H.M., Kidd J.M., Blackwell T., Caron S., Chen W., Emery S., Fritsche L., Fuchsberger C., Jun G., Li B., Lyons R., Scheller C., Sidore C., Song S., Sliwerska E., Taliun D., Tan A., Welch R., Wing M.K., Zhan X., Awadalla P., Hodgkinson A., Li Y., Shi X., Quitadamo A., Lunter G., Marchini J.L., Myers S., Churchhouse C., Delaneau O., Gupta-Hinch A., Kretzschmar W., Iqbal Z., Mathieson I., Menelaou A., Rimmer A., Xifara D.K., Oleksyk T.K., Fu Y., Liu X., Xiong M., Jorde L., Witherspoon D., Xing J., Browning B.L., Browning S.R., Hormozdiari F., Sudmant P.H.,

Khurana E., Tyler-Smith C., Albers C.A., Ayub Q., Chen Y., Colonna V., Jostins L., Walter K., Xue Y., Gerstein M.B., Abyzov A., Balasubramanian S., Chen J., Clarke D., Fu Y., Harmanci A.O., Jin M., Lee D., Liu J., Mu X.J., Zhang J., Zhang Y., Hartl C., Shakir K., Degenhardt J., Meiers S., Raeder B., Casale F.P., Stegle O., Lameijer E.W., Hall I., Bafna V., Michaelson J., Gardner E.J., Mills R.E., Dayama G., Chen K., Fan X., Chong Z., Chen T., Chaisson M.J., Huddleston J., Malig M., Nelson B.J., Parrish N.F., Blackburne B., Lindsay S.J., Ning Z., Zhang Y., Lam H., Sisu C., Challis D., Evani U.S., Lu J., Nagaswamy U., Yu J., Li W., Habegger L., Yu H., Cunningham F., Dunham I., Lage K., Jespersen J.B., Horn H., Kim D., Desalle R., Narechania A., Sayres M.A.W., Mendez F.L., Poznik G.D., Underhill P.A., Mittelman D., Banerjee R., Cerezo M., Fitzgerald T.W., Louzada S., Massaia A., Yang F., Kalra D., Hale W., Dan X., Barnes K.C., Beiswanger C., Cai H., Cao H., Henn B., Jones D., Kaye J.S., Kent A., Kerasidou A., Mathias R., Ossorio P.N., Parker M., Rotimi C.N., Royal C.D., Sandoval K., Su Y., Tian Z., Tishkoff S., Via M., Wang Y., Yang H., Yang L., Zhu J., Bodmer W., Bedoya G., Cai Z., Gao Y., Chu J., Peltonen L., Garcia-Montero A., Orfao A., Dutil J., Martinez-Cruzado J.C., Mathias R.A., Hennis A., Watson H., McKenzie C., Qadri F., LaRocque R., Deng X., Asogun D., Folarin O., Happi C., Omoniwa O., Stremlau M., Tariyal R., Jallow M., Joof F.S., Corrah T., Rockett K., Kwiatkowski D., Kooner J., Hien T.T., Dunstan S.J., ThuyHang N., Fonnie R., Garry R., Kanneh L., Moses L., Schieffelin J., Grant D.S., Gallo C., Poletti G., Saleheen D., Rasheed A., Brooks L.D., Felsenfeld A.L., McEwen J.E., Vaydylevich Y., Duncanson A., Dunn M., Schloss J.A., 2015, A global reference for human genetic variation, *Nature*, 526(7571), pp. 68–74.

73. Jinam T.A., Hong L.C., Phipps M.E., Stoneking M., Ameen M., Edo J., Saitou N., 2012, Evolutionary history of continental southeast Asians: “early train” hypothesis based on genetic analysis of mitochondrial and autosomal DNA data, *Molecular Biology and Evolution*, 29(11), pp. 3513–3527.

74. Delfin F., Min-Shan Ko A., Li M., Gunnarsdóttir E.D., Tabbada K.A., Salvador J.M., Calacal G.C., Sagum M.S., Datar F.A., Padilla S.G., De Ungria M.C.A., Stoneking M., 2014, Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region, *European Journal of Human Genetics : EJHG*, 22(2), pp. 228–237.

75. Gunnarsdo E.D., Li M., Bauchet M., Finstermeier K., Stoneking M., 2011, High-throughput sequencing of complete human mtDNA genomes from the Philippines, *Genome Research*, 21(1), pp. 1–11.

76. Gunnarsdóttir E.D., Nandineni M.R., Li M., Myles S., Gil D., Pakendorf

B., Stoneking M., 2011, Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra, *Nature Communications*, 2(1), pp. 1–7.

77. Ko A.M.S., Chen C.Y., Fu Q., Delfin F., Li M., Chiu H.L., Stoneking M., Ko Y.C., 2014, Early Austronesians: Into and Out Of Taiwan, *American Journal of Human Genetics*, 94(3), pp. 426–36.
78. Maricic T., Whitten M., Pääbo S., 2010, Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products, *PLoS ONE*, 5(11), pp. e14004.
79. Simon A., Accessed,2010.
80. Bolger A.M., Lohse M., Usadel B., 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, 30(15), pp. 2114.
81. Li H., 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
82. Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., 2021, Twelve years of SAMtools and BCFtools, *GigaScience*, 10(2), pp. 1–4.
83. Van der Auwera G., O'Connor B., Safari an O.M.C., **2020**, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st editio, O'Reilly Media, Inc.
84. Katoh K., Standley D.M., 2013, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, *Molecular Biology and Evolution*, 30(4), pp. 772.
85. Excoffier L., Lischer H.E.L., 2010, Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 10: 564-567., *Evolutionary Bioinformatics Online*.
86. Paradis E., Barrett J., 2010, pegas: an R package for population genetics with an integrated–modular approach, *Bioinformatics*, 26(3), pp. 419–420.
87. Paradis E., Schliep K., 2019, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, *Bioinformatics*, 35(3), pp. 526–528.
88. Venables W., Ripley B., **2002**, *Modern Applied Statistics with S*, Fourth, Springer.
89. Lott M.T., Leipzig J.N., Derbeneva O., Michael Xie H., Chalkia D., Sarmady M., Procaccio V., Wallace D.C., 2013, mtDNA Variation and Analysis Using Mitomap and Mitomaster, *Current Protocols in Bioinformatics*, 44(123), pp. 1.23.1-26.

90. Yu X., Li S., Guo Q., Leng J., Ding Y., 2024, The Association Between Mitochondrial tRNAGlu Variants and Hearing Loss: A Case-Control Study, *Pharmacogenomics and Personalized Medicine*, 17, pp. 77–89.

91. Dashti M., Alsaleh H., Eaaswarkhanth M., John S.E., Nizam R., Melhem M., Hebbar P., Sharma P., Al-Mulla F., Thanaraj T.A., 2021, Delineation of Mitochondrial DNA Variants From Exome Sequencing Data and Association of Haplogroups With Obesity in Kuwait, *Frontiers in Genetics*, 12, pp. 626260.

92. Al Asoom L., Khan J., Al Sunni A., Rafique N., Latif R., Alabdali M., Abdulazeez S., Borgio J.F., 2022, A Pilot Mitochondrial Genome-Wide Association on Migraine Among Saudi Arabians, *International Journal of General Medicine*, 15, pp. 6249.

93. Xing S., Jiang S., Wang S., Lin P., Sun H., Peng H., Yang J., Kong H., Wang S., Bai Q., Qiu R., Dai W., Yuan J., Ma Y., Yu X., Yao Y., Su J., 2023, Association of mitochondrial DNA variation with high myopia in a Han Chinese population, *Molecular Genetics and Genomics*, 298(5), pp. 1059.

94. Yang T.L., Guo Y., Shen H., Lei S.F., Liu Y.J., Li J., Liu Y.Z., Yu N., Chen J., Xu T., Cheng Y., Tian Q., Yu P., Papasian C.J., Deng H.W., 2011, Genetic Association Study of Common Mitochondrial Variants on Body Fat Mass, *PLoS ONE*, 6(6), pp. 21595.

95. Knoll N., Jarick I., Volckmar A.L., Klingenspor M., Illig T., Grallert H., Gieger C., Wichmann H.E., Peters A., Wiegand S., Biebermann H., Fischer-Posovszky P., Wabitsch M., Völzke H., Nauck M., Teumer A., Rosskopf D., Rimmbach C., Schreiber S., Jacobs G., Lieb W., Franke A., Hebebrand J., Hinney A., 2014, Mitochondrial DNA Variants in Obesity, *PLoS ONE*, 9(5), pp. e94882.

96. Nardelli C., Labruna G., Liguori R., Mazzaccara C., Ferrigno M., Capobianco V., Pezzuti M., Castaldo G., Farinaro E., Contaldo F., Buono P., Sacchetti L., Pasanisi F., 2013, Haplogroup T Is an Obesity Risk Factor: Mitochondrial DNA Haplotyping in a Morbid Obese Population from Southern Italy, *BioMed Research International*, 2013, pp. 631082.

97. Ebner S., Mangge H., Langhof H., Halle M., Siegrist M., Aigner E., Paulmichl K., Paulweber B., Datz C., Sperl W., Kofler B., Weghuber D., 2015, Mitochondrial Haplogroup T Is Associated with Obesity in Austrian Juveniles and Adults, *PLoS ONE*, 10(8), pp. e0135622.

98. Flaquer A., Baumbach C., Kriebel J., Meitinger T., Peters A., Waldenberger M., Grallert H., Strauch K., 2014, Mitochondrial Genetic Variants Identified to Be Associated with BMI in Adults, *PLoS ONE*, 9(8), pp. e105116.

99. Hwang I.W., Kim K., Choi E.J., Jin H.J., 2019, Association of

mitochondrial haplogroup F with physical performance in korean population, *Genomics & Informatics*, 17(1), pp. e11.

100. Baird P.N., Saw S.M., Lanca C., Guggenheim J.A., Smith E.L., Zhou X., Matsui K.O., Wu P.C., Sankaridurg P., Chia A., Rosman M., Lamoureux E.L., Man R., He M., 2020, Myopia, *Nature Reviews. Disease Primers*, 6(1), pp. 99.
101. Saw S.M., Gazzard G., Shin-Yen E.C., Chua W.H., 2005, Myopia and associated pathological complications, *Ophthalmic & Physiological Optics : The Journal of the British College of Ophthalmic Opticians (Optometrists)*, 25(5), pp. 381–391.
102. Francisco B.M., Salvador M., Amparo N., 2015, Oxidative Stress in Myopia, *Oxidative Medicine and Cellular Longevity*, 2015, pp. 750637.
103. Yang J., Ouyang X., Fu H., Hou X., Liu Y., Xie Y., Yu H., Wang G., 2022, Advances in biomedical study of the myopia-related signaling pathways and mechanisms, *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, 145, pp. 112472.
104. Jafarlou F., Najafi B., Sameni S.J., 2021, Is Newborn Hearing Screening Cost Effective? Economic Consideration for Policy Makers, *International Journal of Preventive Medicine*, 12(1), pp. 155.
105. Ding Y., Leng J., Fan F., Xia B., Xu P., 2013, The role of mitochondrial DNA mutations in hearing loss, *Biochemical Genetics*, 51(7–8), pp. 588–602.
106. Moassass F., Al-Halabi B., Nweder M.S., Al-Achkar W., 2018, Investigation of the mtDNA mutations in Syrian families with non-syndromic sensorineural hearing loss, *International Journal of Pediatric Otorhinolaryngology*, 113, pp. 110–114.
107. Ibrahim I., Dominguez-Valentin M., Segal B., Zeitouni A., da Silva S.D., 2018, Mitochondrial mutations associated with hearing and balance disorders, *Mutation Research*, 810, pp. 39–44.
108. Mutai H., Watabe T., Kosaki K., Ogawa K., Matsunaga T., 2017, Mitochondrial mutations in maternally inherited hearing loss, *BMC Medical Genetics*, 18(1), pp. 32.
109. Yorns W.R., Hardison H.H., 2013, Mitochondrial dysfunction in migraine, *Seminars in Pediatric Neurology*, 20(3), pp. 188–193.
110. Stuart S., Griffiths L.R., 2012, A possible role for mitochondrial dysfunction in migraine, *Molecular Genetics and Genomics : MGG*, 287(11–12), pp. 837–844.
111. Liu C., Fetterman J.L., Liu P., Luo Y., Larson M.G., Vasan R.S., Zhu J., Levy D., 2018, Deep Sequencing of the Mitochondrial Genome Reveals

Common Heteroplasmic Sites in NADH Dehydrogenase Genes, *Human Genetics*, 137(3), pp. 203.

112. Meunier B., Fisher N., Ransac S., Mazat J.P., Brasseur G., 2013, Respiratory complex III dysfunction in humans and the use of yeast as a model organism to study mitochondrial myopathy and associated diseases, *Biochimica et Biophysica Acta*, 1827(11–12), pp. 1346–1361.

113. Nesti C., Meschini M.C., Meunier B., Sacchini M., Doccini S., Romano A., Petrillo S., Pezzini I., Seddiki N., Rubegni A., Piemonte F., Alice Donati M., Brasseur G., Santorelli F.M., 2015, Additive effect of nuclear and mitochondrial mutations in a patient with mitochondrial encephalomyopathy, *Human Molecular Genetics*, 24(11), pp. 3248–3256.

114. Rezvani Z., Didari E., Arastehkani A., Ghodsinejad V., Aryani O., Kamalidehghan B., Houshmand M., 2013, Fifteen novel mutations in the mitochondrial NADH dehydrogenase subunit 1, 2, 3, 4, 4L, 5 and 6 genes from Iranian patients with Leber's hereditary optic neuropathy (LHON), *Molecular Biology Reports*, 40(12), pp. 6837–6841.

115. Ronchi D., Cosi A., Tonduti D., Orcesi S., Bordoni A., Fortunato F., Rizzuti M., Sciacco M., Collotta M., Cagdas S., Capovilla G., Moggio M., Berardinelli A., Veggiotti P., Comi G.P., 2011, Clinical and molecular features of an infant patient affected by Leigh Disease associated to m.14459G > A mitochondrial DNA mutation: a case report, *BMC Neurology*, 11, pp. 85.

116. Stoneking M., **2017**, *Analysis of Genetic Data from Populations*, John Wiley & Sons, Ltd, pp.139–145.

117. Wen B., Hui L., Lu D., Song X., Zhang F., He Y., Li F., Gao Y., Mao X., Zhang L., Qian J., Tan J., Jin J., Huang W., Deka R., Su B., Chakraborty R., Jin L., 2004, Genetic evidence supports demic diffusion of Han culture, *Nature*, 431(7006), pp. 302–305.

118. Yang M.A., Fan X., Sun B., Chen C., Lang J., Ko Y.C., Tsang C.H., Chiu H., Wang T., Bao Q., Wu X., Hajdinjak M., Ko A.M.S., Ding M., Cao P., Yang R., Liu F., Nickel B., Dai Q., Feng X., Zhang L., Sun C., Ning C., Zeng W., Zhao Y., Zhang M., Gao X., Cui Y., Reich D., Stoneking M., Fu Q., 2020, Ancient DNA indicates human population shifts and admixture in northern and southern China, *Science (New York, N.Y.)*, 369(6501), pp. 282–288.

119. Bendl J., Musil M., Štourač J., Zendulka J., Damborský J., Brezovský J., 2016, PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions, *PLoS Computational Biology*, 12(5), pp. e1004962.

120. Alsubaie L.M., Alsuwat H.S., Almandil N.B., AlSulaiman A.,

AbdulAzeez S., Borgio J.F., 2020, Risk Y-haplotypes and pathogenic variants of Arab-ancestry boys with autism by an exome-wide association study, *Molecular Biology Reports*, 47(10), pp. 7623–7632.

121. González-Fernández M., Vázquez-Coto D., Albaiceta G.M., Amado-Rodríguez L., Clemente M.G., Velázquez-Cuervo L., García-Lago C., Gómez J., Coto E., 2024, Chromosome-Y haplogroups in Asturias (Northern Spain) and their association with severe COVID-19, *Molecular Genetics and Genomics*, 299(1), pp. 49.

122. Sezgin E., Lind J.M., Shrestha S., Hendrickson S., Goedert J.J., Donfield S., Kirk G.D., Phair J.P., Troyer J.L., O'Brien S.J., Smith M.W., 2009, Association of Y chromosome haplogroup I with HIV progression, and HAART outcome, *Human Genetics*, 125(3), pp. 281.

PHỤ LỤC

Phụ lục 1. Nồng độ và độ tinh sạch của các mẫu dân tộc Pa Kô, Cơ-tu, Rơ-măm, Kinh miền Trung và Kinh miền Nam

STT	Mã mẫu	Dân tộc	Nồng độ (ng/µl)	OD
1	P01	PAKO	74	1,87
2	P02	PAKO	48	1,87
3	P03	PAKO	48,6	1,87
4	P04	PAKO	47,5	1,89
5	P05	PAKO	50,5	1,83
6	P06	PAKO	77,3	1,86
7	P07	PAKO	56,3	1,89
8	P08	PAKO	35,5	1,79
9	P09	PAKO	67,2	1,87
10	P10	PAKO	46,2	1,88
11	P11	PAKO	80,4	1,87
12	P12	PAKO	56,4	1,87
13	P13	PAKO	97,4	1,87
14	P14	PAKO	55,9	1,87
15	P15	PAKO	32,5	1,9
16	P16	PAKO	40,6	1,85
17	P17	PAKO	40,1	1,88
18	P18	PAKO	32,2	1,75

19	P19	PAKO	53,3	1,91
20	P20	PAKO	43,7	1,85
21	P21	PAKO	40,2	1,86
22	P22	PAKO	78,8	1,87
23	C01	COTU	40,8	1,84
24	C02	COTU	45	1,68
25	C03	COTU	55,1	1,84
26	C04	COTU	30,4	1,75
27	C05	COTU	36,6	1,79
28	C06	COTU	53,6	1,81
29	C07	COTU	54,5	1,78
30	C08	COTU	47,9	1,8
31	C09	COTU	49,3	1,8
32	C10	COTU	45	1,78
33	C11	COTU	55,7	1,75
34	C12	COTU	47,7	1,75
35	C13	COTU	41,5	1,82
36	C14	COTU	59,8	1,74
37	C15	COTU	42,2	1,72
38	C16	COTU	49,5	1,83
39	C17	COTU	46,6	1,83

40	C18	COTU	48,5	1,71
41	C19	COTU	41,3	1,76
42	R01	ROMAM	88,2	1,77
43	R02	ROMAM	84	1,82
44	R03	ROMAM	55,7	1,81
45	R04	ROMAM	77,7	1,85
46	R05	ROMAM	86,0	1,84
47	R06	ROMAM	90,7	1,84
48	R07	ROMAM	81,5	1,85
49	R08	ROMAM	60,2	1,79
50	R09	ROMAM	81,9	1,79
51	R10	ROMAM	104,7	1,84
52	R11	ROMAM	66,0	1,79
53	R12	ROMAM	52,0	1,80
54	R13	ROMAM	56,1	1,82
55	R14	ROMAM	64,3	1,82
56	R15	ROMAM	65,0	1,82
57	R16	ROMAM	62,2	1,82
58	R17	ROMAM	45,9	1,82
59	R18	ROMAM	87,1	1,83
60	R19	ROMAM	68,9	1,83

61	R20	ROMAM	78,0	1,84
62	R21	ROMAM	69,9	1,75
63	R22	ROMAM	70,2	1,82
64	R23	ROMAM	60,3	1,81
65	R24	ROMAM	50,4	1,80
66	KC01	KMT	61,7	1,79
67	KC02	KMT	45,4	1,82
68	KC03	KMT	82,4	1,83
69	KC04	KMT	52,4	1,81
70	KC05	KMT	97,8	1,85
71	KC06	KMT	61,6	1,73
72	KC07	KMT	40,2	1,59
73	KC08	KMT	44,2	1,84
74	KC09	KMT	49,6	1,79
75	KC10	KMT	47,3	1,76
76	KC11	KMT	103,3	1,84
77	KC12	KMT	62,9	1,78
78	KC13	KMT	60,1	1,78
79	KC14	KMT	31,6	1,59
80	KC15	KMT	58,8	1,8
81	KC16	KMT	47,1	1,74

82	KC17	KMT	32,3	1,81
83	KC18	KMT	59	1,85
84	KC19	KMT	59,3	1,84
85	KC20	KMT	77,4	1,83
86	KC21	KMT	51,8	1,79
87	KC22	KMT	80,8	1,84
88	KC23	KMT	33,3	1,81
89	KC24	KMT	30	1,78
90	KC25	KMT	78,1	1,8
91	KC26	KMT	48,3	1,77
92	KC27	KMT	61,1	1,85
93	KC28	KMT	65	1,8
94	KC29	KMT	57,2	1,83
95	KC30	KMT	64,7	1,82
96	KC31	KMT	48,2	1,79
97	KMN01	KMN	30,4	1,64
98	KMN02	KMN	63,9	1,82
99	KMN03	KMN	69,5	1,83
100	KMN04	KMN	89,2	1,86
101	KMN05	KMN	77	1,75
102	KMN06	KMN	95,4	1,75

103	KMN07	KMN	56,3	1,88
104	KMN08	KMN	67,4	1,84
105	KMN09	KMN	74,4	1,77
106	KMN10	KMN	56,5	1,86
107	KMN11	KMN	59,2	1,86
108	KMN12	KMN	48,4	1,73
109	KMN13	KMN	64,2	1,74
110	KMN14	KMN	49,9	1,65
111	KMN15	KMN	39,7	1,84
112	KMN16	KMN	40,5	1,74
113	KMN17	KMN	46,5	1,83
114	KMN18	KMN	64,4	1,77
115	KMN19	KMN	79,4	1,61
116	KMN20	KMN	32,6	1,69
117	KMN21	KMN	78,5	1,77
118	KMN22	KMN	40,9	1,76
119	KMN23	KMN	33,8	1,85
120	KMN24	KMN	49,6	1,54
121	KMN25	KMN	71,3	1,79
122	KMN26	KMN	48,8	1,84
123	KMN27	KMN	66,7	1,59

124	KMN28	KMN	59,7	1,53
125	KMN29	KMN	201,2	1,83
126	KMN30	KMN	61,6	1,81
127	KMN31	KMN	59,3	1,66
128	KMN32	KMN	66,7	1,77
129	KMN33	KMN	75,5	1,87

TẠP CHÍ KHOA HỌC & CÔNG NGHỆ VIỆT NAM
Vietnam Journal of Science and Technology - MOST

Hà Nội, ngày 15 tháng 07 năm 2024

GIẤY CHỨNG NHẬN

Tạp chí Khoa học và Công nghệ Việt Nam đã nhận được bài báo: "Đa dạng di truyền vùng D-loop của hai dân tộc Cơ-tu và Rơ-măm" của các tác giả: Lã Đức Duy[1, 2], Nông Văn Hải[1], Nguyễn Thùy Dương[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*

Bài báo này chúng tôi đã tiến hành biên tập, phản biện và chấp nhận đăng tải trên Tạp chí Khoa học và Công nghệ Việt Nam (Bản B) trong thời gian thích hợp.

Tạp chí làm Giấy chứng nhận này để các tác giả sử dụng khi cần thiết.

TỔNG BIÊN TẬP

TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

TS. Nguyễn Thị Hương Giang

Đa dạng di truyền vùng D-loop ở hai dân tộc Cơ-tu và Rơ-măm

Lã Đức Duy[1, 2], Nông Văn Hải[1], Nguyễn Thùy Dương[1, 2*]

[1]Viện Nghiên cứu Hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 18 Hoàng Quốc Việt, phường Nghĩa Đô, quận Cầu Giấy, Hà Nội, Việt Nam

[2]Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 18 Hoàng Quốc Việt, phường Nghĩa Đô, quận Cầu Giấy, Hà Nội, Việt Nam

Ngày nhận bài 6/6/2024; ngày chuyển phản biện 10/6/2024; ngày nhận phản biện 30/6/2024; ngày chấp nhận đăng 5/7/2024

Tóm tắt:

Vùng D-loop hệ gen ty thể (mitochondrial DNA - mtDNA), vùng có tần suất xuất hiện đột biến cao nhất, thường được dùng để nghiên cứu khảo sát đa dạng di truyền trên các quần thể người khác nhau. Trong nghiên cứu này, chúng tôi thu thập và tách chiết DNA tổng số từ mẫu máu của 19 cá thể dân tộc Cơ-tu và 24 cá thể dân tộc Rơ-măm. Sau khi giải trình tự mtDNA của 43 cá thể và so sánh với trình tự mtDNA tham chiếu RSRS (Reconstructed sapiens reference sequence), 41 điểm đa hình khác nhau được phát hiện trên vùng D-loop, trong đó có 27 và 30 điểm lần lượt được tìm thấy trên dân tộc Cơ-tu và Rơ-măm. Kết quả định danh nhóm đơn bội sử dụng trình tự vùng D-loop và phần mềm Haplogrep3 cho thấy có 12 nhóm khác nhau, trong đó chỉ có F1a1a xuất hiện ở cả hai dân tộc trên. Kết quả phân tích ma trận khoảng cách di truyền giữa các dân tộc trong nghiên cứu và hai dân tộc cùng ngữ hệ Việt Nam khác (Kinh và Mảng) cho thấy Cơ-tu và Rơ-măm có khoảng cách di truyền cao nhất (0,14545). Kết quả là nghiên cứu đầu tiên về vùng D-loop của hai dân tộc Cơ-tu và Rơ-măm sinh sống tại Việt Nam. Các kết quả mới này đã bổ sung thêm thông tin di truyền theo dòng mẹ và các biến thể ở D-loop ở hai dân tộc nghiên cứu.

Từ khóa: **biến thể, D-loop, Việt Nam.**

Chỉ số phân loại: **1.6, 3.1**

Genetic diversity of D-loop region from ethnic groups Co-tu and Ro-mam

Duc Duy La[1, 2], Van Hai Nong[1], Thuy Duong Nguyen[1, 2*]

[1]Institute of Genome Research, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Nghia Do Ward, Cau Giay District, Hanoi, Vietnam

[2]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Nghia Do Ward, Cau Giay District, Hanoi, Vietnam

Received 6 June 2024; revised 30 June 2024; accepted 5 July 2024

**Tác giả liên hệ: Email: tdnguyen@igr.ac.vn*

Abstract:

The D-loop region of mitochondrial DNA (mtDNA), with the highest mutation frequency in the mitochondrial genome, has been frequently used in studies on genetic diversity in different human populations. In this study, total DNA was collected and extracted from whole blood samples of 19 Co-tu and 24 Ro-mam individuals. After sequencing the mitochondrial DNA of 43 study subjects and aligning them to the mitochondrial DNA reconstructed sapiens reference sequence (RSRS), a total of 41 unique variants were obtained in the D-loop region in which 27 and 30 variants were found in the Co-tu and Ro-mam populations, respectively. Haplogroup identification based on the D-loop region revealed 12 haplogroups, among which only haplogroup F1a1a was found in both ethnic groups using Haplogrep3. After analysing the genetic distances of the study populations and other previously published populations (Kinh and Mang), the results showed that the genetic distance between Co-tu and Ro-mam (0.14545) was the highest. This is the first study on the genetic diversity of the D-loop region of the Co-tu and Ro-mam ethnic groups living in Vietnam. These results provide more insights into the maternal genetic background and D-loop A variants of the two studied ethnic groups.

Keywords: **D-loop, variant, Vietnam.**

Classification numbers: **1.6, 3.1**

1. Đặt vấn đề

Ngữ hệ Nam Á (NHNA) là một trong những ngữ hệ lớn nhất trên thế giới bao gồm 167 ngôn ngữ được nói khắp Nam Á (NA), Đông Á (ĐA), và Đông Nam Á (ĐNA) [1]. Tại NA, ngôn ngữ thuộc NHNA được phân loại vào nhánh phụ Munic và là một trong bốn ngữ hệ chính ở miền Nam và miền Trung Ấn Độ. Tại ĐA, NHNA được nói rải rác ở phía Nam Trung Quốc, chẳng hạn như tiếng Bolyu và Bagan [2]. Ở ĐNA, các bằng chứng về khảo cổ và ngôn ngữ học chỉ ra rằng, NHNA là ngữ hệ xuất hiện đầu tiên ở vùng ĐNA lục địa (Mainland Southeast Asia - MSEA), sau đó với sự xuất hiện của các hệ ngôn ngữ khác như Nam Đảo (Austronesian, NHNĐ), Thái - Kadai (Tai - Kadai, NHTK), Hán - Tạng (Sino - Tibetan, NHHT) và Mông - Miền (H'mong - Mien, NHHM), NHNA bị cô lập và phân tán rải rác ở ĐNA [3, 4]. Vào thời điểm hiện tại, ngôn ngữ được nói chủ yếu ở vùng ĐNA hải đảo (island Southeast Asia - ISEA) thuộc về NHNĐ còn ở MSEA thì NHNA vẫn là hệ ngôn ngữ được sử dụng nhiều nhất [1, 5]. Tại MSEA, NHNA là quốc ngữ của Campuchia và Việt Nam và cũng được sử dụng bởi một số dân tộc thiểu số ở Thái Lan, Lào, Myanmar và vùng bán đảo của Malaysia. Với xấp xỉ 126 triệu người sử dụng NHNA [1], ngữ hệ này là ngữ hệ đứng thứ 8 về

số lượng người nói trên thế giới và thứ 3 tại ĐNA, tuy nhiên nguồn gốc và con đường phát tán của NHNA hiện vẫn là vấn đề chưa được giải quyết [5-7].

Với vị trí địa lý chiến lược kết nối bán đảo Đông Dương và ISEA, Việt Nam trở thành khu vực có bề dày lịch sử phong phú và phức tạp được tạo nên bởi nhiều làn sóng di cư của loài người hiện đại từ nhiều hướng khác nhau. Đồng thời, Việt Nam cũng sở hữu số lượng người nói NHNA nhiều nhất trên thế giới với khoảng 90% dân số nói ngôn ngữ thuộc ngữ hệ này (Tổng điều tra dân số và nhà ở năm 2019, www.gso.gov.vn). Trong 25 dân tộc sử dụng NHNA, dân tộc Kinh chiếm khoảng 85% số lượng người sử dụng ngôn ngữ và 24 dân tộc thiểu số còn lại chiếm khoảng 5%. Cho đến nay, các nghiên cứu nhân học phân tử (molecular anthropology) về các dân tộc NHNA ở Việt Nam còn ít, đặc biệt là các nghiên cứu trên các dân tộc thiểu số [8, 9]. Các dân tộc thiểu số thuộc NHNA ở Việt Nam sống thưa thớt, có phần tách biệt trải dài khắp địa bàn cả nước dẫn đến sự đa dạng di truyền của các dân tộc này có thể không được tìm thấy ở quần thể người Kinh, đặc biệt là các dân tộc di cư vào Việt Nam từ các nước lân cận mang theo nguồn gen có thể không có ở quần thể người bản địa.

Để nghiên cứu về đa dạng di truyền, đặc biệt làm sáng tỏ các con đường di cư, sự giao thoa giữa các dân tộc với nhau và sự khác biệt di truyền giữa các dân tộc khác nhau, hệ gen ty thể thường được lựa chọn làm đối tượng nghiên cứu [10, 11]. Trong mtDNA, vùng D-loop (vị trí nucleotide 1-576 và 16.024-16.569 của mtDNA) là vùng không mã hóa và đóng vài trò quan trọng trong việc điều hòa quá trình tái bản và phiên mã của mtDNA [12]. Vùng D-loop cũng là vùng có tần suất xuất hiện đột biến cao nhất trên mtDNA [13], do đó vùng này thường xuyên được sử dụng để nghiên cứu khảo sát đa dạng di truyền theo dòng mẹ trên các quần thể người khác nhau [11]. Vì vậy, trong nghiên cứu này, trình tự vùng D-loop trên mtDNA của 43 cá thể thuộc hai dân tộc nói NHNA tại Việt Nam (Cơ-tu, Rơ-măm) đã được giải trình tự và phân tích. Ngoài ra, vùng D-loop của hai dân tộc Kinh và Mảng thuộc NHNA đã công bố trước đây cũng được sử dụng để phân tích, so sánh sự khác biệt giữa các dân tộc trong cùng ngữ hệ [8].

2. Đối tượng và phương pháp nghiên cứu

2.1. Đối tượng

Mẫu máu ngoại vi được thu từ các cá thể thuộc NHNA người Việt gồm 19 cá thể thuộc dân tộc Cơ-tu (Thừa Thiên Huế) và 24 cá thể thuộc dân tộc Rơ-măm (Kon Tum). Mẫu được chọn là mẫu không có quan hệ huyết thống và có ít nhất ba thế hệ trong gia đình đều thuộc một dân tộc. Các cá thể tham gia nghiên cứu đều ký vào giấy đồng ý tự nguyện cho máu cho nghiên cứu. Nghiên cứu này được thông qua bởi Hội đồng Đạo đức trong nghiên cứu sinh của Viện Nghiên cứu Hệ gen, Viện Hàn lâm Khoa học và Công nghệ Việt Nam (Số 9-2019/NCHG-HĐĐĐ).

2.2. Giải trình tự vùng D-loop hệ gen ty thể

DNA từ bộ gen được chiết xuất bằng GeneJET Whole Blood Genomic DNA Purification Mini Kit (ThermoFisher Scientific, Hoa Kỳ) theo hướng dẫn của nhà sản xuất. Việc xây dựng thư viện bộ gen và làm giàu mtDNA đã được thực hiện như mô tả ở nghiên cứu khác [14]. Các thư viện đã được giải trình tự trên nền tảng NovaSeq 6000 (Illumina, Hoa Kỳ) với các đoạn đọc kép có chiều dài 150 bp. Các đoạn đọc được kiểm tra chất lượng bởi FastQC và được dóng hàng với trình tự tham chiếu Sapiens được sửa đổi (Reconstructed Sapiens Reference Sequence - RSRS) [15] bằng Burrows-Wheeler Alignment (BWA). Sau đó, các trình tự D-loop sẽ được dóng hàng với nhau bằng MAFFT [16].

2.3. Phân tích số liệu

Trình tự vùng D-loop của 43 cá thể thuộc nghiên cứu này kết hợp với 87 trình tự đã được nghiên cứu bao gồm 50 cá thể Kinh và 37 cá thể Mảng [8, 9] được sử dụng để xác định các nhóm đơn bội bằng phần mềm Haplogrep3 [17] và cây PhyloTree mtDNA phiên bản 17.1 [18]. Khoảng cách di truyền giữa các dân tộc Cơ-tu, Rơ-măm, Kinh và Mảng được tính dựa trên chỉ số Φ_{ST} sử dụng phần mềm Arlequin [19]. Sự phân bố của các nhóm đơn bội trong bốn nhóm dân tộc được thể hiện dưới dạng biểu đồ phân tích tương quan (Correspondence analysis - CA) dựa vào tần suất xuất hiện của các nhóm đơn bội. Biểu đồ CA được vẽ trong R [20] sử dụng các gói hỗ trợ "vegan" và "ca". Kiểm định Fisher's exact (2 phía) được sử dụng để so sánh tỷ lệ xuất hiện đa hình giữa hai dân tộc Cơ-tu và Rơ-măm. Tất cả giá trị p nhỏ hơn 0,05 đều được coi là có ý nghĩa.

3. Kết quả và bàn luận

3.1. Tần suất xuất hiện các điểm biến đổi ở vùng D-loop

Trong 43 cá thể tham gia nghiên cứu, chúng tôi phát hiện được 41 điểm biến đổi khác nhau ở vùng D-loop với tất cả các biến đổi đều là điểm đa hình nucleotide đơn (Single nucleotide polymorphism - SNP) (hình 1). Số lượng biến thể tối đa và tối thiểu xuất hiện trong một cá thể của mỗi dân tộc tương đối đồng đều (bảng 1). Hai điểm biến đổi C16311T và G16230A xuất hiện ở trên tất cả các mẫu trong nghiên cứu, đứng thứ hai là C152T được tìm thấy ở tất cả cá thể của dân tộc Rơ-măm và 95% cá thể của dân tộc Cơ-tu. Kết hợp với dữ liệu biến thể của 81.124 trình tự vùng điều khiển mtDNA thu thập từ Mitomap [30], hầu hết những điểm biến đổi xuất hiện với tần suất cao trong nghiên cứu hiện tại cũng được phát hiện ở phần lớn các trình tự trong Mitomap, ví dụ như G16230A (79.785/81.124), A247G (78.859/81.124), C16311T (67.049/81.124) và T16223C (40.381/81.124). Hai biến đổi T489C và T16140C xuất hiện ở ít trình tự hơn lần lượt là 12.479 và 2512. Ngoài ra, chỉ có hai điểm (T489C, T16140C) đã được phát hiện với tần suất cao trong vùng D-loop của các quần thể người Việt Nam thuộc NHNA đã được nghiên cứu trước đó [21].

Bảng 1. Số lượng các điểm biến đổi tìm thấy trong nghiên cứu.

thứ tự	Dân tộc	Số lượng cá thể	Số lượng biến thể ở vùng D-loop		
			Giá trị tối đa giữa các cá thể	*Giá trị tối thiểu giữa các cá thể*	*Giá trị trung bình ± độ lệch chuẩn*
1	Cơ-tu	19	12	9	10,63±0,45
2	Rơ-măm	24	13	8	10,25±0,45

Ở các điểm biến đổi có tần suất thấp hơn, tần suất xuất hiện của điểm biến đổi giữa các quần thể khác nhau có sự khác biệt rõ ràng (hình 1). Cụ thể hơn, điểm biến đổi T16233C xuất hiện ở đa số người Cơ-tu (89%) trong khi chỉ xuất hiện ở 25% người Rơ-măm. Ngược lại, T489C chỉ xuất hiện ở số ít người Cơ-tu (11%) nhưng lại được tìm thấy ở phần lớn người Rơ-măm (62%). Sự khác biệt rõ ràng nhất là hai điểm biến đổi A210G và T16140C, khi hai điểm này có ở phần lớn người Cơ-tu với tần suất đều là 63% nhưng lại không xuất hiện ở người Rơ-măm. Đối với những điểm biến đổi xuất hiện ở tần suất thấp hơn 30% (phân tích trên từng dân tộc), đa số biến đổi chỉ được tìm thấy ở một dân tộc ví dụ như T16298C, C16294T, T16304G, A200G. Kết quả của kiểm định Fisher's exact cho thấy có 7 biến đổi có tần suất xuất hiện khác biệt đáng kể giữa hai dân tộc (T16140C, A210G, T16223C, T489C, T199C, C16294T và A16284G), trong đó T16140C, A210G và T16223C có giá trị p nhỏ hơn 0,0001 (dữ liệu không được thể hiện).

Hình 1. Tần suất xuất hiện của các biến đổi trong hai dân tộc Cơ-tu và Rơ-măm. Mỗi điểm tương ứng với tần suất xuất hiện của các biến đổi. Màu đỏ: biến đổi xuất hiện ở cả hai nhóm dân tộc; màu tím: biến đổi chỉ xuất hiện ở một dân tộc; điểm hình tròn: biến đổi chỉ xuất hiện ở dân tộc Cơ-tu; điểm hình tam giác: biến đổi chỉ xuất hiện ở dân tộc Rơ-măm.

3.2. Sự phân bố nhóm đơn bội

Tổng cộng 12 nhóm đơn bội thuộc 3 nhóm đơn bội lớn (M, N9, R) đã được xác định từ 43 cá thể nam thuộc 2 dân tộc Cơ-tu và Rơ-măm thông qua phần mềm Haplogrep3 (bảng 2). Phần lớn các nhóm đơn bội ở vùng D-loop của hai dân tộc này thuộc nhóm đơn bội lớn R và M, chiếm 40/43 số lượng nhóm đơn bội trong bộ mẫu nghiên cứu. Trong đó, nhóm đơn bội F1a1a là nhóm duy nhất xuất hiện ở cả hai dân tộc. Trong số các nhóm đơn bội chỉ xuất hiện ở một dân tộc, có bốn nhóm đơn bội chỉ được phát hiện trên một cá thể (B5a1d, B4m, B4c2, M7b1a1). Nhóm đơn bội B5a xuất hiện với tần suất cao nhất chiếm 25% số lượng nhóm đơn bội mẫu nghiên cứu và chỉ có ở người Cơ-tu (58%) (bảng 2). Đặc biệt, khác với người Rơ-măm, vốn gen ty thể của người Cơ-tu phần lớn là thuộc macro-nhóm đơn bội R. Đây có thể là kết quả của hiện tượng người sáng lập (Founder effect) hoặc trôi dạt di truyền (Genetic drift) kết hợp với việc sống tách biệt trong thời gian dài. Còn đối với dân tộc Rơ-măm, mặc dù dân tộc này có số lượng nhóm đơn bội ít hơn nhưng lại có số lượng nhóm đơn bội lớn nhiều hơn với sự xuất hiện của nhóm đơn bội lớn N.

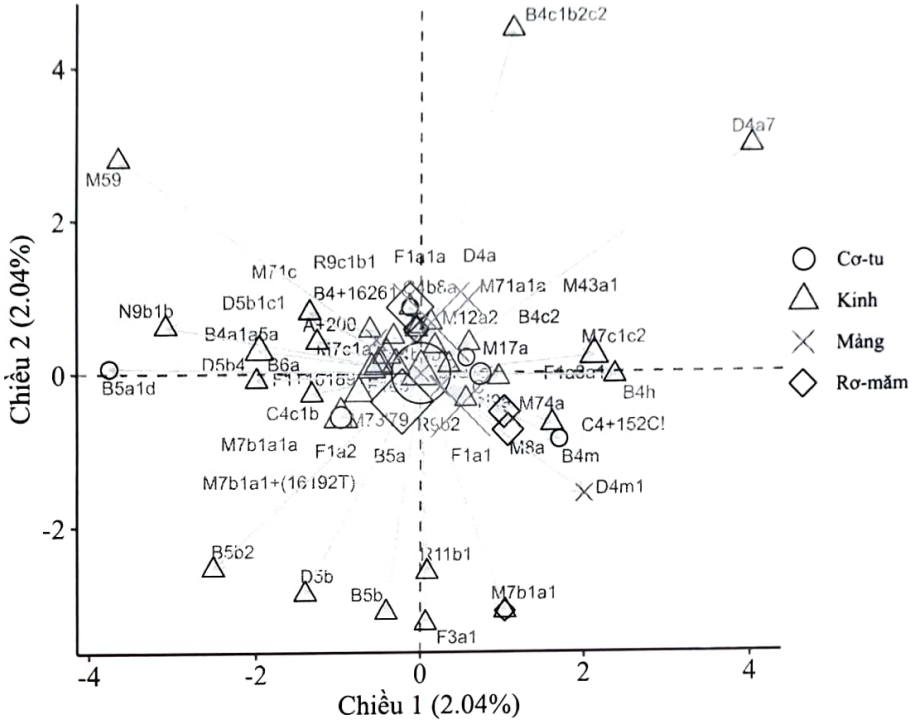
Bảng 2. Tần suất nhóm đơn bội trên hai dân tộc Cơ-tu và Rơ-măm.

Nhóm đơn bội	Nhóm đơn bội lớn	Cơ-tu (n=19)	Rơ-măm (n=24)	Tổng thể (n=43)
	M	17/43		
M17a	M	11% (2/19)	-	5%
M73'79	M	-	38% (9/24)	21%
M8a	M	-	12% (3/24)	7%
M7b1a1	M	-	4% (1/24)	2%
C4b8a	M	-	8% (2/24)	5%
	N	3/43		
N9a	N	-	12% (3/24)	7%
	R	23/43		
B5a	R	58% (11/19)	-	26%
F1a2	R	11% (2/19)	-	5%
F1a1a	R	5% (1/19)	25% (6/24)	16%
B5a1d	R	5% (1/19)	-	2%
B4m	R	5% (1/19)	-	2%
B4c2	R	5% (1/19)	-	2%

n: số lượng cá thể.

Trình tự vùng D-loop của 50 cá thể thuộc dân tộc Kinh và 37 cá thể Mảng được sử dụng để định danh lại các nhóm đơn bội [8, 9] (hình 2). Kết quả được so sánh với các nhóm đơn bội đã được định danh sử dụng toàn bộ trình tự mtDNA công bố trước đây [8, 9] và phát hiện 5/50 mẫu ở dân tộc Kinh và 1/37 mẫu Mảng có nhóm đơn bội bị thay đổi sang nhánh khác,

cụ thể dân tộc Kinh có các nhóm bị thay đổi: C7>C4b8a, M74b2>M43a1, C7a2>C4c1b, F1g>N9b1b và C7a>C4+152C và dân tộc Mảng có nhóm M61 chuyển thành D4m1. Kết quả so sánh cho thấy phần lớn nhóm đơn bội vẫn giữ nguyên nhánh phân loại của mình, do đó sử dụng vùng D-loop cho nghiên cứu quần thể là có độ tin cậy cao.



Hình 2. Biểu đồ tương quan nhóm đơn bội của bốn dân tộc Cơ-tu, Rơ-măm, Kinh và Mảng. Kích cỡ của các điểm trong đồ thị tỷ lệ thuận với số lượng cá thể của mỗi nhóm đơn bội.

Các nhóm đơn bội của hai dân tộc Kinh và Mảng cũng được phân tích, so sánh với các nhóm của dân tộc Cơ-tu và Rơ-măm (hình 2). Kết quả so sánh phát hiện được 8 nhóm đơn bội mới xuất hiện ở hai dân tộc Cơ-tu và Rơ-măm mà không xuất ở Kinh và Mảng [8], cụ thể là nhóm đơn bội F1a2, M17a, B5a1d, B4c2, M73'79, M8a, N9a và C4b8a. Trong đó, M73'79 là nhóm có tần suất xuất hiện cao nhất ở người Rơ-măm (38%), cho thấy tiềm năng trong việc nghiên cứu đa dạng di truyền ở các dân tộc thiểu số. Nhóm đơn bội B5a - nhóm đơn bội xuất hiện với tần suất cao nhất trong nghiên cứu này, cũng xuất hiện ở các nhóm ngữ hệ khác không phải NHNA [8].

Nhóm đơn bội lớn M được tìm thấy ở khắp châu Á với tần suất cao [22, 23], đặc biệt M chiếm đa số quần thể Ấn Độ (60%) [24] và phần lớn ở quần thể người Thái Lan (63/166 [23] và 847/1234 [22]). Trong nghiên cứu này, nhóm đơn bội thuộc nhóm đơn bội lớn M xuất hiện với tần suất lớn nhất là M73'79 (21%), nhóm này cũng được phát hiện ở quần thể PaduangKaren (3/25) thuộc NHHT ở Thái Lan [25]; các nhóm đơn bội M khác cũng được tìm thấy ở các quần thể lân cận như M8a ở một số quần thể người Thái Lan thuộc NHTK [22], M17a ở quần thể NHNA và NHTK ở Thái Lan [22, 25], quần thể NHTK ở Lào [22] và

quần thể NHNA và NHNĐ ở Campuchia [26]. Đại diện của nhóm đơn bội lớn R trong nghiên cứu này có thể được nhóm vào hai nhóm đơn bội B và F. Nhóm đơn bội B4 và B5 là những nhóm đơn bội thường gặp nhất ở Bắc Á và Tây Á [27] với B4 là nhóm đơn bội có tần suất cao thứ hai ở Việt Nam và trải dài ở Thái Lan và Đài Loan (Trung Quốc) [22, 28] còn nhóm đơn bội B5 thì xuất hiện nhiều nhất ở phía bắc Thái Lan [8]. Các nhóm đơn bội thuộc nhóm B được tìm thấy trong nghiên cứu này cũng được phát hiện tại các quần thể thuộc nhiều ngữ hệ khác nhau ở các nước lân cận [24, 25, 28, 29]. Tương tự, nhóm đơn bội F cũng là nhóm đơn bội thường gặp ở châu Á và xuất hiện với tần suất cao ở phía bắc Việt Nam và Thái Lan [8]. Ngoài ra, hai nhóm đơn bội thuộc nhóm F là F1a2 và F1a1a cũng được tìm thấy ở quần thể người Dai Trung Quốc [30] và quần thể người Hakka ở Đài Loan (Trung Quốc) tương ứng [31]. Đối với nhóm đơn bội lớn N, đại diện duy nhất được tìm thấy trong nghiên cứu này là nhóm đơn bội N9a. Nhánh N9a của macro-nhóm đơn bội N được phát hiện ở nhiều quần thể người ĐA, NA, và ĐNA [31, 32], đặc biệt là ở bán đảo của Malaysia (28/86) [8].

3.3. Mối quan hệ di truyền giữa các dân tộc

Để phân tích quan hệ di truyền giữa các dân tộc trong cùng NHNA, hai dân tộc Rơ-măm và Cơ-tu đã được phân tích cùng với hai dân tộc Kinh và Mảng. Phần mềm Arlequin đã được sử dụng để tính ma trận khoảng cách Φ_{ST} dựa vào trình tự vùng D-loop của các cá thể thuộc bốn dân tộc trên (bảng 3). Kết quả phân tích ma trận Φ_{ST} cho thấy giữa người Cơ-tu và người Rơ-măm có khoảng cách di truyền lớn nhất (0,14545). Khoảng cách giữa người Kinh và người Rơ-măm (0,06368) lớn hơn khoảng cách giữa người Kinh với người Cơ-tu (0,03558). Ngoài ra, khoảng cách di truyền giữa dân tộc Kinh đối với dân tộc khác thấp hơn khoảng cách di truyền giữa các dân tộc còn lại với nhau. Điều này phù hợp với thực trạng phân bố khắp cả nước với số lượng cá thể nhiều nhất của dân tộc Kinh, do đó dân tộc Kinh trở thành dân tộc có khả năng tương tác và trao đổi thông tin di truyền cao nhất với các dân tộc khác tại Việt Nam.

Bảng 3. Ma trận khoảng cách Φ_{ST} cho vùng D-loop giữa 4 dân tộc Cơ-tu, Rơ-măm, Mảng và Kinh.

	Cơ-tu	**Rơ-măm**	**Mảng**	**Kinh**
Cơ-tu	0,00000	0,14545 (0,00198±0,0004)	0,13284 (0,00010±0,0001)	0,03558 (0,03693±0,0019)
Rơ-măm	0,14545 (0,00198±0,0004)	0,00000	0,10575 (0,00000±0,0000)	0,06368 (0,00059±0,0002)
Mảng	0,13284 (0,00010±0,0001)	0,10575 (0,00000±0,0000)	0,00000	0,02642 (0,01426±0,0013)
Kinh	0,03558 (0,03693±0,0019)	0,06368 (0,00059±0,0002)	0,02642 (0,01426±0,0013)	0,00000

Giá trị trong ngoặc đơn là giá trị p.

Hơn nữa, phân bố của các nhóm đơn bội trong bốn dân tộc Cơ-tu, Rơ-măm, Kinh và Mảng bằng biểu đồ tương quan CA cho thấy, những nhóm đơn bội xuất hiện với tần suất cao như M73'79 và M71a1a ở các dân tộc tương ứng (Rơ-măm và Mảng) và xuất hiện ở ít nhất hai dân tộc (B5a và F1a1a) sẽ nằm ở gần phía giữa của biểu đồ còn những nhóm đơn bội xuất hiện ở tần suất thấp như M59, D4a7 và B4c1b2c2 sẽ nằm ở phía ngoài của biểu đồ (hình 2). Mặc dù khoảng cách di truyền giữa quần thể Kinh đối với hai quần thể Cơ-tu và Rơ-măm thấp nhưng ba quần thể này lại không có nhiều nhóm đơn bội chung, thể hiện tính đặc trưng của người Cơ-tu và Rơ-măm. Sự đa dạng di truyền của các dân tộc NHNA ở Việt Nam cũng được thể hiện ở biểu đồ này với hầu hết các nhóm đơn bội chỉ tồn tại ở một dân tộc.

4. Kết luận

Trong nghiên cứu này, vùng D-loop của 43 cá thể từ hai quần thể người Cơ-tu và Rơ-măm thuộc NHNA tại Việt Nam đã được giải trình tự và phân tích. Hai điểm biến đổi (C16311T và G16230A) được tìm thấy ở tất cả các cá thể và có 7 SNP có tần suất xuất hiện khác biệt đáng kể giữa hai dân tộc (T16140C, A210G, T16223C, T489C, T199C, C16294T và A16284G). 12 nhóm đơn bội phát hiện ở dân tộc Cơ-tu và Rơ-măm thuộc những nhóm đơn bội lớn thường gặp ở ĐNA gồm M, B và N. Thành phần nhóm đơn bội của giữa hai dân tộc này tương đối khác biệt chỉ có một nhóm đơn bội chung là F1a1a. Điều này phản ảnh mức độ đa dạng di truyền cao giữa hai dân tộc thiểu số Cơ-tu và Rơ-măm. Khi kết hợp với dữ liệu của dân tộc Kinh với Mảng đã công bố, khoảng cách giữa người Cơ-tu và Rơ-măm lớn hơn khoảng cách giữa hai dân tộc này đối với dân tộc Kinh và Mảng. Những phát hiện này sẽ mang đến cái nhìn sâu sắc mới về nhóm NHNA và bổ sung thêm dữ liệu quan trọng cho các nghiên cứu tiếp theo về ngữ hệ này nói riêng và các ngữ hệ khác của dân tộc Việt Nam nói chung.

LỜI CẢM ƠN

Nghiên cứu được hoàn thành với sự đồng ý tham gia của những cá thể cho phép lấy mẫu và được tài trợ bởi nhiệm vụ khoa học và công nghệ quốc gia do Bộ Khoa học và Công nghệ quản lý (mã số ĐTĐL.CN.60/19). Lã Đức Duy được nhận học bổng của Chương trình học bổng đào tạo thạc sỹ, tiến sỹ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), mã số VINIF.2023.ThS.027.

TÀI LIỆU THAM KHẢO

[1] D.M. Eberhard, G.F. Simons, C.D. Fennig (2023), *Ethnologue: Languages of The World*, Twenty-six, SIL International.

[2] P. Sidwell, M. Jenny (2021), "The languages and linguistics of Mainland Southeast Asia: A comprehensive guide", *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, pp.1-968, DOI: 10.1515/9783110558142/EPUB.

[3] P. Sidwell, R. Blench (2011), *The Austroasiatic Urheimat: The Southeastern*

Riverine Hypothesis, Pacific Linguistics, 315pp.

[4] P.S. Bellwood (2005), *First Farmers: The Origins of Agricultural Societies*, Wiley-Blackwell, 384pp.

[5] M. Klamer (2019), "The dispersal of Austronesian languages in island South East Asia: Current findings and debates", *Language and Linguistics Compass*, **13(4)**, DOI: 10.1111/LNC3.12325.

[6] M. Lipson, O. Cheronet, S. Mallick, et al. (2018), "Ancient genomes document multiple waves of migration in Southeast Asian prehistory", *Science*, **361(6397)**, pp.92-95, DOI: 10.1126/SCIENCE.AAT3188.

[7] D. Tagore, F. Aghakhanian, R. Naidu, et al. (2021), "Insights into the demographic history of Asia from common ancestry and admixture in the genomic landscape of present-day Austroasiatic speakers", *BMC Biology*, **19(1)**, DOI: 10.1186/S12915-021-00981-X.

[8] N.T. Duong, E. Macholdt, N.D. Ton, et al. (2018), "Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia", *Scientific Reports*, **8(1)**, pp.1-13, DOI: 10.1038/s41598-018-29989-0.

[9] E. Macholdt, L. Arias, N.T. Duong, et al. (2020), "The paternal and maternal genetic history of Vietnamese populations", *European Journal of Human Genetics*, **28(5)**, pp.636-645, DOI: 10.1038/s41431-019-0557-4.

[10] S. Pischedda, R.B. Arca, A.G. Carballa, et al. (2017), "Phylogeographic and genome-wide investigations of Vietnam ethnic groups reveal signatures of complex historical demographic movements", *Scientific Reports*, **7(1)**, pp.1-15, DOI: 10.1038/s41598-017-12813-6.

[11] N.T. Duong, N.V. Phong, N.T. Ngoc, et al. (2020), "Study on genetic variations of the D-loop region in three Vietnamese ethnic groups Kinh, Lolo and Lahu", *Vietnam Journal of Biotechnology*, **18(2)**, pp.231-238, DOI: 10.15625/1811-4989/18/2/15136.

[12] J.W. Taanman (1999), "The mitochondrial genome: Structure, transcription, translation and replication", *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, **1410(2)**, pp.103-123, DOI: 10.1016/S0005-2728(98)00161-3.

[13] M. Stoneking (2000), "Hypervariable sites in the mtDNA control region are mutational hotspots", *American Journal of Human Genetics*, **67(4)**, pp.1029-1032, DOI: 10.1086/303092.

[14] T. Maricic, M. Whitten, S. Pääbo (2010), "Multiplexed DNA sequence capture of mitochondrial genomes using PCR products", *PLOS ONE*, **5(11)**, DOI: 10.1371/Journal.pone.0014004.

[15] D.M. Behar, M.V. Oven, S. Rosset, et al. (2012), "A "Copernican" reassessment

of the human mitochondrial DNA tree from its root", *American Journal of Human Genetics*, **90(4)**, DOI: 10.1016/J.AJHG.2012.03.002.

[16] K. Katoh, D.M. Standley (2013), "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability", *Molecular Biology and Evolution*, **30(4)**, DOI: 10.1093/MOLBEV/MST010.

[17] S. Schönherr, H. Weissensteiner, F. Kronenberg, et al. (2023), "Haplogrep 3 - An interactive haplogroup classification and analysis platform", *Nucleic Acids Research*, **51(W1)**, pp.W263-W268, DOI: 10.1093/NAR/GKAD284.

[18] M.V. Oven, M. Kayser (2009), "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation", *Human Mutation*, **30(2)**, pp.E386-E394, DOI: 10.1002/HUMU.20921.

[19] L. Excoffier, H.E.L. Lischer (2010), "Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 10: 564-567", *Evolutionary Bioinformatics Online*.

[20] R.C. Team (2021), "R: A language and environment for statistical computing", https://www.r-project.org/, accessed 2021.

[21] N.T. Ngoc, N.B. Trang, N.Q. Huy, et al. (2018), "Single nucleotide polymorphisms in the D-loop region of the mitochondrial genomes of individuals from two ethnic groups Kinh and mang of austro-asiatic language family", *Vietnam Journal of Biotechnology*, **16(2)**, pp.231-240, DOI: 10.15625/1811-4989/16/2/13432.

[22] W. Kutanan, J. Kampuansai, M. Srikummool, et al. (2017), "Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai-Kadai languages", *Human Genetics*, **136(1)**, pp.85-98, DOI: 10.1007/S00439-016-1742-Y.

[23] K. Jaisamut, R. Pitiwararom, P. Sukawutthiya, et al. (2023), "Unraveling the mitochondrial phylogenetic landscape of Thailand reveals complex admixture and demographic dynamics", *Scientific Reports*, **13(1)**, DOI: 10.1038/S41598-023-47762-W.

[24] R. Rajkumar, J. Banerjee, H.B. Gunturi, et al. (2005), "Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages", *BMC Evolutionary Biology*, **5**, DOI: 10.1186/1471-2148-5-26.

[25] W. Kutanan, J. Kampuansai, A. Brunelli, et al. (2018), "New insights from Thailand into the maternal genetic history of Mainland Southeast Asia", *European Journal of Human Genetics*, **26(6)**, DOI: 10.1038/S41431-018-0113-7.

[26] A.K. Brandstätter, M. Summerer, D. Horst, et al. (2021), "An in-depth analysis of the mitochondrial phylogenetic landscape of Cambodia", *Scientific Reports*, **11(1)**, DOI:

10.1038/S41598-021-90145-2.

[27] M. Derenko, B. Malyarchuk, T. Grzybowski, et al. (2007), "Phylogeographic analysis of mitochondrial DNA in northern Asian populations", *American Journal of Human Genetics*, **81(5)**, DOI: 10.1086/522933.

[28] M. Bodner, B. Zimmermann, A. Röck, et al. (2011), "Southeast Asian diversity: First insights into the complex mtDNA structure of Laos", *BMC Evolutionary Biology*, **11(1)**, DOI: 10.1186/1471-2148-11-49.

[29] W. Woravatin, M. Stoneking, M. Srikummool, et al. (2023), "South Asian maternal and paternal lineages in southern Thailand and the role of sex-biased admixture", *PLOS ONE*, **18(9)**, DOI: 10.1371/Journal.pone.0291547.

[30] A. Auton, G.R. Abecasis, D.M. Altshuler, et al. (2015), "A global reference for human genetic variation", *Nature*, **526(7571)**, pp.68-74, DOI: 10.1038/nature15393.

[31] A.M.S. Ko, C.Y. Chen, Q. Fu, et al. (2014), "Early austronesians: Into and out of Taiwan", *American Journal of Human Genetics*, **94(3)**, pp.426-436, DOI: 10.1016/J.AJHG.2014.02.003.

[32] C. Hill, P. Soares, M. Mormina, et al. (2006), "Phylogeography and ethnogenesis of aboriginal Southeast Asians", *Molecular Biology and Evolution*, **23(12)**, pp.2480-2491, DOI: 10.1093/MOLBEV/MSL124.