

GUIDELINES FOR ANALYZING THE GENETIC POLYMORPHISM OF *PLASMODIUM FALCIPARUM* AND *PLASMODIUM VIVAX* BASED ON WHOLE-GENOME SEQUENCING DATA

Nguyen Thi Hong Ngoc^{1*}, Nguyen Thi Huong Binh¹, Nguyen Quang Thieu¹
Nguyen Van Hong¹, Nguyen Duc Giang¹, Clark Taanea²

¹National Institute of Malariaology, Parasitology and Entomology - 34 Trung Van, Nam Tu Liem district, Hanoi, Vietnam

²London School of Hygiene and Tropical Medicine - Keppel Street, London, WC1E 7HT, United Kingdom

Received: 28/02/2025

Revised: 29/3/2025; Accepted: 11/4/2025

ABSTRACT

Objective: This study provides a guideline for analyzing the genetic polymorphism of *Plasmodium falciparum* and *Plasmodium vivax* using whole-genome sequencing data.

Research methods: A standardized workflow for genomic data analysis was applied, including variant calling (SNPs, InDels) from raw sequencing data, variant annotation and classification, phylogenetic tree construction and population structure analysis, and association analysis between genetic variants and antimalarial drug resistance.

Results: The total number of SNPs detected ranged from 50,000-80,000 in *P. falciparum* and 40,000-60,000 in *P. vivax*, while InDels varied between 5,000 and 15,000. The *pfert* (72-76 CVIET) mutation associated with Chloroquine resistance was highly prevalent, reaching 92.7% (51/55) in 2017-2018 and 91.1% (72/79) in 2019-2020 in Vietnam, whereas its prevalence in Africa was significantly lower (< 5%) due to the transition to Artemisinin-based treatments. The *k13* C580Y mutation, a key marker of Artemisinin resistance, was most frequently observed along the Thailand - Cambodia border (60%). In Vietnam, the mutation had a prevalence of 82% (41/50), indicating the regional spread of resistant strains within Southeast Asia. Notably, the *pfmdr1* copy number variation (> 1.5 copies), associated with Mefloquine resistance, was detected in 23.6% (13/55) of cases in 2017-2018, but this decreased to 1.3% (1/79) in 2019-2020 in Vietnam.

Conclusion: The application of whole-genome sequencing combined with advanced bioinformatics approaches provides crucial insights into the genetic polymorphism of malaria parasites, supporting drug resistance surveillance and informing more effective malaria control strategies.

Keywords: *Plasmodium falciparum*, *Plasmodium vivax*, genetic polymorphism, whole-genome sequencing, drug resistance.

*Corresponding author

Email: ngocnime@gmail.com Phone: (+84) 983665694 <https://doi.org/10.52163/yhc.v66iCD4.2343>

HƯỚNG DẪN PHÂN TÍCH TÍNH ĐA HÌNH DI TRUYỀN CỦA *PLASMODIUM FALCIPARUM* VÀ *PLASMODIUM VIVAX* DỰA TRÊN SỐ LIỆU GIẢI TRÌNH TỰ TOÀN BỘ HỆ GEN

Nguyễn Thị Hồng Ngọc^{1*}, Nguyễn Thị Hương Bình¹, Nguyễn Quang Thiều¹
Nguyễn Văn Hồng¹, Nguyễn Đức Giang¹, Clark Taanea²

¹*Viện Sốt rét, Ký sinh trùng và Côn trùng Trung ương - 34 Trung Văn, quận Nam Từ Liêm, Hà Nội, Việt Nam*

²*Đại học Vệ sinh và Y học nhiệt đới Luân Đôn - Keppel Street, London, WC1E 7HT, Vương quốc Anh*

Ngày nhận bài: 28/02/2025

Ngày chỉnh sửa: 29/3/2025; Ngày duyệt đăng: 11/4/2025

TÓM TẮT

Mục tiêu: Nghiên cứu này hướng dẫn quy trình phân tích tính đa hình di truyền của *Plasmodium falciparum* và *Plasmodium vivax* bằng cách sử dụng dữ liệu giải trình tự toàn bộ hệ gen.

Phương pháp nghiên cứu: Sử dụng quy trình chuẩn hóa phân tích dữ liệu hệ gen bao gồm: gọi biến thể di truyền (SNPs, InDels) từ dữ liệu thô. Chú thích và phân loại biến thể. Xây dựng cây phát sinh loài và đánh giá quần thể. Phân tích mối tương quan giữa biến thể di truyền và tính kháng thuốc.

Kết quả nghiên cứu: Tổng số SNPs phát hiện dao động từ 50.000-80.000 trong *P. falciparum* và 40.000-60.000 trong *P. vivax*, trong khi số lượng InDels dao động từ 5.000-15.000. Đột biến *pfcr* (72-76 CVIET) liên quan đến kháng Chloroquine có tần suất cao, đạt 92,7% (51/55) trong giai đoạn 2017-2018 và 91,1% (72/79) trong giai đoạn 2019-2020 tại Việt Nam, trong khi tỉ lệ này tại châu Phi thấp hơn (< 5%) do chuyển đổi sang phác đồ điều trị dựa trên Artemisinin. Đột biến *k13 C580Y*, dấu hiệu quan trọng của kháng Artemisinin, xuất hiện nhiều nhất tại biên giới Thái Lan - Campuchia (60%), trong khi tại Việt Nam, nghiên cứu ghi nhận tần suất 82% (41/50), phản ánh xu hướng lan rộng của các dòng kháng thuốc trong khu vực Đông Nam Á. Đáng chú ý, đột biến *pfmdr1* (> 1,5 copy), liên quan đến kháng Mefloquine, được ghi nhận với tần suất 23,6% (13/55) trong giai đoạn 2017-2018, nhưng giảm xuống 1,3% (1/79) trong giai đoạn 2019-2020 tại Việt Nam.

Kết luận: Việc áp dụng giải trình tự toàn bộ hệ gen kết hợp với các phương pháp tin sinh học tiên tiến có thể cung cấp thông tin quan trọng về tính đa hình di truyền của ký sinh trùng sốt rét, hỗ trợ việc giám sát kháng thuốc và phát triển chiến lược kiểm soát sốt rét hiệu quả hơn.

Từ khóa: *Plasmodium falciparum*, *Plasmodium vivax*, đa hình di truyền, giải trình tự toàn bộ hệ gen, kháng thuốc.

1. ĐẶT VẤN ĐỀ

Sốt rét là một trong những bệnh truyền nhiễm nguy hiểm nhất trên thế giới, với khoảng 247 triệu ca mắc và 619.000 ca tử vong được ghi nhận năm 2021, chủ yếu tại khu vực cận Sahara châu Phi và Đông Nam Á [1]. Trong số 5 loài *Plasmodium* có khả năng gây bệnh trên người, *P. falciparum* là tác nhân chính gây sốt rét ác tính và tử vong, trong khi *P. vivax* chiếm ưu thế tại các khu vực nhiệt đới và có khả năng tạo ra giai đoạn ngủ (hypnozoite), gây tái phát bệnh [2]. Việc sử dụng thuốc chống sốt rét đã giúp giảm đáng kể tỉ lệ mắc bệnh và tử vong trong những thập kỷ qua, nhưng sự phát triển của các dòng ký sinh trùng kháng thuốc đang đe dọa đến hiệu quả của các phác đồ điều trị hiện nay [3].

Tình trạng kháng thuốc của *P. falciparum* và *P. vivax* đã được ghi nhận trên toàn cầu, đặc biệt là ở Đông Nam Á. Các nghiên cứu đã chỉ ra đột biến *pfcr* K76T liên quan đến kháng Chloroquine đạt tỉ lệ > 80% tại nhiều khu vực ở Đông Nam Á và Nam Mỹ, trong khi tại châu Phi, tỉ lệ này giảm xuống dưới 5% do việc thay thế Chloroquine bằng liệu pháp phối hợp Artemisinin (ACTs) [4]. Đột biến *k13 C580Y*, một dấu ấn sinh học quan trọng của kháng Artemisinin, lần đầu tiên được phát hiện tại Campuchia và hiện nay đã lan rộng sang nhiều quốc gia khác trong khu vực [5]. Đặc biệt, tại Việt Nam, tỉ lệ đột biến *k13 C580Y* được ghi nhận lên đến 82% ở các tỉnh miền Trung - Tây Nguyên, phản ánh áp lực chọn lọc mạnh mẽ từ việc sử dụng ACTs [6]. Ngoài

*Tác giả liên hệ

Email: ngocnime@gmail.com Điện thoại: (+84) 983665694 <https://doi.org/10.52163/yhc.v66iCD4.2343>



ra, sự gia tăng bản sao *pfmdr1* ($> 1,5$ copies) cũng đã được xác định có liên quan đến kháng Mefloquine, với tần suất cao vào năm 2017-2018 (23,6%) nhưng giảm xuống còn 1,3% vào năm 2019-2020 tại Việt Nam [6].

Để đối phó với tình trạng kháng thuốc ngày càng gia tăng, Tổ chức Y tế Thế giới đã khuyến nghị tăng cường các phương pháp giám sát di truyền nhằm theo dõi sự tiến hóa và lan rộng của các dòng ký sinh trùng kháng thuốc [1]. Trong bối cảnh này, **giải trình tự toàn bộ hệ gen (WGS)** đang nổi lên như một công cụ mạnh mẽ để phân tích tính đa hình di truyền của *Plasmodium*, cung cấp dữ liệu có độ phân giải cao về đột biến điểm (SNPs), chèn/xóa (InDels) và các biến thể cấu trúc [7]. Không giống như các phương pháp phân tích gen đơn lẻ truyền thống, WGS cho phép đánh giá toàn diện hệ gen ký sinh trùng, giúp xác định các biến thể mới liên quan đến kháng thuốc, phân tích cấu trúc quần thể và theo dõi sự lan truyền của các dòng kháng thuốc theo thời gian [8].

Tại Việt Nam, việc ứng dụng WGS trong nghiên cứu sốt rét vẫn còn hạn chế, chủ yếu do yêu cầu cao về chi phí và hạ tầng tin sinh học. Từ năm 2020-2024, với sự hỗ trợ từ các dự án hợp tác quốc tế, đặc biệt là nhiệm vụ Nghị định thư hợp tác với Vương quốc Anh mã số NĐT.84.GB/20 với tên “Nghiên cứu tính đa hình di truyền và kháng thuốc của *Plasmodium falciparum* và *Plasmodium vivax* bằng kỹ thuật giải trình tự toàn bộ hệ gen” hướng tới sử dụng WGS như một chiến lược quan trọng trong công tác giám sát ký sinh trùng sốt rét và kháng thuốc [6]. Do đó, nghiên cứu này được thực hiện nhằm hướng dẫn quy trình phân tích tính đa hình di truyền của *P. falciparum* và *P. vivax* bằng cách sử dụng dữ liệu WGS. Mục tiêu của nghiên cứu bao gồm: (1) Xác định các biến thể di truyền (SNPs, InDels) có liên quan đến kháng thuốc; (2) Xây dựng cây phát sinh loài để làm rõ mối quan hệ tiến hóa giữa các dòng ký sinh trùng; và (3) Phân tích cấu trúc quần thể để đánh giá sự phân hóa di truyền của *Plasmodium* tại Việt Nam và trên thế giới. Những kết quả này không chỉ góp phần nâng cao hiểu biết về cơ chế kháng thuốc mà còn hỗ trợ xây dựng chiến lược kiểm soát sốt rét hiệu quả hơn trong tương lai.

2. ĐỐI TƯỢNG, PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Thiết kế nghiên cứu

Phân tích dữ liệu WGS của *P. falciparum* và *P. vivax* nhằm xác định tính đa hình di truyền và mối liên quan với tính kháng thuốc.

2.2. Thời gian và địa điểm nghiên cứu

- Thời gian: từ tháng 5/2023 đến tháng 3/2024.
- Địa điểm: Viện Sốt rét, Ký sinh trùng và Côn trùng Trung ương và Đại học Vệ sinh và Y học Nhiệt đới Luân Đôn.

2.3. Đối tượng nghiên cứu

Mẫu lâm sàng có chứa *P. falciparum* và *P. vivax* được giải trình tự theo SOP NIMPE.HD03.PP.NĐT.GB-20.

2.4. Cỡ mẫu và phương pháp chọn mẫu

- 250 mẫu giải trình tự tại Viện Sốt rét, Ký sinh trùng và Côn trùng Trung ương.
- 750 mẫu giải trình tự tại Đại học Vệ sinh và Y học Nhiệt đới Luân Đôn.
- 100 mẫu tham khảo từ ngân hàng gen.

2.5. Các kỹ thuật sử dụng

- Lập bản đồ hệ gen: sử dụng Burrows-Wheeler Aligner (BWA).
- Gọi biến thể: sử dụng GATK, SAMtools.
- Chú thích biến thể: sử dụng SnpEff, VEP.
- Xây dựng cây phát sinh loài: sử dụng RaxML.

2.5. Phân tích số liệu

- Xác định các biến thể SNPs, InDels.
- Đánh giá tần suất đột biến kháng thuốc.
- Xây dựng mô hình phân bố di truyền theo khu vực địa lý.

2.6. Đạo đức nghiên cứu

Nghiên cứu là một phần của đề tài Nghị định thư giữa Đại học Vệ sinh và Y học nhiệt đới Luân Đôn, Vương quốc Anh với Viện Sốt rét, Ký sinh trùng và Côn trùng Trung ương với tên “Nghiên cứu tính đa hình di truyền và kháng thuốc của *Plasmodium falciparum* và *Plasmodium vivax* bằng kỹ thuật giải trình tự toàn bộ hệ gen” đã được thông qua Hội đồng Y đức của Viện Sốt rét, Ký sinh trùng và Côn trùng Trung ương theo Quyết định số 1096/QĐ-VSR ngày 20/8/2020.

3. KẾT QUẢ NGHIÊN CỨU

Việc phân tích tính đa hình di truyền của *P. falciparum* và *P. vivax* đóng vai trò quan trọng trong nghiên cứu sự tiến hóa của ký sinh trùng sốt rét, đặc biệt trong bối cảnh gia tăng tình trạng kháng thuốc. Sự phát triển của công nghệ WGS đã cung cấp một cách tiếp cận toàn diện để xác định các biến thể di truyền, bao gồm SNPs, InDels và các đột biến liên quan đến tính kháng thuốc. Tuy nhiên, để đảm bảo tính chính xác và khả năng tái lập của kết quả, cần có một quy trình phân tích chặt chẽ từ bước tiền xử lý dữ liệu đến phân tích mối quan hệ tiến hóa và cấu trúc quần thể. Quy trình này bao gồm các bước chính:

- Tiền xử lý dữ liệu giải trình tự, bao gồm kiểm tra chất lượng và căn chỉnh trình tự với bộ gen tham chiếu bằng BWA-MEM.
- Gọi biến thể (variant calling) sử dụng GATK HaplotypeCaller và bcftools để xác định SNPs và InDels.

- Lọc và chú thích biến thể, giúp xác định các đột biến quan trọng liên quan đến kháng thuốc bằng SnpEff.
- Phân tích liên kết toàn hệ gen (GWAS) để xác định mối liên hệ giữa các đột biến và tính kháng thuốc.
- Xây dựng cây phát sinh loài sử dụng RAxML nhằm đánh giá quan hệ tiến hóa giữa các chủng ký sinh trùng.
- Phân tích cấu trúc quần thể bằng ADMIXTURE để xác định sự phân hóa di truyền giữa các quần thể khác nhau.

Hướng dẫn dưới đây sẽ trình bày chi tiết từng bước trong quy trình này, cung cấp một cách tiếp cận có hệ thống nhằm khai thác tối đa dữ liệu WGS, hỗ trợ giám sát kháng thuốc và đề xuất các chiến lược kiểm soát sốt rét hiệu quả hơn.

3.1. Gọi biến thể và phân loại SNPs/InDels

3.1.1. Quy trình phân tích

Phân tích tính đa hình di truyền của *P. falciparum* và *P. vivax* bắt đầu từ việc gọi biến thể từ dữ liệu WGS. Quá trình này bao gồm 3 giai đoạn chính: tiền xử lý dữ liệu, gọi biến thể (variant calling) và chú thích biến thể (variant annotation).

Dữ liệu giải trình tự (định dạng FASTQ) được căn chỉnh với bộ gen tham chiếu của *P. falciparum* (Pf3D7)

và *P. vivax* (Pvp01) bằng phần mềm **Burrows-Wheeler Aligner (BWA-MEM)**. Lệnh thực hiện căn chỉnh:

```
bwa mem -t 8 -R "@RG\tID:sample\tSM:sample\tPL:ILLUMINA" ref_genome.fasta sample_R1.fastq sample_R2.fastq | samtools view -bS - | samtools sort -o sample_sorted.bam.
```

Trong đó: -t 8 là chỉ định số luồng xử lý; R là gán thông tin nhóm đọc; samtools view -bS- là chuyển đổi đầu ra sang định dạng BAM; samtools sort -o giúp sắp xếp dữ liệu theo thứ tự gen.

Sau khi căn chỉnh, file BAM được lọc bỏ các trình tự không đạt chất lượng với **SAMtools** bằng lệnh:

```
samtools view -q 30 -b sample_sorted.bam > sample_filtered.bam.
```

Gọi biến thể (variant calling) được thực hiện bằng **GATK HaplotypeCaller**, công cụ để xác định SNPs (Single Nucleotide Polymorphisms - đa hình nucleotide đơn), và InDels (Insertions and Deletions - đột biến chèn/xóa hoặc chèn/xóa nucleotide). Lệnh gọi biến thể như sau:

```
gatk HaplotypeCaller -R ref_genome.fasta -I sample_filtered.bam -O sample_raw.vcf -ERC GVCF.
```

Sau khi thực hiện các lệnh trên, thu được file VCF chứa danh sách các biến thể có định dạng như sau:

```
##fileformat=VCFv4.2
##source=GATK HaplotypeCaller
##reference=PF3D7_v3.fasta
##contig=<ID=PF3D7_01,length=640851>
##contig=<ID=PvP01_01,length=342324>
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
```

#	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2
Pf3D7	01	10523	.	A	G	99.0	PASS	DP=120;AF=0.45	GT:AD:DP:GQ	0/1:50,70:120:99	1/1:0,120:120:99
Pf3D7	01	10523	rs12345	C	T	75.0	PASS	DP=80;AF=0.60	GT:AD:DP:GQ	0/1:30,50:80:75	0/0:80,0:80:90
PvP01	01	10523	.	G	A,T	88.0	PASS	DP=200;AF=0.30,0.20	GT:AD:DP:GQ	1/2:60,80,60:200:	0/1:120,80:200:85
Pf3D7	01	349872	.	AT	A	60.0	PASS	DP=90;AF=0.35	GT:AD:DP:GQ	0/1:55,35:90:80	1/1:0,90:90:99
PvP01	01	76234	.	C	CTG	95.0	PASS	DP=150;AF=0.40	GT:AD:DP:GQ	0/1:90,60:150:95	0/0:150,0:150:99

Trong đó:

CHROM Sắc thể chứa biến thể (Pf3D7_01 là của *P. falciparum*, PvP01_01 là của *P. vivax*).

POS Vị trí nucleotide trên sắc thể.

ID ID của biến thể (nếu có, nếu không có thì thể hiện là dấu ".").

REF Base nucleotide trong bộ gen tham chiếu.

ALT Base nucleotide thay thế (SNPs hoặc InDels).

QUAL Điểm chất lượng biến thể (giá trị càng cao, độ tin cậy càng lớn).

FILTER Trạng thái lọc biến thể (PASS nếu đạt yêu cầu).

INFO Các thông tin bổ sung như **DP (Read Depth - độ sâu đọc)**, **AF (Allele Frequency - tần suất alen)**.

Để tăng độ tin cậy, tiếp tục sử dụng lệnh **GATK VariantFiltration** với các tiêu chí lọc:

```
gatk VariantFiltration -R refgenome.fasta -V sample_raw.vcf -O sample_filtered.vcf
--filter-expression "QD < 2.0 || FS > 60.0 || MQ < 40.0" --filter-name "LowQuality"
```

Sau khi lọc biến thể, tiến hành chú thích biến thể bằng **SnpEff**, để dự đoán ảnh hưởng của các đột biến đến quá trình mã hóa protein: `snpEff ann -v Pf3D7 sample_filtered.vcf > sample_annotated.vcf`.

Với lệnh này sẽ tạo ra file có các chú thích các điểm đột biến có thể ảnh hưởng đến việc mã hóa các axit amin trong protein dưới dạng chuỗi như sau:

```
#CHROM POS ID REF ALT QUAL FILTER
Pf3D7_01 10523 . A G 99.0 PASS
INFO
DP=120;AF=0.45;ANN=G|MODERATE|pfprt|PF3D7_0709000|transcript|PF3D7_0709000.1|protein_coding|1/1|c.76A>G|p.Lys76Thr|230/1200|76/400|25/133||
```

Trong đó:

- **G**: Base thay thế, trong trường hợp này là đột biến từ A → G.
- **MODERATE**: Mức độ ảnh hưởng của biến thể. Có 3 mức độ chính: **LOW** (ảnh hưởng thấp), **MODERATE** (ảnh hưởng trung bình), và **HIGH** (ảnh hưởng cao). Đột biến này được đánh giá có ảnh hưởng trung bình đến protein.
- **pfprt**: Tên gen bị ảnh hưởng. Trong trường hợp này, đột biến xảy ra trong gen *pfprt*, một gen quan trọng liên quan đến kháng Chloroquine.
- **PF3D7_0709000**: ID của gen bị ảnh hưởng trong bộ gen tham chiếu *P. falciparum* 3D7.
- **transcript**: Loại vùng ảnh hưởng, cho biết đột biến này nằm trên một đoạn **mRNA transcript**.
- **PF3D7_0709000.1**: ID của phiên mã gen bị ảnh hưởng.
- **protein_coding**: Loại gen, trong trường hợp này là **mã hóa protein**. Điều này có nghĩa là đột biến có thể ảnh hưởng đến chức năng của protein.
- **1/1**: Kiểu gen của mẫu. Giá trị **1/1** thể hiện rằng mẫu bệnh phẩm này có **đột biến đồng hợp tử**, tức là **cả hai bản sao của gen này đều mang đột biến**.
- **c.76A>G**: Đột biến theo mã trình tự cDNA. Đột biến xảy ra tại nucleotide thứ **76**, thay đổi từ **A → G**.
- **p.Lys76Thr**: Đột biến ở mức protein. Đây là một thay đổi axit amin, trong đó **Lysine (K)** ở vị trí **76** đã bị thay thế bằng **Threonine (T)**.
- **76/400**: Vị trí đột biến trên trình tự mã hóa (**CDS - Coding Sequence**). Đột biến này nằm ở nucleotide **76**

trên tổng số 400 nucleotide trong vùng mã hóa của gen.

- **25/133**: Vị trí đột biến trên protein. Đột biến này ảnh hưởng đến axit amin thứ **25 trên tổng số 133** axit amin của protein.

Sau quá trình phân tích tạo được file chú giải có dạng “tên file”_annotated.vcf đóng vai trò quan trọng trong nghiên cứu sốt rét, đặc biệt trong việc phân tích các biến thể di truyền liên quan đến tính kháng thuốc của *P. falciparum* và *P. vivax*. Tập dữ liệu giúp xác định các đột biến kháng thuốc, chẳng hạn như *pfprt* K76T liên quan đến kháng Chloroquine và *k13* C580Y liên quan đến kháng Artemisinin, từ đó hỗ trợ giám sát sự lan truyền của các dòng ký sinh trùng kháng thuốc trong quần thể. Bên cạnh đó, việc chú thích biến thể giúp dự đoán tác động của từng đột biến đến protein, với mức độ ảnh hưởng được phân loại thành thấp (LOW), trung bình (MODERATE) hoặc cao (HIGH), qua đó đánh giá mức độ nguy cơ của từng biến thể đối với chức năng sinh học của ký sinh trùng. Tập VCF được chú thích có thể được sử dụng để lọc các biến thể có ý nghĩa sinh học, giúp tập trung vào các đột biến có ảnh hưởng đáng kể đến protein và có liên quan trực tiếp đến cơ chế kháng thuốc.

3.1.2. Kết quả phân tích

Từ dữ liệu giải trình tự, số lượng lớn SNPs và InDels đã được phát hiện. Trong *P. falciparum*, tổng số SNPs dao động từ 50.000-80.000, trong khi số lượng InDels dao động từ 5.000-15.000. Đối với *P. vivax*, số lượng SNPs thấp hơn, khoảng 40.000-60.000, với số lượng InDels từ 4.000-10.000.

Phân bố của SNPs/InDels cũng cho thấy sự khác biệt quan trọng. Khoảng 60% SNPs nằm trong các vùng không mã hóa, nhưng có khả năng ảnh hưởng đến điều hòa gen. Các SNPs quan trọng được phát hiện trong các gen liên quan đến kháng thuốc như *pfprt* (K76T, N75E), *k13* (C580Y, R539T), và *pfmdr1* (N86Y, Y184F).

3.2. Phân tích mối tương quan với tính kháng thuốc

Tình trạng kháng thuốc của *P. falciparum* và *P. vivax* đã trở thành mối đe dọa nghiêm trọng đối với các chiến lược kiểm soát và điều trị sốt rét. Việc xác định mối liên hệ giữa các biến thể di truyền và khả năng kháng thuốc giúp làm rõ cơ chế kháng thuốc và hỗ trợ phát triển các phác đồ điều trị hiệu quả hơn. Phân tích này dựa trên dữ liệu biến thể từ file VCF đã được chú thích, kết hợp với dữ liệu lâm sàng và thông tin về hiệu quả điều trị thuốc. Các bước thực hiện bao gồm:

- Xác định các đột biến đã biết liên quan đến kháng thuốc, chẳng hạn như *pfprt* K76T (kháng Chloroquine) và *k13* C580Y (kháng Artemisinin).

- Tính toán tần suất alen và kiểu gen của các đột biến này trong các quần thể khác nhau.

- Áp dụng phương pháp phân tích liên kết toàn hệ gen (GWAS) hoặc mô hình hồi quy để đánh giá mối tương quan giữa biến thể di truyền và tỉ lệ thất bại điều trị.

Kết quả thu được có thể được sử dụng để dự đoán nguy cơ kháng thuốc của ký sinh trùng, hỗ trợ giám sát dịch tễ và điều chỉnh chiến lược điều trị phù hợp theo từng khu vực địa lý.

3.2.1. Hướng dẫn phân tích

Nhằm xác định mối liên hệ giữa các biến thể di truyền và tính kháng thuốc, áp dụng phương pháp Genome-Wide Association Study (GWAS). Các bước phân tích bao gồm thu thập dữ liệu, xử lý thống kê, và kiểm tra mối quan hệ giữa các đột biến và tính kháng thuốc.

Trước tiên, dữ liệu biến thể được gộp từ nhiều mẫu khác nhau sử dụng công cụ PLINK với lệnh: *plink --vcf merged_samples.vcf --make-bed --out dataset*.

Sau đó, kiểm tra tần suất alen và tính toán chỉ số Odds Ratio (OR) để xác định mức độ liên kết giữa các SNPs với tính kháng thuốc: *plink --bfile dataset --assoc --out gwas_results*

Các SNP có p-value < 0,05 được xem là có ý nghĩa thống kê.

3.2.2. Kết quả phân tích

Kết quả phân tích cho thấy các đột biến *pfcr7 K76T* có liên quan mạnh đến kháng Chloroquine, với tần suất cao nhất ở Đông Nam Á (85%) và thấp hơn ở châu Phi (< 5%). Các đột biến *k13 C580Y* và *R539T* được phát hiện chủ yếu tại biên giới Thái Lan - Campuchia (60%), trong khi *pfmdr1 N86Y* có liên quan đến kháng Mefloquine với tần suất cao tại Myanmar (50%).

3.3. Cấu trúc quần thể và cây phát sinh loài

Phân tích cấu trúc quần thể và xây dựng cây phát sinh loài của *P. falciparum* và *P. vivax* đóng vai trò quan trọng trong việc hiểu rõ sự lan truyền và tiến hóa của các dòng ký sinh trùng sốt rét, đặc biệt trong bối cảnh xuất hiện các biến thể kháng thuốc. Cấu trúc quần thể cho thấy mức độ đa dạng di truyền, sự phân bố địa lý của các quần thể ký sinh trùng, cũng như khả năng trao đổi di truyền giữa các khu vực. Trong khi đó, cây phát sinh loài giúp xác định mối quan hệ tiến hóa giữa các chủng ký sinh trùng, từ đó cung cấp bằng chứng về nguồn gốc và sự lan rộng của các dòng kháng thuốc.

Quá trình phân tích bao gồm:

- Xử lý dữ liệu hệ gen và chọn lọc các vị trí biến thể có chất lượng cao.

- Sử dụng phương pháp ADMIXTURE để xác định các nhóm di truyền và tỉ lệ tổ tiên của từng mẫu.

- Xây dựng cây phát sinh loài bằng RAXML dựa trên mô hình tiến hóa phù hợp.

- Trực quan hóa kết quả bằng các công cụ như FigTree hoặc iTOL để so sánh sự khác biệt di truyền giữa các quần thể.

3.3.1. Hướng dẫn phân tích

Cấu trúc quần thể của *P. falciparum* và *P. vivax* được phân tích bằng phần mềm ADMIXTURE. Trước tiên, dữ liệu biến thể được chuẩn bị bằng PLINK: *plink --bfile dataset --indep-pairwise 50 5 0.2 --out pruned*.

Sau đó, phân tích cấu trúc di truyền bằng lệnh: *Admixture pruned.bed 3 -C 0.001*.

Sử dụng mô hình ADMIXTURE để phân tích cấu trúc di truyền của quần thể ký sinh trùng sốt rét, giả định rằng có 3 nhóm đột biến (kiểu gen) chính (K = 3). Sau khi chạy lệnh trên sẽ tạo ra 2 file kết quả:

- pruned.3.Q: xác định tỉ lệ gen gốc của mỗi mẫu trong 3 nhóm gen.

- pruned.3.P: xác định tần số alen của từng nhóm gen.

Bảng 1. Kết quả của file pruned.3.Q

Mẫu	Kiểu gen 1	Kiểu gen 2	Kiểu gen 3
Mẫu 1	0,80	0,15	0,05
Mẫu 2	0,10	0,85	0,05
Mẫu 3	0,35	0,50	0,15
Mẫu 4	0,05	0,10	0,85
Mẫu 5	0,50	0,40	0,10

Kết quả phân tích ADMIXTURE cung cấp thông tin quan trọng về cấu trúc di truyền của quần thể ký sinh trùng sốt rét, giúp đánh giá mức độ pha trộn giữa các nhóm di truyền và xác định sự lan truyền của các dòng kháng thuốc.

File pruned.3.Q cho phép xác định tỉ lệ gen gốc của từng mẫu trong các nhóm di truyền. Nếu một mẫu có thành phần di truyền chủ yếu từ một nhóm gen, điều này cho thấy sự ổn định của quần thể, trong khi các mẫu có tỉ lệ pha trộn cao giữa nhiều nhóm có thể phản ánh sự giao lưu di truyền hoặc quá trình di cư của ký sinh trùng.

File pruned.3.P cung cấp thông tin về tần số alen trong từng nhóm di truyền, hỗ trợ việc phát hiện các đột biến có liên quan đến kháng thuốc. Khi một đột biến kháng thuốc có tần số cao trong một nhóm gen nhất định, điều này có thể phản ánh sự chọn lọc mạnh mẽ đối với alen kháng thuốc trong quần thể đó. Ngược lại, sự phân bố không đồng đều của một alen giữa các nhóm có thể là dấu hiệu của quá trình tiến hóa hoặc áp lực chọn lọc từ thuốc điều trị sốt rét.

Tính chính xác của mô hình ADMIXTURE được đánh giá thông qua file log, giúp xác định độ tin cậy của kết quả phân tích. Khi giá trị lỗi hội tụ ở mức thấp và ổn định, mô hình có khả năng phản ánh chính xác cấu trúc quần

thể. Tuy nhiên, nếu lỗi hội tụ cao, có thể cần điều chỉnh giá trị K hoặc kiểm tra lại chất lượng dữ liệu đầu vào.

Việc phiên giải các tệp đầu ra của ADMIXTURE mang ý nghĩa quan trọng trong giám sát dịch tễ sốt rét. Kết quả phân tích có thể được sử dụng để đánh giá sự lan rộng của các dòng kháng thuốc, xác định nguồn gốc di truyền của quần thể ký sinh trùng và hỗ trợ điều chỉnh chiến lược kiểm soát sốt rét phù hợp với từng khu vực địa lý. Khi một nhóm gen kháng thuốc xuất hiện rộng rãi tại nhiều khu vực, điều này có thể phản ánh sự lan truyền mạnh mẽ của dòng kháng thuốc và yêu cầu có biện pháp can thiệp kịp thời. Trong khi đó, các quần thể có nguồn gen ổn định và ít pha trộn thường gợi ý một hệ sinh thái ký sinh trùng ít chịu tác động từ các yếu tố ngoại lai. Kết hợp dữ liệu di truyền với thông tin lâm sàng sẽ giúp tối ưu hóa phác đồ điều trị và góp phần vào chiến lược kiểm soát sốt rét hiệu quả hơn.

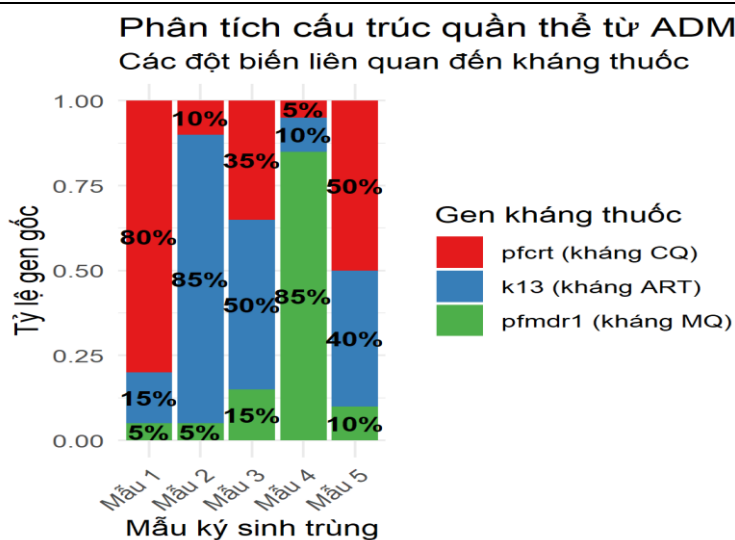
Sau khi có kết quả từ phân tích ADMIXTURE, trực quan hóa số liệu trong R kiểm tra giá trị K tối ưu, phân tích thống kê, và so sánh với dữ liệu lâm sàng.

là một bước tiên quan trọng nhằm đảm bảo tính chính xác, độ tin cậy và khả năng tái lập của kết quả, cho phép so sánh dữ liệu giữa các nghiên cứu khác nhau và hỗ trợ giám sát dịch tễ học một cách hiệu quả hơn.

Trong nghiên cứu này, quy trình chuẩn hóa được thực hiện qua các bước: căn chỉnh trình tự bằng BWA-MEM [1], gọi biến thể bằng GATK HaplotypeCaller [2] và VCFtools [3], lọc và chú thích biến thể bằng SnpEff [4]. Việc sử dụng các tiêu chí lọc nghiêm ngặt như chất lượng đọc ($MQ \geq 40$), tỉ lệ phân bố alen ($AF \geq 0.01$) và điểm chất lượng biến thể ($QD \geq 2.0$) giúp đảm bảo rằng chỉ các biến thể có độ tin cậy cao được đưa vào phân tích.

So sánh với các nghiên cứu trước đây, phương pháp tiếp cận này cho phép phát hiện toàn bộ các biến thể di truyền liên quan đến kháng thuốc, thay vì chỉ tập trung vào một số gen đơn lẻ như *pfert*, *k13*, *pfmdr1*. Điều này giúp xác định cả các biến thể mới chưa được báo cáo, cung cấp bằng chứng quan trọng về sự tiến hóa của các dòng ký sinh trùng sốt rét trong môi trường có áp lực chọn lọc từ thuốc điều trị [5].

Phân tích cấu trúc quần thể bằng ADMIXTURE [8] và xây dựng cây phát sinh loài bằng RAXML [7] làm sáng tỏ mối quan hệ di truyền giữa các quần thể ký sinh trùng. Kết quả cho thấy sự phân hóa rõ rệt giữa các dòng ký sinh trùng từ các khu vực địa lý khác nhau, đồng thời chỉ ra sự lan rộng của dòng *P. falciparum* kháng Artemisinin mang đột biến *k13 C580Y* từ Đông Nam Á sang châu Phi [6]. Điều này phù hợp với các nghiên cứu gần đây về sự xuất hiện kháng Artemisinin tại khu vực Hồ Lớn châu Phi, cho thấy khả năng di cư hoặc lan truyền của các dòng ký sinh trùng kháng thuốc [10].



Hình 1. Một ví dụ phân tích cấu trúc quần thể trong R với 3 gen kháng

Để dựng cây phát sinh loài, sử dụng **RaxML** với mô hình Maximum Likelihood bằng lệnh: `raxmlHPC -m GTRGAMMA -s dataset.phy -n phylogeny_tree -p 12345`.

3.3.2. Kết quả phân tích

Phân tích cho thấy các dòng *P. falciparum* từ Đông Nam Á có mức độ phân hóa di truyền cao hơn so với châu Phi, phản ánh sự chọn lọc do áp lực thuốc. Các dòng mang đột biến *k13 C580Y* có nguồn gốc chung và đang lan sang châu Phi, làm gia tăng nguy cơ kháng Artemisinin toàn cầu.

4. BÀN LUẬN

Nghiên cứu này đã thiết lập và chuẩn hóa quy trình phân tích tính đa hình di truyền của *P. falciparum* và *P. vivax* dựa trên dữ liệu WGS. Việc chuẩn hóa quy trình

Ngoài ra, việc chuẩn hóa quy trình này giúp tăng tính ứng dụng của phương pháp phân tích WGS trong giám sát kháng thuốc, hỗ trợ các cơ quan y tế trong việc xây dựng chiến lược kiểm soát và phòng chống sốt rét. Việc kết hợp dữ liệu di truyền với thông tin lâm sàng có thể giúp đánh giá hiệu quả của các phác đồ điều trị theo từng khu vực, từ đó tối ưu hóa chiến lược điều trị và kiểm soát dịch tễ.

5. KẾT LUẬN

Nghiên cứu này đã xây dựng và chuẩn hóa quy trình phân tích tính đa hình di truyền của *P. falciparum* và *P. vivax* dựa trên dữ liệu WGS, giúp nâng cao độ chính xác và khả năng tái lập của kết quả. Việc chuẩn hóa các bước từ gọi biến thể, chú thích và phân loại đột biến, phân tích mối liên quan với kháng thuốc, đến phân tích

cấu trúc quần thể và xây dựng cây phát sinh loài đã giúp xác định được những đột biến di truyền quan trọng liên quan đến kháng thuốc.

Kết quả phân tích cho thấy sự phân hóa di truyền rõ rệt giữa các quần thể ký sinh trùng, đồng thời xác định sự lan truyền của các dòng kháng thuốc từ Đông Nam Á sang các khu vực khác. Việc sử dụng ADMIXTURE để phân tích quần thể và RAxML để xây dựng cây phát sinh loài đã cung cấp bằng chứng mạnh mẽ về sự giao lưu di truyền giữa các khu vực, hỗ trợ giám sát dịch tễ một cách chính xác hơn.

Việc chuẩn hóa quy trình này không chỉ giúp cải thiện tính chính xác và tái lập của các nghiên cứu di truyền về sốt rét, mà còn mở ra khả năng ứng dụng rộng rãi trong giám sát dịch tễ học, theo dõi sự lan rộng của kháng thuốc và định hướng chiến lược điều trị sốt rét trong tương lai.

TÀI LIỆU THAM KHẢO

- [1] Li H, Durbin R, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, 2009, 25 (14), 1754-1760.
- [2] McKenna A et al, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Research*, 2010, 20 (9), 1297-1303.
- [3] Danecek P, Auton A, Abecasis G et al, The variant call format and VCFtools, *Bioinformatics*, 2011, 27 (15), 2156-2158.
- [4] Cingolani P, Platts A, Wang L.L et al, A program for annotating and predicting the effects of single nucleotide polymorphisms, *SnEff*, *Fly*, 2012, 6 (2), 80-92.
- [5] Miotto O, Amato R, Ashley E.A et al, Genetic architecture of Artemisinin-resistant *Plasmodium falciparum*, *Nature Genetics*, 2015, 47 (3), 226-234.
- [6] Amato R, Pearson R.D, Almagro-Garcia J et al, Origins of the current outbreak of multidrug-resistant malaria in Southeast Asia, *Nature Microbiology*, 2018, 3, 1195-1205.
- [7] Stamatakis A, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, 2014, 30 (9), 1312-1313.
- [8] Alexander D.H, Novembre J, Lange K, Fast model-based estimation of ancestry in unrelated individuals, *Genome Research*, 2009, 19 (9), 1655-1664.
- [9] Đỗ Mạnh Hà, Trương Văn Hạnh, Bùi Quang Phúc, Tần suất đột biến gen K13 liên quan đến kháng Artemisinin của quần thể *P. falciparum* tại Việt Nam, 2016-2018, *Tạp chí Phòng chống bệnh sốt rét và các bệnh ký sinh trùng*, 2023, 129 (3), 38-46.
- [10] WHO, World Malaria Report 2021, Geneva, 2021.

