

DESENVOLVIMENTO DE UMA FERRAMENTA  
COMPUTACIONAL AMIGÁVEL PARA ANÁLISE DE DADOS  
DE SEQUENCIAMENTO GENÔMICO DE BAIXA COBERTURA  
E APLICAÇÃO EM ESTUDO DE GENÉTICA DE POPULAÇÕES

**MARCUS VINICIUS NIZ ALVAREZ**

**BOTUCATU – SP**

**2024**

**UNIVERSIDADE ESTADUAL PAULISTA**  
**“Júlio de Mesquita Filho”**  
**INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU**

**DESENVOLVIMENTO DE UMA FERRAMENTA  
COMPUTACIONAL AMIGÁVEL PARA ANÁLISE DE DADOS  
DE SEQUENCIAMENTO GENÔMICO DE BAIXA COBERTURA  
E APLICAÇÃO EM ESTUDO DE GENÉTICA DE POPULAÇÕES**

**MARCUS VINICIUS NIZ ALVAREZ**  
**ORIENTADOR: PAULO EDUARDO MARTINS RIBOLLA**

Tese submetida ao Programa de Pós-Graduação em Ciências Biológicas (Genética) do Instituto de Biociências de Botucatu da Universidade Estadual Paulista “Júlio de Mesquita Filho” para obtenção do título de Doutor em Ciências Biológicas (Genética).

A473d      Alvarez, Marcus Vinicius Niz  
Desenvolvimento de uma ferramenta computacional amigável para  
análise de dados de sequenciamento genômico de baixa cobertura e  
aplicação em estudo de genética de populações / Marcus Vinicius Niz  
Alvarez. -- Botucatu, 2024  
84 p. : il., tabs.  
  
Tese (doutorado) - Universidade Estadual Paulista (UNESP),  
Instituto de Biociências, Botucatu  
Orientador: Paulo Eduardo Martins Ribolla  
  
1. Genômica. 2. Bioinformática. 3. Software gratuito. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Universidade  
Estadual Paulista (UNESP), Instituto de Biociências, Botucatu. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

**CERTIFICADO DE APROVAÇÃO**

**TÍTULO DA TESE:** Desenvolvimento de uma ferramenta computacional amigável para análise de dados de sequenciamento genômico de baixa cobertura e aplicação em estudo de genética de populações

**AUTOR:** MARCUS VINICIUS NIZ ALVAREZ

**ORIENTADOR:** PAULO EDUARDO MARTINS RIBOLLA

Aprovado como parte das exigências para obtenção do Título de Doutor em Ciências Biológicas (Genética), pela Comissão Examinadora:

Prof. Dr. PAULO EDUARDO MARTINS RIBOLLA (Participação Presencial)  
Departamento de Biodiversidade e Bioestatística / Instituto de Biociências de Botucatu - UNESP

Prof. Dr. JOÃO PEDRO SOARES PINTO (Participação Virtual)  
Vector-borne diseases and pathogens / Instituto de Higiene e Medicina Tropical - Universidade Nova de Lisboa - Portugal

Prof. Dr. MARCELO URBANO FERREIRA (Participação Virtual)  
Departamento de Parasitologia / Institute for Biomedical Sciences - USP

Profa. Dra. MARIA ANICE MUREB SALLUM (Participação Virtual)  
Departamento de Epidemiologia / Faculdade de Saúde Pública - USP

Dr. GUILHERME TARGINO VALENTE (Participação Presencial)  
Departamento de Bioprocessos e Biotecnologia / Faculdade de Medicina de Botucatu - UNESP

Botucatu, 29 de julho de 2024



Documento assinado digitalmente

**FLAVIA DANIELI MARTINS GODINHO**

Data: 06/08/2024 11:09:38-0300

Verifique em <https://validar.iti.gov.br>

**FLÁVIA DANIELI MARTINS GODINHO**  
Assistente Técnico Administrativo I da Seção Técnica de Pós-Graduação  
Instituto de Biociências de Botucatu – UNESP

## DEDICATÓRIA

Dedico este trabalho à minha mãe Cecília e meu pai Raul (*in memoriam*) que sempre acreditaram em mim e me apoiaram até nos momentos mais difíceis. Não há palavras para expressar o tamanho do meu amor e eterna gratidão por vocês. *Pai, a vida te levou pouco antes deste momento, mas seu legado vive em mim, e é com determinação que carrego adiante seus ensinamentos e amor que permearam tua existência.*

À minha irmã Danielle Amanda que amo incondicionalmente. Você é a pessoa que mais confio neste mundo. O laço entre nós é o que me permite voar sem medo de cair. Sei que compartilharemos muitas risadas, conquistas e momentos juntos que estão por vir.

À minha amada Isabelle Tenori, que ilumina minha vida a cada dia que passo com você. Seu sorriso faz meu mundo ficar mais bonito. Juntos, me sinto invencível, completo e profundamente feliz. Meu coração sempre será teu. Nosso amor transcende a compreensão.

Aos meus avós Luzia, Maria, Milciades (*in memoriam*) e Raul (*in memoriam*). Em vocês, sempre encontrei sabedoria, apoio e aconchego. Cada lembrança com vocês é um tesouro inestimável que carregarei para sempre no meu coração.

A toda minha família, tios John e Milciades Jr., tias Ana e Fernanda, primos Ítalo e Igor e prima Lara. Sou grato por todo o carinho e suporte que vocês me proporcionam.

Aos meus amigos Filipe e Daniel que são como verdadeiros irmãos para mim. É como dizem: “Amizade verdadeira é raro encontrar, difícil de acabar e impossível de esquecer”.

Aos cães e gatos que fizeram ou fazem parte da minha vida. “Casa cheia de pelo, coração cheio de amor”.

## **AGRADECIMENTOS**

Agradeço ao programa “O Mundo de Beakman” e a TV Cultura por terem despertado o encanto pela ciência em mim, uma criança que estava apenas começando a entender as maravilhas ao seu redor.

Agradeço aos meus colegas de laboratório do grupo Pangene, em especial ao Diego pelas imensas contribuições acadêmicas ao longo desses anos, mas também pelas valiosas conversas sobre altos e baixos da vida, intercaladas com muita nostalgia e videogames. Agradeço ao Samir não apenas pelas colaborações, mas também pelo compartilhamento de suas experiências de vida e conselhos, que foram fundamentais para me ajudar a trilhar meu caminho. Um agradecimento especial a minha amiga Lorena que desde o mestrado trouxe muita cooperatividade, boas risadas e grandes desabafos nos momentos mais difíceis dessa nossa jornada acadêmica.

Agradeço ao meu orientador Prof. Dr. Paulo Ribolla, com todo meu respeito e gratidão. Obrigado por acreditar em mim, por ter compartilhado seu inestimável conhecimento comigo, mas também por ter me desafiado para que eu buscasse o meu próprio. Você é uma inspiração, mas também um grande amigo que sempre soube como iluminar meu caminho não só como pesquisador, mas como ser humano.

Agradeço imensamente a UNESP por ter sido minha casa nos últimos 11 anos. Sinto orgulho e o privilégio ter me formado e amadurecido aqui. Agradeço a todos os Professores e Servidores e que contribuíram para minha formação acadêmica e possibilitaram meu sonho de ser um cientista.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida processo 88887.817610/2023-00. CAPES sempre promovendo e tornando possível o desenvolvimento científico do nosso país.

## EPÍGRAFE

“Não são nossas habilidades que revelam quem realmente somos, são as nossas escolhas.”

Alvo Dumbledore

## RESUMO

### **Desenvolvimento de uma ferramenta computacional amigável para análise de dados de sequenciamento genômico de baixa cobertura e aplicação em estudo de genética de populações**

Os avanços nas tecnologias de sequenciamento genômico nos últimos anos têm reduzido consideravelmente o custo desta técnica. No entanto, projetos de larga escala continuam apresentando custo elevado, muitas vezes impraticável. Uma das alternativas é o sequenciamento genômico de baixa cobertura (lcWGS). Este estudo teve como objetivo avaliar o desempenho da genotipagem em diversos cenários diferentes usando um programa específico para dados de lcWGS que foi desenvolvido contendo *pipeline* completo, otimizado para a genotipagem por sequenciamento de baixa cobertura. Foram testados parâmetros como profundidade de sequenciamento e tamanho amostral para avaliação do desempenho da genotipagem desta ferramenta. Os resultados demonstraram que dados de lcWGS processados adequadamente tem confiabilidade satisfatória e grande potencial no processo de genotipagem em larga escala. Subsequentemente, a ferramenta desenvolvida intitulada LCSeqTools foi aplicada em um segundo estudo, no qual foram utilizadas amostras de *Nyssorhynchus darlingi* coletados no município de Mâncio Lima, estado do Acre, para detecção de inversões cromossômicas. O objetivo foi testar a capacidade e precisão de detecção de inversões cromossômicas utilizando dados de lcWGS, além de contribuir para melhor entendimento das inversões do genoma de *Ny. darlingi*, bem como a estrutura e dinâmica do genoma dessa espécie. Dez inversões cromossômicas foram detectadas no genoma de *Ny. darlingi* e a análise de genômica comparativa revelou nenhuma inversão em sintonia com inversões conhecidas de *Anopheles gambiae*.

**Palavras-chave:** Genômica; Bioinformática; Genética de populações; Inversões cromossômicas; SNPs; Mosquito;

## **ABSTRACT**

### ***Development of a user-friendly computational tool for analyzing low-coverage genomic sequencing data and application in population genetics studies***

*Advances in genomic sequencing technologies in recent years have significantly reduced the cost of this technique. However, large-scale projects still present high costs, often impractical. An alternative strategy is low-coverage whole-genome sequencing (lcWGS). This study aimed to evaluate the genotyping performance in various scenarios using a specific program for lcWGS data that was developed with a complete workflow optimized for low-coverage sequencing genotyping. Parameters such as sequencing depth and sample size were tested to evaluate the performance of this tool. The results demonstrated that properly processed lcWGS data have satisfactory genotyping reliability and great potential for large-scale genotyping studies. Subsequently, the developed tool, entitled LCSeqTools, was applied in a second study, in which samples of *Nyssorhynchus darlingi* collected in the municipality of Mâncio Lima, Acre state, were used to detect chromosome inversions. The objective was to test the capacity and accuracy of detecting chromosome inversions using lcWGS data, as well as to contribute to a better understanding of the inversions in the genome of *Ny. darlingi*, as well as the structure and dynamics of this species genome. Ten chromosome inversions were detected in the genome of *Ny. darlingi*, and comparative genomic analysis revealed no inversions in synteny with known inversions of *Anopheles gambiae*.*

**Keywords:** *Genomics; Bioinformatics; Population genetics; Chromosome inversions; SNPs; Mosquito;*

## LISTA DE ILUSTRAÇÕES

|                                                                                                                                              |    |
|----------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figura 1. Histórico do custo de sequenciamento de DNA ao longo dos anos.....                                                                 | 16 |
| Figura 2. Representação esquemática de estratégias de sequenciamento genômico.....                                                           | 17 |
| Figura 3. Representação esquemática do processo de imputação.....                                                                            | 18 |
| Figura 4. Descrição das etapas de interação com programas de computador por linhas de comando (esquerda) ou interface gráfica (direita)..... | 19 |
| Figura 5. Número de mortes por malária reportado em 2022.....                                                                                | 21 |
| Figura 6. Classificação de Risco dos municípios segundo a incidência de Malária.....                                                         | 22 |
| Figura 7. <i>Ny. darlingi</i> fêmea adulta.....                                                                                              | 23 |
| Figura 8. Representação esquemática de Inversões Cromossômicas.....                                                                          | 24 |
| Figura 9. Representação esquemática do cariótipo de <i>Ny. Darlingi</i> .....                                                                | 25 |

## CAPÍTULO 1

|                                                                     |    |
|---------------------------------------------------------------------|----|
| Figure 1. Contrast analysis for false positive SNP rate.....        | 35 |
| Figure 2. Contrast analysis for MAF percentage change.....          | 36 |
| Figure 3. Contrast analysis for Homozygosity percentage change..... | 37 |
| Figure 4. Contrast analysis for Genotype concordance rate.....      | 38 |

## CAPÍTULO 2

|                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------|----|
| Figure 1. Chromosome Inversion Detection Analysis for <i>Ny. darlingi</i> dataset.....                        | 55 |
| Figure 2. Inversion Genotyping Analysis.....                                                                  | 57 |
| Figure 3. Chromosome Inversion Detection Analysis for <i>An. gambiae</i> dataset.....                         | 58 |
| Figure 4. Genome Synteny Analysis for <i>Ny. darlingi</i> with <i>An. gambiae</i> and <i>An. albimanus</i> .. | 59 |
| Figure 5. Chromosome Inversion Synteny Analysis.....                                                          | 60 |

## APÊNDICE A

|                                                                                                                                                                  |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure A1. Line graph for false positive and true positive SNP rate (CM dataset). Horizontal facets (legend on the right corner) represent sequencing depth..... | 66 |
| Figure A2. Line graph for false positive and true positive SNP rate (BF dataset). Horizontal facets (legend on the right corner) represent sequencing depth..... | 67 |
| Figure A3. Q-Q Plot and histogram of residuals for false positive SNP rate.....                                                                                  | 68 |
| Figure A4. Boxplots for MAF percentage change (CM dataset).....                                                                                                  | 70 |

|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| Figure A5. Boxplot for MAF percentage change (BF dataset).....                         | 71 |
| Figure A6. Q-Q Plot and histogram of residuals for MAF percentage change.....          | 72 |
| Figure A7. Boxplot for homozygosity percentage change (CM dataset).....                | 73 |
| Figure A8. Boxplot for homozygosity percentage change (BF dataset).....                | 74 |
| Figure A9. Q-Q Plot and histogram of residuals for Homozigosity percentage change..... | 75 |
| Figure A10. Boxplot for genotype concordance rate (CM dataset).....                    | 77 |
| Figure A11. Boxplot for genotype concordance rate (BF dataset).....                    | 78 |
| Figure A12. Q-Q Plot and histogram of residuals for genotype concordance rate.....     | 81 |

## LISTA DE TABELAS

### CAPÍTULO 1

|                                                                                                                      |    |
|----------------------------------------------------------------------------------------------------------------------|----|
| Table 1. Samples from the Anopheles gambiae 1000 Genomes Consortium project genotype panel used for simulations..... | 32 |
|----------------------------------------------------------------------------------------------------------------------|----|

### CAPÍTULO 2

|                                                                                                            |    |
|------------------------------------------------------------------------------------------------------------|----|
| Table 1. Ny. darlingi chromosome inversion coordinate estimates based on Genome-wide association test..... | 56 |
|------------------------------------------------------------------------------------------------------------|----|

### APÊNDICE A

|                                                                                                                 |    |
|-----------------------------------------------------------------------------------------------------------------|----|
| Table A1. Linear regression estimates for false positive SNP rate.....                                          | 68 |
| Table A2. Linear regression estimates for false positive SNP rate, including MAF ranges coefficient.....        | 69 |
| Table A3. Marginal Means and contrasts estimates for false positive SNP rate.....                               | 69 |
| Table A4. Linear regression estimates for MAF percentage change.....                                            | 72 |
| Table A5. Marginal Means and contrasts estimates for MAF percentage change.....                                 | 72 |
| Table A6. Linear regression estimates for homozygosity percentage change.....                                   | 75 |
| Table A7. Marginal Means and contrasts estimates for Homozygosity percentage change.....                        | 76 |
| Table A8. Linear regression estimates for Genotype concordance rate (Minor Allele Homozygotes).....             | 79 |
| Table A9. Linear regression estimates for Genotype concordance rate (Heterozygotes).....                        | 79 |
| Table A10. Linear regression estimates for Genotype concordance rate. (Major Allele Homozygotes).....           | 80 |
| Table A11. Marginal Means and contrasts estimates for Genotype concordance rate (Minor Allele Homozygotes)..... | 82 |
| Table A12. Marginal Means and contrasts estimates for Genotype concordance rate (Heterozygotes).....            | 83 |
| Table A13. Marginal Means and contrasts estimates for Genotype concordance rate (Major Allele Homozygotes)..... | 84 |
| Table A14. Summary of models used in the linear regression analyses.....                                        | 85 |

## LISTA DE SÍMBOLOS

bp: Pares de base. Exemplo: 1kbp equivale a 1.000 pares de base.

p-value: Nível descritivo do teste estatístico.

Q1: Primeiro quartil (ou 25º percentil).

Q3: Terceiro quartil (ou 75º percentil).

Q-Q: Quantil-Quantil.

$r^2$ : Coeficiente de correlação do desequilíbrio de ligação.

$R^2$ : Coeficiente de determinação.

$\rho$ : Coeficiente de correlação de Spearman.

## SUMÁRIO

|                                                                                                 |    |
|-------------------------------------------------------------------------------------------------|----|
| 1. INTRODUÇÃO GERAL.....                                                                        | 16 |
| 1.1. Sequenciamento Genômico de Baixa Cobertura: Aplicações e Desafios.....                     | 16 |
| 1.2. Bioinformática Amigável: Facilitando o Acesso à Análise Genética para Usuários....         | 19 |
| 1.3. Malária no Brasil.....                                                                     | 20 |
| 1.4. Principal vetor de malária no Brasil: <i>Nyssorhynchus darlingi</i> .....                  | 22 |
| 1.5. Inversões cromossômicas no genoma de <i>Ny. darlingi</i> .....                             | 24 |
| 2. OBJETIVOS.....                                                                               | 27 |
| CAPÍTULO 1 – LCSeqTools – A user friendly tool for cost-effective genotyping by sequencing..... | 28 |
| Abstract.....                                                                                   | 28 |
| Background.....                                                                                 | 29 |
| Methods.....                                                                                    | 30 |
| Results.....                                                                                    | 34 |
| Discussion.....                                                                                 | 39 |
| Conclusion.....                                                                                 | 41 |
| References.....                                                                                 | 41 |
| CAPÍTULO 2 – Chromosome Inversion detection using low-coverage sequencing data (lcWGS).....     | 44 |
| Abstract.....                                                                                   | 44 |
| Background.....                                                                                 | 45 |
| Methods.....                                                                                    | 46 |
| Results.....                                                                                    | 49 |
| Discussion.....                                                                                 | 50 |
| Conclusion.....                                                                                 | 51 |
| References.....                                                                                 | 52 |
| Figures and tables.....                                                                         | 55 |
| 3. CONSIDERAÇÕES FINAIS.....                                                                    | 61 |

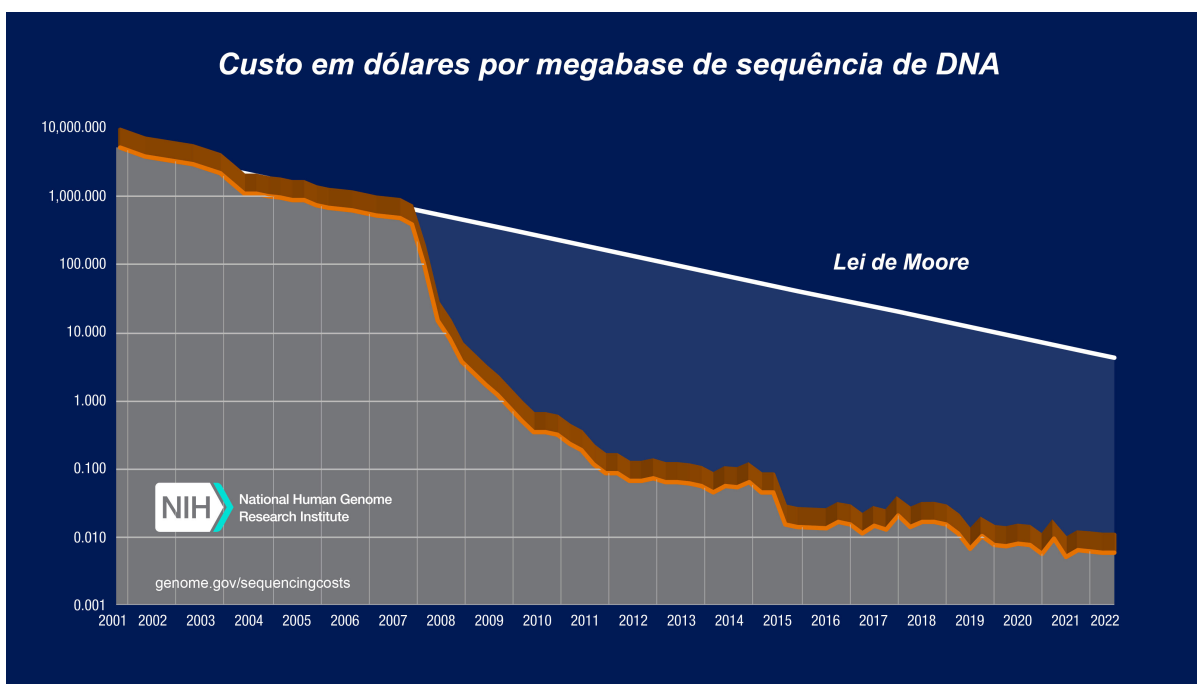
|                                                      |    |
|------------------------------------------------------|----|
| REFERÊNCIAS.....                                     | 62 |
| GLOSSÁRIO.....                                       | 64 |
| APÊNDICE A – Material Suplementar do Capítulo 1..... | 65 |

## 1. INTRODUÇÃO GERAL

### 1.1. Sequenciamento Genômico de Baixa Cobertura: Aplicações e Desafios

Muitos estudos envolvendo sequenciamento genômico tem sido conduzidos nos últimos anos. Com o desenvolvimento das tecnologias de sequenciamento de genoma tem resultado em reduções consideráveis no custo da técnica (Figura 1). No entanto, projetos em larga escala que demandam por grandes quantidades de amostras sequenciadas continuam apresentando custo elevado, muitas vezes, impraticáveis. Por conta disso, alternativas têm sido adotadas para que o uso de dados genômicos em larga escala seja viável, auxiliando no entendimento da estrutura genética de populações.

**Figura 1.** Histórico do custo de sequenciamento de DNA ao longo dos anos.



Fonte: WETTERSTRAND. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program.  
Disponível em: <[www.genome.gov/sequencingcostsdata](http://www.genome.gov/sequencingcostsdata)>. Acesso: 28 mai. 2024. (Adaptado).

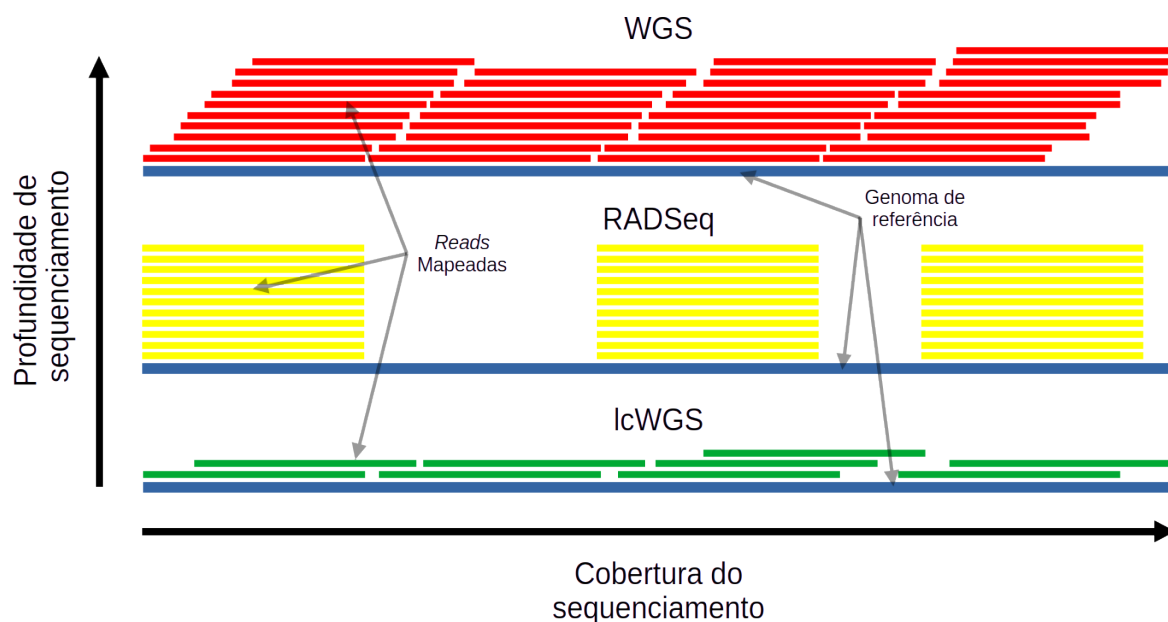
Existem diversas estratégias de sequenciamento do genoma atualmente, por exemplo: sequenciamento de genoma completo (do inglês, *whole genome sequencing*, WGS); sequenciamento de fragmentos de DNA associado a sítios de restrição (do inglês, *Restriction site associated DNA sequencing*, RADseq). Uma das alternativas de baixo custo para sequenciamento em larga escala é o sequenciamento de baixa cobertura do genoma completo

(do inglês, *low-coverage whole genome sequencing*, lcWGS).

Há uma ambiguidade no uso do termo cobertura e profundidade de sequenciamento na literatura. No presente estudo, foi utilizado a padronização dos termos de acordo com a Figura 2, na qual as leituras (do inglês, *reads*) estão representadas em cores de acordo com o tipo de sequenciamento. O termo profundidade de sequenciamento refere-se a quantidade de vezes que uma região do genoma foi lida no processo de sequenciamento e o termo cobertura do sequenciamento refere-se a porcentagem do genoma que foi efetivamente sequenciado, independente do número de vezes.

Quanto menor a profundidade de sequenciamento, menor é a confiabilidade do genótipo identificado (SIMS et al., 2014; RUBINACCI et al., 2020). No WGS tradicional, o processo de sequenciamento resultará em amostras com alta profundidade e alta cobertura. No entanto, o volume de DNA total sequenciado será elevado, aumentando consideravelmente o custo da corrida de sequenciamento, além de outros fatores como número de amostras e o tamanho do genoma da espécie. O RADSeq resultará em um sequenciamento com alta profundidade, mas a cobertura é significativamente menor, consequentemente a densidade de polimorfismos detectados será muito menor.

**Figura 2.** Representação esquemática de estratégias de sequenciamento genômico.

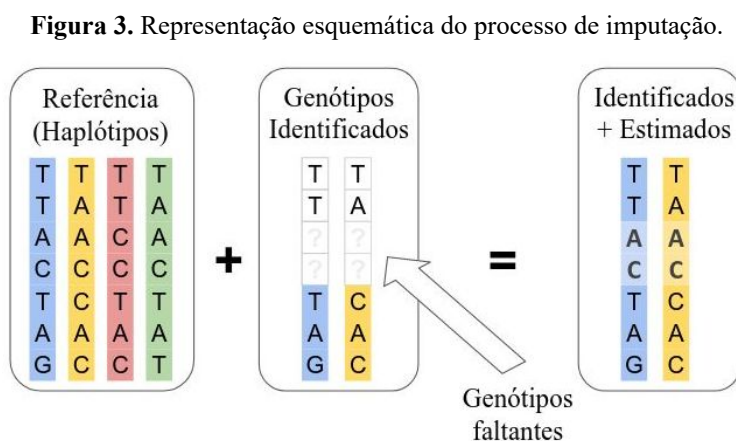


Fonte: Imagem gerada pelo autor.

O lcWGS aliado a técnica de imputação de genótipos confere informações genômicas suficientes para seleção de marcadores com confiabilidade aceitável e custo reduzido

(GORJANC et al., 2017; RUBINACCI et al., 2020). A acurácia na detecção de variantes é reduzida em sequenciamento genômico de baixa cobertura, tendendo 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 et al., 2014; RUBINACCI et al., 2020).

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, os quais se baseiam em recuperar informação de genótipos faltantes utilizando estimativas de haplótipos da população (Figura 3), etapa conhecida como faseamento (do inglês, *Phasing*).



Fonte: Imagem gerada pelo Autor.

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 substancial no custo quando comparada com o WGS padrão (PASANIUC et al., 2012; RUSTAGI et al., 2017)

Li et al. (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.

O presente estudo desenvolveu um programa com interface amigável aos usuários que executa um *pipeline* completo de genotipagem por sequenciamento, tratamento de dados e

imputação otimizados para dados de sequenciamento genômico de baixa cobertura, formando um kit de ferramentas completo e de fácil utilização que contribuirá com o desenvolvimento de futuros projetos que dependam de alternativas de custo reduzido para sequenciamento em larga escala.

## 1.2. Bioinformática Amigável: Facilitando o Acesso à Análise Genética para Usuários

Diversos programas de bioinformática, inclusive para análise de dados de sequenciamento, têm sido desenvolvidos ao longo dos últimos anos. Repositórios de código fonte e binários como GitHub, Sourceforge e LaunchPad, disponibilizam diversos programas e *workflows* para diferentes aplicações bioinformáticas. Na maioria dos casos, são programas desenvolvidos para serem executados em interfaces de linhas de comandos. Embora poderosas, esse tipo de interface não é tão amigável para usuários sem experiência em linguagens de programação, especialmente se comparadas às interfaces gráficas que são projetadas para serem mais acessíveis e convidativas aos usuários.

**Figura 4.** Descrição das etapas de interação com programas de computador por linhas de comando (esquerda) ou interface gráfica (direita).



Fonte: Imagem gerada pelo Autor.

Programas como QIAGEN CLC Genomics Workbench, Geneious e Unipro UGENE são importantes por disponibilizarem um conjunto de ferramentas poderosas para lidar com grande volume de dados biológicos e, ao mesmo tempo, fornecer uma interface amigável que pode economizar tempo e potencializar a acessibilidade da ferramenta por pesquisadores de laboratório experimental sem experiência em interfaces de linha de comando, por exemplo. A

implementação de uma interface gráfica amigável aos usuários é extremamente importante para aproximação da ferramenta e dos usuários sem experiência em linhas de comando. Quanto mais intuitiva e amigável for a interface do programa de computador, maior o potencial para cativar novos usuários.

Existem diversas formas e com diferentes possibilidades para implementar interface gráfica em programas de computador. Uma das formas é através do *framework* Qt (disponível em <https://www.qt.io/>). Um *framework* é um conjunto de códigos de vários projetos que promovem uma funcionalidade genérica, permitindo ao desenvolvedor dedicar seus esforços na resolução de problemas específicos do programa em vez de reescrever códigos comuns para funcionalidades genéricas.

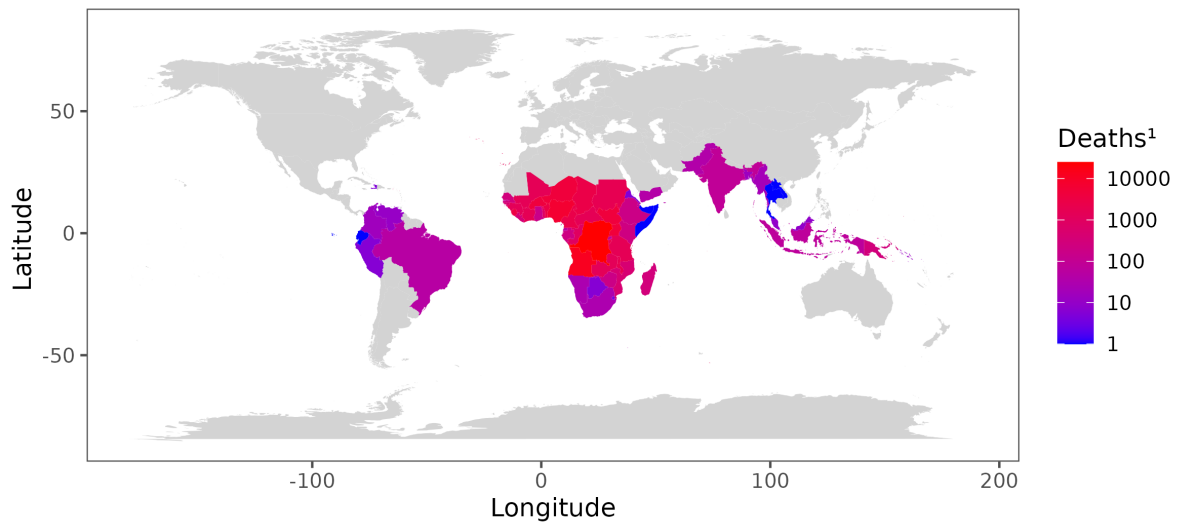
O Qt é um *framework* para desenvolvimento de *software* em linguagem C++, permitindo inclusive o desenvolvimento em múltiplas plataformas como Linux, Windows, macOS, Android, entre outras. Além disso, o Qt é disponibilizado através de licenças gratuitas como GPL e LGPLv3 (disponível em <https://www.qt.io/licensing/>), sendo ideal para casos de projetos com distribuição de código aberto com propósito acadêmico/estudantil.

Com base no descrito, o projeto teve como objetivo desenvolver um programa utilizando o *framework* Qt, implementando uma interface gráfica amigável aos usuários que facilitará a execução do *pipeline* completo para genotipagem por sequenciamento utilizando dados de sequenciamento genômico de baixa cobertura, resultando em um arquivo do tipo VCF (do inglês, *Variant call format*), o qual poderá ser utilizado em diversos programas de análises estatísticas e genômica populacional, inclusive programas que também possuem interface amigável aos usuários.

### **1.3. Malária no Brasil**

A malária é considerada a enfermidade transmitida por artrópode mais impactante em países em desenvolvimento. De acordo com *World Malaria Reports* (2023), a estimativa de casos no mundo é de 249 milhões de casos de malária e 608 mil mortes em 2022, aproximadamente 94% concentradas na África (Figura 5).

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

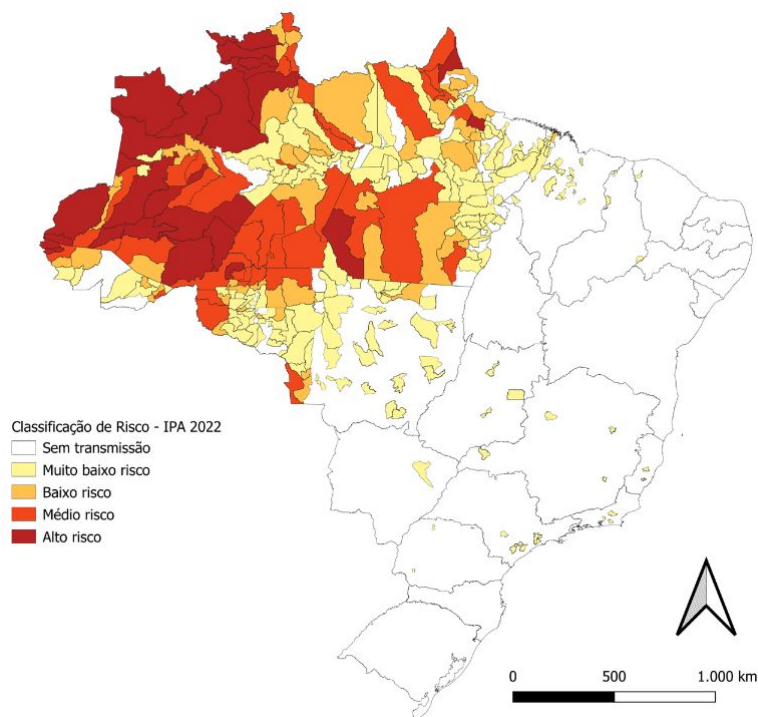
**Figura 5.** Número de mortes por malária reportado em 2022

Fonte: Imagem gerada pelo autor, dados conforme *World Malaria Reports* (WHO, 2023)

O Brasil é um país com alta incidência de malária, foram 131 mil casos registrados em 2022 segundo o Boletim Epidemiológico da Secretaria de Vigilância em Saúde de 2024 (BRASIL, 2024). Cerca de 99,9% dos casos registrados estão concentrados na região amazônica (Figura 6).

Essa doença é caracterizada por desencadear acessos periódicos de febres intensas que debilitam profundamente o indivíduo infectado. 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 cinco espécies e duas subespécies dentro do gênero *Plasmodium* spp que podem causar a malária humana: *Plasmodium vivax*, *P. falciparum*, *P. malariae*, *P. ovale curtisi*, *P. ovale wallikeri* e *P. knowlesi* (SU, 2010).

**Figura 6.** Classificação de Risco dos municípios segundo a incidência de Malária.



Fonte: BRASIL, Boletim Epidemiológico nº55. 2024. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-contenido/publicacoes/boletins/epidemiologicos/edicoes/2024/boletim-epidemiologico-volume-55-no-01/view>. Acesso em: 28 mai. 2024

As principais diferenças entre esses agentes patogênicos são 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, 2023). Aproximadamente 500 espécies de anofelinos já foram identificadas, 70 são vetores de parasitos e destes, cerca de 20 são importantes transmissores da malária ao homem (SERVICE, 2008).

#### 1.4. Principal vetor de malária no Brasil: *Nyssorhynchus darlingi*

*Nyssorhynchus darlingi* é o principal vetor da malária no Brasil, altamente suscetível aos plasmódios humanos e, capaz de transmitir a doença dentro e fora das moradias, mesmo em baixa densidade (Figura 7). Os criadouros deste mosquito são caracteristicamente representados por coleções de águas límpidas, com certa profundidade, sombreadas, dotadas de vegetação pobre em sais e matéria orgânica (FORATTINI, 2002). Além disso, é notavelmente antropofílico, pois na medida em que o ambiente natural se transforma em antrópico, ou desmatado, a população local de *Ny. 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, 1990).

**Figura 7.** *Ny. darlingi* fêmea adulta



Fonte: James Gathany | CDC. Disponível em: <<https://agencia.fapesp.br/discovery-paves-the-way-for-blocking-malaria-transmission-in-brazil/31979>>. Acesso em: 28 de mai. 2024

Na Amazônia, é o vetor que melhor e mais rapidamente se beneficia das alterações causadas pelo ser humano no ambiente silvestre (CONSOLI & LOURENÇO-DE-OLIVEIRA, 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-DASILVA et al., 2019). O uso de inseticida apresenta periculosidade à saúde de habitantes locais e de funcionários que o manejam, além de causar seleção de indivíduos resistentes (VEZENEGHO et al., 2009). Sendo assim, o estudo do comportamento e dinâmica biológica do vetor, sua dispersão e interação com humanos é de grande importância para a entomologia médica e para a saúde pública.

Até o presente momento, poucos estudos foram conduzidos utilizando dados de sequenciamento genômico para *Ny. darlingi*, especialmente WGS. Alvarez et al. (2022) demonstraram que utilizando dados de lcWGS associados à imputação, foi possível identificar SNPs estatisticamente significativos associados ao comportamento de picada do mosquito, além de demonstrar sinais de estratificação populacional em escala microgeográfica.

Mais estudos que utilizem WGS são necessários para caracterizar a nível molecular o processo de adaptação e evolução do mosquito a fim de estabelecer estratégias de alta precisão de controle vetorial, bem como rastrear as mudanças na dinâmica populacional para processos de intervenção. Além disso, uma das vantagens de se utilizar estratégias de

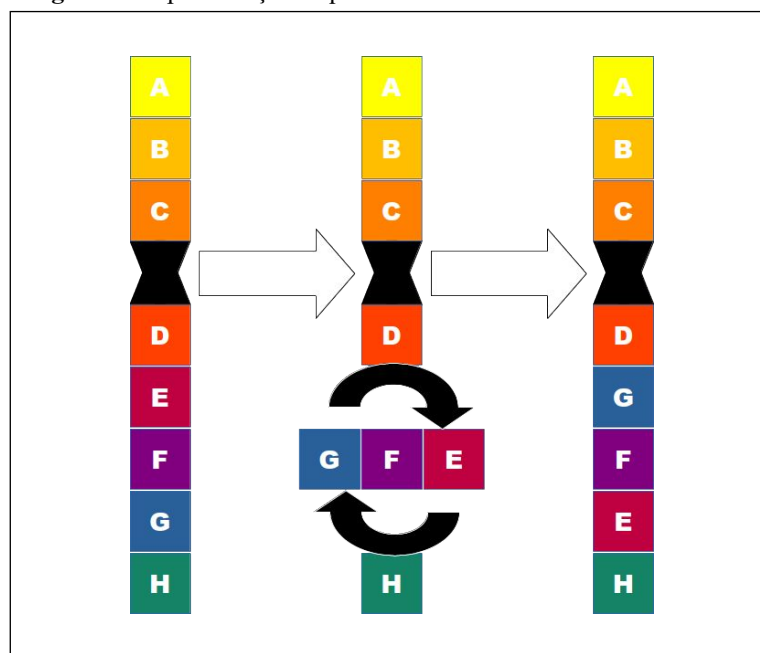
sequenciamento genômico completo é a possibilidade de explorar os dados em busca de testar múltiplas hipóteses biológicas simultaneamente.

### 1.5. Inversões cromossômicas no genoma de *Ny. darlingi*

Inversões cromossômicas são variantes estruturais que ocorrem quando um segmento do cromossomo rotaciona 180° em sua posição cromossômica original (Figura 8). Inversões podem apresentar tamanhos variados, podendo alcançar uma parte substancial de um braço cromossômico, bem como podem ter cerca de apenas alguns milhares de pares de base de comprimento. Regiões adjacentes aos pontos de quebra das inversões cromossômicas apresentam taxa de recombinação fraca ou praticamente nula, sendo regiões particularmente valiosas e importantes para inferir o histórico evolutivo das inversões (CORBETT-DETIG et al, 2019).

São componentes fundamentais no processo de evolução, por exemplo, em *Anopheles gambiae* desempenham importante papel no processo de adaptação a diferentes condições ambientais (CORBETT-DETIG et al., 2019; AYALA et al., 2019), bem como associadas ao tamanho do corpo, resistência ao estresse térmico e de dessecação, consistentes com um papel importante das inversões na tolerância do mosquito à aridez (CHENG et al., 2018).

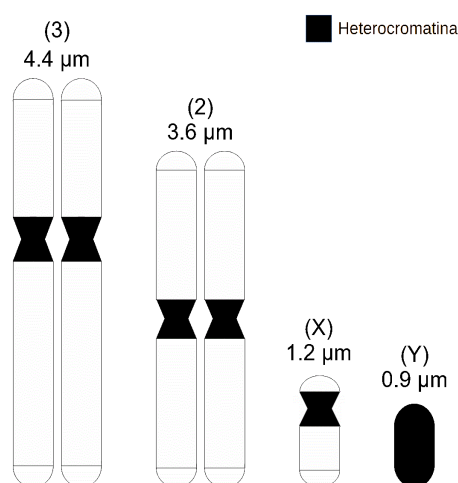
**Figura 8.** Representação esquemática de Inversões Cromossômicas.



Fonte: Imagem gerada pelo Autor.

O genoma de *Ny. darlingi* apresenta uma composição cromossômica de  $2n=6$ , constituído por dois pares de cromossomos autossômicos e um par de cromossomos sexuais (Figura 9). Dentre os cromossomos autossômicos, 2 é metacêntrico e 3 é submetacêntrico. Os cromossomos sexuais são X e Y, sendo acrocêntrico e puntiforme (totalmente heterocromático), respectivamente (RAFAEL & TADEI, 1998; RAFAEL & TADEI, 2000).

**Figura 9.** Representação esquemática do cariótipo de *Ny. darlingi*



Fonte: Imagem gerada pelo Autor, dados conforme Rafael & Tadei (2000)

Em *Ny. darlingi*, já foram descritas inversões em cromossomos politênicos de glândulas salivares de mosquitos do sudeste do Brasil e região amazônica, sendo inversões localizadas nos cromossomos X, braços 2L e 2R do cromossomo 2 e braços 3R e 3L do cromossomo 3 (KREUTZER et al., 1972; TADEI & SANTOS, 1982; RAFAEL et al., 2010; CORNEL et al., 2016).

É possível observar sinais de inversões cromossômicas utilizando dados de SNPs oriundos de sequenciamento genômico completo. Múltiplas abordagens já foram descritas na literatura, possibilitando utilizar diferentes parâmetros genômicos para tal processo (CÁCERES & GONZÁLEZ, 2015). Por exemplo, é possível identificar alterações no padrão de desequilíbrio de ligação entre SNPs por da taxa de recombinação fraca em regiões adjacentes ao ponto de quebra das inversões. Outra maneira utilizável é a redução de dimensões por análise de escalonamento multidimensional (MDS) ou análise de componentes principais (PCA). Ao aplicar esses métodos de redução de dimensões, a variabilidade genética causada pela fixação de alelos por conta do aumento do desequilíbrio de ligação nas regiões das inversões cromossômicas pode ser identificadas através da formação de agrupamentos genéticos ou pela inspeção dos dos autovetores utilizados na transformação dos dados no

PCA, onde é possível relacionar a importância de cada SNP com cada componente principal avaliado.

Até o presente momento, não há estudos que descrevem inversões cromossômicas para a espécie *Ny. darlingi* a nível molecular utilizando tecnologia de sequenciamento genômico completo.

## 2. OBJETIVOS

O objetivo deste projeto foi desenvolver um programa intuitivo e autônomo para a genotipagem de dados de lcWGS, avaliar sua eficiência por meio de simulações computacionais, e aplicar o método desenvolvido a estudos de genômica populacional e comparativa utilizando amostras de *Ny. darlingi* coletadas no município Mâncio Lima no estado do Acre.

## CAPÍTULO 1 – LCSeqTools – A user friendly tool for cost-effective genotyping by sequencing.

**Authors:** Marcus Vinicius Niz Alvarez, Diego Peres Alonso, Paulo Eduardo Martins Ribolla

---

### Abstract

**Introduction:** Over the past decade, low-coverage whole-genome sequencing (lcWGS) has been used as an alternative approach for efficient and cost-effective genotyping strategy. The combination of lcWGS and genotype imputation is a known method for improving data accuracy. We proposed in this study a user-friendly genotyping tool that performs a full genotyping by sequencing pipeline optimized for lcWGS data and does not require haplotype reference panels. This is mainly important for non-model organisms that lack reference data. We also tested the pipeline in different scenarios and compared the genotyping performance with traditional WGS data. **Methods:** We generated *in silico* lcWGS sequencing data using a high coverage publicly available data of *Anopheles gambiae* samples as reference. Different levels of sequencing depth {0.5,1,2,3,4,5,10} and sample size {10,25,50,100,200} were tested. We tested three combinations: traditional genotyping pipeline without imputation and NextSeq500 platform (REF), LCSeqTools pipeline with NextSeq500 (NS5) and LCSeqTools with NovaSeq6000 (NS6). We estimated parameters for each simulation iteration as a benchmark: False positive SNP rate, MAF percentage change, individual homozygosity percentage change and genotype concordance rate. **Results:** Both N5 and N6 pipeline combinations showed significant lower false positive SNPs, lower change in allele frequencies estimates and higher genotype concordance rate in all low coverage scenarios when compared to REF. Lower or equal homozygosity percentage change was also

---

observed, but only at 4x and 5x sequencing depth, for lower depth scenarios, the homozygosity change was greater than REF, but sample size also affects it. **Conclusion:** LCSeqTools pipeline significantly improves the genotyping performance of lcWGS sequencing data and provides an alternative reliable and cost-effective strategy for low budget projects aiming population genomics studies using biallelic SNPs.

**Keywords:** Bioinformatics, Genomics, Biostatistics, Genotyping, Cost-effectiveness

## Background

Population genomics studies over the past decade have used low-coverage whole-genome sequencing (lcWGS) as an alternative approach for efficient and cost-effective genotyping by sequencing. Despite the significant cost reduction of whole-genome sequencing over the last few years (WETTERSTRAND, K.A., 2022; PRESTON, J. et al, 2021), the sequencing cost per base pair is still expensive for some applications, especially for studies demanding a high number of samples or species with large genomes. lcWGS is an alternative approach, in which by sacrificing sequencing depth one can sequence a massive number of samples with the same budget. However, the major lcWGS drawback caused by the reduction in sequencing depth is the lower genotyping reliability, given that the probability of identifying heterozygous *loci* in diploid individuals is lower and the probability of assuming false positives as genuine variants is higher. (NIELSEN, R. et al, 2011).

lcWGS combined with genotype imputation has shown great results to improve genotyping accuracy, as imputation algorithms have been improved in recent years (HUI, R. et al, 2020), surpassing even SNP arrays in terms of performance and accuracy in some scenarios (DAVIES, R.W. et al. 2021; RUBINACCI, S. et al. 2021). Nevertheless, factors such as sample size and average sequencing depth cause significant changes on imputation performance, therefore, they are important variables to be considered in the experimental design stage (LOU, R.N. et al, 2021).

Although the lcWGS-based genotyping accuracy is lower when compared to traditional WGS, many population level parameters can still be inferred with good reliability from the analyzed sample, such as allele frequency and homozygosity. lcWGS combined with imputation has shown to be a highly efficient approach by presenting reliable estimates of population genetic parameters (LOU, R.N. et al, 2021) and also greater statistical power in genome-wide association studies when compared to SNP array genotyping method (ZHANG, Z. et al, 2022; DENG, T. et al, 2021). This is especially important for genome-wide association studies, which often need large sample sizes to obtain sufficient statistical power. Therefore, the lcWGS can be a great alternative for projects that depend on large sample sizes, but reduced budget, such as epidemiological studies.

Currently, several model organisms have haplotype reference panels that are useful to improve imputation performance. However, many other organisms still lack any type of genomic data. In these cases, it is necessary to use an approach in which the haplotypes present in the sample itself are used as a reference in the imputation stage. The goal of this study is to provide a user-friendly tool that executes a complete genotyping pipeline with customizable parameters for lcWGS data of non-model organisms. We also evaluate the genotyping performance under different scenarios of sequencing depth and sample sizes using simulated lcWGS data.

## **Methods**

### **Genotyping by sequencing pipeline**

The LCSeqTools pipeline consists of six main steps: Sequence quality control, sequence alignment, variant calling, pre-imputation variant filtering, imputation and post-imputation variant filtering. The only required input files are the raw sequencing files in modern FASTQ format and a reference genome in FASTA format. Sequencing quality statistics are estimated using FASTQC v0.11.9 (ANDREWS, S., 2010) and displayed with the program graphical interface. The sequence trimming and filtering is performed by

Trimmomatic v0.39 (BOLGER, A. M., LOHSE, M., & USADEL, B., 2014) using predefined or custom parameters by the user. The pre-processed sequences are mapped to the reference genome using Burrows-Wheeler Aligner v0.7.17 program (LI, H., DURBIN, R., 2009). The properly mapped sequences are submitted to biallelic-only SNP calling using SAMtools v1.10 and BCFtools v1.10 programs (LI, H., 2011). Pre-imputation variant filtering is performed by LCSeqTools through the integration of LCVCFtools v1.0.2 (ALVAREZ, M.V.N., 2021; ALVAREZ, M.V.N., 2022). The pre-imputation variant filtering method was implemented specifically for low coverage data. Genotypes were omitted according to genotype quality and sequencing depth parameters, keeping only genotype likelihood data used as *a priori* probability in the imputation process. SNP are removed by allele-depth-based MAF estimate and missing data rate per sample and per SNP. Missing genotypes are imputed using the BEAGLE v4.1 (BROWNING, R., BROWNING, B. L., 2016) with GTGL method. An optional post-imputation variant filtering is performed using BCFtools v1.10 program, removing genotypes according to *a posteriori* probability threshold.

### Genotyping performance evaluation

The pipeline genotyping performance was evaluated through low-coverage whole genome sequencing simulations under different depths and sample sizes scenarios. Public genome-wide genotyping data of *Anopheles gambiae* mosquito was used as a reference to generate *in silico* low-coverage sequencing data. The simulation used a high coverage genome-wide variants panel of the mosquito *An. gambiae* from the first phase of the "The *Anopheles gambiae* 1000 Genomes Consortium" (ANOPHELES GAMBIAE 1000 GENOMES CONSORTIUM et al, 2017) available at EMBL-EBI (ERZ373588). Genotypes from two independent sample groups in the *An. gambiae* data were used as reference in the sequencing simulations. Two datasets were chosen using the sample size criteria, selecting the datasets with the biggest number of contemporaneous individuals, therefore, samples from the Republic of Cameroon (CM) and Burkina Faso (BF), represented in Table 1.

**Table 1.** Samples from the *Anopheles gambiae* 1000 Genomes Consortium project genotype panel used for simulations.

| Country                      | Region        | Sex | N  | Total |
|------------------------------|---------------|-----|----|-------|
| Republic of Cameroon<br>(CM) | Daiguene      | F   | 69 | 232   |
|                              | Gado-Badzere  | F   | 55 |       |
|                              | Mayos         | F   | 87 |       |
|                              | Zembe-Borongo | F   | 21 |       |
| Burkina Faso (BF)            | Bana          | F   | 37 | 127   |
|                              | Pala          | F   | 44 |       |
|                              | Sourukoudinga | F   | 46 |       |

F: Female. N: Region sample size. Total: Country sample size

The *in silico* sample sequencing was generated using the `art_illumina` program from ART software package (HUANG, W. et al., 2012). Illumina NextSeq500 and NovaSeq6000 real sequencing data were used to calibrate `art_illumina` to reproduce sequencing data from these equipment. To calibrate the NextSeq500 quality profile and read error probabilities, publicly available sequencing data from the National Center for Biotechnology Information – NCBI (SAYERS, E.W. et al, 2023) Sequence Read Archive (SRA), identified as Bioproject PRJNA683015, was used. To calibrate the NovaSeq6000 quality profile and reading error probabilities, publicly available sequencing data from the NCBI SRA, identified as Bioproject PRJNA744883, was used. The genome reference used is available at NCBI as AgamP3 from *An. gambiae* (GCF\_000005575.2), the reference used in the “*Anopheles gambiae* 1000 Genomes” project. The Y chromosome, mitochondrial genome and unplaced scaffold genomic regions were discarded for the analysis, using only the 2L, 2R, 3L, 3R and X chromosomes.

The tested parameters were: genotyping method, average sequencing depth and sample size. For genotyping methods, three different pipelines were used: NextSeq500 sequencing data with standard variant calling pipeline without imputation as reference group (REF), NovaSeq6000 sequencing data with default LCSeqTools variant calling pipeline (NS6), NextSeq500 sequencing data with default LCSeqTools variant calling pipeline (NS5). The average sequencing depth (DP) levels generated were:  $DP=\{0.5,1,2,3,4,5,10\}$ . DP=0.5 level was considered as extremely-low coverage and DP=10 as high coverage, both were used only for descriptive analysis. The sample sizes (N) used were:  $N=\{10,25,50,100,200\}$ . Each simulation iteration selects random unique individual genotyping data from the reference.

## Statistical Analysis

Descriptive and inferential analysis were performed using R v4.3.2 with RStudio Server “Ocean Storm” (R CORE TEAM, 2013; TEAM RSTUDIO et al, 2015). Data processing and plotting were performed using packages included in *tidyverse* v2.0.0 meta-package (WICKHAM, H. et al., 2019).

Statistical parameters were calculated for each chromosomal region (2L, 2R, 3R, 3L and X), comparing the resulting genotypes with the reference panel for each simulation iteration. The estimated statistical parameters for descriptive and inferential analyzes were: false positive SNP rate, MAF percentage change, individual homozygosity percentage change and genotype concordance rate. Estimates were summarized in quartiles between chromosomal regions for the parameters MAF percent change, homozygosity percent change, and genotype concordance rate.

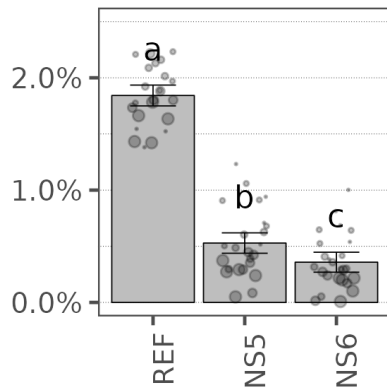
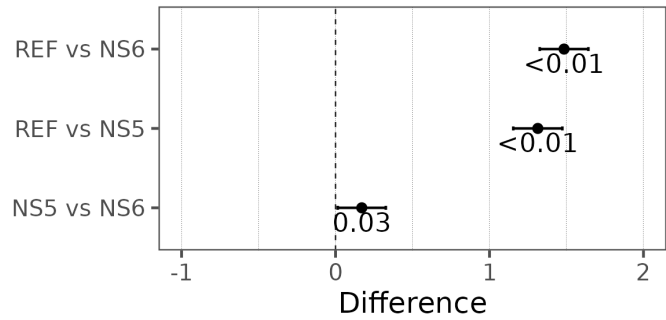
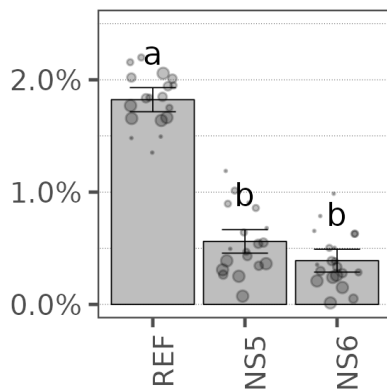
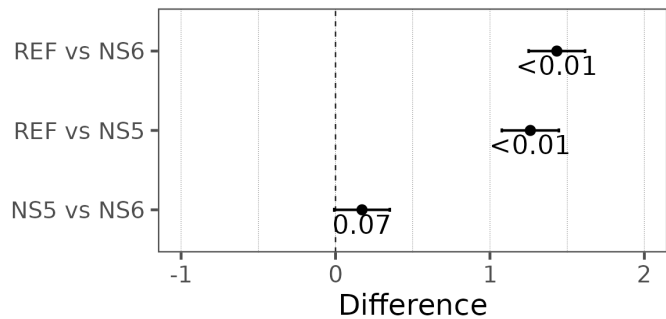
Inferential analysis was performed through statistical tests for all parameters, using median estimates across the chromosomes or sum across chromosomes for false positive SNP rate. A multiple linear regression analysis was performed and marginal means estimate used for group comparison. The predictor variable selection was based on a stepwise regression approach (coefficient estimate  $\neq 0$ , p-value  $\leq 0.05$ ). The inclusion of predictor variables was performed starting with the null model (no variable), adding the variables one by one, including interactions. Model comparison was performed using the Akaike Information Criterion (AIC) and F test. Atypical observations (outliers) were removed when identified by the interquartile range rule, given that if  $x < Q1 - 1.5 \times IQR$  or  $x > Q3 + 1.5 \times IQR$ ,  $x$  is removed from the respective data set. Residuals of the models were evaluated for normality by the Kolmogorov-Smirnov test with Lilliefors correction and homoscedasticity by the Levene test. Data were submitted to logarithmic transformation to reduce deviation from normality when necessary. Comparisons were performed by contrasting the estimates of marginal means. Statistically significant differences were considered when the adjusted p-value for multiple comparisons by the Bonferroni method was  $\leq 0.05$ .

Genotyping performance was evaluated using the following criteria: lowest rate of false positive SNP, lowest percentage variation in MAF, lowest percentage variation in

homozygosity, and highest rate of genotype agreement.

## Results

The genotyping method and sample size parameters showed statistically significant estimates in the linear regression analysis (Table A1) for the false positive SNPs rate, indicating significantly lower false positives in NS5 and NS6 data than REF data (Figure 1). The negative sample size coefficient indicates that the larger the sample size, the lower false positive SNPs rate. A separate model including false positive SNP by MAF range indicated that the higher the frequency of the MAF, the lower the rate of false positive SNPs (Table A2, figure A1 and figure A2).

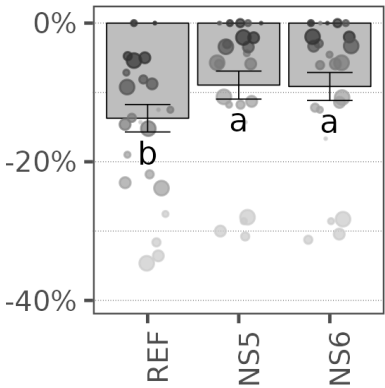
**Estimated Marginal Means (CM)****Marginal contrasts (CM)****Estimated Marginal Means (BF)****Marginal contrasts (BF)**

Sample Size ☐ 10 ☐ 25 ☐ 50 ☐ 100 ☒ 200

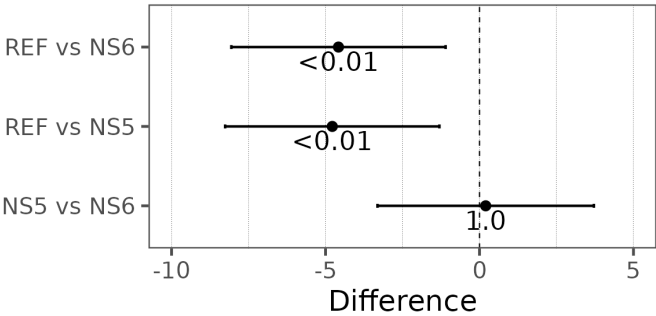
Figure 1. Contrast analysis for false positive SNP rate. Points represent individual data. Error bars indicate the 95% confidence intervals. Different letters represent significant differences between groups (CLD). Values below the horizontal lines indicate Bonferroni-corrected contrasts p-values.

All tested parameters showed statistically significant estimates in the linear regression analysis (Table A4) for MAF percentage change. The allele frequency percent change was significantly lower in the NS5 and NS6 data than in the REF data (Figure 2). As the sequencing depth increases, the MAF percentage change becomes smaller and approaches zero. This is shown by the positive sequencing depth coefficient. Sample size coefficient indicates that there is a slight negative effect on MAF percentage change. However, it is possible to observe in figures A4 and A5 that, the larger the sample size, the greater the correlation of allele frequencies.

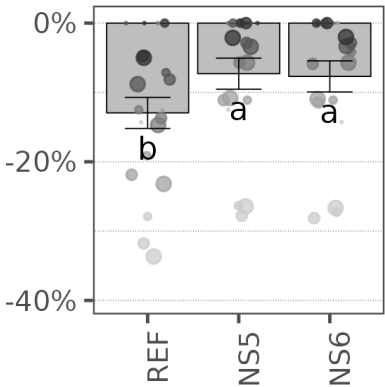
Estimated Marginal Means (CM)



Marginal contrasts (CM)



Estimated Marginal Means (BF)



Marginal contrasts (BF)

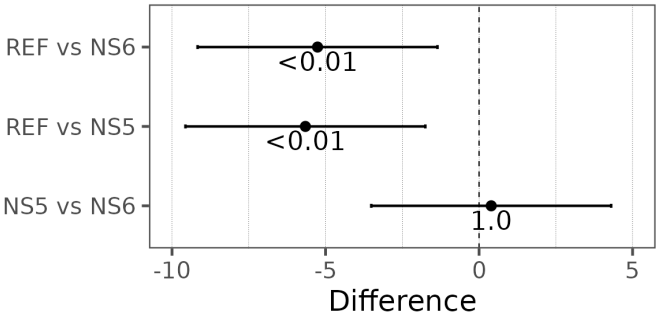


Figure 2. Contrast analysis for MAF percentage change. Points represent individual data. Error bars indicate the 95% confidence intervals. Different letters represent significant differences between groups (CLD). Values below the horizontal lines indicate Bonferroni-corrected contrasts p-values.

All tested parameters showed statistically significant estimates (Table A4). A significant interaction effect was observed between genotyping method and sequencing depth, indicating that data NS5 and NS6 showed homozygosity percentage variation lower than or equal to REF data when the sequencing depth was greater than approximately three times (Figure 3). Negative coefficients for both sequencing depth and sample size indicate that the homozygosity percentage variation is reduced as these variables are incremented.

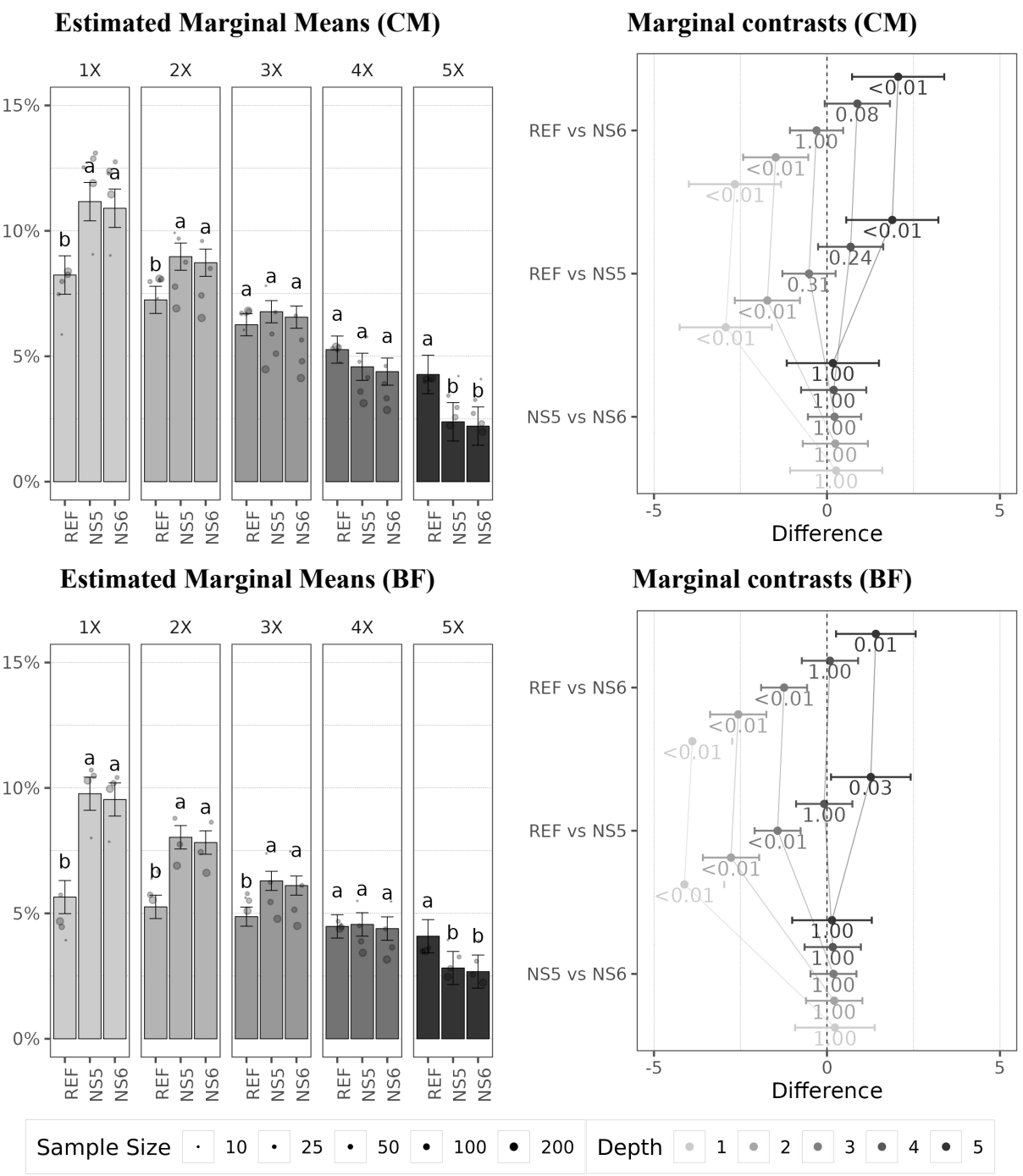


Figure 3. Contrast analysis for Homozygosity percentage change. Points represent individual data. Error bars indicate the 95% confidence intervals. Different letters represent significant differences between groups (CLD). Values below the horizontal lines indicate Bonferroni-corrected contrasts p-values. Facets and solid connected lines indicate interaction effect for sequence depth.

The genotyping method and sequencing depth parameters showed statistically significant estimates in the linear regression analysis (Table A6) for genotype concordance rate. Significant interaction effect was observed between genotyping method and sequencing depth for heterozygotes and homozygotes for the minor allele. The homozygous genotypes

concordance rate was significantly higher in NS5 and NS6 data than REF data (Figure 4). However, REF data showed a lower concordance rate, both for heterozygotes and for homozygotes for the major allele. Linear and quadratic coefficients of sequencing depth indicate that the greater the sequencing depth, the greater the genotype concordance rate.

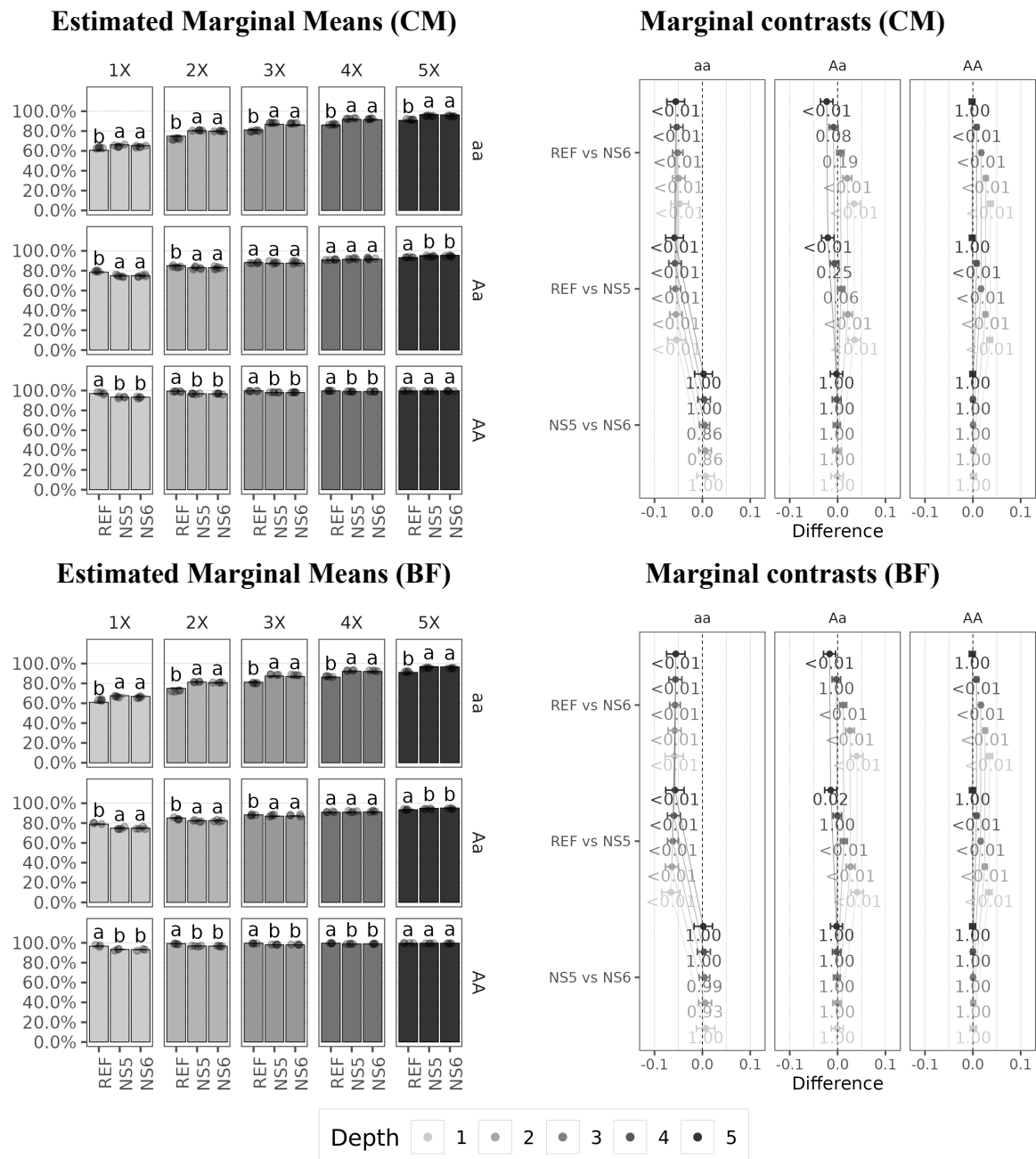


Figure 4. Contrast analysis for Genotype concordance rate. Points represent individual data. Error bars indicate the 95% confidence intervals. Different letters represent significant differences between groups (CLD). Values below the horizontal lines indicate Bonferroni-corrected contrasts p-values. Facets and solid connected lines indicate interaction effect for sequence depth.

## Discussion

One of the main problems lcWGS genotyping is the increase in SNP false positive rate. The smaller the average sequencing depth, the greater the likelihood of false positive SNP identification (SIMS, D. et al, 2014). LCSeqTools pipeline significantly reduces the false positive rate when compared to unprocessed data, resulting in around four times less false positive at average sequencing depth from one to five times (Table A3). Another important factor for significantly reduce false positives is the dataset sample size. Beside of that, false positives occur more frequently in lower allele frequencies, mainly for MAF ranges between 0.1 to 0.2, causing a small noise in most of cases. Nevertheless, researchers can validate a smaller set of candidate SNPs using more accessible and cost-effective genotyping techniques and still reach a good cost-benefit ratio compared to traditional WGS.

The descriptive statistics indicated that an average sequencing depth  $< 1$  of less than one time causes a higher variation in the estimated parameters compared to the observed variation between an average sequencing depth between one to five times. LCSeqTools pipeline does not generate reliable results with extremely-low coverage, so it should not be used for lcWGS data with average DP  $< 1$ . The LCSeqTools pipeline performance was very similar to the traditional genotyping protocol at average DP = 10 times, indicating that the ideal scenario for application of data processing is between one to five times of average sequencing depth.

The variance in allele frequency estimates caused by low-coverage sequencing was significantly reduced using LCSeqTools pipeline, resulting in an average variance of treated data about 60% lower compared to untreated data (Table A5). It is noteworthy that most estimates show negative variation, indicating an underestimation of the real allele frequency of the SNP (Figures A4 and A5). The sequencing depth and sample size have a direct impact on allele frequency estimates, and the greater the sample size and sequencing depth, the smaller the variation in the allele frequency estimate. Homozygosity inflation is another main drawback of genotyping by low coverage sequencing.

---

The mean sequencing depth is directly linked to homozygosity inflation, and this effect can be seen in figures A7 and A8, given that the smaller the mean sequencing depth, the greater the inflation of the homozygosity estimate. The sample size also has an impact on the homozygosity estimate, where the larger the sample size, the lower the homozygosity inflation. LCSeqTools pipeline showed different results in different scenarios, in which there was a reduction in homozygosity inflation when the sequencing depth was greater than three times and amplification of homozygosity inflation when the sequencing depth was less than three times (Table A7). This result suggests that research designs that depend on high reliability estimates for homozygosity, sample size and sequencing depth should be very well structured, given that the reduction in homozygosity inflation is much more associated with sequencing depth than the sample size.

LCSeqTools significantly improved the concordance rate of homozygous genotypes for the smallest allele, which had the greatest impact caused by low coverage sequencing (Table A8). However, divergent results for levels of sequencing depth were observed in the concordance rates of heterozygous and homozygous genotypes for the major allele (Table A9 and Table A10), given that LCSeqTools slightly reduces the concordance rate when the sequencing depth is also reduced. Nevertheless, the reduction caused by data processing keeps the concordance rates sufficiently high, being practically negligible in homozygotes for the major allele.

The lcWGS approach offers an excellent cost-benefit ratio as an alternative strategy, particularly for large-scale sequencing projects. It allows for an increase in the total number of sequenced samples, which is inversely proportional to the reduction in sequencing depth, up to the technical limit of the proposed pipeline. This is particularly crucial for studies requiring large sample sizes, such as epidemiological research, where there is a superior statistical advantage in having more samples than certainty in individual genotypes (LOU, R.N. et al, 2021; ALEX BUERKLE, C. & GOMPERT. Z. et al, 2013).

## Conclusion

LCSeqTools pipeline significantly improves the genotyping performance of lcWGS data, providing a robust alternative for low budget projects aiming population genomics studies using high frequency biallelic SNPs. The number of false positive SNPs is significantly reduced using this pipeline, resulting in an efficient method to identify a large and reliable set of SNPs that can be easily validated by more accessible and less expensive genotyping techniques.

## Availability of data and material

All documentation, including scripts and links necessary to replicate the study's findings or to further extend the research, is available at:

- <https://github.com/marcusnizalvarez/PhD-Scripts>
- <https://github.com/marcusnizalvarez/LCSeqTools>
- <https://doi.org/10.5281/zenodo.11267120>
- <https://doi.org/10.5281/zenodo.13755768>

## References

ALEX BUERKLE, C.; GOMPERT, Zachariah. Population genomics based on low coverage sequencing: how low should we go?. **Molecular ecology**, v. 22, n. 11, p. 3028-3035, 2013.

ALVAREZ, Marcus Vinicius Niz et al. Nyssorhynchus darlingi genome-wide studies related to microgeographic dispersion and blood-seeking behavior. **Parasites & Vectors**. v. 15, n. 1, p. 106. 2022.

ALVAREZ, Marcus Vinicius Niz. LCVCFTools v1.0.2-alpha (v1.0.2-alpha). **Zenodo**. <https://doi.org/10.5281/zenodo.5259931>. 2021.

ANDREWS, Simon et al. FastQC: a quality control tool for high throughput sequence

---

data. **Babraham Bioinformatics**. 2010.

ANOPHELES GAMBIAE 1000 GENOMES CONSORTIUM et al. Genetic diversity of the African malaria vector *Anopheles gambiae*. **Nature**, v. 552, n. 7683, p. 96, 2017.

BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014.

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.

DAVIES, Robert W. et al. Rapid genotype imputation from sequence with reference panels. **Nature Genetics**, v. 53, n. 7, p. 1104-1111, 2021.

DENG, Tianyu et al. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. **Frontiers in genetics**, v. 12, p. 704118-704118, 2021.

HUANG, Weichun et al. ART: a next-generation sequencing read simulator. **Bioinformatics**, v. 28, n. 4, p. 593-594, 2012.

HUI, Ruoyun et al. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. **Scientific reports**, v. 10, n. 1, p. 1-8, 2020.

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.

LOU, Runyang Nicolas et al. A beginner's guide to low-coverage whole genome sequencing for population genomics. **Molecular Ecology**, v. 30, n. 23, p. 5966-5993, 2021.

NIELSEN, Rasmus et al. Genotype and SNP calling from next-generation sequencing data. **Nature Reviews Genetics**, v. 12, n. 6, p. 443-451, 2011.

PRESTON, J.; VANZEELAND, A.; PEIFFER, D. A. Innovation at Illumina: The road to the \$600 human genome. **Nat. Portf**, 2021.

---

R CORE TEAM et al. R development core team. **R: A language and environment for statistical computing**, v. 55, p. 275-286, 2016.

RUBINACCI, Simone et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. **Nature Genetics**, v. 53, n. 1, p. 120-126, 2021.

SAYERS, Eric W. et al. Database resources of the National Center for Biotechnology Information in 2023. **Nucleic acids research**, v. 51, n. D1, p. D29, 2023.

TEAM, RStudio et al. RStudio: integrated development for R. RStudio, Inc. **Boston, MA, USA**, v. 42, p. 14, 2015.

WETTERSTRAND, Kris A. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). **National Human Genome Research Institute**, 2013.

WICKHAM, Hadley. et al. Welcome to the Tidyverse. **Journal of open source software**, v. 4, n. 43, p. 1686, 2019.

ZHANG, Zhe et al. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in Duroc pigs. **Genomics**, v. 114, n. 1, p. 340-350, 2022.

## CAPÍTULO 2 – Chromosome Inversion detection using low-coverage sequencing data (lcWGS).

**Authors:** Marcus Vinicius Niz Alvarez, Filipe Trindade Bozoni, Lorena Leles, Diego Peres Alonso and Paulo Eduardo Martins Ribolla

---

### Abstract

**Introduction:** *Nyssorhynchus darlingi* is the primary malaria vector in the Neotropical region. Previous studies have detected chromosome inversions in salivary glands polytene chromosomes of *Ny. darlingi* samples from Amazon region. Chromosomal inversions have been recognized as important evolutionary mechanisms in diverse organisms. In mosquitoes, structural variants are associated with important ecological adaptation, vectorial capacity and even speciation. A powerful strategy to understand molecular evolution and population dynamics is the use of molecular markers. Although there has been rapid progress in genome sequencing technology, it continues to be an economical challenge for large scale projects. A cost-effective alternative approach is to use low-coverage whole genome sequencing (lcWGS) techniques. In this study, we used genome-wide SNPs to detect chromosome inversions at sequencing level, reliably, with lcWGS data. We detected chromosome inversions in *Ny. darlingi* samples from Amazon region and compared with *Anopheles gambiae* and *Anopheles albimanus* genomes in synteny analysis. **Methods:** We used 321 *Ny. darlingi* samples from Brazil and 200 *An. gambiae* samples from the Republic of Cameroon from publicly available sequencing data. The *An. gambiae* dataset was used to benchmark the inversion detection pipeline with known inversions. Genotyping by sequencing was performed using the LCSeqTools workflow for the lcWGS dataset with an average sequencing depth of 2x. Genetic distance estimates were calculated with PLINK 1.9 software with the pruned SNPs dataset. PCA was applied and each principal component analyzed individually with a sliding window approach for detecting chromosome inversion footprints. A synteny analysis was performed for *Ny. darlingi* inversions regions with *An. gambiae* and *An. albimanus* genomes.

**Results:** The sliding window analysis of PCA components revealed 10 high confidence candidate regions for chromosome inversions in *Ny. darlingi* genome and two known inversions for *An. gambiae* with satisfactory precision of breakpoints. Synteny analysis showed no similar arrangements for *Ny. darlingi* with *An. gambiae* and *An. albimanus* genomes with synteny analysis. **Conclusion:** We demonstrate that lcWGS is a cost-effective and accurate method for detecting chromosome inversions. We reliably detected chromosome inversions in *Ny. darlingi* sample from the Brazilian Amazon that do not share similar inversion arrangements in *An. gambiae* or *An. albimanus* genomes.

**Keywords:** Mosquito, Bioinformatics, lcWGS, Comparative Genomics, Structural Variants

## Background

The *Nyssorhynchus darlingi* specie stands out as the primary vector for malaria transmission in the Neotropical region, including the Amazon basin (Nagaki *et al.*, 2021; Rafael *et al.*, 2021; WHO, 2023). *Ny. darlingi* presents a chromosomal set of  $2n = 6$ , with two pairs of autosomes: the largest (III) is submetacentric and the smallest (II) is metacentric. For sex chromosomes, X is acrocentric and Y is pointed (Tadei *et al.*, 1998).

Previous studies have identified chromosomal inversions in *Ny. darlingi* populations, particularly in the Amazon region, but a comprehensive understanding of their prevalence, distribution, and functional significance is lacking. Chromosome inversions were detected in salivary glands polytene chromosomes from southeastern Brazil and the Amazon region, with inversions located in all chromosomes arms, except Y chromosome (KREUTZER *et al.*, 1972; TADEI & SANTOS, 1982; RAFAEL *et al.*, 2010; CORNEL *et al.*, 2016).

Structural rearrangements like inversions have long been recognized as important evolutionary mechanisms in diverse organisms (Cornel *et al.*, 2016). In mosquitoes, chromosome inversions are associated with ecological adaptation, population divergence, and even speciation. Chromosome inversions can suppress recombination, leading to the maintenance of coadapted gene complexes and potentially facilitating the rapid evolution of traits related to vectorial capacity, such as insecticide resistance, host preference, thermal and desiccation stress resistance (Cornel *et al.*, 2016; CORBETT-DETIG *et al.*, 2019; AYALA *et al.*, 2019; CHENG *et al.*, 2018).

Although the cost of per-base sequencing has decreased significantly due to advances in whole genome sequencing (WGS) technologies, conducting large-scale sequencing projects

can still be financially challenging and may pose significant obstacles in different contexts. An alternative cost-effective strategy is genotyping by sequencing using low-coverage whole genome sequencing (lcWGS) with genotype imputation methods that improves genomic data reliability for precise marker selection (Gorjanc *et al.*, 2017; Sims *et al.*, 2014).

Genotype inference performance has been validated for both panel-based genotyping and sequencing genotyping modalities, enabling the use of lcWGS to identify variants with lower cost compared to traditional WGS method (Pasaniuc *et al.*, 2012; Rustagi *et al.*, 2017; Zheng *et al.*, 2018; Li *et al.*, 2011).

In the present study, we used genome-wide SNPs from lcWGS data to investigate the population of *Ny. darlingi* collected in Mâncio Lima, Acre state, Brazil. We performed a comprehensive genomic analysis of chromosome inversions in *Ny. darlingi* and a comparative genomics analysis with *Anopheles albimanus* and *Anopheles gambiae* genomes. This study examines related species providing insights into the evolutionary dynamics across different mosquito lineages. Our findings demonstrated that lcWGS enables chromosome inversion detection and can contribute significantly to the broader understanding of mosquito genomic evolution.

## Methods

### Sequencing data acquisition

For this study, we used publicly available sequencing data obtained from the National Center for Biotechnology Information – NCBI (Sayers *et al.*, 2023). Sequencing data types were selected based on their relevance to ensure comprehensive analysis. Sequencing data from lcWGS data of 321 *Ny. darlingi* larvae and adults were obtained from Bioproject PRJNA683015. The samples used in this study were collected in the municipality of Mâncio Lima, Acre state of Brazil.

### Genome reference

The *Ny. darlingi*, *An. gambiae* and *An. albimanus* reference genomes used are publicly available at NCBI database with respective accession numbers: GCF\_943734745.1, GCF\_943734735.2 and GCF\_013758885.1.

## Genotyping by sequencing and variant calling

The genotyping by sequencing pipeline was performed using LCSeqTools v0.1.0 (Alvarez, 2024). Through the LCSeqTools, the read alignment was performed with Burrows-Wheeler Aligner v0.7.17 (Li & Durbin 2009) and the variant calling with SamTools v1.10 (Li, 2011) using the default mapping parameters. The LCSeqTools variant filtering parameters applied before genotype imputation were: Minor allele frequency  $< 0.1$ , max missing data  $< 0.5$ , genotype sequencing depth  $< 5$  and genotype quality  $< 20$ . Genotypes were imputed with BEAGLE v4.1 software (Browning & Browning, 2016) using genotype normalized probability field (PL). A final filtering step was applied to the variant dataset with SNPs pruning algorithm from PLINK v1.9 software (Prucel *et al.*, 2007).

## Data processing and statistical analysis

All data manipulation, analysis and plotting was performed using R v4.3.2 and RStudio Server “Ocean Storm” (Team R, 2013; Team Rstudio, 2015), including packages provided by tidyverse v2.0.0 meta-package (Wichkham *et al.*, 2019).

## Chromosome Inversion Identification

Multiple approaches were combined to check the presence of chromosome inversion signals. Principal component analysis (PCA) for each chromosome was performed using PLINK and the sliding windows variance of SNPs weights (eigenvectors) estimates for each principal component were estimated within 10 kbp sliding windows. Based on the absolute PCA SNPs weights, the top 1% SNPs for each component were retained and the Identical by State (IBS) matrix was calculated using the respective SNPs subset. The multidimensional scaling (MDS) technique was applied to the euclidean IBS matrix with `cmdscale` function from R base ( $k=1$  for genotype clustering and  $k=2$  for pairwise comparisons) and, subsequently, a clustering technique with `cmeans` function from `e1071` package for R (Dimitriadou *et al.*, 2006). Only components that clustered into three well defined groups were retained, later identified as AA, AB and BB inversion genotypes.

A genome-wide association test was applied for each candidate chromosome inversion, using the linear model from PLINK, considering AA, AB and BB samples as quantitative 0, 1 and 2 phenotypes, respectively. A Bonferroni post-hoc test was applied for multiple comparisons correction and SNPs were considered statistically significant SNPs

when adjusted p-value  $< 0.05$ .

Alvarez *et al* (2022) showed that the expected  $r^2$  value at 12.57 kbp distance is approximately 0.1 in this *Ny. darlingi* population. A pairwise SNPs linkage disequilibrium estimate analysis was performed for SNPs pairs within ranges of 12 kbp of distance and sliding windows  $r^2$  median was calculated for each chromosome within sliding windows of 0.5% of the respective chromosome length. The Spearman's correlation coefficient  $\rho$  was estimated for  $r^2$  and the mean SNPs absolute weights using `cor.test` function from R and were considered statistically significant when p-value  $< 0.05$ . All chromosome inversion were submitted to a pairwise Spearman correlation test using the estimated sample genotypes data and the p-values were adjusted for multiple comparison using Benjamini-Hochberg False Discovery Rate method (FDR). Correlations were considered statistically significant when FDR-adjusted  $p \leq 0.05$ .

### **Pipeline validation with known variants for *Anopheles gambiae***

High coverage genome-wide variant data from the first phase of the "The *Anopheles gambiae* 1000 Genomes Consortium" (Anopheles Gambiae 1000 Genomes Consortium *et al.*, 2017) publicly available at EMBL-EBI were used as reference (ERZ373588 with GCF\_000005575.2 genome reference). An *in silico* lcWGS dataset was generated for chromosome identification pipeline validation, matching results with known *An. gambiae* structural variants (AgamP4 variant panel publicly available at Ensembl Metazoa). A subset of 200 samples from the Republic of Cameroon group were selected from the full variant panel and lcWGS data were generated with the `art_illumina` program from ART software package (HUANG *et al.*, 2012) with approximately 2x of sequencing depth per sample.

### **Comparative Genomics analysis**

An orthology inference was performed for *An. gambiae* and *An. albimanus* with *Ny. darlingi* as reference using BLASTp (Camacho *et al.*, 2009) and the respective proteomes through the accession codes. The OrthoFinder software (Emms & Kelly, 2019) was also used as a second approach for orthologs inference and convergence check with BLASTp results. A genome synteny analysis was performed between the *An. gambiae*, *An. albimanus* and *Ny. darlingi* genomes using MCScanX (Wang *et al.*, 2012).

## Results

### Chromosome inversions detection

A variant panel containing 4,241,254 SNPs was generated by the LCSeqTools pipeline for *Ny. darlingi* sample. After the SNPs pruning, PLINK retained 626,409 SNPs, around 3.4 SNPs per Kbp.

The PCA components analysis revealed continuous high-shifted regions from SNPs weight sliding windows variance estimates, counting a total of 10 candidate regions: 2, 4 and 4 shifted regions for chromosomes X, 2 and 3, respectively (figure 1). Clustering analysis of each region revealed three well defined genetic clusters with approximately the same 0.5 genetic distance between samples AA at leftmost and AB in the middle, as well as AB and BB at rightmost (figure 2). Genome-wide association test results for each remaining component showed statistically significant SNPs concentrated in converging chromosome regions of its respective PCA shifted variance windows (figure 1). The chromosome inversion coordinate estimates are represented in table 1.

The SNPs pairwise linkage disequilibrium analysis showed that, despite the median  $r^2$  within a distance of 12 to 13 kbp being 0.08, 0.06 and 0.12 for chromosomes 2, 3 and X, respectively, spikes reaching almost double the median  $r^2$  were found in different regions along the chromosomes (figure 1). Significant correlations were observed between  $r^2$  and mean SNPs absolute weight from PCA components, as the  $\rho$  estimates were 0.62, 0.43 and 0.58 for chromosomes 2, 3 and X, respectively.

### *In silico* validation of the pipeline

The *An. gambiae* lcWGS simulated data resulted in a variant panel containing 3,825,409 SNPs and PLINK retained 658,413 SNPs after pruning step, resulting in 2.36 SNPs per Kbp. Two meaningful components were detected with the sliding windows of SNPs weights analysis approach and, subsequently, Genome-wide association test revealed statistically significant SNPs concentrated within two regions, as coordinates estimates were [20466374, 42223401] and [19275187, 26282153] for 2L and 2R (figure 3). Both inversion estimates greatly coincided with well known *An. gambiae* inversions: 2La [20524057, 42165532] and 2Rb [19023924, 26758676]. Significant correlations were observed between  $r^2$  and mean SNPs absolute weight from PCA components, as the  $\rho$

estimates were 0.88 and 0.49 for 2L and 2R, respectively.

### Genome syntenic analysis

Comparative genomics analysis revealed similar arrangements between *Ny. darlingi* and *An. albimanus* genomes and a major rearrangement when comparing chromosome 2 and 3 from *An. gambiae* and *Ny. darlingi* (figure 4). Syntenic analysis for each chromosome inversion region showed that no similarity was found between *Ny. darlingi* detected inversions and *An. gambiae* known chromosome inversions (figure 5). No data was available comparing *An. albimanus* inversions.

### Discussion

First of all, it is important to know that there is a divergence between the Rafael & Tadei (1998) chromosome nomenclature for *Ny. darlingi* and the GCF\_943734745.1 genome assembly, as the chromosome 2 and 3 names are swapped in the genome assembly. The following discussion is based on the genome assembly nomenclature.

This study shows that lcWGS provides a powerful cost-reduced method for detecting chromosome inversions in large sample sizes. The *in silico* experiment reinforces that, besides the higher genotyping error probability of lcWGS, high confidence inversion signal detection is possible because multiple independent SNPs are reduced into a single genotype data with PCA technique, reducing any lcWGS genotyping bias.

Nowling *et al* (2020) performed a similar approach with traditional WGS data for *An. gambiae* dataset and found a comparable performance for PCA-based analysis on chromosome inversion detection. Although there are various pipelines for detecting chromosome inversion using SNPs or different genetic parameters as they summarized, a simple sliding window analysis is enough to detect inversion footprints if an in-depth eigenvectors analysis is performed.

The *An. gambiae* inversion range estimates from lcWGS data in this study were satisfactorily close to the known ranges and ranges estimated from other dedicated tools (Nowling *et al*, 2022). Even on genome drafts in which chromosome data is not well assembled, it is possible to use PCA clustering to detect inversion presence because eigenvectors are estimated essentially with a set of independent SNPs, so contiguity isn't needed for detection. This is an important feature for non-model organisms studies, but the

more shattered the assembly is, the more difficult it will be to estimate inversion coordinate range.

There is a significant correlation between linkage disequilibrium and absolute PCA eigenvectors, although PCA-based seems to be a more precise method (Cáceres & González, 2015). The observed linkage disequilibrium oscillation towards the inversion breakpoints is a known effect caused by the change in the recombination rate (Seich al Basatena *et al*, 2013).

The genome synteny analysis revealed an interesting rearrangement between arms of chromosomes 2 and 3 for *Ny. darlingi* and *An. albimanus* when compared to *An. gambiae* genome. This is a known phenomenon evidenced by other comparative genomics studies focusing on Anophelinae, which have consistently reported analogous chromosomal rearrangements (Wei *et al*, 2017; Jiang *et al*, 2014). The ten chromosome inversions detected in *Ny. darlingi* genome did not show similar ranges or arrangements for known *An. gambiae* inversions in the figure 5. Although *An. albimanus* revealed some interesting equivalent syntenic regions for the chromosome inversions, *An. albimanus* typically lacks chromosome inversions (Artemov *et al*, 2016). Unfortunately, during the period of this study, there was no publicly accessible sequencing data for *An. albimanus* that would allow for a comprehensive analysis of inversion footprints using the methodology employed in this study.

Soboleva *et al* (2024) also found two nested X chromosome inversions in *Anopheles atroparvus* and *Anopheles messeae*, although the ranges seem to be slightly different.

## Conclusion

Chromosomal inversions are an important biological marker related to genome evolution, environment adaptation and speciation. Our study showed that lcWGS is particularly promising for large-scale genomic studies, offering a cost-effective alternative strategy to detect chromosome inversions without compromising accuracy. The sliding window analysis of PCA eigenvectors to detect inversions ensures accurate inversion detection and range estimation with a simple standalone approach. Ten chromosome inversions were found using lcWGS data from *Ny. darlingi* samples collected in one municipality located in the Brazilian Amazon basin. Also, *in-silico* simulations showed that the pipeline has enough sensitivity for accurate inversion detection. This could be an effective strategy for comprehending the population dynamic on a much larger scale.

## Availability of data and material

All documentation, including scripts and links necessary to replicate the study's findings or to further extend the research, is available at:

- <https://github.com/marcusnizalvarez/PhD-Scripts>
- <https://doi.org/10.5281/zenodo.13755768>.

## References

- Alvarez M.V.N. LCSeqTools v0.1.0-alpha. 2024. **Zenodo**. <https://doi.org/10.5281/zenodo.11267121>
- Alvarez, M. V. N., Alonso, D. P., Kadri, S. M., Rufalco-Moutinho, P., Bernardes, I. A. F., de Mello, A. C. F., ... & Ribolla, P. E. (2022). Nyssorhynchus darlingi genome-wide studies related to microgeographic dispersion and blood-seeking behavior. **Parasites & Vectors**, 15(1), 106.
- Anopheles gambiae 1000 Genomes Consortium. (2017). Genetic diversity of the African malaria vector Anopheles gambiae. **Nature**, 552(7683), 96.
- Artemov, G. N., Peery, A. N., Jiang, X., Tu, Z., Stegny, V. N., Sharakhova, M. V., & Sharakhov, I. V. (2017). The physical genome mapping of Anopheles albimanus corrected scaffold misassemblies and identified interarm rearrangements in genus Anopheles. **G3: Genes, Genomes, Genetics**, 7(1), 155-164.
- Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. **The American Journal of Human Genetics**, 98(1), 116-126.
- Cheng, C., Tan, J. C., Hahn, M. W., & Besansky, N. J. (2018). Systems genetic analysis of inversion polymorphisms in the malaria mosquito Anopheles gambiae. **Proceedings of the National Academy of Sciences**, 115(30), E7005-E7014.
- Cáceres, A., & González, J. R. (2015). Following the footprints of polymorphic inversions on SNP data: from detection to association tests. **Nucleic acids research**, 43(8), e53-e53.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. **BMC bioinformatics**, 10, 1-9.
- Cornel, A. J., Brisco, K. K., Tadei, W. P., Secundino, N. F., Rafael, M. S., Galardo, A. K., ... & Lanzaro, G. C. (2016). Anopheles darlingi polytene chromosomes: revised maps including newly described inversions and evidence for population structure in Manaus.

**Memórias do Instituto Oswaldo Cruz**, 111, 335-346.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2006). The e1071 package. Misc Functions of Department of Statistics (e1071), **TU Wien**, 297-304.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome biology**, 20, 1-14.

Gorjanc, G., Dumasy, J. F., Gonen, S., Gaynor, R. C., Antolin, R., & Hickey, J. M. (2017). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. **Crop Science**, 57(3), 1404-1420.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. **Bioinformatics**, 28(4), 593-594.

Jiang, X., Peery, A., Hall, A. B., Sharma, A., Chen, X. G., Waterhouse, R. M., ... & Tu, Z. (2014). Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. **Genome biology**, 15, 1-18.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, 27(21), 2987-2993.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome research*, 21(6), 940-951.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. **bioinformatics**, 25(14), 1754-1760.

MINISTRY OF HEALTH OF BRAZIL. Epidemiological bulletin SVS 35. 2020. Available at: <https://antigo.saude.gov.br/images/pdf/2019/novembro/20/Boletim-epidemiologico-SVS-35.pdf>.

Nagaki, S. S., Chaves, L. S., López, R. V. M., Bergo, E. S., Laporta, G. Z., Conn, J. E., & Sallum, M. A. M. (2021). Host feeding patterns of *Nyssorhynchus darlingi* (Diptera: Culicidae) in the Brazilian Amazon. **Acta tropica**, 213, 105751.

Nowling, R. J., Manke, K. R., & Emrich, S. J. (2020). Detecting inversions with PCA in the presence of population structure. **PLoS One**, 15(10), e0240429.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., ... & Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. **Nature genetics**, 44(6), 631-635.

Pimenta, P. F., Orfano, A. S., Bahia, A. C., Duarte, A. P., Ríos-Velásquez, C. M.,

Melo, F. F., ... & Secundino, N. F. (2015). An overview of malaria transmission from the perspective of Amazon Anopheles vectors. **Memórias do Instituto Oswaldo Cruz**, 110, 23-47.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American journal of human genetics**, 81(3), 559-575.

R Core Team. (2016). R development core team. R: A language and environment for statistical computing, 55, 275-286.

Rafael, M. S., Bridi, L. C., Sharakhov, I. V., Marinotti, O., Sharakhova, M. V., Timoshevskiy, V., ... & Tadei, W. P. (2021). Physical mapping of the Anopheles (Nyssorhynchus) darlingi genomic scaffolds. **Insects**, 12(2), 164.

Rustagi, N., Zhou, A., Watkins, W. S., Gedvilaite, E., Wang, S., Ramesh, N., ... & Xing, J. (2017). Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. **BMC genomics**, 18, 1-12.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., ... & Sherry, S. T. (2023). Database resources of the National Center for Biotechnology Information in 2023. **Nucleic acids research**, 51(D1), D29.

Seich al Basatena, N. K., Hoggart, C. J., Coin, L. J., & O'Reilly, P. F. (2013). The effect of genomic inversions on estimation of population genetic parameters from SNP data. **Genetics**, 193(1), 243-253.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, 15(2), 121-132.

Soboleva, E. S., Kirilenko, K. M., Fedorova, V. S., Kokhanenko, A. A., Artemov, G. N., & Sharakhov, I. V. (2024). Two Nested Inversions in the X Chromosome Differentiate the Dominant Malaria Vectors in Europe, Anopheles atroparvus and Anopheles messeae. **Insects**, 15(5), 312.

Tadei, W. P., & Santos, J. M. M. D. (1982). Biologia de anofelinós amazônicos. VII. Estudo da variação de frequências das inversões cromossômicas de Anopheles darlingi Root (Diptera, Culicidae). **Acta Amazonica**, 12, 759-785.

Tadei, W. P., Thatcher, B. D., Santos, J. M. M., Scarpassa, V. M., Rodrigues, I. B., & Rafael, M. S. (1998). Ecologic observations on anopheline vectors of malaria in the Brazilian Amazon. **Journal Of Tropical Medicine And Hygiene**. 59(2), 325-335.

Team, R. (2015). RStudio: integrated development for R. RStudio. Inc., Boston, MA,

700, 879.

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., ... & Paterson, A. H. (2012). MCSscanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*, 40(7), e49-e49.

Wei, Y., Cheng, B., Zhu, G., Shen, D., Liang, J., Wang, C., ... & Xia, A. (2017). Comparative physical genome mapping of malaria vectors *Anopheles sinensis* and *Anopheles gambiae*. *Malaria Journal*, 16, 1-10.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software*, 4(43), 1686.

WORLD HEALTH ORGANIZATION et al. World malaria report 2023. *World Health Organization*, 2023.

Zheng, C., Boer, M. P., & van Eeuwijk, F. A. (2018). Accurate genotype imputation in multiparental populations from low-coverage sequence. *Genetics*, 210(1), 71-82.

Figures and tables

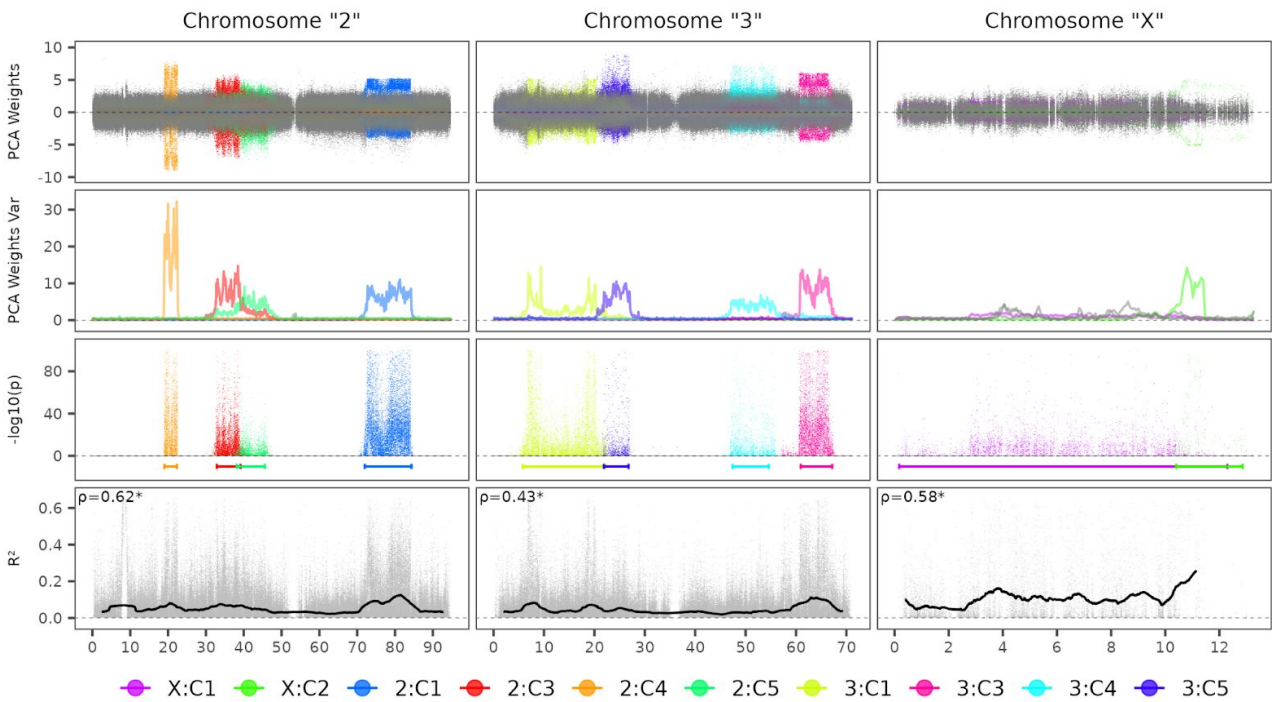


**Figure 1. Chromosome Inversion Detection Analysis for *Ny. darlingi* dataset.** PCA Weights Var: Sliding window of SNPs weights variance estimate.  $\rho$ : Spearman correlation coefficient for  $r^2$  and mean absolute SNPs weights. \*: Correlation test p value < 0.05. Legend stands for Chromosome: Principal Component. Horizontal colored segments represent the inversion coordinate estimates described in table 1. Horizontal scale in Mbp.

**Table 1.** *Ny. darlingi* chromosome inversion coordinate estimates based on Genome-wide association test.

| Chromosome | Accession Code | Component | Start    | End      | Length   |
|------------|----------------|-----------|----------|----------|----------|
| 2          | NC_064874.1    | C1        | 71998919 | 84371386 | 12372467 |
| 2          | NC_064874.1    | C3        | 32875093 | 39167094 | 6292001  |
| 2          | NC_064874.1    | C4        | 19048046 | 22354438 | 3306392  |
| 2          | NC_064874.1    | C5        | 38207551 | 45614642 | 7407091  |
| 3          | NC_064875.1    | C1        | 5786821  | 22248013 | 16461192 |
| 3          | NC_064875.1    | C3        | 60944934 | 67169663 | 6224729  |
| 3          | NC_064875.1    | C4        | 47375561 | 54594956 | 7219395  |
| 3          | NC_064875.1    | C5        | 21879305 | 26791351 | 4912046  |
| X          | NC_064873.1    | C1        | 169219   | 12292796 | 12123577 |
| X          | NC_064873.1    | C2        | 10405329 | 12856594 | 2451265  |

Component: Principal component from PCA analysis.

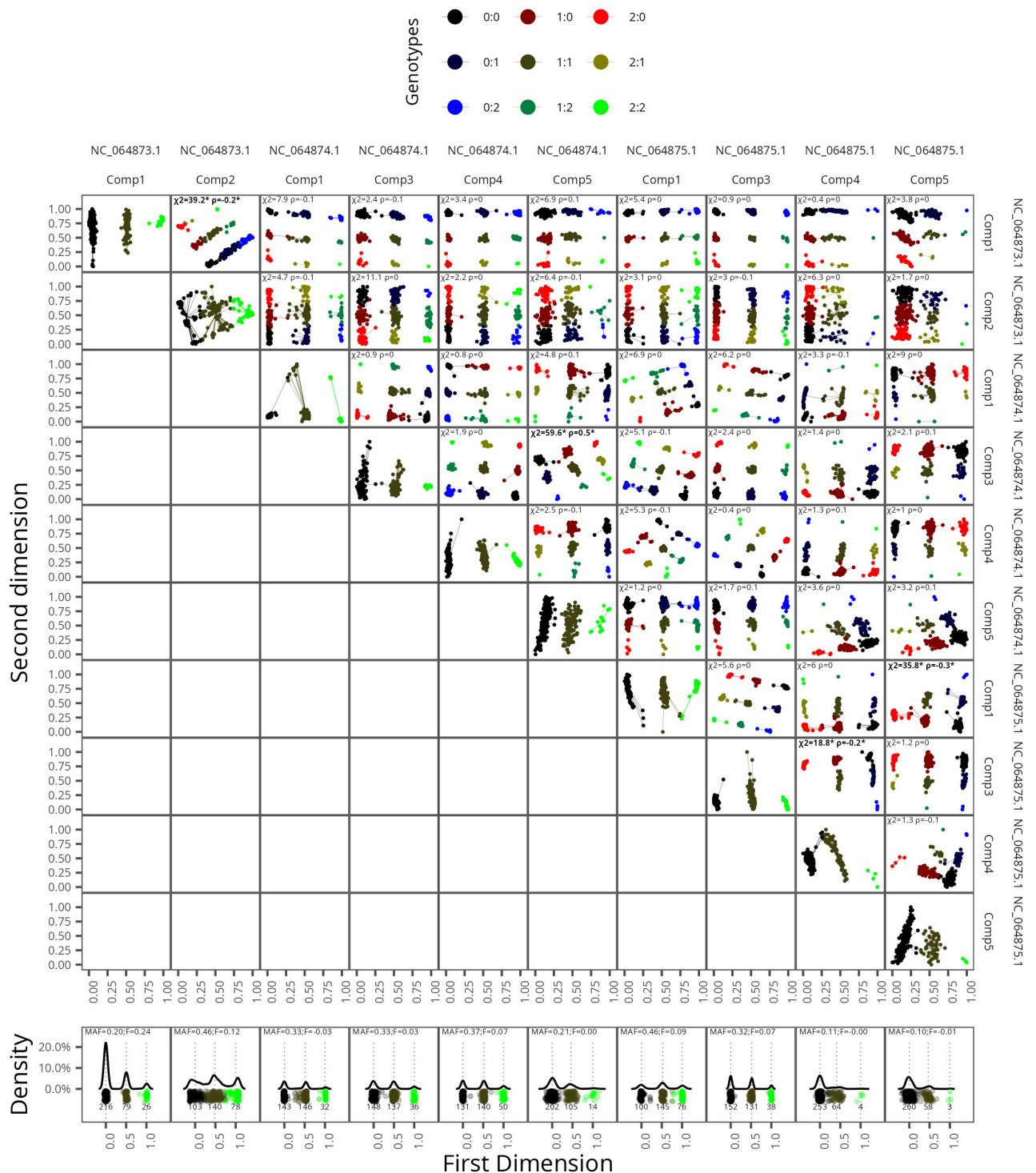


**Figure 2. Inversion Genotyping Analysis.** Top biplots display in two dimensions (MDS with k=2) the relative distances between samples based on genetic similarity from the top 1% most representative SNPs for each respective inversion.  $\chi^2$ : Chi-square test between genotypes from two respective inversions.  $\rho$ : Spearman correlation coefficient for genotypes states of both inversions. Bottom line plots represent in one dimension the density of samples grouping for each genetic cluster for each inversion. MAF: Minor inversion frequency. F: Inversion fixation index.

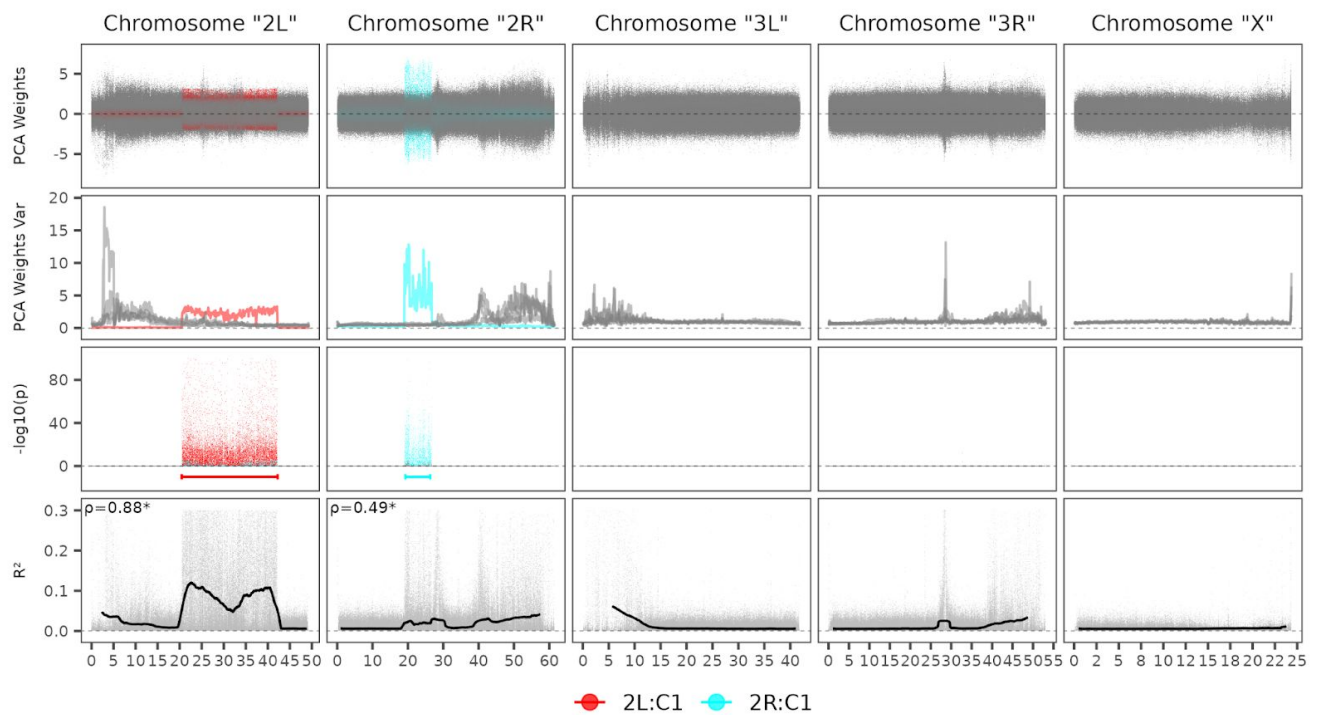


**Figure 3. Chromosome Inversion Detection Analysis for *An. gambiae* dataset.** PCA Weights Var: Sliding window of SNPs weights variance estimate.  $p$ : Spearman correlation coefficient for  $r^2$  and mean absolute SNPs weights. \*: Correlation test  $p$  value  $< 0.05$ . Legend stands for Chromosome: Principal Component. Horizontal scale in Mbp.

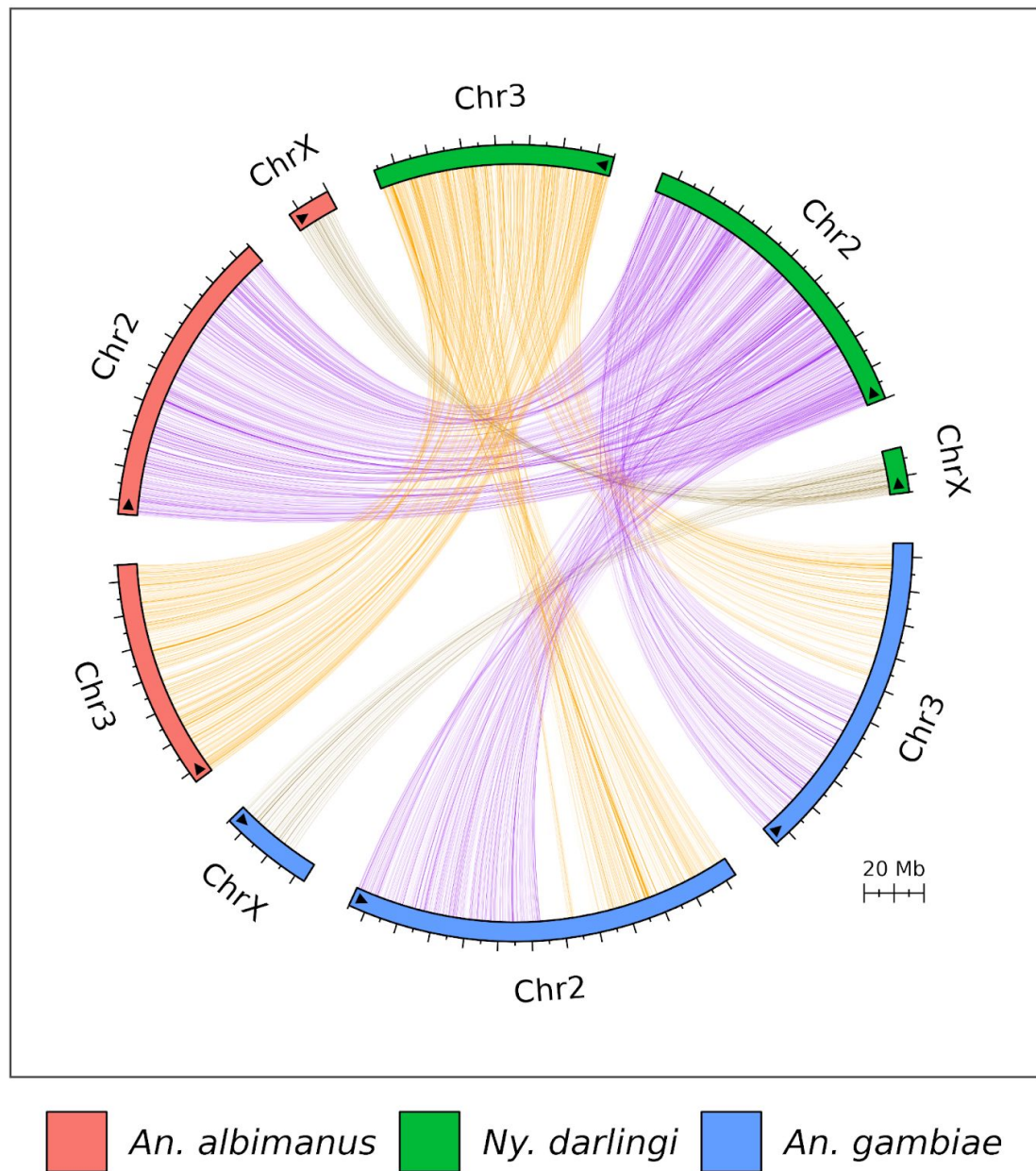


**Figure 4. Genome Synteny Analysis for *Ny. darlingi* with *An. gambiae* and *An. albimanus*.** Black arrow inside each segment represents the original reference orientation.

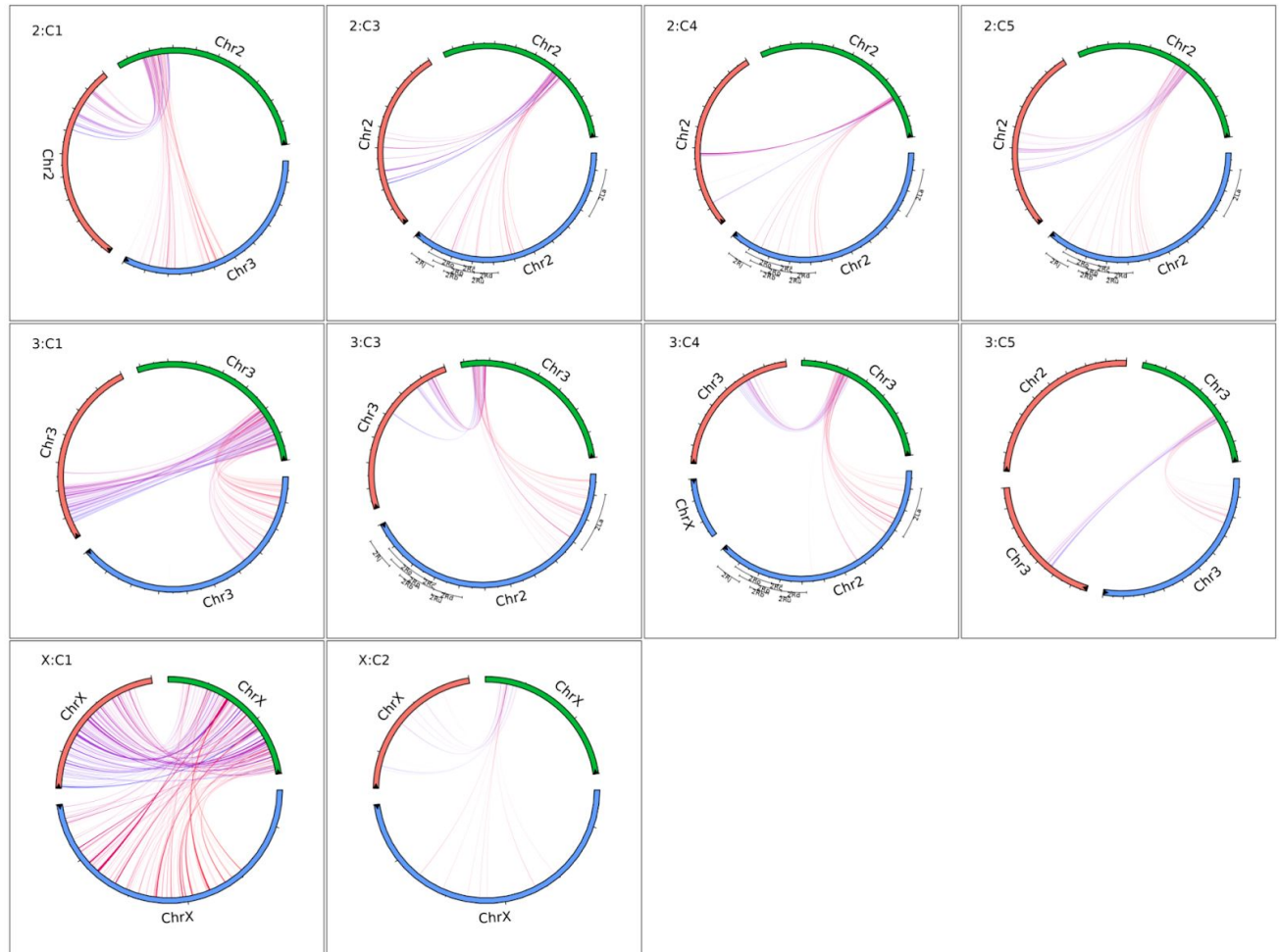


**Figure 5. Chromosome Inversion Synteny Analysis.** Green color: *Ny. darlingi*. Blue color: *An. gambiae*. Red color: *An. albimanus*. Upper left label indicates the chromosome and the i-th principal component in which chromosome inversion was detected for *Ny. darlingi*. Known *An. gambiae* inversions for chromosome 2 are represented with the black line segments outside the circles. Black arrow inside each segment represents the original reference orientation. Only chromosomes with at least one link were retained in each plot.

### 3. CONSIDERAÇÕES FINAIS

O sequenciamento genômico completo de baixa cobertura aliado ao tratamento de dados adequado e imputação de genótipos, como é realizado pelo programa LCSeqTools, produz resultados confiáveis e satisfatórios para projetos em larga escala. Ganhos significativos na confiabilidade da genotipagem permitem que estratégia de genotipagem lcWGS seja uma alternativa robusta de genotipagem com orçamento reduzido e confiabilidade satisfatória. A depender do desenho experimental e da hipótese biológica a ser testada, pode ser uma abordagem com mais vantagens que o WGS tradicional, pois pelo mesmo orçamento, o maior número de amostras sequenciadas no lcWGS permite a possibilidade do uso de desenhos experimentais de larga escala e viabiliza a exploração de diversas hipóteses simultaneamente.

O primeiro capítulo demonstrou que, o viés nas estimativas dos parâmetros de genotipagem causados pela baixa cobertura pode ser compensados se o projeto for corretamente dimensionado, levando em consideração os principais fatores que são tamanho amostral, profundidade média de sequenciamento e tratamento de dados adequado. Se o objetivo de um projeto for detecção de polimorfismos de alta frequência, há mais vantagens do que desvantagens nessa abordagem. Por exemplo, no contexto de estudos epidemiológicos que o processo de coleta pode ser massivo, múltiplos fatores podem ser avaliados por um orçamento reduzido.

No segundo capítulo, a estratégia de detecção de inversões cromossômicas por lcWGS demonstrou ser altamente eficaz. Além de produzir resultados confiáveis e que contribuem para a compreensão sobre a estrutura e evolução genômica da espécie em questão, reforça o poder da genotipagem de alta densidade. O estudo atesta a possibilidade de explorar mais de uma hipótese biológica partindo do mesmo conjunto de dados, mesmo que o desenho experimental inicial tenha sido dimensionado para testar uma hipótese diferente em outro estudo publicado anteriormente.

Estamos vivendo a era dos dados. Fica claro que há muito mais conhecimento que pode ser extraído de dados genômicos. A viabilização de estudos em larga escala a um custo acessível pelo lcWGS mostra o potencial dessa abordagem, nos permitindo explorar de maneira ainda mais ampla os dados genético para compreender a complexidade da vida em sua essência mais profunda.

## REFERÊNCIAS

- ALVAREZ, Marcus Vinicius Niz et al. Nyssorhynchus darlingi genome-wide studies related to microgeographic dispersion and blood-seeking behavior. **Parasites & Vectors**, v. 15, n. 1, p. 106, 2022.
- AYALA, Diego et al. Chromosome inversions and ecological plasticity in the main African malaria mosquitoes. **Evolution**, v. 71, n. 3, p. 686-701, 2017.
- 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.
- BRASIL. Ministério da Saúde. Boletim Epidemiológico. **Secretaria de Vigilância em Saúde e Ambiente**, v. 55, n. 1, p. 1-15, 2024.
- CÁCERES, Alejandro; GONZÁLEZ, Juan R. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. **Nucleic acids research**, v. 43, n. 8, p. e53-e53, 2015.
- CHENG, Changde et al. Systems genetic analysis of inversion polymorphisms in the malaria mosquito Anopheles gambiae. **Proceedings of the National Academy of Sciences**, v. 115, n. 30, p. E7005-E7014, 2018.
- CONSOLI, Rotraut AGB; LOURENÇO-DE-OLIVEIRA, Ricardo. **Principais mosquitos de importância sanitária no Brasil**. SciELO-Editora FIOCRUZ, 1994.
- CORBETT-DETIG, Russell B. et al. Fine-mapping complex inversion breakpoints and investigating somatic pairing in the Anopheles gambiae species complex using proximity-ligation sequencing. **Genetics**, v. 213, n. 4, p. 1495-1511, 2019.
- 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, p. 335-346, 2016.
- FORATTINI, O. P. **Culicidologia médica**. Editora USP. São Paulo. 2002.
- 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.
- KREUTZER RD, Kitzmiller JB, Ferreira E. Inversion polymorphism in the salivary gland chromosomes of Anopheles darlingi root. **Mosquito News**, v. 32, n. 4, p. 555-565, 1972.
- 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.

- PASANIUC, Bogdan et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. **Nature genetics**, v. 44, n. 6, p. 631-635, 2012.
- 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, 2010.
- RAFAEL, Míriam Silva; TADEI, Wanderli Pedro. Metaphase karyotypes of *Anopheles* (*Nyssorhynchus*) *darlingi* Root and *A.(N.) nuneztovari* Gabaldón (Diptera; Culicidae). **Genetics and molecular biology**, v. 21, p. 351-354, 1998.
- RAFAEL, Míriam Silva; TADEI, Wanderli Pedro. Heterochromatin variation in chromosomes of *Anopheles* (*Nyssorhynchus*) *darlingi* Root and *A.(N.) nuneztovari* Gabaldón (Diptera: Culicidae). **Genetics and Molecular Biology**, v. 23, p. 67-70, 2000.
- 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, p. 221-234, 1990.
- RUBINACCI, Simone et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. **BioRxiv**, 2020.
- RUSTAGI, Navin et al. Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. **BMC genomics**, v. 18, n. 1, p. 1-12, 2017.
- SERVICE, M. W. Medical entomology for students. **Cambridge University Press**, 2008.
- SIMS, David et al. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, v. 15, n. 2, p. 121-132, 2014.
- SNOW, Robert W. et al. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. **Nature**, v. 434, n. 7030, p. 214-217, 2005.
- 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.
- TADEI, Wanderli Pedro; SANTOS, Joselita Maria Mendes dos; RABBANI, Mohammad Ghulan. Biologia de anofelinos amazônicos. V. Polimorfismo cromossômico de *Anopheles darlingi* Root (Diptera, Culicidae). **Acta Amazonica**, v. 12, p. 353-369, 1982.
- 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.
- WORLD HEALTH ORGANIZATION. World Malaria Report 2023 (World Health Organization, Geneva), **World Health Organization**, v. 1, n. 1, p. 365, 2023.

## GLOSSÁRIO

|                                          |                                                                                                                                                                                                                                                                                                                                                                                                                  |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b><i>Compact Letter Display</i></b>     | Em estatística, CLD é um método utilizado para apresentar os resultados de comparações pareadas entre médias de grupos. Letras são exibidas ao lado das médias dos grupos, onde médias que não compartilham uma mesma letra são consideradas significativamente diferentes. É uma forma eficiente de visualizar quais tratamentos diferem entre si após testes de hipóteses múltiplas.                           |
| <b><i>Framework</i></b>                  | Em engenharia de software, um <i>framework</i> é uma estrutura de códigos genérica que tem o objetivo de prover uma nova função dentro de outro código. A utilização de um <i>framework</i> torna desnecessário a implementação de funções genéricas já implementadas no próprio <i>framework</i> , auxiliando o programador com ferramentas que vão além do que é oferecido pela linguagem de programação base. |
| <b>Frequência do menor alelo</b>         | Em genética, MAF (do inglês, <i>Minor Allele Frequency</i> ) refere-se à frequência com que o segundo alelo mais comum ocorre em uma população.                                                                                                                                                                                                                                                                  |
| <b>Iteração</b>                          | Em informática, iteração representa a repetição de um conjunto de instruções por uma quantidade finita de vezes.                                                                                                                                                                                                                                                                                                 |
| <b><i>Pipeline</i></b>                   | Um <i>pipeline</i> é um conjunto de etapas ou processos sequenciais usados para transformar dados ou executar tarefas.                                                                                                                                                                                                                                                                                           |
| <b>Polimorfismo de Nucleotídeo Único</b> | Em genética, SNP (do inglês, <i>Single Nucleotide Polymorphism</i> ) é uma variação na sequência de DNA que ocorre quando um único nucleotídeo em uma posição específica do genoma é substituído por outro.                                                                                                                                                                                                      |
| <b><i>Software</i></b>                   | Em português, programa de computador. É uma coleção ou sequência de instruções a serem seguidas por um computador para executar uma determinada tarefa.                                                                                                                                                                                                                                                          |

**APÊNDICE A – Material Suplementar do Capítulo 1**

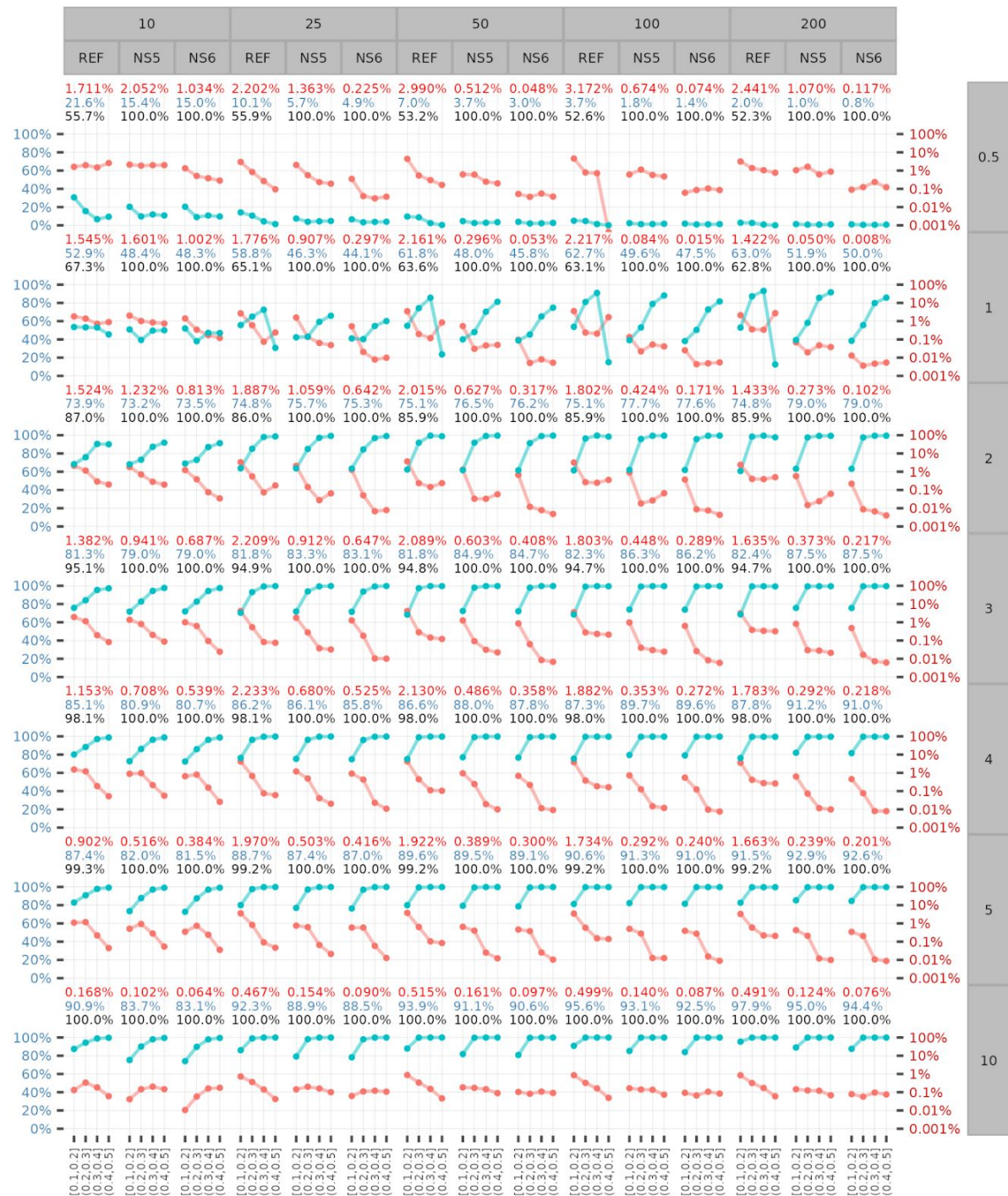


**Figure A1.** Line graph for false positive and true positive SNP rate (CM dataset). Horizontal facets (legend on the right corner) represent sequencing depth. Vertical facets (legend on the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represents the minor allele frequency ranges. The vertical axis on the left represents a linear scale of the true positive SNP rate (blue). The vertical axis on the right represents a logarithmic scale of the false positive SNP rate (red). Top red values: Total false positive SNPs. Top blue values: Total true positive SNPs. Top black values: Non-missing rate.

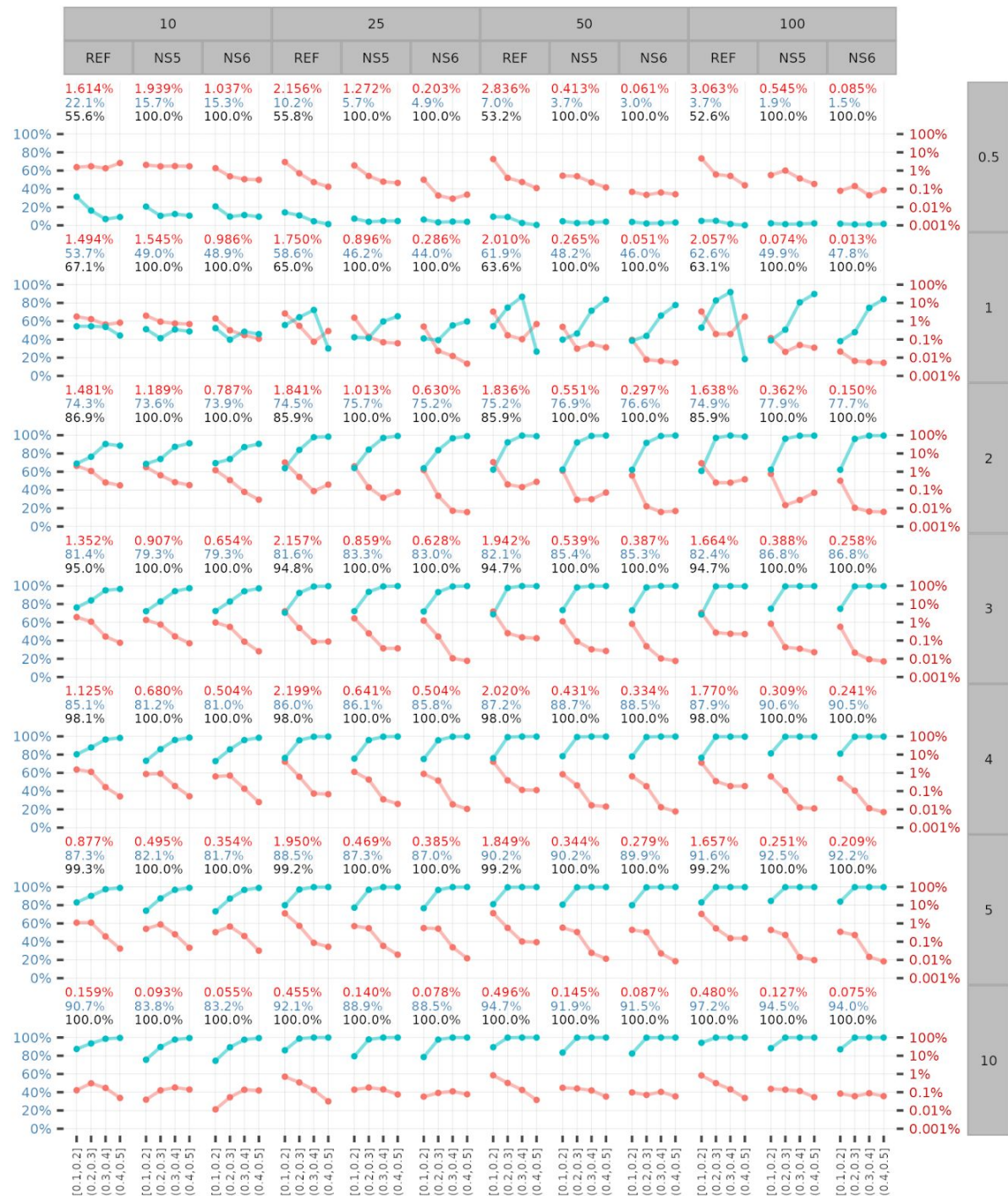


**Figure A2.** Line graph for false positive and true positive SNP rate (BF dataset). Horizontal facets (legend on the right corner) represent sequencing depth. Vertical facets (legend on the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represents the minor allele frequency ranges. The vertical axis on the left represents a linear scale of the true positive SNP rate (blue). The vertical axis on the right represents a logarithmic scale of the false positive SNP rate (red). Top red values: Total false positive SNPs. Top blue values: Total true positive SNPs. Top black values: Non-missing rate.

**Table A1.** Linear regression estimates for false positive SNP rate.

| Dataset | term        | estimate | std.error | statistic | p.value |
|---------|-------------|----------|-----------|-----------|---------|
| CM      | (Intercept) | 2.0282   | 0.0562    | 36.1029   | <0.01   |
|         | DatasetNS6  | -1.4842  | 0.0644    | -23.0635  | <0.01   |
|         | DatasetNS5  | -1.3137  | 0.0650    | -20.2233  | <0.01   |
|         | Sample.Size | -0.0023  | 0.0004    | -6.0930   | <0.01   |
| BF      | (Intercept) | 2.0045   | 0.0696    | 28.7889   | <0.01   |
|         | DatasetNS6  | -1.4331  | 0.0739    | -19.4020  | <0.01   |
|         | DatasetNS5  | -1.2617  | 0.0747    | -16.8860  | <0.01   |
|         | Sample.Size | -0.0038  | 0.0009    | -4.2598   | <0.01   |

CM: Residual standard error: 0.2226 on 68 degrees of freedom, Multiple R-squared: 0.9059, Adjusted R-squared: 0.9017, F-statistic: 218.1 on 3 and 68 DF, p-value: < 2.2e-16;

BF: Residual standard error: 0.2271 on 53 degrees of freedom, Multiple R-squared: 0.8948, Adjusted R-squared: 0.8888, F-statistic: 150.2 on 3 and 53 DF, p-value: < 2.2e-16

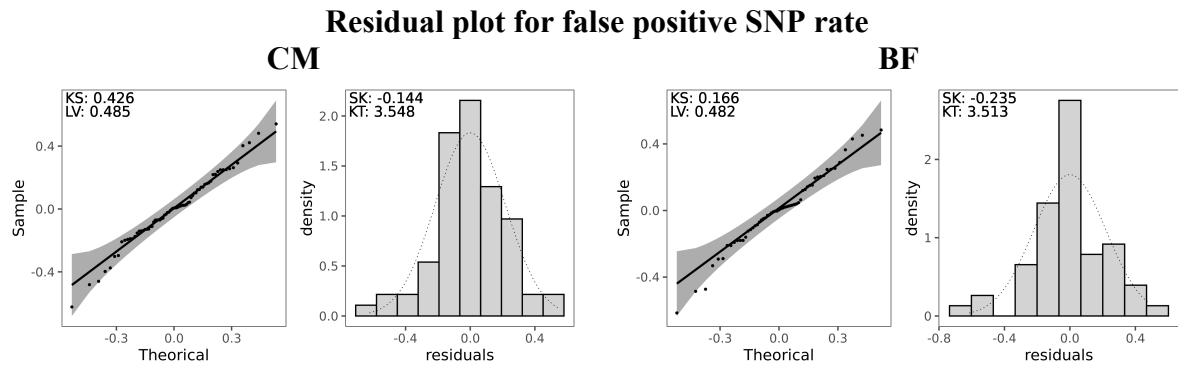


**Figure A3.** Q-Q Plot and histogram of residuals for false positive SNP rate. KS: Kolmogorov-Smirnov normality test p-value. LV: Levene test p-value. SK: Skewness estimate. KT: Kurtosis estimate. Grey bands represent the 95% confidence interval based on the Gaussian distribution.

**Table A2.** Linear regression estimates for false positive SNP rate, including MAF ranges coefficient.

| Dataset | term              | estimate | std.error | statistic | p.value |
|---------|-------------------|----------|-----------|-----------|---------|
| CM      | (Intercept)       | 0.7176   | 0.0604    | 11.8739   | <0.01   |
|         | DatasetNS6        | -0.9143  | 0.0557    | -16.4286  | <0.01   |
|         | DatasetNS5        | -0.6121  | 0.0554    | -11.0572  | <0.01   |
|         | Sample.Size       | -0.0023  | 0.0003    | -6.8826   | <0.01   |
|         | MafRange(0.2,0.3] | -0.7514  | 0.0641    | -11.7171  | <0.01   |
|         | MafRange(0.3,0.4] | -1.3367  | 0.0639    | -20.9104  | <0.01   |
|         | MafRange(0.4,0.5] | -1.4665  | 0.0644    | -22.7835  | <0.01   |
| BF      | (Intercept)       | 0.7419   | 0.0693    | 10.7027   | <0.01   |
|         | DatasetNS6        | -0.8549  | 0.0609    | -14.0462  | <0.01   |
|         | DatasetNS5        | -0.4887  | 0.0609    | -8.0294   | <0.01   |
|         | Sample.Size       | -0.0057  | 0.0007    | -7.7527   | <0.01   |
|         | MafRange(0.2,0.3] | -0.7218  | 0.0701    | -10.3042  | <0.01   |
|         | MafRange(0.3,0.4] | -1.2899  | 0.0701    | -18.4136  | <0.01   |
|         | MafRange(0.4,0.5] | -1.4345  | 0.0704    | -20.3888  | <0.01   |

CM: Residual standard error: 0.3874 on 285 degrees of freedom, Multiple R-squared: 0.7787, Adjusted R-squared: 0.774, F-statistic: 167.1 on 6 and 285 DF, p-value: < 2.2e-16; BF: Residual standard error: 0.3837 on 232 degrees of freedom, Multiple R-squared: 0.7704, Adjusted R-squared: 0.7645, F-statistic: 129.7 on 6 and 232 DF, p-value: < 2.2e-16

**Table A3.** Marginal Means and contrasts estimates for false positive SNP rate.

| Group      | CM           |         | BF           |         |
|------------|--------------|---------|--------------|---------|
|            | Mean         | Signif  | Mean         | Signif  |
| REF        | 1.84 ± 0.05  | a       | 1.82 ± 0.05  | a       |
| NS6        | 0.36 ± 0.04  | c       | 0.39 ± 0.05  | b       |
| NS5        | 0.53 ± 0.05  | b       | 0.56 ± 0.05  | b       |
| Contrast   | Difference   | p-value | Difference   | p-value |
| REF vs NS5 | 1.31 ± 0.06  | <0.01   | 1.26 ± 0.07  | <0.01   |
| REF vs NS6 | 1.48 ± 0.06  | <0.01   | 1.43 ± 0.07  | <0.01   |
| NS6 vs NS5 | -0.17 ± 0.06 | 0.02    | -0.17 ± 0.07 | 0.06    |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means.

**Figure A4.** Boxplots for MAF percentage change (CM dataset). Horizontal facets (legend on the right side) represent sequencing depth. Vertical facets (legend at the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represent percentage change. Grey values: Spearman's correlation coefficient. Top black values: average across chromosomes.

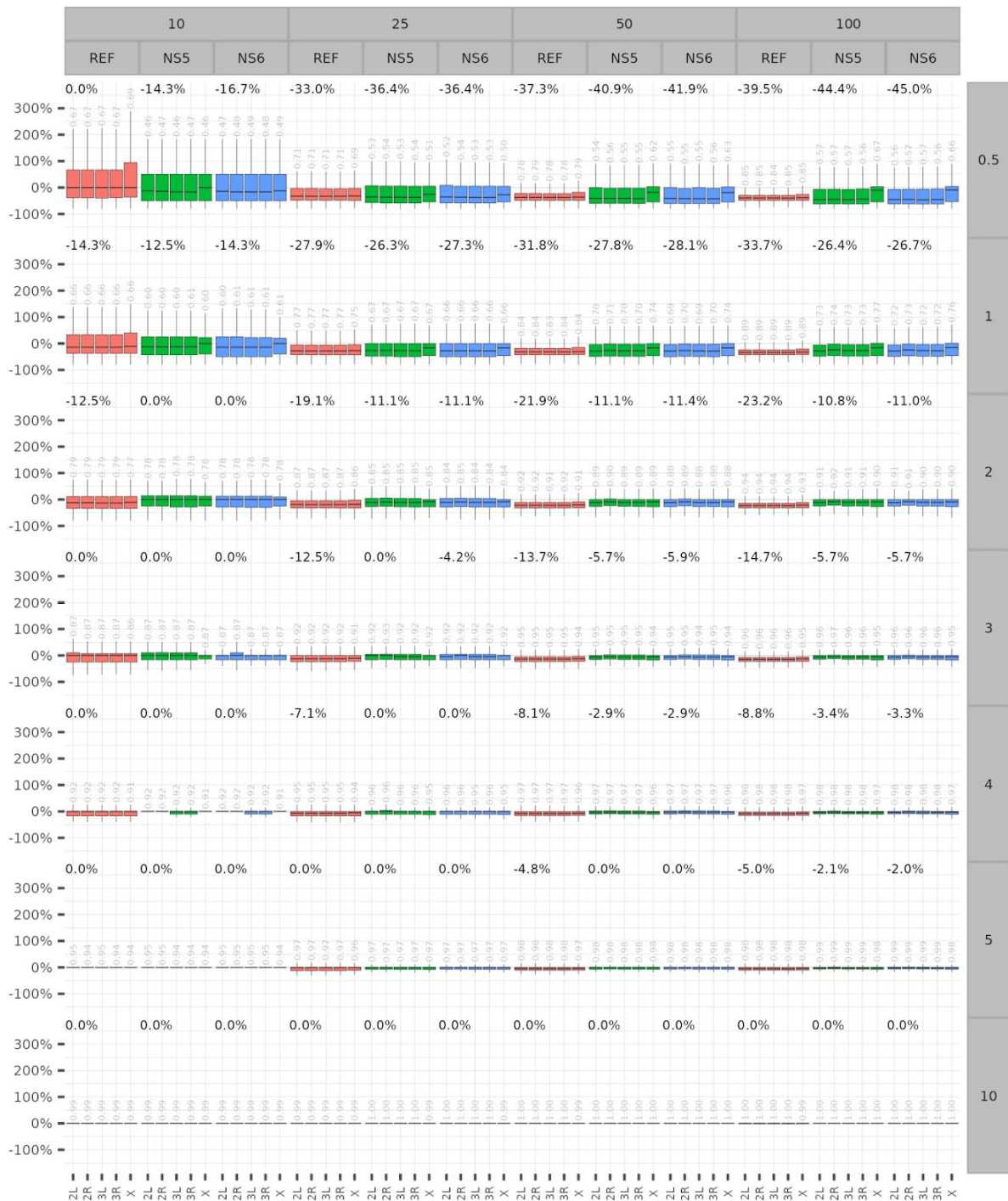


**Figure A5.** Boxplot for MAF percentage change (BF dataset). Horizontal facets (legend on the right side) represent sequencing depth. Vertical facets (legend at the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represent percentage change. Grey values: Spearman's correlation coefficient. Top black values: average across chromosomes.

**Table A4.** Linear regression estimates for MAF percentage change.

| Dataset | term        | estimate | std.error | statistic | p.value |
|---------|-------------|----------|-----------|-----------|---------|
| CM      | (Intercept) | -30.7030 | 1.7131    | -17.9230  | <0.01   |
|         | DatasetNS6  | 4.5802   | 1.4173    | 3.2317    | <0.01   |
|         | DatasetNS5  | 4.7813   | 1.4173    | 3.3735    | <0.01   |
|         | Seq.Depth   | 6.2508   | 0.4078    | 15.3292   | <0.01   |
|         | Sample.Size | -0.0246  | 0.0085    | -2.9140   | <0.01   |
| BF      | (Intercept) | -26.1980 | 1.9730    | -13.2786  | <0.01   |
|         | DatasetNS6  | 5.2607   | 1.5819    | 3.3255    | <0.01   |
|         | DatasetNS5  | 5.6565   | 1.5819    | 3.5757    | <0.01   |
|         | Seq.Depth   | 5.6074   | 0.4567    | 12.2790   | <0.01   |
|         | Sample.Size | -0.0773  | 0.0189    | -4.0891   | <0.01   |

CM: Residual standard error: 4.958 on 68 degrees of freedom, Multiple R-squared: 0.7934, Adjusted R-squared: 0.7813, F-statistic: 65.3 on 4 and 68 DF, p-value: < 2.2e-16;

BF: Residual standard error: 5.003 on 55 degrees of freedom, Multiple R-squared: 0.7693, Adjusted R-squared: 0.7526, F-statistic: 45.86 on 4 and 55 DF, p-value: < 2.2e-16

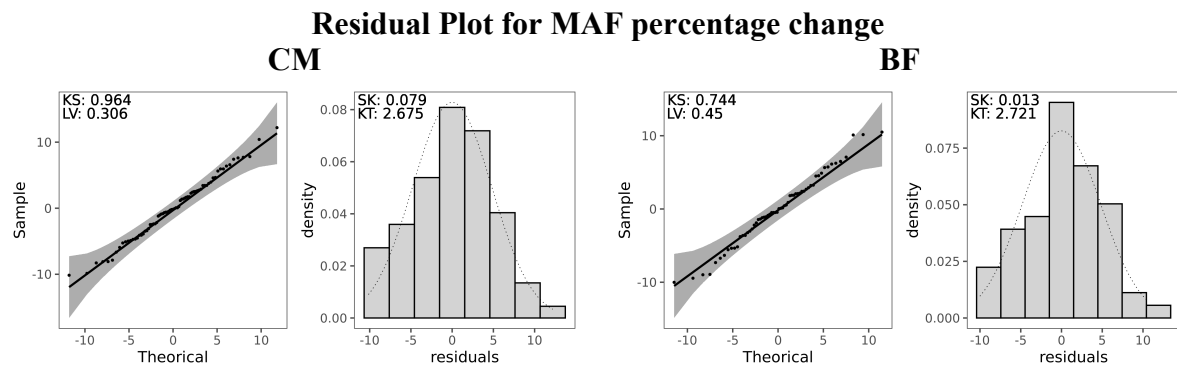


**Figure A6.** Q-Q Plot and histogram of residuals for MAF percentage change. KS: Kolmogorov-Smirnov normality test p-value. LV: Levene test p-value. SK: Skewness estimate. KT: Kurtosis estimate. Grey bands represent the 95% confidence interval based on the Gaussian distribution.

**Table A5.** Marginal Means and contrasts estimates for MAF percentage change.

| Group      | CM            |         | BF            |         |
|------------|---------------|---------|---------------|---------|
|            | Mean          | Signif  | Mean          | Signif  |
| REF        | -13.72 ± 0.99 | b       | -12.95 ± 1.12 | b       |
| NS6        | -9.14 ± 1.01  | a       | -7.69 ± 1.12  | a       |
| NS5        | -8.94 ± 1.01  | a       | -7.29 ± 1.12  | a       |
| Contrast   | Difference    | p-value | Difference    | p-value |
| REF vs NS5 | -4.78 ± 1.42  | <0.01   | -5.66 ± 1.58  | <0.01   |
| REF vs NS6 | -4.58 ± 1.42  | <0.01   | -5.26 ± 1.58  | <0.01   |
| NS6 vs NS5 | -0.20 ± 1.43  | 1.0     | -0.40 ± 1.58  | 1.0     |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means.

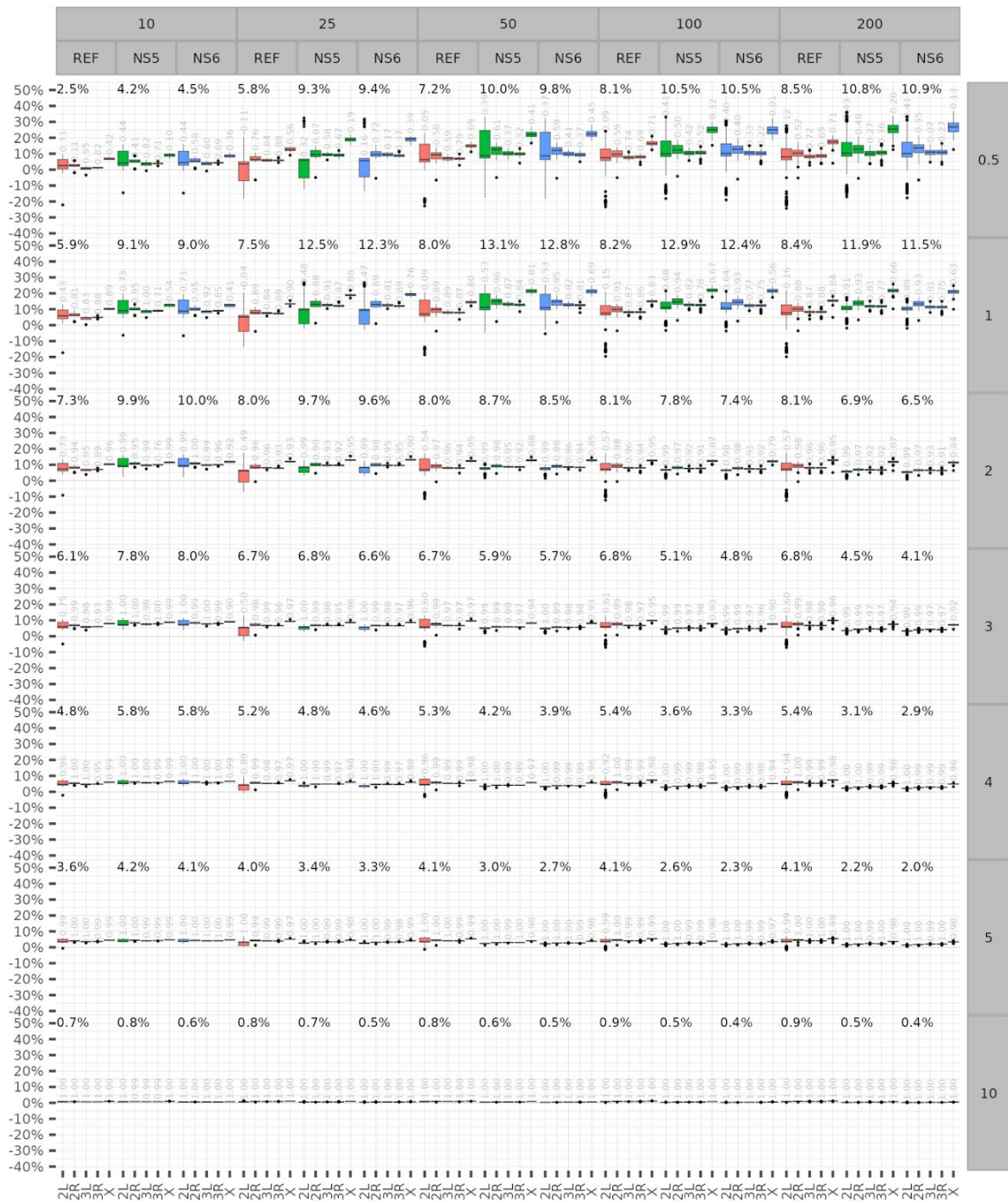


**Figure A7.** Boxplot for homozygosity percentage change (CM dataset). Horizontal facets (legend on the right side) represent sequencing depth. Vertical facets (legend at the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represents percentage change. Grey values: Spearman's correlation coefficient. Top black values: average across chromosomes.

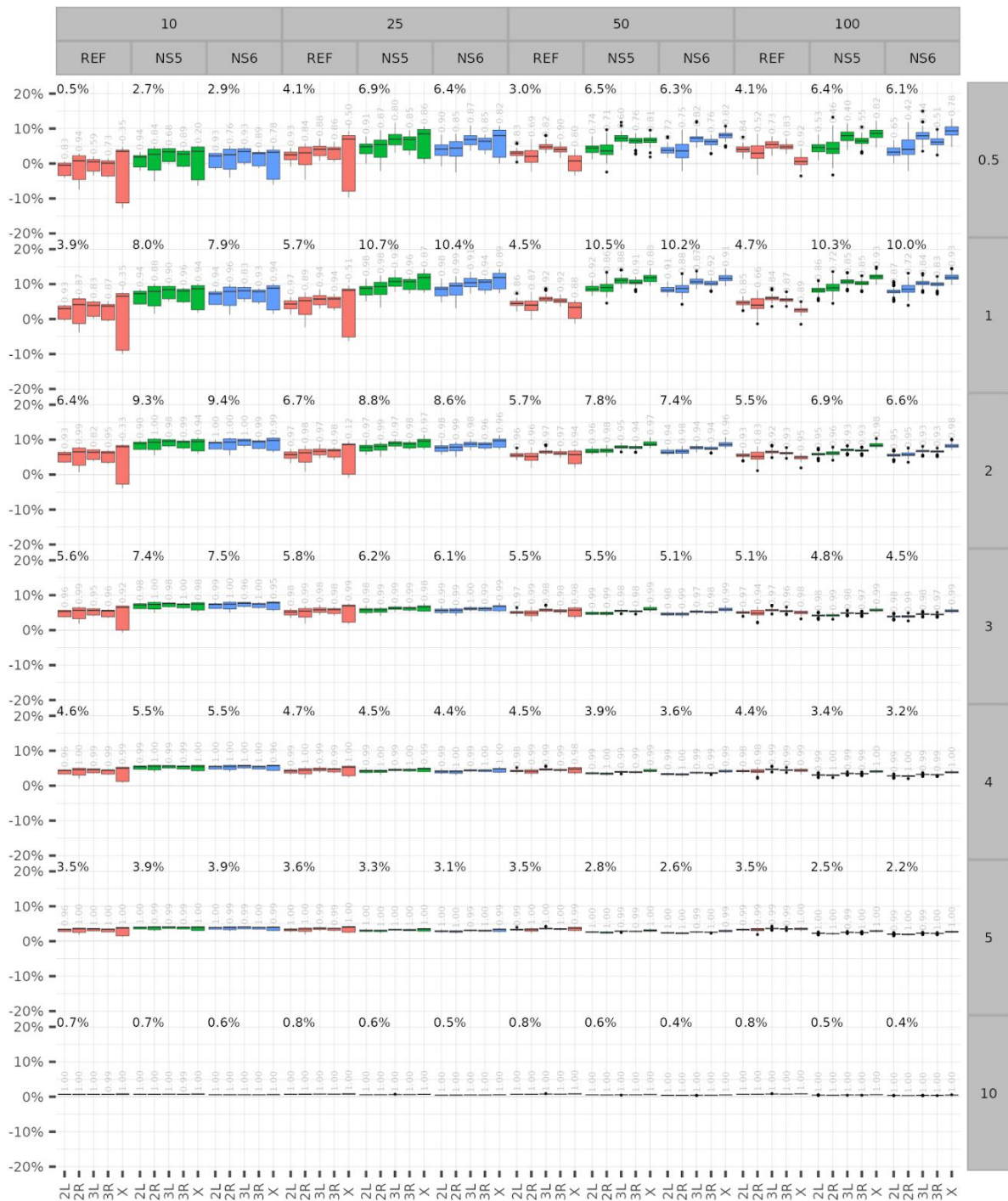


**Figure A8.** Boxplot for homozygosity percentage change (BF dataset). Horizontal facets (legend on the right side) represent sequencing depth. Vertical facets (legend at the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represents percentage change. Grey values: Spearman's correlation coefficient. Top black values: average across chromosomes.

**Table A6.** Linear regression estimates for homozygosity percentage change.

| Dataset | term                 | estimate | std.error | statistic | p.value |
|---------|----------------------|----------|-----------|-----------|---------|
| CM      | (Intercept)          | 9.6441   | 0.5390    | 17.8924   | <0.01   |
|         | DatasetNS6           | 3.8407   | 0.7348    | 5.2269    | <0.01   |
|         | DatasetNS5           | 4.1295   | 0.7348    | 5.6200    | <0.01   |
|         | Seq.Depth            | -0.9913  | 0.1567    | -6.3277   | <0.01   |
|         | Sample.Size          | -0.0054  | 0.0019    | -2.8813   | <0.01   |
|         | DatasetNS6:Seq.Depth | -1.1796  | 0.2215    | -5.3243   | <0.01   |
|         | DatasetNS5:Seq.Depth | -1.2034  | 0.2215    | -5.4317   | <0.01   |
| BF      | (Intercept)          | 6.5887   | 0.4697    | 14.0262   | <0.01   |
|         | DatasetNS6           | 5.2166   | 0.6302    | 8.2772    | <0.01   |
|         | DatasetNS5           | 5.4679   | 0.6302    | 8.6760    | <0.01   |
|         | Seq.Depth            | -0.3895  | 0.1344    | -2.8989   | <0.01   |
|         | Sample.Size          | -0.0119  | 0.0032    | -3.7068   | <0.01   |
|         | DatasetNS6:Seq.Depth | -1.3258  | 0.1900    | -6.9772   | <0.01   |
|         | DatasetNS5:Seq.Depth | -1.3471  | 0.1900    | -7.0893   | <0.01   |

CM: Residual standard error: 1.108 on 68 degrees of freedom, Multiple R-squared: 0.866, Adjusted R-squared: 0.8542, F-statistic: 73.23 on 6 and 68 DF, p-value: < 2.2e-16;

BF: Residual standard error: 0.8498 on 53 degrees of freedom, Multiple R-squared: 0.8791, Adjusted R-squared: 0.8654, F-statistic: 64.24 on 6 and 53 DF, p-value: < 2.2e-16

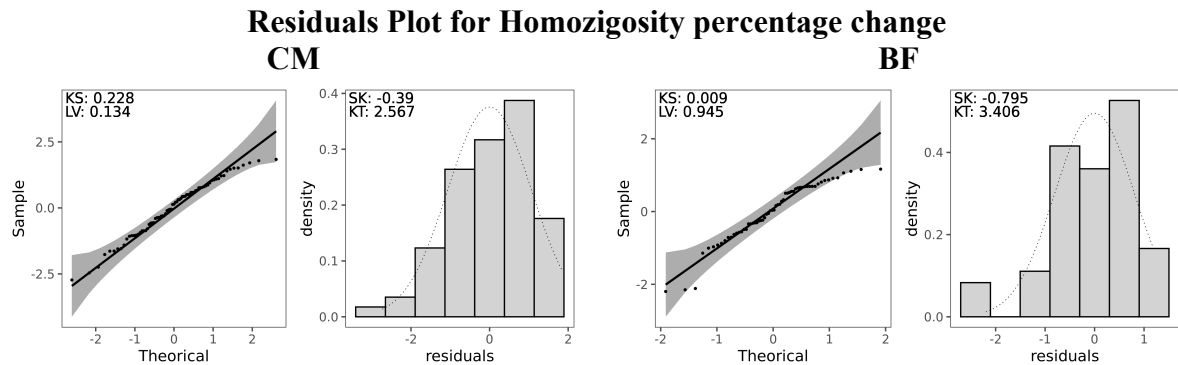


**Figure A9.** Q-Q Plot and histogram of residuals for Homozygosity percentage change. KS: Kolmogorov-Smirnov normality test p-value. LV: Levene test p-value. SK: Skewness estimate. KT: Kurtosis estimate. Grey bands represent the 95% confidence interval based on the Gaussian distribution.

**Table A7.** Marginal Means and contrasts estimates for Homozigosity percentage change.

| Group      | DP | CM           |         | BF           |         |
|------------|----|--------------|---------|--------------|---------|
|            |    | Mean         | Signif  | Mean         | Signif  |
| REF        | 1  | 8.24 ± 0.38  | b       | 5.65 ± 0.33  | b       |
| NS6        | 1  | 10.90 ± 0.38 | a       | 9.54 ± 0.33  | a       |
| NS5        | 1  | 11.17 ± 0.38 | a       | 9.77 ± 0.33  | a       |
| REF        | 2  | 7.25 ± 0.27  | b       | 5.26 ± 0.23  | b       |
| NS6        | 2  | 8.73 ± 0.27  | a       | 7.82 ± 0.23  | a       |
| NS5        | 2  | 8.97 ± 0.27  | a       | 8.03 ± 0.23  | a       |
| REF        | 3  | 6.26 ± 0.22  | a       | 4.87 ± 0.19  | b       |
| NS6        | 3  | 6.56 ± 0.22  | a       | 6.11 ± 0.19  | a       |
| NS5        | 3  | 6.78 ± 0.22  | a       | 6.30 ± 0.19  | a       |
| REF        | 4  | 5.27 ± 0.27  | a       | 4.48 ± 0.23  | a       |
| NS6        | 4  | 4.39 ± 0.27  | a       | 4.39 ± 0.23  | a       |
| NS5        | 4  | 4.58 ± 0.27  | a       | 4.56 ± 0.23  | a       |
| REF        | 5  | 4.27 ± 0.38  | a       | 4.09 ± 0.33  | a       |
| NS6        | 5  | 2.22 ± 0.38  | b       | 2.68 ± 0.33  | b       |
| NS5        | 5  | 2.39 ± 0.38  | b       | 2.82 ± 0.33  | b       |
| Contrast   | DP | Difference   | p-value | Difference   | p-value |
| REF vs NS5 | 1  | -2.93 ± 0.54 | <0.01   | -4.12 ± 0.47 | <0.01   |
| REF vs NS6 | 1  | -2.66 ± 0.54 | <0.01   | -3.89 ± 0.47 | <0.01   |
| NS6 vs NS5 | 1  | -0.27 ± 0.54 | 0.88    | -0.23 ± 0.47 | 0.87    |
| REF vs NS5 | 2  | -1.72 ± 0.38 | <0.01   | -2.77 ± 0.33 | <0.01   |
| REF vs NS6 | 2  | -1.48 ± 0.38 | <0.01   | -2.56 ± 0.33 | <0.01   |
| NS6 vs NS5 | 2  | -0.24 ± 0.38 | 0.80    | -0.21 ± 0.33 | 0.80    |
| REF vs NS5 | 3  | -0.52 ± 0.31 | 0.23    | -1.43 ± 0.27 | <0.01   |
| REF vs NS6 | 3  | -0.30 ± 0.31 | 0.60    | -1.24 ± 0.27 | <0.01   |
| NS6 vs NS5 | 3  | -0.22 ± 0.31 | 0.77    | -0.19 ± 0.27 | 0.77    |
| REF vs NS5 | 4  | 0.68 ± 0.38  | 0.18    | -0.08 ± 0.33 | 0.97    |
| REF vs NS6 | 4  | 0.88 ± 0.38  | 0.06    | 0.09 ± 0.33  | 0.96    |
| NS6 vs NS5 | 4  | -0.19 ± 0.38 | 0.87    | -0.17 ± 0.33 | 0.87    |
| REF vs NS5 | 5  | 1.89 ± 0.54  | <0.01   | 1.27 ± 0.47  | 0.02    |
| REF vs NS6 | 5  | 2.06 ± 0.54  | <0.01   | 1.41 ± 0.47  | 0.01    |
| NS6 vs NS5 | 5  | -0.17 ± 0.54 | 0.95    | -0.14 ± 0.47 | 0.95    |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means. DP: Depth.

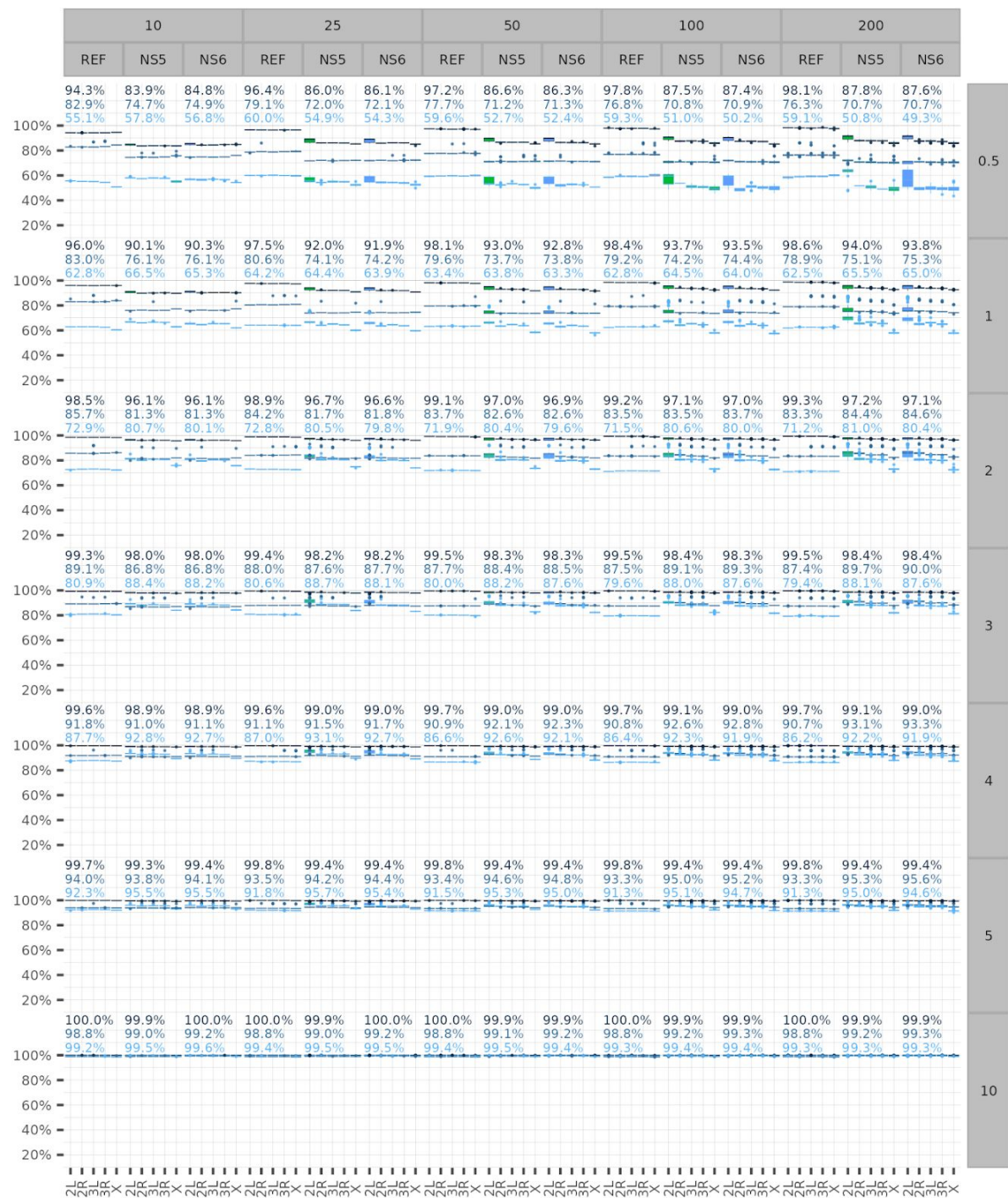


**Figure A10.** Boxplot for genotype concordance rate (CM dataset). Horizontal facets (legend on the right corner) represent sequencing depth. Vertical facets (legend on the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represents the percentage rate. Dark blue value: Homozygotes for the major allele. Blue values: heterozygotes. Light Blue values: Homozygotes for the minor allele.

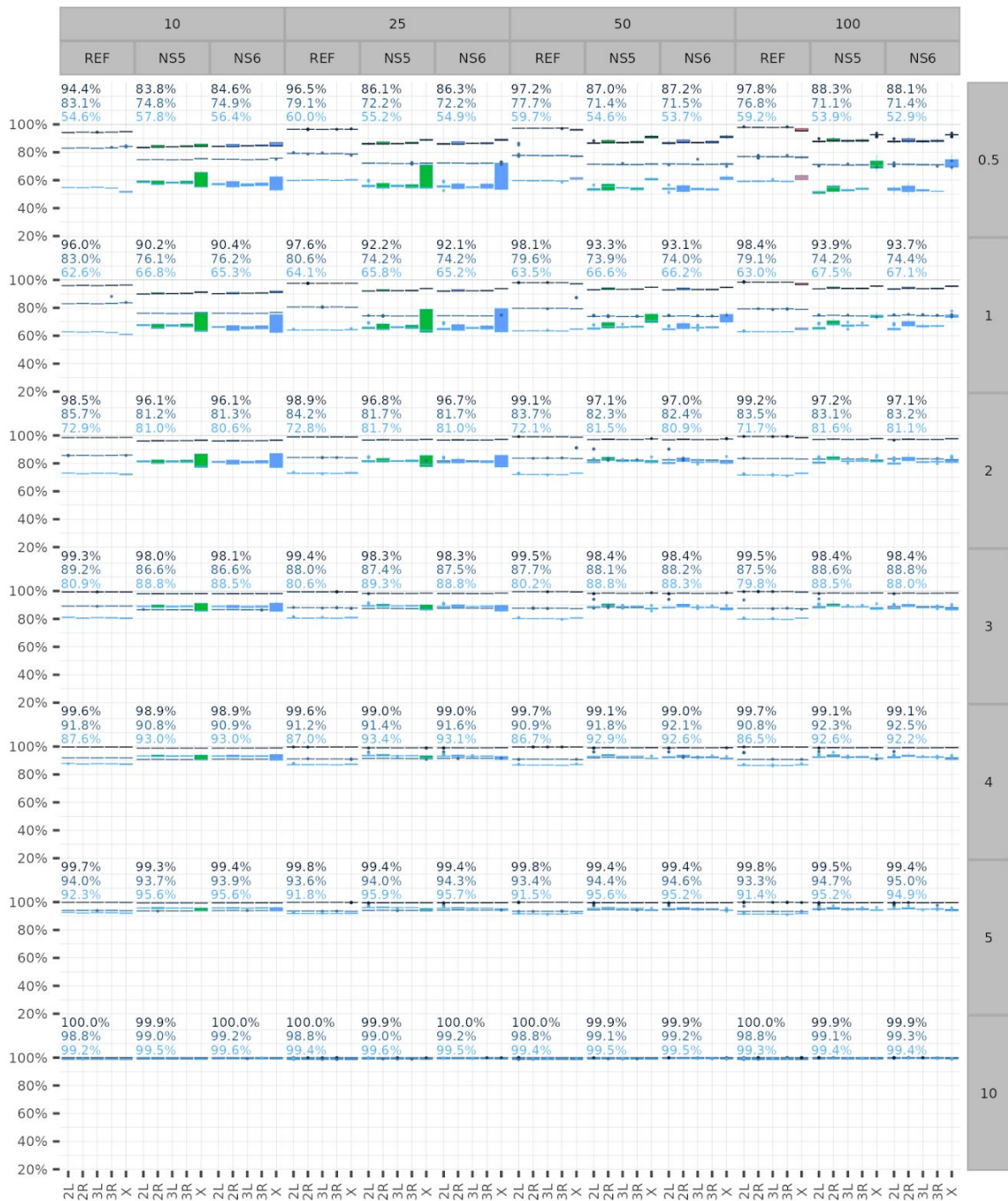


**Figure A11.** Boxplot for genotype concordance rate (BF dataset). Horizontal facets (legend on the right corner) represent sequencing depth. Vertical facets (legend on the top) represent sample size (1st level) and genotyping method (2nd level). The horizontal axis represent chromosome SNP set. The vertical axis represents the percentage rate. Dark blue value: Homozygotes for the major allele. Blue values: heterozygotes. Light Blue values: Homozygotes for the minor allele.

**Table A8.** Linear regression estimates for Genotype concordance rate (Minor Allele Homozygotes).

| dataset | term                 | estimate | std.error | statistic | p.value |
|---------|----------------------|----------|-----------|-----------|---------|
| CM      | (Intercept)          | 0.6946   | 0.0108    | 64.1531   | <0.01   |
|         | DatasetNS6           | 0.0451   | 0.0102    | 4.4142    | <0.01   |
|         | DatasetNS5           | 0.0530   | 0.0102    | 5.1890    | <0.01   |
|         | Seq.Depth            | 0.0438   | 0.0028    | 15.3811   | <0.01   |
|         | I(1/Seq.Depth^2)     | -0.1320  | 0.0087    | -15.1388  | <0.01   |
|         | DatasetNS6:Seq.Depth | 0.0021   | 0.0031    | 0.6859    | 0.5     |
|         | DatasetNS5:Seq.Depth | 0.0010   | 0.0031    | 0.3339    | 0.7     |
| BF      | (Intercept)          | 0.6921   | 0.0112    | 62.0677   | <0.01   |
|         | DatasetNS6           | 0.0597   | 0.0105    | 5.6773    | <0.01   |
|         | DatasetNS5           | 0.0678   | 0.0105    | 6.4508    | <0.01   |
|         | Seq.Depth            | 0.0448   | 0.0029    | 15.2700   | <0.01   |
|         | I(1/Seq.Depth^2)     | -0.1265  | 0.0090    | -14.0835  | <0.01   |
|         | DatasetNS6:Seq.Depth | -0.0008  | 0.0032    | -0.2639   | 0.8     |
|         | DatasetNS5:Seq.Depth | -0.0021  | 0.0032    | -0.6571   | 0.5     |

CM: Residual standard error: 0.0154 on 68 degrees of freedom, Multiple R-squared: 0.9823, Adjusted R-squared: 0.9808, F-statistic: 629.7 on 6 and 68 DF, p-value: < 2.2e-16;

BF: Residual standard error: 0.01418 on 53 degrees of freedom, Multiple R-squared: 0.9846, Adjusted R-squared: 0.9829, F-statistic: 566.4 on 6 and 53 DF, p-value: < 2.2e-16

**Table A9.** Linear regression estimates for Genotype concordance rate (Heterozygotes).

| dataset | term                 | estimate | std.error | statistic | p.value |
|---------|----------------------|----------|-----------|-----------|---------|
| CM      | (Intercept)          | 0.8193   | 0.0076    | 107.2749  | <0.01   |
|         | DatasetNS6           | -0.0486  | 0.0074    | -6.5775   | <0.01   |
|         | DatasetNS5           | -0.0492  | 0.0074    | -6.6552   | <0.01   |
|         | Seq.Depth            | 0.0231   | 0.0020    | 11.5094   | <0.01   |
|         | I(1/Seq.Depth^2)     | -0.0574  | 0.0062    | -9.2491   | <0.01   |
|         | DatasetNS6:Seq.Depth | 0.0143   | 0.0022    | 6.4678    | <0.01   |
|         | DatasetNS5:Seq.Depth | 0.0139   | 0.0022    | 6.3141    | <0.01   |
| BF      | (Intercept)          | 0.8156   | 0.0073    | 111.3733  | <0.01   |
|         | DatasetNS6           | -0.0543  | 0.0071    | -7.6010   | <0.01   |
|         | DatasetNS5           | -0.0546  | 0.0071    | -7.6456   | <0.01   |
|         | Seq.Depth            | 0.0242   | 0.0019    | 12.5444   | <0.01   |
|         | I(1/Seq.Depth^2)     | -0.0507  | 0.0060    | -8.5088   | <0.01   |
|         | DatasetNS6:Seq.Depth | 0.0142   | 0.0021    | 6.6902    | <0.01   |
|         | DatasetNS5:Seq.Depth | 0.0138   | 0.0021    | 6.5005    | <0.01   |

CM: Residual standard error: 0.01079 on 67 degrees of freedom, Multiple R-squared: 0.9754, Adjusted R-squared: 0.9732, F-statistic: 443.6 on 6 and 67 DF, p-value: < 2.2e-16;

BF: Residual standard error: 0.009234 on 52 degrees of freedom, Multiple R-squared: 0.9821, Adjusted R-squared: 0.98, F-statistic: 474.5 on 6 and 52 DF, p-value: < 2.2e-16

**Table A10.** Linear regression estimates for Genotype concordance rate. (Major Allele Homozygotes).

| dataset | term                 | estimate | std.error | statistic | p.value |
|---------|----------------------|----------|-----------|-----------|---------|
| CM      | (Intercept)          | 1.0073   | 0.0035    | 290.2246  | <0.01   |
|         | DatasetNS6           | -0.0462  | 0.0034    | -13.5691  | <0.01   |
|         | DatasetNS5           | -0.0447  | 0.0034    | -13.1261  | <0.01   |
|         | Seq.Depth            | -0.0021  | 0.0009    | -2.3398   | 0.02    |
|         | I(1/Seq.Depth^2)     | -0.0358  | 0.0029    | -12.5024  | <0.01   |
|         | DatasetNS6:Seq.Depth | 0.0096   | 0.0010    | 9.4833    | <0.01   |
|         | DatasetNS5:Seq.Depth | 0.0092   | 0.0010    | 9.1298    | <0.01   |
| BF      | (Intercept)          | 0.6921   | 0.0112    | 62.0677   | <0.01   |
|         | DatasetNS6           | 0.0597   | 0.0105    | 5.6773    | <0.01   |
|         | DatasetNS5           | 0.0678   | 0.0105    | 6.4508    | <0.01   |
|         | Seq.Depth            | 0.0448   | 0.0029    | 15.2700   | <0.01   |
|         | I(1/Seq.Depth^2)     | -0.1265  | 0.0090    | -14.0835  | <0.01   |
|         | DatasetNS6:Seq.Depth | -0.0008  | 0.0032    | -0.2639   | 0.8     |
|         | DatasetNS5:Seq.Depth | -0.0021  | 0.0032    | -0.6571   | 0.5     |

CM: Residual standard error: 0.004824 on 65 degrees of freedom, Multiple R-squared: 0.9487, Adjusted R-squared: 0.9439, F-statistic: 200.2 on 6 and 65 DF, p-value: < 2.2e-16;

BF: Residual standard error: 0.01418 on 53 degrees of freedom, Multiple R-squared: 0.9846, Adjusted R-squared: 0.9829, F-statistic: 566.4 on 6 and 53 DF, p-value: < 2.2e-16

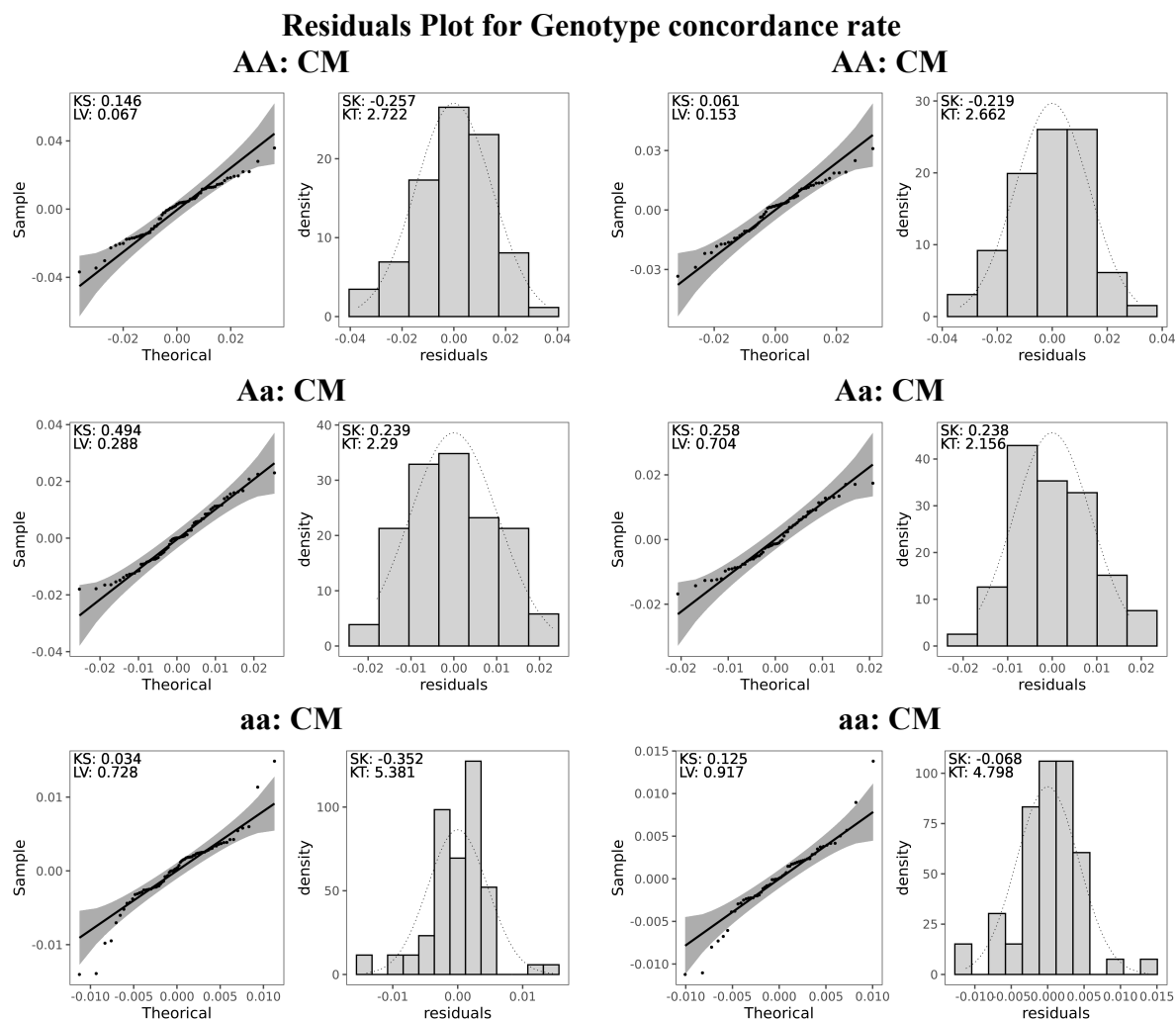


**Figure A12.** Q-Q Plot and histogram of residuals for genotype concordance rate. KS: Kolmogorov-Smirnov normality test p-value. LV: Levene test p-value. SK: Skewness estimate. KT: Kurtosis estimate. Grey bands represent the 95% confidence interval based on the Gaussian distribution. AA, Aa and aa: Major allele homozygotes, Heterozygotes and Minor Allele Homozygotes, respectively.

**Table A11.** Marginal Means and contrasts estimates for Genotype concordance rate (Minor Allele Homozygotes).

| Group      | DP | CM               |         | BF               |         |
|------------|----|------------------|---------|------------------|---------|
|            |    | Mean             | Signif  | Mean             | Signif  |
| REF        | 1  | 0.6064 ± 0.0059  | b       | 0.6104 ± 0.0061  | b       |
| NS6        | 1  | 0.6536 ± 0.0059  | a       | 0.6693 ± 0.0061  | a       |
| NS5        | 1  | 0.6604 ± 0.0059  | a       | 0.6762 ± 0.0061  | a       |
| REF        | 2  | 0.7492 ± 0.0044  | b       | 0.7501 ± 0.0045  | b       |
| NS6        | 2  | 0.7985 ± 0.0044  | a       | 0.8081 ± 0.0045  | a       |
| NS5        | 2  | 0.8043 ± 0.0044  | a       | 0.8138 ± 0.0045  | a       |
| REF        | 3  | 0.8114 ± 0.0035  | b       | 0.8125 ± 0.0036  | b       |
| NS6        | 3  | 0.8628 ± 0.0035  | a       | 0.8697 ± 0.0036  | a       |
| NS5        | 3  | 0.8675 ± 0.0035  | a       | 0.8741 ± 0.0036  | a       |
| REF        | 4  | 0.8616 ± 0.0038  | b       | 0.8634 ± 0.0039  | b       |
| NS6        | 4  | 0.9152 ± 0.0038  | a       | 0.9198 ± 0.0039  | a       |
| NS5        | 4  | 0.9187 ± 0.0038  | a       | 0.9229 ± 0.0039  | a       |
| REF        | 5  | 0.9084 ± 0.0055  | b       | 0.9111 ± 0.0057  | b       |
| NS6        | 5  | 0.9641 ± 0.0055  | a       | 0.9666 ± 0.0057  | a       |
| NS5        | 5  | 0.9666 ± 0.0055  | a       | 0.9685 ± 0.0057  | a       |
| Contrast   | DP | Difference       | p-value | Difference       | p-value |
| REF vs NS5 | 1  | -0.0540 ± 0.0075 | <0.01   | -0.0658 ± 0.0078 | <0.01   |
| REF vs NS6 | 1  | -0.0472 ± 0.0075 | <0.01   | -0.0589 ± 0.0078 | <0.01   |
| NS6 vs NS5 | 1  | -0.0068 ± 0.0075 | 0.6     | -0.0069 ± 0.0078 | 0.7     |
| REF vs NS5 | 2  | -0.0550 ± 0.0053 | <0.01   | -0.0637 ± 0.0055 | <0.01   |
| REF vs NS6 | 2  | -0.0493 ± 0.0053 | <0.01   | -0.0580 ± 0.0055 | <0.01   |
| NS6 vs NS5 | 2  | -0.0057 ± 0.0053 | 0.5     | -0.0056 ± 0.0055 | 0.6     |
| REF vs NS5 | 3  | -0.0561 ± 0.0044 | <0.01   | -0.0616 ± 0.0045 | <0.01   |
| REF vs NS6 | 3  | -0.0514 ± 0.0044 | <0.01   | -0.0572 ± 0.0045 | <0.01   |
| NS6 vs NS5 | 3  | -0.0047 ± 0.0044 | 0.5     | -0.0044 ± 0.0045 | 0.6     |
| REF vs NS5 | 4  | -0.0571 ± 0.0053 | <0.01   | -0.0595 ± 0.0055 | <0.01   |
| REF vs NS6 | 4  | -0.0535 ± 0.0053 | <0.01   | -0.0564 ± 0.0055 | <0.01   |
| NS6 vs NS5 | 4  | -0.0036 ± 0.0053 | 0.8     | -0.0031 ± 0.0055 | 0.8     |
| REF vs NS5 | 5  | -0.0581 ± 0.0075 | <0.01   | -0.0574 ± 0.0078 | <0.01   |
| REF vs NS6 | 5  | -0.0556 ± 0.0075 | <0.01   | -0.0555 ± 0.0078 | <0.01   |
| NS6 vs NS5 | 5  | -0.0025 ± 0.0075 | 0.9     | -0.0019 ± 0.0078 | 1.0     |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means. DP: Depth.

**Table A12.** Marginal Means and contrasts estimates for Genotype concordance rate (Heterozygotes).

| Group      | DP | CM               |         | BF               |         |
|------------|----|------------------|---------|------------------|---------|
|            |    | Mean             | Signif  | Mean             | Signif  |
| REF        | 1  | 0.7850 ± 0.0045  | a       | 0.7891 ± 0.0044  | a       |
| NS6        | 1  | 0.7507 ± 0.0041  | b       | 0.7490 ± 0.0040  | b       |
| NS5        | 1  | 0.7498 ± 0.0041  | b       | 0.7483 ± 0.0040  | b       |
| REF        | 2  | 0.8512 ± 0.0031  | a       | 0.8513 ± 0.0030  | a       |
| NS6        | 2  | 0.8311 ± 0.0031  | b       | 0.8255 ± 0.0029  | b       |
| NS5        | 2  | 0.8299 ± 0.0031  | b       | 0.8243 ± 0.0029  | b       |
| REF        | 3  | 0.8823 ± 0.0024  | a       | 0.8826 ± 0.0023  | a       |
| NS6        | 3  | 0.8765 ± 0.0024  | ab      | 0.8710 ± 0.0023  | b       |
| NS5        | 3  | 0.8749 ± 0.0024  | b       | 0.8694 ± 0.0023  | b       |
| REF        | 4  | 0.9083 ± 0.0026  | a       | 0.9092 ± 0.0025  | a       |
| NS6        | 4  | 0.9167 ± 0.0026  | a       | 0.9118 ± 0.0025  | a       |
| NS5        | 4  | 0.9148 ± 0.0026  | a       | 0.9099 ± 0.0025  | a       |
| REF        | 5  | 0.9327 ± 0.0039  | b       | 0.9346 ± 0.0037  | b       |
| NS6        | 5  | 0.9555 ± 0.0039  | a       | 0.9514 ± 0.0037  | a       |
| NS5        | 5  | 0.9532 ± 0.0039  | a       | 0.9491 ± 0.0037  | a       |
| Contrast   | DP | Difference       | p-value | Difference       | p-value |
| REF vs NS5 | 1  | 0.0353 ± 0.0055  | <0.01   | 0.0408 ± 0.0053  | <0.01   |
| REF vs NS6 | 1  | 0.0344 ± 0.0055  | <0.01   | 0.0401 ± 0.0053  | <0.01   |
| NS6 vs NS5 | 1  | 0.0009 ± 0.0053  | 0.98    | 0.0007 ± 0.0051  | 0.99    |
| REF vs NS5 | 2  | 0.0213 ± 0.0039  | <0.01   | 0.0270 ± 0.0037  | <0.01   |
| REF vs NS6 | 2  | 0.0201 ± 0.0039  | <0.01   | 0.0259 ± 0.0037  | <0.01   |
| NS6 vs NS5 | 2  | 0.0013 ± 0.0037  | 0.94    | 0.0011 ± 0.0036  | 0.95    |
| REF vs NS5 | 3  | 0.0074 ± 0.0031  | 0.05    | 0.0132 ± 0.0030  | <0.01   |
| REF vs NS6 | 3  | 0.0058 ± 0.0031  | 0.15    | 0.0116 ± 0.0030  | <0.01   |
| NS6 vs NS5 | 3  | 0.0016 ± 0.0031  | 0.86    | 0.0015 ± 0.0029  | 0.86    |
| REF vs NS5 | 4  | -0.0065 ± 0.0037 | 0.19    | -0.0007 ± 0.0036 | 0.98    |
| REF vs NS6 | 4  | -0.0085 ± 0.0037 | 0.07    | -0.0026 ± 0.0036 | 0.75    |
| NS6 vs NS5 | 4  | 0.0019 ± 0.0037  | 0.86    | 0.0019 ± 0.0036  | 0.85    |
| REF vs NS5 | 5  | -0.0205 ± 0.0053 | <0.01   | -0.0145 ± 0.0051 | 0.02    |
| REF vs NS6 | 5  | -0.0227 ± 0.0053 | <0.01   | -0.0168 ± 0.0051 | <0.01   |
| NS6 vs NS5 | 5  | 0.0023 ± 0.0053  | 0.90    | 0.0023 ± 0.0051  | 0.89    |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means. DP: Depth.

**Table A13.** Marginal Means and contrasts estimates for Genotype concordance rate (Major Allele Homozygotes).

| Group      | DP | CM               |         | BF               |         |
|------------|----|------------------|---------|------------------|---------|
|            |    | Mean             | Signif  | Mean             | Signif  |
| REF        | 1  | 0.9694 ± 0.0020  | a       | 0.9669 ± 0.0022  | a       |
| NS6        | 1  | 0.9327 ± 0.0020  | b       | 0.9323 ± 0.0022  | b       |
| NS5        | 1  | 0.9339 ± 0.0020  | b       | 0.9334 ± 0.0022  | b       |
| REF        | 2  | 0.9941 ± 0.0014  | a       | 0.9931 ± 0.0015  | a       |
| NS6        | 2  | 0.9670 ± 0.0014  | b       | 0.9675 ± 0.0015  | b       |
| NS5        | 2  | 0.9678 ± 0.0014  | b       | 0.9682 ± 0.0015  | b       |
| REF        | 3  | 0.9969 ± 0.0011  | a       | 0.9965 ± 0.0012  | a       |
| NS6        | 3  | 0.9794 ± 0.0011  | b       | 0.9798 ± 0.0012  | b       |
| NS5        | 3  | 0.9798 ± 0.0011  | b       | 0.9802 ± 0.0012  | b       |
| REF        | 4  | 0.9965 ± 0.0012  | a       | 0.9965 ± 0.0012  | a       |
| NS6        | 4  | 0.9886 ± 0.0012  | b       | 0.9888 ± 0.0012  | b       |
| NS5        | 4  | 0.9887 ± 0.0012  | b       | 0.9888 ± 0.0012  | b       |
| REF        | 5  | 0.9952 ± 0.0017  | a       | 0.9956 ± 0.0018  | a       |
| NS6        | 5  | 0.9968 ± 0.0017  | a       | 0.9968 ± 0.0018  | a       |
| NS5        | 5  | 0.9965 ± 0.0017  | a       | 0.9964 ± 0.0018  | a       |
| Contrast   | DP | Difference       | p-value | Difference       | p-value |
| REF vs NS5 | 1  | 0.0355 ± 0.0025  | <0.01   | 0.0335 ± 0.0027  | <0.01   |
| REF vs NS6 | 1  | 0.0366 ± 0.0025  | <0.01   | 0.0346 ± 0.0027  | <0.01   |
| NS6 vs NS5 | 1  | -0.0012 ± 0.0025 | 0.9     | -0.0011 ± 0.0027 | 0.9     |
| REF vs NS5 | 2  | 0.0263 ± 0.0018  | <0.01   | 0.0249 ± 0.0019  | <0.01   |
| REF vs NS6 | 2  | 0.0271 ± 0.0018  | <0.01   | 0.0256 ± 0.0019  | <0.01   |
| NS6 vs NS5 | 2  | -0.0008 ± 0.0018 | 0.9     | -0.0007 ± 0.0019 | 0.9     |
| REF vs NS5 | 3  | 0.0171 ± 0.0014  | <0.01   | 0.0163 ± 0.0015  | <0.01   |
| REF vs NS6 | 3  | 0.0175 ± 0.0014  | <0.01   | 0.0167 ± 0.0015  | <0.01   |
| NS6 vs NS5 | 3  | -0.0004 ± 0.0014 | 0.9     | -0.0004 ± 0.0015 | 1.0     |
| REF vs NS5 | 4  | 0.0079 ± 0.0017  | <0.01   | 0.0077 ± 0.0018  | <0.01   |
| REF vs NS6 | 4  | 0.0080 ± 0.0017  | <0.01   | 0.0077 ± 0.0018  | <0.01   |
| NS6 vs NS5 | 4  | -0.0001 ± 0.0017 | 1.0     | -0.0000 ± 0.0018 | 1.0     |
| REF vs NS5 | 5  | -0.0013 ± 0.0024 | 0.8     | -0.0009 ± 0.0025 | 0.9     |
| REF vs NS6 | 5  | -0.0016 ± 0.0024 | 0.8     | -0.0012 ± 0.0025 | 0.9     |
| NS6 vs NS5 | 5  | 0.0003 ± 0.0024  | 1.0     | 0.0003 ± 0.0025  | 1.0     |

Mean: marginal mean ± standard error. Difference: Difference between marginal means ± standard error. Signif: Groups sharing the same letter are not statistically different from each other and groups with different letters indicate significant differences between marginal means. DP: Depth.

**Table A14.** Summary of models used in the linear regression analyses.

| Dependent Variable | Model                                                                                                | Dropped Variables | R <sup>2</sup> | p-value |
|--------------------|------------------------------------------------------------------------------------------------------|-------------------|----------------|---------|
| False positive     | $y_i \sim \beta_0 + \beta_1 T_i + \beta_2 SS_i + \epsilon_i$                                         | DP                | 0.90           | <0.01   |
| SNP rate           |                                                                                                      |                   | 0.88           | <0.01   |
| False positive     | $y_i \sim \beta_0 + \beta_1 T_i + \beta_2 SS_i + \beta_3 MR_i + \epsilon_i$                          | DP                | 0.77           | <0.01   |
| with MAF range     |                                                                                                      |                   | 0.76           | <0.01   |
| MAF                |                                                                                                      |                   | 0.78           | <0.01   |
| percentage change  | $y_i \sim \beta_0 + \beta_1 T_i + \beta_2 SS_i + \beta_3 DP_i + \epsilon_i$                          | -                 | 0.75           | <0.01   |
| Homozygosity       | $y_i \sim \beta_0 + \beta_1 T_i + \beta_2 SS_i + \beta_3 DP_i + \beta_4 T_i DP_i + \epsilon_i$       | -                 | 0.85           | <0.01   |
| percentage change  |                                                                                                      |                   | 0.86           | <0.01   |
| Genotype           | $y_i \sim \beta_0 + \beta_1 T_i + \beta_2 DP_i + \beta_3 (1/DP_i^2) + \beta_4 T_i DP_i + \epsilon_i$ | SS                | 0.98           | <0.01   |
| concordance rate   |                                                                                                      |                   | 0.98           | <0.01   |

$\beta_0$ : intercept based on REF; T: Genotyping methodology. SS: Sample Size; DP: Sequencing Depth. MR: Binned MAF range as following: (0.2,0.3] as intercept, (0.3,0.4] and (0.4,0.5].  $\epsilon$ : Random Error. R<sup>2</sup>: Coefficient of determination. p-value: F-test p-value for Model and a null model test. Dropped Variables: Independent variables dropped during the modeling process due to statistical insignificance.