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*).

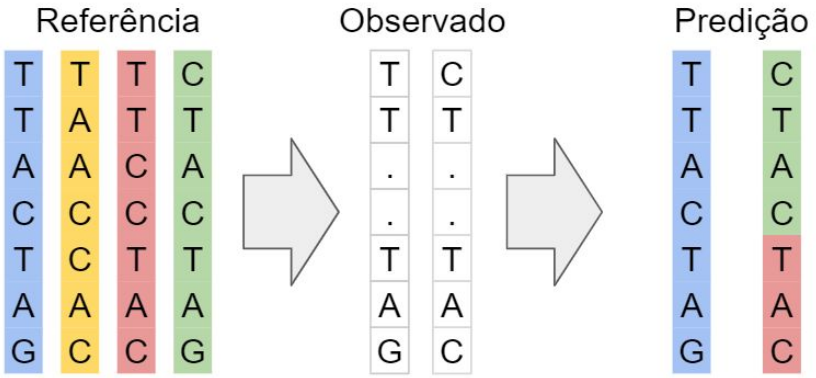

Figura 1 - Representação de imputação de genótipos. Representação esquemática do processo de imputação de genótipos baseado em blocos de SNP altamente correlacionados, ou seja, forte desequilíbrio de ligação. Fonte: Imagem elaborada pelo autor utilizando Google Docs.

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.

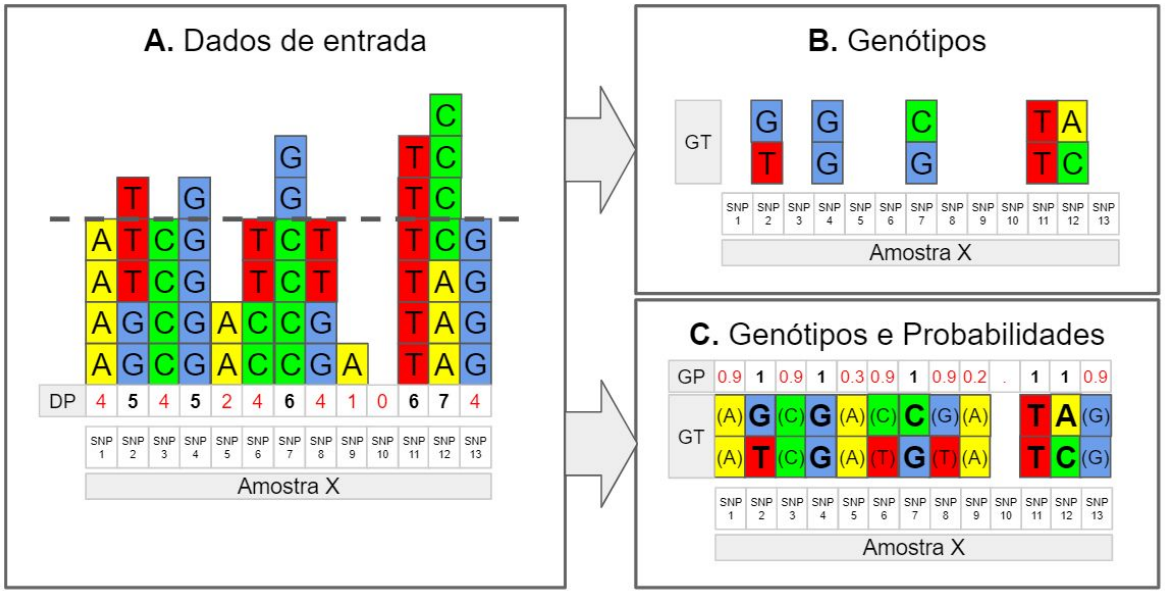

Figura 2 - Representação esquemática do controle de qualidade de painel de variante. A. Painel de variantes hipotético, sem controle de qualidade aplicado. B. Painel de variantes hipotético após o controle de qualidade com DP mínimo igual a cinco. C. Painel de variantes hipotético após o controle de 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.

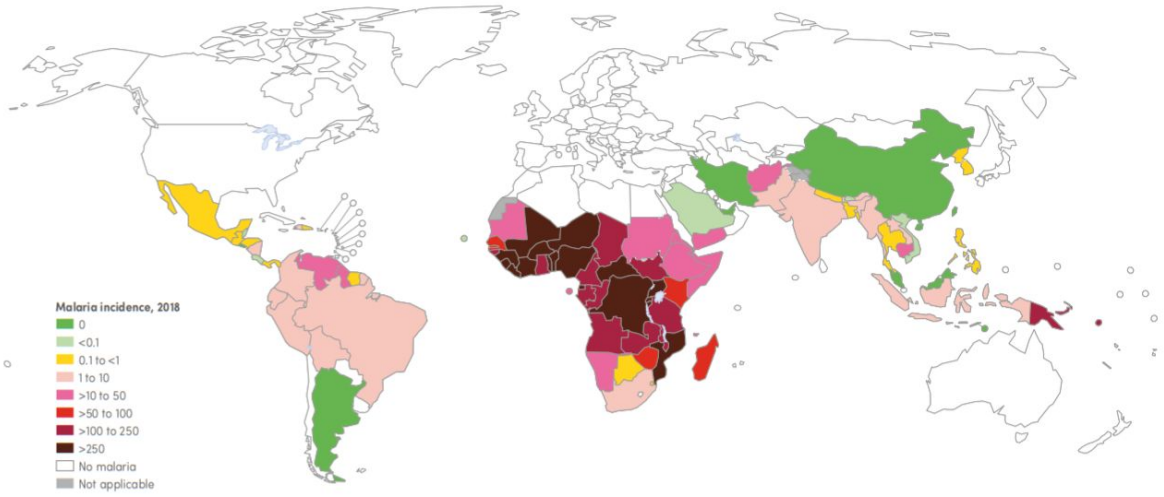

Figura 3 - Taxa global de incidência de casos de malária de 2018. Estimativa da taxa de incidência de malária no ano de 2018 por *World Malaria Reports* de 2019. Fonte: WHO, 2019.

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.

2. OBJETIVOS

2.1. Objetivo Geral

O objetivo geral do presente estudo foi desenvolver uma estratégia para analisar dados de sequenciamento WGS de baixa cobertura, com o propósito de verificar interações entre dados genéticos e epidemiológicos da população de Anofelinos coletados no município de Mâncio Lima - AC.

2.2. Objetivos Específicos

- ❑ Estabelecer estratégia para manipulação de dados de sequenciamento de baixa cobertura.
- ❑ Verificar viabilidade do uso de imputação de genótipos em dados de sequenciamento de baixa cobertura.
- ❑ Avaliar sinais de estratificação e agrupamento na população.
- ❑ Avaliar parâmetros genéticos populacionais das amostras coletadas nas diferentes localidades.
- ❑ Verificar associação entre dados genéticos e comportamento de picada.
- ❑ Verificar associação entre dados genéticos e horário de atividade.

3. MATERIAL E MÉTODOS

3.1. Coleta de amostras

As amostras de larvas e adultos foram coletadas em torno de três casas no município de Mâncio Lima (Acre) em quatro períodos de coleta: dezembro de 2016, fevereiro de 2017, maio de 2017 e setembro de 2017. Outras amostras de larvas foram coletadas na cidade de Iquitos, província do Peru localizada na região de Loreto, no ano de 2016.

Os anofelinos adultos foram coletados por quatro voluntários através do método de *Human Landing Catch* (HLC), em coletas de 12 horas, iniciando às 18h e terminando às 6h do dia seguinte, durante dois dias para cada ponto de coleta, sendo dois voluntários dentro e dois fora das casas. As três casas consistem em três diferentes perfis considerando o desmatamento: meio urbano (URB) em 7°37'12.9"S 72°53'06.7"W, meio peri-urbano (PURB) em 7°38'02.1"S 72°52'26.5"W e meio rural (RUR) em 7°39'05.3"S 72°53'20.9"W. Foram coletadas também informações sobre comportamento de picada, sendo classificadas como intradomiciliar e peridomiciliar. A imagem de satélite da região estudada está representada na Figura 4. As distâncias lineares aproximadas entre os pontos de coleta são: 3,39 Km de RUR até PURB, 1,96 Km de PURB até URB e 2,51 Km de URB até RUR. A Distância linear aproximada entre a região estudada de Mâncio Lima e Loreto no Peru é de 398 Km.

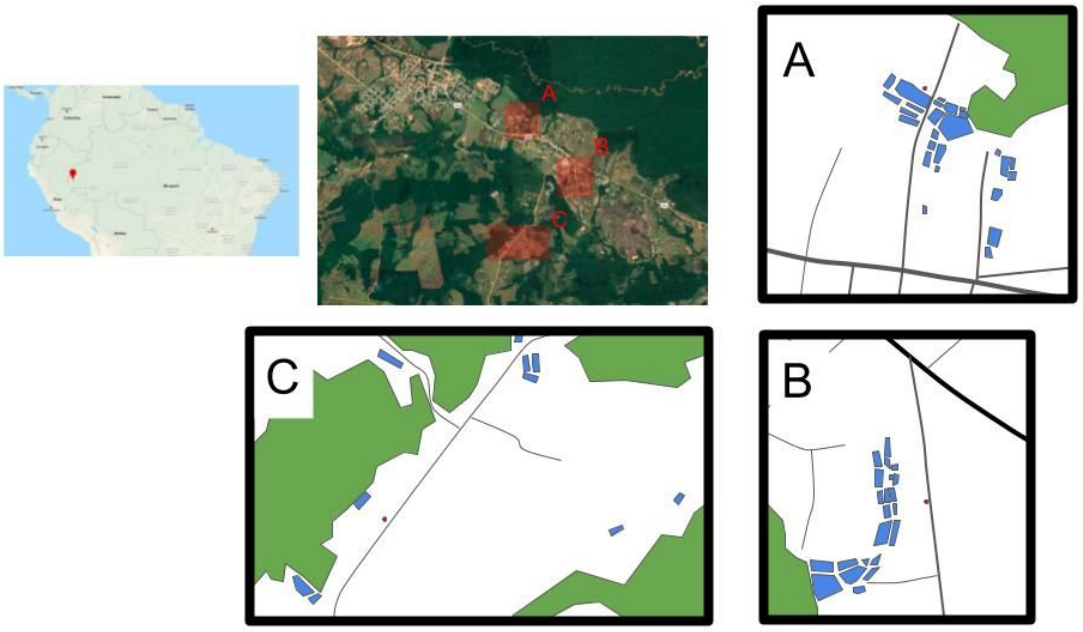

Figura 4 - Mapa esquemático dos pontos de coleta de *Anopheles darlingi* no no município de Mâncio Lima no estado do Acre. São os pontos de coleta: Urbano (URB) em A, Peri-urbano (PURB) em B e Rural (RUR) em C. As imagens dos cantos superior central e superior esquerdo foram tiradas do serviço gratuito de imagens de satélite e mapas do *Google Maps* - AC. Fonte: GOOGLE MAPS, 2018.

3.2. Preparação das amostras e sequenciamento

Os mosquitos foram separados em duas seções, uma contendo cabeça e tórax, e outra abdômen. A separação foi feita usando o bisturi, o abdômen foi armazenado, e o complexo cabeça/tórax foi extraído. Cada mosquito foi extraído individualmente, usando o Kit Glass Fiber Plate DNA Extraction, da CCDB (Canadian Centre for DNA Barcoding), realizada de acordo com as recomendações do fabricante. A quantificação de DNA foi feita por quantificação de fluorometria do Kit QuBit dsDNA HS Assay, da Thermo Fisher Scientific, de acordo com as recomendações do fabricante.

A preparação das bibliotecas foi feita segundo as recomendações da Nextera™ DNA Library Prep Reference Guide, utilizando um quinto do volume total de DNA recomendado. Inicialmente foi feito a tagmentação do DNA, usando a Tagment DNA Enzyme, transposase que adiciona fragmentos curtos de DNA conhecidos nas extremidades, com adaptador para futura ligação do index. Houve a purificação do DNA tagmentado, para que a transposase não continuasse a se ligar ao DNA. Em sequência, houve purificação do DNA tagmentado e adição de adaptadores index por um programa de PCR de 5 ciclos. As amostras foram reunidas de cinco em cinco para purificação usando AMPure XP beads. Os pools de DNA foram agrupados até um total de 60 amostras e submetidos a sequenciamento individual na plataforma NextSeq500 (Illumina) em programa *single-read* de 151 ciclos.

O programa FASTQC (ANDREWS, S., 2010) foi utilizado para análise de qualidade do sequenciamento. Para isso, 100.000 sequências foram amostradas aleatoriamente dos arquivos de sequenciamento de cada amostra e concatenados em um arquivo para avaliação do programa. O subconjunto representa cerca de 10% do volume total de dados, sendo 55,7 GB de informação, aproximadamente, considerando todos os arquivos de sequenciamento do estudo. Foram considerados principais critérios para controle de qualidade: qualidade por base, qualidade média por sequência, conteúdo nucleotídico por sequência e conteúdo guanina-citosina por sequência.

A cobertura e profundidade de sequenciamento do genoma foi estimado pelo programa BedTools (QUINLAN, A. R., HALL, I. M., 2010), considerando o genoma de referência de *Anopheles darlingi*, versão *AdarC3*, disponível no banco de dados VectorBase (GIRALDO-CALDERÓN, G.L., *et al*., 2015).

3.3. Comparação dos protocolos de genotipagem

Dois programas alinhadores de fragmentos curtos (em inglês, *short-reads*) amplamente utilizados foram combinados com dois programas de chamada de variantes. As quatro combinações foram realizadas com objetivo de comparar o desempenho das diferentes combinações de protocolos de genotipagem por sequenciamento quando utilizado dados de sequenciamento de baixa cobertura. Os programas de alinhamento de sequências Burrows-Wheeler Aligner 0.7.17 (LI, H., DURBIN, R., 2009) e Bowtie2 (LANGMEAD, B., SALZBERG, S. L., 2012) foram combinados com os programas de chamada de variantes SAMtools 1.8 (LI, H., 2011) e GATK 4 (MCKENNA, A., *et al*., 2010). Foram comparados entre si os painéis de variantes resultantes da chamada de variantes de cada combinação. Todos os programas foram executados com os parâmetros padrões descritos nos respectivos manuais de usuário.

O genoma de *Anopheles darlingi* foi utilizado como referência nos programas de alinhamento. A versão da montagem do genoma (em inglês, *genome assembly*) utilizado foi GCA_000211455.3, disponível em NCBI (BENSON, D. A., *et al*., 2013).

O critério de comparação foi o número de variantes total chamadas por cada combinação de programas, seguindo os critérios de controle de qualidade: qualidade de genótipo em *Phred Scale* (GQ) $\geq$ 20 e profundidade de sequenciamento (DP) $\geq$ 5. A combinação que apresentou melhor desempenho foi escolhida para realização da genotipagem final para ser utilizada nas etapas posteriores do presente estudo.

Foram utilizados dois computadores com configurações iguais, tanto na questão de componentes eletrônicos (em inglês, *hardware*), como versões de programas e sistema operacional. Os computadores foram equipadas com processadores Intel® Core™ i7-7700 com velocidade de *clock* em 3,6 GHz e 64 GB de memória RAM DDR4 2133 MHz. O sistema operacional utilizado foi Ubuntu 16.04 LTS 64 bits e os programas utilizados foram instalados na mesma versão.

3.4. Genotipagem por Sequenciamento (GBS)

Os dados do sequenciamento das amostras foram alinhados com o genoma de referência de *Anopheles darlingi*, versão *AdarC3*, disponível no banco de dados VectorBase (GIRALDO-CALDERÓN, G.L., *et al*., 2015). O alinhamento foi realizado com o programa Burrows-Wheeler Aligner (BWA) e a chamada de variantes pelo programa SAMtools. O processo de genotipagem foi realizado e e formatado da seguinte forma:

- *GT*: Genótipo (em inglês, *Genotype*). Código que representa o genótipo da amostra para respectiva variante.
- *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*). Os valores da lista de probabilidades podem ser calculados da seguinte forma: $PL_{ij} = P_{ij} - max(P_j)$, em que, o valor de PL do i-ésimo genótipo para a j-ésima variante é igual ao escore em escala *Phred* do i-ésimo genótipo da j-ésima variante menos o maior valor de escore em escala *Phred* do dos genótipos observados para a j-ésima variante.
- *AD*: Profundidade de sequenciamento do alelo (em inglês, *Allele Depth*). Número de vezes que sequências alinharam para o respectivo alelo.
- *DP*: Profundidade de sequenciamento (em inglês, *Depth*). Número de vezes que sequências alinharam para cada alelo chamado no respectivo genótipo. Pode ser calculada como a somatória dos valores de AD.
- *GQ*: Qualidade do genótipo (em inglês, *Genotype Quality*). Probabilidade de erro de genotipagem, na escala *Phred Scale*. A qualidade do genótipo pode ser calculada na escala *Phred Scaled* de acordo com a seguinte fórmula: $P(err) = 10^{\frac{-P}{10}}$, dado que, quanto maior o valor do escore *Phred* representado por *P*, menor a probabilidade de erro de genotipagem P(*err*).

O painel de variantes foi formatado como VCF (em inglês, *Variant Call Format*) na versão 4.2 (DANECEK, P., *et al*., 2011) pelo programa BCFtools do pacote de programas SAMtools.

3.5. Identificação Taxonômica

A sequência nucleotídica do gene citocromo oxidase subunidade I (KP193458.1) foi utilizada como referência para identificação das espécies das amostras coletadas. As sequências do sequenciamento das amostras foram alinhadas com a referência pelo programa BWA. As sequências consenso foram geradas pelo programa BCFtools do pacote de programas SAMtools.

A identificação taxonômica das amostras foi realizada com a ferramenta BLASTn (CAMACHO C., *et al.*, 2009). Foram consideradas identificações significativas quando *e* valor (em inglês, *e value*) ≤ 1e-50, identidade máxima (em inglês, *max identity*) ≥ 70.

Foram descartadas das próximas etapas amostras que não identificadas significativamente como *Anopheles darlingi*. Amostras que apresentaram menos de 25% do genoma com profundidade de sequenciamento DP ≥ 1 foram descartadas.

3.6. Teste de parâmetros para filtragem pré-imputação

Foram selecionadas 10 amostras com maior profundidade de sequenciamento e cobertura do genoma para verificar o desempenho da imputação de dados em diferentes configurações no controle de qualidade. As versões de baixa cobertura das respectivas amostras foram geradas por subamostragem aleatória das sequências presentes no arquivo FASTq do sequenciamento, mantendo apenas, aproximadamente, 10% do sequenciamento original. Para reduzir a demanda computacional dos testes, apenas a região "*scaffold_1*" (1.087.660 pares de bases) do genoma de referência versão *AdarC3* foi considerada.

A chamada de variantes do painel de referência de alta confiabilidade foi realizada com os programas BWA e SamTools, considerando GQ ≥ 30, DP ≥ 12 e MAF ≥ 10%. A chamada de variantes do painel de subamostragem foi realizada com diferentes configurações dos parâmetros de controle de qualidade foram definidos a cada iteração:

- QG mínimo: 10, 13, 20 e 30, sendo a probabilidade de erro de sequenciamento 10%, 5%, 1% e 0,1%, respectivamente.
- DP mínimo: 4, 5 e 12.
- *CallRate* (em português, taxa de chamada) mínimo: 0%, 5% e 10%.
- *NonMissingRate* (em portugês, taxa de não-faltantes) mínimo: 25%, 50% e 75%.
- MAF mínimo: 10%.

A imputação de dados foi executada pelo programa Beagle 4.1 (BROWNINGS. R., BROWNING B. L., 2016). O programa Beagle foi escolhido porque atualmente é uma das poucas opções para imputação de painéis de genótipos que suportam entrada de dados no formato lista de probabilidades de genótipo (PL). O programa foi executado com o argumento GTGL, permitindo a utilização dos dados de genótipos de alta confiabilidade não-omitidos (GT), bem como as probabilidades dos genótipos omitidos (PL). Após a imputação de cada iteração, foram removidos genótipos com probabilidade do genótipo imputado (GP) < 95%.

O fluxograma da Figura 5 a seguir representa, de forma resumida, as etapas da simulação.

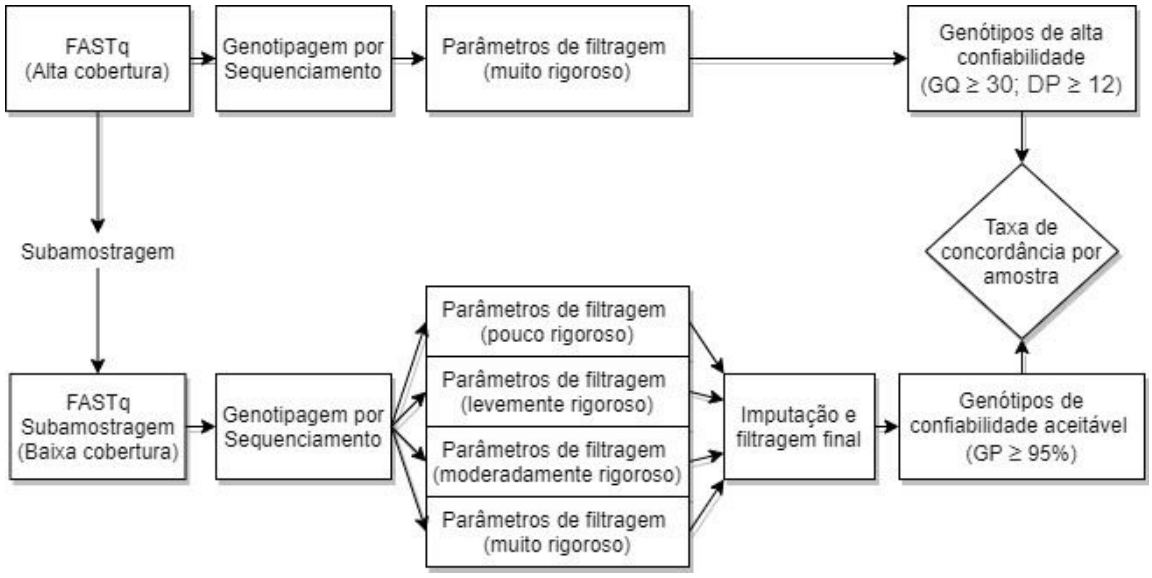

Figura 5 - Desenho esquemático do teste de parâmetros para imputação dos dados. Os parâmetros de filtragem descritos resumidamente após a etapa de GBS somam 108 configurações, sendo combinações das seguintes configurações: GQ≥{10;13;20;30}, DP≥{4;5;12}, *CallRate*≥{0%;5%;10%} e *NonMissingRate*≥{25%;50%;75%}. Fonte: Imagem elaborada pelo autor utilizando a ferramenta Online gratuita www.draw.io.

O programa utilizado para filtragem do painel de genótipos de referência foi o VCFtools (DANECEK, P. *et al.,* 2011). Para o painel de genótipos de baixa cobertura, um programa escrito na linguagem de programação C++ foi desenvolvido, permitindo selecionar alguns parâmetros específicos que não estão disponíveis em programas convencionais de manipulação de painéis de variantes. Mais detalhes do programa estão descritos no Apêndice X.

A cada iteração, foram anotados os valores finais de taxa de concordância, número de variantes e taxa de genotipagem por variante. O critério de comparação entre as configurações

foi um sistema de pontuação, no qual a pontuação (Po) foi calculada pela seguinte fórmula: $Po_i = 100\left(\frac{TC_i - 0.95}{0.05}\right)^2 \frac{N_i}{Max(N)}$, dado que, TC_i é a taxa de concordância e N_i, número de genótipos não-faltantes final da i-ésima configuração, respectivamente. A função *Max*(N) representa o maior valor de N observado no conjunto de testes. A pontuação foi calculada apenas quando TC > 95%, caso contrário a pontuação é automaticamente zero. As configurações que resultaram na melhor pontuação foram utilizadas para gerar o painel de genótipos final.

3.7. Finalização do painel de genótipos

Dois painéis de genótipos foram preparados para realização das análises estatísticas, um painel contendo genótipos não imputados (NIM) e um painel contendo genótipos imputados (IMP). NIM e IMP foram gerados pelos programas BWA e SAMtools utilizando as mesmas configurações citadas nas etapas anteriores.

NIM foi preparado com o programa VCFTools, mantendo apenas variantes seguindo os critérios: GQ ≥ 13, DP ≥ 4, MAF ≥ 10% e mínimo de genótipos não faltantes ≥ 30. IMP foi gerado pelo programa em C++ aplicando os critérios: GQ ≥ 20, DP ≥ 5, MAF ≥ 10%, *CallRate* ≥ 5% e *NonMissingRate* ≥ 50%, sendo submetido à imputação de genótipos omitidos e faltantes pelo programa Beagle 4.1 no modo GTGL. Após a imputação, foram mantidas variantes seguindo os critérios: GP ≥ 95%, MAF ≥ 10% e Taxa de genotipagem ≥ 70%.

Variantes que apresentam desequilíbrio de Hardy-Weinberg (HWD) frequentemente são descartadas por conta de possíveis erros de sequenciamento e genotipagem. No entanto, em um conjunto de amostras de uma população possivelmente estratificada, o efeito de Wahlund pode ser confundido com erros esses tipos de erros, causando descarte equivocado de variantes com sinais de estratificação importantes. Uma das hipóteses testadas no presente estudo é presença de sinal estratificante em escala microgeográfica. Por conta disso, o teste de equilíbrio de Hardy-Weinberg (HWE) foi aplicado em cada subgrupo, agrupados por local de coleta, para que não houvesse confusão sobre o efeito de Wahlund.

O procedimento descrito acima pode ser representado como na figura 6. O efeito de Wahlund é observado numa população estratificada quando há HWE nos subgrupos mas não há HWE na população.

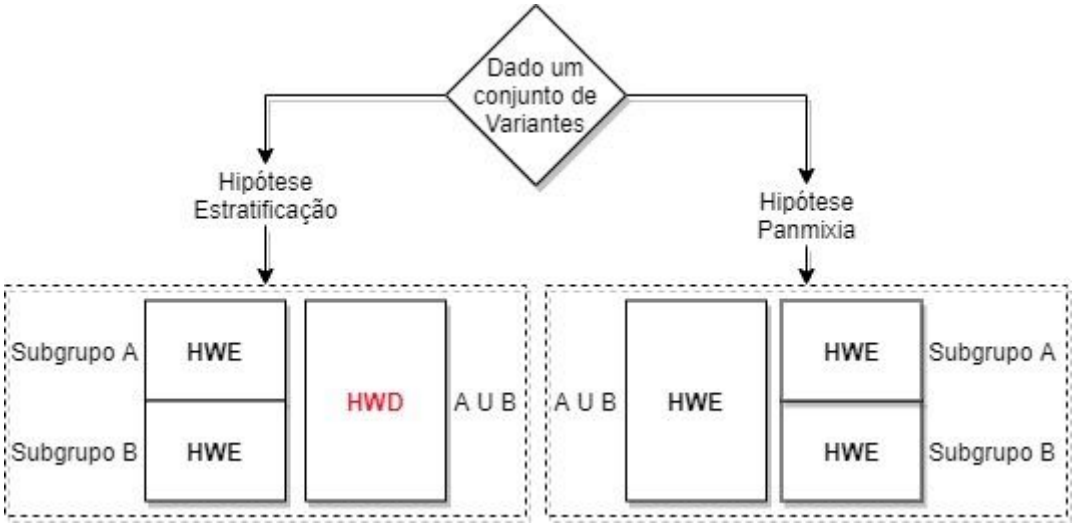

Figura 6 - Representação esquemática do efeito Wahlund. São dois cenários hipotéticos: População estratificada à esquerda e Panmixia à direita. O conjunto de variantes hipotéticos da imagem representa um conjunto de variantes que supostamente apresentariam ou não sinal estratificante entre as populações. Fonte: Imagem elaborada pelo autor utilizando a ferramenta Online gratuita www.draw.io.

Sendo assim, para cada subgrupo, agrupados por local de coleta, foram descartadas variantes em HWD, considerando *p-valor* < 1e-4 para teste de HWE no conjunto de dados IMP e NIM.

O painel NIM foi gerado apenas para comparação dos resultados de estratificação da população, verificando a concordância de resultados. Todas as outras análises estatísticas foram conduzidas apenas com o painel IMP.

3.8. Seleção por *Pruning* baseado em Desequilíbrio de Ligação

As análises estatísticas foram conduzidas pelo programa PLINK 1.9 (PURCELL, S., *et al*., 2007). A curva de decaimento do desequilíbrio de ligação foi estimada com base nos valores de desequilíbrio de ligação entre marcadores par a par, dado a distância em pares de base dos marcadores.

Desequilíbrio de ligação é uma medida de não-independência entre alelos em diferentes *loci*. Quando um alelo em particular é encontrado junto a outro em *loci* diferentes mais frequentemente que o esperado, estão em desequilíbrio de ligação (ZHENG, G. *et al*., 2012). O coeficiente de desequilíbrio pode ser calculado por $D_{AB} = p_{AB} - p_A p_B$. Uma medida de associação que é muito menos sensitiva à frequência alélica é r² (GILLESPIE, J. H., 2004), e pode ser calculada como $r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$.

Foram calculadas as médias de desequilíbrio de ligação (r²) para janelas de 500 pares de base de tamanho, por 40 janelas adjacentes, totalizando 20 kb de distância total. A predição da função de decaimento do desequilíbrio de ligação $\widehat{Y}$ foi calculada segundo o modelo não-linear: $\widehat{Y} = \beta_0 + \beta_1 \frac{1}{Log(x)} + e_{res}$, em que, β_0 é o valor de intercepto, β_1 é o coeficiente para a variável um sobre logaritmo da distância dos marcadores em pares de base e e_{res} valor de erro residual.

Após estimado a distância em pares de base aproximada para que médio r² < 0,1, foi realizado seleção de marcadores por *pruning*, considerando janelas de 10 kb de tamanho, r² < 0,1 a cada 10 marcadores adjacentes.

3.9. Análise de estratificação da população

Sinal de estratificação na população foi investigado com base na estimativa do parâmetro F_{ST} de acordo com o modelo matemático de Weir, B. S. e Cockerham, C. C. (1984). F_{ST} é visto como uma medida de diferença entre a probabilidade de que dois alelos escolhidos aleatoriamente entre as subdivisões e a probabilidade de que dois alelos escolhidos aleatoriamente na população sejam iguais (GILLESPIE, J. H., 2004). O valor de F_{ST} pode ser calculado de acordo com a seguinte fórmula: $F_{ST} = \frac{GS - GT}{1 - GT}$, dado que, GT e GS são as probabilidades de dois alelos escolhidos aleatoriamente na população e entre os subgrupos, respectivamente, sejam idênticos por estado.

O programa PLINK 1.9 foi utilizado para estimativa dos valores de F_{ST}. Comparações par a par foram realizadas e os valores médios de F_{ST} amplo no genoma, estimados. Sinais de estratificação entre subgrupos foram calculados, considerando amostras agrupadas por localização em escala microgeográfica (RUR, URB e PURB), comportamento de picada (Peri

e Intra, sendo peridomiciliar e intradomiciliar, respectivamente), horário de atividade (Dia e Noite, sendo 02h às 06h e 18h às 21h, respectivamente) e por localização geográfica (Brasil e Peru, sendo Brasil considerado apenas amostras da região RUR).

Foram realizados testes de permutação para verificar a significância dos valores médios de F_{ST} estimados pelo PLINK, sendo aplicado 10.000 permutações e considerando estatisticamente significativo apenas estimativas de F_{ST} médio quando $p_{valor} \leq 0{,}05$. As estimativas dos valores de F_{ST} por variante foram calculadas em comparação tripla e 1.000.000 de permutações aplicadas para verificar a significância do sinal estratificante ao longo do genoma. Variantes informativas foram considerados estatisticamente significativas quando $p_{valor} \leq 0{,}05$.

Análise de componentes principais (PCA) e análise discriminante dos componentes principais (DAPC) foram utilizados para avaliação do agrupamento entre as amostras. PCA foi executado para visualização do agrupamento das amostras, baseado na matriz de parentesco genômico (em inglês, *genomic variance-standardized relationship matrix*) calculado pelo programa PLINK. DAPC foi calculado pela função *dapc* do pacote *adegenet 2.1.2* para R (JOMBART, T.; AHMED, I., 2011).

3.10. Diversidade populacional e endogamia

Os parâmetros diversidade nucleotídica e coeficiente endogamia foram estimados para os grupos agrupados por local de coleta (RUR, URB e PURB). Para diversidade nucleotídica, foi calculado o valor de $\widehat{\pi}$ com o programa Arlequin 3.5 (EXCOFFIER, L., LISCHER, H.E.L., 2010), considerando o número de diferenças médias entre pares de sequências de DNA, segundo o modelo $\widehat{\pi} = \frac{n}{n-1} \sum_{i=1}^{k} \sum_{j>i}^{k} p_i p_j d_{ij}$, em que p_i e p_j são as frequências do alelo i e j em uma amostra de n sequências diferentes e d_{ij} o número de diferenças entre as sequências de DNA de i e j (TAJIMA, F., 1993).

O coeficiente de endogamia foi calculado individualmente para as amostras dentro dos subgrupos separados por local de coleta (RUR, URB e PURB) com o modelo: $f_i = 1 - \frac{HO_i}{HE_i}$, em que, *HO* é o número de genótipos homozigotos observados para o

i-ésimo indivíduo e *HE* é o valor esperado de genótipos homozigotos para o i-ésimo indivíduo, dado que $HE_i = L_i - \sum^{j} \frac{2p_j q_j T_j}{(T_j - 1)}$, em que L_i é o total de genótipos *non-missing* para o i-ésimo indivíduo e T_j é o dobro do número de genótipos *non-missing* para o j-ésimo SNP (PURCELL, S., *et al*., 2007).

3.11. Estudo de Associação ampla de Genoma

Estudo de associação ampla de genoma (GWAS) foi realizado com o painel de genótipos IMP utilizando o modelo estatístico de Cochran-Mantel-Haenszel Test (MANTEL, N., HAENSZEL, W. 1959). O teste assume caso controle 2x2xK para *k strata* sob hipótese nula H_0 que $MH \curvearrowleft \chi^2$ (chi-quadrado) com um grau de liberdade. O valor de *MH* pode ser calculado como: $\chi^2_{MH} = \frac{\left(|\sum_{i=1}^{k} [a_i - \frac{(a_i+b_i)(a_i+c_i)}{n_i}]| - \frac{1}{2} \right)^2}{\sum_{i=1}^{k} \frac{(a_i+b_i)(a_i+c_i)(b_i+d_i)(c_i+d_i)}{(n_i^3 - n_i^2)}}$, dado que, em sítios bialélicos (alelos A e B) para *k*-ésimo *stratum*, a e c igual ao número total de alelos A para caso e controle, respectivamente. Assim como b e d igual ao número total de alelos B para caso e controle, respectivamente. n é o total de alelos observados para a *k*-ésimo *stratum*, em que n = a + b + c + d.

Foram consideradas grupos caso/controle: intradomiciliar (Intra) e peridomiciliar (Peri) para comportamento de picada e grupo amostras coletadas entre as 18h00 à 22h00 (18h-22h) e amostras coletadas entre 02h00 à 06h00 (02h-06h) para horário de atividade. Foram considerados *stratum* os grupos por local de coleta (RUR, URB e PURB).

O método de correção de múltiplos testes de *FDR* de Benjamini e Hochberg (1995) foi aplicado para controle de falsos-positivos. As imagens do tipo *Manhattan Plot* foram geradas por script em linguagem R no RStudio (TEAM, R., 2013; RStudio TEAM, *et al*, 2015).

Genes adjacentes em até 10kb das variantes significativamente associadas foram investigados baseado na referência de anotação do genoma de *Anopheles darlingi*, versão AdarC3 disponível no formato gff3 em *VectorBase* (GIRALDO-CALDERÓN, G.L., *et al*., 2015).

4. RESULTADOS

4.1. Estatísticas do sequenciamento, GBS e construção do painel

4.1.1. Relatório de qualidade do sequenciamento

O relatório emitido pelo programa FASTQC está representado na Figura 7. Nenhuma sequência foi detectada com baixa qualidade. O total de sequências amostradas e analisadas no programa foi de 191.641.554 sequências. A qualidade média global por sequência foi de 31,4 em escala *Phred Score*.

O relatório notificou oscilações no conteúdo das 10 primeiras bases das sequências, como representado pela figura 7 superior direita. De acordo com o manual do FASTQC, dados de Illumina normalmente apresentam oscilações nas primeiras bases, no entanto a oscilação observada está entre 10% a 20%, portanto o módulo de controle de qualidade não apresentou falha e os dados foram mantidos.

O conteúdo GC observado de 48% foi semelhante a 48,15% descrito por Marinotti e colaboradores (2013) para *Anopheles darlingi*. A distribuição dos escores de qualidade em *Phred Scale* representados na imagem superior esquerda da figura 7 mostram que Q1 foi superior ao limiar de 20 ao longo de todas as 151 bases representadas. A queda observada do escore de qualidade a partir de 70 bases é frequentemente observado para esse tipo de sequenciamento, como relatado por Manley *et al* (2016).

Nenhuma contaminação por adaptadores foi encontrada. O tamanho das sequências avaliadas oscilou entre 35 pb e 151 pb, sendo predominante sequências de aproximadamente 151 pb. O nível de duplicação de sequências foi muito baixo, sendo que, aproximadamente, 94,33% das sequências seriam mantidas se os dados fossem desduplicados, ou seja, 5,67% de sequências duplicadas.

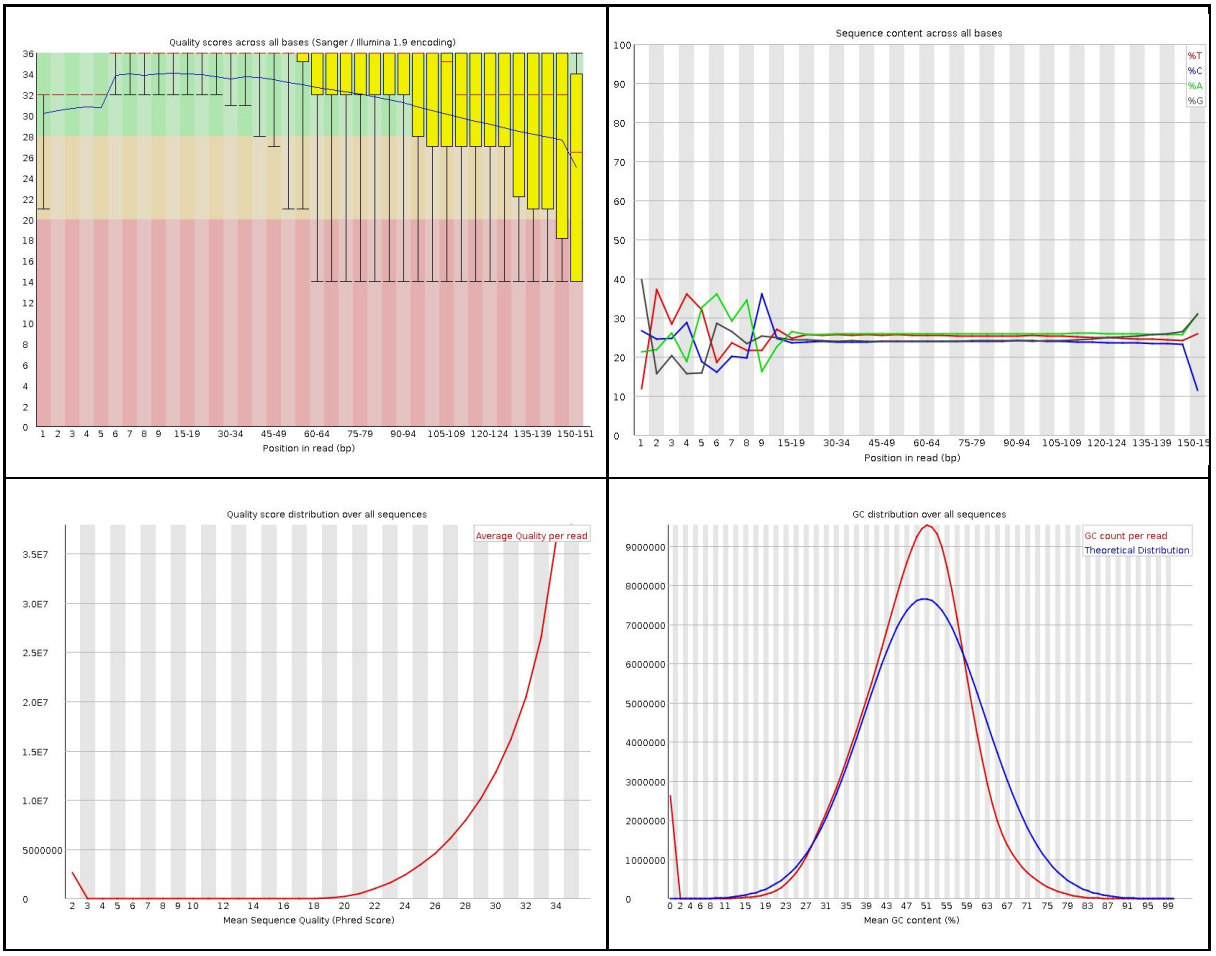

Figura 7 - Relatório do sequenciamento pelo programa FASTQC. Superior esquerdo: Escore de qualidade distribuídos ao longo das base, a linha azul representa média. Superior direito: Conteúdo das sequências ao longo das bases. Inferior esquerdo: Escore de qualidade ao longo das sequências. Inferior direito: Conteúdo GC em porcentagem ao longo das sequências. Fonte: Imagem gerada pelo programa FASTQC.

A cobertura média global das 524 amostras foi de 1,43 vezes, considerando genoma de referência versão *AdarC3* do banco VectorBase. A distribuição da cobertura média observada foi de: Min=0; Q1=0,3; Q2=0,8; Q3=2,0; Max=9,5. A Figura 8 representa a cobertura do genoma dado a profundidade de sequenciamento com base no alinhamento gerado pela combinação BWA + SamTools.

Figura 8 - Distribuição da profundidade do sequenciamento. Representação da cobertura do genoma de 524 amostras considerando sequência de referência *AdarC3*. Linha tracejada azul representa o valor médio de cobertura. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

4.1.2. Resultados da Genotipagem por Sequenciamento

Os resultados das combinações entre os programas de alinhamento e programas para chamada de variantes estão representados na Tabela 1. O total de genótipos chamados de cada configuração está descrito na segunda coluna, considerando apenas genótipos de acordo com os critérios de controle de qualidade estabelecidos para as 524 amostras. Cerca de 495,6 *Gigabytes* de dados foram processados em cada iteração. As combinações em que o processo de chamada de variantes foi realizado pelo programa SAMtools apresentaram taxa de genotipagem aproximadamente quatro vezes superior em relação às combinações com GATK.

Tabela 1 - Comparação do desempenho entre diferentes programas de alinhamento e chamada de variantes. Foram considerados somente genótipos que apresentaram GQ ≥ 20 e DP ≥ 5.

Configuração	Genótipos* (total)	Variantes* (total)	Taxa Genotipagem (% por amostra)	Tempo de operação (horas)
BWA + SamTools	**812.197.769**	**23.004.067**	**6,73%**	**60**
Bowtie2 + SamTools	429.186.962	15.182.371	5,39%	73
BWA + GATK	190.649.091	21.184.584	1,71%	189
Bowtie2 + GATK	86.902.786	13.491.693	1,22%	184

* nenhuma variante foi descartada por MAF, HWE ou MD para o cálculo da taxa de genotipagem. Melhores resultados de cada coluna estão destacados em negrito. Fonte: Elaborada pelo autor.

O protocolo de genotipagem escolhido para gerar o painel de variantes utilizado nas análises estatísticas do estudo foi o BWA+SamTools e sequência de referência *AdarC3*. O desempenho da configuração está descrito na tabela 2.

Tabela 2 - Desempenho da genotipagem por sequenciamento da configuração escolhida. BWA+SAMtools, considerando apenas genótipos que apresentaram GQ ≥ 20 e DP ≥ 5 e sequência de referência *AdarC3* disponível em VectorBase.

Workflow	Genotipados* (total)	Variantes* (total)	Taxa Genotipagem (% por amostra)	Tempo de operação (horas)
BWA+ SamTools	812.202.176	23.004.125	6,73%	53

* nenhum dado foi descartado por MAF, HWE ou MD nessa etapa. Fonte: Elaborada pelo autor.

4.1.3. Resultados do BLASTn utilizando sequência de COI

Os resultados do BLASTn estão descrito na tabela 3, totalizando 501 amostras identificadas taxonomicamente com resultados considerados estatisticamente significativos. Do total de 524 amostras, 23 amostras foram descartadas por não apresentarem resultados devido a cobertura extremamente baixa de 0,02% ± 0,1 para profundidade de sequenciamento igual a um.

Tabela 3 - Número de amostras identificadas taxonomicamente. Contagem considerando apenas amostras identificadas significativamente.

Genus	*Species*	Brasil	Peru
Anopheles	*albitarsis*	8	0
Anopheles	*benarrochi*	1	0
Anopheles	*costai*	0	1
Anopheles	*darlingi*	394	42
Anopheles	*dunhami*	0	1
Anopheles	*konderi*	1	9
Anopheles	*mattogrossensis*	0	3
Anopheles	*rangeli*	0	19
Anopheles	*triannulatus*	1	15
Resultado inconclusivo		6	0
TOTAL		411	90

Fonte: Elaborada pelo Autor

Foram descartadas 23 amostras sem identificação taxonômica, 17 amostras coletadas no Brasil e 48 amostras coletadas no Peru que não foram identificadas taxonomicamente como *Anopheles darlingi*. Também foram desconsideradas amostras que apresentaram menos de 25% do genoma com profundidade de sequenciamento de ao menos uma vez, sendo 73 amostras coletadas no Brasil e 3 amostras coletadas no Peru descartadas.

O total de amostras efetivas para as próximas etapas foi de 321 amostras coletadas no Brasil (22 adultos e 38 larvas de URB; 133 adultos e 33 larvas de RUR; 47 adultos e 48 larvas de PURB) e 39 amostras de larvas coletadas no Peru. A distribuição da cobertura do genoma das amostras selecionadas está representada na figura 9.

Figura 9 - Distribuição da profundidade do sequenciamento do conjunto amostral final. Distribuição da cobertura do genoma dado a profundidade do sequenciamento entre as 321 amostras considerando sequência de referência *AdarC3*. Linha tracejada azul representa o valor médio de cobertura. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

A cobertura média das 321 amostras foi de 1,79 vezes, com distribuição da cobertura média por amostra observada de: Min=0,3; Q1=0,74; Q2=1,3; Q3=2,3; Max=9,5.

4.1.4. Resultados do teste de parâmetros pré-imputação

A cobertura média do genoma das 10 amostras selecionadas para o teste de parâmetros pré-imputação foi de 7,03 ± 0,46 *e.p.*. A cobertura média do genoma das 10 amostras após a subamostragem foi de 0,71 ± 0,04 *e.p.*, cerca de 10 vezes menos, exatamente como o esperado.

A configuração utilizada em cada iteração e os respectivos resultados para número de variantes finais, concordância dos dados, taxa de genotipagem, escore final e outros estão descritos no apêndice A. A configuração que apresentou o melhor escore baseado na relação entre concordância de dados e total de genotipados final foi: GQ ≥ 20, DP ≥ 5, *CallRate* ≥ 5% e *NonMissingRate* ≥ 50%.

Essa configuração foi considerada, especificamente para o conjunto de dados do presente estudo, a melhor configuração para manter boa relação entre total de variantes chamadas e a confiabilidade que do painel imputado final. A configuração resultou no terceiro maior valor concordância (CONC ≅ 98,3%) e aproximadamente 2,75 vezes mais genótipos finais (total).

A relação entre dados faltantes por variante, número de variantes e concordância dos dados estão representados na figura 10.

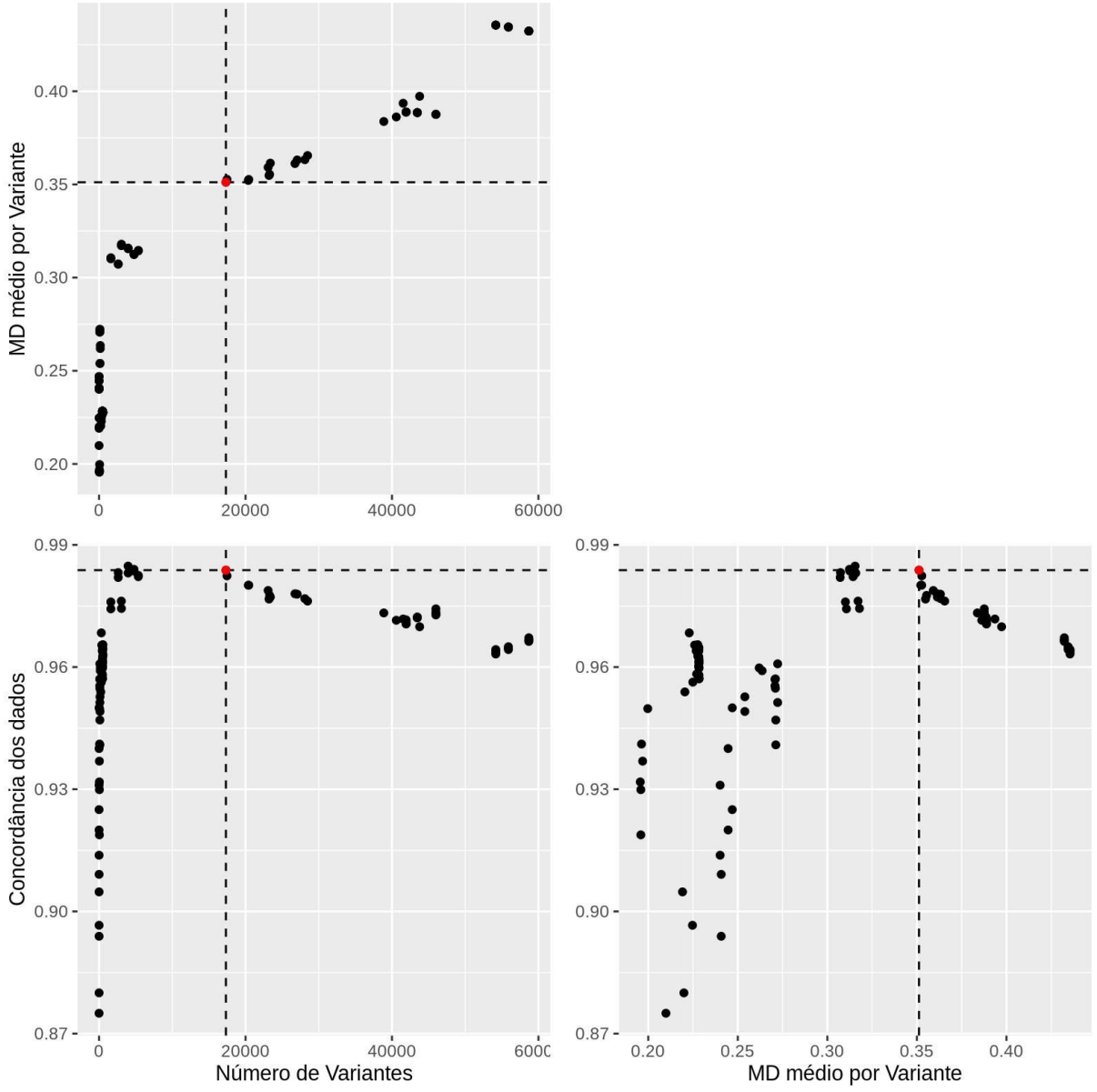

Figura 10 - Relação entre número de Variantes chamadas, média de dados faltantes por variante chamada e concordância dos dados imputados. Dados ausentes representado pela abreviatura MD, em inglês, *Missing-Data*. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com os pacotes GGPLOT2 e gridExtra.

4.1.5. Resultados da Imputação e Finalização do painel de Genótipos

Após submetido à imputação, o painel de variantes IMP resultou em 1.441.560 variantes e taxa de genotipagem de 83,61%, considerando que foram mantidas apenas variantes com MAF ≥ 10%, ou seja, número de cromossomos observados do alelo menor ≥ 32. O número de variantes para cada grupo por local de coleta considerando HWE e o número de variantes final utilizado para as próximas etapas estão descritos na tabela 4.

Tabela 4 - Número de variantes observados no painel de variantes final. São parâmetros utilizados como controle de qualidade: probabilidade de genótipo imputado GP ≥ 95%, *callRate* ≥ 0.7, MAF ≥ 10%, *HWE* < 1e-4.

Grupo	Número de Amostras	Não-MD Mínimo	Taxa de Genotipagem (%)	MAF ≥10%	MAF ≥20%	MAF ≥30%
RUR	166	45	83,7%	1.261.706	358.866	140.160
PURB	95	45	83,6%	1.261.646	358.571	140.180
URB	60	45	89,2%	1.111.793	270.816	93.614
FINAL	**321**	**135**	**89,3%**	**1.122.309**	**280.492**	**98.857**

Não-MD Mínimo: Mínimo de amostras com genótipos não ausentes. Fonte: Elaborada pelo Autor

4.2. Resultados das Análises estatísticas

4.2.1. Estimativa do decaimento de LD e seleção de marcadores

A curva de decaimento do desequilíbrio de ligação entre marcadores par a par foi estimada e os coeficientes da função não-linear observados foram -0,25 (p_{valor} < 0,001) e 3,35 (p_{valor} < 0,001) para o β_0 e β_1, respectivamente. O valor de R^2 ajustado da função foi de 0,95. Para o intervalo de confiança de 95%, os coeficientes para o intervalo inferior foram -0,31 e 3,8 para o β_0 e β_1, respectivamente. Em aproximadamente 12,57 kb de distância em pares de base, o desequilíbrio de ligação médio esperado estimado foi de 0,1 de acordo com a curva. Para fins práticos o valor considerado no processo de *pruning* foi de 10 kb de distância.

A Figura 11 representa os valores médios de desequilíbrio de ligação observados, dado a distância em pares de base, bem como a curva estimada de decaimento do desequilíbrio de ligação.

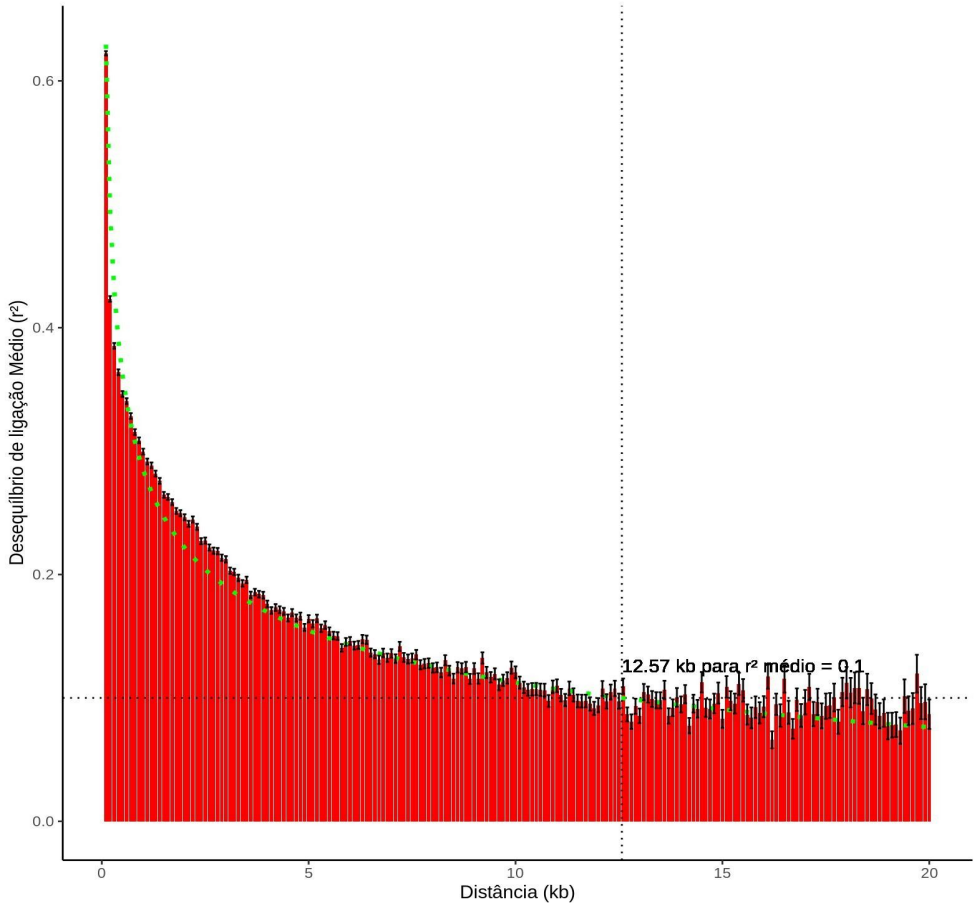

Figura 11 - Estimativa da curva de decaimento do desequilíbrio de ligação médio. A linha preta representa o valor para r² = 0,1. A linha verde tracejada representa a curva estimada do decaimento de desequilíbrio de ligação. Linhas pretas sobre as barras reproduzem o valor para de erro padrão da média. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

A seleção de variantes por *pruning* resultou em 294.881 variantes distribuídos em, aproximadamente, 2,04 variantes/kb e taxa de genotipagem de 89,33%. Houve redução do valor médio global de LD de 0,257 para 0,124, aproximadamente 48,24% de redução. O conjunto de variantes resultantes desse processo foi utilizado nas análises estatísticas das próximas etapas.

4.2.2. Resultados da análise de estratificação da população

Os valores F_{ST} médio do genoma obtidos com o modelo de *Weir* e *Cockerham* em comparações par a par pelo programa PLINK 1.9 estão descritos na tabela 5, considerando dados imputados (IMP) e não-imputados (NIM).

Tabela 5 - Valores de F_{ST} médio do genoma observado em comparações par a par.

Painel	Grupo A	Grupo B	N	Geno	F_{ST}
IMP	RUR	URB	294.881	194 (226)	**0,0008** *
	RUR	PURB	294.881	224 (261)	**0,0012** ***
	PURB	URB	294.881	133 (155)	**0,0015** **
	Peri	Intra	294.881	117 (133)	0,0005
	18h-21h	02h-06h	294.881	68 (73)	0,0005
NIM	RUR	URB	34.775	90 (226)	**0,0008** **
	RUR	PURB	204.413	70 (261)	**0,0003** *
	PURB	URB	23.143	51 (155)	**0,0026** ***
	Peri	Intra	15. 791	73 (133)	0,0005
	18h-22h	02h-06h	5.989	52 (73)	0,0000
	Brasil	Peru	92.503	64 (208)	**0,0406** ***

N: Número de variantes (SNP e INDEL) utilizados na análise. MD: Número médio de de genótipos não-faltantes por variante (total de amostras entre parênteses). MAF ≥ 10%. Valores de F_{ST} foram destacados quando $p_{VALOR} < 0,05$. *$p_{VALOR} < 0,05$. **$p_{VALOR} < 0,01$. ***$p_{VALOR} < 0,001$. Grupo Brasil é o mesmo que RUR. Fonte: Elaborada pelo Autor.

Os valores de F_{ST} estimados para cada variante nas comparações par a par (IMPUT) dos grupos URB, PURB e RUR estão representados na figura 12. Total de variantes com estimativas de $F_{ST} \geq 0,05$: 4.223 (1,43%) para PURB *vs* URB na figura da esquerda, 975 (0,331%) RUR *vs* PURB na figura do centro e 1721 (0,584%) para RUR *vs* URB na figura da direita.

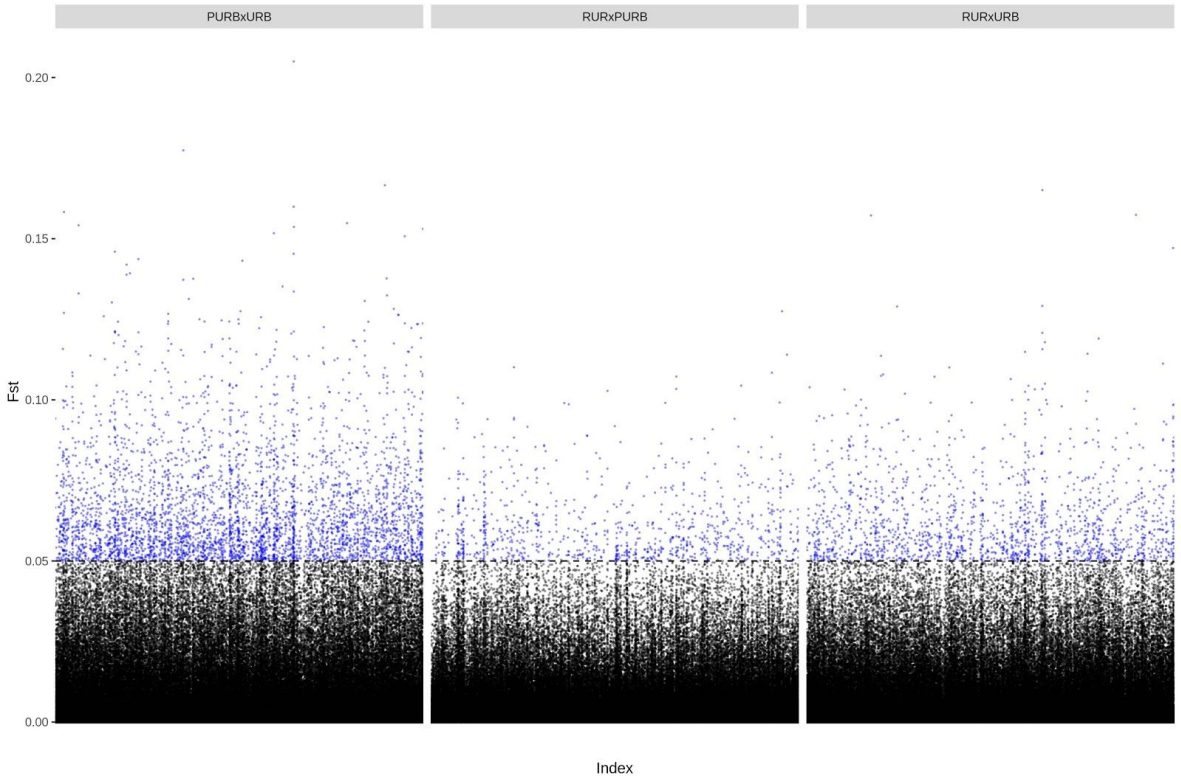

Figura 12 - Estimativas do parâmetro de F_{ST} em comparações par a par por variante ao longo do genoma. Estão destacados em azul variantes com estimativa de $F_{ST} \geq 0{,}05$. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

Os valores de F_{ST} estimados para cada variante na comparação tripla (URB, PURB e RUR) e seus respectivos níveis descritivos estão representados na figura 13. O valor médio de F_{ST} genômico foi 0,0012. Foram 5.509 variantes significativas no teste de permutação, cerca de 1,86% do total de variantes da análise. O valor médio de F_{ST} das variantes significativas foi de 0,030, distribuídos em: Min=0,016; Q1=0,025; Q2=0,028; Q3=0,033; Max=0,105.

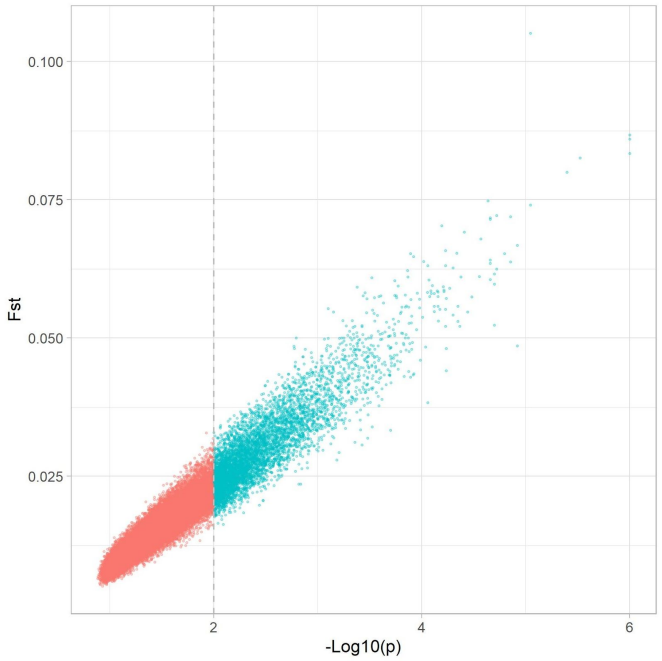

Figura 13 - Estimativas de F_{ST} e nível descritivo para estratificação em comparação tripla (URB, PURB e RUR). Os valores de p_{VALOR} foram calculados por teste de permutação (1 milhão de permutações). 41.903 variantes com $p_{VALOR} < 0{,}1$ estão representadas na figura de um total de 294.881 variantes. Linha horizontal tracejada representa o limiar de significância adotado, sendo $p_{VALOR} < 0{,}01$. Pontos com cor azul representam variantes significativas para o teste de permutação. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

Os valores F_{ST} médio obtidos utilizando variantes informativas para estratificação estão descritos na tabela 6.

Tabela 6 - Valores de F_{ST} Médio observado em comparações par a par utilizando variantes informativas.

Grupo A	Grupo B	N	Tx.Geno	F_{ST}
RUR	URB	5.509	85,98%	**0,026** ***
RUR	PURB	5.509	85,98%	**0,031** ***
PURB	URB	5.509	85,98%	**0,038** ***

N: Número de variantes (SNP e INDEL) utilizados na análise. MD: Número médio de de genótipos não-faltantes por variante (total de amostras entre parênteses). MAF ≥ 10%. Valores de F_{ST} foram destacados quando $p_{VALOR} < 0{,}05$. *$p_{VALOR} < 0{,}05$. **$p_{VALOR} < 0{,}01$. ***$p_{VALOR} < 0{,}001$. Grupo Brasil é o mesmo que RUR. Fonte: Elaborada pelo Autor.

A análise de componentes principais (PCA) calculada com base nas 5.509 variantes significativas no teste de permutação para estratificação foi conduzida e os dois primeiros componentes estão representados no canto esquerdo da figura 14. Cerca de 47,5% da variância da matriz de relacionamento genômico calculada pelo conjunto de 5.509 variantes foi descrita pelos dois primeiros componentes. O R^2 médio entre as 5.509 variantes é de 0,21 sendo: Min=0,00; Q1=0,02; Q2=0,11; Q3=0,34; Max=1,00. A porcentagem de amostras em que o grupos atribuído pelo DAPC foi igual ao observado foi de 97,19%.

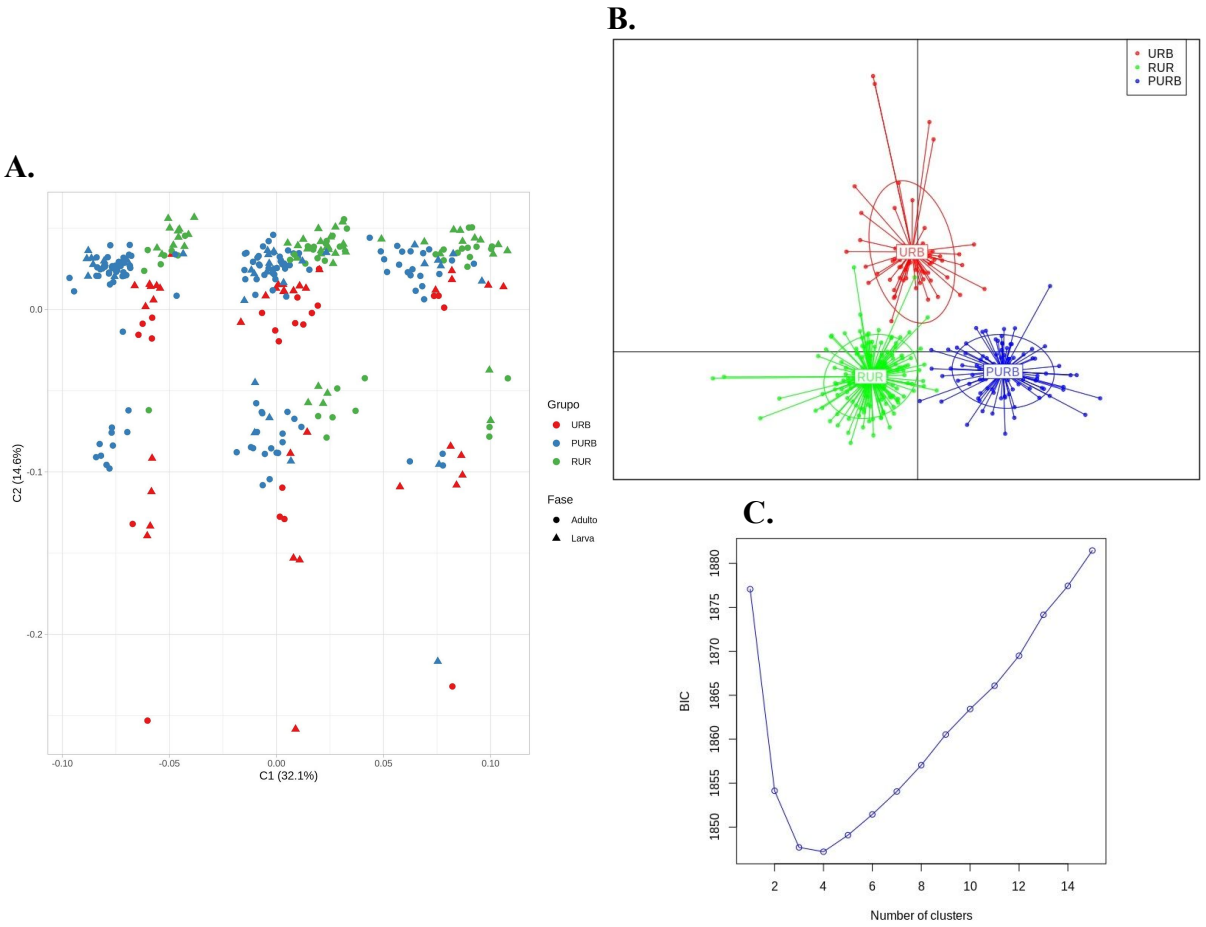

Figura 14 - Resultado da análise de componentes principais PCA (A), Análise discriminante dos componentes principais DAPC (B) e Valores de BIC para número de clusters (C). Foram utilizadas 5.509 variantes com sinal estratificante significativas no teste de permutação. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

4.2.3. Resultados das análises de diversidade nucleotídica e endogamia

As estimativas do coeficiente de endogamia e diversidade nucleotídica para os três grupos estão representados na tabela 7.

Tabela 7 - Estimativa do coeficiente de endogamia $\widehat{f}$ e diversidade molecular π. π calculado por média dos *loci*. Parâmetros estimados por grupos de local de coleta.

Grupo	N	$\widehat{f}$	$\widehat{f}$ *d.p.*	π	π *d.p.*
RUR	166	0,088	0,138	0,228	0,108
PURB	95	0,126	0,188	0,219	0,104
URB	60	0,143	0,138	0,215	0,102

N: Número de amostras dentro da categoria. Fonte: Elaborada pelo Autor.

4.2.4. Resultados das análises de GWAS

O número de amostras por grupo no teste de associação ampla de genoma para comportamento de picada foi: 40, 12 e 7 amostras intradomiciliares (caso) e 93, 35 e 15 amostras peridomiciliares (controle) para RUR, PURB e URB, respectivamente.

O resultado do teste de associação ampla de genoma pelo modelo Cochran-Mantel-Haenszel para comportamento de picada está representado na figura 15. Os valores de p_{FDR} estão representados na escala -$\text{Log}_{10}(p)$ no eixo Y do gráfico. O grau de inflação genômica estimada do teste foi de 1,18882.

GWAS para Comportamento de Picada (Intradomiciliar e Peridomiciliar)

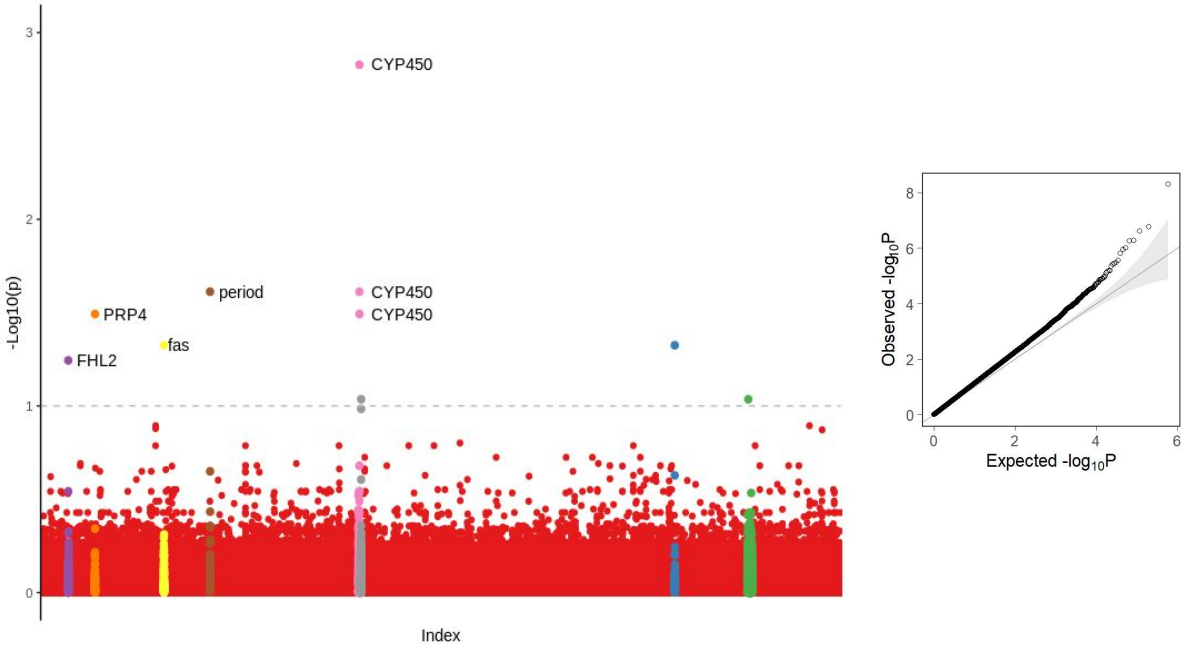

Figura 15 - Resultado da análise de GWAS para comportamento de picada. Linha tracejada representa limiar de significância estatística, considerando $p_{FDR} < 0.1$. Cores em destaque representam *Scaffolds* com ao menos uma variante significativa para o teste. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

O número de amostras por grupo no teste de associação ampla de genoma para horário de atividade foi: 34, 12 e 1 amostras coletadas das 18h00 às 22h00 (caso) e 39, 23 e 18 amostras coletadas das 02h00 às 06h00 (controle) para RUR, PURB e URB, respectivamente.

O resultado do teste de associação ampla de genoma pelo modelo Cochran-Mantel-Haenszel para horário de atividade está representado na figura 16. Os valores de p_{FDR} estão representados na escala -$\text{Log}_{10}(p)$ no eixo Y do gráfico. O grau de inflação genômica do teste foi de 1,14115.

GWAS para Horário de coleta (18h-22h e 02h-06h)

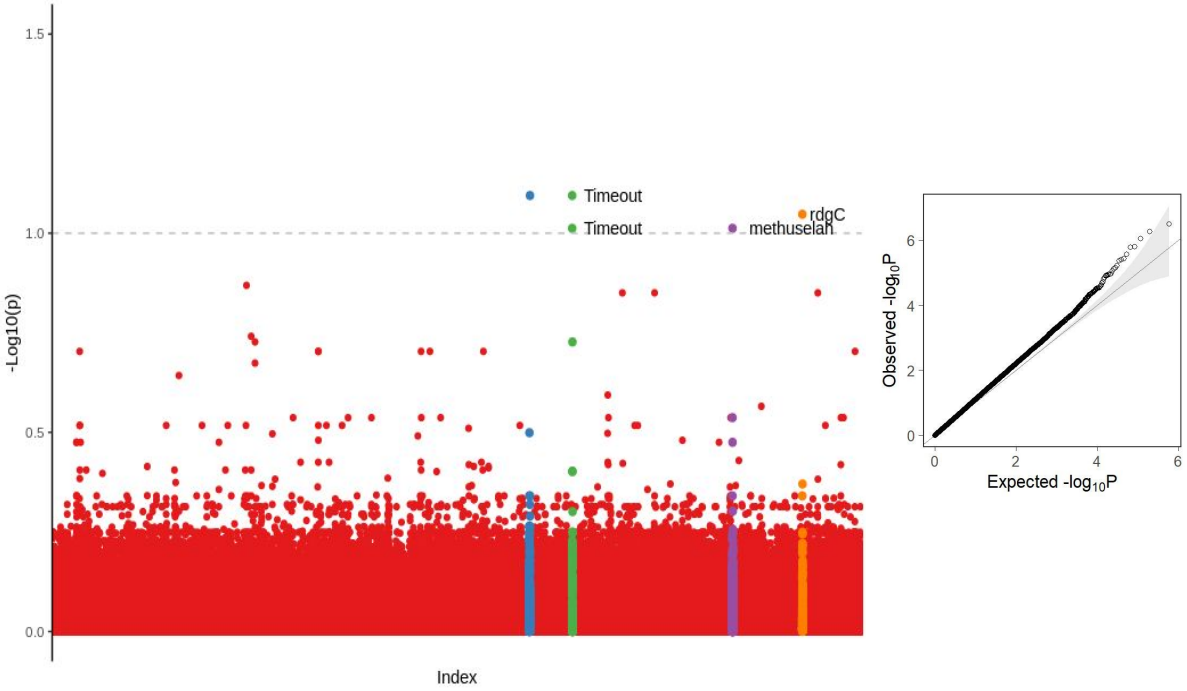

Figura 16 - Resultado da análise de GWAS para horário de atividade. Linha tracejada representa limiar de significância estatística, considerando $p_{FDR} < 0{,}1$. Cores em destaque representam *Scaffolds* com ao menos uma variante significativa para o teste. Fonte: Imagem elaborada pelo autor utilizando o programa RStudio Server com o pacote GGPLOT2.

5. DISCUSSÃO

5.1. Sequenciamento, GBS e imputação

A caracterização do sequenciamento como sequenciamento de baixa cobertura é evidente ao observar a distribuição da cobertura do genoma na figura 8. A baixa cobertura das 524 amostras de cerca de 1,43 vezes e, até mesmo considerando a cobertura de 1,79 vezes das 321 amostras efetivas do estudo representadas na figura 9, deixa nítido a importância do processamento adequado dos dados, principalmente na etapa de genotipagem por sequenciamento. De acordo com Yun Li *et al* (2011), mesmo com 2 a 4 vezes de cobertura é possível descobrir e genotipar com confiabilidade aceitável SNPs com MAF de até 0.5% e o presente estudo objetivou avaliar apenas SNPs de MAF ≥ 10%.

Outra característica importante da cobertura do sequenciamento é a distribuição da profundidade de sequenciamento entre amostras. Na figura 9, das 321 amostras efetivas e utilizadas na etapa da imputação, metade apresentou mais de 62,5% do genoma sequenciado ao menos uma vez e outra a metade, de 25% a 62,5%. Se a estratégia de preparo do painel de genótipos e imputação não fosse aplicada, cerca de 48,8% dos dados do sequenciamento seriam completamente descartados, considerando um DP mínimo de cinco vezes.

A combinação de programas BWA e SamTools apresentou melhor taxa de genotipagem o entre as comparações realizadas descritas na Tabela 1, portanto essa configuração foi escolhida para construção do painel de genótipos final. A comparação foi feita utilizando os parâmetros e configurações padrões dos manuais para que houvesse neutralidade com relação aos possíveis ajustes manuais dos programas. Existem diferenças para total de variantes chamadas e taxa de genotipagem entre os resultados das quatro combinações, mas tal comportamento é esperado, como já foi descrito na literatura (O'RAWE, J., *et al*., 2013).

Berry & Kearney (2011) mostraram que o número de genótipos chamados está relacionado com a acurácia na predição genômica no processo de imputação de painéis comerciais de baixa densidade usando painéis comerciais de alta intensidade. Apesar da confiabilidade dos genótipos ser reduzida em *loci* de baixa cobertura adotando a estratégia descrita na figura 2 (A para C), a informação recuperada como probabilidade de genótipos aumenta consideravelmente o número de genótipos e variantes na etapa de imputação, podendo ter efeito positivo e desempenho mais robusto por conta da maior quantidade de informação para imputação.

Considerando que a taxa de genotipagem pode influenciar na eficiência da imputação, um dos critérios para escolha da combinação BWA e SamTools para construção do painel de genótipos utilizado na imputação e, posteriormente, nas análises estatísticas do projeto, foi por apresentar a melhor taxa de genotipagem. No entanto, não houve conclusão sobre qual combinação de programas desempenharia o resultado mais acurado, a escolha foi baseada simplesmente na praticidade do protocolo, na taxa de genotipagem obtida utilizando parâmetros e configurações padrões descritas nos manuais e tempo de operação.

Os resultados do teste de parâmetros para preparação do painel para imputação está descrito no apêndice A, na tabela de resultados. A figura 10 representa a relação entre a proporção de dados ausentes, número de variantes chamadas e a concordância entre dos genótipos da simulação de baixa cobertura com a referência de cada iteração do teste.

O efeito descrito a seguir talvez seja o **ponto-chave** da estratégia para otimização de painéis de genótipos de baixa cobertura para imputação, adotando estratégia como da figura 2. É possível observar na figura 10 que há um ponto ótimo entre proporção de dados ausentes e concordância, bem como para número de variantes e concordância. O comportamento do resultado pode ser resumidamente descrito da seguinte forma:

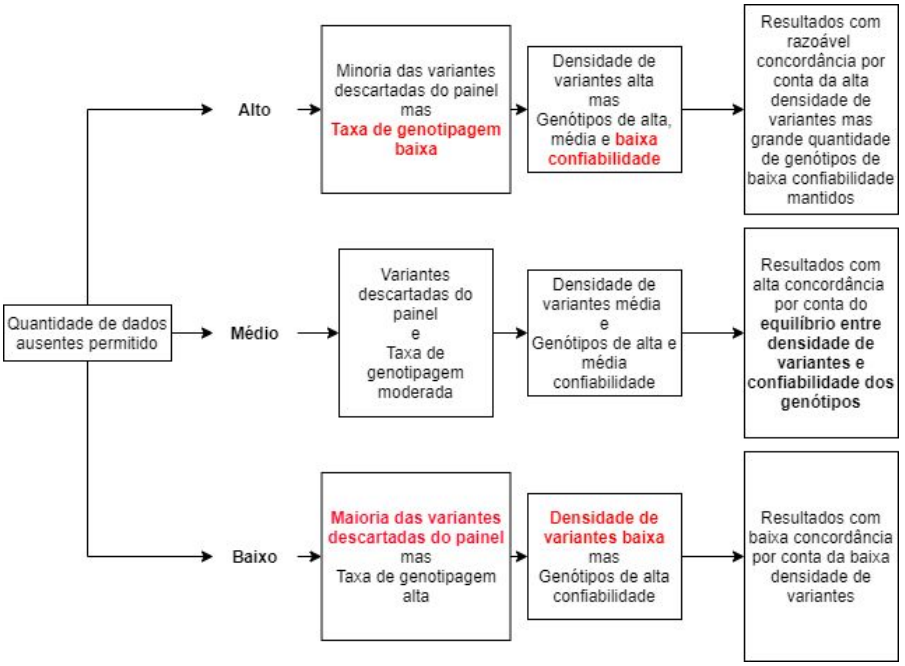

Figura 17 - Representação esquemática do impacto da quantidade de dados ausentes permitidos no painel de genótipos na imputação. Fonte: Imagem elaborada pelo autor utilizando a ferramenta Online gratuita www.draw.io.

Ou seja, controlar a quantidade e qualidade dos dados no painel de genótipos têm impacto relevante na acurácia da imputação. Roshyara *et al* (2014) também observaram que os principais fatores que influenciam na qualidade da imputação são a quantidade de dados ausentes e número de SNPs usados para a imputação, mesmo quando utilizado painel de referência de haplótipos.

Por isso é importante definir adequadamente parâmetros de profundidade de sequenciamento mínima, taxa de genotipagem (em inglês, *Call Rate*) e taxa de dados não faltantes (em inglês, *NonMissing Rate*). Os parâmetros definidos e utilizados no presente estudo foram estimados com base no teste e são específicos para o conjunto de dados do trabalho, portanto são necessários mais estudos para dimensionamento da estratégia como um todo, definindo as limitações, vantagens e desvantagens da aplicação em relação à outras estratégias de baixo custo.

5.2. Análises estatísticas

A curva de decaimento do desequilíbrio de ligação está descrito na figura 11. Em apenas alguns pares de base é possível notar declínio acentuado, estabilizando em r^2 médio de 0,1 em aproximadamente 12,57 kbp.

Harris *et al* (2010) descreveram baixo desequilíbrio de ligação em quatro genes relacionados à imunidade em população selvagem de *Anopheles gambiae* coletados na república dos camarões. O declínio acentuado do r^2 médio observado pelos autores é semelhante ao observado no presente estudo, até mesmo distâncias curtas como 200 bp apresentaram r^2 médio < 0,3. O'Loughlin *et al* (2016) também observaram estimativas de desequilíbrio de ligação semelhantes em *Anopheles gambie* coletados no Quênia, no distrito de Kilifi. Distância entre SNPs de 100 bp foram suficientes para r^2 médio observado ser muito baixo.

Mais descrições sobre o decaimento do desequilíbrio de ligação de *Anopheles darlingi* na região amazônica precisam ser realizadas para comparações diretas, no entanto os resultados obtidos nas estimativas foram suficientes para definir tamanho das janelas para o processo de seleção de marcadores por *pruning*. Apesar do desequilíbrio de ligação ser relativamente baixo em apenas alguns pares de base de distância, o r^2 de 0,1 em janelas de 10

kbp foi suficiente para que as 294.881 variantes após o *pruning* apresentassem metade do valor de r² médio em relação ao conjunto inicial de 1.122.309 variantes.

As estimativas de F_{ST} médio do genoma nas comparações par a par dos grupos por local de coleta foram significativos tanto para o painel não imputado (NIM) quanto para o imputado (IMP), representado na tabela 5. Os resultados de NIM são relativamente muito próximos de IMP e o único objetivo da utilização das estimativas de NIM foi comparar e verificar a concordância de resultados das comparações par a par. Apesar do valor F_{ST} médio do genoma ser classificado como muito fraco, a estimativa é significativa nas três comparações par a par representados na tabela 5, portanto o valor médio está significativamente atribuído aos grupos em escala microgeográfica. Além disso, é possível observar na figura 12 diversas variantes com potencial para sinal de estratificação para todas as comparações par a par.

As estimativas de F_{ST} médio do genoma nos grupos separados em escala microgeográfica apresentaram aproximadamente 3% do valor médio para Brasil e Peru, no entanto a distância média entre os grupos de Mâncio Lima é cerca de 150 vezes menor em comparação à distância do grupo Brasil até o grupo Peru.

Gélin e colaboradores (2016) avaliaram estratificação entre populações de *Anopheles funestus* e *Anopheles gambiae* em Muheza na Tanzânia. O estudo foi conduzido com utilizando microssatélites e as distâncias lineares entre as localidades estudadas variaram entre 5 e 10 Km. Foram observados valores de F_{ST} de 0,001, 0,003 e 0,009 em distâncias de 6,5 Km, 9,2 Km e 3,5 Km, respectivamente, mas nenhum dos resultados foram significativos nos testes de permutação. Embora marcadores SNP terem demonstrado superioridade em relação à microssatélites para predição de estratificação populacional (TELFER, E., *et al*., 2015; JEFFRIES, D., *et al*., 2016), as estimativas de Gélin e colaboradores são bem próximas das estimativas observadas no presente, dado que as distâncias entre grupos são semelhantes.

Considerando as variantes informativas para estratificação representadas na figura 13, apenas pequena porção das variantes apresentou significância nas estimativas de F_{ST} na comparação tripla. No entanto, é possível observar estimativas em máximas de até aproximadamente 0,2 para comparações par a par e até aproximadamente 0,1 para comparação tripla. Também é possível notar agrupamentos no PCA e DAPC da figura 14 utilizando variantes significativas no teste de permutação da comparação tripla da tabela 6

Campos *et al* (2017) também observaram grau de estratificação em população de *Anopheles darlingi* comparados em escala microgeográfica em uma amostra coletada na região amazônica no estado do Acre do Brasil utilizando SNPs de sequenciamento ddRADseq. As regiões de coleta descritas por Campos apresentavam-se cerca de 60 km distantes entre si e F_{ST} aproximado de 0,072.

Na análise de PCA da figura 14, é possível notar agrupamentos das amostras por local de coleta, mas também é possível notar um segundo padrão de agrupamento maior, formando subgrupos separados por local de coleta. Não houve conclusão significativa sobre a formação dos agrupamentos maiores. No entanto, uma hipótese é que esse efeito de agrupamento seja causado por inversões cromossômicas presentes na população, considerando que inversões cromossômicas já foram descritas em *Anopheles darlingi* (RAFAEL, M., *et al.*, 2010), inclusive como indicativo de estratificação populacional na região Amazônica (CORNEL, A., *et al.*, 2016).

Nowling *et al* (2018) demonstraram que inversões cromossômicas podem ser detectadas a nível molecular por genotipagem de alta densidade e análises de associação, bem como análise de PCA. Os padrões e pontos de inversão na sequência de referência avaliados em *Anopheles gambiae* e *Anopheles coluzzi* são capturados de forma eficiente por estudo de associação ampla do genoma demonstrado pelos autores, bem como são notáveis os agrupamentos no PCA separados por conta do genótipo da inversão.

O principal problema no momento do desenvolvimento das análises de diagnóstico de inversão cromossômica do presente estudo foi a disponibilidade de uma sequência de referência para *Anopheles darlingi* que apresentasse continuidade na sequência de referência. A sequência de referência do genoma de *Anopheles darlingi* disponível no momento das análises era muito fragmentada, tanto na versão disponível em NCBI quanto VectorBase. O genoma de referência *AdarC3*, por exemplo, está fragmentado em 2221 sequências (*contigs*), dificultando a identificação e orientação das inversões.

O coeficiente de endogamia médio descrito na tabela 7 sugere possível relação com o processo de urbanização, bem como a diversidade nucleotídica. É possível observar que o coeficiente de endogamia aumenta e a diversidade nucleotídica diminui conforme o nível de urbanização do local aumenta: RUR, PURB e URB. Wilke e colaboradores (2017) observaram que populações de *Aedes aegypti* em escala microgeográfica de não mais que 30 Km de distância, na cidade de São Paulo, apresentaram aumento do coeficiente de endogamia

de acordo com o aumento da urbanização do local de coleta, sendo que os valores mais altos das estimativas do coeficiente de endogamia foram observados no ambiente urbanizado, seguido por áreas intermediárias e os valores mais baixos nas áreas mais conservadas.

A análise de associação ampla de genoma para comportamento de picada apresentou 12 variantes estatisticamente significativas, representadas na figura 15. A investigação de genes adjacentes aos *loci* significativamente associados identificou alguns genes importantes para serem discutidos: *Cytochrome P450 4C1* (*cyp450*) e *prp4*.

A superfamília de genes citocromo P450 são extremamente importantes no metabolismo de compostos endógenos como hormônios, ácidos graxos e esteróides, além de catabolismo e anabolismo de xenobióticos como drogas, pesticidas, e toxinas de plantas. O gene *cyp450* é bem conhecido como importante componente na resistência a inseticidas em insetos (SCOTT, J. G., 1999), inclusive entre os anofelinos (BALADANIDOU, V., *et al*., 2016; IBRAHIM, S. S., *et al.*, 2016; DONNELLY, M. J., *et al*., 2016). É evidente a relação entre uso de inseticidas piretróides no interior das casas como forma de controle do vetor e a função biológica do gene *cyp450*. A presença de marcadores adjacentes ao *cyp450* sugere fortemente que existem polimorfismos nos indivíduos que apresentam maior resistência a inseticidas e proporcionam maior chance de sobrevivência em ambiente com exposição a inseticida.

Gao e colaboradores (2018) estudaram o perfil de expressão gênica em resposta a cinco diferentes tipos de inseticidas em *Plutella xylostella* baseado em análise de transcriptoma. A expressão diferencial de *cyp450* foi notada, indicando novamente a importância funcional desse gene na resistência a inseticidas. Interessantemente, o gene *Serine/threonine-protein kinase* (*prp4*) foi observado entre os genes diferencialmente expressos para todos os tratamentos, inclusive estando entre os genes com maior *Fold Change*. A associação de variantes adjacentes ao *prp4* vista no presente estudo também pode indicar relação entre o gene e algum mecanismo de resistência a inseticidas, mas o papel do *prp4* ainda não é claro e os resultados proporcionam pistas para que novos estudos sejam conduzidos para estabelecer relação entre *prp4* e resistência a inseticidas.

A análise de associação ampla de genoma para para horário de atividade apresentou 5 variantes estatisticamente significativas, representadas na figura 16. A investigação de genes adjacentes aos *loci* significativamente associados identificou alguns genes importantes para

serem discutidos: *Retinal Degeneration C Protein* (*rdgC*) e *Timeout/Timeless-2* (*tim2* ou *Timeout*).

A rodopsina-fosfatase *rdgC* tem importante papel na desfosforilação da rodopsina *Rh1*, proteína fotossensorial mais abundante em *Drosophila melanogaster*. *Rh1* é requerido para sincronismo molecular com a luz e comportamento de ritmos circadianos (OGUETA, M., *et al*., 2018). A perda de função do gene *rdgC* está associada com a hiperfosforilação de Rh1, levando a degeneração de fotorreceptores na presença de luz em em adultos de *D. melanogaster* (XIONG, B., BELLEN, H. J., 2013; VOOLSTRA, O. *et al*., 2018). Adewoye (2011) descreveu *loci* de traços quantitativos (QTL) associados a resposta a variação da luz em *Drosophila melanogaster* e encontrou diversos genes associados aos intervalos dos QTL, dos quais muitos relacionados a regulação do ciclo circadiano, inclusive o gene *rdgC*.

Timeout é um gene parálogo de *Timeless1* (*tim1*), ambos descritos como componentes do mecanismo de regulação do ciclo circadiano em *Drosophila melanogaster*. *Timeout* está envolvido principalmente com a percepção de luminosidade e fotorrecepção circadiana em adultos de *Drosophila melanogaster* (BENNA, C,. *et al*., 2010). Honnen *et al* (2016) estudaram genes sexo-específico diferencialmente expressos em resposta a tratamento com luz artificial durante a noite em mosquitos *Culex pipiens* e *Timeout* foi observado, inclusive apresentando tendência para machos. Considerando que todas as amostras utilizadas na análise de GWAS do presente estudo eram *Anopheles darlingi* adultos, é notável a associação entre o papel biológico de *Timeout* e o contraste entre os estímulos luminosos nos diferentes momentos em que anofelinos foram coletados.

É importante destacar que os estudos de associação ampla de genoma só foram possíveis utilizando o painel de variantes IMP, pois os dados NIM não permitem devido a grande quantidade de *missing-data* que acaba sendo gerada ao descartar genótipos com qualidade muito próxima mas abaixo do limiar mínimo. Por isso, novas abordagens e estudos para verificar a eficiência dessa metodologia em diferentes cenários são necessários para verificar as limitações da utilização de dados de baixa cobertura.

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.

APÊNDICES

APÊNDICE A - Tabela de resultados do teste de parâmetros pré-imputação

Tabela A - Resultados do teste de parâmetros pré-imputação. São parâmetros definidos em cada iteração: qualidade de genótipo mínima (GQ), profundidade de sequenciamento mínimo (DP), CallRate mínimo (CR) e *NonMissingRate* mínimo (NMD). São resultados de cada iteração: Número de variantes final (NVAR), taxa de genotipagem (Tx.GT), porcentagem de dados faltante ou *Missing-data* (MD), total de genótipos chamados e imputados considerando probabilidade de genótipo mínima de 95% (GT), concordância entre genótipos da referência (GQ ≥ 20 e DP ≥ 12) e imputados (CONC), pontuação ou escore final (P).

GQ	DP	CR	NMD	NVAR	Tx.GT	MD	GT	CONC	P
20	5	0.05	0.5	17326	7.78%	64.88%	23782	98.38%	30.84
20	5	0.05	0.25	17474	7.75%	64.72%	23956	98.24%	28.57
13	4	0.1	0.25	20445	13.70%	35.28%	26048	98.01%	26.83
13	4	0.1	0.5	20382	13.70%	35.22%	26017	98.01%	26.82
13	5	0.05	0.25	27056	8.02%	36.33%	29483	97.79%	26.07
13	5	0.05	0.5	26754	8.02%	36.11%	29345	97.80%	26
20	4	0.05	0.5	23088	7.96%	35.93%	26879	97.88%	25.37
10	5	0.05	0.5	28105	8.15%	36.33%	29592	97.68%	24.05
10	4	0.1	0.25	23291	13.83%	35.56%	27442	97.76%	23.69
20	12	0	0.5	45998	0.22%	38.77%	35065	97.43%	23.55
10	5	0.05	0.25	28477	8.12%	36.54%	29687	97.62%	23.15
20	4	0.05	0.25	23389	7.93%	36.14%	26967	97.72%	22.65
10	4	0.1	0.5	23209	13.83%	35.49%	27385	97.67%	22.15
10	12	0	0.5	45998	0.22%	38.77%	35176	97.35%	22.07
30	12	0	0.5	45998	0.19%	38.77%	35209	97.30%	21.16
13	12	0	0.5	45998	0.22%	38.77%	35267	97.28%	20.83
13	4	0.05	0.5	38886	10.90%	38.36%	30324	97.33%	18.65
13	5	0	0.5	43461	6.36%	38.86%	31960	97.23%	17.97
10	5	0	0.5	43461	6.60%	38.86%	31877	97.22%	17.85
20	5	0	0.5	43461	4.97%	38.86%	31983	97.21%	17.71
30	5	0	0.5	43461	1.98%	38.86%	31926	97.21%	17.7
13	4	0.05	0.25	41528	10.62%	39.35%	30371	97.18%	16.32
30	4	0	0.5	41928	2.65%	38.89%	30184	97.16%	15.96
10	4	0.05	0.5	40592	11.33%	38.61%	30425	97.15%	15.93
20	4	0	0.5	41928	5.99%	38.89%	30280	97.10%	15.08

Continua

GQ	DP	CR	NMD	NVAR	Tx.GT	MD	GT	CONC	P
13	4	0	0.5	41928	10.40%	38.89%	30384	97.06%	14.57
10	4	0	0.5	41928	11.08%	38.89%	30286	97.06%	14.53
13	5	0.1	0.5	4775	12.56%	31.23%	10489	98.40%	13.78
10	4	0.05	0.25	43777	10.96%	39.72%	30300	96.99%	13.66
10	5	0.1	0.25	5369	12.62%	31.45%	11248	98.25%	13.46
13	5	0.1	0.25	4779	12.56%	31.27%	10477	98.36%	13.4
10	5	0.1	0.5	5362	12.62%	31.42%	11290	98.22%	13.31
20	4	0.1	0.5	3966	12.59%	31.54%	8637	98.48%	11.85
20	12	0	0.25	58688	0.15%	43.24%	34760	96.72%	11.61
30	12	0	0.25	58688	0.15%	43.24%	34646	96.68%	11.14
20	4	0.1	0.25	3973	12.59%	31.57%	8688	98.31%	10.82
10	12	0	0.25	58688	0.19%	43.24%	34822	96.65%	10.72
13	12	0	0.25	58688	0.19%	43.24%	34761	96.63%	10.51
20	5	0.1	0.25	2618	12.56%	30.74%	6612	98.32%	8.27
13	5	0	0.25	55901	5.37%	43.46%	31567	96.50%	8.1
20	5	0.1	0.5	2616	12.56%	30.71%	6628	98.20%	7.72
20	5	0	0.25	55901	4.17%	43.46%	31631	96.47%	7.7
10	5	0	0.25	55901	5.59%	43.46%	31518	96.46%	7.63
30	5	0	0.25	55901	1.64%	43.46%	31541	96.43%	7.28
13	4	0	0.25	54176	8.83%	43.55%	29875	96.43%	6.88
30	4	0	0.25	54176	2.22%	43.55%	29753	96.41%	6.68
10	4	0	0.25	54176	9.44%	43.55%	29922	96.36%	6.27
20	4	0	0.25	54176	5.03%	43.55%	29862	96.32%	5.93
30	4	0.05	0.5	3046	6.73%	31.73%	6614	97.62%	5.14
30	4	0.05	0.25	3057	6.73%	31.79%	6656	97.44%	4.49
30	5	0.05	0.5	1610	6.94%	31.02%	4336	97.60%	3.33
30	5	0.05	0.25	1614	6.94%	31.08%	4295	97.43%	2.87
30	12	0	0.75	531	1.48%	22.78%	1376	96.55%	0.37
20	5	0.1	0.75	312	15.31%	22.28%	933	96.84%	0.36
13	12	0	0.75	531	1.60%	22.78%	1379	96.45%	0.33
30	5	0	0.75	487	6.02%	22.84%	1264	96.48%	0.31
10	5	0.05	0.75	487	15.65%	22.84%	1253	96.45%	0.3

Continua

GQ	DP	CR	NMD	NVAR	Tx.GT	MD	GT	CONC	P
13	5	0.05	0.75	487	15.12%	22.84%	1249	96.40%	0.28
20	4	0.1	0.75	373	15.74%	22.59%	1041	96.54%	0.28
10	12	0	0.75	531	1.64%	22.78%	1394	96.31%	0.27
13	5	0.1	0.75	435	15.86%	22.69%	1166	96.40%	0.26
20	12	0	0.75	531	1.54%	22.78%	1406	96.27%	0.26
20	5	0	0.75	487	12.59%	22.84%	1239	96.25%	0.22
10	5	0	0.75	487	15.65%	22.84%	1269	96.14%	0.19
13	4	0.1	0.75	465	21.73%	22.84%	1208	96.19%	0.19
13	4	0	0.75	465	21.73%	22.84%	1206	96.14%	0.18
13	5	0	0.75	487	15.12%	22.84%	1261	96.11%	0.18
20	4	0	0.75	465	14.35%	22.84%	1216	96.09%	0.16
20	5	0.05	0.75	479	12.75%	22.81%	1245	96.02%	0.15
10	4	0.05	0.75	465	22.84%	22.84%	1201	96.00%	0.14
20	4	0.05	0.75	464	14.35%	22.84%	1204	95.97%	0.13
10	4	0	0.75	465	22.84%	22.84%	1214	95.80%	0.09
10	5	0.1	0.75	446	16.27%	22.72%	1186	95.83%	0.09
10	4	0.1	0.75	465	22.84%	22.84%	1224	95.71%	0.07
13	4	0.05	0.75	465	21.73%	22.84%	1202	95.72%	0.07
30	4	0	0.75	465	7.01%	22.84%	1216	95.72%	0.07
30	4	0.1	0.5	171	12.22%	26.20%	659	95.98%	0.07
10	12	0.05	0.5	114	7.10%	27.22%	446	96.08%	0.06
30	4	0.1	0.25	172	12.22%	26.36%	660	95.91%	0.06
30	4	0.05	0.75	314	8.43%	22.50%	904	95.63%	0.04
20	12	0.05	0.5	107	7.01%	27.10%	431	95.71%	0.02
30	12	0.05	0.5	99	6.91%	27.07%	407	95.70%	0.02
20	12	0.05	0.25	107	7.01%	27.10%	431	95.48%	0.01
30	5	0.05	0.75	234	8.58%	22.04%	760	95.39%	0.01
30	12	0.05	0.25	99	6.91%	27.07%	404	95.54%	0.01
10	12	0.05	0.25	114	7.10%	27.22%	462	95.13%	0
10	12	0.05	0.75	48	7.65%	19.60%	221	92.99%	0
10	12	0.1	0.25	6	12.28%	24.07%	33	89.39%	0
10	12	0.1	0.5	6	12.28%	24.07%	33	90.91%	0

Continua

GQ	DP	CR	NMD	NVAR	Tx.GT	MD	GT	CONC	P
10	12	0.1	0.75	5	12.35%	22.47%	29	89.66%	0
13	12	0.05	0.25	112	7.07%	27.13%	457	94.09%	0
13	12	0.05	0.5	112	7.07%	27.13%	462	94.70%	0
13	12	0.05	0.75	48	7.56%	19.60%	240	91.88%	0
13	12	0.1	0.25	5	12.53%	24.01%	29	91.38%	0
13	12	0.1	0.5	5	12.53%	24.01%	29	93.10%	0
13	12	0.1	0.75	4	12.75%	21.98%	25	88.00%	0
20	12	0.05	0.75	46	7.31%	19.69%	214	93.69%	0
20	12	0.1	0.25	4	13.06%	24.44%	25	94.00%	0
20	12	0.1	0.5	4	13.06%	24.44%	25	92.00%	0
20	12	0.1	0.75	3	13.49%	21.91%	21	90.48%	0
30	4	0.1	0.75	75	13.27%	19.97%	299	94.98%	0
30	5	0.1	0.25	131	12.25%	25.40%	521	94.91%	0
30	5	0.1	0.5	131	12.25%	25.40%	518	95.27%	0
30	5	0.1	0.75	67	13.15%	19.63%	263	94.11%	0
30	12	0.05	0.75	42	7.16%	19.57%	198	93.18%	0
30	12	0.1	0.25	3	13.49%	24.69%	20	92.50%	0
30	12	0.1	0.5	3	13.49%	24.69%	20	95.00%	0
30	12	0.1	0.75	2	14.51%	20.99%	16	87.50%	0

Fonte: Elaborada pelo Autor.

APÊNDICE B - Lista de genes adjacentes aos marcadores significativos nas análises de GWAS para comportamento de picada

Tabela B - Lista de genes adjacentes aos marcadores estatisticamente significativos no GWAS para comportamento de picada. Estão descritos genes adjacentes localizados em um intervalo máximo de 10 kb, sendo 5 kb *upstream* e *downstream*. Os intervalos das regiões gênicas estão representadas entre chaves da seguinte forma: [inicial:final].

Scaffold	Posição	Lista de genes adjacentes	p_{VALOR}	p_{FDR}
241	85279	FMRFamide receptor [81054:82517]; cytochrome P450 4C1 [87272:89456];	5.04E-09	1.49E-03
241	86788	FMRFamide receptor [81054:82517]; cytochrome P450 4C1 [87272:89456]; DNA-J [91017:92150];	1.73E-07	2.45E-02
1524	14688	period circadian protein [8167:14819];	2.49E-07	2.45E-02
1116	31454	prp4 [28738:32685];	5.29E-07	3.22E-02
241	93296	cytochrome P450 4C1 [87272:89456]; DNA-J [91017:92150];	5.46E-07	3.22E-02
1336	22031	faint sausage [174:20291];	1.00E-06	4.74E-02
654	12333	-	1.12E-06	4.74E-02
105	168412	four and a half lim [163053:167286];	1.55E-06	5.71E-02
34	385878	-	2.71E-06	8.87E-02
400	82196	-	3.24E-06	9.21E-02
243	54278	-	3.59E-06	9.21E-02
8	154691	-	3.75E-06	9.21E-02

Fonte: Elaborada pelo Autor

APÊNDICE C - Lista de genes adjacentes aos marcadores significativos nas análises de GWAS para horário de atividade

Tabela C - Lista de genes adjacentes aos marcadores estatisticamente significativos no GWAS para horário de atividade. Estão descritos genes adjacentes localizados em um intervalo máximo de 10 kb, sendo 5 kb *upstream* e *downstream*. Os intervalos das regiões gênicas estão representadas entre chaves da seguinte forma: [inicial:final].

Scaffold	Posição	Lista de genes adjacentes	p_{VALOR}	p_{FDR}
404	77474	-	3.17E-07	8.04E-02
462	12409	timeout/timeless-2 [10017:14540]; transmembrane protein 53-B [15487:16608];	5.46E-07	8.04E-02
867	14323	retinal degeneration C protein [15781:45295];	9.12E-07	8.96E-02
462	14409	timeout/timeless-2 [10017:14540]; transmembrane protein 53-B [15487:16608]; ap endonuclease [17694:20540];	1.61E-06	9.72E-02
73	219145	methuselah [216568:225247];	1.65E-06	9.72E-02

Fonte: Elaborada pelo Autor