RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta dissertação será disponibilizado somente a partir de 27/04/2021.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Instituto de Biociências de Botucatu - Unesp

ANÁLISE DE DADOS POR IMPUTAÇÃO DE SEQUENCIAMENTO DE BAIXA COBERTURA: SELEÇÃO DE MARCADORES E GENÉTICA POPULACIONAL

MARCUS VINICIUS NIZ ALVAREZ

BOTUCATU – SP

2020

UNIVERSIDADE ESTADUAL PAULISTA

"Júlio de Mesquita Filho"

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

ANÁLISE DE DADOS POR IMPUTAÇÃO DE SEQUENCIAMENTO DE BAIXA COBERTURA: SELEÇÃO DE MARCADORES E GENÉTICA POPULACIONAL

MARCUS VINICIUS NIZ ALVAREZ

ORIENTADOR: PAULO EDUARDO MARTINS RIBOLLA

Dissertação apresentada ao Instituto de Biociências, Câmpus de Botucatu, UNESP, para obtenção do título de Mestre no Programa de Pós- Graduação em Ciências Biológicas (Genética) Genética.

BOTUCATU – SP

2020

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Instituto de Biociências de Botucatu - Unesp

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Alvarez, Marcus Vinicius Niz.
Análise de dados por imputação de sequenciamento de baixa cobertura : seleção de marcadores e genética populacional / Marcus Vinicius Niz Alvarez. - Botucatu, 2020

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Paulo Eduardo Martins Ribolla
Capes: 20204000

1. Sequenciamento completo do genoma. 2. Bioinformática. 3. Genômica. 4. Mosquitos.

Palavras-chave: GWAS; baixa-cobertura; bioinformática; genômica; mosquito.

AGRADECIMENTOS

Aos meus pais, Raul e Cecília, pelo amor incondicional e por todo o apoio e incentivo em tudo na minha vida, se abdicando de tantas coisas pelo meu futuro e da minha irmã. À minha querida irmã, Danielle Amanda, por sempre confiar em mim e pelo seu amor eternamente especial. À minha família, pelo apoio e incentivo que sempre me deram durante toda a minha vida.

À Isabelle, minha namorada, que sempre está comigo nos momentos de alegria, mas também de dificuldade. Agradeço pelo carinho, companheirismo e dedicação. Seu amor me dá forças e coragem. Sou grato por ter conhecido uma pessoa tão maravilhosa como você.

Ao meu amigo e quase irmão Filipe, pela amizade de tantos anos, por todas as longas conversas, apoio e mesmo distante, sempre presente na minha vida.

Ao Prof. Paulo Ribolla, meu orientador e, sobretudo, um grande amigo, pela dedicação e confiança depositada na minha proposta de projeto. Agradeço por acreditar no meu potencial e por me manter sempre motivado durante todo o processo.

A todo o grupo Pangene, em especial Diego Alonso, pelas valiosas contribuições e Heitor Troca, pelas boas conversas e grande amizade além dos momentos de trabalho.

À Universidade Estadual Paulista (Unesp) Câmpus Botucatu, e todos os professores e que sempre proporcionaram ensino de alta qualidade. Agradeço também aos funcionários, em especial da Seção técnica de Pós-graduação, pela competência e colaboração na resolução de problemas.

À Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pelo apoio financeiro com a concessão de bolsa de mestrado (processo 2018/07406-6).

RESUMO

Introdução: O desenvolvimento de estratégias para redução no custo do sequenciamento de genoma completo (WGS) é importante para projetos que demandam por grandes quantidades de amostras. Uma estratégia de baixo custo é o sequenciamento de baixa cobertura aliado a técnicas de imputação para genotipagem eficiente e de confiabilidade adequada. A malária é uma das principais doenças transmitidas por artrópodes no mundo e o Brasil é considerado um país com alta incidência de malária, principalmente na região Amazônica, sendo principal vetor o mosquito *Anopheles darlingi*. **Objetivo:** O objetivo do presente estudo foi desenvolver estratégia para analisar dados de WGS de baixa cobertura de mosquitos *Anopheles darlingi* coletados no município de Mâncio Lima no Acre e verificar associação entre dados genéticos e dados de importância epidemiológica, tais como comportamento de picada, horário de atividade e distanciamento em escala microgeográfica. **Materiais e métodos:** Amostras de mosquitos *Anopheles darlingi* foram coletadas no município de Mâncio Lima - AC, entre 2016 e 2017. As bibliotecas foram preparadas com Nextera™ XT e sequenciadas no NextSeq500 da Illumina. Foi realizado genotipagem por sequenciamento e aplicado imputação. Estudos de associação ampla do genoma foram realizados com comportamento de picada e horário de atividade. Sinais de estratificação na população foram investigados por F_{ST} amplo no genoma e teste de permutação para significância. **Resultados:** Sinais fracos porém significativos para estratificação foram encontrados considerando distâncias de 2 a 3 km entre os grupos. Associações significativas foram observadas entre comportamento de picada e polimorfismos de nucleotídeo único (SNP), principalmente SNPs adjacentes ao gene *Cyp450*. Associações significativas foram observadas entre horário de atividade e SNPs adjacente aos genes *timeless-2* e *rdgC*. **Conclusões:** A utilização de dados de WGS de baixa cobertura aliado à imputação de dados é uma estratégia viável para redução do custo em projetos de sequenciamento genômico com grandes quantidades de amostras. Os resultados das análises de estratificação sustentam a hipótese de que a população de *Anopheles darlingi* está em processo de estratificação genética em escala microgeográfica no município de Mâncio Lima. Os resultados dos estudos de associação ampla genômica sugerem que SNPs significativos para comportamento de picada podem estar associados a genes de resistência de inseticidas e SNPs significativos para horário de atividade sugerem associação com genes relacionados a regulação do ciclo circadiano.

Palavras-chave: Genômica; Mosquito; GWAS; Citocromo P450; Ciclo circadiano; Estratificação;

ABSTRACT

***Introduction**: Strategy development to reduce the cost of whole genome sequencing (WGS) is important for projects that demand large quantities of samples. A low-cost strategy is low-coverage sequencing combined with imputation techniques for efficient genotyping and sufficient confiability. Malaria is one of the main diseases transmitted by arthropods in the world and Brazil is considered a country with a high incidence of malaria, especially in the Amazon region with the main vector being the Anopheles darlingi mosquito. **Objective**: The objective of the present study was to develop a strategy to analyze low-coverage WGS data from Anopheles darlingi mosquitoes collected in the municipality of Mâncio Lima in Acre State and verify associations between genetic data and data of epidemiological importance, such as biting behavior, time of activity and distance on a microgeographic scale. **Materials and methods**: Samples of Anopheles darlingi mosquitoes were collected in the municipality of Mâncio Lima - AC, between 2016 and 2017. The libraries were prepared with Nextera ™ XT and sequenced on Illumina's NextSeq500. Genotyping by sequencing was performed and imputation was applied. Genome wide association studies were performed with biting behavior and time of activity. Population stratification signals were investigated by genome-wide F_{ST} and permutation test applied for significance. **Results**: Weak but significant stratification signals were identified considering distances of 2 to 3 km between the groups. Significant associations were observed between biting behavior and single nucleotide polymorphisms (SNP), mainly in SNP adjacent to the Cyp450 gene. Significant associations were observed between time of activity and SNP, including SNP adjacent to the timeless-2 and rdgC genes. **Conclusions**: The use of low coverage WGS data and data imputation is a viable strategy for cost reduction in genomic sequencing projects with large amounts of samples. The results of the stratification analyzes support the hypothesis that the population of Anopheles darlingi is in genetic stratification process on a microgeographic scale in the municipality of Mâncio Lima. The results of genome wide association studies suggest that significant SNPs for biting behavior may be associated with insecticide resistance genes and significant SNPs for time of activity suggest an association with genes related to circadian cycle regulation.*

***Keywords:** Genomics; Mosquito; GWAS; Cytochrome P450; Circadian Rhythm; Stratification;*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD: Profundidade de sequenciamento do alelo (em inglês, *Allele Depth*).

BLAST: Ferramenta básica de localização de alinhamento local (em inglês, *Basic Local Alignment Search Tool*)

DNA: Ácido desoxirribonucleico, também abreviado como ADN (em inglês, *deoxyribonucleic acid*)

DP: Profundidade de sequenciamento (em inglês, *Depth*).

GBS: Genotipagem por sequenciamento (em inglês, *Genotyping by sequencing*)

GP: Probabilidade posterior do genótipo (em inglês, *Genotype Probability*).

GQ: Qualidade do genótipo (em inglês, *Genotype Quality*).

GT: Genótipo.

GWAS: Estudo de associação ampla de genoma (em inglês, *Genome Wide Association Studies*)

HWD: Desequilíbrio de Hardy-Weinberg.

HWE: Equilíbrio de Hardy-Weinberg.

INDEL: Polimorfismo de Inserção ou deleções de nucleotídeos (em inglês, *Insertions-deletions polymorphism*)

MAF: Frequência do alelo menor (em inglês, *Minor allele frequency*).

MD: Dados ausentes (em inglês, *Missing Data*).

NMD: Dados não-ausentes (em inglês, *Non-Missing Data*).

PL: Lista de probabilidades dos genótipos em escala *Phred*, arredondada para o número inteiro mais próximo (em inglês, *phred-scaled genotype likelihoods rounded to the closest integer*).

SNP: Polimorfismo de nucleotídeo único (em inglês, *Single Nucleotide Polymorphism*).

VCF: Formato de chamada de variantes (em inglês, *Variant Call Format*).

WGS: Sequenciamento de genoma completo (em inglês, *Whole genome sequencing*).

LISTA DE SÍMBOLOS

bp: Pares de base.

kb (equivalente a kbp): quilo pares de base (ou 1.000 pares de base).

Km: Quilômetro (ou 1.000 metros).

GB: *GigaBytes* de informação (8×10^9 *bits* de informação).

Min: Mínimo.

Q1: Primeiro quartil (ou 25º percentil).

Q2: Segundo quartil, coincide com o valor da mediana (ou 50º percentil).

Q3: Terceiro quartil (ou 75º percentil).

Max: Máximo.

p_{VALOR} : Nível descritivo do teste estatístico.

p_{FDR} : Nível descritivo do teste estatístico, corrigido para múltiplas comparações pelo método taxa de falso positivo (em inglês, *False Discovery Rate*).

R^2: Coeficiente de determinação.

r^2: Desequilíbrio de ligação.

F_{ST}: Índice de fixação da população.

f : Coeficiente de endogamia.

e.p.: Erro padrão da média.

$\widehat{\pi}$: Parâmetro de diversidade nucleotídica.

d.p.: Desvio padrão.

1. INTRODUÇÃO

1.1. Sequenciamento de nova geração e estratégias envolvidas.

O acelerado desenvolvimento de tecnologias envolvidas no sequenciamento de genoma completo (WGS) tem resultado em reduções notáveis no custo da técnica. No entanto, projetos que demandam por sequenciamento de grandes quantidades de amostras continuam apresentando custo elevado, muitas vezes, impraticáveis. Por isso, estratégias são adotadas para que seja possível explorar informações genômicas em larga escala que auxiliam no entendimento da estrutura genética de populações de forma viável.

Uma estratégia de baixo custo é o uso de sequenciamento de genoma completo de baixa cobertura para genotipagem por sequenciamento (GBS), aliado a técnica de imputação que confere informações genômicas suficientes para seleção marcadores com menor custo e de forma acurada (GORJANC, G., *et al*., 2017).

A acurácia na detecção de variantes é reduzida em sequenciamento genômico com baixa profundidade de sequenciamento e tendem a apresentar taxa de falso-positivo elevada, mas isso é atenuado quando a informação entre as amostras é combinada, proporcionando bom poder de identificação de variantes comuns (SIMS, D., *et al*, 2014).

A inferência de genótipos por imputação, tanto para painéis de genotipagem quanto para genotipagem por sequenciamento, demonstra ser uma técnica com bons resultados acurados, possibilitando o uso de sequenciamento de genoma completo de baixa cobertura para descoberta de variantes com uma redução dramática no custo quando comparada com o WGS padrão (PASANIUC, B., *et al*., 2012; RUSTAGI, N., *et al*., 2017)

Li e colaboradores (2011) demonstraram que variantes raras em amostras de WGS de baixa cobertura apresentam maior dificuldade de serem detectadas por conta da dificuldade de distinguir alelos raros genuínos de erros de sequenciamento. A quantidade de variantes identificadas é superior quando a proporção de polimorfismos na população que segregou entre os indivíduos sequenciados é maior.

Considerando que diferentes abordagens podem ser aplicadas em análises de sequenciamento de baixa cobertura, a sensibilidade de cada método deve ser cuidadosamente ajustada, pois a redução na cobertura inevitavelmente amplifica a probabilidade de detecção de falsos positivos.

1.2. Marcadores moleculares e Genotipagem por Sequenciamento

Marcadores moleculares são polimorfismos genéticos entre indivíduos que podem ser utilizados amplamente em estudos com seres vivos. A genotipagem é o processo de identificação de polimorfismos genéticos, os quais podem ser utilizados como marcadores moleculares. Uma das aplicações mais comuns do sequenciamento de nova geração (NGS) é a detecção de variação genômica entre indivíduos de uma população.

Localizar variações no genoma e correlacionar com características biológicas tem sido um dos principais focos de muitos estudos de NGS. Atualmente existem diferentes algoritmos, programas e métodos para genotipagem. São comumente utilizados em painéis de genótipos informações de marcadores moleculares do tipo polimorfismos de nucleotídeo único (SNP) e inserções e deleções de nucleotídeos (INDEL). No entanto, o desafio primário é diferenciar verdadeiros polimorfismos de erros causados pelo sequenciamento e alinhamento de sequências.

De forma geral, o desempenho da chamada de variantes pode ser influenciado por diversos fatores, principalmente: qualidade da chamada de base, qualidade do alinhamento, sequenciamento *single-end* ou *pair-end*, comprimento dos fragmentos e cobertura do sequenciamento (WANG, X., 2016). Além disso, há divergências até mesmo entre programas de chamada de variantes. Yu e Sun (2013) demonstraram que o uso de diferentes algoritmos em dados WGS de baixa cobertura apresentaram pouca concordância entre si e, por isso, se faz necessário a aplicação e comparação de mais de um algoritmo e uso de métricas para controle de qualidade da chamada e cobertura dos dados.

1.3. Imputação de Genótipos

A técnica de imputação de genótipos é uma técnica que apresenta ótimo custo-benefício para aumentar o poder em análises genômicas, tais como estudos de associação ampla de genoma, seleção genômica, entre outros. Existem diversos programas e algoritmos de imputação disponíveis atualmente.

De maneira geral, a lógica por trás da técnica de imputação se resume em recuperar informação de genótipos faltantes baseado na estimativa de haplótipos da população, etapa conhecida como faseamento (em inglês, *Phasing*).

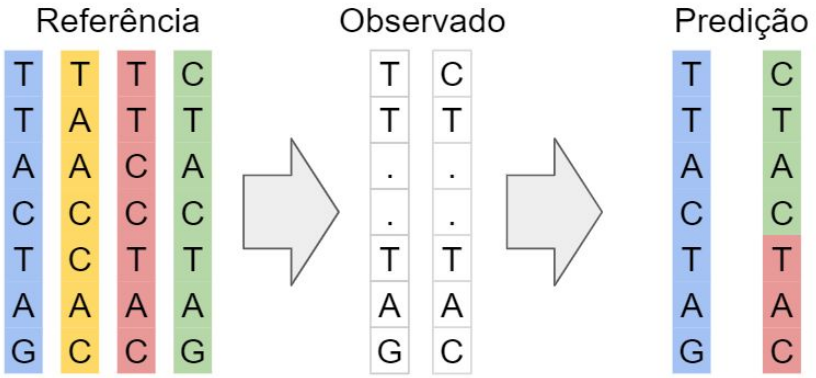

Figura 1 - Representação de imputação de genótipos. Representação esquemática do processo de imputação de genótipos baseado em blocos de SNP altamente correlacionados, ou seja, forte desequilíbrio de ligação. Fonte: Imagem elaborada pelo autor utilizando Google Docs.

Um dos métodos mais utilizados de imputação é o método baseado em desequilíbrio de ligação. Programas de imputação, como por exemplo o programa BEAGLE (BROWNINGS. R., BROWNING B. L., 2016), utilizam estimativas de desequilíbrio de ligação para resolver o faseamento dos haplótipos e, por fim, predizer os genótipos faltantes baseados nos haplótipos disponíveis do conjunto amostral ou até mesmo painéis de referências disponíveis em banco de dados online.

Diversos fatores podem influenciar na acurácia tanto do faseamento quanto da imputação. São fatores: densidade de genótipos, total de indivíduos genotipados, parentesco entre indivíduos, número e distribuição de marcadores, frequência alélica e até mesmo oscilações no desequilíbrio de ligação local (GONDRO, C; VAN DER WERF, J; HAYES, B, 2013).

1.4. Estratégia para dados de WGS de baixa cobertura

Um dos principais desafios em trabalhos com dados de sequenciamento de baixa cobertura é estabelecer um equilíbrio entre confiabilidade da genotipagem e descarte de dados. As práticas de controle de qualidade comuns para dados de WGS muitas vezes não se aplicam adequadamente em dados de baixa cobertura.

Um dos parâmetros de controle de qualidade mais importantes aplicados frequentemente é a profundidade de sequenciamento DP (em inglês, *Depth*). Um valor de DP

mínimo comumente utilizado em sequenciamentos convencionais é de 10 vezes. Suponha que um valor suficientemente conservador frequentemente utilizado para DP seja cinco, ou seja, o número mínimo de vezes que uma base em uma determinada posição na referência precisa ter sido sequenciada e alinhada para chamada de genótipo seja cinco vezes.

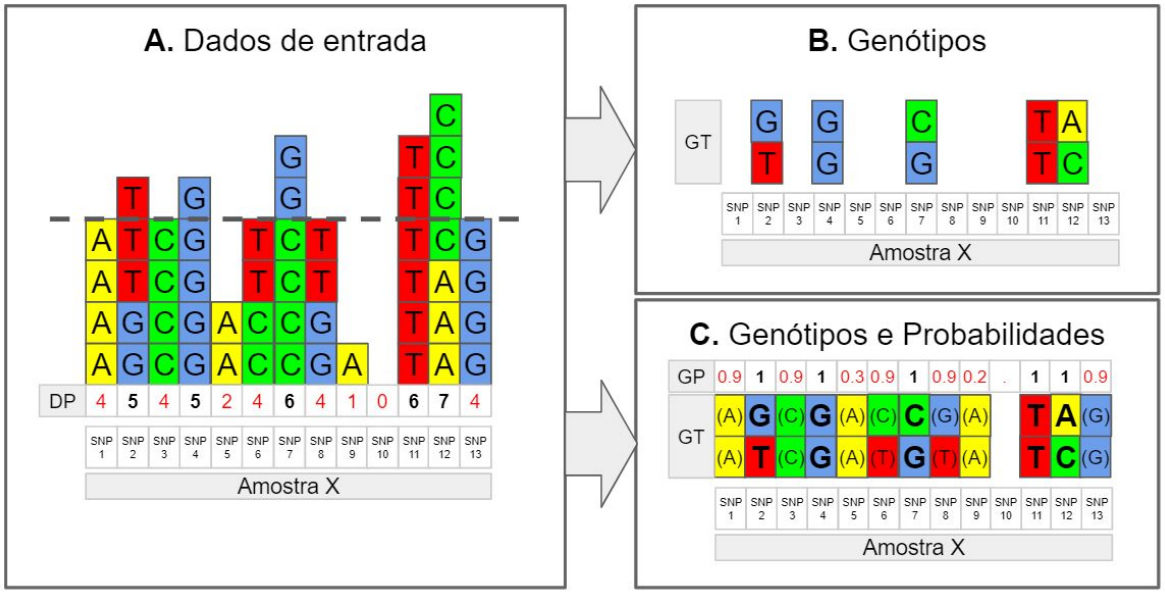

Figura 2 - Representação esquemática do controle de qualidade de painel de variante. A. Painel de variantes hipotético, sem controle de qualidade aplicado. B. Painel de variantes hipotético após o controle de qualidade com DP mínimo igual a cinco. C. Painel de variantes hipotético após o controle de qualidade com DP mínimo igual a cinco e genótipos convertidos em probabilidade de genótipos. Linha tracejada em A representa o limiar de DP igual a cinco. DP = profundidade de sequenciamento. GT = genótipo. GP = probabilidade do genótipo. Fonte: Imagem elaborada pelo autor utilizando Google Docs.

Uma situação que pode ocorrer no controle de qualidade de sequenciamento de baixa cobertura é o descarte massivo de dados. Por exemplo, suponha a seguinte situação: Em uma determinada amostra, apenas 40% do genoma apresentou DP superior a cinco vezes. No entanto, a mesma amostra apresentou menos de 10% do genoma DP igual a zero. Ou seja, estabelecendo DP mínimo de cinco vezes, cerca de 50% dos dados seriam completamente descartados, mesmo se as sequências apresentassem altíssima qualidade de sequenciamento. O cenário está representado na figura 2, seguindo o fluxo de A para B. Do total de 13 possíveis variantes encontradas, apenas 5 foram mantidas.

Uma forma de contornar o descarte massivo de dados é a reutilização dos dados descartados como probabilidade *a priori* do genótipo em modelos de imputação de genótipos. O programa de imputação BEAGLE 4.1, por exemplo, é um programa que permite utilizar

informação dos genótipos, bem como a probabilidade do genótipo, para imputação de genótipos faltantes. A estratégia está representada na figura 2, seguindo o fluxo de A para C. A estratégia pode ser descrita da seguinte forma:

- ❏ Se uma determinada variante em uma determinada amostra apresentou DP maior que o mínimo estabelecido, o genótipo é chamado. A probabilidade do genótipo é definida como 100% (probabilidade real normalmente muito próxima de 100%).
- ❏ Se uma determinada variante em uma determinada amostra apresentou DP menor que o mínimo estabelecido e maior que zero, o genótipo é omitido e uma lista de probabilidades para cada genótipo provável é calculado.
- ❏ Se uma determinada variante em uma determinada amostra apresentou DP igual a zero, é considerado dado ausente (MD).
- ❏ Variantes que apresentem excesso de MD são removidas do painel de variantes. O parâmetro utilizado é a taxa de dados faltantes NMD, em inglês, *Non-missing data*. NMD é calculado como a proporção de amostras com DP igual a zero em uma determinada variante.
- ❏ O painel é submetido à imputação de genótipos utilizando as probabilidades dos genótipos. Exemplo: (Imputação pelo modelo GTGL disponível no programa BEAGLE, versão 4.1)

Após a imputação ser executada, apenas genótipos de confiabilidade aceitável devem ser mantidos para as etapas da análise. Portanto, como forma de controle de qualidade, os dados pós-imputação podem ser filtrados por probabilidade posterior do genótipo (GP) maior que, por exemplo, 95%. Além disso, outros parâmetros frequentemente utilizados podem ser aplicados, por exemplo: frequência do alelo menor mínima, equilíbrio de Hardy-Weinberg e excesso de dados ausentes após imputação.

1.5. Modelo de estudo: Malária, *Anopheles darlingi* e região amazônica

A malária é considerada a enfermidade transmitida por artrópode mais impactante em países em desenvolvimento. De acordo com *World Malaria Reports* (2019), a estimativa de casos no mundo é de 228 milhões de casos de malária e 405 mil mortes em 2018, aproximadamente 93% concentradas na África.

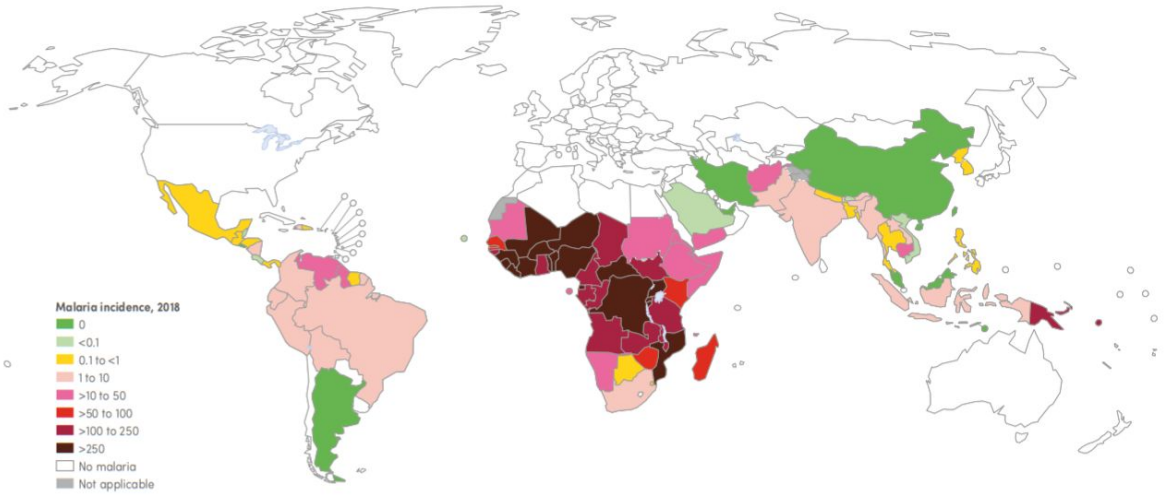

Figura 3 - Taxa global de incidência de casos de malária de 2018. Estimativa da taxa de incidência de malária no ano de 2018 por *World Malaria Reports* de 2019. Fonte: WHO, 2019.

Além da África, esta doença afeta outras populações pobres em áreas tropicais e subtropicais, devido às condições ambientais serem favoráveis para o desenvolvimento do agente causador da doença bem como do seu transmissor (SNOW, R. W. *et al*., 2005).

O Brasil é um país com alta incidência de malária, foram 194 mil casos registrados em 2018 segundo o Boletim Epidemiológico da Secretaria de Vigilância em Saúde de 2019 – Ministério da Saúde (2019), sendo a maioria dos casos registrados concentrados na Amazônia brasileira, cerca de 47% dos casos no estado do Amazonas e 22% no Acre. Segundo *World Malaria Reports* (2019), há estimativa de 217 mil casos em 2018 no Brasil.

Essa doença é caracterizada por desencadear acessos periódicos de febres intensas que debilitam profundamente o doente. O ciclo de transmissão é composto por protozoários (Reino Protozoa) do gênero *Plasmodium* spp. que são transmitidos ao homem através da picada de mosquitos do gênero *Anopheles* spp. Há seis espécies dentro do gênero *Plasmodium* que podem causar a malária humana: *Plasmodium vivax, P. falciparum, P. malariae, P. ovale curtisi, P. ovale wallikeri e P. knowlesi* (SU, X.Z., 2010).

As diferenças entre esses agentes patogênicos são quanto à infecção, sintomas e terapias utilizadas para o tratamento. Do ponto de vista epidemiológico, a principal diferença é a mortalidade: casos de *P. falciparum* não tratados possuem elevados índices de óbitos, sendo a espécie mais letal (WORLD HEALTH ORGANIZATION, 2019). Existem quase 500

espécies de anofelinos, sendo que apenas 70 são vetores do parasita e destes, cerca de 20 são importantes transmissores da malária ao homem (SERVICE, M. W., 2008).

Anopheles darlingi é o principal vetor de malária no Brasil, altamente suscetível aos plasmódios humanos e, capaz de transmitir a doença dentro e fora das moradias, mesmo quando sua densidade é baixa. Os criadouros deste anofelino são caracteristicamente representados por coleções de águas límpidas, com certa profundidade, sombreadas, dotadas de vegetação pobres em sais e matéria orgânica (FORATTINI, O. P., 2002). Além disso, é notavelmente antropofílico, pois na medida em que o ambiente natural se transforma em antrópico, ou desmatado, a população local de *Anopheles darlingi* tenderá a coabitar com o homem, invadindo-lhe os domicílios, traduzindo a capacidade de adaptação do mosquito ali presente e potencializando seu papel de vetor (ROZENDAAL, J. A., 1990).

Na Amazônia, é o vetor anofelino que melhor e mais rapidamente se beneficia das alterações que o homem produz no ambiente silvestre (CONSOLI, R., LOURENÇO-DE-OLIVEIRA, R., 1994). O controle do vetor é realizado com borrifação de inseticida e, uso de mosquiteiros ao entardecer e durante a noite, horário de pico da atividade hematofágica do vetor (BAIA-DA-SILVA, D. C., *et al*., 2019). O uso de inseticida possui periculosidade à saúde de habitantes do local e de funcionários que o manejam, além de produzir a seleção de indivíduos resistentes (VEZENEGHO, S. B., *et al*., 2009). Sendo assim, o estudo do comportamento e biologia do vetor, sua dispersão e interação com humanos é de grande importância para entomologia médica.

O estado do Acre está inserido no grupo que compõe a Amazônia Legal. O seu histórico de urbanização é semelhante a outras regiões do bioma Amazônico, iniciado entre o final do século XIX e o começo do século XX, com o ciclo da borracha, forte extrativismo econômico que atraiu intensamente mão de obra. Em 1977, o Projeto de Assentamento Dirigido de Pedro Peixoto (PAD Peixoto), do Governo Federal, direcionou a migração para a região causando a segunda grande colonização do Acre (SOUZA, A. *et al*., 2017). Uma grande porção da região oeste desse estado, que engloba os municípios de Acrelândia, Plácido de Castro, Senador Guiomar e a capital Rio Branco foi loteada e distribuída durante o PAD. O objetivo era instalar pequenos agricultores em lotes com menos de 100 hectares por família, em meio à mata nativa da região.

O resultado foi, a continuidade da exploração, desde os próprios recursos da mata, até o extrativismo mineral e a agro-pecuária, causando gradativas e constantes alterações do espaço

físico. Esses lotes adentram a floresta formando vias paralelas, perpendiculares a uma estrada principal, conhecidas popularmente como Ramais. Os Ramais se diferenciam entre outros fatores, principalmente pelo grau de desflorestação, número de residências e de habitantes. Esses fatores são de grande importância para se compreender a epidemiologia da malária, pois alteram a composição de anofelinos e consequentemente, a transmissão dos protozoários causadores da malária (WALSH, J. F., *et al*., 1993; TAIPE-LAGOS, D. A. C. C. B., 1994). No ano de 2007, foram registrados 9.410 casos de malária, sendo que 8.595 desse total ocorreram nestes assentamentos rurais. A região é alvo de diversos projetos científicos e epidemiológicos referentes à malária, com espécies de *Plasmodium* (BASTOS, M. S. *et al*., 2007; SILVA-NUNES, M., FERREIRA, M. U., 2007) e de *Anopheles* (MARRELLI, M. T. *et al*., 1998; CAMPOS, M. *et al*., 2017).

Os espécimes de *Anopheles darlingi* coletados no Brasil e outros países da América do Sul apresentam heterogeneidade, tanto genética quanto de comportamento. Mediante análise por RFLP do DNA mitocondrial (mtDNA) foi possível demonstrar isolamento por distância (CONN, J. E. *et al*., 1999). A análise de sequências da região ITS2 dos agrupamentos de regiões ribossomais apresentou aproximadamente 5% de divergência quando populações da Região Sudeste foram comparadas com populações do Norte do Brasil (MALAFRONTE, R. S., *et al*., 1999). Voorham (2002) mostrou que populações de mosquitos coletadas no Amapá apresentaram diferenças quanto ao horário para a hematofagia.

Esta heterogeneidade é de grande importância epidemiológica, pois pode refletir diferentes capacidades vetoriais nas populações de *Anopheles darlingi* (WHITE, G. B., 1982). Recentes estudos demonstram que *Anopheles darlingi* deve representar um complexo de espécies, sendo que os mosquitos presentes na Amazônia representam uma linhagem deste complexo (EMERSON, K. J. *et al*. 2015). No entanto, estudos em escala microgeográfica com marcadores espalhados pelo genoma de *Anopheles darlingi* mostram diferenças genéticas nesta escala (CAMPOS, M. *et al*., 2017), o que poderia representar diferenças fenotípicas importantes para a epidemiologia desta doença.

6. CONCLUSÕES

A otimização e imputação do painel de variantes de um grupo de amostras sequenciadas em baixa cobertura mostrou ser uma estratégia viável para analisar dados de WGS de baixa cobertura, tanto para estudos populacionais com número amostral consideravelmente representativo, bem como para estudos de associação ampla de genoma. No entanto, é importante salientar que os procedimentos e parâmetros utilizados na otimização dos dados do presente estudo são específicos para o conjunto de dados do mesmo, sendo importante novos estudos para dimensionamento e refinamento da estratégia.

O aprimoramento da estratégia desenvolvido no presente estudo apresenta potencial como alternativa para futuros projetos que dependam de menor custo para sequenciamento em larga-escala.

Os resultados das análises de estratificação significativo, o agrupamento observado nas figuras do PCA e DAPC e o fato do modelo estratificado para análise de GWAS ter sido bem sucedido sustentam a hipótese de que a população de *Anopheles darlingi* está em processo de estratificação genética em escala microgeográfica no município de Mâncio Lima.

Os genes adjacentes aos SNP dos estudos de associação são evidências que comportamentos de importância epidemiológica podem ser influenciados por fatores genéticos.

A relação entre comportamento de picada e o gene *CYP450* que está associado à resistência a inseticidas sugere que a aplicação de inseticida como forma de controle do vetor pode estar causando adaptação comportamental do mosquito.

Os genes adjacentes aos SNP estatisticamente associados ao horário de atividade sugerem que esse comportamento está relacionado principalmente com genes associados ao mecanismo de regulação do ciclo circadiano dos mosquitos.

REFERÊNCIAS BIBLIOGRÁFICAS

ADEWOYE, Adeolu Badi. Genetic architecture and molecular mechanisms underlying light entrainment of the Drosophila circadian clock. **Tese de Doutorado**. University of Leicester, Inglaterra. 252 p. 2011.

ANDREWS, Simon et al. FastQC: a quality control tool for high throughput sequence data. **Babraham Bioinformatics**. 2010.

BAIA-DA-SILVA, Djane Clarys et al. Current vector control challenges in the fight against malaria in Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 52, 2019.

BALABANIDOU, Vasileia. et al. Cytochrome P450 associated with insecticide resistance catalyzes cuticular hydrocarbon production in Anopheles gambiae. **Proceedings of the National Academy of Sciences**, v. 113, n. 33, p. 9268-9273, 2016.

BASTOS, Melissa S. et al. Antigenic polymorphism and naturally acquired antibodies to Plasmodium vivax merozoite surface protein 1 in rural Amazonians. **Clin. Vaccine Immunol.**, v. 14, n. 10, p. 1249-1259, 2007.

BENJAMINI, Yoav; HOCHBERG, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal statistical society: series B (Methodological)**, v. 57, n. 1, p. 289-300, 1995.

BENNA, Clara et al. Drosophila timeless2 is required for chromosome stability and circadian photoreception. **Current Biology**, v. 20, n. 4, p. 346-352, 2010.

BENSON, Dennis A. et al. GenBank. **Nucleic acids research**, v. 42, n. D1, p. D32-D37, 2013.

BERRY, Donagh P.; KEARNEY, J. F. Imputation of genotypes from low-to high-density genotyping platforms and implications for genomic selection. **Animal**, v. 5, n. 8, p. 1162-1169, 2011.

BROWNING, Brian L.; BROWNING, Sharon R. Genotype imputation with millions of reference samples. **The American Journal of Human Genetics**, v. 98, n. 1, p. 116-126, 2016.

CAMACHO, Christiam et al. BLAST+: architecture and applications. **BMC bioinformatics**, v. 10, n. 1, p. 421, 2009.

CAMPOS, Melina et al. Microgeographical structure in the major Neotropical malaria vector Anopheles darlingi using microsatellites and SNP markers. **Parasites & vectors**, v. 10, n. 1, p. 76, 2017.

CONN, J. E. et al. Molecular population genetics of the primary neotropical malaria vector Anopheles darlingi using mtDNA. **Journal of the American Mosquito Control Association**, v. 15, n. 4, p. 468-474, 1999.

CONSOLI, Rotraut AGB; LOURENÇO-DE-OLIVEIRA, Ricardo. **Principais mosquitos de importância sanitária no Brasil**. SciELO-Editora FIOCRUZ, 1994.

CORNEL, Anthony J. et al. Anopheles darlingi polytene chromosomes: revised maps including newly described inversions and evidence for population structure in Manaus. **Memórias do Instituto Oswaldo Cruz**, v. 111, n. 5, p. 335-346, 2016.

DANECEK, Petr et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156-2158, 2011.

DONNELLY, Martin J.; ISAACS, Alison T.; WEETMAN, David. Identification, validation, and application of molecular diagnostics for insecticide resistance in malaria vectors. **Trends in Parasitology**, v. 32, n. 3, p. 197-206, 2016.

EMERSON, Kevin J. et al. Brazilian Anopheles darlingi Root (Diptera: Culicidae) clusters by major biogeographical region. **PLoS One**, v. 10, n. 7, p. e0130773, 2015.

EXCOFFIER, Laurent; LISCHER, Heidi EL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Molecular ecology resources**, v. 10, n. 3, p. 564-567, 2010.

FORATTINI, O. P. **Culicidologia médica**. Editora USP. São Paulo. 2002.

GAO, Yue et al. Transcriptome-based identification and characterization of genes commonly responding to five different insecticides in the diamondback moth, Plutella xylostella. **Pesticide biochemistry and physiology**, v. 144, p. 1-9, 2018.

GÉLIN, Pauline et al. The fine-scale genetic structure of the malaria vectors Anopheles funestus and Anopheles gambiae (Diptera: Culicidae) in the north-eastern part of Tanzania. **International journal of tropical insect science**, v. 36, n. 4, p. 161-170, 2016.

GILLESPIE, John H. **Population genetics: a concise guide**. JHU Press, 2004.

GIRALDO-CALDERÓN, Gloria I. et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. **Nucleic acids research**, v. 43, n. D1, p. D707-D713, 2015.

GONDRO, Cedric; VAN DER WERF, Julius; HAYES, Ben (Ed.). **Genome-wide association studies and genomic prediction**. Totowa, NJ, USA: Humana Press, 2013.

GOOGLE. GOOGLE MAPS. Version 7.3. 2018. Mâncio Lima - AC. Disponível em: <https://earth.google.com/web/@-7.45219694,-73.38693502,207.08502036a,196105.8713598 1d,35y,0h,0t,0r/data=ClEaTxJHCiUweDkxOTgyZmZmOTIwOTc5M2Y6MHg0YTJmZjIxYz cxM2ZjNDExGXcq4J7nXx7AIfEEDNxLO1LAKgxNw6JuY2lvIExpbWEYAiABKAI>. Acesso em: 6 de novembro de 2018.

GORJANC, Gregor et al. Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. **Crop Science**, v. 57, n. 3, p. 1404-1420, 2017.

HARRIS, Caroline et al. Low linkage disequilibrium in wild Anopheles gambiae sl. populations. **BMC genetics**, v. 11, n. 1, p. 81, 2010.

HONNEN, Ann-Christin; JOHNSTON, Paul R.; MONAGHAN, Michael T. Sex-specific gene expression in the mosquito Culex pipiens f. molestus in response to artificial light at night. **BMC genomics**, v. 17, n. 1, p. 22, 2016.

IBRAHIM, Sulaiman S. et al. The cytochrome P450 CYP6P4 is responsible for the high pyrethroid resistance in knockdown resistance-free Anopheles arabiensis. **Insect biochemistry and molecular biology**, v. 68, p. 23-32, 2016.

JEFFRIES, Daniel L. et al. Comparing RAD seq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, Carassius carassius, L. **Molecular ecology**, v. 25, n. 13, p. 2997-3018, 2016.

JOMBART, Thibaut; AHMED, Ismaïl. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. **Bioinformatics**, v. 27, n. 21, p. 3070-3071, 2011.

LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature methods**, v. 9, n. 4, p. 357, 2012.

LI, Heng; DURBIN, Richard. Fast and accurate short read alignment with Burrows–Wheeler transform. **bioinformatics**, v. 25, n. 14, p. 1754-1760, 2009.

LI, Heng. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, v. 27, n. 21, p. 2987-2993, 2011.

LI, Yun et al. Low-coverage sequencing: implications for design of complex trait association studies. **Genome research**, v. 21, n. 6, p. 940-951, 2011.

LI, Yun et al. Low-coverage sequencing: implications for design of complex trait association studies. **Genome research**, v. 21, n. 6, p. 940-951, 2011.

MALAFRONTE, Rosely dos Santos; MARRELLI, Mauro Toledo; MARINOTTI, Osvaldo. Analysis of ITS2 DNA sequences from Brazilian Anopheles darlingi (Diptera: Culicidae). **Journal of medical entomology**, v. 36, n. 5, p. 631-634, 1999.

MANLEY, Leigh J.; MA, Duanduan; LEVINE, Stuart S. Monitoring error rates in Illumina sequencing. **Journal of biomolecular techniques**, v. 27, n. 4, p. 125, 2016.

MANTEL, Nathan; HAENSZEL, William. Statistical aspects of the analysis of data from retrospective studies of disease. **Journal of the national cancer institute**, v. 22, n. 4, p. 719-748, 1959.

MARINOTTI, Osvaldo et al. The genome of Anopheles darlingi, the main neotropical malaria vector. **Nucleic acids research**, v. 41, n. 15, p. 7387-7400, 2013.

MARRELLI, Mauro Toledo et al. Correlation between positive serology for Plasmodium vivax-like/Plasmodium simiovale malaria parasites in the human and anopheline populations in the State of Acre, Brazil. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 92, n. 2, p. 149-151, 1998.

MCKENNA, Aaron et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome research**, v. 20, n. 9, p. 1297-1303, 2010.

MINISTÉRIO DA SAÚDE. Secretaria de Vigilância em Saúde no Brasil 2003|2009 da criação da Secretaria de Vigilância em Saúde aos dias atuais. **Boletim Epidemiológico.** 156 p. 2019.

NOWLING, Ronald J.; EMRICH, Scott J. Detecting chromosomal inversions from dense snps by combining pca and association tests. In: **Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics**. p. 270-276. 2018

O'RAWE, Jason et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. **Genome medicine**, v. 5, n. 3, p. 28, 2013.

OGUETA, Maite; HARDIE, Roger C.; STANEWSKY, Ralf. Non-canonical phototransduction mediates synchronization of the Drosophila melanogaster circadian clock and retinal light responses. **Current Biology**, v. 28, n. 11, p. 1725-1735. e3, 2018.

O'LOUGHLIN, Samantha M. et al. Genomic signatures of population decline in the malaria mosquito Anopheles gambiae. **Malaria journal**, v. 15, n. 1, p. 182, 2016.

PASANIUC, Bogdan et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. **Nature genetics**, v. 44, n. 6, p. 631, 2012.

PURCELL, Shaun et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American journal of human genetics**, v. 81, n. 3, p. 559-575, 2007.

QUINLAN, Aaron R.; HALL, Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841-842, 2010.

RAFAEL, Míriam S. et al. Salivary polytene chromosome map of Anopheles darlingi, the main vector of neotropical malaria. **The American journal of tropical medicine and hygiene**, v. 83, n. 2, p. 241-249, 2010.

ROSHYARA, Nab Raj et al. Impact of pre-imputation SNP-filtering on genotype imputation results. **BMC genetics**, v. 15, n. 1, p. 88, 2014.

ROZENDAAL, J. A. Observations on the distribution of anophelines in Suriname with particular reference to the malaria vector Anopheles darlingi. **Memórias do Instituto Oswaldo Cruz**, v. 85, n. 2, p. 221-234, 1990.

RUSTAGI, Navin et al. Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. **BMC genomics**, v. 18, n. 1, p. 396, 2017.

SCOTT, Jeffrey G. Cytochromes P450 and insecticide resistance. **Insect biochemistry and molecular biology**, v. 29, n. 9, p. 757-777, 1999.

SERVICE, M. W. **Medical entomology for students.** Cambridge University Press, 2008.

SILVA-NUNES, Mônica da; FERREIRA, Marcelo U. Clinical spectrum of uncomplicated malaria in semi-immune Amazonians: beyond the" symptomatic" vs" asymptomatic" dichotomy. **Memórias do Instituto Oswaldo Cruz**, v. 102, n. 3, p. 341-348, 2007.

SIMS, David et al. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, v. 15, n. 2, p. 121, 2014.

SNOW, Robert W. et al. The global distribution of clinical episodes of Plasmodium falciparum malaria. **Nature**, v. 434, n. 7030, p. 214, 2005.

SOUZA, Anderson et al. Situação epidemiológica da malária no Brasil. **Seminários de Biomedicina do Univag**, v. 1, 2017.

SU, Xin-zhuan. Human malaria parasites: are we ready for a new species?. **The Journal of infectious diseases**, v. 201, n. 10, p. 1453-1454, 2010.

TAIPE-LAGOS, DA COSTA CB. Caracterização epidemiológica da malária no Projeto de colonização agrícola Pedro Peixoto Gomide, Estado do Acre, Brasil. **Dissertação de Mestrado.** USP. São Paulo. 1994.

TAJIMA, F. Measurement of DNA polymorphism. Mechanisms of Molecular Evolution, **Introduction to Molecular Paleopopulation Biology**, p. 37-59, 1993.

TEAM, R. Core. R development core team. **RA Lang Environ Stat Comput**, v. 55, p. 275-286, 2013.

TEAM, RStudio et al. RStudio: integrated development for R. **RStudio, Inc., Boston, MA URL http://www. rstudio. com**, v. 42, p. 14, 2015.

TELFER, Emily J. et al. Parentage reconstruction in Eucalyptus nitens using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. **PloS one**, v. 10, n. 7, 2015.

VEZENEGHO, S. B. et al. Malaria vector composition and insecticide susceptibility status in Guinea Conakry, West Africa. **Medical and veterinary entomology**, v. 23, n. 4, p. 326-334, 2009.

VOOLSTRA, Olaf et al. Functional characterization of the three Drosophila retinal degeneration C (RDGC) protein phosphatase isoforms. **PloS one**, v. 13, n. 9, 2018.

VOORHAM, Jaco. Intra-population plasticity of Anopheles darlingi's (Diptera, Culicidae) biting activity patterns in the state of Amapá, Brazil. **Revista de Saúde Pública**, v. 36, p. 75-80, 2002.

WALSH, J. F.; MOLYNEUX, D. H.; BIRLEY, M. H. Deforestation: effects on vector-borne disease. **Parasitology**, v. 106, n. S1, p. S55-S75, 1993.

WANG, Xinkun. **Next-generation sequencing data analysis**. CRC Press, 2016.

WEIR, Bruce S.; COCKERHAM, C. Clark. Estimating F‐statistics for the analysis of population structure. **evolution**, v. 38, n. 6, p. 1358-1370, 1984.

WHITE, G. B. Malaria vector ecology and genetics. **British Medical Bulletin**, v. 38, n. 2, p. 207-212, 1982.

WILKE, Andre Barretto Bruno; WILK-DA-SILVA, Ramon; MARRELLI, Mauro Toledo. Microgeographic population structuring of Aedes aegypti (Diptera: Culicidae). **PloS one**, v. 12, n. 9, 2017.

WORLD HEALTH ORGANIZATION. World Malaria Report 2019 (World Health Organization, Geneva), **World Health Organization**, v. 1, n. 1, p. 210, 2019.

XIONG, Bo; BELLEN, Hugo J. Rhodopsin homeostasis and retinal degeneration: lessons from the fly. **Trends in neurosciences**, v. 36, n. 11, p. 652-660, 2013.

YU, Xiaoqing; SUN, Shuying. Comparing a few SNP calling algorithms using low-coverage sequencing data. **BMC bioinformatics**, v. 14, n. 1, p. 274, 2013.

ZHENG, Gang *et al*. Analysis of genetic association studies. **Springer Science & Business Media**, 2012.