UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM GENÉTICA

Lucca Viana Aguiar

ESTUDOS DE ASSOCIAÇÃO ENTRE GENÓTIPOS E NÍVEIS DE MEDIADORES INFLAMATÓRIOS NA POPULAÇÃO BRASILEIRA MISCIGENADA

BELO HORIZONTE
2025

Lucca Viana Aguiar

ESTUDOS DE ASSOCIAÇÃO ENTRE GENÓTIPOS E NÍVEIS DE MEDIADORES INFLAMATÓRIOS NA POPULAÇÃO BRASILEIRA MISCIGENADA

Dissertação apresentada ao Programa Interunidades de Pós-Graduação em Genética do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Mestre em Genética.

Orientador: Prof. Eduardo Martin Tarazona Santos

BELO HORIZONTE
2025

043 Aguiar, Lucca Viana.
Estudos de associação entre genótipos e níveis de mediadores inflamatórios na população brasileira miscigenada [manuscrito] / Lucca Viana Aguiar. – 2025.
73 f. : il. ; 29,5 cm.

Orientador: Prof. Eduardo Martin Tarazona Santos.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Inflamação. 3. Citocinas. 4. Interleucinas. 5. Quimiocinas. 6. Proteína C-Reativa. I. Santos, Eduardo Martin Tarazona. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575

Ficha catalográfica elaborada pela Bibliotecária Fabiane C M Reis. CRB/6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS

"Estudos de Associação Entre Genótipos e Níveis de Mediadores Inflamatórios na População Brasileira Miscigenada"

Lucca Viana Aguiar

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Martin Tarazona Santos - Orientador
UFMG

Profa Ana Lucia Brunialti Godard
UFMG

Prof. Fabio Nogueira Demarqui
UFMG

Belo Horizonte, 16 de setembro de 2025.

Documento assinado eletronicamente por **Eduardo Martin Tarazona Santos**, **Professor do Magistério Superior**, em 16/09/2025, às 13:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Lucia Brunialti Godard**, **Professora do Magistério Superior**, em 17/09/2025, às 12:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabio Nogueira Demarqui**, **Professor do**

Magistério Superior, em 30/09/2025, às 15:23, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4555669** e o código CRC **FA63F04D**.

Agradecimentos

Agradeço profundamente à minha companheira, Taiana Gomes Rodrigues Bezerra, pelo apoio incondicional durante esta etapa da minha vida. Expresso, igualmente, sincera gratidão à minha família, em especial aos meus pais, Maurício Aguiar do Nascimento e Luciana de Gouvêa Viana.

Ao Prof. Eduardo Tarazona, toda minha gratidão pelos ensinamentos, apoio e incentivo. Aos amigos do Laboratório de Diversidade Genética Humana registro meus agradecimentos pela parceria e solidariedade nos momentos mais desafiadores dessa jornada.

Aos pesquisadores da Coorte de Envelhecimento de Bambuí, registro minha profunda deferência pelo trabalho competente, do qual resultaram inúmeros produtos científicos de grande relevância e que possibilitou a realização desta dissertação. Estendo igualmente meu agradecimento aos participantes que confiaram na iniciativa da Coorte e aceitaram dela fazer parte, fornecendo dados na linha de base e nos acompanhamentos subsequentes, que seguem servindo de fundamento para pesquisas, como a presente dissertação.

Ao Programa de Pós-graduação em Genética da UFMG e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), meus agradecimentos por terem me acolhido e financiado meu projeto.

“NÃO ENTRE EM PÂNICO”

(Douglas Adams, O Guia do Mochileiro das Galáxias, 1979)

Resumo

Os Estudos de Associação de Varredura Genômica (GWAS) têm-se mostrado eficazes na identificação da base genética de diversos fenótipos, como pressão arterial, glicemia e, no presente caso, concentrações séricas de mediadores inflamatórios. Os GWAS realizam uma varredura ao longo do genoma para testar se posições (loci) específicas estão associadas ao fenótipo de interesse. Os mediadores inflamatórios, por sua vez, desempenham papel central na regulação da resposta imune e configuram importantes biomarcadores de saúde, doença e risco para condições crônicas. Suas concentrações séricas representam fenótipos de natureza complexa e multifatorial, nos quais fatores genéticos exercem influência direta sobre a variabilidade individual de suas expressões. A elucidação de genes e variantes genéticas que influenciam vias bioquímicas associadas à expressão torna-se fundamental para a compreensão dos mecanismos moleculares subjacentes. Para isso, foram analisados dados genéticos de 1442 integrantes da Coorte de Envelhecimento de Bambuí com o objetivo de investigar associações de genótipo–fenótipo. Sete mediadores inflamatórios foram considerados no estudo: quimiocinas CCL2, CCL5, CXCL8, CXCL9 e CXCL10, além de interleucina IL-6 e proteína C reativa (PCR). Treze loci, localizados no cromossomo 1, região 1q23, demonstraram associação significativa, ultrapassando o limiar de significância estabelecido ($p < 5\times10^{-8}$). Todas as variantes genéticas foram associadas às concentrações séricas da quimiocina CCL2, na qual rs12075 apresentou menor estatística de teste (p = 4,18E - 20). O polimorfismo rs12075 corresponde a uma substituição nucleotídica no gene *ACKR1*, resultando em uma mutação de sítio trocado. Esse gene codifica uma proteína de membrana que atua como receptor atípico, modulando a biodisponibilidade de múltiplas quimiocinas, incluindo CCL2. Os achados são biologicamente coerentes e compatíveis com evidências prévias de associação neste locus. Contudo, entre as variantes associadas às concentrações séricas de CCL2, oito não apresentam associação descrita com qualquer fenótipo e constituem alvos promissores para investigações futuras. Este estudo é o primeiro a relatar, na população brasileira, a associação entre concentrações séricas de CCL2 e variações no gene *ACKR1*.

Palavras-chave: GWAS; Resposta Inflamatória; Citocinas; Interleucinas; Quimiocinas; Proteína C Reativa Ultrassensível.

Abstract

Genome-wide association studies (GWAS) have proven effective in identifying the genetic basis of various phenotypes, such as blood pressure, blood glucose, and — in the present case — serum concentrations of inflammatory mediators. GWAS performs a scan across the genome to test whether specific positions (loci) are associated with the phenotype of interest. Inflammatory mediators, in turn, play a central role in regulating the immune response and serve as important biomarkers of health, disease, and risk for chronic conditions. Their serum concentrations represent complex, multifactorial phenotypes in which genetic factors directly influence interindividual variability in expression. Elucidating genes and genetic variants that affect biochemical pathways related to expression is therefore crucial to understanding the underlying molecular mechanisms. To this end, genetic data from 1,442 members of the Bambuí Aging Cohort were analyzed to investigate genotype–phenotype associations. Seven inflammatory mediators were considered in the study: the chemokines CCL2, CCL5, CXCL8, CXCL9 and CXCL10, as well as the interleukin IL-6 and C-reactive protein (CRP). Thirteen loci located on chromosome 1, region 1q23, showed significant association, surpassing the established significance threshold ($p < 5\times10^{-8}$). All the genetic variants were associated with serum concentrations of the chemokine CCL2, with rs12075 showing the smallest test statistic ($p = 4.18E - 20$). The polymorphism rs12075 corresponds to a nucleotide substitution in the *ACKR1* gene, resulting in a missense change. This gene encodes a membrane protein that functions as an atypical receptor, modulating the bioavailability of multiple chemokines, including CCL2. The findings are biologically coherent and consistent with previous evidence of association at this locus. However, among the variants associated with serum CCL2 concentrations, eight have no previously described phenotype associations and thus represent promising targets for future investigation. This study is the first to report, in the Brazilian population, an association between serum CCL2 concentrations and variants in the *ACKR1* gene.

Keywords: GWAS; Inflammatory Response; Cytokines; Interleukins; Chemokines; High-Sensitivity C-Reactive Protein.

Lista de Figuras

Lista de tabelas

Lista de abreviaturas e siglas

1KGP3: *1000 Genomes Project Phase 3*
BIC: *Bayesian Information Criterion* (Critério de Informação Bayesiano)
CCR: *Chemokine Receptor* (Receptores de Quimiocinas)
CCL: *Chemokine Ligand* (Ligantes de Quimiocinas)
cCKRs: *conventional chemokine receptors* (Receptores Convencionais de Quimiocinas)
CNV: *Copy Number Variation* (Variação de Número de Cópias)
CXCL: *C-X-C Motif Chemokine Ligand* (Ligantes de Quimiocinas com Motivo C-X-C)
EHW: Equilíbrio de Hardy-Weinberg
EPIGEN: Epidemiologia Genômica de Coortes Brasileiras
eQTLS: Expression quantitative trait loci (Expressão de loci de características quantitativas)
GLM: *Generalized Linear Models* (Modelos Lineares Generalizados)
GLMM: *Generalized Linear Mixed Models* (Modelos Lineares Mistos Generalizados)
GPCR: *G Protein-Coupled Receptor* (Receptor Acoplado à Proteína G)
GRCh37: *Genome Reference Consortium Human Build 37*
GRM: Matriz de Relacionamento Genético
GWAS: *Genome-Wide Association Studies* (Estudos de varredura de associação genômica)
HDL: *High-Density Lipoprotein* (Lipoproteína de Alta Densidade)
hg19: *Human Genome Version 19* (Genoma Humano Versão 19)
HCT: *Hematocrit* (Hematócrito)
HGB: *Hemoglobin* (Hemoglobina)
IBGE: Instituto Brasileiro de Geografia e Estatística
IL: Interleucina
IMC: Índice de Massa Corporal
IP-10: *Interferon Gamma-Induced Protein 10* (Proteína 10 Induzida por Interferon Gama)
IQR: *Interquartile Range* (Intervalo Interquartil)
Kg: Quilograma
LD: *Linkage Disequilibrium* (Desequilíbrio de Ligação)
LDL: *Low-Density Lipoprotein* (Lipoproteína de Baixa Densidade)
LM: *Linear Models* (Modelos Lineares)
LMM: *Linear Mixed Models* (Modelos Lineares Mistos)
LRT: *Likelihood Ratio Test* (Teste da Razão de Verossimilhança)
MAF: *Minor Allele Frequency* (Frequência do Menor Alelo)
MCH: *Mean Corpuscular Hemoglobin* (Hemoglobina Corpuscular Média)
MCP-1: *Monocyte Chemoattractant Protein 1* (Proteína Quimioatraente de Monócitos-1)
m: Metro
ml: Mililitro
MIG: *Monokine Induced by Interferon-Gamma* (Monocina Induzida pelo Interferon Gama)
ML: *Maximum Likelihood* (Máxima Verossimilhança)
PCR: Proteína C Reativa
PCR-us: Proteína C Reativa Ultrassensível
PC: *Principal Component* (Componente Principal)
PCA: *Principal Component Analysis* (Análise da Componente Principal)

pg: Picograma
QQ: *Quantile-Quantile*
RANTES: *Regulated On Activation, Normal T-Cell Expressed and Secreted* (Regulada sob Ativação, Expressa e Secretada por Células T Normais)
RBC: *Red Blood Cells* (Glóbulos Vermelhos / Hemácias)
REML: *Restricted Maximum Likelihood* (Máxima Verossimilhança Restrita)
SABE: Saúde, Bem-Estar e Envelhecimento
SCAALA: *Social Changes, Asthma and Allergy in Latin America Programme* (Programa Mudanças Sociais, Asma e Alergia na América Latina)
SD: *Standard Deviation* (Desvio Padrão)
SE: *Standard Error* (Erro Padrão)
SNP: *Single Nucleotide Polymorphism* (Polimorfismo de Base Única)
SNV: *Single Nucleotide Variants* (Variantes de Base Única)
VLDL: *Very Low-Density Lipoprotein* (Lipoproteína de Muito Baixa Densidade)
WBC: *White Blood Cells* (Glóbulos Brancos / Leucócitos)

Sumário

1. Introdução

Com o avanço das tecnologias de sequenciamento e genotipagem, aliado ao consequente aumento da disponibilidade de dados genômicos, os estudos de associação tornaram-se cada vez mais frequentes nas últimas décadas. O objetivo desses estudos é identificar a base genética de traços e doenças, em uma população. Os *Genome-wide Association Studies* (GWAS), realizam uma varredura ao longo do genoma para testar se posições (*loci*) específicas estão associadas ao fenótipo de interesse. A abordagem mais comum para esses testes são os modelos de regressão, nos quais cada locus é avaliado individualmente em busca de variantes que expliquem parte da variabilidade fenotípica observada na coorte (HIRSCHHORN; DALY, 2005; UFFELMANN *et al.*, 2021).

1.1 Genética de populações humanas e estudos de associação

Os modelos de regressão mais utilizados em estudos de associação genótipo-fenótipo são os modelos lineares (LM) e logísticos. Modelos lineares são aplicados quando a variável dependente é contínua (como pressão arterial, glicemia, ou no presente caso, concentrações séricas de mediadores inflamatórios). As regressões logísticas são empregadas em estudos do tipo caso-controle. Nos modelos de regressão, busca-se explicar a variabilidade do fenótipo de interesse (variável dependente) a partir de covariáveis e genótipos (variáveis independentes), sendo os genótipos codificados conforme o número de cópias de um alelo, como 0, 1 ou 2. Sendo assim, a hipótese nula pressupõe que o alelo ou variante genética não está associada ao fenótipo, ou seja, possui efeito igual a zero, enquanto a hipótese alternativa postula que o alelo ou variante genética exerce efeito significativo sobre o fenótipo, diferente de zero.

Ao realizar estudos de associação em bases genéticas, é fundamental considerar fatores que podem enviesar a análise. Os modelos lineares utilizados em regressões assumem determinadas condições sobre os dados e sobre os erros, que representam variáveis não observáveis incluídas na equação do modelo. Entre essas suposições, destacam-se, por exemplo: a independência estatística dos erros, a relação linear entre variáveis independentes e dependentes, a homocedasticidade (igualdade de variâncias dos erros), a normalidade da distribuição dos erros e a ausência de multicolinearidade entre as covariáveis (MONTGOMERY; PECK; VINING, 2013).

No contexto de dados genéticos, esta dissertação se refere a “variantes” ou loci (plural de locus) do tipo Variante de Nucleotídeo Único (SNV). Um SNV corresponde a uma alteração de um único nucleotídeo em relação a um genoma de referência. A maioria dos SNVs apresenta apenas dois estados possíveis, denominados alelos, e, considerando que os humanos são diplóides, cada indivíduo pode possuir 0, 1 ou 2 cópias de um determinado alelo. Vale destacar que os alelos em diferentes loci distribuídos pelo genoma não são necessariamente independentes. O Desequilíbrio de Ligação (LD) é a associação estatística de alelos em diferentes loci (LEWONTIN, 1974). Devido ao LD, em GWAS, não podemos inferir diretamente que a variante responsável pela associação estatística seja a variante biologicamente causal. Uma variante que ultrapassa o limiar de significância pode estar sendo carregada, em LD, por uma variante causal, que não está presente na análise (efeito caroneiro).

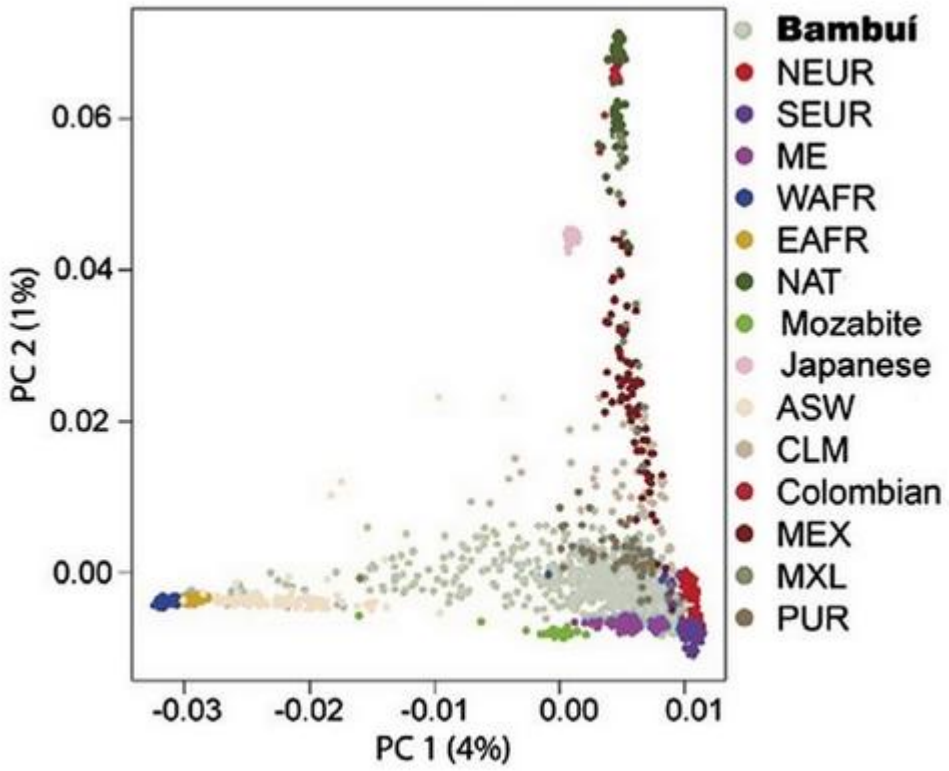
Nos GWAS estamos realizando milhares ou até milhões de testes estatísticos, parte dos quais são independentes e parte não. Assim, é realizada uma correção para múltiplas comparações, em que foi adotado o nível de significância para as análises de GWAS de 5×10^{-8}, conforme proposto originalmente pelo *The International HapMap Consortium* (2005). Esse limiar foi posteriormente validado por estudos independentes, como os de Pe’er *et al.* (2008) e Dudbridge e Gusnanto (2008), sendo levados em consideração simulações e o número efetivo de testes de associação independentes, que depende também do LD entre variantes. Esse limiar genômico continua sendo amplamente utilizado em GWAS, sendo considerado adequado para o controle de erros do tipo I (falsos positivos), especialmente em coortes com número reduzido de indivíduos (CHEN, Z. *et al.*, 2021).

Quando trabalhamos com genética populacional, devemos tomar cuidado com a estrutura da população. Estrutura da população se refere à presença de diferenças sistemáticas na frequência dos alelos entre populações, subpopulações ou núcleos familiares que pode causar associações espúrias e distorções nas estimativas de efeito entre variantes genéticas e traços/doenças (BYUN *et al.*, 2017). A estrutura populacional está relacionada com a ancestralidade, ou seja, com a origem dos alelos, particularmente relevante no Brasil devido à natureza miscigenada da população, com componentes europeus, africanos e indígenas. A estrutura populacional pode ser dividida em dois níveis: estruturação populacional e relações familiares (PRICE *et al.*, 2010).

Para lidar com a estruturação populacional, a abordagem mais utilizada consiste na Análise de Componentes Principais (PCA), realizada a partir dos genótipos dos indivíduos. Essa técnica permite avaliar a variância genética presente na coorte e identificar padrões de estrutura populacional (REICH; PRICE; PATTERSON, 2008), como ilustrado na Erro! Fonte d

e referência não encontrada., podendo ainda ser incorporada como covariável nos modelos de regressão. Em estudos de associação, o uso das Componentes Principais (PCs) não deve ser apenas para corrigir a estruturação populacional em si, mas sim para ajustar fatores confundidores correlacionados com componentes específicos de ancestralidade. Isso ocorre porque parte da variabilidade genotípica que influencia os resultados está diretamente ligada à estrutura populacional, e é essa estrutura que pode gerar sinais espúrios de associação.

Figura 1. Análise de Componentes Principais da Coorte de Envelhecimento de Bambuí, junto com populações de referência.



Plot dos Componentes Principais 1 e 2, com a variância explicada entre parênteses. A partir dos genótipos dos indivíduos, é possível capturar a estrutura genética da população e visualizar agrupamentos relacionados à ancestralidade e similaridade genética. Fonte: Kehdy et al., 2015.

A estrutura genética em nível familiar gera dependência entre as observações, resultando em não independência estatística dos erros. Em populações humanas, essa situação é comum, pois os genótipos de indivíduos aparentados tendem a ser altamente correlacionados. Em uma população grande, com pouco parentesco entre os pares de indivíduos, a suposição de independência estatística dos resíduos consegue ser atendida.

Uma forma de avaliar o grau de parentesco entre indivíduos é por meio da medida de Identidade por Descendência (IBD). Segundo Hedrick, em *Genetics of Populations* (2005), IBD se refere aos alelos derivados de um único alelo portado por um ancestral em comum. Já a Identidade por Estado (IBS), conceito complementar ao IBD, refere-se a alelos que apresentam a mesma forma (ou sequência), mas que não foram necessariamente herdados de um ancestral comum em um passado recente.

Ambos os conceitos servem de base para o cálculo de coeficiente de *Kinship*. Ainda de acordo com Hedrick, em *Genetics of Populations*, coeficiente de *Kinship* é definido como a probabilidade que alelos do mesmo locus amostrados aleatoriamente entre dois indivíduos estejam em IBD. Este coeficiente é importante para análise da coorte, pois, por meio dele, conseguimos ter uma ideia da relação dos pares de indivíduos, para identificar qualquer relação de parentesco que existe na amostra, entender e corrigir para os testes de associação, sendo essenciais para prevenir associações espúrias.

Outra estimativa importante é a Matriz de Relacionamento Genético (GRM), que quantifica a similaridade genética entre pares de indivíduos com base nos Polimorfismos de Nucleotídeo Único (SNPs). A GRM é obtida calculando-se, para cada par de indivíduos ***j*** e ***k***, uma correlação agregada ao longo de todos os loci ***i***, codificando os genótipos, ***x***, como: 0 (homozigoto referência), 1 (heterozigoto) e 2 (homozigoto alternativo). Cada elemento da matriz representa a proporção de variação genética compartilhada entre dois indivíduos (YANG et al., 2011), considerando simultaneamente a frequência alélica (***p***) e o número total de marcadores, SNVs, utilizados (***N***).

$$A_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

A GRM e o coeficiente de *kinship* são conceitos relacionados, mas diferentes: o *kinship* estima a probabilidade de alelos serem IBD, enquanto a GRM mede a similaridade genotípica observada (IBS), sem diferenciar origem ancestral.

Para os estudos de associação com parentes presentes na análise, em que se deseja preservar o número amostral, é necessário contabilizar o parentesco para evitar associações espúrias. Para isso, utilizamos de Modelos Lineares Mistos (LMM), onde incorporamos a GRM como efeito aleatório. Os efeitos aleatórios contabilizam a não-independência dos indivíduos ao modelar a estrutura de correlação genética entre os pares de indivíduos, permitindo

decompor a variância do fenótipo em um componente associado à similaridade genética e um componente residual. Com os modelos mistos conseguimos preservar indivíduos em nossa amostra, possibilitando uma análise mais detalhada de variantes menos comuns na população, e com um maior poder estatístico.

Mesmo ao controlar a estrutura populacional com LMM, é essencial verificar possíveis vieses sistemáticos nos resultados. Nesse contexto, destaca-se o cálculo da inflação genômica (λ) (THE GIANT CONSORTIUM *et al.*, 2011; VAN DEN BERG *et al.*, 2019). Esse cálculo consiste em comparar os valores observados das estatísticas de teste de cada variante com os valores esperados sob a distribuição qui-quadrado, conforme a **Fórmula 1**.

$$\lambda_{GC} = \frac{\text{median}\left(\chi^2_{obs}\right)}{\text{median}\left(\chi^2_{esp}\right)}$$

Fórmula 1

O parâmetro λ atua como um controle genômico, permitindo avaliar se há desvios sistemáticos na distribuição das estatísticas de associação ao longo do genoma. Esse controle auxilia na identificação de potenciais artefatos, como erros de genotipagem, efeitos de lote (*batch effect*), problemas localizados de qualidade dos dados, entre outros. Quando $\lambda \approx 1$, não há indícios de inflação nas estatísticas de teste; valores de $\lambda > 1$ indicam inflação (possível superestimação dos testes), enquanto $\lambda < 1$ sugerem deflação (possível subestimação). De forma prática, valores entre 0,95 e 1,05 são geralmente considerados aceitáveis.

Uma outra ferramenta essencial em estudos de associação é o cálculo do poder estatístico, que representa a probabilidade de rejeitar corretamente a hipótese nula (COHEN, 2013). Em geral, adota-se como referência um poder estatístico mínimo de 80% para garantir confiabilidade na detecção de associações em GWAS (WANG, M.; XU, 2019). No caso de fenótipos contínuos, diversos fatores influenciam essa detecção, incluindo: a frequência alélica na população, o tamanho amostral, o tamanho de efeito detectável, o desvio padrão populacional da variável dependente e o modelo estatístico empregado. Outros elementos também exercem impacto, como o nível de significância adotado, definido, geralmente, como $\alpha=5\times10^{-8}$, e o modelo genético assumido. O modelo genético define como os alelos interagem em um locus. No modelo aditivo, o efeito da variante cresce de forma linear com o número de cópias do alelo, enquanto os modelos dominante e recessivo pressupõem efeito com apenas uma ou com duas cópias, respectivamente (TAM *et al.*, 2019). Neste estudo, utilizou-se o

modelo aditivo, amplamente empregado para fenótipos contínuos por capturar grande parte da variabilidade fenotípica, apresentar menor custo computacional (XU, 2023) e manter bom poder estatístico, mesmo quando o padrão real é dominante. Além disso, evita a necessidade de múltiplos testes para diferentes modelos de herança, embora seu desempenho seja reduzido em casos de herança recessiva (LETTRE; LANGE; HIRSCHHORN, 2007).

Após a realização de todos os procedimentos de controle de qualidade, os resultados do GWAS podem ser interpretados com maior confiança. O *Summary Statistics* reúne, para cada variante testada, as informações essenciais como a frequência alélica, estimativa de efeito (*Odds Ratio* ou β), o erro padrão associado e p valor (**Tabela 1**). Dependendo do software utilizado, também podem estar disponíveis dados adicionais, como o número de amostras incluídas na regressão, o valor da estatística T (em regressões lineares) ou o escore Z de Wald (em regressões logísticas).

Tabela 1. Exemplo de GWAS *Summary Statistics*.

Chr	SNV	bp	A1	A2	Freq	b	se	p
1	rs144434834	723918	A	G	0,03	0,17	0,11	0,15
1	rs3094315	752566	C	T	0,21	-0,01	0,05	0,90
1	rs3131972	752721	T	C	0,23	0,03	0,05	0,58
1	rs12184312	754063	T	G	0,04	0,09	0,09	0,33
1	rs74045212	757691	C	T	0,02	-0,25	0,14	0,08
1	rs114525117	759036	A	G	0,04	0,09	0,10	0,35
1	rs144708130	761764	A	G	0,01	-0,18	0,24	0,45

As colunas são: cromossomo, rsID, posição física em pares de base, alelo de referência (A1), alelo alternativo (A2), frequência do alelo A1 (Freq), tamanho de efeito (β) do alelo de referência, erro padrão associado (SE) e p valor.

A interpretação biológica de um sinal estatístico identificado no *Summary Statistics* é consolidada por meio de análises pós-estudo de associação, baseadas em anotação funcional. Nessa etapa, são considerados aspectos como os blocos de LD da população, comparação das estatísticas de teste e dos tamanhos de efeito com outras populações, avaliação do poder estatístico, consulta a bancos de dados públicos, caracterização da região gênica, análises e estatísticas diagnósticas de regressão, além do estudo dos resíduos. Essas análises permitem interpretar cada associação de forma mais precisa, levando em conta como a estrutura de LD da população pode influenciar os resultados, fornecendo uma visão mais clara das relações genéticas possivelmente causais.

1.2 Fenótipos de estudo

A coorte de Bambuí tem dados para as Neste trabalho, foram analisadas as seguintes quimiocinas e proteínas inflamatórias: CCL2 (proteína quimioatraente de monócitos-1, MCP-1), CCL5 (regulada por ativação, expressa e secretada por células T normais, RANTES), CXCL8 (interleucina-8, IL-8), CXCL9 (monocina induzida por interferon-gama, MIG), CXCL10 (proteína 10 induzida por interferon-gama, IP-10), interleucina-6 (IL-6) e proteína C reativa (PCR). Nesta dissertação, focaremos na associação envolvendo a quimiocina CCL2, que apresentou o sinal genótipo-fenótipo mais consistente e menor estatística de teste.

As quimiocinas são pequenas proteínas envolvidas principalmente na resposta inflamatória, atuando na ativação e no recrutamento (quimiotaxia) de células do sistema imune, além de contribuir para a manutenção da homeostase imunológica. São classificadas em famílias de acordo com a disposição relativa dos resíduos de cisteína próximos à extremidade N-terminal da molécula. No presente estudo temos representantes das famílias CC (β-quimiocinas) e CXC (α-quimiocinas). As quimiocinas CXC possuem um resíduo de aminoácido variado entre os resíduos de cisteínas, conferindo diferenças uma estrutura mais rígida para os dímeros de CXC, garantindo interações estáveis com seus receptores. Em comparação, os dímeros de CC possuem uma interface mais flexível, tornando-as mais versáteis possibilitando a interações com diferentes receptores, podendo atrair uma gama maior de células do sistema imune.

De modo geral, as quimiocinas induzem quimiotaxia ao se ligarem aos seus receptores de membrana convencionais (*conventional chemokine receptors*, cCKRs), que pertencem à superfamília dos Receptores Acoplados à Proteína G (*G protein-coupled receptors*, GPCRs). A ligação ativa a proteína G desencadeia uma cascata intracelular que resulta em alterações funcionais, além de mecanismos de retroalimentação que regulam a intensidade e a duração do sinal. Uma característica importante é que diferentes quimiocinas podem se ligar a diferentes receptores, e cada receptor consegue mediar respostas específicas de acordo com a quimiocina. Essa redundância confere ao sistema imune flexibilidade, robustez e constância na resposta.

Existe uma outra classe de receptores, os chamados Receptores Atípicos de Quimiocinas (*Atypical Chemokine Receptors*, ACKRs). Esses receptores atuam como receptor sequestrador, eles possuem alta afinidade com as quimiocinas, mas não possuem a proteína de membrana G, não ativando a cascata de sinalização. ACKRs regulam a biodisponibilidade das quimiocinas, degradando os seus ligantes, indiretamente modulando as respostas de atividade das cCKRs e ajustando a resposta imunológica geral.

As quimiocinas desempenham múltiplas funções no organismo, atuando de forma coordenada frente a diferentes estímulos. Por isso, são amplamente estudadas e reconhecidas como importantes alvos terapêuticos em diversas condições clínicas. As concentrações séricas das quimiocinas e proteínas inflamatórias representam fenótipos de natureza complexa e multifatorial, nos quais fatores genéticos exercem influência direta sobre a variabilidade individual de suas expressões. Por isso, a elucidação de genes e variantes genéticas que influenciam vias bioquímicas associadas à expressão torna-se fundamental para a compreensão dos mecanismos moleculares subjacentes.

1.3 Coorte de Envelhecimento de Bambuí

A população brasileira é tri-híbrida, sendo produto de miscigenação intensa das populações parentais europeias, africanas e nativo americanas, o que gerou grande heterogeneidade alélica e padrões de LD distintos daqueles observados em populações europeias (SAWYER *et al.*, 2005). Essas estruturas de LD e, consequentemente, os haplótipos, podem modular a expressão fenotípica e, em alguns casos, influenciar respostas a fármacos (RODRIGUES-SOARES *et al.*, 2020). Por esse motivo, identificar as regiões em LD numa população-alvo é crucial para interpretar resultados de GWAS: marcadores associados raramente são a variante causal em si, sendo mais provável que sinalizem uma variante causal em forte LD com eles (HIRSCHHORN; DALY, 2005; IOANNIDIS; PATSOPOULOS; EVANGELOU, 2007).

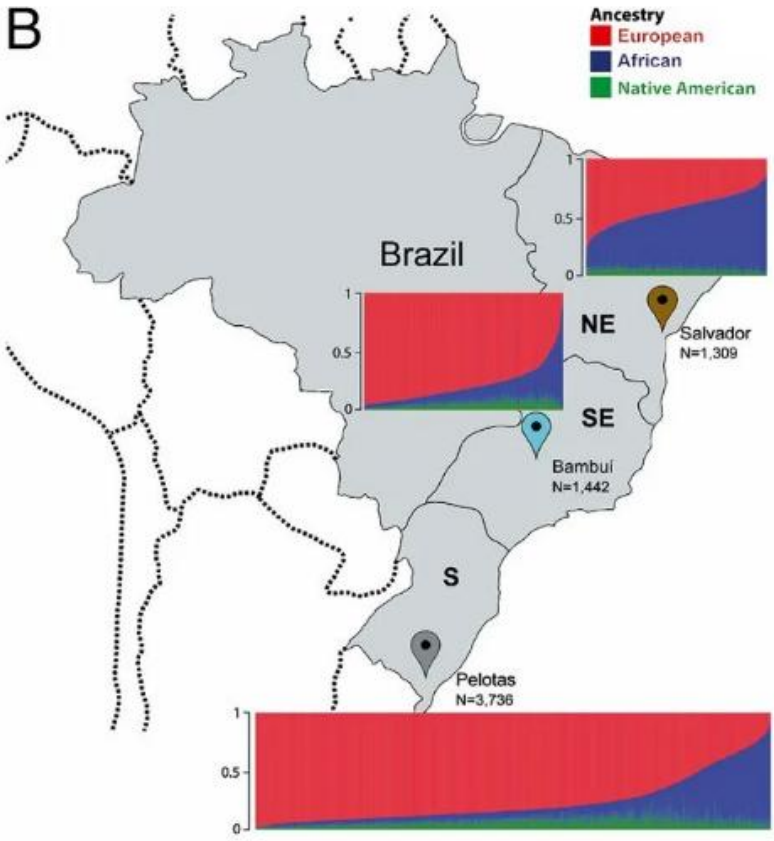
O presente estudo foi conduzido na Coorte de Envelhecimento de Bambuí, localizado na região centro-oeste do estado de Minas Gerais. Até 2022, a cidade de Bambuí abrigava 23.546 pessoas, predominantemente com baixa escolaridade e baixa renda (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2023). A escolha da cidade de Bambuí para a criação da coorte foi baseada em quatro critérios principais: baixa taxa de migração, características sociodemográficas, perfil de mortalidade e viabilidade do estudo em termos de cooperação com os residentes (LIMA-COSTA, M. F.; FIRMO; UCHOA, 2011).

Com o objetivo de compreender melhor a relação entre doenças complexas e a população brasileira, bem como sua prevalência e associação com a ancestralidade, a coorte de Bambuí foi integrada à iniciativa de Epidemiologia de Doenças Complexas em Coortes de Populações Brasileiras (EPIGEN). Esta iniciativa também incluiu a Coorte de Nascimento de Pelotas (VICTORA; BARROS, 2006) e o estudo populacional de Salvador-SCAALA (*Social Changes, Asthma and Allergy in Latin America Programme*; BARRETO *et al.*, 2006).

Kehdy *et al*. (2015) estimaram que a população de Bambuí apresenta 78,5% (SE = 0,4) de ancestralidade Europeia, 14,7% (SE = 0,4) de ancestralidade Africana e 6,7% (SE = 0,1) de ancestralidade Nativo Americana (**Figura 2**). Entre as coortes do EPIGEN, Bambuí se destaca por ter a maior proporção de ancestralidade europeia, além da maior taxa de *inbreeding* (0,010; SE = 0,0008).

Dada a ampla diversidade genética observada no Brasil, é necessário cautela ao extrapolar achados de uma população para outra; validação e replicação em coortes distintas ou em bases de dados públicas são essenciais. Em âmbito global, comunidades latino-americanas permanecem sub-representadas em estudos genômicos: estima-se que indivíduos hispânicos/latinos constituam menos de 1% da diversidade amostral em investigações internacionais (MILLS; RAHAL, 2020). Como enfatizado por Alvim *et al*. (2024) e Wojcik *et al*. (2019), ampliar a inclusão dessas populações é urgente para avançar na compreensão da etiologia de fenótipos complexos e nas frequências alélicas específicas.

Figura 2. Distribuição da ancestralidade continental individual nas coortes do EPIGEN.



Ancestralidades individuais de cada coorte, onde cada barra corresponde a um indivíduo; as cores indicam as contribuições de cada componente continental à ancestralidade individual. Fonte: Kehdy *et al*., 2015.

2. Justificativa

Compreender o desenvolvimento da inflamação e como as citocinas atuam nas vias de sinalização permite caracterizar as respostas a diferentes estímulos e identificar as citocinas envolvidas. Esse conhecimento é essencial para elucidar a dinâmica do processo inflamatório e, por meio de estudos de associação genética, detectar variantes que modulam as respostas imunológicas, especialmente na população miscigenada brasileira, contribuindo para esclarecer mecanismos patogênicos e aperfeiçoar estratégias de prevenção e tratamento.

3. Objetivos

3.1 Objetivo geral

- Estudar a arquitetura genética das concentrações séricas de mediadores inflamatórias em idosos brasileiros, no contexto de variáveis não genéticas e da ancestralidade da população brasileira.

3.2 Objetivos específicos

- Modelar os níveis de concentrações séricas de CCL2, ajustando para covariáveis ambientais e estimar a parcela da variância fenotípica explicada por variáveis não-genéticas, pelo parentesco e pela ancestralidade continental.
- Quantificar as correlações entre CCL2 e os demais mediadores inflamatórios.
- Mapear desequilíbrio de ligação na região associada e caracterizar as variantes presentes.
- Caracterizar a região genômica das variantes significativamente associadas a níveis séricos de CCL2, incluindo consequências funcionais.
- Comparar frequências alélicas e sinais de associação com populações de referência.

4. Materiais e Métodos

4.1 População de estudo: Coorte de Envelhecimento Bambuí

A Coorte de Envelhecimento de Bambuí é uma coorte de base populacional. Teve início em 1997 com os objetivos de "examinar separadamente os efeitos articulares da infecção crônica por Trypanosoma cruzi e doenças não transmissíveis nos resultados de saúde na velhice" além de "avaliar a incidência e os determinantes de eventos em saúde em uma população idosa de baixo nível socioeconômico" (LIMA-COSTA, MARIA FERNANDA; FIRMO; UCHÔA, 2011).

Foram elegíveis indivíduos com 60 anos ou mais em 1997; dos 1.742 residentes aptos, 1.606 foram recrutados (idade média = 69,3 anos; 904 mulheres e 642 homens). Além da linha de base (1997), a coorte teve quatro acompanhamentos até 2008, nos quais foram realizados entrevistas, medidas antropométricas, eletrocardiograma, coleta de sangue e outros procedimentos. Informações mais detalhadas a respeito das metodologias aplicadas na linha de base e acompanhamentos subsequentes, consultar: Lima-Costa; Firmo; Uchôa (2011).

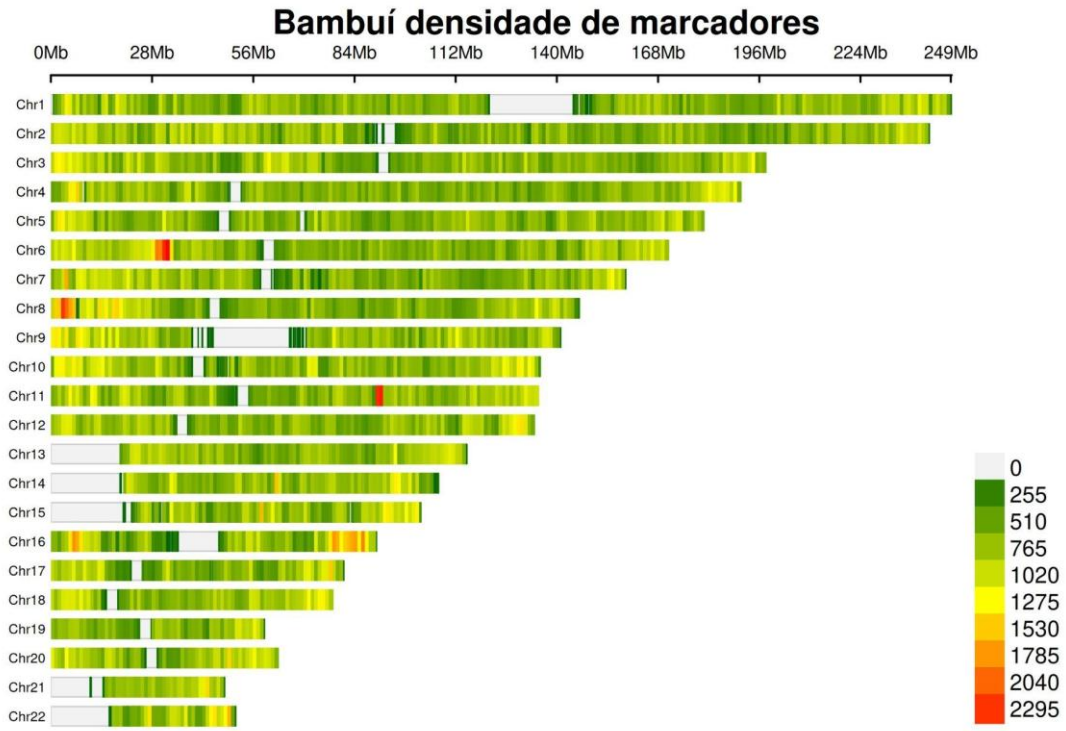
4.2 Controle de Qualidade para dados Genômicos

Dos 1.606, foram genotipados 1.502 indivíduos da coorte de Bambuí utilizando o a*rray Illumina HumanOmni* 2.5-8 v1, com uma densidade de aproximadamente 2,5 milhões de variantes. A anotação das variantes foi realizada no genoma de referência GRCh37/hg19 (*Genome Reference Consortium Human Build 37, Human Genome version 19*). Os procedimentos iniciais de controle de qualidade integrados ao array, bem como etapas adicionais de filtragem, chamada de variantes e processamento dos dados genotípicos da coorte, estão detalhados no material suplementar do estudo de Kehdy *et al.* (2015), na seção *"Quality Control and Data Cleaning for Genotyping Data"*.

Posteriormente, os dados genômicos foram submetidos pelo controle de qualidade realizado com o *software* MosaiQC (https://github.com/ldgh/Smart-cleaning-public), ferramenta que automatiza o pipeline de qualidade e limpeza dos dados do Laboratório de Diversidade Genética Humana. Esse *software* foi utilizado para verificar o sexo das amostras, remover variantes cujas as sondas do *array* não foram mapeadas adequadamente, remover dados incompletos (variantes sem genótipo para mais de 10% dos indivíduos e indivíduos sem genótipo para mais de 10% das variantes), remover SNVs ambíguos (A/T e C/G), remover

variantes 100% heterozigotas, adicionar a identificação de variantes de acordo com o dbSNP, remover dados duplicados e separar os dados por cromossomos autossômicos, cromossomo X, cromossomo Y, região cromossômica pseudoautossômica X e Y e DNA mitocondrial. Após o processamento do MosaiQC, comparou-se o sexo autodeclarado com o sexo inferido geneticamente; não houve discrepâncias e nenhum indivíduo foi excluído. No final de todos os controles de qualidades, obtiveram-se informações genéticas para 1442 indivíduos da coorte, com densidade de 2.186.849 SNVs para a região autossômica, **Figura 3**.

Figura 3. Densidade de SNVs por região cromossômica, obtido com array Illumina HumanOmni 2.5-8 v1.



Densidade de SNVs por janelas de 1 Mb (*megabase pair*).

Os coeficientes de *kinship* entre indivíduos Bambuí (Φij) foram estimados utilizando o método implementado no *software* REAP (*Relatedness Estimation in Admixed Populations;* THORNTON *et al.*, 2012). As estimativas de Φij pelo REAP são condicionadas às proporções de ancestralidades continentais individuais estimadas com o software ADMIXTURE (ALEXANDER; NOVEMBRE; LANGE, 2009), em modo não supervisionado para três populações parentais (europeia, africana e nativa americana; anexo **Tabela A1**). A execução do ADMIXTURE foi realizada com 811.442 SNVs independentes (sem desequilíbrio de ligação, $R^2 < 0,4$).

Para criar um conjunto de dados sem indivíduos aparentados, foi utilizado o NAToRA, um método de exclusão de parentesco para minimizar a perda de conjunto de dados em análises

genéticas e ômicas (LEAL *et al.*, 2022), utilizando o ponto de corte de parentesco de Φij ≥ 0,1, assim como Kehdy *et al*. (2015).

Para reduzir a dimensionalidade dos dados genéticos observados, e obter variáveis indicadoras da estrutura genética populacional foi conduzida a Análise de Componentes Principais (*Principal Components Analysis*, PCA) a partir dos genótipos autossômicos dos indivíduos. Foram calculados os Componentes Principais apenas para os indivíduos de Bambuí, sendo utilizados 792.321 SNVs independentes dos 1.442 indivíduos, com o *software* PLINK v2.0.0.

4.3 Controle de Qualidade para dados Fenotípicos

Os mediadores inflamatórios foram quantificados empregando-se o *Cytometric Bead Array Assay (CBA immunoassay kit; Becton Dickinson Biosciences Pharmingen, San Diego, EUA)* e analisados em citômetro de fluxo FACSVerse (Becton Dickinson, EUA) com o *software BD FCAP Array 3.0*. As amostras de sangue utilizadas para esses exames são provenientes da linha de base de 1997 da Coorte de Envelhecimento de Bambuí, armazenadas a -80ºC, assim como descrito por Cosso *et al.* (2019).

Para identificar possíveis medições atípicas, ou errôneas, dos mediadores inflamatórios foi realizada revisão da literatura, priorizando estudos em populações latinas. Se o valor da observação individual excedesse qualquer outro previamente reportado, a amostra era removida da análise final. Caso a medição estivesse dentro de valores já descritos e considerados biologicamente plausíveis, a amostra era mantida para análise.

A distribuição dos mediadores inflamatórios foi explorada por meio de histogramas e avaliada quanto à sua distribuição normal utilizando o teste de Shapiro-Wilk (SOKAL; ROHLF, 2010). Para caracterizar a forma das distribuições, foram calculados o coeficiente de assimetria ajustado de Fisher-Pearson e o coeficiente de curtose acompanhado de seu p valor (SOKAL; ROHLF, 2010), permitindo a identificação de possíveis desvios de simetria e variações na espessura das caudas. Visto que os mediadores inflamatórios não estavam normalmente distribuídos, com estatísticas do teste de normalidade, assimetria e curtose muito extremos, foi decidido aplicar uma Transformação Normal Inversa Baseada em Ranking (*Rank-based inverse normal transformation*) em todos os fenótipos (MCCAW *et al.*, 2020).

Para investigar as interações entre os mediadores inflamatórios, foi gerada uma matriz de correlação de Spearman (SOKAL; ROHLF, 2010), apresentada na seção Resultados. Estas

análises foram conduzidas em Python 3.10.16, com o uso das bibliotecas SciPy 1.15.2 (VIRTANEN *et al.*, 2020) e Seaborn 0.13.2 (WASKOM, 2021).

4.4 Controle de Qualidade para Covariáveis

Neste estudo, foram incluídas 33 covariáveis, abrangendo aspectos sociodemográficos, de estilo de vida, clínicos e laboratoriais (**Tabela 2**).

Tabela 2. Variáveis de estudo na coorte de idosos de Bambuí.

Categoria	Variável	Tipo	Níveis/Categorização
Variáveis sociodemográficas	Idade	Quantitativa discreta	-
	Sexo	Qualitativa nominal	1 - Masculino, 2 - Feminino
	Cor de pele	Qualitativa nominal	1 - Branca, 2 - Parda, 3 - Negra
	Escolaridade	Qualitativa nominal	0 (Nunca estudou) a 15 (Pós-graduação)
	Renda pessoal	Qualitativa nominal	1 (Menos de 1 S.M.) a 7 (Mais que 20 S.M.)
Variáveis Clínicas	Angina	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Artrite/Reumatismo	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Doença de Chagas	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Diabetes	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Hipertensão	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Infarto	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Osteoporose	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
	Problemas de Coluna	Qualitativa nominal	1 - Sim (já diagnosticado), 2 - Não
Variáveis de estilo de vida	Consumo de álcool	Qualitativa nominal	1 - Sim (≥12 drinques no último ano), 2 - Não
	Cigarros fumados	Qualitativa nominal	1 - Sim (mais de 100 cigarros fumados), 2 - Não
Variáveis Antropométricas	Índice de massa corporal (IMC)	Quantitativa contínua	-
Variáveis Bioquímicas	Albumina	Quantitativa contínua	-
	Cálcio	Quantitativa contínua	-
	Colesterol total	Quantitativa contínua	-
	Colesterol HDL	Quantitativa contínua	-
	Colesterol LDL	Quantitativa contínua	-
	Lipoproteína de densidade muito baixa (VLDL)	Quantitativa contínua	-

	Triglicerídeos	Quantitativa contínua	-
	Creatina	Quantitativa contínua	-
	Glicose	Quantitativa contínua	-
	Ureia	Quantitativa contínua	-
	Magnésio	Quantitativa contínua	-
	Proteína total	Quantitativa contínua	-
Variáveis Hematológicas	Hematócrito (HCT)	Quantitativa contínua	-
	Hemoglobina (HGB)	Quantitativa contínua	-
	Hemoglobina Corpuscular Média (MCH)	Quantitativa contínua	-
	Contagem de hemácias (RBC)	Quantitativa contínua	-
	Contagem de leucócitos (WBC)	Quantitativa contínua	-

As variáveis sociodemográficas, clínicas e de estilo de vida foram obtidas por meio de um questionário aplicado em 1997, assim como descrito em Costa *et al*. (2000). A variável idade foi determinada com base nos dados do censo realizado pela própria equipe de pesquisa em 1997, com o objetivo de identificar os indivíduos elegíveis para o estudo (≥ 60 anos) (LIMA-COSTA, MARIA FERNANDA; FIRMO; UCHÔA, 2011). A presença das condições relacionadas às variáveis clínicas foi considerada apenas quando o participante relatou ter recebido diagnóstico médico da respectiva doença. O índice de massa corporal (IMC) foi calculado de acordo com a fórmula:

$$IMC = \frac{Kg}{m^2}$$

No caso da variável "cor da pele", o questionário original, de 1997 apresentava as seguintes opções de resposta: (1) Branca, (2) Morena, (3) Mulata e (4) Negra. Para o trabalho atual, as categorias "Morena" e "Mulata" foram unificadas sob a designação "Parda", visto que o Instituto Brasileiro de Geografia e Estatística (IBGE) adota essa nomenclatura em seus censos atuais (ANJOS, 2013).

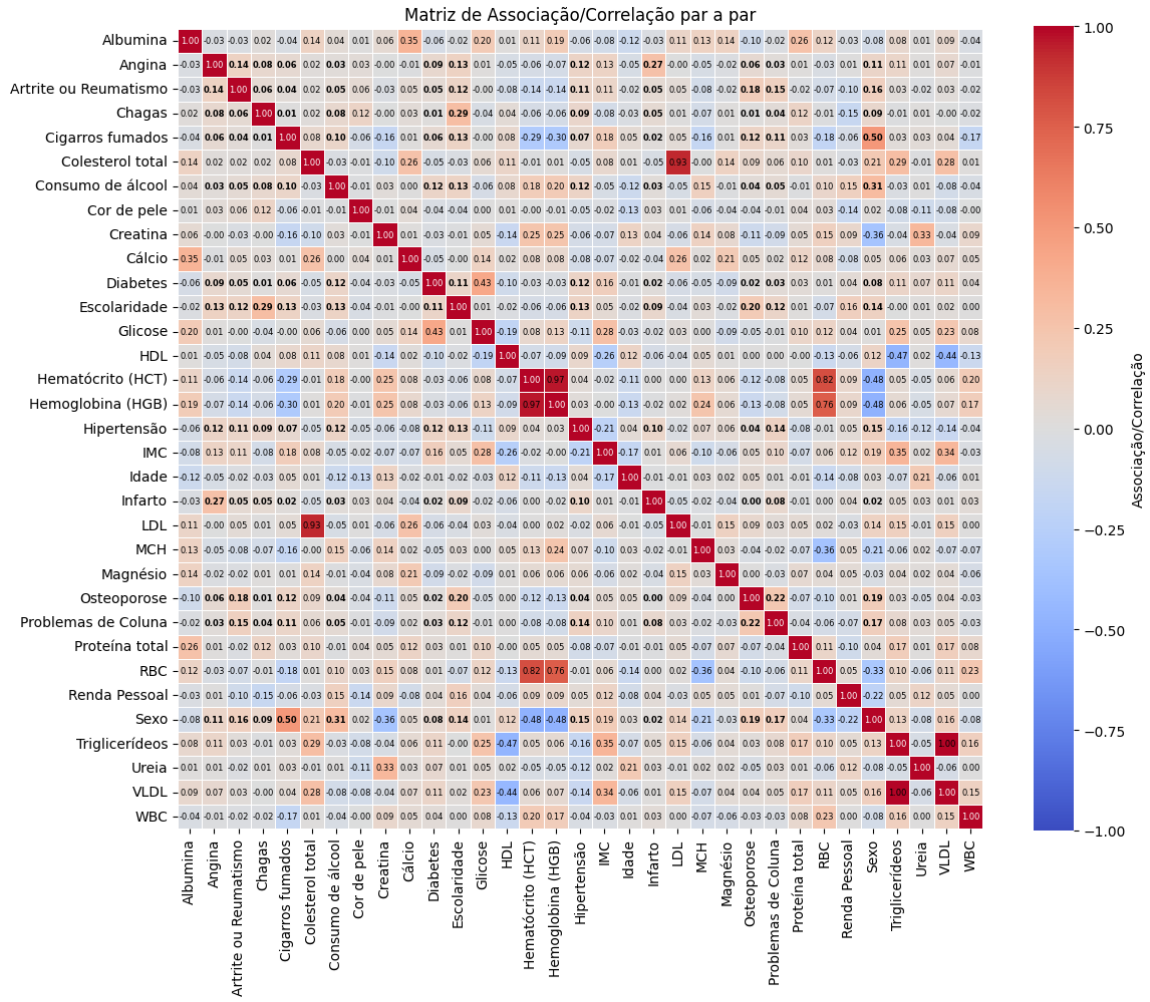
As análises bioquímicas da **Tabela 2** foram realizadas por métodos enzimáticos padronizados em equipamento automatizado, enquanto as análises hematológicas foram conduzidas com contador eletrônico (Coulter Counter T890; Beckman Coulter, EUA). Ambas

as metodologias, assim como o procedimento de coleta sanguínea, estão descritas em Costa *et al*. (2000).

Para avaliar a distribuição das variáveis contínuas, foram aplicados o teste de normalidade de Shapiro-Wilk e a construção de histogramas e *Boxplots*. No caso das variáveis categóricas foi avaliado a distribuição dos indivíduos pelas categorias.

Para investigar as associações entre as covariáveis, foi construído um mapa de calor baseado na análise de correlação entre cada par de variáveis. Para avaliar associações entre variáveis categóricas, foi utilizado o V de Cramér (*Cramér's* V) (SIEGEL, 2021). Nas correlações entre variáveis contínuas e categóricas, utilizou-se o coeficiente de correlação ponto-bisserial (SIEGEL, 2021). Para avaliar as correlações entre variáveis contínuas, utilizou-se o coeficiente de Pearson (SOKAL; ROHLF, 2010) para quantificar relações lineares entre os pares de variáveis. Já o coeficiente de Spearman (SIEGEL, 2021) foi empregado nos casos com observações discrepantes ou quando a relação entre os pares de variáveis se mostrava monótona. Estas análises foram conduzidas em Python 3.10.16, utilizando a biblioteca SciPy 1.13.1 (VIRTANEN *et al.*, 2020).

Figura 4. Mapa de calor para a correlação par a par entre as covariáveis.



Foi utilizado os coeficientes de correlação ponto-bisserial, de Pearson e de Spearman a depender da natureza dos dados. Em negrito as associações realizadas com V de Cramér, variando de 0 (nenhuma associação) a 1 (associação perfeita).

4.5 Análise de Regressões Múltiplas

Com o objetivo de compreender as relações entre covariáveis não genéticas e sua capacidade preditiva sobre os mediadores inflamatórios, foi conduzida uma série de análises de regressão. Inicialmente, foram aplicados modelos lineares mistos de regressão, utilizando o pacote *lme4qtl* (ZIYATDINOV *et al.*, 2018) com a versão 4.4.1 do R. Cada covariável foi testada individualmente para avaliar seu poder preditivo sobre os mediadores inflamatórios normalizados. Como os mediadores inflamatórios foram submetidos a múltiplos testes, aplicou-se a correção de Bonferroni. Considerando um nível de significância inicial de 0,05 e um total de 27 covariáveis disponíveis para testes, o limiar de significância ajustado foi estabelecido em ≈ 0,0018.

Foi aplicado um controle de qualidade ao arquivo de covariáveis, assegurando que todos os indivíduos tivessem informações completas para a análise de regressões múltiplas. Para isso, foram excluídas as variáveis com número excessivo de dados ausentes, sendo elas: consumo de álcool (635), renda pessoal (124), artrite ou reumatismo (95), colesterol LDL (31) e Lipoproteína de Muito Baixa Densidade (*Very-low-density lipoprotein* - VLDL) (31).

Com o objetivo de evitar multicolinearidade, foram analisadas as variáveis altamente correlacionadas a partir da matriz de associação/correlação previamente construída. As variáveis HCT (Hematócrito), HGB (Hemoglobina) e RBC (Contagem de glóbulos vermelhos) estavam fortemente correlacionados (coeficiente de correlação de Spearman > 0,76; $p < 1\times10^{-16}$ para as três comparações). Para evitar a exclusão das variáveis na análise e manter o máximo de informação possível, elas foram substituídas pela Primeira Componente principal construída a partir destas 3 variáveis transformadas com z-scores, chamada PC1 Hemograma, que explica o 90,5% da variância observada. Para gerar as Componentes Principais foi utilizado Python 3.10.16 com a biblioteca Scikit-learn versão 1.5.1 (PEDREGOSA *et al.*, 2012).

Apesar de o colesterol total refletir, em grande parte, as concentrações de LDL, HDL e triglicerídeos, esta variável só demonstrou forte correlação (Correlação Spearman: 0,93, valor p < 1e-16) com o colesterol LDL. Como o LDL foi excluído por excesso de dados faltantes, optou-se, portanto, por manter as frações lipídicas específicas, que fornecem informações mais detalhadas, junto com a medição de colesterol total.

Feito o controle de qualidade das covariáveis, os indivíduos que apresentavam dados faltantes em pelo menos uma covariável foram removidos. Como resultado, obteve-se um conjunto final contendo 1358 indivíduos com dados completos para 27 variáveis, sendo elas: Sexo, Idade, PC1 Genético, IMC, Albumina, Cálcio, Colesterol Total, Creatina, Glicose, PC1 Hemograma, HDL, Magnésio, Proteína Total, Triglicerídeos, Ureia, WBC, MCH, Hipertensão, Angina, Infarto, Osteoporose, Problemas de Coluna, Chagas, Diabetes, Escolaridade, Cor de Pele e Cigarros Fumados.

Com o objetivo de identificar o conjunto mais parcimonioso de covariáveis que melhor explica a variabilidade das medições dos mediadores inflamatórios e que contribua de forma mais eficiente para o estudo de associação, foi adotada a abordagem de seleção de variáveis do tipo *stepwise.* Esse método consiste na comparação de modelos de regressão que incluem diferentes covariáveis como efeitos fixos, acrescentando ao modelo, de forma iterativa, aquelas que mais contribuem para a explicação da variável dependente, neste caso, os mediadores inflamatórios. Todas as 27 covariáveis do conjunto final foram utilizadas na análise *stepwise*.

Para permitir a comparação entre modelos de regressão com diferentes conjuntos de efeitos fixos (modelo mais simples vs. modelo mais complexo) e viabilizar a incorporação de indivíduos aparentados, as regressões foram ajustadas por Máxima Verossimilhança (*Maximum Likelihood*, ML) utilizando o pacote *lme4qtl* (Ziyatdinov et al., 2018) na versão 4.4.1 do R. A seleção entre os modelos foi baseada nos critérios estatísticos de *Bayesian Information Criterion* (BIC) e Teste de razão de verossimilhança (*Likelihood Ratio Test*, LRT). Como controle de qualidade destas regressões para evitar multicolinearidade, foi calculado a correlação dos efeitos fixos.

O modelo mínimo incluiu, obrigatoriamente, as variáveis Idade, Sexo e PC1 como efeitos fixos.

O primeiro passo do procedimento *stepwise* é o *forward step*, que consiste na identificação da covariável mais adequada para ser incluída no modelo. A partir de um conjunto de covariáveis disponíveis, cada uma é adicionada individualmente ao modelo inicial e avaliada. Para cada modelo gerado, são considerados dois critérios: o valor de p do teste de LRT e o BIC. A covariável associada ao modelo que apresentar p valor$< 0,05$ no LRT e o menor valor de BIC entre todos os modelos testados (desde que também seja inferior ao BIC do modelo inicial) é incorporada ao modelo e levada para a etapa seguinte.

Após a inclusão da nova covariável, inicia-se o segundo passo do processo: *backward step*. Nessa etapa, a principal questão a ser avaliada é: com a entrada da nova covariável, alguma das demais, que não sejam permanentes, ainda contribui significativamente para o modelo? Ou sua relevância foi reduzida com a nova composição?

No *backward step*, cada covariável não permanente é removida individualmente do modelo completo. O modelo reduzido (sem a covariável) é então comparado ao modelo completo também com base nos dois critérios anteriores (valor de p do teste de LRT e o BIC). Caso a exclusão da covariável resulte em um $p > 0,10$ no LRT (indicando que sua retirada não compromete significativamente o ajuste do modelo) e em um BIC menor (indicando melhor desempenho), opta-se pelo modelo mais simples, sem essa covariável. Por outro lado, se a covariável estiver associada a um valor de p significativo e a um BIC menor, ela é mantida no modelo por ainda contribuir para a explicação da variabilidade da variável resposta. Todas as covariáveis são avaliadas individualmente e, se alguma for removida nessa etapa, ela permanece em "tempo de espera" (*timeout*) por uma iteração, evitando que a mesma covariável saia e entre no modelo em uma estrutura repetitiva.

Os dois passos - para frente (*forward step*) e para trás (*backward step*) - são repetidos iterativamente até que se encontre o modelo mais parcimonioso, ou seja, o conjunto mais

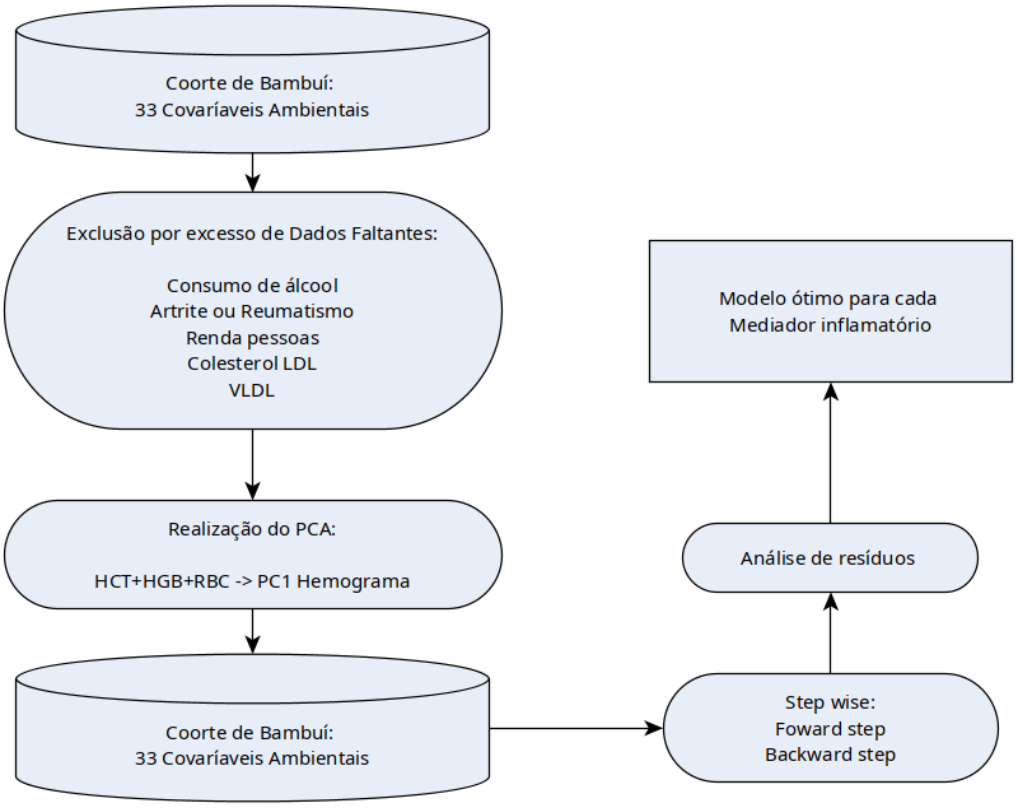
simples e informativo de covariáveis que melhor explica a variabilidade dos dados da variável resposta (mediadores inflamatórios). Como critério de parada, foi definido um limite máximo de 50 iterações, evitando demora excessiva e ciclos repetitivos de inclusão e remoção de variáveis no modelo stepwise. Caso o critério de parada do modelo não fosse satisfeito dentro desse número de passos, o processo era interrompido, e o último modelo gerado, após o passo para trás, era considerado o modelo final. Além disso, se fossem detectados altos níveis de multicolinearidade no modelo final através da matriz de correlação dos efeitos fixos, a covariável com menor contribuição era removida.

Nenhum modelo atingiu o limite máximo de iterações, sendo todos os modelos concluídos antes. Também não foi detectado nenhuma correlação preocupante entre os efeitos fixos nos modelos gerados.

Após a seleção dos modelos alvos para cada mediador inflamatório, os seus parâmetros foram reestimados utilizando Máxima Verossimilhança Restrita (*Restricted Maximum Likelihood*, REML) com o objetivo de obter estimativas e estatísticas de teste mais robustas. Foi calculado o R^2 marginal (R^2 somente dos efeitos fixos) e condicional (R^2 do modelo completo, efeitos fixos + efeitos aleatórios) do modelo, utilizando metodologia proposta por (NAKAGAWA; SCHIELZETH, 2013), disposta no pacote MuMIn (BARTOŃ, 2010) com a versão 4.4.1 do R.

O diagnóstico dos resíduos dos modelos finais foi realizado para verificar pressupostos de linearidade, homocedasticidade, normalidade e independência dos erros. Foram utilizados os gráficos diagnósticos padrão do R: *Quantile-Quantile* (QQ plot), Resíduos vs. Valores Ajustados, Resíduos vs. Alavancagem e Escala-Local (*Scale-Location plot*).

Figura 5. Fluxograma do controle de qualidade e seleção de modelo para os Mediadores Inflamatórios.



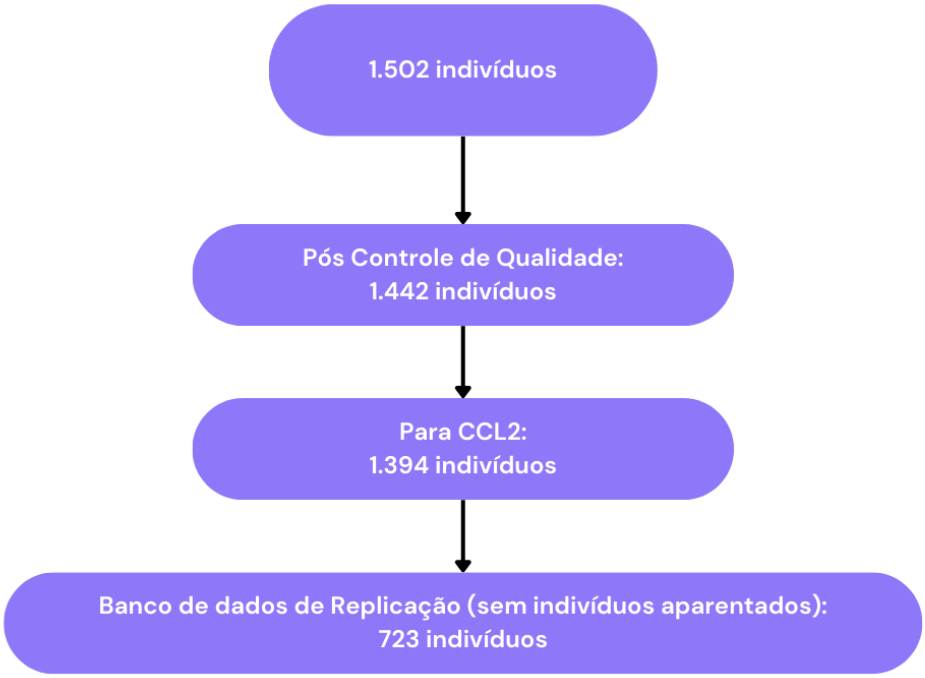
4.6 Estudos de associação de varredura genômica

Realizou-se um controle de qualidade dos dados genômicos específico para os testes de associação genótipo fenótipo. As etapas incluíram de um limiar de $p < 1\times10^{-19}$ para o teste de equilíbrio de Hardy-Weinberg e um limiar de frequência alélica mínima (*Minor Allele Frequency*, MAF) de 0,001, que na prática elimina variantes com dois ou menos alelos alternativos presentes na coorte. Após o controle de qualidade, restaram 1.973.299 variantes em 1394 indivíduos para níveis séricos de CCL2 (**Figura 6**).

O GWAS foi realizado utilizando o software GCTA, adequado para coortes com indivíduos aparentados (YANG *et al.*, 2011).

Como forma de avaliar a robustez do GWAS e verificar a ausência de vieses na análise, as associações genéticas que ultrapassaram o limiar de significância foram testadas em uma amostra independente da Coorte de Bambuí, composta apenas por indivíduos não aparentados. Esse grupo reúne 723 participantes sem parentesco entre si, conforme descrito na seção de Controle de Qualidade para Dados Genômicos. Nesta análise, foi utilizado o *software* PLINK v2.0.0 para as associações. A associação encontrada com as variantes significativas foi replicada utilizando o pacote *lme4qtl* na versão 4.4.1 do R.

Figura 6. Fluxograma do número de indivíduos em cada etapa.



Para se obter a estimativa de β e erro padrão nas unidades originais (pg/ml), foi realizado um novo GWAS do mediador inflamatório sem a transformação, como feito por Meng *et al.* (2024), utilizando o *software* GCTA.

Como etapa de controle de qualidade após o estudo de associação, a frequência alélica das variantes foi comparada com as frequências observadas em populações parentais (europeias, africanas e nativo americanas), utilizando os bancos de dados 1000 *Genomes Phase* 3 (1KGP3; BYRSKA-BISHOP *et al.*, 2022), *Genome Aggregation Database* (gnomAD; CHEN, S. *et al.*, 2024) e *Allele Frequency Aggregator v2* (ALFA; KATTMAN, B. L. *et al.*, 2020) com o objetivo de verificar a coerência das frequências.

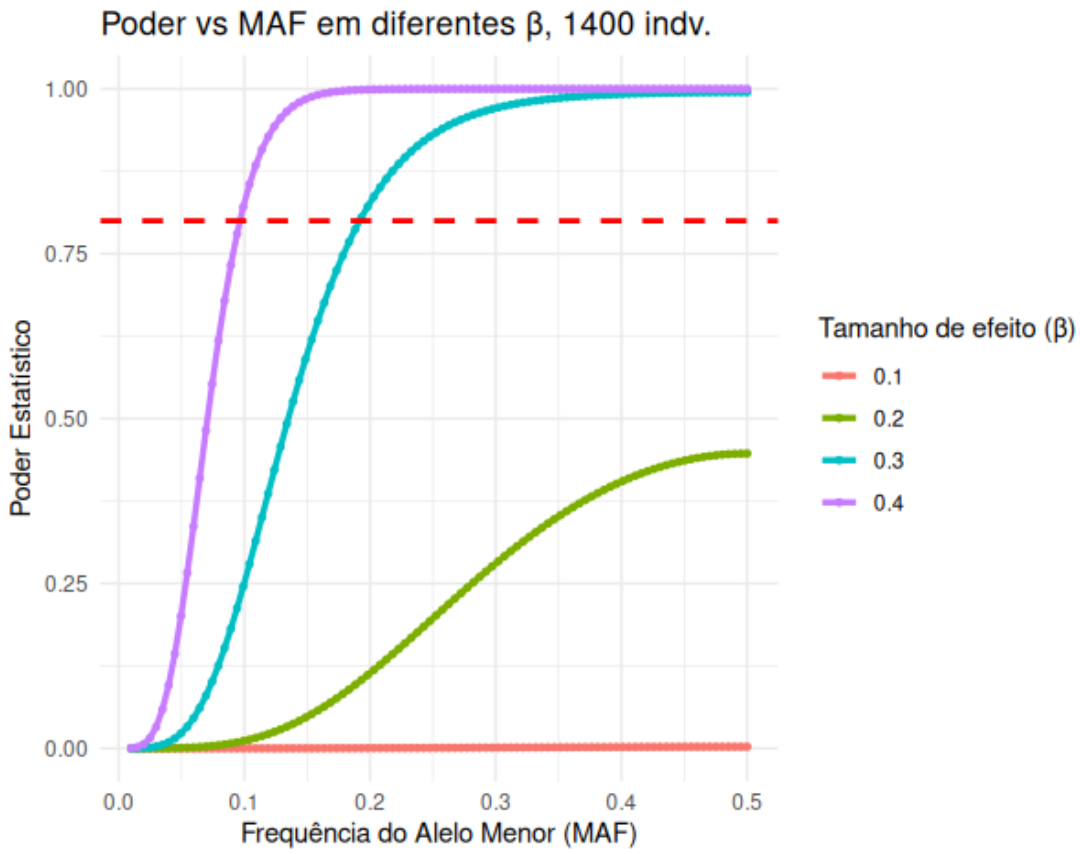
Para todas os GWAS realizados, foram gerados Manhattan e QQ plot, além de calcular o fator de inflação genômica (λ).

A região genômica onde foi identificada associação significativa foi analisada com a plataforma FUMA (WATANABE *et al.*, 2017) em conjunto com o banco de dados do NCBI SNP (SAYERS *et al.*, 2025), com o objetivo de visualizar com mais detalhes o sinal de associação e identificar genes próximos. Para comparar as associações identificadas neste estudo com aquelas previamente descritas, foi utilizado o banco de dados GWAS *Catalog* (BUNIELLO *et al.*, 2019), acessado em 29 de agosto de 2025 (https://www.ebi.ac.uk/gwas/).

Adicionalmente, utilizou-se o software Haploview (BARRETT *et al.*, 2005) para avaliar o desequilíbrio de ligação (LD) entre as variantes na região de interesse da coorte de Bambuí. Com o Haploview foram calculados os coeficientes de determinação R^2 (*pairwise*) e inferidos os haplótipos mais frequentes na população estudada. Indivíduos aparentados foram excluídos das análises para minimizar vieses decorrentes da estrutura familiar.

Para estimar o poder de detecção da análise, poder estatístico, foram realizadas simulações. Partindo de um coorte de 1.400 indivíduos, variando valores de efeito (β) e frequências alélicas, observou-se que, considerando o limiar convencional de ≥80% de poder estatístico em estudos de GWAS, apenas variantes com efeitos moderados a grandes ($\beta \gtrsim 0{,}3$) e frequência alélica ≥10% atingiram poder estatístico > 80% (**Figura 7**). Especificamente, calculamos o poder estatístico das variantes significativamente associadas, utilizando o pacote *genpwr* versão 1.0.4 (MOORE; JACOBSON; FINGERLIN, 2019), na versão 4.4.1 do R.

Figura 7. Análise inicial de poder estatístico para as variantes presentes na Coorte de Envelhecimento de Bambuí.



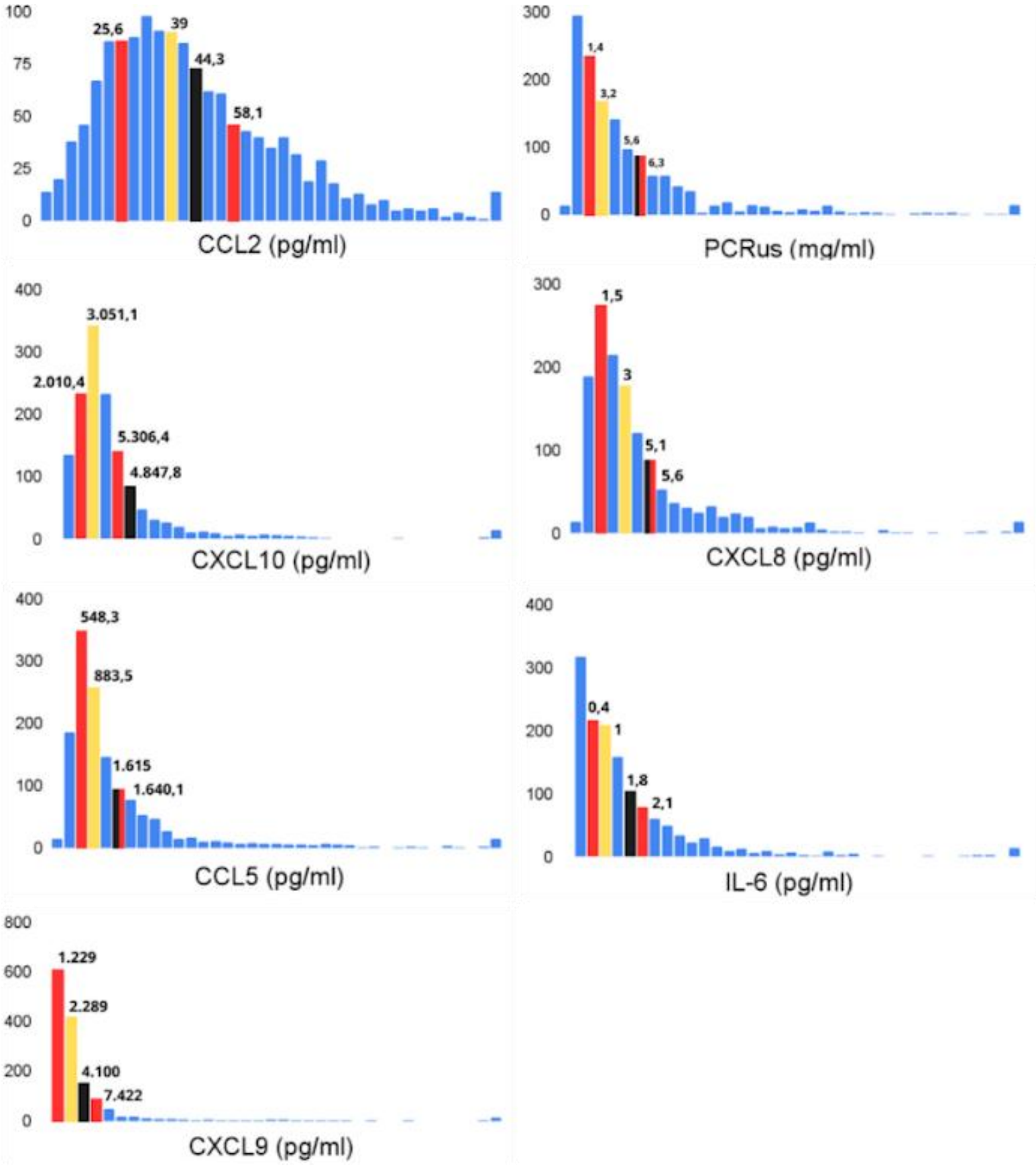
Os tamanhos de efeito (β) estão expressos em unidades de Desvio Padrão (DP).

5. Resultados

5.1 Distribuições dos mediadores inflamatórios

A distribuição dos mediadores inflamatórios está apresentada na **Figura 8**, acompanhada de suas respectivas medidas de tendência central (média e mediana) e medida de dispersão (intervalo interquartil). Todas as distribuições analisadas apresentaram desvios significativos da normalidade (**Tabela 3**, $p < 1\times 10^{-16}$). Como mencionado, na presente dissertação focaremos na quimiocina CCL2, deixando a análise da arquitetura genética dos outros marcadores para outro trabalho.

Figura 8. Gráfico de barras dos mediadores inflamatórios: CCL2 (MCP-1), proteína C reativa ultrassensível (PCRus), CXCL10 (IP-10), CXCL8 (IL-8), CCL5 (RANTES), IL-6 e CXCL9 (MIG).



O eixo horizontal representa os intervalos (*bins*) de concentração, enquanto o eixo vertical indica a frequência absoluta de indivíduos. A distribuição dos dados é destacada por estatísticas descritivas: intervalos interquartis (Q1 e Q3) nos *bins* em vermelho, mediana no *bin* em amarelo e média no *bin* em preto. O último *bin* à direita corresponde aos 1% dos valores mais extremos, permitindo melhor visualização da dispersão dos dados.

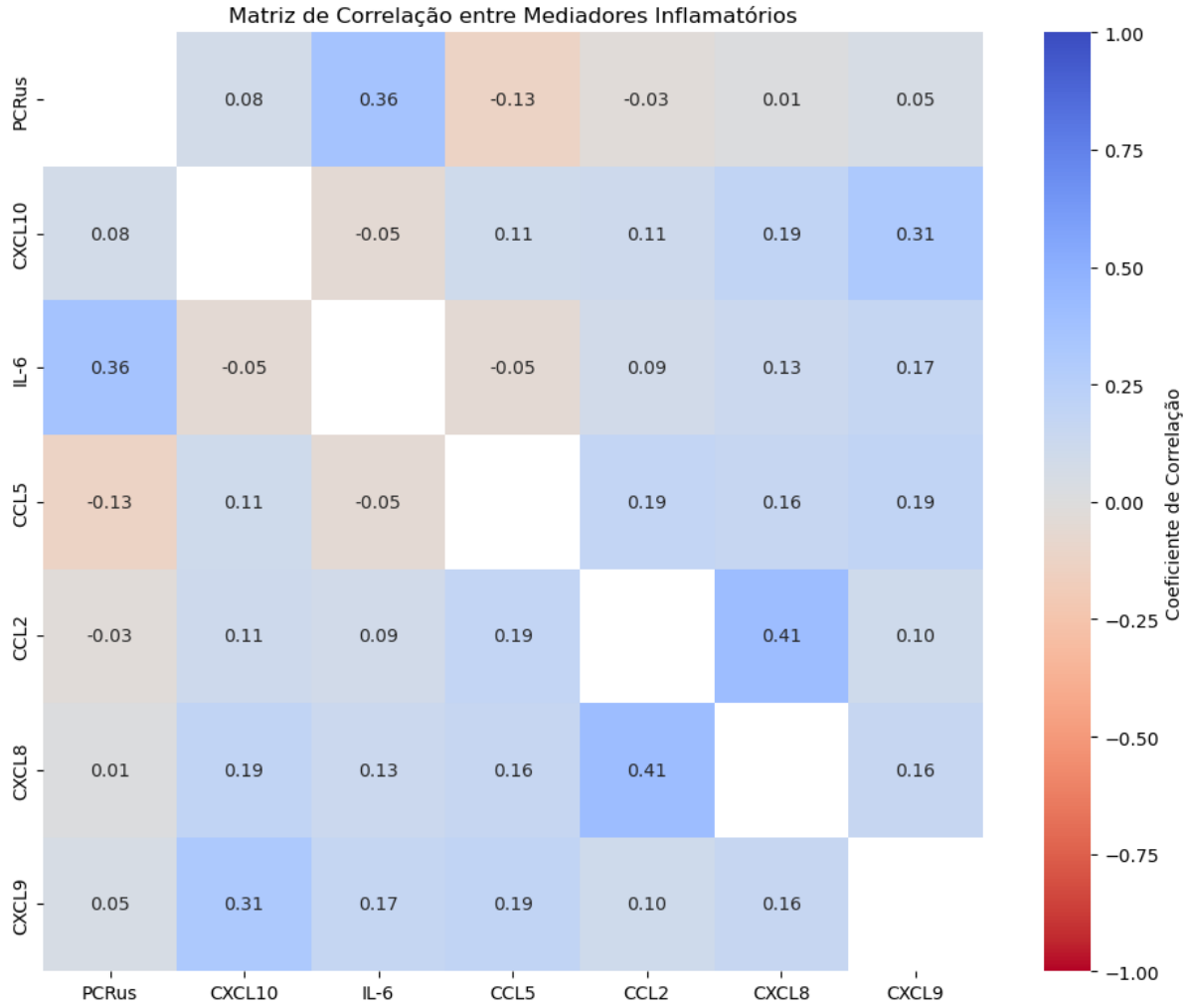
Tabela 3. Estatísticas descritivas: teste de normalidade e estatísticas de distribuição para os mediadores inflamatórios.

Fenótipo	Valor máximo (pg/ml)	Estatística W (Shapiro-Wilk)	p valor (Shapiro-Wilk)	Assimetria (Fisher-Pearson)	Coeficiente de curtose	p-valor (curtose)	Desvio-padrão
PCRus	201,00	0,41	$< 1\times10^{-16}$	9,43	24,00	$< 1\times10^{-16}$	10,10
CXCL8	183,5	0,2	$< 1\times10^{-16}$	25,0	28,0	$< 1\times10^{-16}$	18,0
CCL5	81.752,82	0,0091	$< 1\times10^{-16}$	37,3	28	$< 1\times10^{-16}$	3.520,70
CXCL9	945.788,31	0,0953	$< 1\times10^{-16}$	15,6	26	$< 1\times10^{-16}$	45.743,00
CCL2	284,5	0,9009	$< 1\times10^{-16}$	1,68	14	$< 1\times10^{-16}$	26,9
CXCL10	685.526,56	0,0493	$< 1\times10^{-16}$	29,89	28	$< 1\times10^{-16}$	48.440,90
IL-6	72,11	0,0257	$< 1\times10^{-16}$	36,48	28	$< 1\times10^{-16}$	3,3

Para contextualizar a distribuição de CCL2, os valores observados foram comparados com outros estudos. Na coorte de Bambuí, a mediana de CCL2 foi 39 pg/mL, contrastando com um estudo de 164 indivíduos saudáveis, com idades entre 10 e 79 anos, no qual foram observadas elevadas concentrações, superiores a 559 pg/mL (ANTONELLI *et al.*, 2006). Em coortes de idosos, como o WHICAP (*Washington Heights-Inwood Community Aging Project*, EUA), relatou-se mediana de 590 pg/mL (GUO *et al.*, 2020). Essas discrepâncias podem refletir diferenças populacionais, de história demográfica, de fatores ambientais e sociais.

A respeito da matriz de correlação entre os mediadores (**Figura 9**), observou-se uma correlação moderada entre CCL2 e CXCL8 (r de Spearman = 0,41) e uma correlação mais fraca entre CCL2 e CCL5 (r de Spearman = 0,19).

Figura 9. Mapa de calor para a correlação de Spearman entre os mediadores inflamatórios.



5.2 Modelagem da concentração dos mediadores inflamatórios a partir de variáveis ambientais, parentesco e ancestralidade

Como mencionado, para corrigir a estrutura populacional presente em nossa coorte, utilizamos os Componentes Principais (PCs) proveniente da Análise de Componentes Principais (PCA) e, para as relações familiares, a Matriz de Relacionamento Genético. Como apenas o primeiro PC refletia a estruturação populacional, enquanto os demais capturavam a segregação de núcleos familiares (KEHDY et al., 2015), apenas o PC1 foi incluído como covariável. A **Figura 10** mostra como os PCs refletem a segregação familiar. Incluir os demais PCs seria um sobreajuste, pois a Matriz de Relacionamento Genético já ajusta o modelo para a presença dos núcleos familiares.

Figura 10. Análise de Componentes Principais realizado com o genótipo dos indivíduos da Coorte de Envelhecimento de Bambuí.

Análise de Componentes Principais realizado com o genótipo dos indivíduos da Coorte de Envelhecimento de Bambuí. Cada cor representa um núcleo familiar. Foi utilizado $\Phi ij \geq 0,1$ como limiar de parentesco.

Os melhores modelos para cada mediador inflamatório, identificados pela metodologia *stepwise* descrita em Materiais e Métodos, são:

CCL2 ~ Idade + Sexo + PC1 + (GRM)

CCL5 ~ Idade + Sexo + PC1 + MCH + (GRM)

CXCL8~ Idade + Sexo + PC1 + Cálcio + (GRM)

CXCL9 ~ Idade + Sexo + PC1 + Chagas + (GRM)

CXCL10 ~ Idade + Sexo + PC1 + (GRM)

IL-6 ~ Idade + Sexo + PC1 + Magnésio + Hipertensão + (GRM)

PCRus ~ Idade + Sexo + PC1 + IMC + WBC + Albumina + Proteína total + HDL + Cigarros fumados + (GRM)

Os resultados dos testes individuais das covariáveis que atingiram significância estão apresentados nas **Tabelas 4** do Anexo. Detalhes completos das regressões, incluindo estatísticas de teste, análise da dispersão dos resíduos, bem como efeitos fixos e aleatórios, encontram-se nas **Tabelas A6–A12** do Anexo.

Foi calculado o coeficiente de determinação (R^2) em duas métricas: R^2 marginal e R^2 condicional. O R^2 marginal considera apenas os efeitos fixos e quantifica a proporção da variabilidade fenotípica explicada pelas covariáveis do modelo. No presente estudo, as covariáveis incluem fatores não genéticos (por exemplo, idade e sexo) e componentes de ancestralidade (PC1).

O R^2 condicional, por sua vez, inclui efeitos fixos e aleatórios, representando a variabilidade total explicada pelo modelo. Importante ressaltar que a diferença entre ambos não pode ser interpretada de forma direta como a contribuição exclusiva dos efeitos aleatórios, uma vez que estes interagem com os efeitos fixos, impossibilitando a separação precisa de suas influências. Em nossos modelos, o efeito aleatório é representado pela Matriz de Relacionamento Genético.

Nos testes de regressão de covariáveis individuais, apenas a Albumina superou o limiar de significância (**Tabela A4**) na regressão com a concentração sérica de CCL2. No entanto, ao inserir Albumina no modelo *stepwise,* junto com as demais covariáveis (idade, sexo, PC1), ela

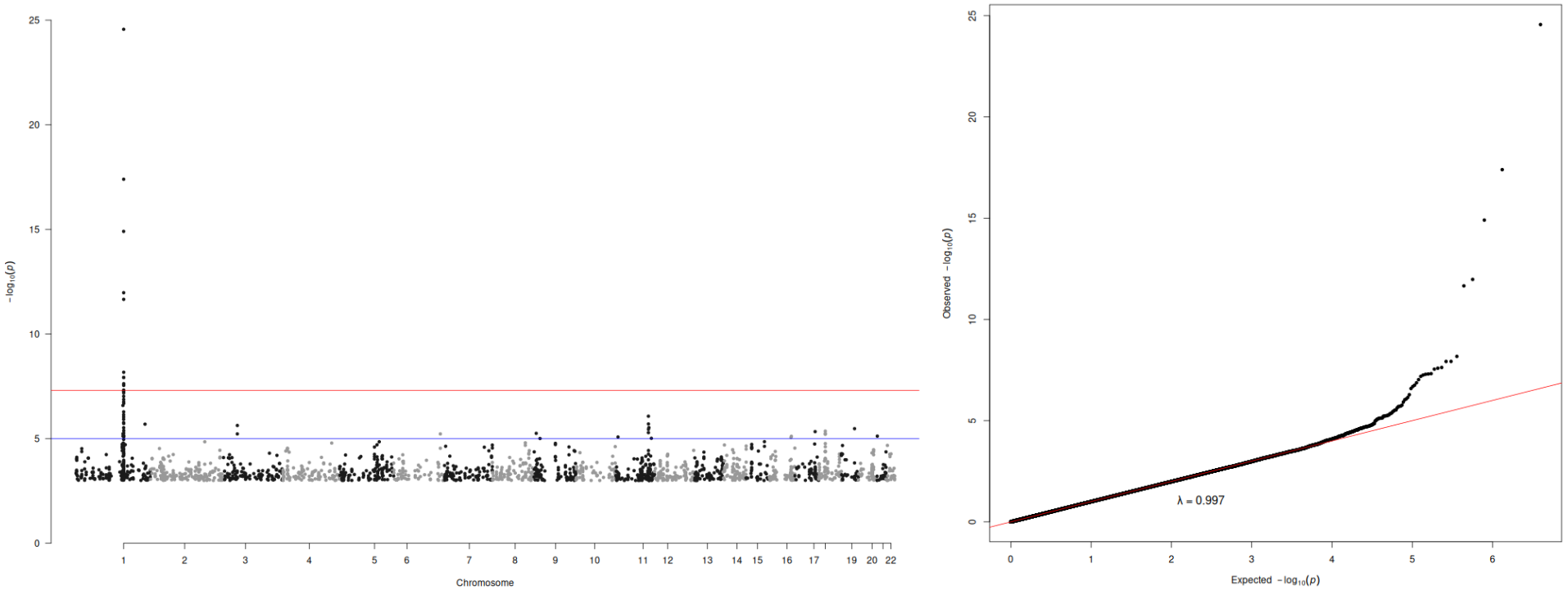
deixou de ser significativa, e não entrou no modelo. Esse resultado sugere que parte da variabilidade fenotípica que Albumina explica passou a ser explicada por outra covariável fixa.

Entre os fenótipos avaliados, CCL2 apresentou o maior R^2 condicional (0,4476), em contraste com um R^2 marginal de apenas 0,0206 (**Tabela A6**). Esse padrão indica que as covariáveis fixas (idade, sexo e componentes de ancestralidade) explicam pouca variação, enquanto uma parcela substancial da variabilidade de CCL2 é capturada pelo componente poligênico representado pela Matriz de Relacionamento Genético. Ainda assim, um R^2 condicional de 0,4476 significa que mais da metade da variabilidade fenotípica permanece não explicada pelo modelo, possivelmente devido a fatores ambientais não incluídos, variantes genéticas não modeladas, interações gene–ambiente, entre outros. Com a posterior inclusão (no GWAS) das variantes genéticas no modelo, se espera que o R^2 marginal aumente, enquanto que R^2 condicional permaneça similar, e que possamos comparar a contribuição de variantes genéticas com as variáveis ambientais.

5.3 Análises de Associação Genômica (GWAS) de Mediadores Inflamatórios

Com base nas covariáveis escolhidas para o modelo de regressão, foi conduzido um GWAS para CCL2 (**Figura 11**). Foram encontradas 13 variantes significativamente associadas ($p < 5\times10^{-8}$) à CCL2. Destas, 8 não estão registradas no GWAS *Catalog* como associadas a qualquer fenótipo, nem mesmo a CCL2 (**Tabela 4**). Todas as 13 variantes com associação significativa ($p < 1x10^{-8}$) concentram-se na região citogenética 1q23.2 (chr1: 159.203.354 - 159.440.094), exibindo frequência alélica superior a 10 % na população de Bambuí e estando em equilíbrio de Hardy-Weinberg ($p > 0,09$).

Figura 11. Manhattan plot da associação de genótipos com medições séricas de CCL2 e respectivo QQ plot, com o valor de inflação genômica (λ = 0,997).



O pico em 1q23.2 corresponde à região 1q23.2 (chr1: 159.203.354 - 159.440.094). No Manhattan plot, a linha azul representa o limiar de associações sugestivas ($p<5x10^{-5}$), e em vermelho o limiar de associações significativas ($p<5x10^{-8}$). No QQ plot, a linha vermelha indica a distribuição esperada dos p-valores sob a hipótese nula.

Tabela 4. Variantes genéticas associadas com a concentração sérica de CCL2 na Coorte de Envelhecimento de Bambuí.

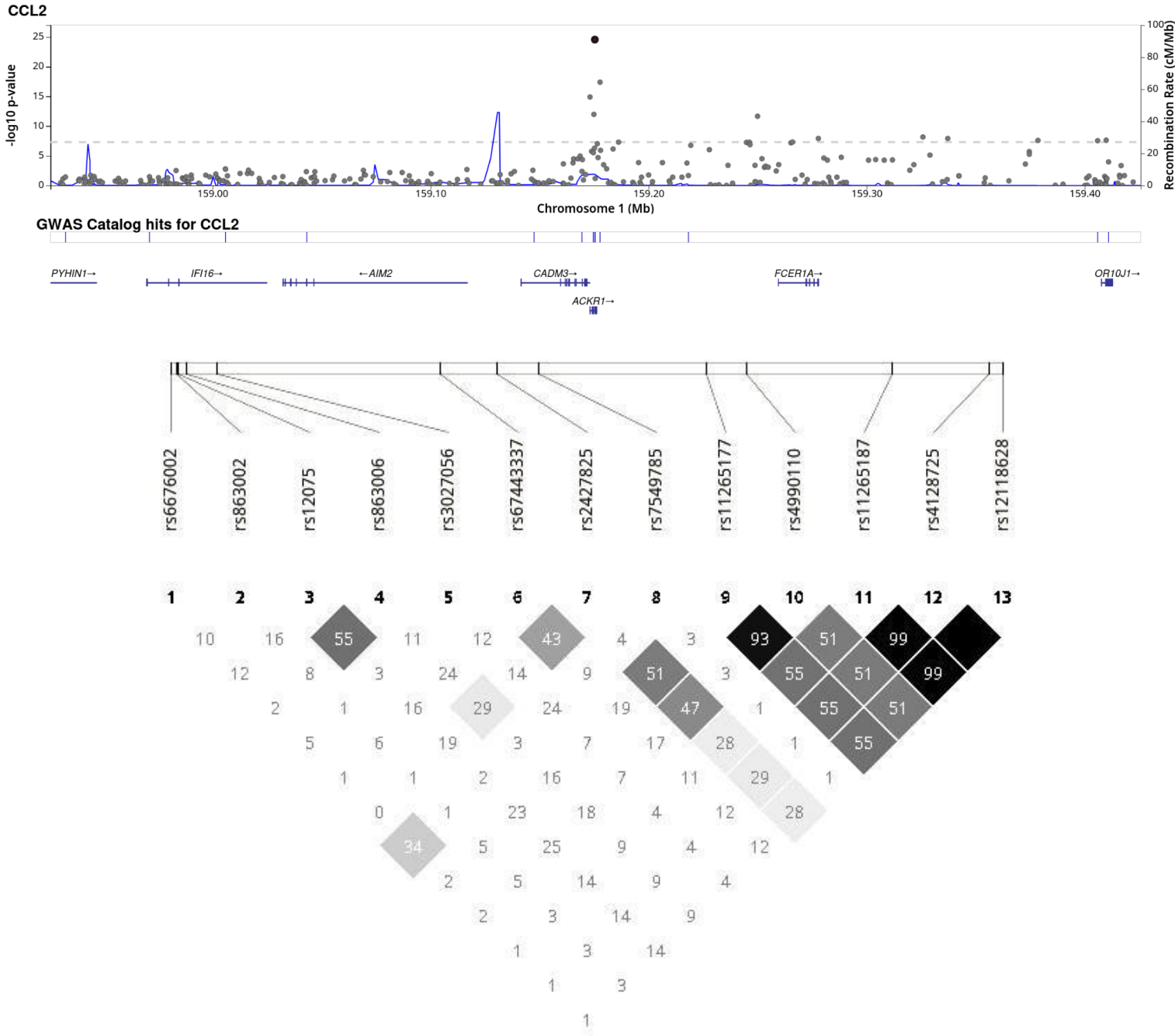
SNV	CHR	BP	A1	A2	FREQ	B (DP)	EP (DP)	P	Poder estatístico	HWE (p valor)	Anotação Funcional
rs6676002	1	159203354	C	T	0,81	-0,40	0,05	1,25E-15	99,75%	0,93	Região intergênica
rs863002	1	159205130	C	T	0,68	-0,29	0,04	1,06E-12	96,22%	0,09	Região intergênica
rs12075	1	159205564	G	A	0,35	-0,43	0,04	2,74E-25	100,00%	0,77	Gly 42 Asp, *ACKR1*
rs863006	1	159207958	G	A	0,42	-0,34	0,04	4,02E-18	99,97%	0,70	Região intergênica
rs3027056	1	159216441	G	A	0,64	-0,22	0,04	4,99E-08	59,28%	0,73	Região intergênica
rs67443337	1	159280188	G	A	0,39	-0,28	0,04	2,21E-12	95,68%	0,25	Região intergênica
rs2427825	1	159296276	T	C	0,24	-0,25	0,05	4,79E-08	54,58%	0,94	Região intergênica
rs7549785	1	159308078	G	A	0,87	-0,33	0,06	1,19E-08	69,16%	0,14	3’ UTR do gene *FCER1A*
rs11265177	1	159356044	G	A	0,17	-0,30	0,05	6,73E-09	66,29%	0,40	Região intrônica de lncRNA (LOC124904433)
rs4990110	1	159367439	G	T	0,16	-0,29	0,05	1,20E-08	62,69%	0,25	Região intergênica
rs11265187	1	159408804	T	C	0,10	-0,36	0,06	2,57E-08	60,00%	1,00	Região intrônica (*OR10J1*)
rs4128725	1	159436169	C	T	0,10	-0,35	0,06	2,86E-08	59,66%	1,00	Região intergênica
rs12118628	1	159440094	A	G	0,10	-0,36	0,06	2,37E-08	60,52%	1,00	Met 101 Ile, *OR10J1*

Em amarelo, SNV líder, com menor p valor associado ao GWAS de CCL2; em roxo, variantes que não haviam sido associadas a nenhum traço no GWAS *Catalog*. As posições genômicas seguem o GRCh38. Valores de frequência alélica na coorte de Bambuí do alelo A1(FREQ), tamanho de efeito (β) do alelo A1, erro padrão (EP), p valor (P), poder estatístico e Equilíbrio de Hardy-Weinberg do alelo A1. β e EP estão em unidades de desvio padrão (DP) do fenótipo normalizado. Anotação funcional de cada variante associada disponível no banco de dados NCBI e ENSEMBL.

As variantes identificadas associadas com CCL2, estão localizadas na região que compreende os genes *CADM3* (*Cell Adhesion Molecule* 3) e *ACKR1* (*Atypical Chemokine Receptor* 1, também conhecido como *Duffy Antigen Receptor for Chemokines*, DARC), **Figura 12**. A SNV líder (aquele com menor p valor da região) é rs12075 (**Tabela 4**). Em estudos de associação genótipo-fenótipo, foca-se em estudar a possível causalidade da SNV líder, mas sempre levando em consideração que ela pode estar em LD com uma variante verdadeiramente causal.

rs12075 está localizada no gene *ACKR1* (**Figura 12**), que codifica um tipo de receptor atípico, para diversas quimiocinas, como CCL2 e CXCL8. ACKR1 atua como receptor sequestrador, sequestrando e degradando quimiocinas, regulando sua biodisponibilidade e prevenindo a hiperativação do sistema imune (Stone, *et al*. 2017). A variante é uma mutação de sítio trocado, onde a mudança de Guanina para Adenina, resulta em uma mudança de aminoácido, de uma Glicina para um Ácido Aspártico no aminoácido 42 (NM_002036.4(*ACKR1*):c.125G > A (p. Gly42Asp), NCBI database).

Figura 12. Associação da variante líder rs12075 com os níveis séricos de CCL2 e o padrão de desequilíbrio de ligação na população de Bambuí.



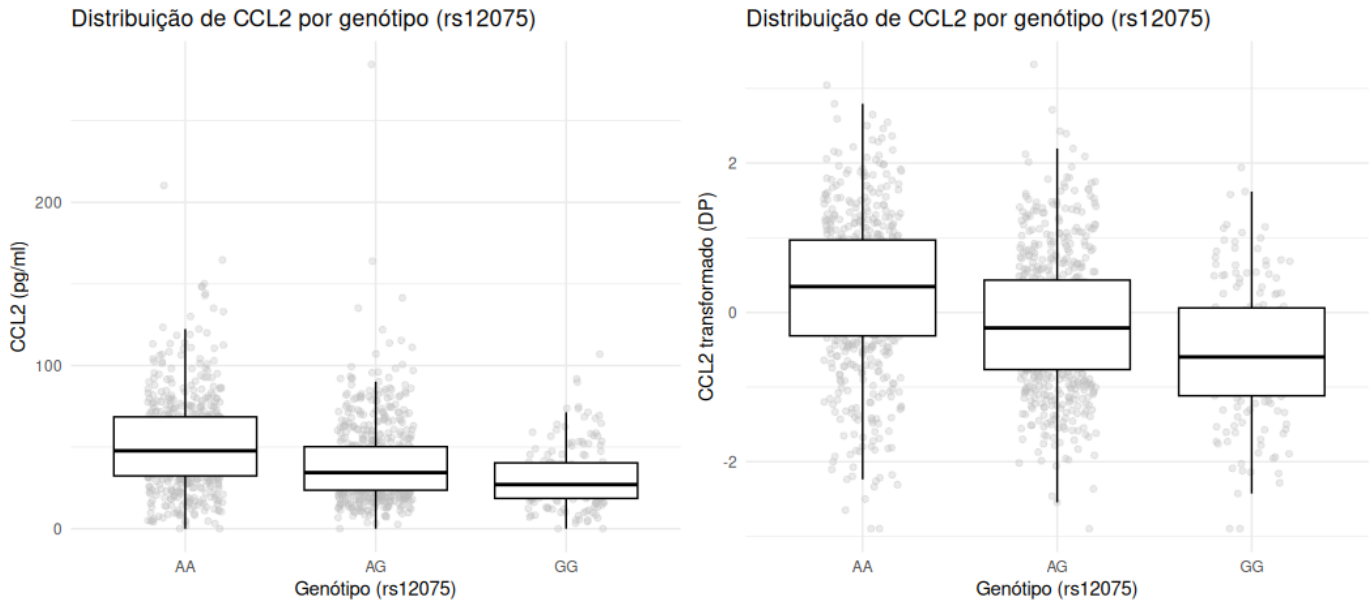
Painel superior, rs12075 associado a medições séricas de CCL2 destacado em preto. Logo abaixo, os genes próximos ao pico de associação. Painel inferior, Desequilíbrio de ligação (R^2) estimado para a população de Bambuí exibindo as 13 variantes associadas à CCL2, em uma janela genômica de 236 kb. Quanto mais escura a célula, maior a correlação entre variantes.

Para elucidar o pico de associação e as variantes que o compõem, avaliou-se o padrão de desequilíbrio de ligação (LD) entre os alelos, utilizando o coeficiente R^2 como medida de associação estatística (**Figura 12**).

Os SNVs rs12075 e rs863002 apresentaram LD moderado ($R^2 = 0,55$). Notavelmente, a SNV líder rs12075 não mostrou LD com o conjunto de variantes rs11265177, rs4990110, rs11265187, rs4128725 e rs12118628, as quais estão em forte LD entre si. Esse padrão sugere que a associação observada para esse grupo de SNVs pode representar um sinal independente daquele liderado por rs12075.

Sobre a distribuição dos genótipos da variante rs12075, foram observados 555 indivíduos com genótipo AA, 614 com AG e 166 com GG (N = 1.335). Foi comparado a distribuição da quimiocina CCL2 entre os genótipos utilizando tanto os valores transformados pela Transformação Normal Inversa Baseada em Ranqueamento, quanto os valores originais em pg/mL. A tendência permaneceu inalterada após a transformação (**Figura 13**). A **Tabela 5** apresenta médias e medianas de CCL2 por genótipo.

Figura 13. Boxplots das distribuições dos valores de CCL2 estratificados de acordo com os genótipos da variante rs12075.



Estão representados em duas escalas: (esquerda) valores transformados com Transformação Normal Inversa Baseada em Ranking, expressos em unidades de desvio padrão (DP), e (direita) valores originais, expressos em pg/ml.

A **Tabela 5** apresenta médias e medianas de CCL2 por genótipo. A transformação do fenótipo (concentrações séricas de CCL2) melhora a adequação dos pressupostos do modelo e a confiabilidade das inferências, em particular dos *p* valores, mas reduz a interpretabilidade do coeficiente β, que passa a estar em unidades padronizadas (desvios-padrão). Por isso, após o

teste com o fenótipo transformado efetuou-se também a regressão nos valores originais, com o objetivo de reportar β e seus erros padrão nas unidades naturais: picogramas por mililitro.

Tabela 5. Média e mediana dos valores de CCL2, transformados e não transformados (pg/ml), de acordo com os genótipos de rs12075.

rs12075	Média (CCL2 transformado)	Mediana (CCL2 transformado)	Média (pg/ml)	Mediana (pg/ml)	n
0 (AA)	0,284	0,346	52	47,8	555
1 (AG)	–0,132	–0,209	40,4	34,4	614
2 (GG)	–0,529	–0,597	31,5	27,1	166

Na análise com os valores de concentração de CCL2 não transformados para normalidade, ajustada por idade, sexo, PC1 e incluindo a Matriz de Relacionamento Genética (GRM) como efeito aleatório (modelo: CCL2 (pg/mL) ~ Idade + Sexo + PC1 + GRM), a estimativa do efeito por cópia do alelo G foi β = −10,95 pg/mL (SE = 1,12), ou seja, cada cópia do alelo G na posição genômica 159.205.564 associa-se a uma redução média de ≈10,95 pg/mL para CCL2. A direção do efeito é consistente com estudos reportados no GWAS *Catalog* para esta variante (**Tabela 6**).

Tabela 6. Associação da variante rs12075 com CCL2 disponível na literatura.

Alelo de Risco	P valor	Freq	β	Referência
rs12075-A	1,00E-21	0,43	-	(VORUGANTI *et al.*, 2012)
rs12075-A	1,00E-21	0,44	aumento de 0,1 pg/mL	(COMUZZIE *et al.*, 2012)
rs12075-A	4,00E-51	0,49	-	(NAITZA *et al.*, 2012)
rs12075-G	1,00E-44	-	diminuição de 0,2185 unidades de desvio padrão (DP)	(AHOLA-OLLI *et al.*, 2017)
rs12075-A	7,00E-149	0,58	aumento de 0,225 unidades	(JIANG *et al.*, 2024)
rs12075-A	2,00E-22	0,56	aumento de 0,0779 unidades	(FOLKERSEN *et al.*, 2020)
rs12075-A	1,00E-134	-	aumento de 0,1166 unidades	(WANG, Y. *et al.*, 2020)
rs12075-G	3,00E-168	0,44	diminuição de 0,492103 unidades	(GUDJONSSON *et al.*, 2022)
rs12075-G	2,00E-24	0,56	diminuição de 0,125 unidades	(SLIZ *et al.*, 2019)
rs12075-A	8,00E-183	-	aumento de 0,6918 unidades	(PNG *et al.*, 2023)

Associação da variante rs12075 com CCL2. Nas colunas: Alelo de risco, p valor, frequência do alelo de risco, β (tamanho do efeito), e referência. Importante destacar o alelo de risco considerado em cada estudo: o alelo A está associado a valores mais elevados da medida, enquanto o alelo G está associado a valores mais baixos.

A respeito do modelo estatístico e da variabilidade fenotípica explicada pelo modelo, temos que a inclusão da variante rs12075 no modelo estatístico aumentou o R^2 marginal de 0,0206 para 0,1056, o que significa que a variante genética rs12075 explica uma parte substancial da variância fenotípica observada.

Ao comparar as frequências alélicas de rs12075 (*ACKR1*) entre as populações europeias e africanas e a população brasileira, observa-se elevada heterogeneidade. Por ser resultado de intensa miscigenação entre componentes ancestrais ameríndia, europeia e africana, a população brasileira tende a apresentar frequências intermediárias (**Tabela 7**).

Tabela 7. Frequência alélica das variantes significativamente associadas a concentrações séricas de CCL2 na coorte de Bambuí, em populações europeias, africanas e americanas

SNP	A1	A2	FREQ Bambuí	EUR	AFR	AMR
rs6676002	T	C	0,19	0,18	0,01	0,23
rs863002	T	C	0,32	0,40	0,01	0,29
rs12075	G	A	0,35	0,40	0,02	0,47
rs863006	G	A	0,42	0,41	0,35	0,50
rs3027056	A	G	0,36	0,36	0,07	0,43
rs67443337	G	A	0,39	0,29	0,97	0,46
rs2427825	T	C	0,24	0,24	0,03	0,26
rs7549785	A	G	0,13	0,16	0,02	0,20
rs11265177	G	A	0,17	0,18	0,02	0,22
rs4990110	G	T	0,16	0,16	0,03	0,20
rs11265187	T	C	0,10	0,10	0,02	0,17
rs4128725	C	T	0,10	0,10	0,02	0,17
rs12118628	A	G	0,10	0,10	0,02	0,17

As frequências alélicas utilizadas são provenientes do *The genome Aggregation Database* v.4 (*gnomAD* v4), 1000 *Genomes Project phase* 3 e *Allele Frequency Aggregator v2.*

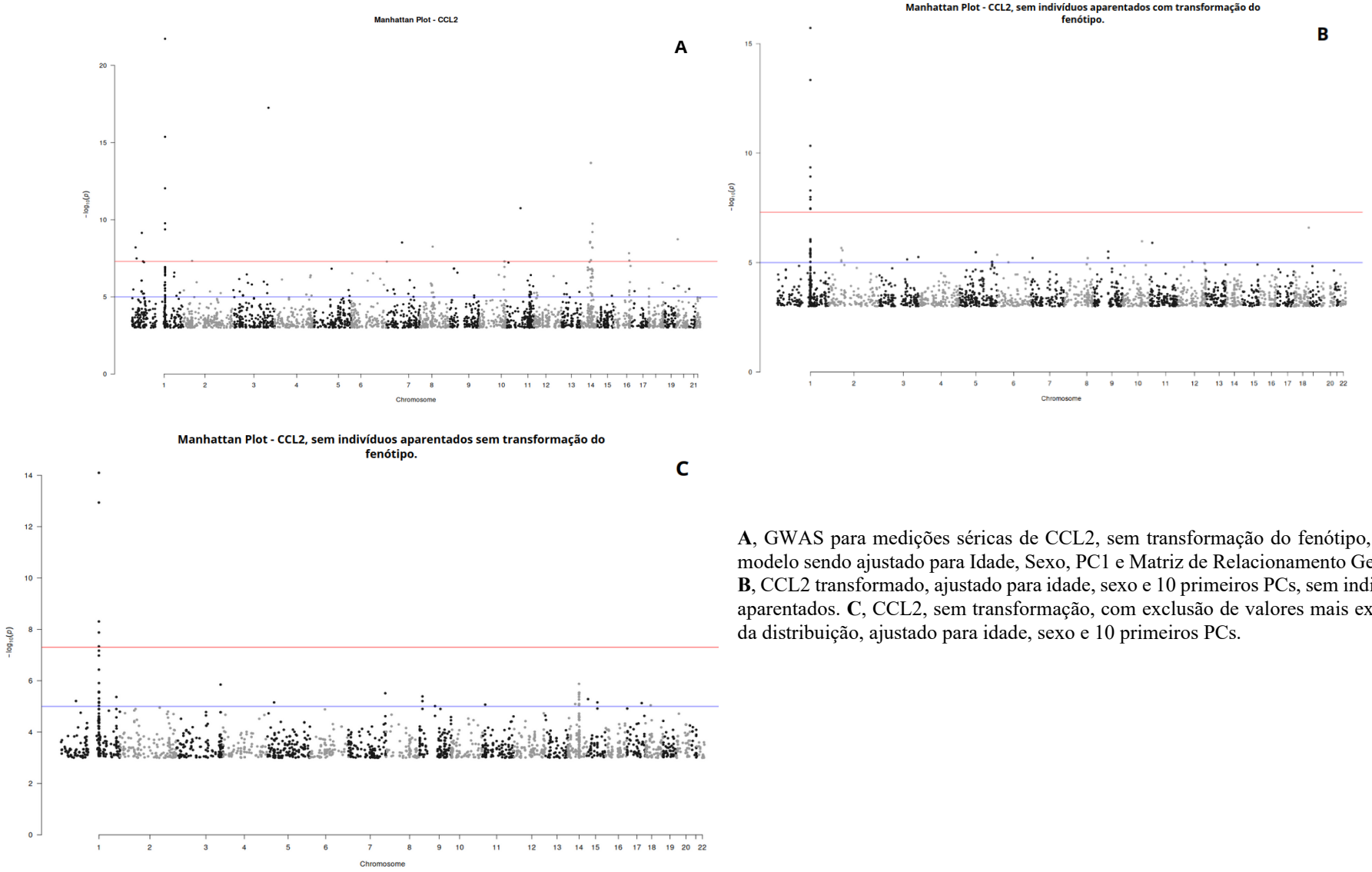
Para avaliar a robustez dos nossos resultados, realizamos GWAS em três cenários distintos:

1. Análise em 1.387 indivíduos da coorte de Bambuí, utilizando o fenótipo bruto (sem normalização), ajustado por idade, sexo e PC1, e incluindo a Matriz de Relacionamento Genético como efeito aleatório. O objetivo foi avaliar se a normalização do fenótipo altera os sinais de associação em relação à distribuição original.
2. Análise em 723 indivíduos não aparentados da coorte de Bambuí, com fenótipo transformado e covariáveis idade, sexo e 10 PCs.
3. Análise em 699 indivíduos não aparentados coorte de Bambuí, sem transformação do fenótipo, e remoção das 24 medições mais extremas, utilizando idade, sexo e 10 PCs como covariáveis.

Essas análises têm como objetivo investigar três aspectos centrais: (i) se a normalização do fenótipo influencia os sinais de associação detectados; (ii) se a inclusão de indivíduos aparentados introduz uma estrutura populacional residual não totalmente corrigida apenas pelo uso da Matriz de Relacionamento Genético combinada com um único componente principal genético (PC1) como covariável; (iii) se a presença de valores mais extremos na distribuição, decorrente da distribuição não normal e da assimetria (*skewness*), afeta os resultados do GWAS.

Os resultados indicam que o sinal de associação foi preservado em todas as análises: apesar da distribuição não normal e assimetria das concentrações de CCL2 (**Figura 14A**), o pico de associação foi replicado mantendo magnitude e direção, e permaneceu presente mesmo após a redução da amostra com a retirada dos indivíduos aparentados (**Figura 14B**). A manutenção do sinal após a aplicação de normalização e remoção de valores mais extremos (**Figura 14C**), sugere que essas características não foram determinantes e que a normalização não produziu associações espúrias. Além disso, a inclusão de indivíduos aparentados não alterou substancialmente os resultados. Em conjunto, esses achados sustentam que o pico de associação reflete um efeito real da variante sobre os níveis de CCL2.

Figura 14. GWAS de concentrações séricas de CCL2 testada em 3 cenários diferentes.



A, GWAS para medições séricas de CCL2, sem transformação do fenótipo, com o modelo sendo ajustado para Idade, Sexo, PC1 e Matriz de Relacionamento Genético. **B**, CCL2 transformado, ajustado para idade, sexo e 10 primeiros PCs, sem indivíduos aparentados. **C**, CCL2, sem transformação, com exclusão de valores mais extremos da distribuição, ajustado para idade, sexo e 10 primeiros PCs.

6. Discussão

A variante rs12075 (p.Gly42Asp) promove a substituição de glicina por ácido aspártico no aminoácido 42 da proteína ACKR1, originando isoformas com propriedades distintas. O alelo rs12075-G (guanina) codifica glicina, enquanto o alelo rs12075-A (adenina) codifica ácido aspártico. Essa substituição implica mudança relevante nas características locais da proteína, uma vez que a glicina é o aminoácido mais simples, apolar e flexível, ao passo que o ácido aspártico é maior, polar e carregado negativamente em pH fisiológico (NELSON; COX; NELSON, 2013).

Estudos de associação sugerem que a isoforma com glicina no aminoácido 42 apresenta maior eficiência na interação com quimiocinas, como CCL2, favorecendo sua ligação. Como consequência, indivíduos portadores do alelo G exibem níveis séricos mais baixos de CCL2. Já a presença de ácido aspártico no aminoácido 42 reduz essa capacidade de sequestro, resultando em concentrações mais elevadas de CCL2 no soro sanguíneo, assim como observado por Schnabel et al. (2010).

Das duas correlações observadas envolvendo CCL2, a correlação entre CCL2 e CXCL8 (IL-8) pode dever-se ao fato de ambas as quimiocinas interagirem com o mesmo receptor atípico, ACKR1. Além disso, a variante funcional de ACKR1 (rs12075) foi encontrada como associada a CXCL8 em um estudo, de 17 ao total relacionados a medições CXCL8. Neste estudo a direção do tamanho de efeito se mostrou igual ao apresentado, com rs12075-G diminuindo 0,12 desvio padrão de CXCL8 (LOYA et al., 2025).

Por sua vez, a correlação entre CCL2 e CCL5 reflete a pertença de ambas ao subgrupo CC das quimiocinas. Portanto, apresentam forte atividade quimiotática para monócitos/macrófagos em processos inflamatórios e agem de forma cooperativa como mediadores pró-inflamatórios, onde compartilham vias de sinalização do sistema imune (GSCHWANDTNER; DERLER; MIDWOOD, 2019).

Com base no GWAS Catalog, a SNV líder rs12075 possuí 10 registros associados com níveis séricos de CCL2. Desses, oito foram conduzidos em populações europeias, e duas provêm do mesmo estudo (*Viva La Familia Study*), realizado em indivíduos hispânicos nos EUA (VORUGANTI et al., 2012). Os achados são concordantes do ponto de vista biológico: em nossa coorte, o pico de associação no gene ACKR1 se mantém robusto em diferentes modelagens (com e sem ajuste para parentesco; com e sem transformação fenotípica), e todas as investigações anteriores que reportaram associação dessa variante com concentrações

séricas indicam a mesma direção de efeito (**Tabela 6**). A arquitetura da região gênica é compatível com a função da proteína, o que reforça a plausibilidade biológica de que rs12075 seja uma variante causal influenciando a homeostase das quimiocinas.

Em termos funcionais, a proteína codificada pelo gene ACKR1, que era chamada anteriormente de DARC (Duffy Antigen Receptor for Chemokines), é também de grande interesse epidemiológico, pois sua expressão na membrana dos eritrócitos define o antígeno do grupo sanguíneo Duffy (Fy), que atua também como o principal receptor de entrada do *Plasmodium vivax* nos eritrócitos.

Dois polimorfismos principais determinam o fenótipo do antígeno Duffy:

- rs2814778 (c.-67T>C): a mutação na região promotora elimina o sítio de ligação de GATA1 nos eritrócitos, silenciando a expressão de ACKR1 na superfície das hemácias, sendo assim denominado Duffy-negativo (Fy-). Esse alelo (conhecido como FYO ou FYBES) é praticamente fixo em muitas populações da África subsaariana, sendo exemplo de seleção natural positiva por resistência à malária (HOWES et al., 2011).
- A nossa SNV líder rs12075 (G125A; p.Gly42Asp): diferencia o fenótipo FyA (guanina - glicina) e FyB (adenina - ácido aspártico) e alteram a interação com a proteína de ligação do parasita (KING et al., 2011). O fenótipo Duffy-negativo confere resistência ao *P. vivax* e, por pressão seletiva, é altamente prevalente em populações asiáticas e é fixado em populações da África ocidental, central e oriental (HOWES et al., 2011).

Essa distribuição de alelos influência a composição genética de populações miscigenadas como a latino-americana (HAMBLIN; THOMPSON; DIRIENZO, 2002). King, et al. (2011) também mostraram que o fenótipo FyA (rs12075-G, o mesmo alelo associado a menor nível de CCL2) apresenta uma resistência a infecção de *P. vivax*, expressando uma ligação inferior de 41–50% quando comparado com FyB (rs12075-A).

Como mostra a **Tabela 7**, a frequência de rs12075-G varia entre populações — sendo provavelmente seleção natural devido à malária. rs12075-G é rara em populações africanas, aproximadamente 0,35 em Bambuí, e encontra-se em frequência mais elevada em europeus e em populações das Américas (0,40 e 0,47 respectivamente). Em 2022, foram notificados no Brasil 131.224 casos de malária, principalmente por causados por *P. vivax* (84,2%), sendo 99,9% desses casos registrados na região Amazônica (SECRETARIA DE VIGILÂNCIA EM SAÚDE E AMBIENTE, 2024). Dada a diferenciação alélica pronunciada entre continentes, a alta prevalência de *P. vivax* no Brasil, diante dos experimentos que mostram a interação do patógeno com a proteína ACKR1, é possível que este locus esteja sob seleção.

Como ilustrado na **Figura 11**, em um intervalo de 236 kb, o desequilíbrio de ligação (LD) entre rs12075 e as variantes associadas com níveis séricos de CCL2: rs11265177, rs4990110, rs11265187, rs4128725 e rs12118628, é fraco. Isso sugere a possibilidade da presença de um segundo sinal de associação.

Essas variantes marginalizadas apresentam forte LD entre si. Dentre elas, apenas a rs4128725 possui associações previamente registradas no GWAS Catalog, sendo uma com os níveis de CCL2. Nesse estudo, conduzido em uma coorte de 1.000 indivíduos europeus, a associação observada para rs4128725 provavelmente refletia um efeito de hitchhiking, mediado por outra variante, rs2494250, localizada em 1:159.308.461, ausente em nosso conjunto de dados (BENJAMIN et al., 2007).

Em nossa análise, identificamos a variante rs11265177 como associada a CCL2, localizada em 1:159.356.044, que apresentou o menor p valor dentro do bloco de LD das variantes marginalizadas. Esse achado levanta a hipótese de que se trata de dois sinais independentes de associação na coorte de Bambuí possivelmente relacionado a uma variante não presente em nosso GWAS.

Das 13 variantes encontradas como associadas à CCL2, oito não possuem registros de associação com CCL2 no GWAS Catalog (**Tabela 4**). Na coorte de Bambuí (≈1.400 indivíduos), identificamos um pico robusto de associação, representado por rs12075. A ausência de sinal em estudos externos pode refletir diferenças na frequência alélica e na estrutura de LD.

7. Conclusão

Diante dos resultados apresentados, e da literatura existente, a variante rs12075 está associada às concentrações séricas de CCL2. Este estudo identificou um sinal robusto de associação entre o gene *ACKR1* e os níveis séricos de CCL2, reforçando o papel funcional de ACKR1 na regulação dessa quimiocina, em populações brasileiras.

É importante enfatizar que associação não implica causalidade. Estudos de GWAS fornecem um mapa para priorização genômica e estudos de validação funcional fornecem o próximo passo para estabelecer a causalidade das variantes.

Os padrões de LD observados na população de Bambuí sugerem a existência de um segundo sinal de associação, atualmente mascarado pelo pico gerado pela SNV líder rs12075. Para clarificar esse segundo pico independente, uma abordagem promissora é utilizar de dados imputados, que consegue ampliar o número de variantes disponíveis para análise, incluindo variantes de baixa frequência, e permite realizar análises condicionais e mapeamento fino com maior poder estatístico. Assim, será possível avaliar com mais precisão a hipótese de um segundo pico de associação nos níveis séricos de CCL2.

Estudos de seleção natural seriam úteis para caracterizar a distribuição e a frequência da variante rs12075 no Brasil. Considerando seu efeito funcional na afinidade com *Plasmodium vivax* e as evidências epidemiológicas que a relacionam a menor risco de infecção, é plausível que esse alelo esteja sendo favorecido por seleção positiva. Além disso, rs12075 pode constituir um biomarcador promissor para os níveis séricos de CCL2.

Como limitação inerente ao estudo, o poder estatístico da amostra não é suficiente para detectar associações com variantes raras. Apesar das limitações, este é o primeiro estudo realizado na população brasileira sobre as concentrações séricas de CCL2, onde foram identificadas 8 novas associações nunca descritas com qualquer fenótipo. Ressalta-se a relevância de investigar populações miscigenadas. Ao analisar a Coorte de Envelhecimento de Bambuí, identificamos variantes novas que provavelmente não seriam detectadas em amostras homogêneas, com pouca diversidade alélica, demonstrando que a diversidade populacional amplia a capacidade de revelar sinais genéticos relevantes.

Referências

AHOLA-OLLI, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *American Journal of Human Genetics*, v. 100, n. 1, p. 40–50, 5 jan. 2017.

ALEXANDER, D. H.; NOVEMBRE, J.; LANGE, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, v. 19, n. 9, p. 1655–1664, set. 2009.

ALVIM, I. *et al.* The need to diversify genomic studies: Insights from Andean highlanders and Amazonians. *Cell*, v. 187, n. 18, p. 4819–4823, set. 2024.

ANJOS. A questão "cor" ou "raça" nos censos nacionais. *A questão "cor" ou "raça" nos censos nacionais.*, jan. 2013. , p. 103–108.

ANTONELLI, A. *et al.* Increase of CXC chemokine CXCL10 and CC chemokine CCL2 serum levels in normal ageing. *Cytokine*, v. 34, n. 1–2, p. 32–38, 21 abr. 2006.

B. L. KATTMAN, L. P. *ALFA: Allele Frequency Aggregator*. . [S.l.]: National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine, 10 mar. 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>.

BARRETO, M. L. *et al.* Risk factors and immunological pathways for asthma and other allergic diseases in children: background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study). *BMC Pulmonary Medicine*, v. 6, n. 1, p. 15, dez. 2006.

BARRETT, J. C. *et al.* Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, v. 21, n. 2, p. 263–265, 15 jan. 2005.

BARTOŃ, K. *MuMIn: Multi-Model Inference*. . [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=MuMIn>. Acesso em: 1 set. 2025. , 28 maio 2010

BENJAMIN, E. J. *et al.* Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC medical genetics*, v. 8 Suppl 1, n. Suppl 1, p. S11, 19 set. 2007.

BUNIELLO, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, v. 47, n. D1, p. D1005–D1012, 8 jan. 2019.

BYRSKA-BISHOP, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, v. 185, n. 18, p. 3426- 3440.e19, set. 2022.

BYUN, J. *et al.* Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. *BMC Genomics*, v. 18, n. 1, p. 789, dez. 2017.

CHEN, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, v. 625, n. 7993, p. 92–100, 4 jan. 2024.

CHEN, Z. *et al.* Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes|Genomes|Genetics*, v. 11, n. 2, p. jkaa056, 12 abr. 2021.

COHEN, J. *Statistical Power Analysis for the Behavioral Sciences*. 0. ed. [S.l.]: Routledge, 2013. Disponível em: <https://www.taylorfrancis.com/books/9781134742707>. Acesso em: 31 ago. 2025.

COMUZZIE, A. G. *et al.* Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS One*, v. 7, n. 12, p. e51954, 2012.
COSSO, R. M. G. *et al.* Associação entre marcadores inflamatórios e ocorrência de hospitalizações: evidências da linha de base da coorte de idosos de Bambuí. *Revista Brasileira de Epidemiologia*, v. 22, p. e190039, 2019.
COSTA, M. F. F. L. E. *et al.* The Bambuí health and ageing study (BHAS): methodological approach and preliminary results of a population-based cohort study of the elderly in Brazil. *Revista de Saúde Pública*, v. 34, n. 2, p. 126–135, abr. 2000.
DUDBRIDGE, F.; GUSNANTO, A. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, v. 32, n. 3, p. 227–234, abr. 2008.
FOLKERSEN, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism*, v. 2, n. 10, p. 1135–1148, out. 2020.
GSCHWANDTNER, M.; DERLER, R.; MIDWOOD, K. S. More Than Just Attractive: How CCL2 Influences Myeloid Cell Behavior Beyond Chemotaxis. *Frontiers in Immunology*, v. 10, p. 2759, 13 dez. 2019.
GUDJONSSON, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nature Communications*, v. 13, n. 1, p. 480, 25 jan. 2022.
GUO, J. *et al.* Reproducibility of serum cytokines in an elderly population. *Immunity & Ageing*, v. 17, n. 1, p. 29, dez. 2020.
HAMBLIN, M. T.; THOMPSON, E. E.; DI RIENZO, A. Complex Signatures of Natural Selection at the Duffy Blood Group Locus. *The American Journal of Human Genetics*, v. 70, n. 2, p. 369–383, fev. 2002.
HEDRICK, P. W. *Genetics of populations*. 3rd ed ed. Boston: Jones and Bartlett Publishers, 2005.
HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, v. 6, n. 2, p. 95–108, fev. 2005.
HOWES, R. E. *et al.* The global distribution of the Duffy blood group. *Nature Communications*, v. 2, n. 1, p. 266, 5 abr. 2011.
INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Censo Demográfico 2022*. . Rio de Janeiro: [s.n.], 2023.
IOANNIDIS, J. P. A.; PATSOPOULOS, N. A.; EVANGELOU, E. Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. *PLoS ONE*, v. 2, n. 9, p. e841, 5 set. 2007.
JIANG, M.-Z. *et al.* Whole genome sequencing based analysis of inflammation biomarkers in the Trans-Omics for Precision Medicine (TOPMed) consortium. *Human Molecular Genetics*, v. 33, n. 16, p. 1429–1441, 6 ago. 2024.
KEHDY, F. S. G. *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, v. 112, n. 28, p. 8696–8701, 14 jul. 2015.
KING, C. L. *et al.* Fy^a /Fy^b antigen polymorphism in human erythrocyte Duffy antigen affects susceptibility to *Plasmodium vivax* malaria. *Proceedings of the National Academy of Sciences*, v. 108, n. 50, p. 20113–20118, 13 dez. 2011.

LEAL, T. P. *et al.* NAToRA, a relatedness-pruning method to minimize the loss of dataset size in genetic and omics analyses. *Computational and Structural Biotechnology Journal*, v. 20, p. 1821–1828, 2022.
LETTRE, G.; LANGE, C.; HIRSCHHORN, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, v. 31, n. 4, p. 358–362, maio 2007.
LEWONTIN, R. C. *The genetic basis of evolutionary change*. New York: Columbia University Press, 1974. (Columbia biological series, no. 25).
LIMA-COSTA, M. F.; FIRMO, J. O.; UCHOA, E. Cohort Profile: The Bambui (Brazil) Cohort Study of Ageing. *International Journal of Epidemiology*, v. 40, n. 4, p. 862–867, 1 ago. 2011.
LIMA-COSTA, MARIA FERNANDA; FIRMO, J. O. A.; UCHÔA, E. The Bambuí Cohort Study of Aging: methodology and health profile of participants at baseline. *Cadernos de Saúde Pública*, v. 27, n. suppl 3, p. s327–s335, 2011.
LOYA, H. *et al.* A scalable variational inference approach for increased mixed-model association power. *Nature Genetics*, v. 57, n. 2, p. 461–468, fev. 2025.
MCCAW, Z. R. *et al.* Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, v. 76, n. 4, p. 1262–1272, dez. 2020.
MENG, X. *et al.* Multi-ancestry genome-wide association study of major depression aids locus discovery, fine mapping, gene prioritization and causal inference. *Nature Genetics*, v. 56, n. 2, p. 222–233, fev. 2024.
MILLS, M. C.; RAHAL, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics*, v. 52, n. 3, p. 242–243, mar. 2020.
MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5. Aufl ed. s.l.: Wiley, 2013. (Wiley Series in Probability and Statistics).
MOORE, C. M.; JACOBSON, S. A.; FINGERLIN, T. E. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Human Heredity*, v. 84, n. 6, p. 256–271, 2019.
NAITZA, S. *et al.* A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS genetics*, v. 8, n. 1, p. e1002480, jan. 2012.
NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, v. 4, n. 2, p. 133–142, fev. 2013.
NELSON, D. L.; COX, M. M.; NELSON, D. L. Lehninger principles of biochemistry. Sixth edition ed. Basingstoke: Macmillan Higher Education, 2013.
PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. 2012. Disponível em: <https://arxiv.org/abs/1201.0490>. Acesso em: 1 set. 2025.
PE'ER, I. *et al.* Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, v. 32, n. 4, p. 381–385, maio 2008.
PNG, G. *et al.* Identifying causal serum protein-cardiometabolic trait relationships using whole genome sequencing. *Human Molecular Genetics*, v. 32, n. 8, p. 1266–1275, 6 abr. 2023.

PRICE, A. L. *et al.* New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, v. 11, n. 7, p. 459–463, jul. 2010.
REICH, D.; PRICE, A. L.; PATTERSON, N. Principal component analysis of genetic data. *Nature Genetics*, v. 40, n. 5, p. 491–492, maio 2008.
RODRIGUES-SOARES, F. *et al.* Genomic Ancestry, *CYP 2D6* , *CYP 2C9* , and *CYP 2C19* Among Latin Americans. *Clinical Pharmacology & Therapeutics*, v. 107, n. 1, p. 257–268, jan. 2020.
SAWYER, S. L. *et al.* Linkage disequilibrium patterns vary substantially among populations. *European Journal of Human Genetics*, v. 13, n. 5, p. 677–686, maio 2005.
SAYERS, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, v. 53, n. D1, p. D20–D29, 6 jan. 2025.
SCHNABEL, R. B. *et al.* Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. *Blood*, v. 115, n. 26, p. 5289–5299, 1 jul. 2010.
SECRETARIA DE VIGILÂNCIA EM SAÚDE E AMBIENTE. *Dia da Malária nas Américas – um panorama da malária no Brasil em 2022 e no primeiro semestre de 2023*. Boletim Epidemiológico. Brasília: Ministério da Saúde, 18 jan. 2024. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2024/boletim-epidemiologico-volume-55-no-01/>. Acesso em: 1 set. 2025.
SIEGEL, S. *Estatística Não-Paramétrica Para Ciências Do Comportamento*. 2. ed. Porto Alegre, RS: Artmed, 2021.
SLIZ, E. *et al.* Genome-wide association study identifies seven novel loci associating with circulating cytokines and cell adhesion molecules in Finns. *Journal of Medical Genetics*, v. 56, n. 9, p. 607–616, set. 2019.
SOKAL, R. R.; ROHLF, F. J. *Biometry: the principles and practice of statistics in biological research*. 3. ed., 11. print ed. New York, NY: Freeman, 2010.
SPROSTON, N. R.; ASHWORTH, J. J. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Frontiers in Immunology*, v. 9, p. 754, 13 abr. 2018.
TAM, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, v. 20, n. 8, p. 467–484, ago. 2019.
TANAKA, T.; NARAZAKI, M.; KISHIMOTO, T. IL-6 in Inflammation, Immunity, and Disease. *Cold Spring Harbor Perspectives in Biology*, v. 6, n. 10, p. a016295–a016295, 1 out. 2014.
THE GIANT CONSORTIUM *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, v. 19, n. 7, p. 807–812, jul. 2011.
THE INTERNATIONAL HAPMAP CONSORTIUM. A haplotype map of the human genome. *Nature*, v. 437, n. 7063, p. 1299–1320, out. 2005.
THORNTON, T. *et al.* Estimating Kinship in Admixed Populations. *The American Journal of Human Genetics*, v. 91, n. 1, p. 122–138, jul. 2012.
UFFELMANN, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers*, v. 1, n. 1, p. 59, 26 ago. 2021.

VAN DEN BERG, S. *et al.* Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *Journal of Animal Breeding and Genetics*, v. 136, n. 6, p. 418–429, nov. 2019.
VICTORA, C. G.; BARROS, F. C. Cohort Profile: The 1982 Pelotas (Brazil) Birth Cohort Study. *International Journal of Epidemiology*, v. 35, n. 2, p. 237–242, 1 abr. 2006.
VIRTANEN, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, v. 17, n. 3, p. 261–272, 2 mar. 2020.
VORUGANTI, V. S. *et al.* Genome-wide association replicates the association of Duffy antigen receptor for chemokines (DARC) polymorphisms with serum monocyte chemoattractant protein-1 (MCP-1) levels in Hispanic children. *Cytokine*, v. 60, n. 3, p. 634–638, dez. 2012.
WANG, M.; XU, S. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*, v. 123, n. 3, p. 287–306, set. 2019.
WANG, Y. *et al.* Genome-wide association study identifies 16 genomic regions associated with circulating cytokines at birth. *PLoS genetics*, v. 16, n. 11, p. e1009163, nov. 2020.
WASKOM, M. seaborn: statistical data visualization. *Journal of Open Source Software*, v. 6, n. 60, p. 3021, 6 abr. 2021.
WATANABE, K. *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, v. 8, n. 1, p. 1826, 28 nov. 2017.
WOJCIK, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, v. 570, n. 7762, p. 514–518, jun. 2019.
XU, Z. Association Testing of a Group of Genetic Markers Based on Next-Generation Sequencing Data and Continuous Response Using a Linear Model Framework. *Mathematics*, v. 11, n. 6, p. 1285, 7 mar. 2023.
YANG, J. *et al.* GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, v. 88, n. 1, p. 76–82, jan. 2011.
ZIYATDINOV, A. *et al.* lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*, v. 19, n. 1, p. 68, dez. 2018.

Anexo

Tabela A1, Análise de Componentes Principais (PCA) realizada com os parâmetros hematológicos: Hematócrito (HCT), Hemoglobina (HGB) e Contagem de Hemácias (RBC).

Variável	PC1 (90,5%)	PC2 (8,7%)	PC3 (0,8%)
HCT	0,5965	-0,283	0,751
RBC	0,5499	0,8257	-0,1256
HGB	0,5846	-0,4879	-0,6482

Tabela A2, Mediadores inflamatórios e suas observações excluídas durante controle de qualidade.

Fenótipo	Medição (pg/ml)
CXCL8	579,24
CCL5	309802,25
CCL5	3664791,25
CCL5	1480460544,00
CXCL10	1642394,75
IL-6	101,46
IL-6	1286,54

Tabela A3, Testes individuais de regressão, utilizando modelo misto linear. Valor p adquirido através do teste de Wald com nível de significância corrigido para ≈ 0,0018.

Mediador Inflamatório	Variável	Valor p
CXCL8	Idade	8,47E-01
CXCL8	Albumina	0,00E+00
CXCL8	Cálcio	3,11E-09
CXCL8	LDL	4,80E+02
IL-6	Idade	3,48E-04
IL-6	Albumina	0,00E+00
IL-6	Magnésio	2,26E+02
IL-6	WBC	3,70E-06
PCRus	Hipertensão	2,79E+01
PCRus	Albumina	1,58E-01
PCRus	HDL	8,75E-03
PCRus	Proteína total	7,54E+01
PCRus	Triglicerídeos	8,06E-01
PCRus	VLDL	7,41E-03
PCRus	WBC	1,13E-08
PCRus	MCH	1,74E+01
CXCL10	Sexo	2,73E+02
CXCL10	Chagas	1,36E+00
CXCL9	Idade	0,00E+00
CXCL9	Chagas	6,48E-01
CXCL9	Proteína total	1,21E+00
CXCL9	Ureia	9,27E+02
CCL2	Albumina	4,72E-03
CCL5	Sexo	1,66E+00
CCL5	Albumina	4,57E-03
CCL5	Creatina	2,98E+01
CCL5	MCH	1,30E+01

Tabela A4, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para CCL2 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: ccl2mcp1_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: ccl2mcp1_transformed ~ age0197 + sexo + PC1 + (1 | IID)
Data: data

REML criterion at convergence: 3793.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.53964	-0.51719	-0.00889	0.50681	3.06924

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.4273	0.6537
Residual		0.5528	0.7435

Number of obs: 1356, groups: IID, 1356

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.472797	0.261185	-1.810
age0197	0.007580	0.003764	2.014
sexo2	-0.083453	0.053918	-1.548
PC1	-4.650980	3.374415	-1.378

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2
age0197	-0.989		
sexo2	-0.098	-0.027	
PC1	-0.011	0.012	-0.007

R^2 marginal (fixed effects): 0.0206
R^2 conditional (fixed + random): 0.4476
ICC (proportion of variance due to individual): 0.436

Tabela A5, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para CCL5 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: ccl5rantes_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: ccl5rantes_transformed ~ age0197 + sexo + PC1 + mch + (1 | IID)
Data: data

REML criterion at convergence: 3811.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.5201	-0.6059	0.0156	0.5895	2.9763

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.1984	0.4454
Residual		0.7741	0.8798

Number of obs: 1353, groups: IID, 1353

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-1.402017	0.515568	-2.719
age0197	0.002223	0.003822	0.582
sexo2	-0.233509	0.056069	-4.165
PC1	-2.750873	2.419244	-1.137
mch	0.044961	0.013954	3.222

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2	PC1
age0197	-0.510			
sexo2	-0.235	-0.027		
PC1	-0.055	0.019	0.002	
mch	-0.857	0.001	0.218	0.053

R^2 marginal (fixed effects): 0.0332
R^2 conditional (fixed + random): 0.2304
ICC (proportion of variance due to individual): 0.204

Tabela A6, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para CXCL8 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: cxcl8il8_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: cxcl8il8_transformed ~ age0197 + sexo + PC1 + calc + (1 | IID)
Data: data

REML criterion at convergence: 3747.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.97570	-0.54811	0.00526	0.56913	3.13582

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.2380	0.4879
Residual		0.6904	0.8309

Number of obs: 1355, groups: IID, 1355

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.659727	0.458057	3.623
age0197	0.018560	0.003723	4.986
sexo2	-0.054452	0.053389	-1.020
PC1	-0.771176	2.597894	-0.297
calc	-0.288789	0.037068	-7.791

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2	PC1
age0197	-0.578			
sexo2	-0.014	-0.029		
PC1	-0.012	0.017	-0.009	
calc	-0.826	0.024	-0.050	0.003

R^2 marginal (fixed effects): 0.0622
R^2 conditional (fixed + random): 0.3026
ICC (proportion of variance due to individual): 0.2564

Tabela A7, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para CXCL9 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: cxcl9mig_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: cxcl9mig_transformed ~ age0197 + sexo + PC1 + Chagas + (1 | IID)
Data: data

REML criterion at convergence: 3719.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.59795	-0.54252	0.03292	0.57733	3.13185

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.2020	0.4495
Residual		0.7044	0.8393

Number of obs: 1356, groups: IID, 1356

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-1.901726	0.260320	-7.305
age0197	0.030152	0.003684	8.184
sexo2	0.093702	0.052937	1.770
PC1	-0.461181	2.423315	-0.190
Chagas2	-0.297663	0.061767	-4.819

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2	PC1
age0197	-0.970			
sexo2	-0.111	-0.028		
PC1	-0.027	0.018	-0.005	
Chagas2	-0.183	-0.006	0.084	0.057

R^2 marginal (fixed effects): 0.0653
R^2 conditional (fixed + random): 0.2737
ICC (proportion of variance due to individual): 0.2229

Tabela A8, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para CXCL10 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: cxcl10ip10_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: cxcl10ip10_transformed ~ age0197 + sexo + PC1 + (1 | IID)
Data: data

REML criterion at convergence: 3752.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.11896	-0.58492	0.00102	0.58225	2.92682

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.2595	0.5094
Residual		0.6769	0.8228

Number of obs: 1355, groups: IID, 1355

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.749382	0.259540	-2.887
age0197	0.009409	0.003737	2.517
sexo2	0.193854	0.053453	3.627
PC1	0.740111	2.697723	0.274

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2
age0197	-0.988		
sexo2	-0.098	-0.026	
PC1	-0.015	0.016	-0.009

R^2 marginal (fixed effects): 0.0147
R^2 conditional (fixed + random): 0.2877
ICC (proportion of variance due to individual): 0.2771

Tabela A9, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para IL-6 submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: il6_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: il6_transformed ~ age0197 + sexo + PC1 + magn + Hipertensao + (1 | IID)
Data: data

REML criterion at convergence: 3764.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.1783	-0.5999	-0.0373	0.5975	3.3539

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.1879	0.4334
Residual		0.7496	0.8658

Number of obs: 1354, groups: IID, 1354

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.915234	0.320766	-2.853
age0197	0.024508	0.003762	6.515
sexo2	-0.057071	0.054446	-1.048
PC1	-0.349494	2.355608	-0.148
magn	-0.322601	0.090070	-3.582
Hipertensao2	-0.149291	0.053921	-2.769

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2	PC1	magn
age0197	-0.796				
sexo2	-0.104	-0.032			
PC1	-0.031	0.017	-0.008		
magn	-0.579	-0.009	0.027	0.028	
Hipertensa2	-0.023	-0.046	0.155	0.013	-0.047

R^2 marginal (fixed effects): 0.0433
R^2 conditional (fixed + random): 0.235
ICC (proportion of variance due to individual): 0.2004

Tabela A10, Resultados do modelo linear misto ajustado pelo método de máxima verossimilhança restrita (REML) para PCRus submetido à Transformação Normal Inversa Baseada em Ranking.

Pheno: pcrus_transformed
Linear mixed model fit by REML ['lmerMod']
Formula: pcrus_transformed ~ age0197 + sexo + PC1 + BMI + wbc + albu + prot + hdl + Cigarros_fumados + (1 | IID)
Data: data

REML criterion at convergence: 3607.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.4181	-0.5551	0.0112	0.5756	3.0515

Random effects:

Groups	Name	Variance	Std.Dev.
IID	(Intercept)	0.2282	0.4777
Residual		0.6305	0.7940

Number of obs: 1332, groups: IID, 1332

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-1.9546087	0.5481525	-3.566
age0197	-0.0004672	0.0037363	-0.125
sexo2	0.2111921	0.0613582	3.442
PC1	-0.6178959	2.5564917	-0.242
BMI	0.0401503	0.0056176	7.147
wbc	0.0782575	0.0121226	6.456
albu	-0.2942539	0.0524317	-5.612
prot	0.2636391	0.0508183	5.188
hdl	-0.0060856	0.0017532	-3.471
Cigarros_fumados2	-0.1941830	0.0612556	-3.170

Correlation of Fixed Effects:

	(Intr)	ag0197	sexo2	PC1	BMI	wbc	albu	prot	hdl
age0197	-0.543								
sexo2	0.035	-0.008							
PC1	0.025	0.022	-0.013						
BMI	-0.421	0.165	-0.152	0.004					
wbc	-0.098	-0.023	-0.008	0.024	-0.011				
albu	-0.337	0.156	0.067	0.015	0.059	0.029			

prot -0.606 -0.022 -0.048 -0.060 0.054 -0.098 -0.268
hdl -0.194 -0.090 -0.135 -0.024 0.252 0.084 -0.030 0.033
Cgrrs_fmds2 0.033 -0.059 -0.477 0.028 -0.104 0.142 -0.002 -0.039 -0.027

R^2 marginal (fixed effects): 0.1399
R^2 conditional (fixed + random): 0.3685
ICC (proportion of variance due to individual): 0.2657

Tabela A11, Estatísticas de Qualidade e Frequência Alélica variante rs12075.

Arquivo	Informação	Valor
Bambui_filter_status.frq	Alelos	G / A
	Frequência do alelo menor (MAF)	0,3548
	Número total de alelos observados	2.762
Bambui_filter_status.hwe	Contagem genotípica (GG/GA/AA)	171 / 638 / 572
	Frequência esperada do alelo (HWE)	0,462
	Frequência observada	0,4578
	p-valor do teste de Hardy-Weinberg	0,769
Bambui_filter_status.lmiss	Número de indivíduos com dados ausentes	61
	Número total de indivíduos	1442
	Proporção de dados ausentes	0,04230