UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA EVOLUTIVA

Cainã Max Couto da Silva

Influência da seleção natural em populações nativas de diferentes ecorregiões americanas

Influence of natural selection on native populations from distinct American ecoregions

São Paulo
2021

Cainã Max Couto da Silva

Influência da seleção natural em populações nativas de diferentes ecorregiões americanas

Influence of natural selection on native populations from distinct American ecoregions

Tese apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Doutor em Ciências, na Área de Genética e Biologia Evolutiva.

Orientadora: Tábita Hünemeier

São Paulo
2021

Ficha catalográfica elaborada pelo Serviço de Biblioteca do Instituto de Biociências da USP,
com os dados fornecidos pelo (a) autor (a) no formulário:
'https://biblioteca.ib.usp.br/ficha-catalografica/src/ficha.php'

Couto-Silva, Cainã Max
Influência da seleção natural em populações nativas de diferentes ecorregiões americanas / Cainã Max Couto-Silva ; orientadora Tábita Hünemeier -- São Paulo, 2021.
124 p.

Tese (Doutorado) -- Instituto de Biociências da Universidade de São Paulo. Programa de Pós-Graduação em Genética e Biologia Evolutiva.

1. seleção natural. 2. nativos americanos. 3. adaptação local. 4. genômica. I. Hünemeier, Tábita, orient. II. Título.

Bibliotecária responsável pela catalogação:
Elisabete da Cruz Neves - CRB - 8/6228

Dedico este trabalho ao meu filho de quatro patas, meu melhor amigo e fiel companheiro,
que encheu meu coração de amor e ternura, Darwin.

Agradecimentos

À minha orientadora, Tábita Hünemeier, pela orientação, pelas discussões científicas, pela compreensão, e por todas as oportunidades incríveis de aprendizagem dentro e fora do país.

Ao Prof. Dr. David Comas, por me receber de braços abertos para estágio em seu laboratório em Barcelona. E ao pós-doutorando André Flores, que esteve comigo no laboratório de Comas e contribuiu grandemente para o meu desenvolvimento acadêmico.

À todos os pesquisadores de campo que coletaram os dados, principalmente o Prof. Dr. Francisco Maurício Salzano (*in memoriam*), que coletou a maioria das amostras analisadas neste trabalho, e teve um papel fundamental na genética de populações no Brasil.

À toda equipe envolvida no Departamento de Genética e Biologia Evolutiva do Instituto de Biociências da Universidade de São Paulo, pela infraestrutura e serviços prestados.

À CAPES e FAPESP (processo 2017/14916-8) pelas bolsas concedidas e auxílio financeiro para execução desta pesquisa.

À todos os meus colegas de laboratório, Renan Lemes, Marcos Silva, Tiago Silva, Maíra Rodrigues e Gabrielle Rizzato, pelas discussões científicas e todos os bons momentos ao longo do percurso do meu doutorado. Destaco a contribuição especial de Renan Lemes tanto para meu desenvolvimento acadêmico-profissional, como para a elaboração desta tese.

À toda equipe de docentes e discentes do "porão", local principal onde foi desenvolvida minha pesquisa. Em especial ao André Fonseca, pelas discussões e bons momentos de confraternização.

Aos meus amigos que sempre me apoiaram e fazem parte da minha família, Rafael Santos, Rafael Pinheiro, Diego Silva, Leonardo Jackson, Caio Mendonça, Júlio Afonso e César Henrique.

À minha irmã de consideração, Magna Magalhães, por estar presente em minha vida desde o mestrado, e ter me auxiliado nos momentos mais difíceis.

À Sandra Caballero Salas, pelas conversas maravilhosas, momentos de reflexão, desconstrução, alegria, e por ter me acolhido tão bem em Barcelona.

À minha mãe, Porancy Couto, que sempre batalhou muito para nos manter em casa, e que me ajuda com muito carinho na criação da minha filha.

À minha filha, por me apoiar, compreender, ensinar, e pelo amor incondicional que sinto como pai.

À Talita, mãe do meu filho de quatro patas, Darwin, que esteve comigo do início ao fim do doutorado, me apoiando, apoiando minha família, e proporcionando muitos momentos de alegria e reflexão.

Ao Darwin (*in memoriam*), meu cachorro, meu filho, e meu melhor amigo, que fez dos meus dias, dias mais felizes, e trouxe harmonia para a casa. Sinto muito a sua falta, e quero acreditar que está bem, feliz e correndo em algum lugar além daqui.

À todas as pessoas não citadas aqui, mas que passaram pela minha vida durante o meu doutorado e contribuíram para meu desenvolvimento pessoal.

Sobretudo agradeço às populações indígenas pela participação nesta e em tantas outras pesquisas, contribuindo para o desenvolvimento científico.

1500, o homem branco em Pindorama chegou
Muita riqueza natural foi o que encontrou
Um clima quente, um belo dia e um povo que vivia em harmonia
Ticuna, Caiagangue, Guarani-kwoa, Juruna, Caetés, Xavantes e Tupinambá
Do Oiapoque ao Chuí, no Brasil, testemunhos do maior crime que se viu
Muito ódio, muita maldade, a coroa mandou pra cá a escória da humanidade
Muito sangue, muita matança, esvaindo com toda esperança
Com sentimento de justiça o índio ficou, se levantando bem mais forte contra o opressor
Agora isso é o que importa na sua vida, usando a lança pra curar sua ferida

Direitos iguais e justiça para o povo Tupi e Guarani
E todas as etnias remanescentes daqui

Direitos Iguais – Ponto de Equilíbrio

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas

aDNA	ancient DNA
ANC-A	Ancestral A
ANC-B	Ancestral B
AP	Antes do presente
DNA	Deoxyribonucleic Acid
EHH	Extended Haplotype Homozygosity
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
FUMA GWAS	Functional Mapping and Annotation of Genome-Wide Association Studies
FUNASA	Fundação Nacional da Saúde
Fst	Fixation index
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Studies
HGDP	Human Genome Diversity Project
IFC	Ice-free corridor
iHS	Integrated Haplotype Score
KEGG	Kyoto Encyclopedia of Genes and Genomes
LKT	Lewontin-Krakauer Test
LSBL	Locus-Specific Branch Length
MAF	Minor Allele Frequency
ORA	Over-Representation Analysis
PBS	Population Branch Statistics

PCA	Principal Component Analysis
Rsb	iHS across populations
SGDP	Simons Genome Diversity Project
SNP	Single Nucleotyde Polymorphism
XP-EHH	Cross-population Extended haplotype homozygosity

Sumário

Capítulo 1

INTRODUÇÃO GERAL

1.1 TEORIA EVOLUTIVA

1.1.1 Evolução por meio da Seleção Natural

A teoria evolutiva por meio da seleção natural foi proposta inicialmente por Charles Darwin e Alfred Russel Wallace em 1858 (DARWIN; WALLACE, 1858) em uma conferência realizada em Londres, um ano antes da publicação da obra seminal "A Origem das Espécies por meio da Seleção Natural" de Charles Darwin (DARWIN, 1859). Em seu próprio livro, Darwin reconheceu que autores anteriores já haviam feito menção à mesma, embora ele tenha o mérito de reunir um vasto conjunto de evidências que suportam especificamente a teoria evolutiva por meio da seleção natural. Assim sendo, Darwin e Wallace propuseram primeiro que as espécies se modificam ao longo do tempo (evolução), onde os descendentes diferem de seus ancestrais – descendência com modificação –, e tais pequenas mudanças, ao longo de um amplo tempo, podem culminar em grandes mudanças. E, segundo, que esta evolução ocorre por meio da seleção natural, em que indivíduos com características que contribuem para a sobrevivência e reprodução, portanto adaptados ao ambiente, tendem a perpetuar, enquanto os demais tendem a ser eliminados (REECE et al., 2013).

A priori, tanto Darwin quanto Wallace não abordaram diretamente a evolução humana, sendo somente em 1871 que Darwin discutiu publicamente sobre a evolução humana em seu livro "A Descendência do Homem" (DARWIN, 1871). Na época, no entanto, Darwin não compreendia, assumidamente, o mecanismo de herança biológica, e tudo indica que não teve conhecimento dos trabalhos de Gregor Mendel, pesquisador contemporâneo de Darwin (embora o contrário não fosse verdade, pois Mendel conhecia o trabalho de Darwin) (FAIRBANKS, 2020). Mendel publicou seu trabalho que hoje compreende os fundamentos das leis de herança em 1866, ainda que tenha sido reconhecido

somente 34 anos depois, após sua morte. A partir da intersecção da evolução e a genética, numerosos avanços e descobertas foram realizadas principalmente ao longo do século XX, sobretudo com contribuições de grandes nomes como Ronald Fisher, Sewall Wright e John Haldane, culminando na base da genética de populações e genética evolutiva, cerne deste trabalho. A seleção natural, no entanto, não é o único mecanismo da evolução, e segundo a teoria neutra (KIMURA, 1969), tampouco o principal. Portanto, compreender e distinguir os mecanismos que promovem a evolução, além da seleção, é uma etapa crucial nos estudos evolutivos.

1.1.2 Mecanismos Evolutivos

A discussão da evolução dentro do contexto de uma mesma espécie é chamada de microevolução, em contraste com o estudo da evolução entre espécies, denominada macroevolução (JOBLING et al., 2014). Nesta tese, discute-se apenas a microevolução da nossa espécie (*Homo sapiens sapiens*) e, portanto, deste ponto em diante, sempre que for mencionado evolução (mecanismos e descrições), lê-se microevolução.

Trazendo para os dias de hoje, do ponto de vista genético, compreende-se por evolução o processo de mudança na frequência alélica ao longo das gerações (VITTI; GROSSMAN; SABETI, 2013). As vias pelas quais este processo evolutivo ocorre são conhecidas como mecanismos evolutivos, que podem ser separados em quatro principais: 1) deriva genética, 2) migração, 3) mutação e, por fim, 4) a seleção natural (HANCOCK; RIENZO, 2008). A ocorrência de qualquer um destes eventos quebra as premissas de um modelo onde não há evolução, denominado Equilíbrio de Hardy-Weinberg (TEMPLETON, 2006).

A mutação consiste na fonte primária de evolução, ou seja, é ela que introduz alterações diretas no DNA, matéria-prima para a evolução. Uma vez que ela introduz variação no DNA, ela tende a contribuir para a diversidade genética da população (LOSOS et al., 2014). A deriva genética pode ser considerada simplesmente como o acaso do processo evolutivo. Em uma população finita, alguns alelos podem mudar de frequência na população de forma aleatória por uma questão puramente amostral (TEMPLETON, 2006). Em contraste com a mutação, que tende a aumentar a diversidade genética, a deriva tende a diminuir a diversidade genética, e é mais forte em populações menores. A migração consiste no movimento de indivíduos ou gametas de uma população para uma área ocupada por outra população, que resulta em um fluxo gênico, ou seja, na troca genética entre as populações (TEMPLETON, 2006). Na genética de populações, ambos os termos migração e fluxo gênico são utilizados de forma equivalente.

Por fim, há a seleção natural, que pode atuar de forma direcional (positiva/negativa) ou balanceadora. A seleção positiva consiste no aumento da frequência de um ou mais alelos, que por sua vez promovem o aumento da sobrevivência, da fertilidade, ou ambos.

A seleção negativa ocorre quando um alelo se torna prejudicial ao organismo, diminuindo a sobrevivência ou fertilidade, e sua frequência diminui e tende a desaparecer rapidamente dentro de uma população. Sugere-se que a seleção negativa seja a mais comum dos tipos de seleção que ocorre nas populações (POUYET et al., 2018; SALAS, 2019). A seleção balanceadora ocorre quando um genótipo heterozigoto é vantajoso, onde o equilíbrio entre diferentes alelos em um gene propicia um fenótipo adaptativo.

Atualmente, dispõe-se de mais métodos estatísticos para detecção da seleção positiva, em contraste com a seleção negativa ou balanceadora. Parece haver também um maior interesse (ou em consequência disso) na seleção positiva, uma vez que alelos que tenham sido selecionados possuem uma conexão forte com regiões funcionais do genoma que também são de interesse da área médica.

1.1.3 Métodos de Detecção da Seleção Natural

Conforme discutido acima, tanto os alelos produzidos por novas mutações, quanto alelos pré-existentes, podem estar sujeitos à seleção natural. Quando isso acontece, ambos deixam marcas no DNA que nos permite inferir quais alelos foram alvo de seleção em um passado recente ou distante.

Dentre as diversas técnicas para inferência de seleção, algumas das abordagens mais comuns incluem os métodos baseados em sítios segregantes (*e.g.* D de Tajima), no desequilíbrio de ligação (*e.g.* iHS, XP-EHH), na distância genética (*e.g.* Fst, PBS), bem como métodos compostos (VITTI; GROSSMAN; SABETI, 2013). Tais métodos são eficazes, quando utilizados em conjunto e também com simulações demográficas ou permutações, para detecção de seleção positiva principalmente monogênica (NIELSEN et al., 2007). Para detecção de adaptação poligênica, no entanto, outras abordagens são utilizadas, como, por exemplo, análises de enriquecimento gênico e de super-representação (WANG et al., 2013; WATANABE et al., 2017; LIAO et al., 2019).

1.1.3.1 Métodos baseados no desequilíbrio de ligação

Os métodos baseados no desequilíbrio de ligação, usualmente denominados como *selective sweeps* (em português, varreduras seletivas), possuem como princípio o efeito *genetic hitchhiking* (ou "efeito carona"), proposto inicialmente em 1974 (SMITH; HAIGH, 1974). De forma resumida, compreende-se por *genetic hitchhiking* quando um alelo sob forte pressão seletiva, bem como seus alelos neutros "vizinhos", aumentam de frequência na população em decorrência da seleção. Como resultado, há uma perda da diversidade genética (heterozigosidade) no haplótipo, gerando uma extensão de homozigose do haplótipo. O primeiro método baseado no desequilíbrio de ligação – *Extended Homozygosity Haplotype* (EHH) *Statistics* – explora essa extensão de homozigosidade para inferir seleção (SABETI et al., 2002). Este tipo de abordagem é extremamente importante na genética evolutiva,

visto que através das variantes neutras ligadas aos alelos sob seleção, torna-se possível detectar estes últimos, mesmo sem saber previamente sua localização genômica (nesta era da genômica, este evento também é referido como varredura adaptativa) (NOVEMBRE; HAN, 2012; STEPHAN, 2019).

A partir da estatística EHH, outros métodos derivados dela foram implementados ao longo dos anos que a sucederam, como Integrated Homozygosity Statistics (iHS) (VOIGHT et al., 2006), cross-population EHH (XP-EHH) (SABETI et al., 2007), e iHS across populations (Rsb) (TANG; THORNTON; STONEKING, 2007).

Cada um dos métodos mencionados acima possui suas vantagens e desvantagens, e todos utilizam diretamente ou indiretamente uma variação da estatística EHH, diferindo na abordagem da respectiva aplicação. O método iHS, por exemplo, identifica regiões de extensão de homozigose, diferindo os alelos ancestrais e derivados, em uma mesma população (intra-populacional), considerado robusto para *sweeps* incompletos (VITTI; GROSSMAN; SABETI, 2013). Ambos os métodos XP-EHH e Rsb são inter-populacionais, e por meio da comparação com a EEH de uma população próxima, tem maior poder para identificar *sweeps* próximos ou já fixados (SUZUKI, 2010). Normalmente, iHS e XP-EHH são utilizados em conjunto nas análises de seleção.

1.1.3.2 Métodos baseados na diferenciação populacional

A métrica mais simples e difundida para mensurar a diferenciação populacional entre duas populações consiste no índice de fixação (Fst). Esta também é uma das métricas mais utilizadas na genética de populações, sendo desenvolvida independentemente por Sewall Wright e Gustave Malécot no final da década de 40 (MALÉCOT, 1948; WRIGHT, 1949). O Fst, como originalmente proposto, mensura a distribuição de variação genética entre subpopulações, comparando a diversidade genética dentro das subpopulações à diversidade genética da população total. Pode-se, contudo, adaptar a fórmula para comparação entre duas populações (Fst par-a-par, ou *pairwise Fst*) para utilizá-la como medida de distância genética (BHATIA et al., 2013; JOBLING et al., 2014).

O princípio dos métodos baseados na distância genética parte do pressuposto de que, quando a seleção está atuando em um locus dentro uma população, mas não dentro de outra população próxima, a frequência alélica entre estas populações deve variar e mostrar-se maior na população sob pressão seletiva (VITTI; GROSSMAN; SABETI, 2013). Partindo deste princípio, surgiu o primeiro teste com o objetivo de detectar sinais de seleção natural utilizando o Fst – o Lewontin-Krakauer test (LKT) –, que foi publicado em 1973 (LEWONTIN; KRAKAUER, 1973). A partir de então, principalmente com o avanço da tecnologia e aumento dos dados disponíveis, diversos outros testes foram e continuam surgindo (REES; CASTELLANO; ANDRÉS, 2020).

Dentre os métodos baseados na distância genética, destaca-se aqui os métodos

Locus-Specific Branch Length metric (LSBL) (SHRIVER et al., 2004) e Population Branch Statistics (PBS) (YI et al., 2010). Neles, calcula-se o Fst par-a-par entre pelo menos três populações: a população-alvo e uma população próxima (ditas populações-irmãs), bem como uma população externa (outgroup), permitindo deste modo isolar o ramo de diferenciação na população-alvo.

1.1.3.3 Adaptação poligênica

Apesar dos testes para detecção de seleção positiva discutidos acima serem muito eficientes para atingir seus objetivos, cabe ressaltar que os mesmos possuem como foco sítios monogênicos, uma vez que identificam loci independentes entre si e com fortes sinais de seleção (PRITCHARD; RIENZO, 2010). Contudo, compreende-se que a maioria dos fenótipos biológicos são poligênicos, ou seja, vários loci atuam em conjunto a fim de contribuir para um dado fenótipo. Nesse contexto, quando considerados de forma individual, apresentam baixa frequência alélica em cada loci (PRITCHARD; PICKRELL; COOP, 2010; VISSCHER et al., 2017). Exemplos de fenótipos poligênicos incluem altura, peso, pigmentação e fertilidade.

Alguns dos métodos para detecção da interação entre genes, que possivelmente atuam em conjunto em um mesmo fenótipo (*i.e.* epistasia), atuam através da verificação de interação estatística entre variantes genéticas intra- ou inter-populacional, quer seja por métodos convencionais de correlação que utilizam regressão linear ou logística, respectivamente (WAN et al., 2010; UEKI; CORDELL, 2012), ou mesmo estatísticas de correlação desenvolvidas especificamente para estudos genéticos (CLIMER et al., 2014; CLIMER; TEMPLETON; ZHANG, 2014). Contudo, adaptação poligênica não pode nem deve ser deduzida apenas por meio da identificação de sinais de epistasia. Diferentemente da detecção de sinais de seleção positiva monogênica, ainda não existem métodos convencionais estabelecidos para detecção de seleção poligênica (PRITCHARD; PICKRELL; COOP, 2010) e, ainda, boa parte da aplicação dos métodos desenvolvidos até a presente data têm sido alvos de críticas (SOHAIL et al., 2019; REFOYO-MARTÍNEZ et al., 2020). Métodos para detecção de adaptação poligênica frequentemente têm utilizado bancos de dados funcionais e fenotípicos, como GWAS Catalog (BUNIELLO et al., 2019), GO (ASHBURNER et al., 2000; The Gene Ontology Consortium, 2019), KEGG (KANEHISA; GOTO, 2000; KANEHISA et al., 2019), entre outros. Outro método baseado em redes (subgrupos), que também adota informações das categorias biológicas presentes em banco de dados públicos, foi desenvolvido a fim de verificar de forma mais sensível sinais de adaptação poligênica (GOUY; DAUB; EXCOFFIER, 2017).

Em contraste com outros métodos que requerem dados fenotípicos ou ambientais associados à amostra (*e.g.* HANCOCK et al., 2010; GÜNTHER; COOP, 2013), ou mesmo demográficos (*e.g.* RACIMO; BERG; PICKRELL, 2018), os métodos que utilizam

categorias biológicas, como a análise de enriquecimento gênico (GSEA, do inglês, *gene set enrichment analysis*) (SUBRAMANIAN et al., 2005) ou como a análise de super representação (ORA, do inglês, *over representation analysis*) (KHATRI; SIROTA; BUTTE, 2012). Por meio dessa abordagem, atribui-se um valor a cada gene (*e.g.* Fst, ou índices de teste de seleção), e testa-se para verificar se um conjunto de genes com os valores mais altos estão super representados em vias ou fenótipos de banco de dados, considerando o pool gênico da amostra (BARGHI; HERMISSON; SCHLÖTTERER, 2020).

Dada a reconhecida escassez de estudos de adaptação poligênica em nativos americanos (MENDES et al., 2020), e que tais métodos têm sido frequentemente aplicados em artigos recentes de adaptação local (BERGEY et al., 2018; HARRISON et al., 2019; HSIEH et al., 2017; LOPEZ et al., 2019; REYNOLDS et al., 2019), optamos por também aplicar métodos de adaptação poligênica no presente estudo.

1.2 ADAPTAÇÃO LOCAL NA AMÉRICA DO SUL

O povoamento da América é um tema amplamente estudado e, desde o início, tem sido alvo de intenso debate na comunidade científica. As diversas áreas da ciência como arqueologia, paleontologia, paleoantropologia, paleoclimatologia, bem como a genética evolutiva, frequentemente se contrapõem. Logo, há um esforço recorrente da comunidade em conciliar os achados dessas diferentes áreas.

Ao passo que a ciência evolui, como, por exemplo, com o advento de novas metodologias, aquisição e disponibilização de dados, principalmente dados de DNA antigo (aDNA, do inglês, *ancient DNA*), nossa compreensão sobre o povoamento torna-se um pouco mais clara, ainda que provavelmente estejamos longe de um consenso geral. Para discorrer sobre este tema, portanto, é importante salientar que este é um campo com mudanças recorrentes, bem como concepções diferentes entre os próprios pesquisadores da área. Sabendo disso, no texto a seguir adotamos como consenso o delineamento realizado por Sutter em um extensivo artigo de revisão (SUTTER, 2020), apresentado aqui de forma a atingir dois objetivos principais: i) introduzir de forma simples e direta os principais pontos para compreensão do povoamento da América, com enfoque na América do Sul, e ii) discorrer sobre suas consequências para os estudos de adaptação local.

A hipótese mais plausível para o povoamento das Américas atribui sua entrada pela província da Beríngia, uma região que conectava a Sibéria ao Alasca e noroeste do Canadá, hoje submersa (Estreito de Bering), mas que ficou exposta durante o último máximo glacial (entre 17 e 24 mil anos atrás), formando uma ponte que não apenas possibilitou a passagem dos seres humanos e outros animais, como provavelmente também serviu de abrigo durante esse período tão hostil (HOFFECKER et al., 2016; SUTTER, 2020).

Este abrigo se explica pelas características climáticas e vegetativas propícias para o refúgio humano, apontadas por estudos paleoambientais (WOOLLER et al., 2018), e também porque duas grandes geleiras na entrada da América bloqueavam a passagem para o continente. Estima-se que os povos ali presentes permaneceram nesta região por aproximadamente 5 a 8 mil anos AP (antes do presente) (FAGUNDES et al., 2008). Neste período, provavelmente acumularam mutações que hoje estão presentes em boa parte dos nativo-americanos, como, por exemplo, novos haplogrupos mitocondriais (TAMM et al., 2007), ou novas variantes com possível maior valor adaptativo, como identificado no gene *FADS* (AMORIM et al., 2017).

Partindo do noroeste asiático e da Sibéria, os primeiros nativos-americanos então se diversificaram de uma população ancestral homogênea, separando-se por volta de 22 a 18 mil anos AP em dois ramos principais: o grupo dos beringianos antigos (AB, do inglês, *Ancient Beringians*) (MORENO-MAYAR et al., 2018a; MORENO-MAYAR et al., 2018b), e o grupo dos nativos-americanos ancestrais, que por sua vez se dividiram em nativos-americanos do Sul, também denominados como ANC-A (Ancestral A), e nos nativos-americanos do Norte, ou ANC-B (Ancestral-B) (RAGHAVAN et al., 2015; MORENO-MAYAR et al., 2018b; POSTH et al., 2018; SCHEIB et al., 2018).

Os primeiros estudos com múltiplas amostras de aDNA da América do Sul com alta cobertura de sequenciamento surgiram em 2018 (POSTH et al., 2018; MORENO-MAYAR et al., 2018b). Apesar de algumas diferenças, ambos concordam que o povoamento da América do Sul ocorreu de forma extremamente rápida, como já anteriormente proposto (LLAMAS et al., 2016). As rotas mais prováveis para migração adentro da América do Sul percorridas pelos nativos-americanos foram pela costa do Pacífico Norte (NPC, do inglês, *North Pacific coast*) e o corredor livre de gelo (IFC, do inglês, *ice-free corridor*), formada entre as duas grandes geleiras (POTTER et al., 2018), sendo que esta última tornou-se viável como rota apenas por volta de 13 mil anos AP (PEDERSEN et al., 2016).

Apesar dos avanços no estudo do povoamento da América, ainda há muito a ser estudado para maior compreensão deste tema, principalmente na América do Sul. Posth et al. (2018) reconheceram na conclusão do próprio artigo, por exemplo, que uma das limitações do trabalho foi a ausência de dados (aDNA) proveniente de populações amazônicas. Assim como as terras altas andinas, a Amazônia faz parte das principais ecorregiões americanas (ANTONELLI et al., 2018), e abrigam ainda hoje povos nativos-americanos, que, em conjunto com os dados de aDNA, contribuem tanto para o estudo do povoamento e demografia dos nativos-americanos, quanto para os estudos de adaptação local a ambientes que poderiam ser considerados como inóspitos.

Desde a saída da África até a entrada na América, as populações passaram por numerosos eventos gargalos de garrafa seguidos de efeitos fundadores, haja visto o evidente decréscimo da diversidade genética ao longo deste percurso (PRUGNOLLE; MANICA;

BALLOUX, 2005). Estima-se que a população ancestral dos nativos-americanos passou por um gargalo de garrafa profundo (FAGUNDES et al., 2008), e o mesmo provavelmente ocorreu em todas as divisões que se sucederam dentro da América. Portanto, torna-se mais do que fundamental analisar de forma cautelosa as análises de seleção natural em populações nativo-americanas, não só para descartar os vieses potenciais da deriva genética, como também o fluxo gênico entre populações de ecorregiões diferentes.

Desta forma, tivemos como objetivo investigar sinais de seleção natural em populações nativas de duas das principais ecorregiões Americanas: Amazônia e Andes, visto que ambas as regiões consistem em "laboratórios naturais" para estudo de adaptação local. Para isso, discorremos sobre as características de cada uma destas regiões, e aplicamos os métodos de detecção de seleção natural aqui mencionados, em conjunto outras abordagens mais específicas (e.g. metanálise, simulação demográfica).

Capítulo 2

ADAPTAÇÃO LOCAL À FLORESTA AMAZÔNICA

2.1 INTRODUÇÃO

2.1.1 Características e Populações Nativas

A Floresta Amazônica é reconhecida pela vasta biodiversidade que apresenta, bem como sua importância econômica e biológica, sendo a maior e mais biodiversa floresta tropical do planeta (LEAL, 2019; FEARNSIDE, 2020). Ela possui mais de 5 milhões de quilômetros quadrados, e abrange oito nações independentes: Brasil, Bolívia, Peru, Equador, Colômbia, Venezuela, Guiana e Suriname, bem como uma colônia (Guiana Francesa). Sua maior parte, entretanto, reside dentro do território brasileiro (68%) (MOREIRA, 2009; LEAL, 2019), e é também nesta parte que se concentra a maioria dos povos indígenas brasileiros (FUNAI, [s.d.]) (Figura 2.1).

As florestas tropicais, apesar de serem representarem um ecossistema rico em biodiversidade, apresentam poucos recursos para subsistência humana. Nestes ambientes, por exemplo, as plantas investem a maioria da sua energia para a manutenção de suas estruturas, gerando poucos frutos, que por sua vez serviriam de alimento para humanos e outros animais (BAILEY et al., 1989).

A alta diversidade de plantas com troncos altos e copas largas, faz com que haja pouca penetração solar (RATNAM et al., 2011), que, aliado com a alta umidade e temperatura característica das florestas tropicais, atua como um fator de complicação para termorregulação (PERRY; DOMINY, 2009). Ademais, a alta sazonalidade das florestas tropicais pode atrapalhar a caça, complicando ainda mais a subsistência humana (HART; HART, 1986; BAILEY et al., 1989).

Além das complicações para obtenção de alimento e termorregulação, constata-

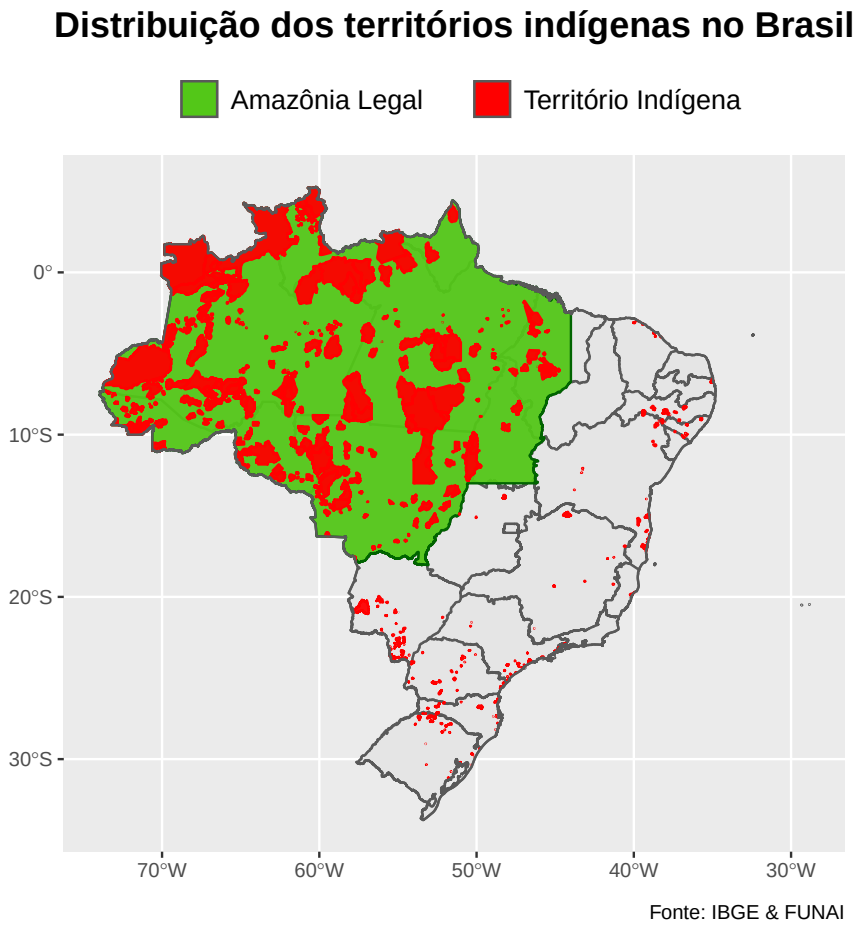

Figura 2.1 – Distribuição dos territórios indígenas no Brasil.

se uma elevada patogenicidade nas florestas tropicais. Estima-se, por exemplo, que as florestas tropicais contenham cerca de 70% de patógenos a mais que as florestas temperadas (GUERNIER; HOCHBERG; GUÉGAN, 2004). Como consequência disso e da falta de sistema de saúde adequado, constata-se uma alta mortalidade em povos indígenas, principalmente em crianças (OHENJO et al., 2006).

Dado às diversas dificuldades para subsistência humana em florestas tropicais, somado à rápida expansão América do Sul adentro, questiona-se se é sequer possível que os habitantes destes ambientes tenham sobrevivido como caçadores-coletores, sem a dependência de recursos externos, como agricultura (BAILEY et al., 1989).

Tradicionalmente, compreende-se que as populações nativas da Amazônia foram constituídas por pequenas populações de caçadores-coletores. Cabe ressaltar, no entanto, que tem havido um crescente número de evidências que suportam a existência grandes sociedades (HECKENBERGER; NEVES, 2009; SOUZA et al., 2018), centros de domesticação de plantas (CLEMENT et al., 2010; CLEMENT et al., 2015), bem como um papel crucial dos rios como rotas primárias de migração (ARIAS et al., 2018). Ainda assim, o hábito caçador-coletor não só esteve presente, como foi essencial na história das populações

nativas amazônicas, e ainda hoje encontram-se populações voluntariamente isoladas que vivem desta forma (OLIVEIRA, 2007).

Não se sabe ainda ao certo qual a data aproximada do povoamento inicial da floresta amazônica, embora cada vez mais tenham surgido novos estudos que contribuem diretamente para esta área. Como mencionado acima, já na América do Sul, tem-se evidências concretas da presença humana há aproximadamente 14.000AP em Monte Verde, Chile (DILLEHAY, 2009). Considerando que nativos-americanos adentraram a América do Sul pelo Istmo do Panamá, podemos considerar uma entrada bem anterior a isso, quer seja pela Amazônia ou pela região costeira do oceano pacífico (POTTER et al., 2018). No referente à Amazônia, um estudo publicado ainda em 2020 mostra evidências da presença humana há aproximadamente 10 mil anos na planície beniana (Llanos de Moxos), localizada nas terras baixas da Bolívia, no sudoeste da Amazônia (LOMBARDO et al., 2020). Em acordo com outros estudos, esse artigo mostra um evento claro de nicho construído por populações nativo-americanas (HÜNEMEIER et al., 2012; WATLING et al., 2018).

2.1.2 Adaptação Local à Floresta Amazônica

Os estudos de seleção na América do Sul são demasiadamente escassos na literatura. Se considerarmos a adaptação local à Floresta Amazônica, por exemplo, em nosso conhecimento, existem apenas dois estudos publicados até a presente data (AMORIM et al., 2015; BORDA et al., 2020).

Amorim et al. (2015) avaliaram as populações Suruí e Karitiana disponíveis no banco de dados do Human Genome Diversity Project (HGDP) (CANN et al., 2002; ROSENBERG et al., 2002) a fim de comparar evolução convergente com outras populações de floresta tropical que vivem na África (*e.g.* pigmeus). Com isso, identificaram sinais de seleção em genes relacionados ao sistema imunológico, metabolismo de lipídio, desenvolvimento corporal e resposta ao estresse. Em estudo recente, Borda et al. (2020), utilizando uma amostra restrita de populações peruanas, realizaram testes de seleção em populações da Amazônia Peruana (Yunga) e das Terras Baixas Amazônicas, identificando como gene-candidato *PTPRC*, responsável pela produção da proteína CD45, que por sua vez atua diretamente no reconhecimento de patógenos, sobretudo virais (BORDA et al., 2020). Assim, conforme esperado em ambientes com altos índices de patógenos (GUERNIER; HOCHBERG; GUÉGAN, 2004), ambos os estudos publicados até o momento convergiram em seus achados, identificando com os maiores sinais seletivos genes-candidatos responsáveis pela resposta imunológica.

A maioria dos estudos de adaptação local na América tem focado nos Andes (vide Capítulo 3). Demais estudos têm focado na Mesoamérica (HÜNEMEIER et al., 2012), na Groenlândia (FUMAGALLI et al., 2015), ou mesmo na América como um todo, no contexto

do período de parada na Beríngia e povoamento das Américas (ACUÑA-ALONZO et al., 2010; AMORIM et al., 2017). Contrariamente à escassez de pesquisa sobre adaptação local à Floresta Tropical Amazônica, existem numerosas pesquisas avaliando adaptação local a outras florestas tropicais, sobretudo na África. Estes estudos tornam-se valiosos para esta tese, uma vez que servem como base de comparação.

Este trabalho tem como objetivo investigar a influência da seleção natural na populações nativas americanas da floresta amazônica, por meio (1) da identificação de SNPs e genes candidatos à seleção positiva; (2) da análise de vias e de redes de genes candidatos à seleção positiva; e (3) da caracterização das possíveis pressões seletivas exercidas sobre os genes e vias identificados.

2.2 MATERIAL E MÉTODOS

2.2.1 Populações

Para o estudo de adaptação local à Floresta Tropical Amazônica, utilizamos dois conjuntos de dados principais, provenientes de diferentes arranjos de genotipagem (Axiom Human Origins e Illumina), aqui nomeados como dataset S1 e dataset S2, respectivamente. O primeiro e principal conjunto de dados contém amostras de 118 nativos Amazônicos, resultados da junção de 37 indivíduos genotipados para este projeto (27 Xavantes, 7 Xikrin, 2 Munduruku e 1 Asurini), 48 indivíduos do estudo de (SKOGLUND et al., 2015), 12 indivíduos de (SILVA et al., 2020), e 21 indivíduos do dataset 11 provenientes do projeto HGDP (<http://www.cephb.fr/hgdp/>). Aproveitamos também este último para aquisição de indivíduos nativos da Mesoamérica (n = 35) e do leste da asiático (n = 231) para fins de análises de seleção, bem como populações de outros continentes para análise de estruturação populacional. Todas as populações nativas americanas, bem como os grupos utilizados para as análises de seleção podem ser verificadas nas tabelas 2.1 e 2.2, respectivamente.

O segundo conjunto de dados contém 104 indivíduos nativos amazônicos e 81 nativos andinos, provenientes do estudo de Gnecchi-Ruscone et al. (2019), somados aos dados de sequenciamento completo do Simons Genome Diversity Project – SGDP (MALLICK et al., 2016) e de Bergström et al. (2020), incluindo 63 indivíduos nativos da Mesoamérica, e aproximadamente mil indivíduos dos demais continentes. Os indivíduos nativos amazônicos e os grupos utilizados nas análises de seleção são apresentados nas Tabelas 2.3 e 2.4, respectivamente.

2.2.2 Filtros de Qualidade e Faseamento do Genoma

Para ambos os conjunto de dados, aplicamos os mesmos filtros de qualidade utilizando o software PLINK v1.9 (CHANG et al., 2015), que consistiram na remoção de indivíduos com mais de 10% de dados faltantes (flag –mind 0.1) e remoção de SNPs com MAF menor que 5% ou com mais de 1% de dados faltantes nos indivíduos (flags –maf 0.05 e –geno 0.01, respectivamente), quando considerado o conjunto de dados como um todo. Adicionalmente, para os dados de sequenciamento utilizados no dataset S2, removemos todas as variantes cujos alelos eram A e T ou C e G, bem como aquelas com *Phred score* menor que 40. Como resultado, os conjuntos de dados finais apresentaram 523.319 e 612.293 SNPs, respectivamente. A fim de realizar as análises de varredura genômica, utilizamos o software SHAPEIT (DELANEAU; MARCHINI; ZAGURY, 2012) com os argumentos-padrão para fasear os conjuntos de dados, utilizando como mapa de recombinação os dados do Projeto 1000 Genomas (1000 Genomes Project Consortium et al., 2015). Adicionalmente, uma vez que abordagens específicas de seleção requerem

a diferenciação dos alelos ancestrais (*e.g.* iHS), utilizamos também os dados do Projeto 1000 Genomas (<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/>) para ajustar os alelos ancestrais, descartando as variantes sem anotação disponível, o que resultou em 517.984 e 601.648 SNPs para as análises derivadas da EHH (vide tópico 2.2.5).

Tabela 2.1 – Nativos Americanos – Dataset S1.

População	Grupo Linguístico	Região	Latitude	Longitude	Referência	N
Apalai	Karib	Brasil	-1.33	-54.67	Skoglund	4
Arara	Karib	Brasil	-3.91	-53.59	Skoglund	4
Asurini	Tupi	Brasil	-3.89	-51.05	Este	1
Gavião	Jê	Brasil	-10.17	-61.13	Castro e Silva	2
Guarani_GN	Tupi	Brasil	-23.33	-54.50	Skoglund	7
Guarani_KW	Tupi	Brasil	-23.33	-55.20	Skoglund	10
Guarani_Mbya	Tupi	Brasil	-23.10	-55.00	Castro e Silva	4
Karitiana	Tupi	Brasil	-8.75	-63.84	Skoglund/HGDP	17
Munduruku	Tupi	Brasil	-6.38	-59.15	Este	2
Parakanã	Tupi	Brasil	-5.37	-51.28	Castro e Silva	3
Surui	Tupi	Brasil	-11.00	-62.00	Skoglund/HGDP	12
Tupiniquim	Tupi	Brasil	-19.88	-40.18	Castro e Silva	1
Urubu_Kaapor	Tupi	Brasil	-2.50	-46.50	Skoglund	3
Wajãpi	Tupi	Brasil	1.05	-52.83	Castro e Silva	2
Xavante	Jê	Brasil	-14.00	-52.50	Skoglund/Este	38
Xikrin	Jê	Brasil	-5.92	-51.00	Este	7
Zoro	Tupi	Brasil	-10.33	-60.33	Skoglund	1
Maya	Mayan	México	19.00	-91.00	HGDP	21
Pima	Uto-Aztecan	México	29.00	-108.00	HGDP	14

Tabela 2.2 – Grupos utilizados nas análises de seleção – Dataset S1.

Grupo	Macrorregião	N
Amazônia	América do Sul	118
Lesta da Ásia	Ásia	231
Mesoamérica	América Central	35

Tabela 2.3 – Nativos Americanos – Dataset S2.

População	País	Ecorregião	Referência	N
América do Sul				
Ashaninka	Peru	Amazonia	GnecchiRuscone	9
Cashibo	Peru	Amazonia	GnecchiRuscone	9
Huambisa	Peru	Amazonia	GnecchiRuscone	6
Karitiana	Brazil	Amazonia	HGDP3	12
Shipibo	Peru	Amazonia	GnecchiRuscone	14
Surui	Brazil	Amazonia	HGDP3	8
Yanesha	Peru	Amazonia	GnecchiRuscone	46
Mesoamérica				
Maya	Mexico	Mesoamerica	HGDP3	20
Mixe	Mexico	Mesoamerica	SGDP	3
Mixtec	Mexico	Mesoamerica	SGDP	2
Pima	Mexico	Mesoamerica	HGDP3	13
Tzotzil	Mexico	Mesoamerica	GnecchiRuscone	23
Zapotec	Mexico	Mesoamerica	SGDP	2

Tabela 2.4 – Grupos utilizados nas análises de seleção – Dataset S2.

Grupo	Macrorregião	N
Amazônia	América do Sul	104
Leste da Ásia	Ásia	225
Mesoamérica	América Central	63

2.2.3 Anotação do Genoma

Realizamos a anotação do genoma com o software ANNOVAR (WANG; LI; HAKONARSON, 2010), bem como uso de scripts personalizados. Os genes foram atribuídos a cada SNP utilizando como referência o banco RefGene, fornecido pelo próprio software ANNOVAR. Para anotação da identificação dos SNPs (dbSNP), utilizamos as identificações providas pelo site comercial do arranjo Axiom Human Origins (<https://www.thermofisher.com/br/en/home/life-science/microarray-analysis/microarray-data-analysis/genechip-array-annotation-files.html>) para o dataset S1, e as aquelas providas pelo banco de dados do 1000G para o dataset S2. Adicionalmente, removemos a anotação dos genes alocados a mais de 10 kb do SNP-alvo.

2.2.4 Estrutura Populacional

Para análise da estrutura populacional, aplicamos análise de PCA em ambos os datasets, utilizando o software Plink v1.9 (CHANG et al., 2015), após filtro de desequilíbrio de ligação com 50 SNPs para cada janela, deslizando de 5 em 5 SNPs, mantendo apenas os SNPs com correlação inferior a 20%, utilizando o mesmo programa.

A análise de miscigenação foi realizada com o software ADMIXTURE (ALEXANDER; NOVEMBRE; LANGE, 2009), adotando uma abordagem não-supervisionada com dez corridas independentes, utilizando amazônicos, europeus e africanos subsaarianos. Para identificar exclusivamente se há componentes europeus ou africanos nos nativos amazônicos (e não subestruturação dentro da Amazônia ou nos continentes analisados), reportamos a análise com os três componentes. Utilizamos o software pong (BEHR et al., 2016) para pós-processamento das corridas e produção do gráfico.

2.2.5 Varreduras Seletivas

Foram aplicados três testes para detecção de sinais de seleção positiva, que diferem em sua abordagem estatística principal. O primeiro, PBS, possui como princípio estatístico a subestruturação populacional por meio do cálculo de Fst. Os dois seguintes, iHS e XP-EHH, baseiam-se no decaimento da extensão de homozigose do haplótipo (EHH). Enquanto iHS, como método intra-populacional, tem-se mostrado mais apropriado na detecção de sweeps incompletos, XP-EHH compara a população-alvo com uma população-irmã e detecta de forma mais robusta sweeps completos (ou muitos próximos da fixação) (SUZUKI, 2010).

Em ambos os datasets, estruturamos a análise de PBS utilizando os amazônicos como população focal, mesoamericanos como população-irmã, e asiáticos do leste como grupo externo. Realizamos a inferência de PBS como originalmente descrito por (YI et al., 2010), utilizando o índice Fst de (REYNOLDS; WEIR; COCKERHAM, 1983) nos dados previamente faseados. Adicionalmente, rodamos uma média móvel por janela utilizando 20 SNPs em cada janela, ao passo que 5 SNPs. Para cada janela, apenas o SNP com maior valor de PBS, bem como suas informações associadas (*e.g.* cromossomo, posição, dbSNP e gene) foram mantidos. Para identificar os SNPs e genes candidatos, utilizamos duas abordagens distintas: 1) por SNP e 2) por gene. Na abordagem por SNP, selecionamos os SNPs cujo valor de PBS é igual ou superior ao quantil de 0,999, ou seja, os top 0,1% da distribuição (p-valor unicaudal empírico $< 0{,}001$). Na análise por gene, calculamos a média de PBS por gene e, nesta distribuição, de maneira similar à abordagem por SNP, selecionamos como top 0,1% dos genes na distribuição como genes candidatos. Em ambas as abordagens, calculamos o p-valor empírico unilateral dos dados considerando a proporção de valores maiores ou iguais ao valor-alvo, corrigindo para amostra finita. Uma maneira de fazer isso, é utilizando a fórmula $p = (1 + \sum s \geq s_i)/(N + 1)$, onde s é o

vetor de valores, s_i o valor-alvo, e N o tamanho do vetor. Outra maneira, utilizando o mesmo princípio, consiste em primeiro ranquear os valores-alvos, possibilitando o uso da vetorização com o pacote data.table em R, através da fórmula $p = 1 - r/(N + 1)$, onde r consiste no valor ranqueado proveniente da função frank do pacote data.table do R. Optamos pela segunda opção, visto que é mais de 4.000 vezes mais rápida que a primeira. Adicionalmente, computamos o logaritmo negativo dos p-valores, na base 10, onde um p-valor de 0,01 equivale a 2.

Nas análises derivadas de EHH, utilizamos os dados previamente faseados e polarizados (corrigidos para o alelo ancestral) como entrada para as funções do pacote rehh do R (GAUTIER; VITALIS, 2012; GAUTIER; KLASSMANN; VITALIS, 2017), utilizando os valores-padrão das funções *ihh2ihs* e *ies2xpehh* para realizar as inferências dos valores normalizados de iHS e XP-EHH, respectivamente. Em ambos os datasets, a análise de XP-EHH foi estruturada utilizando os amazônicos como população-alvo e os mesoamericanos como população-irmã, ao passo que iHS, como método intra-populacional, utilizou-se apenas os amazônicos. Adicionalmente, para o método iHS, transformamos os valores inferidos em valores absolutos, uma vez que tanto valores negativos quanto positivos são indicativos de seleção. De maneira similar ao método PBS, calculamos p-valores empíricos e aplicamos a abordagem de detecção por SNP e por gene para identificar SNPs e genes candidatos, focando no extremo 0,1% da distribuição (*i.e.* p-valor de 0,001).

Aqui, optamos por realizar tanto a abordagem por SNP quanto por gene como meio de identificar os SNPs e genes candidatos porque ambas possuem suas vantagens e desvantagens. Ao passo que a abordagem por SNP – segundo (WENG et al., 2011), a mais utilizada nas análises de enriquecimento gênico – pode representar com maior fidelidade a relevância dos SNPs identificados dado seus valores extremos, os SNPs identificados representam apenas uma pequena fração das variantes genéticas que contribuem para fenótipos complexos (SHRINER et al., 2007), além de que, genes maiores podem apresentar SNPs com altos valores por acaso. Na abordagem por gene, por sua vez, ao passo que captura todos os valores de SNPs por gene, genes com SNPs que apresentam alta variância tendem a ser ignorados. De forma resumida, a abordagem por SNP pode resultar em falsos positivos, ao passo que a abordagem por gene tende a apresentar falsos negativos. Os ruídos dos SNPs falsos positivos, contudo, são drasticamente diminuídos pela abordagem da média em janelas móveis.

Os resultados provenientes das varreduras genômicas foram utilizados também para construir uma lista de genes para as análises de super-representação gênica (ORA) e enriquecimento gênico (GSEA). Nestes casos, como uma quantidade maior de genes candidatos foi encontrada, selecionamos nestas análises apenas aqueles cujos p-valores correspondentes (corrigidos por FDR) eram significativos a um nível de 5%, definimos, também, diferentes valores de corte da distribuição além dos extremos 0,1%, como 0,5% e

1%, resultando em diferentes conjuntos de genes candidatos (vide item 2.2.6). Adicionalmente, para cada conjunto de SNPs e genes candidatos, realizamos o teste não-paramétrico de Wilcoxon (WILCOXON, 1945) a fim de verificar se estes candidatos estão significativamente distantes da mediana da distribuição, visto que, em todos os casos (ambos datasets para três métodos), os dados possuem distribuição não-normal (teste de Shapiro-Wilk $< 2,2^{-16}$).

2.2.6 Anotação Funcional

A fim de compreender as funções biológicas dos SNPs e genes candidatos, utilizamos três banco de dados - Ensembl (YATES et al., 2015), GWAS Catalog (BUNIELLO et al., 2019; MACARTHUR et al., 2017) e GeneCards (STELZER et al., 2016) - para mapear os fenótipos associados aos genes candidatos. Para isso, utilizamos a REST API tanto do Ensembl quanto do GWAS Catalog para realizar as nossas buscas.

Além da anotação de fenótipos derivados dos genes candidatos, realizamos uma busca mais específica focando nos alelos-alvo dos SNPs candidatos utilizando o banco de dados GTEx (GTEx Consortium, 2013). Nesta análise, selecionamos apenas os alelos dos SNPs candidatos com maior frequência na população amazônica quando comparado aos mesoamericanos, previamente identificados pelos métodos PBS e XP-EHH (p-valor $< 0{,}01$), e então buscamos seus valores eQTL para os respectivos genes utilizando a REST API do banco GTEx (GTEx Consortium, 2013), para cada um dos 48 tecidos disponíveis na versão "gtex_v7" (referente à versão do genoma Grch37), mantendo apenas aqueles cujo p-valor de eQTL é menor que 0,01. Uma vez que não é possível distinguir programaticamente o papel biológico do aumento ou diminuição da expressão dos genes no tecidos, reportamos tanto a contagem de alelos-alvo para o aumento (*up*) ou diminuição (*down*) da expressão nos tecidos, como também somamos a contagem do aumento e diminuição por tecido para verificar em quais tecidos os alelos candidatos estão mais expressos, independentemente da maneira de expressão (promovendo ou reprimindo).

2.2.7 Enriquecimento Gênico

Posteriormente às análises de varreduras genômicas, realizamos análises de super-representação gênica (ORA) e enriquecimento gênico (GSEA) utilizando quatro plataformas: Enrichr (CHEN et al., 2013; KULESHOV et al., 2016), FUMAGWAS (WATANABE et al., 2017), WebGestalt (ZHANG; KIROV; SNODDY, 2005; WANG; LIAO, 2020), e GOATOOLS (KLOPFENSTEIN et al., 2018), focando em três bancos de dados: GWAS Catalog (MACARTHUR et al., 2017), KEGG Pathway (KANEHISA; GOTO, 2000) e Gene Ontology (ASHBURNER et al., 2000; The Gene Ontology Consortium, 2019). Optamos primeiramente por Enrichr para analisar os bancos de dados GWAS Catalog e KEGG, uma vez que o mesmo contém uma API em Python, GSEApy (FANG et al.,

2020), possibilitando a análise em conjunto dos diferentes grupos de genes candidatos mais facilmente. Contudo, uma vez que o mesmo não possibilita o uso de uma lista de população gênica personalizada (utilizando os genes do SNP-Array), utilizamos também os softwares FUMAGWAS e WebGestalt para análise nos bancos GWAS Catalog e KEGG Pathway, respectivamente, especificando como população gênica os genes disponíveis em nossos dados (Affymetrix e Illumina para os datasets S1 e S2, respectivamente). Para análise de enriquecimento de termos GO, utilizamos a ferramenta GOATOOLS. Dado que esta biblioteca possibilita múltiplas manipulações e agrupamentos dos termos GO, analisamos tanto o conjunto completo dos termos GO relacionados aos processos biológicos, bem como sua versão não-redundante (GO slim), e termos específicos do sistema imunológico. Neste programa, selecionamos como método de correção dos p-valores a implementação de FDR baseada em reamostragem (*resampling-based FDR*), disponibilizado pelo próprio programa, enquanto para as demais plataformas utilizamos a correção padrão de FDR. Para todos os softwares, selecionamos como genes candidatos aqueles identificados no extremo 0,1%, 0,5% e 1% superior à distribuição de ambas as abordagens por SNP e por gene.

2.2.8 Convergência Evolutiva

Para análise de convergência evolutiva, selecionamos todos os artigos de seleção em populações nativas de Floresta Tropicais disponíveis, totalizando 13 estudos. Para cada estudo, anotamos as populações-alvo e os métodos aplicados, e selecionamos os genes candidatos reportados (outliers), geralmente disponíveis na seção de material suplementar. Verificamos e reportamos então a interseção dos genes candidatos levantados nestes trabalhos e dos genes candidatos levantados em nossa pesquisa, tanto de forma individual (por artigo) como de forma unificada (todos artigos em conjunto). Para cada gene em intersecção anotamos os fenótipos associados como descrito no item 2.2.6, e agrupamos estes fenótipos para poder ter uma melhor compreensão daqueles mais frequentes, bem como as informações associadas aos mesmos (quantidade de estudos, genes e positividade nos métodos aplicados).

2.2.9 Disponibilização dos Dados e Códigos

Considerando os conjuntos de dados filtrados e faseados, todos os scripts utilizados nesta pesquisa se encontra disponível no link <https://github.com/cmcouto-silva/tese>, organizados para reprodutibilidade via Snakemake (KOSTER; RAHMANN, 2012), um sistema de gerenciamento de fluxo de trabalho (*workflow*) baseado em Python comumente utilizado na área da bioinformática. Todas as dependências para execução do *workflow* (*e.g.* R, Python e softwares de bioinformática) estão disponíveis tanto via Docker container (<https://hub.docker.com/repository/docker/cmcoutosilva/tese>) quanto arquivo de configuração para ambiente virtual conda (ANACONDA SOFTWARE DISTRIBUTION,

2021). As figuras e tabelas resultantes desta pesquisa estão disponíveis no repositório de dados do Mendeley (<http://dx.doi.org/10.17632/gztff7wmjt.1>).

2.3 RESULTADOS

2.3.1 Estrutura Populacional

A distribuição atual das populações nativas americanas presentes no dataset S1 pode ser verificada na figura 2.2. Cabe ressaltar que mesmo as populações que já não se encontram atualmente na região Amazônica passaram e muito provavelmente permaneceram nela por um longo tempo formativo. Pode-se observar que estas populações se agrupam de maneira distante das populações européias e africanas, estando mais próximas às populações Mesoamericanas e do leste Asiático, padrão que se repete no dataset S2 (Figura 2.3A-B).

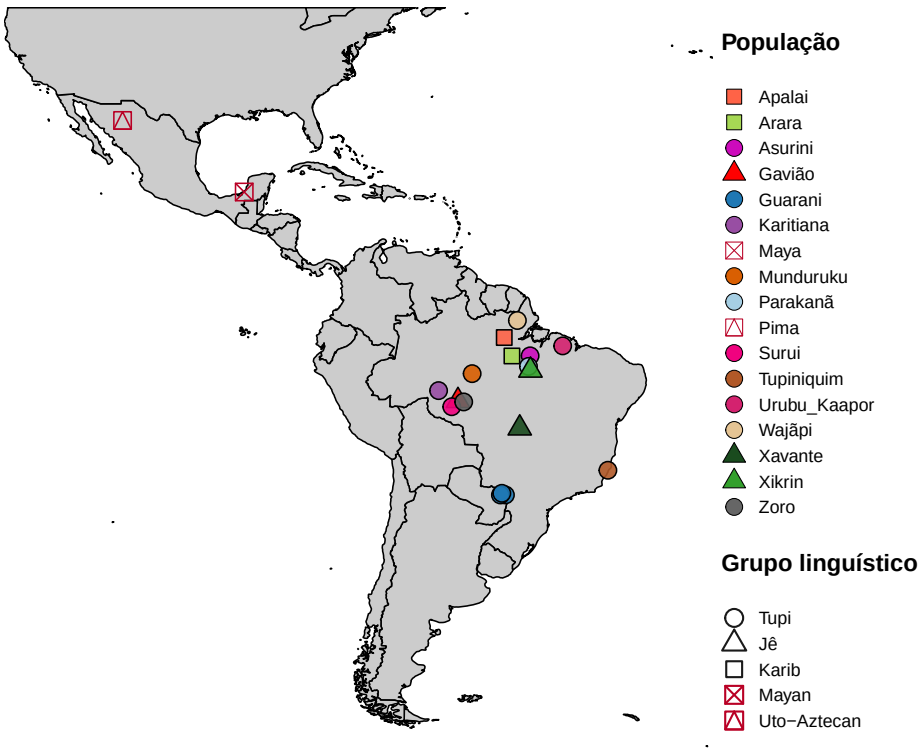

Figura 2.2 – Distribuição das populações nativas Americanas analisadas.

Para avaliar possível miscigenação dos nativos Amazônicos em nosso estudo, utilizamos ADMIXTURE com populações da África Subsaariana e Europa, dividindo em três componentes (Figura 2.3C). Com exceção de possíveis ruídos nos primeiros nativos amazônicos, não se observam componentes presentes em Africanos e Europeus. A miscigenação e dispersão dos nativos Amazônicos do dataset S2 pode ser verificada diretamente no artigo de Gnecchi-Ruscone et al. (2019).

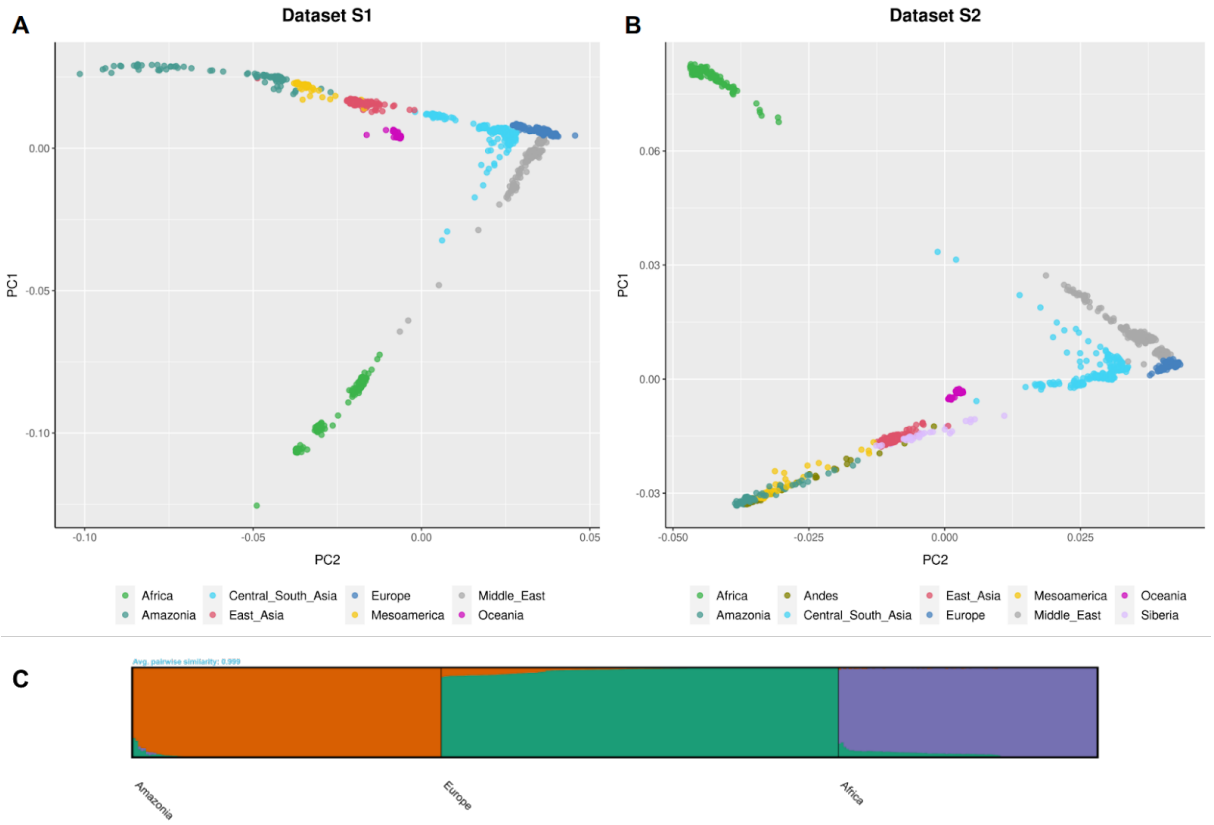

Figura 2.3 – Análise de PCA e ADMIXTURE. (A) PCA para o dataset S1 e (B) dataset S2, englobando todas as populações disponíveis nos respectivos datasets. (C) Análise não-supervisionada de miscigenação no dataset S1, utilizando três componentes (K = 3).

2.3.2 Varreduras Seletivas

Nós computamos três estatísticas inferenciais de seleção que contém propriedades complementares para detecção de eventos adaptativos: PBS, iHS e XP-EHH. Partindo pela análise de PBS, estruturamos a análise com os nativos amazônicos como população focal, nativos da Mesoamerica como população próxima, e populações do Leste Asiático como grupo externo, calculando uma média móvel por janelas deslizantes de 20 SNPs movendo a cada 5 SNPs (vide item 2.2.5). Através dessa abordagem, identificamos 29 SNPs com p-valor empírico $<$ 0,001, dentro ou próximos de 31 genes no dataset S1 (Tabela 2.5). Os genes correspondentes aos SNPs com valores mais altos dentro dessa distribuição, por cromossomo, estão representados na Figura 2.4A. Adicionalmente, calculamos a média dos valores de PBS para cada gene, e recalculamos p-valores empíricos para a nova distribuição, resultando em 12 genes com p-valor menor que 0,01 (Figura 2.4B). Cabe ressaltar que, dado que selecionamos valores extremos na distribuição, todos os valores extremos identificados tanto nas análises de PBS como nas seguintes (*i.e.* iHS e XP-EHH) são significativamente diferentes da mediana observada para a respectiva distribuição (maior p-valor para o teste Wilcoxon foi de $< 2,2^{-16}$).

Repetimos as análises de PBS para o dataset S2, identificando 26 genes na

análise por SNP, e 13 genes na análise por gene (Tabelas 2.6-2.7, Figura 2.6). Destes, encontramos uma intersecção do gene *ACACA* abordagem por SNP (considerando *ACACA* no dataset S2), e *PDK4* na abordagem por gene. Se permitimos a identificação de *outliers* acima do quantil 0,995 ao invés de 0,999, identificamos então 16 genes em intersecção na abordagem por SNP (incluindo *ACACA*, *DOCK2*, *MEGF11*, *PLXNA4* e *NEB*), e 6 genes na abordagem por gene, incluindo *ACACA* e *PDK4* (Figura 2.6).

Tabela 2.5 – Resultado da análise de PBS por SNP no dataset S1.

CHR	POS	SNP	GENE	FUNÇÃO	PBS	-$\log_{10}(p)$
1	162060864	rs10918593	NOS1AP	intergenic	0.2458	3.7893
2	25447908	rs13390436	DTNB	intronic	0.2481	3.8437
2	152438741	rs76206843	FMNL2	intronic	0.2346	3.5726
2	162146324	rs11897425	LOC101929532	ncRNA_intronic	0.1975	3.2136
2	162342778	rs2033300	GCA	intronic	0.2273	3.5284
3	10260685	rs451952	TATDN2	intronic	0.1956	3.1625
4	101198533	rs2659540	PPP3CA	intronic	0.3074	4.7188
4	109860572	rs12498239	LRIT3	intronic	0.1911	3.0955
5	7897170	Affx-27022154	MTRR	exonic	0.2617	3.9058
5	10263613	rs2244964	CCT5	intronic	0.1910	3.0904
5	78395121	rs1864172	SCAMP1	intronic	0.1904	3.0853
5	169742629	rs12514018	DOCK2	intronic	0.2861	4.2416
6	46279855	rs1442227	LOC101926915	ncRNA_intronic	0.1946	3.1506
6	154229719	rs9397697	IPCEF1,OPRM1	intronic,intronic	0.2475	3.8157
6	154400423	rs6917661	CNKSR3	UTR3	0.2347	3.5884
6	170578141	rs7745933	PDCD2	intronic	0.1854	3.0375
7	69664093	rs2533434	AUTS2	intronic	0.2376	3.7188
7	95872168	rs12535988	DYNC1I1	intronic	0.2107	3.3385
7	135973421	rs7801598	LUZP6,MTPN	intronic,intronic	0.1976	3.2204
8	118416705	rs4077747	SAMD12	intronic	0.2920	4.3208
10	77465512	rs607483	KCNMA1	intronic	0.2366	3.6581
11	125204544	rs585974	PKNOX2	intronic	0.1961	3.1873
13	31904154	rs3848084	EEF1DP3	ncRNA_intronic	0.2142	3.3666
13	110395003	rs7326145	COL4A2	intronic	0.1932	3.1333
15	66135032	rs55648467	MEGF11	intronic	0.1840	3.0112
15	78266167	rs2037347	DNAJA4	intronic	0.2008	3.2639
16	23060611	rs11074541	USP31	downstream	0.2195	3.4516
17	37175051	rs2898659	ACACA	intronic	0.3265	5.0198
17	72467730	rs12948544	LINC00673	ncRNA_intronic	0.2070	3.3038

Nas análises derivadas da estatística de EHH, selecionamos os SNPs com p-valores empíricos menores que 0,001 para os métodos XP-EHH e iHS, que por sua vez mapearam para 59 e 70 genes no dataset S1, respectivamente (tabelas disponíveis no repositório Mendeley). De forma similar à análise de PBS, calculamos a média por gene para cada uma destas estatísticas, resultando em 21 e 17 genes. Os genes mais representativos no método XP-EHH estão representados na figura 2.7, ao passo que os resultados do método iHS podem ser verificados no repositório Mendeley.

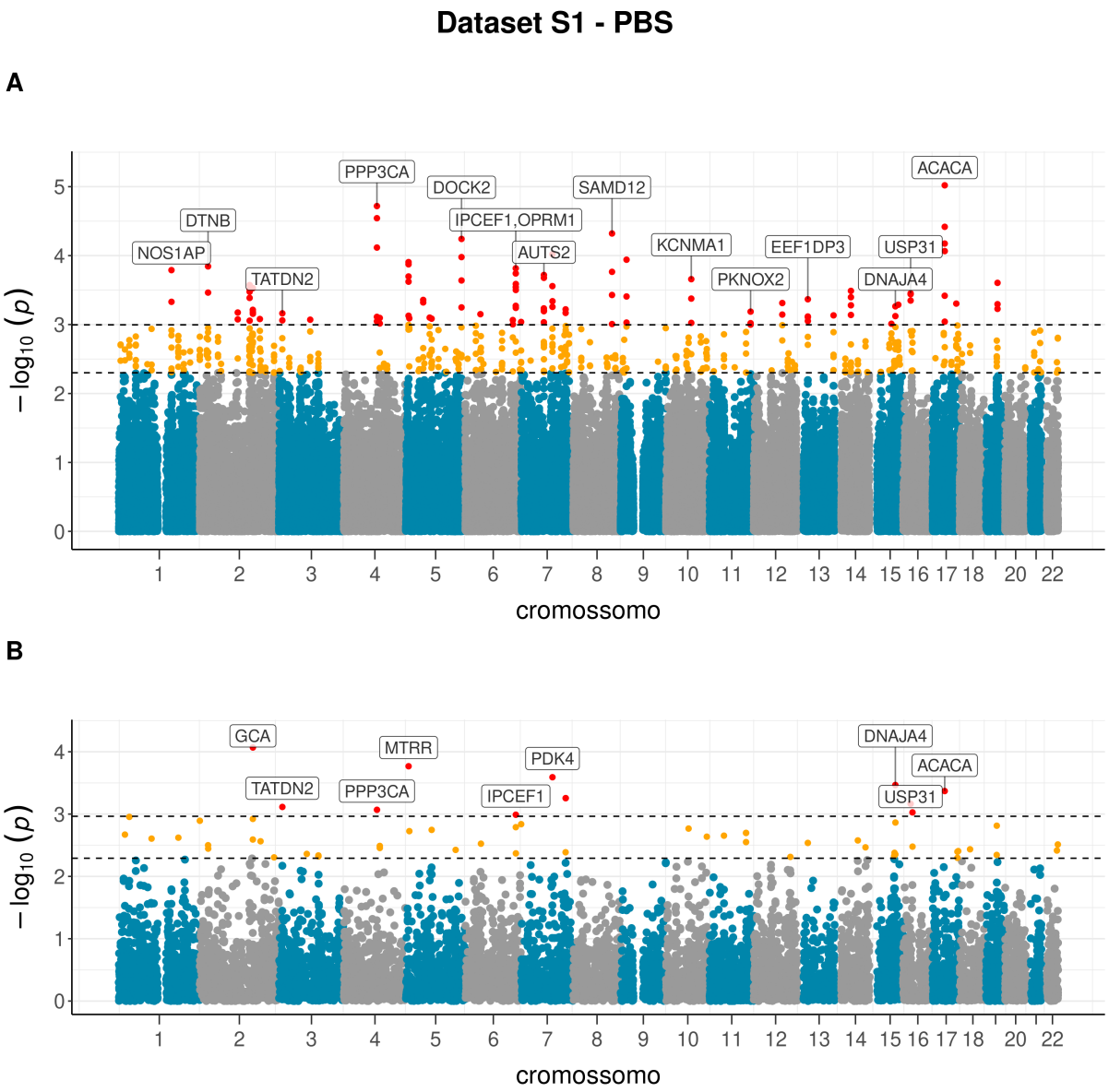

Figura 2.4 – Resultados da análise de PBS para o dataset S1. (A) Abordagem por SNP. (B) Abordagem por gene. Na abordagem por SNP cada ponto corresponde ao SNP com o valor mais alto dentro da média móvel de PBS. Na abordagem por gene cada ponto representa a média dos valores de PBS de todos SNPs para cada gene. As linhas horizontais pontilhadas representam os quantis 0,995 e 0,999, enquanto as cores laranja e vermelha denotam os (A) SNPs e (B) genes distribuídos acima destes quantis, respectivamente. Para cada cromossomo, (A) apontamos os genes correspondentes ao SNP com o valor mais alto na abordagem por SNP e (B) reportamos diretamente o gene com maior valor de PBS, considerando em ambos os casos apenas os pontos com valores iguais ou superiores ao quantil 0.999.

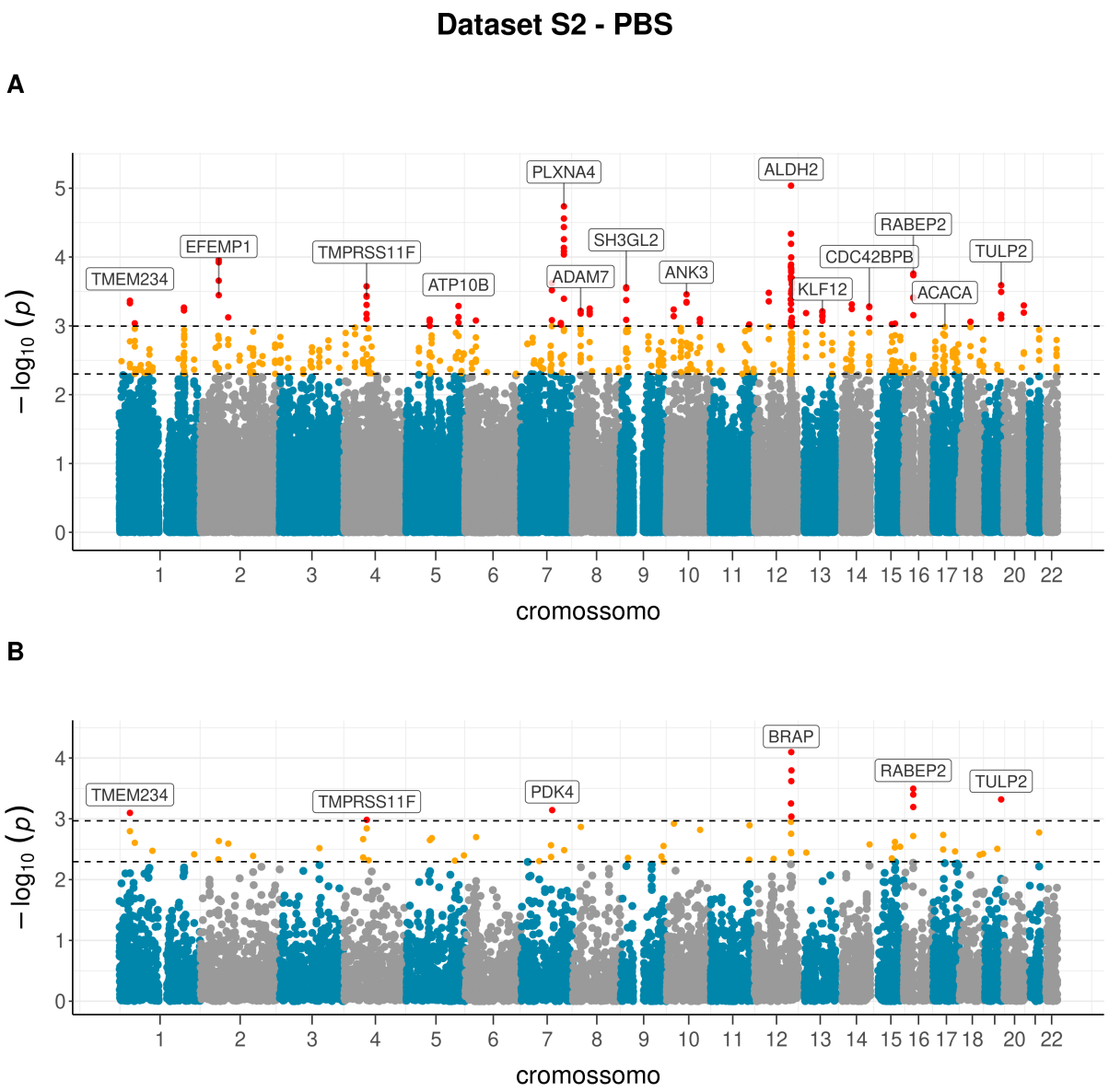

Figura 2.5 – Resultados da análise de PBS para o dataset S2. (A) Abordagem por SNP. (B) Abordagem por gene. Na abordagem por SNP cada ponto corresponde ao SNP com o valor mais alto dentro da média móvel de PBS. Na abordagem por gene cada ponto representa a média dos valores de PBS de todos SNPs para cada gene. As linhas horizontais pontilhadas representam os quantis 0,995 e 0,999, enquanto as cores laranja e vermelha denotam os (A) SNPs e (B) genes distribuídos acima destes quantis, respectivamente. Para cada cromossomo, (A) apontamos os genes correspondentes ao SNP com o valor mais alto na abordagem por SNP e (B) reportamos diretamente o gene com maior valor de PBS, considerando em ambos os casos apenas os pontos com valores iguais ou superiores ao quantil 0.999.

Tabela 2.6 – Resultado da análise de PBS por SNP no dataset S2.

CHR	POS	SNP	GENE	FUNÇÃO	PBS	-$\log_{10}(p)$
1	32686594	rs3738001	TMEM234	intronic	0.2227	3.3645
1	47571391	rs6690005	CYP4Z1	intronic	0.1862	3.0366
2	85115111	rs6728804	TRABD2A	intergenic	0.1958	3.1228
5	71744199	rs10068109	ZNF366	intronic	0.1936	3.0921
5	160134369	rs2033462	ATP10B	intronic	0.2160	3.2884
7	95216262	rs4729201	PDK4	intronic	0.2486	3.6217
7	132094453	rs2341825	PLXNA4	intronic	0.3609	4.7356
8	24339679	rs13255694	ADAM7	exonic	0.2060	3.2171
9	17801555	rs11791520	SH3GL2	intergenic	0.2384	3.5595
10	22504189	rs2148959	EBLN1	intergenic	0.2099	3.2373
10	61801739	rs10821662	ANK3	intronic	0.2325	3.4568
10	102689217	rs2273654	SLF2	intronic	0.1939	3.0971
11	118825636	rs1790191	UPK2	intergenic	0.1849	3.0196
12	111614736	rs6489979	CUX2	intronic	0.2853	3.8605
12	112233018	rs10744777	ALDH2	intronic	0.3860	5.0366
12	112463296	rs4767293	ERP29,NAA25	intergenic,intergenic	0.1916	3.0681
12	112667675	rs1005902	HECTD4	exonic	0.2924	3.9952
12	114254860	rs1043811	RBM19	UTR3	0.1830	3.0032
13	24525604	rs2147995	ANKRD20A19P	intergenic	0.2022	3.1854
13	74532811	rs1570744	KLF12	intronic	0.2049	3.2106
14	103427555	rs10136828	CDC42BPB	intronic	0.2151	3.2808
15	65626219	rs894491	IGDCC3	intronic	0.1856	3.0238
15	75242155	rs2415251	RPP25	intergenic	0.1861	3.0323
16	28922149	rs11646653	RABEP2	intronic	0.2636	3.7579
19	49399821	rs12461075	TULP2	intronic	0.2454	3.5895

Aplicando os mesmos métodos no segundo conjunto de dados (*i.e.* XP-EHH e iHS), identificamos 54 e 135 genes na abordagem de *outliers* por SNP, e 22 e 21 genes na abordagem de genes *outliers* por gene, respectivamente. A figura 2.8 apresenta os *outliers* em ambas as abordagens para o método XP-EHH, enquanto as tabelas para ambos os métodos e as figuras para o método iHS podem ser acessadas via repositório Mendeley.

Pode-se observar que, na abordagem tradicional por SNP, mais genes são mapeados no método iHS em ambos os datasets, quando comparado às estatísticas XP-EHH e PBS (Figuras 2.11-2.12). Logo, verificamos a distribuição dos valores nestes métodos, em ambos os datasets, apresentados nas Figuras 2.9 e 2.10, respectivamente.

Tabela 2.7 – Resultado da análise de PBS por gene nos datasets S1 e S2.

DATASET	CHR	GENE	PBS	-$\log_{10}(p)$
Dataset S1	2	GCA	0.2060	4.0687
	3	TATDN2	0.1636	3.1144
	4	PPP3CA	0.1623	3.0687
	5	MTRR	0.2021	3.7676
	6	IPCEF1	0.1485	2.9895
	7	PDK4	0.2005	3.5915
	7	LUZP6,MTPN	0.1667	3.2558
	15	DNAJA4	0.1719	3.4666
	16	USP31	0.1658	3.1656
	16	ATXN2L	0.1609	3.0273
	17	ACACA	0.1713	3.3697
Dataset S2	1	TMEM234	0.1942	3.0968
	4	TMPRSS11F	0.1916	2.9828
	7	PDK4	0.1953	3.1425
	12	LINC01405	0.2163	3.2517
	12	BRAP	0.2925	4.0968
	12	ALDH2	0.2781	3.6196
	12	ERP29,NAA25	0.1934	3.0361
	12	HECTD4	0.2821	3.7957
	16	RABEP2	0.2621	3.4947
	16	NFATC2IP	0.2270	3.3978
	16	SPNS1	0.1980	3.1937
	19	TULP2	0.2182	3.3186

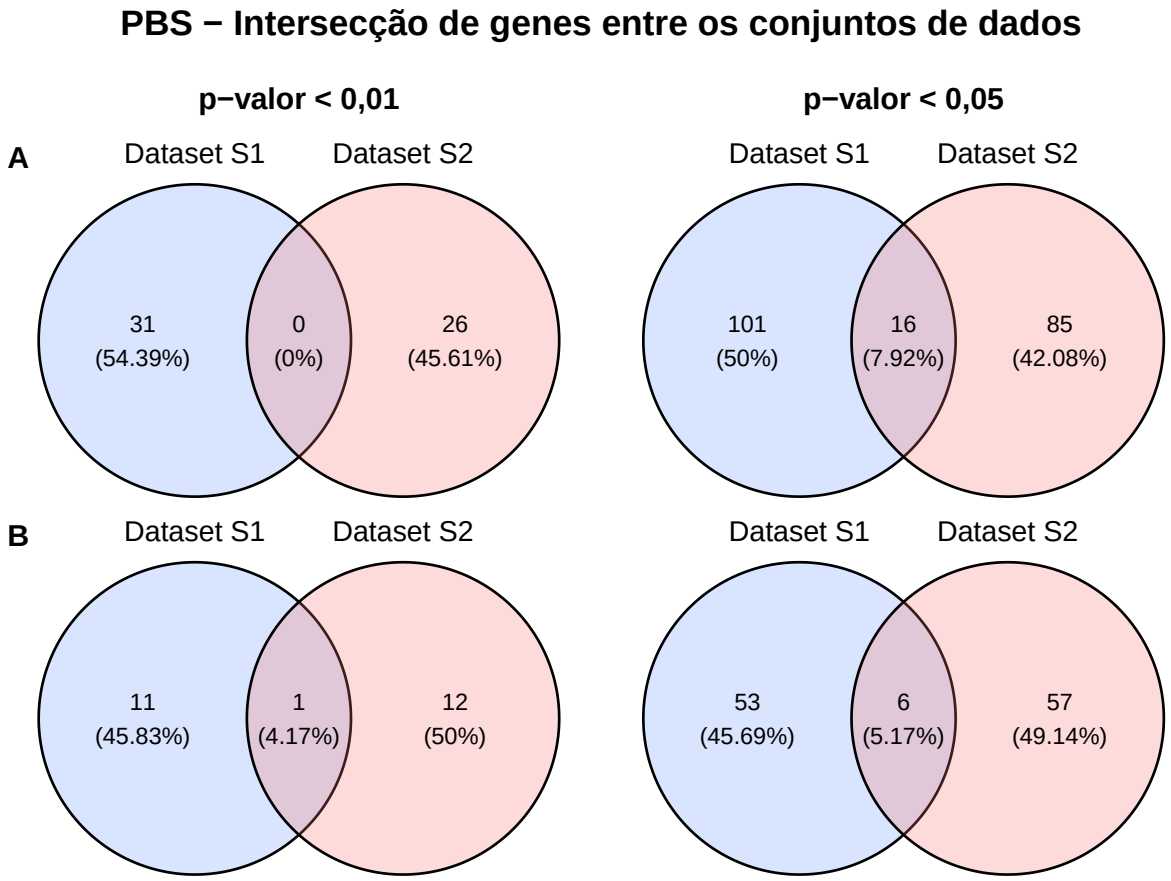

Figura 2.6 – Intersecção dos genes candidatos na análise de PBS entre os datasets. (A) Abordagem por SNP. (B) Abordagem por gene. Os genes foram identificados utilizando como valores de corte os SNPs/genes com p-valor abaixo de 0,01 e 0,05 (colunas 1 e 2), respectivamente.

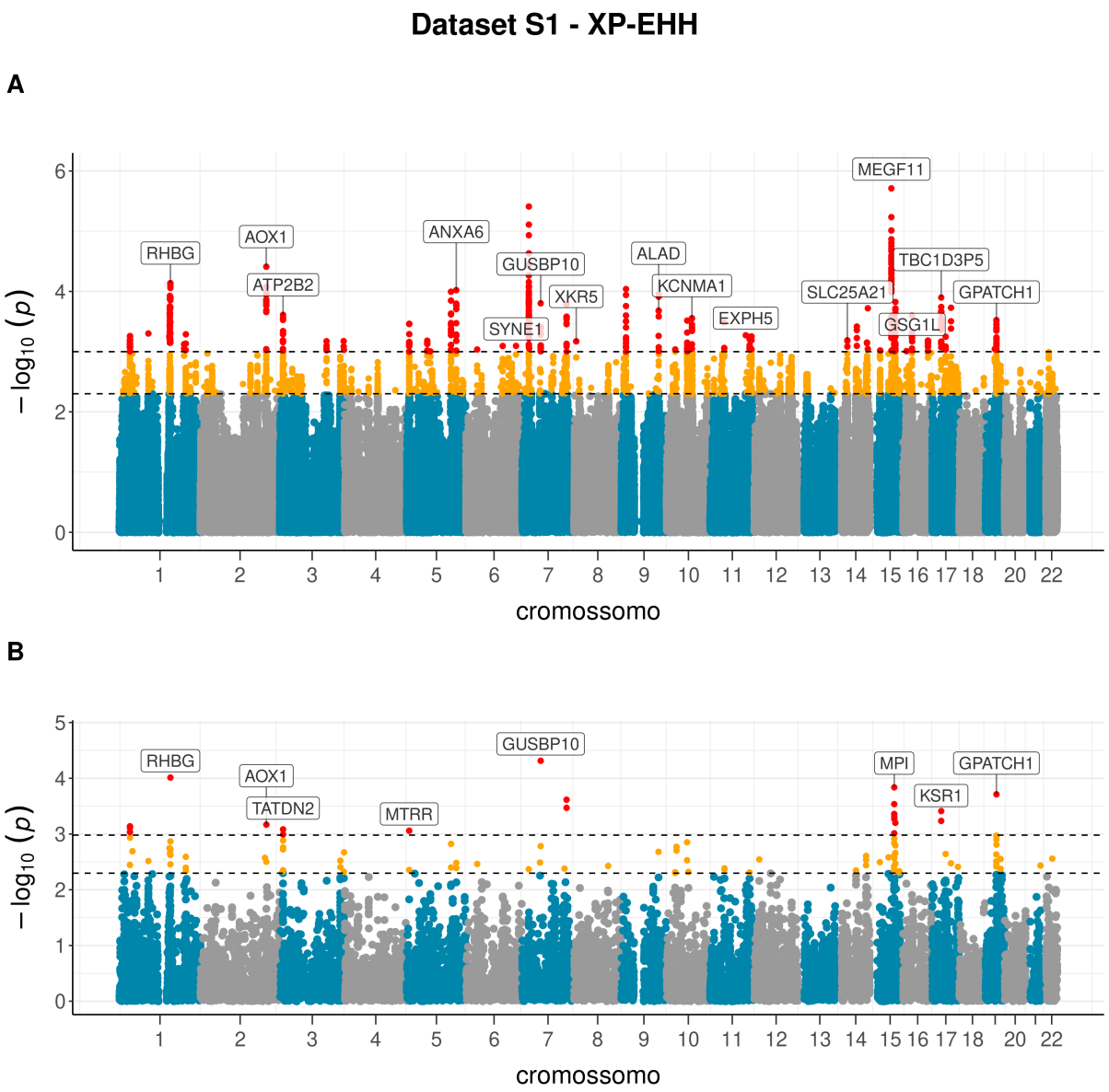

Figura 2.7 – Resultados da análise de XP-EHH para o dataset S1. (A) Abordagem por SNP. (B) Abordagem por gene. As linhas horizontais pontilhadas representam os quantis 0,995 e 0,999, enquanto as cores laranja e vermelha denotam os (A) SNPs e (B) genes distribuídos acima destes quantis, respectivamente. Para cada cromossomo, (A) apontamos os genes correspondentes ao SNP com o valor mais alto na abordagem por SNP e (B) reportamos diretamente o gene com maior valor de XP-EHH, considerando em ambos os casos apenas os pontos com valores iguais ou superiores ao quantil 0.999. Nota-se um acentuado valor no cromossomo 15 na abordagem por SNP, mapeando para o gene *MEGF11*. Os SNPs com valores acentuados no cromossomo 7 não mapeiam para nenhum gene ao redor de 10kb dos mesmos.

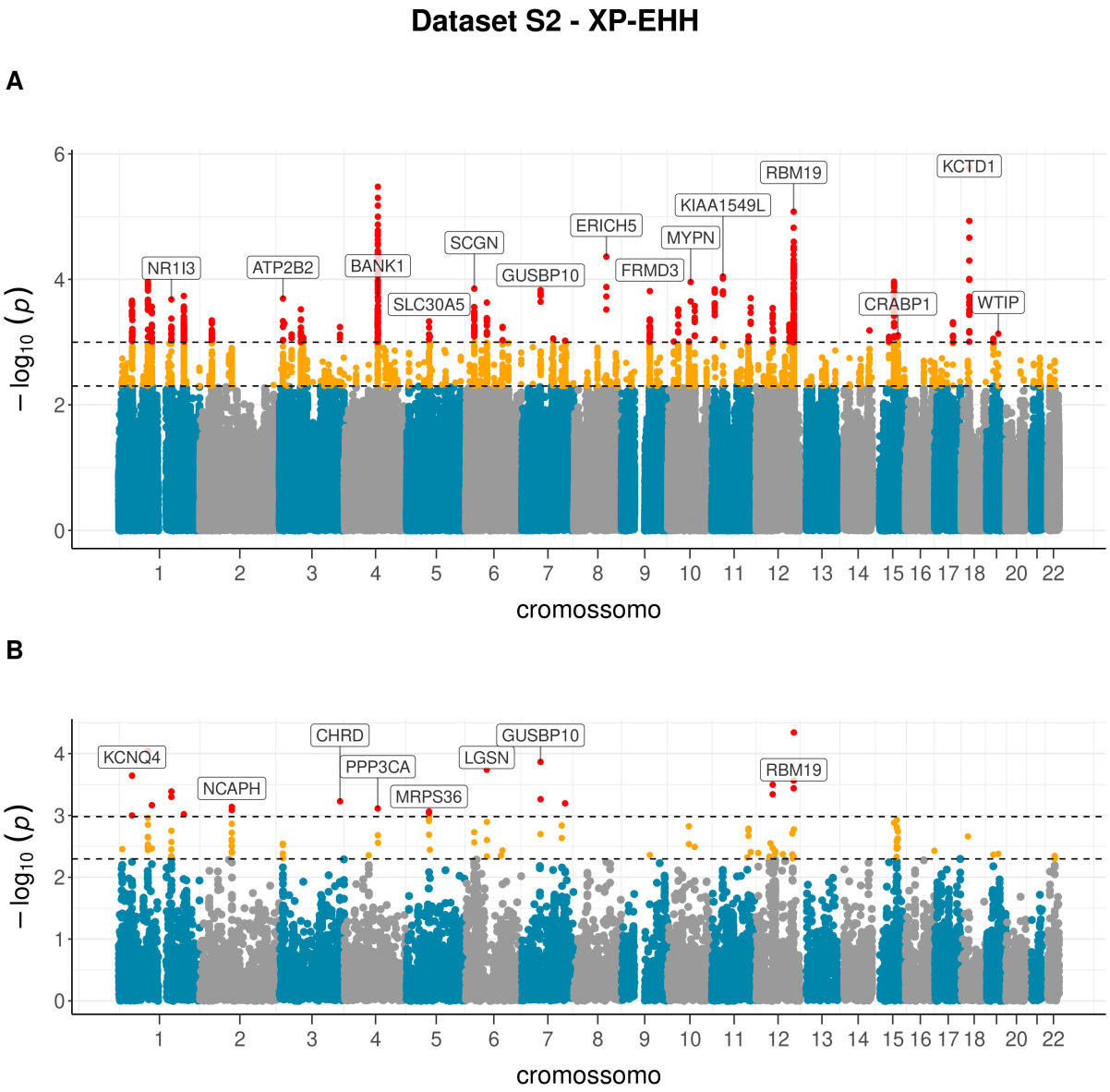

Figura 2.8 – Resultados da análise de XP-EHH para o dataset S2. (A) Abordagem por SNP. (B) Abordagem por gene. As linhas horizontais pontilhadas representam os quantis 0,995 e 0,999, enquanto as cores laranja e vermelha denotam os (A) SNPs e (B) genes distribuídos acima destes quantis, respectivamente. Para cada cromossomo, (A) apontamos os genes correspondentes ao SNP com o valor mais alto na abordagem por SNP e (B) reportamos diretamente o gene com maior valor de XP-EHH, considerando em ambos os casos apenas os pontos com valores iguais ou superiores ao quantil 0.999. Os SNPs com valores mais acentuados no cromossomo 4, na abordagem por SNP, não mapeiam para nenhum gene ao redor de 10kb dos mesmos.

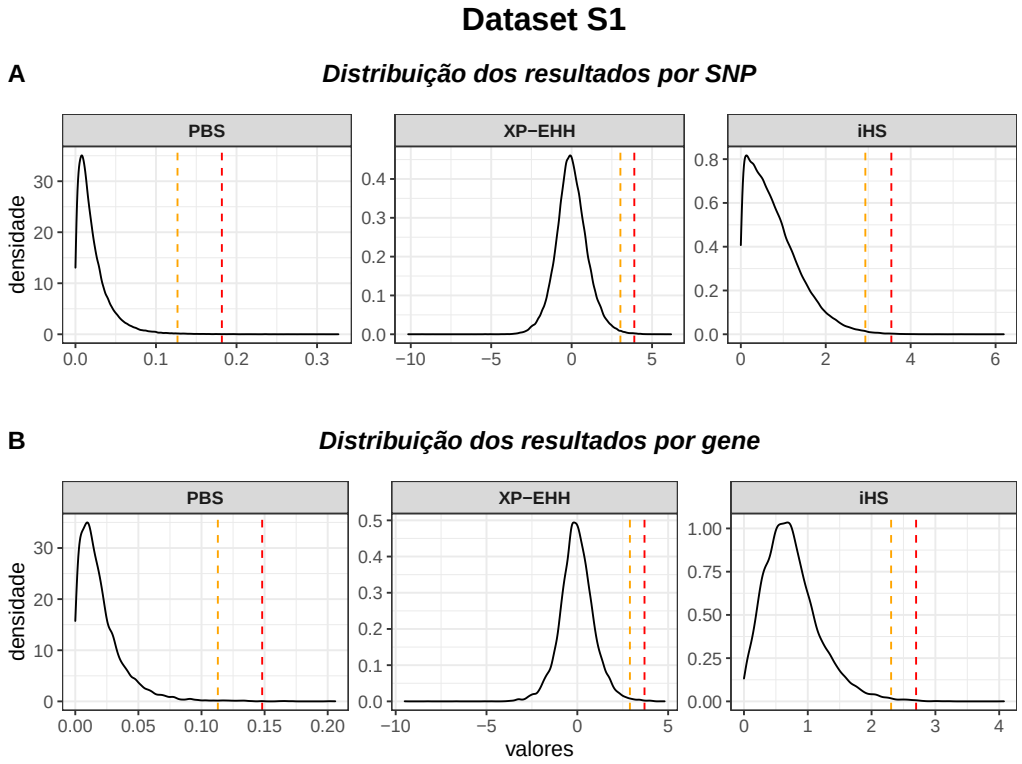

Figura 2.9 – Distribuição dos valores dos testes de seleção no Dataset S1. (A) Abordagem por SNP. (B) Abordagem por gene. As linhas verticais amarelas e vermelhas representam os quantis 0.995 e 0.999, respectivamente.

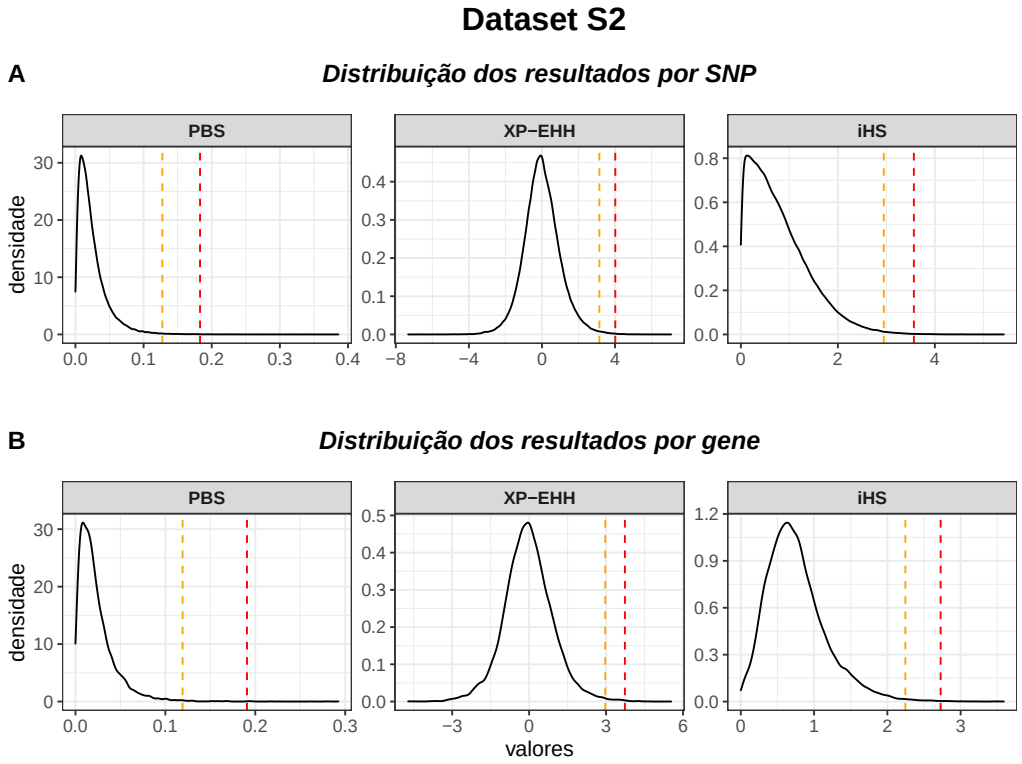

Figura 2.10 – Distribuição dos valores dos testes de seleção no Dataset S2. (A) Abordagem por SNP. (B) Abordagem por gene. As linhas verticais amarelas e vermelhas representam os quantis 0.995 e 0.999, respectivamente.

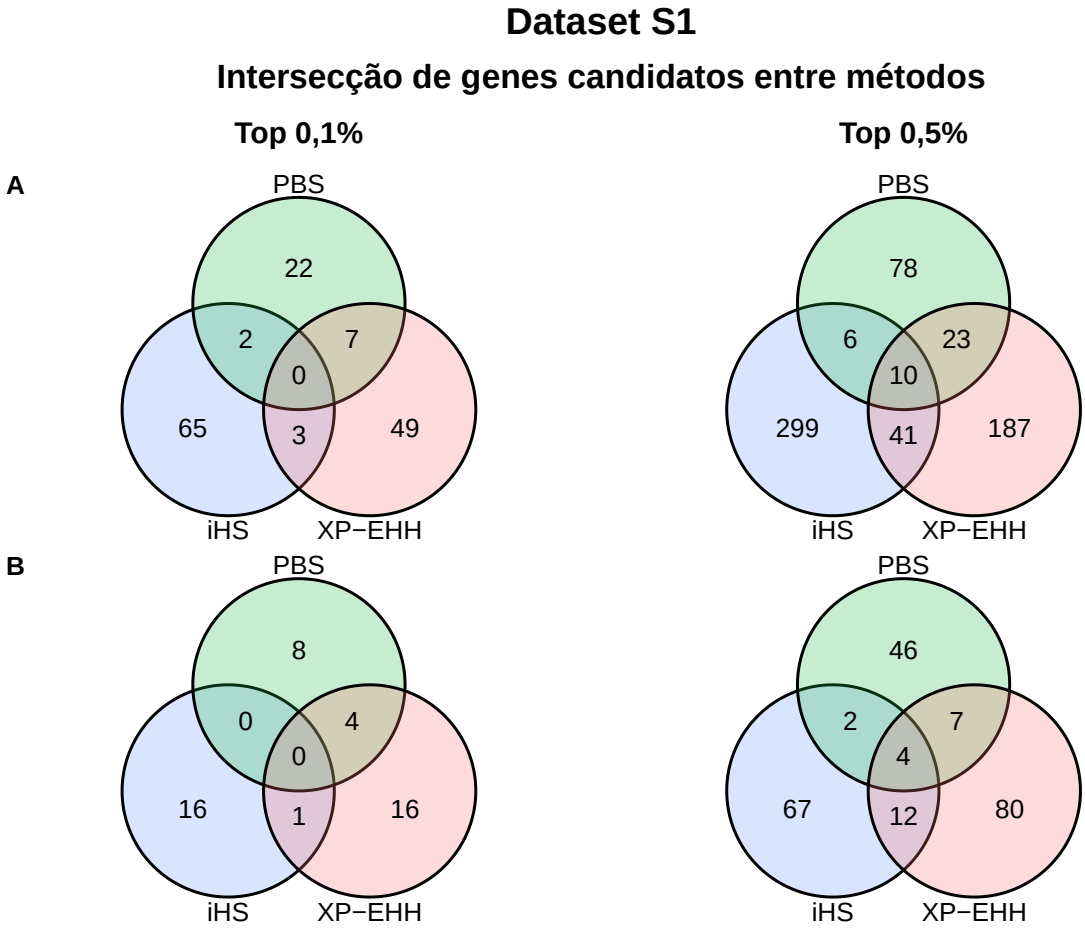

Figura 2.11 – Intersecção dos genes candidatos no dataset S1. (A) Abordagem por SNP. (B) Abordagem por gene. No painel esquerdo foi considerado o extremo 0,1% da distribuição das estatísticas para identificação dos genes candidatos, ao passo que, no painel direito, 0,5%.

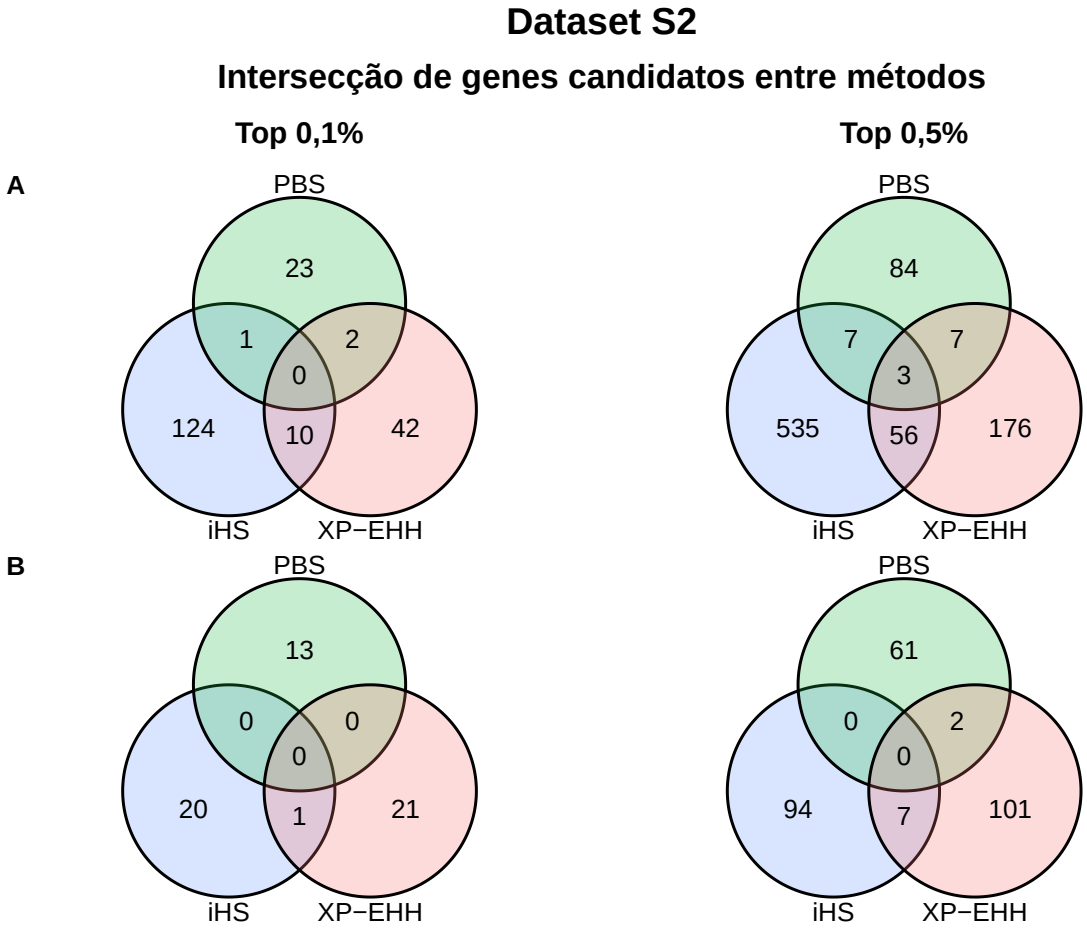

Figura 2.12 – Intersecção dos genes candidatos no dataset S1. (A) Abordagem por SNP. (B) Abordagem por gene. No painel esquerdo foi considerado o extremo 0,1% da distribuição das estatísticas para identificação dos genes candidatos, ao passo que, no painel direito, 0,5%.

2.3.3 Anotação Funcional

Para compreender melhor os fenótipos resultantes dos genes levantados pelas análises de seleção acima apontadas, buscamos os fenótipos associados aos mesmos em dois bancos de dados públicos: Ensembl e GWAS Catalog. As tabelas 2.8-2.11 contém os resultados destas da abordagem por SNP nos dois datasets analisados. Nestas tabelas, ordenamos os fenótipos primeiramente pela quantidade de genes candidatos mapeados para os mesmos (colunas "N"), e então por ordem alfabética, restringindo cada tabela aos trinta primeiros fenótipos.

Adicionalmente, verificamos também a expressão dos alelos candidatos provenientes das análises de PBS e XP-EHH (p-valor $< 0{,}01$). Para cada SNP, definimos como alelo-alvo aquele com maior frequência na população amazônica quando comparado à Mesoamérica, e buscamos no banco de dados GTEx sua influência na expressão dos seus respectivos genes, mantendo apenas os alelos cujo eQTL possui p-valor menor que 0,01 no tecido (Figuras 2.13 e 2.14). Reportamos aqui a contagem dos alelos com eQTL significativo considerando ambos os métodos e datasets em conjunto, ao passo que as figuras separadas e as tabelas completas podem ser verificadas no repositório Mendeley.

Tabela 2.8 – Anotação dos fenótipos na abordagem por SNP no dataset S1, utilizando o banco de dados do GWAS Catalog.

PBS		XP-EHH		iHS	
Fenótipo	N	Fenótipo	N	Fenótipo	N
body mass index	9	body height	11	body height	26
body height	8	body mass index	8	FEV/FEC ratio	14
heel bone mineral density	8	self reported educational attainment	8	acute myeloid leukemia	12
acute myeloid leukemia	6	monocyte count	7	mathematical ability	12
type II diabetes mellitus	6	smoking status measurement	7	schizophrenia	12
self reported educational attainment	5	heel bone mineral density	6	adolescent idiopathic scoliosis	11
asthma	4	lymphocyte count	6	body mass index	11
chronotype measurement	4	mean corpuscular hemoglobin	6	breast carcinoma	11
diastolic blood pressure	4	platelet count	6	gut microbiome measurement	11
lymphocyte count	4	red blood cell distribution width	6	self reported educational attainment	11
mathematical ability	4	reticulocyte measurement	6	smoking behavior	10
platelet count	4	blood protein measurement	5	type II diabetes mellitus	10
Abnormality of refraction	3	diastolic blood pressure	5	colorectal cancer	9
BMI-adjusted waist circumference	3	eosinophil count	5	erythrocyte count	9
DNA methylation	3	lymphocyte percentage of leukocytes	5	intelligence	9
balding measurement	3	mean corpuscular volume	5	monocyte count	9
bone density	3	schizophrenia	5	smoking status measurement	9
chronic lymphocytic leukemia	3	adolescent idiopathic scoliosis	4	cognitive function measurement	8
hematocrit	3	breast carcinoma	4	forced expiratory volume	8
leukocyte count	3	colorectal cancer	4	heel bone mineral density	8
mean corpuscular hemoglobin	3	eosinophil percentage of leukocytes	4	hemoglobin measurement	8
monocyte count	3	erythrocyte count	4	high density lipoprotein cholesterol measurement	8
platelet crit	3	gut microbiome measurement	4	triglyceride measurement	8
serum IgG glycosylation measurement	3	hemoglobin measurement	4	Alzheimer's disease	7
testosterone measurement	3	mathematical ability	4	BMI-adjusted waist circumference	7
unipolar depression	3	mean arterial pressure	4	alcohol consumption measurement	7
vital capacity	3	monocyte percentage of leukocytes	4	bipolar disorder	7
waist-hip ratio	3	myeloid white cell count	4	chronic obstructive pulmonary disease	7
Alzheimer's disease	2	reticulocyte count	4	response to bronchodilator	7
Trypanosoma cruzi seropositivity	2	BMI-adjusted waist-hip ratio	3	unipolar depression	7

Tabela 2.9 – Anotação dos fenótipos na abordagem por SNP no dataset S2, utilizando o banco de dados do Ensembl.

PBS		XP-EHH		iHS	
Fenótipo	N	Fenótipo	N	Fenótipo	N
Body Mass Index	10	Body Mass Index	7	Body Mass Index	28
Echocardiography	6	Hypertension	6	Metabolite levels	20
Refractive error	6	Metabolite levels	6	Myocardial infarction	18
Diastolic blood pressure	5	Systolic blood pressure	6	Hip	16
Systolic blood pressure x alcohol consumption	5	Blood pressure	5	Blood pressure	15
Alcohol consumption (max-drinks)	4	Diastolic blood pressure	5	Cholesterol, LDL	15
Blood pressure	4	Refractive error	5	Body Weight	14
Chronotype	4	Body Height	4	Coronary Artery Disease	14
Coffee consumption	4	Coronary Artery Disease	4	Refractive error	14
Crohn's disease	4	Echocardiography	4	Stroke	14
Diastolic blood pressure x alcohol consumption	4	Heel bone mineral density	4	Cholesterol, HDL	12
Estimated glomerular filtration rate	4	Height	4	Blood protein levels	11
Inflammatory bowel disease	4	Lipids	4	Cholesterol	11
Mean arterial pressure x alcohol consumption	4	Triglycerides	4	Echocardiography	11
Response to alcohol consumption (flushing response)	4	Atrial fibrillation	3	Heel bone mineral density	11
Triglycerides	4	Autism spectrum disorder or schizophrenia	3	Intraocular pressure	11
Type 2 diabetes	4	Cardiovascular phenotype	3	Waist circumference	11
Urate levels	4	Cholesterol, LDL	3	Heart Rate	10
Alcohol dependence symptom count	3	Estimated glomerular filtration rate	3	Obesity-related traits	10
Asthma	3	Hemoglobin S	3	Body Height	9
Body Weight	3	Long QT syndrome	3	Heart failure	9
Cholesterol, LDL	3	Myocardial infarction	3	Iron	9
Coronary Artery Disease	3	Obesity-related traits	3	Schizophrenia	9
Heel bone mineral density	3	Primary dilated cardiomyopathy	3	Tunica Media	9
Hip	3	Pulse pressure	3	Height	8
Mean arterial pressure	3	Tunica Media	3	Inflammatory bowel disease	8
Myocardial infarction	3	Alcohol consumption (max-drinks)	2	Small cell lung carcinoma	8
Serum uric acid levels	3	Amyotrophic lateral sclerosis (sporadic)	2	Systolic blood pressure	8
Systolic blood pressure	3	Apolipoprotein B levels	2	Triglycerides	8
Adventurousness	2	Arrhythmogenic right ventricular cardiomyopathy	2	squamous cell lung carcinoma	8

Tabela 2.10 – Anotação dos fenótipos na abordagem por gene no dataset S1, utilizando o banco de dados do GWAS Catalog.

PBS		XP-EHH		iHS	
Fenótipo	N	Fenótipo	N	Fenótipo	N
body mass index	4	red blood cell distribution width	5	mean platelet volume	3
chronotype measurement	3	body mass index	4	systolic blood pressure	3
self reported educational attainment	3	diastolic blood pressure	4	mean arterial pressure	2
celiac disease	2	monocyte count	4	mean corpuscular volume	2
hematocrit	2	self reported educational attainment	4	platelet count	2
platelet count	2	smoking status measurement	4	triglyceride measurement	2
type II diabetes mellitus	2	hemoglobin measurement	3	Corneal astigmatism	1
Common variable immunodeficiency	1	low density lipoprotein cholesterol measurement	3	FEV/FEC ratio	1
Crohn's disease	1	mean corpuscular hemoglobin	3	Moderate albuminuria	1
DNA methylation	1	platelet count	3	PR interval	1
Keratoconus	1	systolic blood pressure	3	QT interval	1
Trypanosoma cruzi seropositivity	1	alcohol consumption measurement	2	alcohol drinking	1
Vitiligo	1	alcohol drinking	2	allergen exposure measurement	1
acute myeloid leukemia	1	body height	2	asthma	1
albumin:globulin ratio measurement	1	breast carcinoma	2	atrial fibrillation	1
allergy	1	coffee consumption	2	atypical femoral fracture	1
ankylosing spondylitis	1	hypertension	2	basal cell carcinoma	1
anorexia nervosa	1	mean arterial pressure	2	blood protein measurement	1
arterial stiffness measurement	1	mean corpuscular hemoglobin concentration	2	body height	1
asthma	1	mean corpuscular volume	2	celiac disease	1
atopic asthma	1	monocyte percentage of leukocytes	2	cocaine dependence	1
autoimmune disease	1	myeloid white cell count	2	coffee consumption	1
autoimmune thyroid disease	1	response to antineoplastic agent	2	corneal topography	1
bitter non-alcoholic beverage consumption measurement	1	risk-taking behaviour	2	diastolic blood pressure	1
body height	1	systemic lupus erythematosus	2	electrocardiography	1
body weight	1	BMI-adjusted hip circumference	1	erythrocyte count	1
brain aneurysm	1	BMI-adjusted waist circumference	1	glomerular filtration rate	1
brain measurement	1	Crohn's disease	1	hormone measurement	1
chronic lymphocytic leukemia	1	Eczema	1	keratinocyte carcinoma	1
cognitive function measurement	1	albuminuria	1	leukocyte count	1

Tabela 2.11 – Anotação dos fenótipos na abordagem por gene no dataset S2, utilizando o banco de dados do GWAS Catalog.

PBS		XP-EHH		iHS	
Fenótipo	N	Fenótipo	N	Fenótipo	N
body mass index	7	PHF-tau measurement	3	platelet count	2
alcohol drinking	6	body mass index	3	red blood cell density measurement	2
coronary artery disease	5	electrocardiography	3	type II diabetes mellitus	2
high density lipoprotein cholesterol measurement	5	CCL2 measurement	2	3-hydroxy-1-methylpropylmercapturic acid	1
myocardial infarction	5	DNA methylation	2	Agents acting on the renin-angiotensin system	1
parental longevity	5	FEV/FEC ratio	2	FEV/FEC ratio	1
systolic blood pressure	5	PR interval	2	Hyperhidrosis	1
uric acid measurement	5	Trypanosoma cruzi seropositivity	2	acute myeloid leukemia	1
alcohol consumption measurement	4	acute myeloid leukemia	2	age-related macular degeneration	1
diastolic blood pressure	4	blood metabolite measurement	2	albuminuria	1
mean arterial pressure	4	body height	2	blood osmolality measurement	1
reaction time measurement	4	chronic kidney disease	2	cardiovascular disease	1
eosinophil count	3	chronotype measurement	2	diastolic blood pressure	1
gout	3	corneal endothelial cell measurement	2	disease progression measurement	1
hemoglobin measurement	3	eye colour measurement	2	erythrocyte count	1
intelligence	3	generalised epilepsy	2	generalised epilepsy	1
low density lipoprotein cholesterol measurement	3	monocyte count	2	generalized anxiety disorder	1
mean corpuscular hemoglobin concentration	3	platelet count	2	glucose measurement	1
mean corpuscular volume	3	response to anticonvulsant	2	hemoglobin measurement	1
metabolic syndrome	3	schizophrenia	2	hypertension	1
platelet count	3	sex hormone-binding globulin measurement	2	iron biomarker measurement	1
platelet crit	3	smoking behavior	2	lip morphology measurement	1
serum gamma-glutamyl transferase measurement	3	testosterone measurement	2	lymphocyte count	1
stroke	3	type II diabetes mellitus	2	lymphocyte percentage of leukocytes	1
type II diabetes mellitus	3	vital capacity	2	mathematical ability	1
urate measurement	3	Alzheimer's disease	1	mean corpuscular hemoglobin	1
C-reactive protein measurement	2	BMI-adjusted waist circumference	1	mean corpuscular hemoglobin concentration	1
Cleft palate	2	Barrett's esophagus	1	mean corpuscular volume	1
alcohol dependence	2	C-reactive protein measurement	1	monocyte count	1
aspartate aminotransferase measurement	2	Keratoconus	1	red blood cell distribution width	1

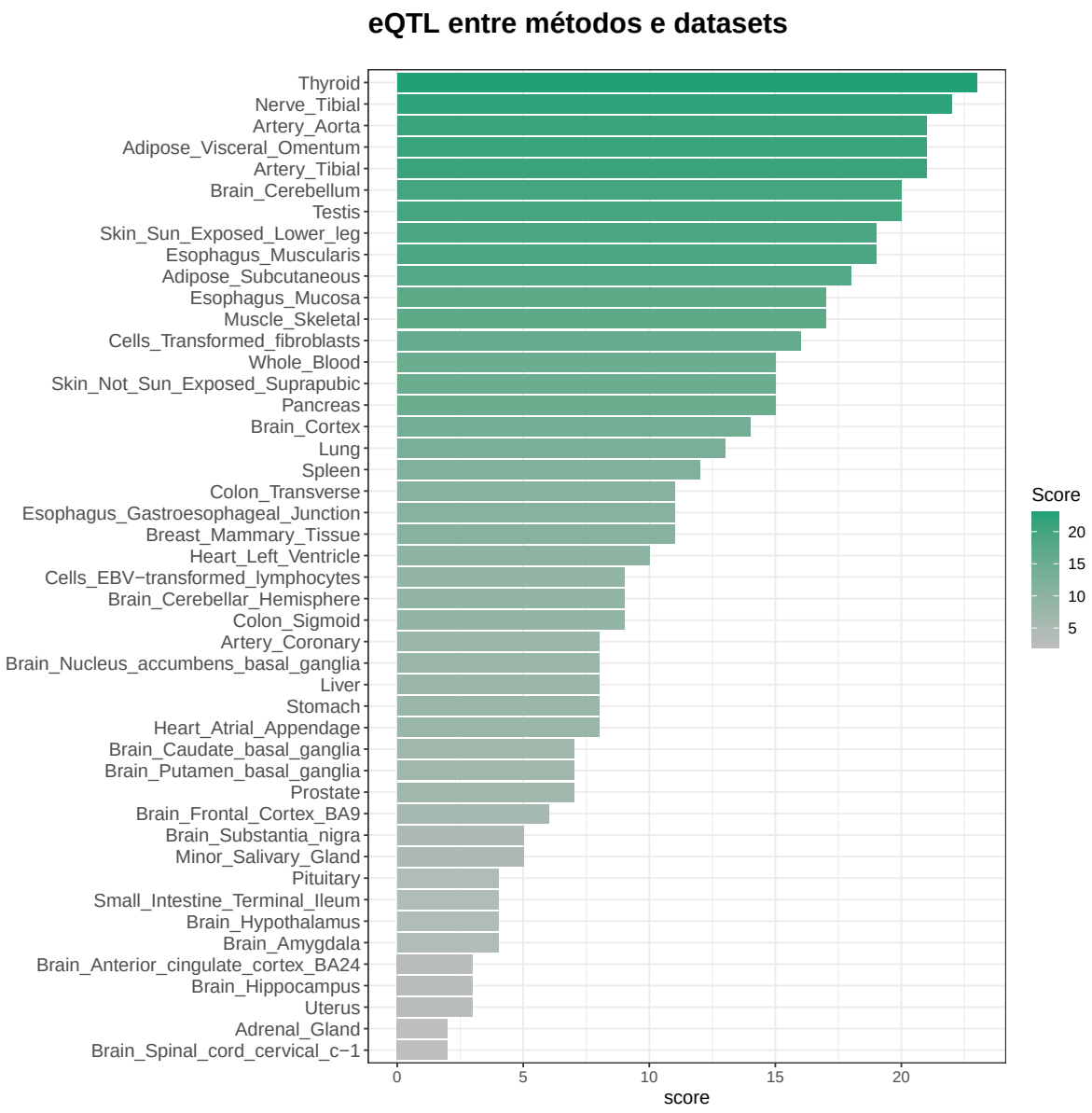

Figura 2.13 – Valores totais de eQTL para os alelos candidatos. Nesta abordagem, compilamos os resultados de eQTL para os métodos PBS e XP-EHH em ambos os datasets, e atribuímos um valor (*score*) para cada tecido somando o total de alelos candidatos que contribuem de forma significativa para a expressão de seu respectivo gene (positiva ou negativamente).

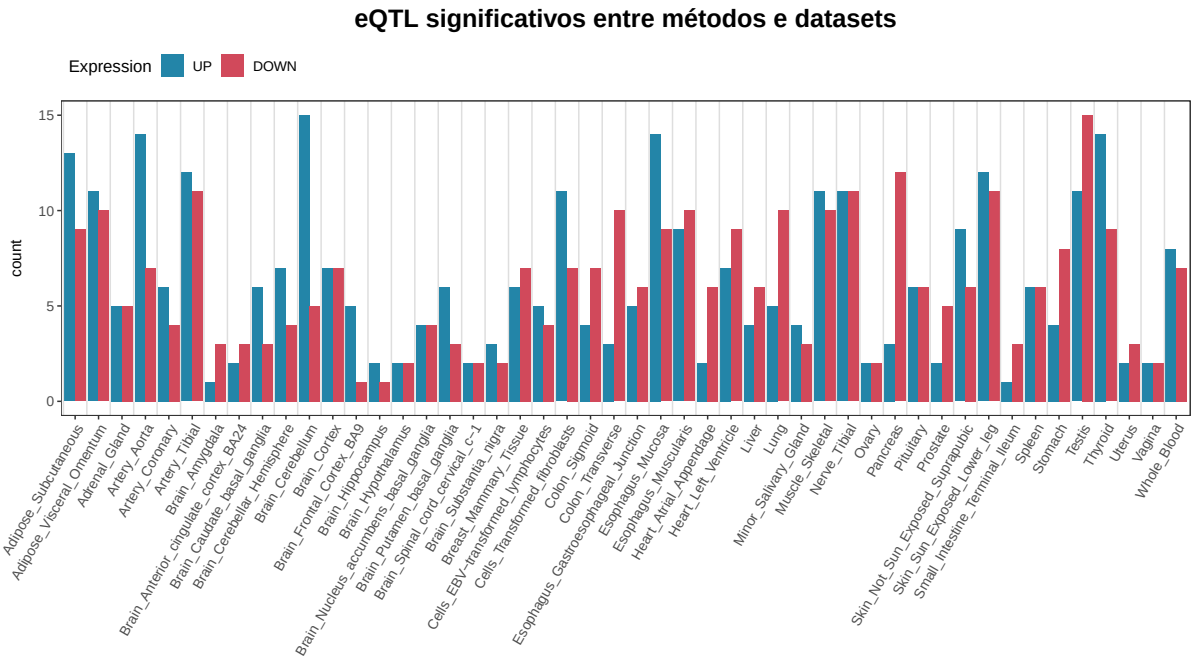

Figura 2.14 – Valores diferenciais de eQTL para os alelos candidatos. Nesta abordagem, compilamos os resultados de eQTL para os métodos PBS e XP-EHH em ambos os datasets e somamos o total de alelos candidatos que contribuem de forma significativa para a expressão de seu respectivo gene, agrupando por categoria de expressão.

2.3.4 Enriquecimento Gênico

Outra maneira de realizar o levantamento de fenótipos dado uma lista de genes são as análises de super-representação (ORA) e enriquecimento gênico (GSEA). Aplicamos estas análises nos bancos de dados públicos GWAS Catalog, KEGG e Gene Ontology. Primeiramente, utilizamos a plataforma Enrichr para análise ORA nos bancos GWAS Catalog e KEGG. Posteriormente, uma vez que esta plataforma não possibilita o uso de uma segunda lista de genes que sirvam de parâmetro (por padrão utiliza todos os genes humanos do banco de dados Ensembl), utilizamos duas outras ferramentas – FUMAGWAS e WebGestalt – que possibilitam o uso dessa lista para os bancos GWAS Catalog e KEGG, respectivamente. Adicionalmente, também rodamos a análise GSEA no banco de dados KEGG (WebGestalt).

Focando primeiramente no banco de dados do GWAS Catalog, no dataset S1, o método PBS resultou nos fenótipos "consumo de álcool", "consumo de café" e "traços relacionados à obesidade" na abordagem por SNP quando considerado os SNPs do extremo 1% da distribuição na plataforma Enrichr. Na plataforma FUMAGWAS, estes mesmos fenótipos foram detectados em conjunto com "consumo de café" e "traços relacionados à obesidade". Este último também foi detectado no top 0,5% da distribuição, em conjunto com "tolerância geral ao risco" (FDR < 0.038). No dataset S2, em ambas as plataformas, os termos "consumo de café" e "consumo de álcool", de forma literal ou derivada, foram

identificados em ambas as plataformas. Demais fenótipos identificados pelo método PBS no dataset S2 podem ser verificados na tabela 2.12. Na abordagem por gene, o método PBS resultou nos mesmos fenótipos que na abordagem por SNP quando utilizada a plataforma Enrichr no dataset S1, enquanto o dataset S2 apresentou os mesmos termos em ambas as plataformas, mesmo quando selecionado o extremo 0,1% dos genes (Tabela 2.13). Adicionalmente, quando consideramos o top 0,5%, detectamos outros fenótipos como "mensuração de imunoglobulina", "contagem de linfócitos", "contagem de glóbulos brancos", todos relacionados com o sistema imune, além de traços relacionados ao metabolismo energético (dados não mostrados).

Tabela 2.12 – Fenótipos enriquecidos no banco GWASCatalog (FUMAGWAS), utilizando a abordagem por SNP no extremo 0,1% da distribuição de PBS do dataset S2.

Fenótipo	-$\log_{10}(p)$	FDR
Response to alcohol consumption (flushing response)	8.9030	0.0000
Alcohol consumption (max-drinks)	8.3391	0.0000
Systolic blood pressure x alcohol consumption interaction (2df test)	7.2454	0.0000
Mean arterial pressure x alcohol consumption interaction (2df test)	5.9580	0.0005
Alcohol dependence symptom count	5.7618	0.0006
Diastolic blood pressure x alcohol consumption interaction (2df test)	5.6333	0.0007
Coffee consumption	5.0736	0.0022
Mean arterial pressure	4.6972	0.0046
Alcohol consumption (drinkers vs non-drinkers)	4.5856	0.0050
Aspartate aminotransferase levels	4.5615	0.0050
Coronary heart disease	3.9928	0.0168
Esophageal cancer	3.7438	0.0273

No método XP-EHH, o fenótipo sobressalente no dataset S1 foi "consumo de café", em ambas as plataformas e nos três valores de corte de distribuição aplicados. Neste método, como no anterior, a plataforma FUMAGWAS apresentou mais termos significativamente enriquecidos quando comparado à plataforma Enrichr, mostrando uma importante influência da população gênica selecionada na análise. Ao considerar valores de corte mais liberais, como 0,5% e 1%, detectamos traços como "tecido adiposo subcutâneo", "mortalidade por falha cardíaca", "tolerância ao risco geral" e outros relacionados com o sistema imunológico, destacando-se doenças inflamatórias (*e.g.* alergias, psoríase), além dos traços de consumo de café, álcool e cigarro. No dataset S2, os termos mais salientes foram termos relacionados a pressão sanguínea e contagem de leucótios, em conjunto com traços relacionaos ao consumo de café e álcool. De forma similar ao dataset S1, mais termos significativos foram identificados pela plataforma FUMAGWAS. Na análise por gene, identifica-se termos semelhantes ou idênticos àqueles identificados na abordagem por SNP.

Tabela 2.13 – Fenótipos enriquecidos no banco GWASCatalog (FUMAGWAS), utilizando a abordagem por gene no extremo 0,1% da distribuição de PBS do dataset S2.

Fenótipo	-$\log_{10}(p)$	FDR
Response to alcohol consumption (flushing response)	10.3086	0.0000
Alcohol consumption (max-drinks)	9.7437	0.0000
Alcohol dependence symptom count	9.5183	0.0000
Mean arterial pressure x alcohol consumption interaction	7.3521	0.0000
Esophageal cancer	7.0309	0.0000
Diastolic blood pressure x alcohol consumption interaction	7.0245	0.0000
Systolic blood pressure x alcohol consumption interaction	6.8588	0.0000
Coffee consumption	6.4583	0.0001
Crohn's disease	6.1587	0.0001
Diastolic blood pressure x alcohol consumption (light vs heavy)	6.0828	0.0001
Alcohol consumption (drinkers vs non-drinkers)	5.2424	0.0009
Hypothyroidism	5.1186	0.0012
Inflammatory bowel disease	4.4465	0.0050
Serum uric acid levels	4.3010	0.0065
Mean arterial pressure	4.2357	0.0070
Alcohol consumption	4.0574	0.0099
C-reactive protein levels or LDL-cholesterol levels (pleiotropy)	3.9443	0.0121
Hematological and biochemical traits	3.6485	0.0227
Aspartate aminotransferase levels	3.4485	0.0340
Metabolic syndrome	3.3372	0.0417
Body mass index	3.2689	0.0465

No método iHS, cinco termos enriquecidos apareceram quando considerado o extremo 0,1% da distribuição pela plataforma FUMAGWAS, dataset S1 (Tabela 2.14) e, destes, três – "níveis de urato em indivíduos obesos", "hipertrofia cardíaca" e "concentração de tiroxina livre" – apareceram também na plataforma Enrichr. Ao selecionar os valores menos restritivos da distribuição, observa-se todos os termos identificados pelos métodos anteriores. Na abordagem por gene, apenas "consumo de café" está em convergência entre as plataformas, com demais termos possivelmente relacionados (*e.g.* consumo de álcool e cigarro) presentes na plataforma Enrichr. No dataset S2, o termo "nascimento prematuro espontâneo moderado-tardio", encontrado no dataset S1, foi detectado no extremo 0,5% da distribuição, em conjunto com uma variedade de fatores. Para o método iHS, observa-se múltiplos termos significativos quando consideramos os top 0,5% ou 1% da distribuição por SNP. Dentre os fenótipos identificados, incluem-se diabetes tipo II, metabolismo, densidade mineral óssea, altura, pressão sistólica e diastólica, respota viral (*e.g.* HIV, Epstein-Barr), contagem de leucócitos, e traços relacionados à picada de mosquito (também identificados no dataset S1).

Tabela 2.14 – Fenótipos enriquecidos no banco GWASCatalog (FUMAGWAS) no extremo 0,1% da distribuição de iHS na abordagem por SNP, dataset S1.

Fenótipo	-$\log_{10}(p)$	FDR
Urate levels in obese individuals	4.6378	0.0245
Renal cell carcinoma	4.5014	0.0245
Moderate-to-late spontaneous preterm birth	4.3584	0.0245
Cardiac hypertrophy	4.2299	0.0245
Free thyroxine concentration	4.1703	0.0245

Na análise ORA realizada com o banco KEGG Pathway pelo WebGestalt, nenhuma via significativa apareceu em ambos os datasets, em nenhum dos métodos. A mesma análise realizada na plataforma Enrichr resultou apenas no termo "síntese e secreção de aldosterona" no dataset S1, quando considerado o top 1% da distribuição de iHS na abordagem por SNP. No dataset S2, este mesmo termo apareceu no método PBS, também utilizando 1% como valor de corte da distribuição (tabela 2.15). Dentre as vias de interesse identificadas nestas distribuições, destacam-se: vício à nicotina e morfina; sinapses glutamatérgicas e colinérgicas; arrastamento circadiano; sinalização adrenérgica em cardiomiócitos; síntese, secreção e ação do hormônio da paratireóide; síntese e secreção de cortisol e via de sinalização do estrogênio, discutidas no tópico 2.4. Adicionalmente, a abordagem por gene, pelo método XP-EHH, identificou três vias no dataset S2, sendo elas "via de sinalização IL-17", que atua na mediação de respostas inflamatórias, "via de sinalização de receptores *NOD-like*", que atua em resposta à patógenos, e "lúpus eritematoso sistêmico". A análise de GSEA, por sua vez, que considera todos os genes e seus respectivos valores, resultou em vias relacionadas ao metabolismo em ambos os datasets, nos métodos XP-EHH e iHS, respectivamente (Tabela 2.16).

Por fim, utilizamos o banco de dados da Gene Ontology, focando exclusivamente nos termos pertencentes a processos biológicos. Realizamos uma análise ampla abrangendo todos os termos disponíveis, bem como reduzindo a quantidade de termos a fim de reduzir redundância de processos similares (utilizando termos GO slim, vide material e métodos). Adicionalmente, dado a alta virulência dos ambientes de floresta tropicais, verificamos especificamente os termos GO pertencentes ao sistema imunológico (GO:0002376).

Nas análises utilizando todos os termos GO, na abordagem por SNP, poucos ou nenhum termos surgiram no dataset S1 quando utilizando os extremos 0,1% e 0,5% dos três métodos aplicados. Contudo, utilizando os top 1% da distribuição, encontramos termos relacionados ao crescimento no método XP-EHH e múltiplos termos para o método iHS, incluindo vias cardiovasculares (Tabela 2.17). No dataset S2, o extremo 0,5% das distribuições de iHS e XP-EHH já possibilitou identificar termos significativos, sobressaindo termos relacionados ao sistema imunológico pelo método XP-EHH (Tabela 2.18). Na abordagem por gene, os termos "regeneração do tecido muscular esquelético" e "metilação

Tabela 2.15 – ORA utilizando o banco de dados KEGG - Dataset S2.

Termo	iHS					
	Top 0,5%			Top 1%		
	Score	$-log_{10}(p)$	FDR	Score	$-log_{10}(p)$	FDR
Nicotine addiction	59.9289	3.77	0.0224	27.3208	2.8188	0.0151
Circadian entrainment	36.5269	3.7954	0.0224	54.9613	5.7767	1e-04
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	37.7809	3.5226	0.0264	93.5812	7.185	0
Glutamatergic synapse	25.5608	3.1838	0.0346	60.5433	6.4693	0
Cholinergic synapse	26.6032	3.249	0.0346	73.4442	7.2541	0
Oxytocin signaling pathway	-	-	-	58.3875	6.9434	0
Adrenergic signaling in cardiomyocytes	-	-	-	51.2635	6.2798	0
Morphine addiction	-	-	-	53.6082	5.5698	1e-04
GABAergic synapse	-	-	-	46.7865	5.0808	3e-04
Dilated cardiomyopathy (DCM)	-	-	-	44.3123	4.9463	4e-04
Parathyroid hormone synthesis, secretion and action	-	-	-	36.7548	4.6071	7e-04
Aldosterone synthesis and secretion	-	-	-	36.8837	4.5082	8e-04
Calcium signaling pathway	-	-	-	25.4194	4.2394	0.0014
Hypertrophic cardiomyopathy (HCM)	-	-	-	34.5165	4.1371	0.0016
Long-term depression	-	-	-	38.4978	3.9611	0.0023
Relaxin signaling pathway	-	-	-	26.0753	3.907	0.0024
cAMP signaling pathway	-	-	-	20.7287	3.7944	0.0029
Cortisol synthesis and secretion	-	-	-	31.8089	3.6148	0.004
Estrogen signaling pathway	-	-	-	22.6164	3.6038	0.004
Insulin secretion	-	-	-	26.6092	3.5183	0.0046
Serotonergic synapse	-	-	-	20.72	3.228	0.0086
Taste transduction	-	-	-	22.6849	3.1429	0.01
Gap junction	-	-	-	19.5286	2.8997	0.0141
Long-term potentiation	-	-	-	22.2306	2.915	0.0141
Vascular smooth muscle contraction	-	-	-	16.7855	2.9181	0.0141
Cushing syndrome	-	-	-	15.9446	2.931	0.0141
Dopaminergic synapse	-	-	-	17.1484	2.9551	0.0141
Rap1 signaling pathway	-	-	-	14.0141	2.8604	0.0148
Inflammatory mediator regulation of TRP channels	-	-	-	18.0156	2.8474	0.0148
Salivary secretion	-	-	-	18.4205	2.8084	0.0151
Retrograde endocannabinoid signaling	-	-	-	14.9165	2.7683	0.016
GnRH signaling pathway	-	-	-	16.8997	2.6773	0.0191
Thyroid hormone synthesis	-	-	-	17.3365	2.5584	0.0244
Gastric acid secretion	-	-	-	16.7532	2.5117	0.0263
Cell adhesion molecules (CAMs)	-	-	-	12.7955	2.4777	0.0272
Pancreatic secretion	-	-	-	14.6909	2.4735	0.0272
HIF-1 signaling pathway	-	-	-	13.906	2.3967	0.0315
Apelin signaling pathway	-	-	-	12.0653	2.3478	0.0344
cGMP-PKG signaling pathway	-	-	-	10.5422	2.2295	0.044
Axon guidance	-	-	-	9.9084	2.1727	0.0489

Tabela 2.16 – GSEA utilizando o banco de dados KEGG e resultados de XP-EHH no dataset S2.

Via de Sinalização	Enriquecimento	p-valor	FDR
Ascorbate and aldarate metabolism	2.3517	0.0000	0.0000
Pentose and glucuronate interconversions	2.0084	0.0021	0.0087
Porphyrin and chlorophyll metabolism	1.9648	0.0000	0.0102
Retinol metabolism	1.8890	0.0000	0.0195
Vitamin B6 metabolism	1.7817	0.0020	0.0386
Steroid hormone biosynthesis	1.7879	0.0022	0.0444

ou desmetilação de DNA" foram identificados no dataset S1 (método PBS), e apenas o método XP-EHH retornou termos significativos no dataset S2, sobressaindo termos do sistema imune (Tabela 2.18). Na análise de termos GO não-redundantes, na abordagem por SNP, em ambos os datasets aparecerem os termos "crescimento", "locomoção" e "processo do sistema nervoso", enquanto que no dataset S2 apareceu também, dentre termos relacionados a processos celulares, o termo "processo do sistema circulatório". Adicionalmente, na abordagem por gene apareceu "desenvolvimento embrionário" no dataset S1, e os termos "metabolismo de carboidratos" e "resposta ao estresse" no dataset S2.

Tabela 2.17 – Termos enriquecidos do banco Gene Ontology no dataset S1, abordagem por SNP.

Método	Termo	-$\log_{10}(p)$	FDR
XP-EHH	developmental growth	6.002302	0.002
	growth	5.923484	0.002
	multicellular organismal process	4.799597	0.040
	regulation of localization	4.758747	0.040
iHS	regulation of blood circulation	6.291592	0.000
	cardiac conduction	5.964111	0.000
	inorganic cation transmembrane transport	5.887296	0.000
	regulation of heart contraction	5.851375	0.000
	cation transmembrane transport	5.796583	0.000
	regulation of system process	5.696401	0.000
	modulation of chemical synaptic transmission	5.616438	0.002
	regulation of trans-synaptic signaling	5.592960	0.002
	sensory perception	5.474760	0.008
	regulation of biological quality	5.431910	0.010
	regulation of localization	5.336457	0.012
	multicellular organismal signaling	5.272507	0.018
	system process	5.222667	0.018
	ion transmembrane transport	5.203898	0.018
	inorganic ion transmembrane transport	5.032140	0.020
	cation transport	4.976156	0.026
	multicellular organismal process	4.936052	0.028

Ao analisar especificamente termos GO do sistema imune, exceto pela "ativação da célula T envolvida na resposta imune" pelo método iHS na abordagem por SNP (p-valor = 0,000131; FDR = 0,04), nenhum outro termo foi detectado como significativo após correção dos p-valores. Contudo, no dataset S2, múltiplos termos aparecem para os métodos PBS e XP-EHH na abordagem por SNP (Tabela 2.19), bem como no método XP-EHH na abordagem por gene 2.20) (mesmos termos identificados na abordagem por SNP).

Tabela 2.18 – Termos enriquecidos do banco Gene Ontology no dataset S2, abordagem por gene.

Método	Termo	-log10(p)	FDR
XP-EHH (top 0,5%)	cellular response to interferon-gamma	6.802781	0.000
	response to interferon-gamma	6.116870	0.004
	defense response to protozoan	5.332370	0.012
	response to protozoan	5.332370	0.012
XP-EHH (top 1%)	protein-DNA complex assembly	6.110807	0.002
	cytokine-mediated signaling pathway	5.758698	0.004
	regulation of multicellular organismal process	5.705118	0.004
	protein-DNA complex subunit organization	5.286646	0.016
	DNA replication-dependent nucleosome assembly	5.073446	0.026
	DNA replication-dependent nucleosome organization	5.073446	0.026
	rDNA heterochromatin assembly	5.000396	0.028
	nucleosome assembly	4.970933	0.030
	regulation of immune system process	4.850650	0.038
	cellular response to interferon-gamma	4.792833	0.038

Tabela 2.19 – Termos enriquecidos do banco Gene Ontology no dataset S2, abordagem por SNP.

Método	Term	$-\log_{10}(p)$	FDR
PBS	positive thymic T cell selection	4.802562	0.022
	negative thymic T cell selection	4.802562	0.022
	negative T cell selection	4.679098	0.034
XP-EHH	response to biotic stimulus	5.748423	0.006
	response to external biotic stimulus	5.477667	0.012
	negative regulation of cell communication	5.368043	0.014
	negative regulation of signaling	5.356821	0.016

Em conjunto, as análises de enriquecimento gênico revelaram fenótipos relacionados ao desenvolvimento corporal (*e.g.* crescimento, massa corporal, obesidade, nascimento prematuro), vias cardíacas (e.g. pressão sistólica, diastólica e arterial, hipertrofia cardíaca, doença coronariana, sinalização adrenérgica em cardiomiócitos), a diabetes (*e.g.* diabetes tipo II, traços glicêmicos e secreção de insulina), à tireoide (*e.g.* síntese do hormônio tireoidiano, diferenciação das células Th1 e Th2, hipotireoidismo, concentração de tiroxina livre), ao metabolismo, sistema imunológico, sobretudo na quantidade, resposta e diferenciação de leucócitos (especialmente linfócito T), bem como comportamentos de vício, como alcoolismo e consumo de café.

Tabela 2.20 – Termos GO do sistema imune identificados na abordagem por gene, dataset S2.

Termo	-$\log_{10}(p)$	FDR
PBS		
positive thymic T cell selection	2.817464	0.050
negative thymic T cell selection	2.817464	0.050
XP-EHH		
granulocyte migration	4.661809	0.000
granulocyte chemotaxis	4.077900	0.006
myeloid leukocyte migration	3.924738	0.010
neutrophil migration	3.864530	0.012
leukocyte chemotaxis	3.722748	0.012
neutrophil chemotaxis	3.193541	0.040

2.3.5 Convergência Evolutiva (Metanálise)

A fim de comparar nossos resultados com aqueles já publicados na literatura sobre seleção em populações nativas de Floresta Tropical, comparamos os resultados de 13 artigos de nosso conhecimento com os nossos resultados. A tabela 2.21 sumariza as populações e métodos utilizados para detecção de seleção positiva nos artigos avaliados.

Tabela 2.21 – Estudos avaliados na metanálise para convergência evolutiva.

Autores	Populações	Métodos de seleção
Amorim et al. (2015)	Biaka, Mbuti, Surui, Karitiana	BayeScan
Bergey et al. (2018)	Batwa, Andamanese	PBS, PBSi, Bayenv
Harrison et al. (2019)	Batwa	PBS, iHS
Hsieh et al. (2016)	Biaka, Baka	iHS, G2D
Jarvis et al. (2012)	Bakola, Baka, Bedzan	LSBL, XP-EHH, iHS
Lachance et al. (2012)	Western Pygmy,l Hadza, Sandawe	LSBL, AIMs
López Herráez et al. (2009)	Pygmy, Surui, Karitiana	lnRsb
Miligliano et al. (2013)	Biaka, Mbuti, Aeta, Batak, Agta + 2	XP-EHH, iHS
Perry et al. (2014)	Baka, Batwa	Fst, iHS, BayeScan
Scheinfeldt et al. (2019)	Hadza, Sandawe, Dahalo, Sabue + 6 outras	D, iHS, XP-CLR

Para cada artigo, consideramos a intersecção dos genes candidatos que constam nos artigos com a união dos genes candidatos levantados na presente pesquisa, pelos três métodos de seleção (*i.e.* PBS, XP-EHH e iHS), em ambos os datasets. De forma similar

ao item 2.3.3, realizamos a anotação funcional dos genes de intersecção, utilizando tanto os bancos Ensembl como GWAS Catalog. A Tabela 2.22 apresenta os resultados dos trinta primeiros fenótipos com mais genes candidatos, levantados pelo banco de dados Ensembl, com interseção entre os métodos e estudos avaliados. As colunas "Genes" e "Estudos" denotam quantos únicos genes e estudos foram mapeados para um dado fenótipo, respectivamente, enquanto as demais colunas denotam quantos genes mapeados para o respectivo fenótipo apareceram entre os métodos e datasets, podendo (e com provável) repetição de gene-fenótipo entre métodos e datasets (*e.g.* um fenótipo cujo gene apareceu em mais de um método). Para evitar redundância de resultados, não incluímos os resultados do banco GWAS Catalog, visto que os resultados foram similares (dados não mostrados).

Tabela 2.22 – Fenótipos mapeados aos genes candidatos com sinais de convergência evolutiva em populações caçadoras-coletoras de florestas tropicais (banco de dados Ensembl).

			Dataset S1			Dataset S2		
Fenótipo	Genes	Estudos	PBS	XP-EHH	iHS	PBS	XP-EHH	iHS
Body Mass Index	35	7	5	4	11	4	2	15
Blood pressure	30	9	4	3	8	1	3	16
Cholesterol, HDL	24	7	3	2	7	4	0	14
Metabolite levels	23	8	5	3	6	2	1	12
Body Height	22	7	3	3	5	1	3	9
Blood protein levels	21	8	2	4	5	1	1	7
Height	21	8	2	1	4	1	1	16
Triglycerides	20	8	1	1	2	0	1	14
Asthma	19	7	1	1	4	2	4	8
Cholesterol, LDL	19	7	1	2	9	1	2	10
Refractive error	19	7	2	4	5	1	0	12
Myocardial infarction	18	7	3	2	5	1	1	7
Autism spectrum disorder or schizophrenia	17	6	1	5	2	1	2	8
Body Weight	17	7	1	2	7	2	0	9
Cholesterol	17	7	1	3	5	0	2	11
Echocardiography	17	8	2	0	7	2	0	9
Stroke	17	8	1	2	3	0	3	13
Heart Rate	16	6	2	3	3	0	1	9
Iron	16	7	1	2	0	0	2	13
Heel bone mineral density	15	6	2	3	2	2	0	7
Obesity-related traits	15	7	2	1	3	0	2	11
Pulse pressure	14	6	1	1	3	0	0	10
Chronotype	13	6	1	0	3	1	0	9
Erythrocyte Count	13	6	3	3	8	1	0	4
Crohn's disease	12	7	0	3	4	0	4	4
Estimated glomerular filtration rate	12	6	1	2	0	3	1	4
Respiratory Function Tests	12	7	1	2	5	1	0	9
Coronary Artery Disease	11	6	0	3	3	1	0	7
Hemoglobin S	11	5	1	2	4	1	1	5
Lipids	11	6	1	2	4	1	0	6

2.3.6 Estrutura de Arquivos no Repositório Mendeley

Devido à estrutura metodológica deste trabalho, utilizando dois conjunto de dados, três testes de neutralidade (PBS, XP-EHH, iHS), duas abordagens (por SNP e por gene), quatro plataformas de enriquecimento gênico (Enrichr, FUMAGWAS, WebGestalt, GOATOOLS), análise em três banco de dados públicos (GWASCatalog, KEGG e GO), sendo este último banco dividido em termos não redundantes ou exclusivos do sistema imune, bem como anotação funcional e metanálise, uma variedade de gráficos e tabelas resultaram destas análises, tanto separadamente quanto em conjunto (intersecção dos resultados). Ainda, nas análises de enriquecimento gênico, aplicamos diferentes valores de corte para identificar genes candidatos (e.g. top 0,5% e 1%), uma vez que, utilizando o padrão nas análises de varredura genômica (*i.e.* top 0,1%), não encontramos termos significativos dada a quantidade insuficiente de genes para este tipo de análise. Assim, exceto as análises de estrutura populacional, todos os gráficos e tabelas das demais análises foram depositadas no banco de dados da Mendeley (<http://dx.doi.org/10.17632/gztff7wmjt.1>) com dois objetivos principais: 1) fornecer figuras em alta resolução nos formatos .png e .pdf e 2) promover a transparência da pesquisa científica.

As figuras e tabelas estão alocadas nas pastas figures e tables, respectivamente. A convenção de nomenclatura dos arquivos é *dataset* (ds1 ou ds2) + *método* + *modo*+ *descrição* (*e.g.* ds1_pbsw_persnp.pdf). O modo corresponde à abordagem (por SNP ou por gene). Caso não tenha informação sobre o dataset no início, pode-se supor que se trata dos dois em conjunto. A descrição pode ou não estar presente. Por fim, para verificar qual script gerou uma figura/tabela, pode-se procurar pelo nome da figura/table (sem extensão) no arquivo Snakefile do repositório da tese.

2.4 DISCUSSÃO

2.4.1 Varreduras Adaptativas e Enriquecimento Gênico

Em conjunto, os resultados das varreduras adaptativas e de enriquecimento gênico identificaram vias relacionadas à estrutura corporal, sistema cardiovascular, sistema imune, comportamentos de risco, bem como metabolismo, sobressaindo-se fenótipos ligados à diabetes e tireoide. Cada um destes fenótipos será discutido separadamente a seguir, após uma breve análise da distribuição dos resultados.

Pode-se observar que, nas análises de varredura adaptativa, ambas as abordagens – por SNP e por gene – identificaram genes candidatos com valores extremos de seleção em diferentes cromossomos (Figuras 2.4-2.5, 2.7-2.8). Contudo, na abordagem por SNP (Figuras 2.4A, 2.5A, 2.7A e 2.8A) se identifica mais cromossomos com genes candidatos quando comparado à abordagem por gene (Figuras 2.4B, 2.5B, 2.7B e 2.8B), em ambos os datasets (dataset S1 e dataset S2) e nos três métodos aplicados (*i.e.* PBS, XP-EHH e iHS). De modo geral, mais genes candidatos foram identificados na abordagem por SNP do que na abordagem por gene (Figuras 2.6, 2.11 e 2.12).

Ao considerarmos os testes de seleção separadamente, observou-se menos genes candidatos identificados pelo método PBS quando comparado aos métodos XP-EHH e iHS (figuras 2.11 e 2.12). Esta diferença se explica principalmente pela distribuição mais dispersa dos valores de XP-EHH e iHS (figuras 2.9 e 2.10) em ambos os datasets. Contudo, cabe ressaltar que, para o método PBS, dois filtros adicionais foram aplicados: 1) média móvel por janelas e 2) remoção das janelas consecutivas após adquirir o subconjunto dos resultados com valores acima de determinado limite (*e.g.* p-valor $\leq 0{,}01$). Tais filtros explicam a diferença de valores observados entre os métodos, visto que, sem a aplicação destes filtros, mais genes candidatos são identificados no método PBS (171 e 22 nas abordagem por SNP e por gene, respectivamente; dados não mostrados). De fato, uma abordagem comum também aos métodos XP-EHH e iHS consiste no cálculo de médias móveis por janelas (SZPIECH; HERNANDEZ, 2014; GAUTIER; KLASSMANN; VITALIS, 2017).

Adicionalmente, pode-se notar que há maior intersecção de genes candidatos entre os métodos XP-EHH e iHS do que estes com PBS (Figuras 2.11 e 2.12), o que é esperado uma vez que estas estatísticas, apesar de diferirem no foco da seleção (iHS é voltado para sweeps incompletos, e XP-EHH para sweeps quase ou completamente fixados) (SUZUKI, 2010), possuem como princípio em comum a análise da extensão de homozigose dos haplótipos, ao passo que PBS possui como princípio a variação da frequência alélica entre populações.

A maioria dos SNPs identificados nas análises de seleção são provenientes de regiões não-codificantes (Tabelas 2.5 e 2.6). Estes resultados são esperados, uma vez que,

de maneira similar, grande parte dos loci identificados em análises de GWAS são também provenientes de regiões não-codificantes (CROTEAU-CHONKA et al., 2015; HILL et al., 2019). De fato, observamos em nossa análise *in silico* no banco GTEx que a maioria SNPs candidatos, mesmo em regiões não-codificantes, contribuem de forma significativa para o aumento ou diminuição da expressão de seus respectivos genes nos tecidos (Figuras 2.13 e 2.14).

A anotação dos fenótipos mapeados aos genes candidatos levantaram fenótipos comuns entre os bancos de dados do Ensembl e do GWAS Catalog, em ambos os datasets (Tabelas 2.8-2.11). Entre estes, incluem-se fenótipos relacionados à 1) estrutura corporal, como altura e peso (*e.g.* "body height", "body mass index", "heel bone mineral density", "body fat distribution", "obesity-related traits"), 2) sistema cardiovascular (*e.g.* "systolic and diastolic blood pressure", "hypertension", "echocardiography", "myocardial infarction", "heart failure", "coronary artery disease"), 3) sistema imune (*e.g.* "*Trypanosoma cruzi* seropositivity", "white blood cell count", "leukocyte count", "lymphocyte count", "neutrophil count", "eosinophil count", "macrophage inflammatory protein 1b measurement"), 4) metabolismo (*e.g.* "metabolite levels", "metabolic syndrome", "cholesterol", "lipids", "glucose", "glucose homeostasis traits"), e 5) comportamentos de risco ("alcohol drinking", "drinking and smoking behavior", "alcohol dependence symptom count", "smoking status measurement", "smoking initiation", "caffeine consumption", "risk-taking behavior", "general risk tolerance (MTAG)", "adventurousness"), além de traços como "type II diabetes mellitus". Apesar destes fenótipos apresentarem um forte potencial de contribuição na discussão da saúde e história indígena (vide próximos itens da discussão), os mesmos são majoritariamente provenientes de estudos de associação realizados em outras populações, usualmente européias. Ademais, para este mapeamento não foi calculada significância estatística. Para isso, recorremos a algoritmos já implementados para enriquecimento gênico.

A análise de enriquecimento gênico utilizando o banco de dados GWAS Catalog resultou em fenótipos convergentes tanto pela plataforma Enrichr quanto FUMAGWAS, sobressaindo fenótipos relacionados ao comportamento de vício ou de risco. Pode-se observar que mais termos significativos foram identificados na plataforma FUMAGWAS, que utiliza uma lista personalizada de genes como *background* (Tabelas 2.13-2.14). A menor diversidade detectada pela plataforma Enrichr, contudo, pode ser compensada por uma análise programática extremamente rápida dos resultados via API GSEApy, e mesmo em sua versão online, apresenta uma outra vantagem ao retornar os resultados de enriquecimento para todos os 170 bancos de dados disponíveis na plataforma, não se restringindo ao banco GWAS Catalog, cabendo então ao usuário a escolha da plataforma.

Uma maior diversidade fenotípica foi identificada no dataset S2 quando comparado ao dataset S1, destacando-se fenótipos relacionados ao sistema imune, que apareceram a

partir da seleção dos genes candidatos dentro do extremo 0,5% da distribuição. Interessantemente, estes fenótipos relacionados ao sistema imunológico também apareceram com maior representatividade no dataset S2 nas análises do banco Gene Ontology (Tabelas 2.18-2.20).

Na análise de convergência evolutiva, observa-se que a maioria dos genes candidatos identificados entre populações caçadoras-coletoras de distintas regiões de florestas tropicais apresentam fenótipos relacionados à estrutura corporal, como altura e peso, bem como traços relacionados ao metabolismo, vias cardíacas e neurológicas. O fenótipo mais evidente, altura, se justifica pelo fato da maioria dos estudos utilizados na comparação estarem concentrados em populações de pigmeus africanos. Os demais fenótipos que apareceram, assim como nas análises de enriquecimento gênico, refletem aqueles levantados pela análise de anotação funcional dos genes candidatos.

Interessantemente, em um dos mais famosos livros sobre epidemiologia indígena brasileiro, Coimbra Jr et al. (2003) apontam que, apesar das doenças infecciosas possuírem um papel central no perfil epidemiológico dos indígenas no Brasil, tem-se ampliado o número de evidências de ocorrência de doenças crônicas não transmissíveis, como, por exemplo, alcoolismo, obesidade, hipertensão e diabetes mellitus (COIMBRA JR; SANTOS; ESCOBAR, 2003; GUIMARÃES; GRUBITS, 2007).

Dada a intersecção entre as vias identificadas e as crescentes evidências de fenótipos associados às mesmas nas populações indígenas brasileiras, hipotetizamos que genes outrora importantes para o processo adaptativo na floresta tropical amazônica, hoje podem estar associados com o perfil epidemiológico atual destas populações. A fim de discutir melhor esta hipótese, discorremos primeiramente sobre os resultados dos métodos neste trabalho, e então discorremos - por tópico - sobre cada uma das principais vias identificadas, sobre o porquê terem sido alvo de seleção, e seu reflexo nas populações atuais.

2.4.2 Estrutura Corporal

As discussões mais frequentes na literatura em relação à estrutura corporal, quando se trata de caçadores-coletores nativos de floresta tropical, focam na estatura, uma vez que tais populações são amplamente conhecidas por apresentarem baixa estatura, sendo denominadas como pigmeus (VERDU, 2016). Apesar do termo "pigmeu" ser originalmente aplicado para populações de caçadores-coletores africanas, o mesmo tem sido utilizado para se referir também a outras populações de floresta com baixa estatura (altura masculina média < 160 cm) provenientes do continente asiático e sul-americano (PERRY; DOMINY, 2009).

Acredita-se que a baixa estatura observada em diversas populações nativas de floresta tropical é resultante de um processo adaptativo a estes ambientes. As hipóteses que visam explicar este fenômeno argumentam que a baixa estatura proporciona maior

facilidade para mobilidade e termorregulação em ambientes de floresta densa, bem como redução dos requerimentos calóricos. Também argumenta-se que a baixa estatura surgiu como consequência da reprodução precoce, a fim de compensar a baixa expectativa de vida (MIGLIANO; VINICIUS; LAHR, 2007; PERRY; DOMINY, 2009; VERDU, 2016).

Há poucos estudos sobre a estatura indígena brasileira, e estes apontam para uma estatura média entre homens adultos que varia de 152 cm a 168,1 cm, como no caso das populações Yanomami e Xavante, respectivamente (COIMBRA JR; SANTOS, 2001; PERRY; DOMINY, 2009). Em estudo com populações de idosos indígenas de Roraima, abrangendo seis etnias distintas (Makuxí, Wapixána, Taurepáng, Ingarikó, Patamona e Waiwái), por exemplo, Coimbra Jr et al. (2003) identificaram 155 cm como estatura média masculina. Ferreira (2016), por sua vez, estudando o perfil indígena dos Mura de Autazes - Amazonas, verificou como estatura média 155,29 cm ($\pm$ 7,91) (homens e mulheres adultos, sem dados separando os sexos) (FERREIRA, 2016).

Apesar da estatura dos nativos Amazônicos ser aparentemente maior que a dos pigmeus africanos, ela é menor que a da população brasileira como um todo, cuja estatura média masculina adulta é de 174 cm. De fato, há populações indígenas brasileiras classificadas como pigmeus em artigo de revisão sobre os mesmos (PERRY; DOMINY, 2009). Logo, dado os diversos genes relacionados à altura que foram identificados neste trabalho, é também plausível que as hipóteses sobre seleção de baixa estatura em pigmeus africanos também se apliquem às populações nativas amazônicas.

Nas populações nativas amazônicas, no entanto, como mencionado anteriormente, a estatura pode variar consideravelmente entre os grupos. Considerando onze destas populações pelas quais temos dados de estatura (dados não apresentados), seis delas possuem estatura média (homem adulto) inferior a 160 cm (Apalai, Urubu-Kaapor, Ticuna, Yanomami e Baniwa), e as outras cinco variam entre 160 e 170 cm (Kaingang, Parakanã, Arara, Kayapó e Xavante). A média geral da estatura destas populações é de 159 cm, podendo ser, portanto, classificadas como pigmeus, dado que diversos autores consideram como pigmeus populações com estatura inferior a 160 cm (CAVALLI-SFORZA, 1986; PERRY; DOMINY, 2009; VERDU, 2016). De fato, os Yanomami já foram citados como pigmeus (PERRY; DOMINY, 2009). Logo, pode-se então discutir hipóteses evolutivas para o fenótipo de baixa estatura nestas populações, considerando a variação dada a especificidade de cada população indígena brasileira, reconhecendo que tais hipóteses podem e provavelmente não se aplicam para todas as populações.

Uma das hipóteses com destaque potencial para as populações indígenas brasileiras consiste na "life history" (história de vida), que prediz uma correlação indireta entre o tamanho do corpo e a expectatividade de vida, onde, a fim de maximizar a expectativa de vida, o corpo cessa precocemente o crescimento para promover o início também precoce da fase reprodutiva (WALKER et al., 2006; MIGLIANO; VINICIUS; LAHR, 2007).

Destacamos esta hipótese porque há inúmeras evidências que apontam um início precoce da fase reprodutiva em aldeias indígenas brasileiras (IGANSI; ZATTI, 2018; LIMA et al., 2018) e de alta mortalidade infantil indígena (COIMBRA JR; SANTOS; ESCOBAR, 2003). Verificou-se, por exemplo, que pelo 24% das adolescentes entre 11 e 15 anos da etnia Javaé engravidaram e foram atendidas pela prefeitura local (Tocantins), número provavelmente subestimado dado aos casos de gravidez e abortos por métodos provocados ou desconhecidos (SANTIAGO, 2017). Referente à taxa de mortalidade infantil, que se correlaciona diretamente com a expectativa de vida, observa-se, segundo censo demográfico de 2010, que a mortalidade de crianças indígenas com menos de um ano de idade é 60% maior que a mortalidade não-indígena (MARINHO et al., 2019).

Adicionalmente, como hipóteses alternativas, destaca-se a hipótese originalmente proposta por Cavalli-Sforza (1986), em que a menor estatura teria um papel vantajoso na termorregulação, naturalmente dificultada em ambientes de florestas tropicais (PERRY; DOMINY, 2009). Interessantemente, os maiores e-QTLS dos SNPs candidatos identificados neste trabalho foram para perna exposta ao sol e tireoide (Figuras 2.13 e 2.14), órgão amplamente conhecido por atuar na termorregulação (IWEN; OELKRUG; BRABANT, 2018), além de possuir relação com a estatura e fase reprodutiva (TARIM, 2011; SILVA; OCARINO; SERAKIDES, 2018). A baixa exposição à luz solar acarreta deficiência na síntese de vitamina D, que em consequência pode levar a diminuição na estatura e aumento da gordura corporal (KREMER et al., 2009; MISSAGGIA et al., 2020). Neste contexto, as duas rotas metabólicas indicadas em nossas análises como sob seleção natural podem estar atuando em sinergia para a manutenção da baixa estatura em nativos amazônicos.

2.4.3 Metabolismo Energético

Sabe-se que a caça, pesca e a coleta de frutos e raízes tiveram uma papel essencial na dieta indígena, ainda que haja também evidências de práticas agrícolas (SOARES, 2015). Cordain et al. (2000) utilizaram os dados do Atlas Etnográfico (MURDOCK, 1967) para avaliar a proporção de macronutrientes em diversas populações caçadoras-coletoras ao redor do mundo e chegaram à conclusão que estas populações consomem uma quantidade elevada de alimentos de origem animal (45 - 65% do total de energia), caracterizando assim maior consumo proteico e menor consumo de carboidratos. Nesta linha de raciocínio, Miller e Colagiuri propuseram uma hipótese denominada “conexão carnívora”, alegando que, no passado, dietas ricas em proteínas e pobre em carboidratos teriam atuado como pressão seletiva ao fenótipo de resistência à insulina, tornando-se prejudicial após a mudança para dietas ricas em carboidratos associadas à agricultura (MILLER; COLAGIURI, 1994; COLAGIURI; MILLER, 2002). Esta hipótese foi discutida e considerada plausível por um grupo de pesquisadores que trabalham especificamente com seleção em populações de floresta tropical há muitos anos (LUCA; PERRY; RIENZO, 2010).

Dentre as consequências da resistência à insulina promovida pela mudança nos mecanismos de subsistência, destaca-se a diabetes mellitus tipo 2. Cabe ressaltar também que, além dos fatores genéticos, fatores fisiopatológicos também contribuem para resistência à insulina, como atividade física e obesidade (KAHN; HULL; UTZSCHNEIDER, 2006; RAVEL et al., 2012). A obesidade, por sua vez, pode ser causada tanto por fatores fisiopatológicos quanto por influência genética (HASLAM; JAMES, 2005; SERAVALLE; GRASSI, 2017), de forma que observamos uma rede de interação entre dieta, metabolismo, obesidade, resistência à insulina, diabetes tipo 2 e consequências cardiovasculares (esta última discutida no próximo item).

Interessantemente, nossos resultados identificaram genes candidatos diretamente associados à síndrome metabólica (Tabela 2.13) e diabetes tipo II (Tabelas 2.8-2.11), além de variados termos indiretos associados às mesmas (*e.g.* níveis de metabólitos, traços glicêmicos, secreção de insulina, etc). A síndrome metabólica consiste em um grupo de doenças caracterizadas pela resistência à insulina que encontram-se intimamente associadas à obesidade, doenças cardiovasculares e diabetes (SAMSON; GARBER, 2014), fenótipos consistentemente identificados na presente pesquisa por distintos métodos, abordagens e plataformas de genotipagem.

Os estudos com populações indígenas brasileiras têm mostrado altos índices de obesidade (SOARES, 2015). Soares et al. (2015), por exemplo, avaliaram a prevalência da síndrome metabólica nos Xavantes e verificaram que 66% dos indivíduos nesta população sofrem de obesidade, diabetes e doença coronariana. Outros estudos já demonstraram altos índices de sobrepeso e obesidade tanto em Xavantes (LEITE et al., 2006; WELCH et al., 2009), quanto em outras populações indígenas brasileiras, como Parkatêjê (CAPELLI; KOIFMAN, 2001), Guarani-Mbyá (CARDOSO; MATTOS; KOIFMAN, 2001), Suruí (LOURENÇO et al., 2008) e Suyá (SALVO et al., 2009).

Dada a convergência dos genes candidatos mapeados aos fenótipos de massa corporal, incluindo obesidade, diabetes e doença coronariana, e a alta prevalência de obesidade em populações indígenas brasileiras, acreditamos que esta prevalência, além de ser promovida por mudanças sócio-culturais (*e.g.* dieta e menor taxa de exercícios físicos), pode também ser explicada do ponto de vista genético, onde genes responsáveis por um melhor aproveitamento do teor glicêmico em dietas pobres em carboidrato teriam sido alvo de seleção, tornando-se prejudiciais após inclusão de alimentos industrializados e redução da atividade física.

2.4.4 Vias Cardiovasculares

Foram encontrados indícios de enriquecimento para vias cardiovasculares (*e.g.* pressão sistólica e diastólica, hipertrofia cardíaca, entre outros) nos três bancos de dados analisados: GWAS Catalog, KEGG e GO. Termos similares, relacionados ao desenvolvi-

mento do coração, já foram identificados em outro trabalho com populações nativas de floresta tropical da África (Batwa) e da Ásia (Andamanese), utilizando o banco de dados GO (BERGEY et al., 2018). Neste trabalho, os autores argumentam que a evolução da resposta ao hormônio de crescimento, que sabidamente desempenha um papel para baixa estatura destas populações, pode ter resultado em uma forte pressão seletiva para efeitos compensatórios nas vias cardíacas, uma vez que o hormônio de crescimento é importante para o desenvolvimento cardiovascular.

Conforme discutido no tópico anterior, a estatura média das populações indígenas brasileiras varia consideravelmente, e apenas algumas delas poderiam ser consideradas como pigmeus, cuja estatura média do homem adulto é inferior a 160 cm (CAVALLI-SFORZA, 1986). Embora a argumentação de Bergey et al. (2018) possa também ser aplicada aos nativos caçadores-coletores brasileiros, diferentemente dos pigmeus Africanos, não temos dados de seleção ao gene do hormônio de crescimento (GH) (BECKER et al., 2013). Desta forma, hipóteses alternativas são importantes para compreender os sinais de seleção em vias cardiovasculares em nativos americanos.

Nos ambientes de floresta tropical, é provável que populações caçadoras-coletoras tenham enfrentado altos níveis de estresse calórico e nutricional dada a sazonalidade dos recursos alimentícios (PERRY; DOMINY, 2009). De fato, por este motivo se questiona se realmente houve ocupação integral destes ambientes sem o auxílio direto ou indireto da agricultura (BAILEY et al., 1989). É possível, portanto, que o estresse causado pela escassez sazonal de alimento possa ter atuado como pressão seletiva para vias cardíacas, uma vez que a restrição calórica tem impacto direto tanto no metabolismo quanto na função cardíaca (HAN et al., 2004; ALBAKRI, 2019). Adicionalmente, outra explicação, também levantada por Bergey et al. (2018), relacionada ao sistema imunológico, é que a elevada exposição a protozoários e helmintos, comuns em ambientes de floresta tropical, possa ter atuado como pressão evolutiva, uma vez que estes microrganismos impactam direta ou indiretamente a saúde cardíaca (HIDRON et al., 2010).

2.4.5 Resposta Imunológica a Doenças Infecciosas

A população indígena brasileira tem passado por uma transição de perfil epidemiológico, apresentando uma diminuição de doenças infecciosas e parasitárias, em contraste com um aumento de morbidade associado a doenças crônicas como diabetes, hipertensão e obesidade (COIMBRA JR; SANTOS, 2001). Coimbra Jr et al. (2003) argumentam que esta transição está relacionada à alteração do modo de vida indígena, como, por exemplo, diminuição da atividade física e inserção de comportamentos de vício como o consumo de bebidas alcoólicas e o fumo (COIMBRA JR; SANTOS; ESCOBAR, 2003). De fato, têm surgido diversas evidências de um aumento elevado de casos de obesidade e hipertensão (FILHO et al., 2015), bem como alcoolismo (Ministério da Saúde, 2001).

Apesar da diminuição das doenças infecciosas e parasitárias, as populações indígenas permanecem marcadas por uma prevalência acentuada destas doenças (CARVALHO; OLIVEIRA; GUIMARÃES, 2014). Sabe-se que as regiões tropicais, como a floresta amazônica, possuem maior diversidade patogênica para humanos do que regiões temperadas (GUERNIER; HOCHBERG; GUÉGAN, 2004). Deste modo, pode-se esperar que algum mecanismo de defesa tenha sido selecionado proveniente de uma adaptação local à floresta amazônica, que se apresenta como um ambiente hostil à subsistência humana.

Nossos resultados apontam para uma pressão seletiva no mecanismo de defesa imunológica, haja visto a presença de termos como "contagem de leucócitos", "contagem de linfócitos" e "contagem de eosinófilos" mapeados aos genes candidatos em ambos os datasets, ao passo que múltiplos termos relacionados aos linfócitos T foram identificados no dataset S2, em conjunto com migração e quimiotaxia de leucócitos, via de sinalização mediada por citocinas e resposta humoral antimicrobiana. Ainda neste dataset, foram identificados outros termos como resposta ao estímulo biótico, resposta e defesa a organismos externos e resposta de defesa a protozoários.

Apesar da análise de enriquecimento gênico ter apresentado termos enriquecidos apenas no dataset S2, os genes candidatos mais diferenciados no dataset S1 também estão associados com fenótipos de contagem de linfócitos, como *ACACA*, *AUTS2*, *FMNL2* e *SAMD12* pelo método PBS, bem como *ANXA6*, *EXPH5* e *SLC7A10* método XP-EHH, e seis outros no método iHS, incluindo *FMNL2*. Cabe ressaltar que outros artigos de seleção em populações caçadoras-coletoras também identificaram genes candidatos com os valores mais extremos envolvidos na produção e diferenciação de células T (JARVIS et al., 2012; SCHEINFELDT et al., 2012). Adicionalmente, assim como no presente trabalho, Scheinfeldt et al. (2012) identificaram vias enriquecidas na sinalização de quimiocinas e citocinas, responsáveis pela migração e diferenciação dos leucócitos, mediando assim as respostas imunológicas.

Em um dos únicos artigos publicados sobre adaptação local de nativos americanos à floresta tropical, Amorim et al. (2015) encontraram os termos "Cholesterol Biosynthesis" e "Chemokine receptors bind chemokines" enriquecidos ao submeter os genes candidatos à análise de enriquecimento gênico. Entre os 14 genes que encontramos em convergência com este trabalho – que aparecem no extremo 0,5% da distribuição de pelo menos um dos métodos aplicados (PBS, XP-EHH ou iHS) – encontra-se *SCP2*, destacado pelos autores por desempenhar um papel tanto na nutrição como também, potencialmente, no mecanismo de resposta imune, uma vez que o colesterol desempenha um papel importante em doenças infecciosas (LEE et al., 2008). A hipótese destes autores se torna ainda mais consistente pelo fato de que, quatro anos mais tarde, Harrison et al. (2019) realizaram um estudo funcional, submetendo células mononucleares do sangue periférico (*PBMC*, do inglês, peripheral blood mononuclear cells) coletadas de populações africanas caçadoras-coletoras

e agriculturalistas a dois agentes que simulam infecções virais e bacterianas e verificaram que há maior enriquecimento gênico na via de homeostase do colesterol em populações caçadoras-coletoras no grupo que simula infecção viral quando comparado à população agriculturalista próxima (HARRISON et al., 2019). Esta mesma análise também revelou que, em ambos os grupos (exposição a agentes viral e bacteriano), mais o grupo controle (sem exposição a nenhum dos agentes), as células provenientes de caçadores-coletores apresentam genes enriquecidos em vias de resposta à interferon quando comparadas aquelas provenientes de agriculturalistas.

Interessantemente, observamos múltiplos termos associados à reação a picada de mosquitos, totalizando sete genes no dataset S1 (*e.g. AUTS2* e *NOS1AP*) e oito genes no dataset S2, incluindo *CTNNA2* e *KCTD1*. Ainda mais notório se faz o fenótipo "perceived unattractiveness to mosquitos measurement", que pode ser compreendido como "medida da autopercepção à falta de atração aos mosquitos", associado aos genes *CTNNA2* e *LINC02789*, identificados no extremo 0,01% da distribuição de XP-EHH e iHS no dataset S2. Sabe-se que os mosquitos são um dos vetores mais importantes na transmissão de doenças ao redor do mundo, com alta prevalência nos ambientes de floresta tropical, sobretudo em períodos chuvosos (ARAÚJO et al., 2020).

Além dos fenótipos relacionados à picada e não-atratividade de mosquitos, em ambos os datasets identificamos genes associados ao fenótipo de soropositividade ao *Trypanosoma cruzi* (*T. cruzi*), agente responsável pela doença de Chagas. Dentre os genes levantados, destacam-se *PPP3CA* e *DYNC1I1*, com valores extremos em ambos os datasets. A identificação deste fenótipo é intrigante porque, ao passo que se tem evidências concretas da doença na América pré-colombiana na região dos Andes, muitas pesquisas falharam em identificar sua ocorrência na região Amazônica, apesar da ampla distribuição de seus vetores e reservatórios (COIMBRA JR, 1988; AGUILAR et al., 2007). Testes foram realizados em distintas populações nativas Amazônicas do Alto Xingu, Asuriní do Pará, Karitiana e Suruí de Rondônia, e Xavante do Mato Grosso, e todos resultaram em sinais negativos da doença (SANTOS; COIMBRA JR, 1994). Apesar disso, outros trabalhos identificaram alta soropositividade para *T. cruzi* em populações indígenas Amazônicas isoladas, no Equador, e sugerem que este agente está presente nestas áreas tropicais há muitos anos (CHICO et al., 1997). De fato, há indícios da ocorrência de *T. cruzi* nas Américas muito antes da presença humana no continente (ZELEDÓN; RABINOVICH, 1981).

Há também relatos de múmias com doença de Chagas em outras regiões de terras baixas na América, incluindo o Brasil (ARAÚJO et al., 2009). Dada a escassez de resultados sorológicos positivos para *T. cruzi* nas populações nativas Amazônicas, Santos e Coimbra Jr. (1994) argumentam que os motivos pelo qual a doença não é endêmica na Amazônia (em contraste com os Andes) são o tipo de moradia, mobilidade e domesticação

de animais. Contudo, considerando os resultados aqui apresentados, podemos inferir que a diferença existente entre Amazônia e Andes no que se refere à infecção por *T. cruzi* pode ser resultado de um processo seletivo que resultou em um fator protetivo nos habitantes das terras baixas amazônicas.

2.4.6 *Novelty seeking*

Nossos resultados identificaram múltiplos genes associados ao consumo de substâncias aditivas, em ambos os datasets, e pelos três testes de seleção aplicados (PBS, XP-EHH e iHS), conforme análises de enriquecimento gênico realizadas pelas plataformas Enrichr e FUMAGWAS (Tabelas 2.12, 2.13). Além do consumo de álcool, também foi identificado o fenótipo de consumo de café em convergência entre os datasets. Logo, somado a outros fenótipos como "aventura" (Tabela 2.9), é possível que genes que outrora contribuíram para uma ansiedade necessária para subsistência no passado, auxiliando por exemplo maior exploração de ambientes em busca de novidade, ou *novelty seeking*, promovido pela caça e coleta. No entanto, após a transição sociocultural pós-contato, esses mecanismos podem estar contribuindo para maior incidência de dependência química nessas populações. Estudos com o receptor de dopamina D4 (DRD4) mostram que indivíduos que apresentam o alelo 7R são propensos a consumirem tabaco e álcool em excesso, no entanto, também são propensos a apresentarem o temperamento novelty seeking. Notavelmente, nativos americanos caçadores-coletores apresentam uma maior frequência do alelo 7R do que populações agriculturalistas (TOVO-RODRIGUES et al., 2010). Considerando o exposto, nossos achados podem indicar que os fatores antes necessários para um maior sucesso em ambiente nômade e de caça, hoje, devido à sedentarização e perda de contextos culturais, se tornam nocivos aos indivíduos.

O consumo de álcool nas populações indígenas tem sido alvo de múltiplos estudos, uma vez constatado um aumento considerável em sua prevalência (COIMBRA JR; SANTOS; ESCOBAR, 2003). A Fundação Nacional da Saúde (FUNASA) indicou que o alcoolismo está entre as enfermidades mais comuns nos grupos indígenas brasileiros (FUNASA, 2002). Como exemplo da relevância do alcoolismo nestes grupos, no ano 2000 houve um evento científico voltado para a questão do alcoolismo e vulnerabilidade às DST/AIDS entre os povos indígenas, que contou com o reconhecimento do problema e narrativas de representantes indígenas e oito artigos específicos sobre o alcoolismo nestas populações (Ministério da Saúde, 2001).

Guimarães e Grubits (2007), com base no estudo da FUNASA, ainda afirmam que o alcoolismo tem sido considerado uma das principais causas de mortalidade, quer seja por doenças causadas pelo abuso excessivo do mesmo (*e.g.* cirrose, diabetes, estresse) ou por fatores externos, como acidentes, brigas, quedas, atropelamentos, entre outros. Os autores, em geral, associam a prevalência do alcoolismo em populações indígenas a

questões sócio-culturais complexas, decorrentes do processo de colonização, salientando sempre a necessidade de entender a especificidade cultural e histórica de cada grupo, bem como o significado de beber (dado que existem rituais religiosos que incluem bebida) (Ministério da Saúde, 2001; LANGDON, 2005).

Em uma tentativa de compreender a relação do alcoolismo com a cultura, Horton (1943) postulou que o grau de embriaguez é proporcional ao grau de ansiedade da cultura, sendo que a ansiedade causada por (1) nível de subsistência, (2) presença ou ausência de risco à subsistência, e (3) grau de aculturação (GUIMARÃES; GRUBITS, 2007). Neste sentido, dado os riscos eminentes do subsistência predominantemente caçador-coletor em regiões de Floresta Tropical, a consequente alta mortalidade nestes ambientes, e o processo de aculturação decorrente do contato com os brancos, as populações indígenas brasileiras estudadas neste trabalho podem ser encaixadas na hipótese de Horton.

2.5 CONCLUSÃO

Através de diferentes abordagens de varreduras seletivas e testes de enriquecimento gênico, identificamos múltiplos genes candidatos que contribuem para resposta imunológica a doenças infecciosas, estrutura corporal, metabolismo energético e comportamento de risco. Para cada um destes fenótipos, discutimos sua relevância biológica no contexto das populações caçadoras-coletoras nativas da floresta amazônica. Em suma, destacam-se três achados: a. genes que foram outrora importantes para subsistência como caçador-coletor em ambientes de floresta tropical, promovendo maior aproveitamento energético e exposição ao risco, passaram a contribuir para o perfil epidemiológico atual dos indígenas brasileiros, com alta prevalência de diabetes, obesidade e alcoolismo; b. vias relacionadas à exposição à luz solar e função tireoidal encontradas sob seleção que podem ter sido importantes para promoção da termorregulação e do desenvolvimento reprodutivo precoce, contribuindo para diminuição da estatura nas populações indígenas ao longo de sua história evolutiva; e c. a resposta a patógenos moldou o repertório imunológicos das populações de floresta, o que pode ser evidenciado tanto em nível macro, como a seleção em genes que levam a uma menor resposta a picadas de insetos em indivíduos que habitam a floresta tropical, como a evidência de genes relacionados à infecção pelo *T. cruzi* em nativos amazônicos, que pode estar relacionado a um fator protetivo nessas populações, e que é inexistente nas populações de altitude, onde a doença de Chagas é endêmica.

Os resultados aqui apresentados configuram uma visão abrangente de diferentes processos que podem ter sido fundamentais na adaptação ao ecossistema amazônico ao longo dos últimos milhares de anos.

Capítulo 3

ADAPTAÇÃO LOCAL AOS ANDES

3.1 INTRODUÇÃO

3.1.1 Características e Populações Nativas

A Cordilheira dos Andes, ou simplesmente Andes, trata-se de uma vasta cadeia montanhosa que se estende desde a Venezuela até a Patagônia (aproximadamente 8.000 km), e corresponde a uma das principais ecorregiões da América do Sul. Sua altitude média é de 4.000 m acima do nível do mar. Na região central com elevada altitude, há o planalto Altiplano, abrangendo três países: Peru, Bolívia e Chile. De modo geral, a região ao nordeste dos Andes é mais baixa, úmida e fértil, enquanto existe uma região ao sudeste mais alta e árida.

As diferenças nas latitudes andinas são importantes para os estudos de estrutura populacional, migração e interações culturais e socioeconômicas entre outras regiões, como a Amazônia (BORDA et al., 2020).

As populações mais estudadas e também as mais abundantes nas terras altas andinas são os povos indígenas Quéchua e Aymara, com aproximadamente 6 milhões de habitantes que residem majoritariamente no Peru e Bolívia (JULIAN; MOORE, 2019). Há evidências arqueológicas da ocupação humana neste ambiente há cerca de 14.000 AP (RADEMAKER et al., 2014).

Estudos fisiológicos revelaram as populações das terras altas possuem maior concentração de hemoglobina no sangue, em contraste com populações próximas de terras baixas, ou mesmo em populações de terras altas da Etiópia e Tibete, que só apresentaram tal aumento de concentração quando acima de 4.000 m de altitude (JEONG; RIENZO, 2014). Outros fenótipos identificados na população nativa dos Andes são aumento da pressão pulmonar arterial, aumento da ventilação em repouso, diminuição da saturação

arterial de oxigênio e presença de vasoconstrição devido à hipóxia (JEONG; RIENZO, 2014).

Ao passo que o aumento da concentração de hemoglobina no sangue fornece uma vantagem para circulação e distribuição de oxigênio aos tecidos do corpo, há como consequência maior viscosidade sanguínea, que por sua vez pode gerar complicações para o sistema circulatório e o período de gestação.

3.1.2 Adaptação Local aos Andes

Considera-se "terras altas" ou "elevadas altitudes" as regiões com mais de 2.500 metros acima do nível do mar (MOORE, 2017). Além das terras altas Andinas, estudos de seleção são frequentemente conduzidos nas terras altas do Himalaia (China/Tibete) e Etiópia (África), uma vez que a pressão seletiva nestas regiões é similar, além de apresentarem populações nativas isoladas durante milhares de anos, tornando-os "laboratórios" excelentes para análise de seleção (FAN et al., 2016).

Assim como os estudos com pigmeus e outras populações nativas de floresta tropical podem ser utilizadas para comparação com os estudos na Amazônia (AMORIM et al., 2015), estudos realizados nas terras altas do Tibete e Etiópia costumam ser consideradas em conjunto no estudo de adaptação local aos Andes (BIGHAM; LEE, 2014), quer seja para verificar convergência evolutiva, ou evolução em diferentes loci que contribuem para um mesmo fenótipo, ou mesmo distintos fenótipos para lidar com uma mesma pressão evolutiva. Embora existam mais estudos nos Andes do que na Amazônia, o mesmo não é verdade para regiões com pressões seletivas similares, como nas terras altas do Tibete (JULIAN; MOORE, 2019).

A maioria das análises de seleção realizadas nas populações nativas dos Andes tem identificado genes pertencentes a via dos fatores indutores de hipóxia (HIF, do inglês, *hypoxia-inducible factors*) – uma via essencial na homeostase do oxigênio –, ainda que estes não sejam os únicos responsáveis pela homeostase do oxigênio, tampouco os únicos alvos de seleção nestes ambientes (revisado por JULIAN; MOORE, 2019). De fato, Bigham et al. (2009) verificaram que a via HIF, como um todo, não apresenta sinal de seleção (BIGHAM et al., 2009).

Se comparado com outros povos nativos de terras altas, como, por exemplo, as populações do Himalaia, os nativos Andinos apresentam tanto sinais de convergência evolutiva, como, majoritariamente, sinais de seleção particulares dos Andes. Um dos sinais de convergência evolutiva mais fortes foi encontrado no gene *EGLN1* (BIGHAM et al., 2010), que codifica uma molécula sensitiva ao oxigênio responsável pela regulação da transcrição de HIF (TO; HUANG, 2005). Pelo menos outros 18 genes foram detectados em convergência evolutiva com populações nativas das terras altas do Himalaia, conforme revisto por Moore (2017).

Dentre os genes candidatos identificados em populações Andinas, destacam-se genes envolvidos na angiogênese (*VEGF*, *ELTD1*) (EICHSTAEDT et al., 2014), vasoreatividade (*EDNRA*, *BRINP3*, *NOS2*, *PRKAA1*, *TBX5*) (BIGHAM et al., 2009; BIGHAM et al., 2010; CRAWFORD et al., 2017) e defesa oxidativa (*FAM213A*) (VALVERDE et al., 2015). Ademais, foram detectados sinais de seleção positiva no gene *AS3MT*, responsável pela metabolização de arsênico, uma substância tóxica comum em fontes de água em regiões Andinas (*e.g.* Puna) (EICHSTAEDT et al., 2015), bem como no gene *TMEM38B*, que possivelmente atua na redução da policitemia (excesso de eritrócitos, com consequente aumento da viscosidade e prejuízo no fluxo sanguíneo) (CRAWFORD et al., 2017).

A fim de detectar possíveis novos candidatos para adaptação local nas terras elevadas andinas, utilizamos populações nativas publicadas no artigo de Reich et al. (2012) e aplicamos estatísticas inferenciais de seleção, como XP-EHH e PBS, simulação de cenários demográficos e análise de expressão gênica in silico. Como resultado, publicamos um artigo na revista *Scientific Reports*, sendo a primeira autoria dividida entre Vanessa Jacovas, Cainã Max Couto-Silva e Kelly Nunes (JACOVAS et al., 2018). Atualmente, o artigo conta com sete citações e um dos genes candidatos (*DUOX2*), identificado pela primeira vez em nosso artigo, também foi detectado sob seleção nos Andes em artigo posterior (BORDA et al., 2020).

3.2 ARTIGO

Abstract

The Andean Altiplano has been occupied continuously since the late Pleistocene, 12,000 years ago, which places the Andean natives as one of the most ancient populations living at high altitudes. In the present study, we analyzed genomic data from Native Americans living a long-time at Andean high altitude and at Amazonia and Mesoamerica lowland areas. We have identified three new candidate genes - *SP100*, *DUOX2* and *CLC* - with evidence of positive selection for altitude adaptation in Andeans. These genes are involved in the *TP53* pathway and are related to physiological routes important for high-altitude hypoxia response, such as those linked to increased angiogenesis, skeletal muscle adaptations, and immune functions at the fetus-maternal interface. Our results, combined with other studies, showed that Andeans have adapted to the Altiplano in different ways and using distinct molecular strategies as compared to those of other natives living at high altitudes.

3.2.1 Introduction

Along their great expansion, humans have inhabited almost all environments in the five continents. Among several harsh environments that were occupied, the highlands are probably the ones that needed more adaptations for survival (ESPINOZA-NAVARRO et al., 2011). At least in three geographically distinct locations have this evolutionary adaptation been studied: Andean Altiplano (South America), Himalaya (China/Tibet, Asia) and Semien Mountain (northern Ethiopia, Africa) Plateaus. Andes have been peopled continuously since the late Pleistocene, 12,000 yBP2 while the time of settlement and permanent occupation of both Tibet and Ethiopia remain a topic of debate, varying widely (ALDENDERFER, 2011; LU et al., 2016). Despite some uncertainties in the permanent occupation dating, it is certain that humans have inhabited these regions of hostile climates for thousands of years.

Several physiologic factors are associated with living at high altitude ($\geq$ 2,500 meters where only 75% of the oxygen available at sea level occurs; (<http://www.altitude.org/air_pressure.php>), including adaptations for high ultraviolet radiation index, thermal amplitude, and changes in the pulmonary capacity due to hypoxia (MOORE, 2001; SABETI et al., 2007). High altitude leads to a rapid physiologic/adaptive response in individuals from lowlands; however, prolonged exposure to environmental-related factors might have harmful outcomes. Remarkable features such as increased pulmonary function, hypoxia tolerance, and increased hemoglobin levels have been observed in Andean populations (BIGHAM; LEE, 2014). How such adaptations took place is still not clear, and just a few genes have been associated with the high altitude adaptation phenotype in human populations (SCHEINFELDT et al., 2012; HUERTA-SÁNCHEZ et al., 2013; SIMONSON et al., 2015; VALVERDE et al., 2015; FEHREN-SCHMITZ; GEORGES, 2016; CRAWFORD et al., 2017).

Interestingly, the set of genes presenting signs of natural selection changes according to high altitude, indicating that under an analogous selective pressure, different genetic solutions have emerged. For instance, genomic scans for selection have revealed at least 40 candidate genes related to the Hypoxia Inducible Factor (HIF), such as *EPAS1* in populations from Tibet, *EGLN1* in Andeans and Tibetans and *THRB* and *ARNT2* in Ethiopians (BEALL et al., 2010; BIGHAM et al., 2010; SIMONSON et al., 2010; PENG et al., 2011; XU et al., 2011; SCHEINFELDT et al., 2012). The populations from the Andean plateau also presented signs of natural selection in other genes, such as *BRINP3*, *NOS2*, and *TBX5*, involved in the nitric oxide pathway (NOS) and related to cardiovascular health (FEHREN-SCHMITZ; GEORGES, 2016). In addition, Jacovas et al. (2015) using the candidate gene approach inferred that a combination of some derived and ancestral alleles of *USP7*, *LIF* and *MDM2* genes, all three in the *TP53* pathway, could have been essential for the successful establishment of Native American populations in the Andean

highlands.

Since different investigations pointed to distinct sets of genes involved in high altitude adaptation, more studies are necessary to fully understand the different genetic landscapes present in highland populations around the world. In the present study, we compared genomic data from Native American populations living for a long-time at high altitude (Andean Altiplano) with those living at lowlands (Amazon and Mesoamerica), with the purpose of expanding our knowledge about the genetic repertoire responsible for the successful human colonization of the Andes.

3.2.2 Results

3.2.2.1 Natural selection analysis

Population Branch Statistic (PBS) values were estimated for each individual SNP. To avoid spurious results due to single SNPs, windows of 20 SNPs were used to estimate the mean PBS values for a given region. Then, we checked the outliers' peaks, above the 99.5th and 99.9th percentiles, to identify in each outlier window the SNPs with the highest PBS value and assigned the gene to which it belonged (or the nearest gene). Based on this approach, five candidate genes were identified: *SP100* (SP100 Nuclear Antigen), *TMEM38B* (Transmembrane Protein 38B), *AS3MT* (Arsenite 3 Methyltransferase), *DUOX2* (Dual Oxidase 2) and *CLC* (Charcot-Leyden Crystal Galectin, also known as Galectin-10) (Table 3.1 and Figure 3.1). Among these candidate genes, *AS3MT* and *TMEM38B* have been identified in previous scans for natural selection in Andeans (EICHSTAEDT et al., 2015; CRAWFORD et al., 2017).

Table 3.1 – Population Branch Statistic (PBS) individual values and Cross-Population Extended Haplotype Homozygosity (XP-EHH) for all SNPs found under selection in Native Andean populations.

SNP	Allele		Gene	Position	PBS	XP-EHH			
	Ancestral	Derived				Andean vs. Mesoamerican	*p*-value	Andean vs. Amazonian	*p*-value
rs13411586	C*	T	*SP100*	230988046	0.5846	2.3789	0.0037	2.1703	0.0065
rs9678342	C*	T	*SP100*	230991955	0.5547	2.3193	0.0044	2.1356	0.0071
rs7582700	T*	C	*SP100*	231024349	0.4644	2.2704	0.0050	2.1074	0.0076
rs7039618	A	G*	*TMEM38B*	107497627	0.3618	0.0842	0.3312	0.5672	0.1458
rs3817141	T*	C	*TMEM38B*	107507950	0.3906	0.0255	0.3099	0.6205	0.1351
rs10978213	G*	A	*TMEM38B*	107511706	0.3618	0.0235	0.3092	0.6171	0.1358
rs10816302	A*	G	*TMEM38B*	107526354	0.3835	0.0937	0.2697	0.6664	0.1264
rs10978240	A	G*	*TMEM38B*	107575093	0.3923	0.0764	0.2753	0.6307	0.1331
rs1046778	T	C*	*AS3MT*	104651474	0.3124	0.5023	0.5118	0.4008	0.4631
rs269866	G*	A	*DUOX2*	43181698	0.6185	2.0599	0.0086	2.5865	0.0021
rs440191	A	G*	*CLC*	44913483	0.3039	1.6207	0.0234	0.3166	0.2046

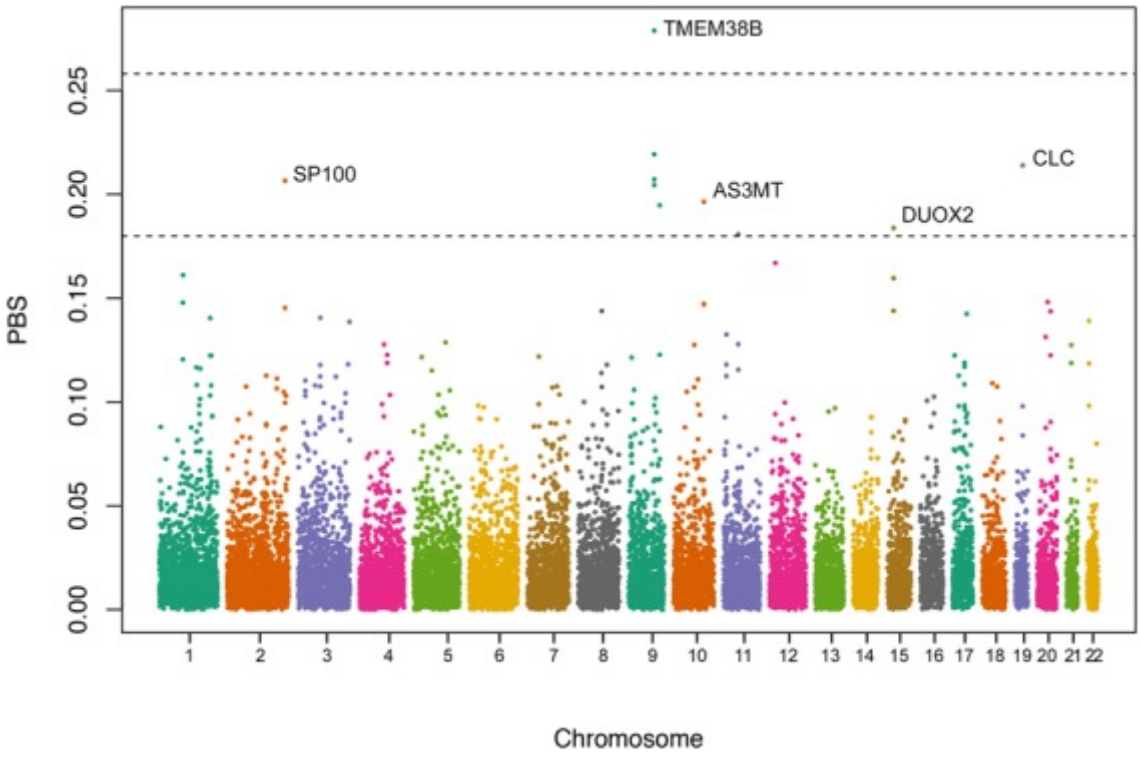

Figure 3.1 – Average PBS values in windows of 20 SNPs, using a step size of 5 SNPs. The 99.5th and 99.9th percentiles of the empirical distribution are shown as black dashed horizontal lines. Names of genes associated with the highest peaks are shown.

Neutral coalescent simulations indicated that these deviations were statistically significant (p-values ranging between 0.03 and 0.0001; Figure 3.2, Table 3.3), consistent with the action of positive selection as opposed to genetic drift in increasing the frequency of the putative selected alleles at all five tested loci. In addition, we applied the Cross-Population Extended Haplotype Homozygosity (XP-EHH) test to the same regions. The XP-EHH results also show significant differences between the Andean and Mesoamerican groups in three SNPs (rs13411586, rs9678342, rs7582700) of *SP100* and one SNP (rs269866) of *DUOX2* (Table 3.1). These SNPs, which are under putative selection in the PBS analysis with the most extreme values (0.46 to 0.62), also present significant XP-EHH values ≥ 2 in both Andean vs Mesoamerican and Andean vs Amazonian groups.

The observed allele density provided by the iHS test showed a notable Gaussian distribution pattern for all three groups (Supplementary Figure 3.4), with homozygosity decaying according to the distance from the focal markers.

It should be noted that the distribution of alleles C (rs13411586, *SP100*), G (rs269866, *DUOX2*) and G (rs440191, *CLC*), which presented the highest PBS values (Table 3.1), showed their highest values in areas of very high Andean altitudes (Table 3.2 and Figure 3.3).

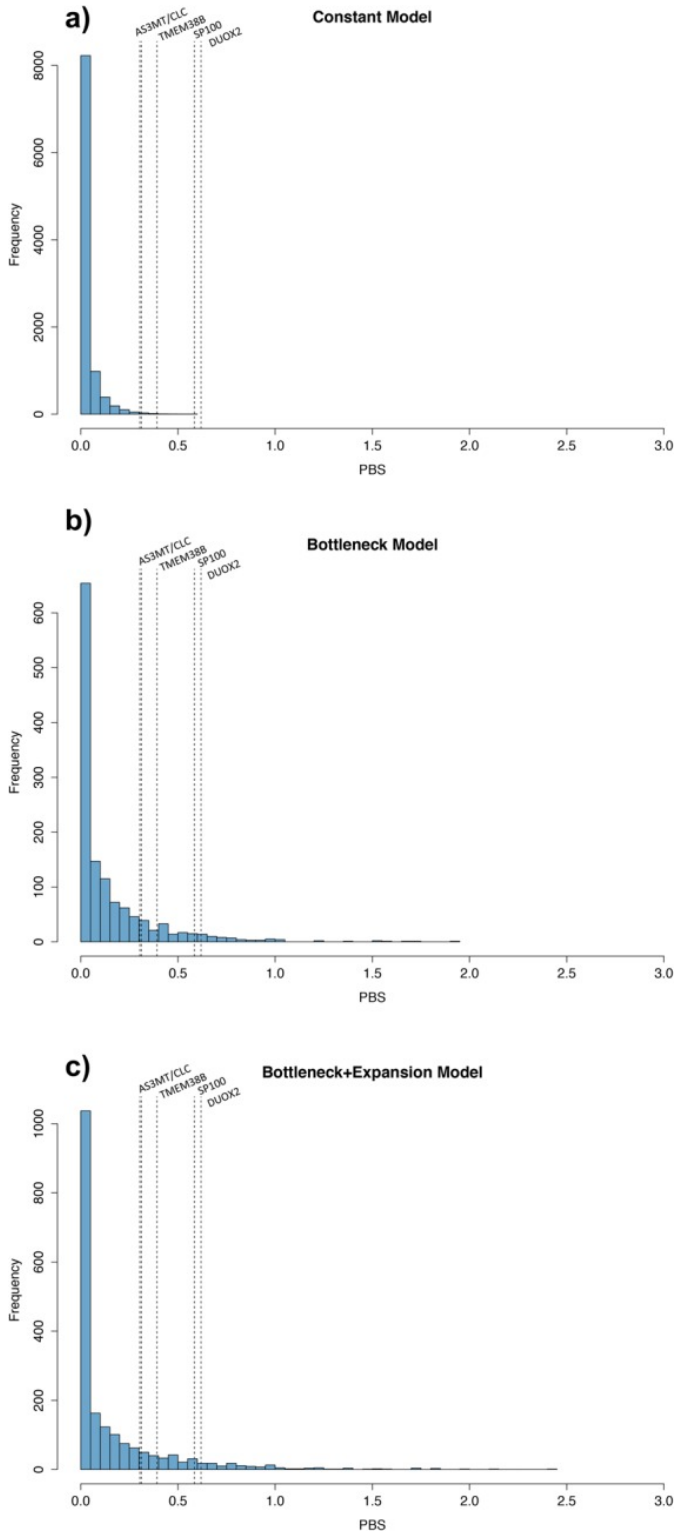

Figure 3.2 – Distribution of 10,000 simulated PBS values under three neutral coalescent models. **(a)** Constant population model. **(b)** Population bottleneck model; and **(c)** Population bottleneck followed by expansion model. The dashed line represents the top observed PBS SNP values in the empirical datasets.

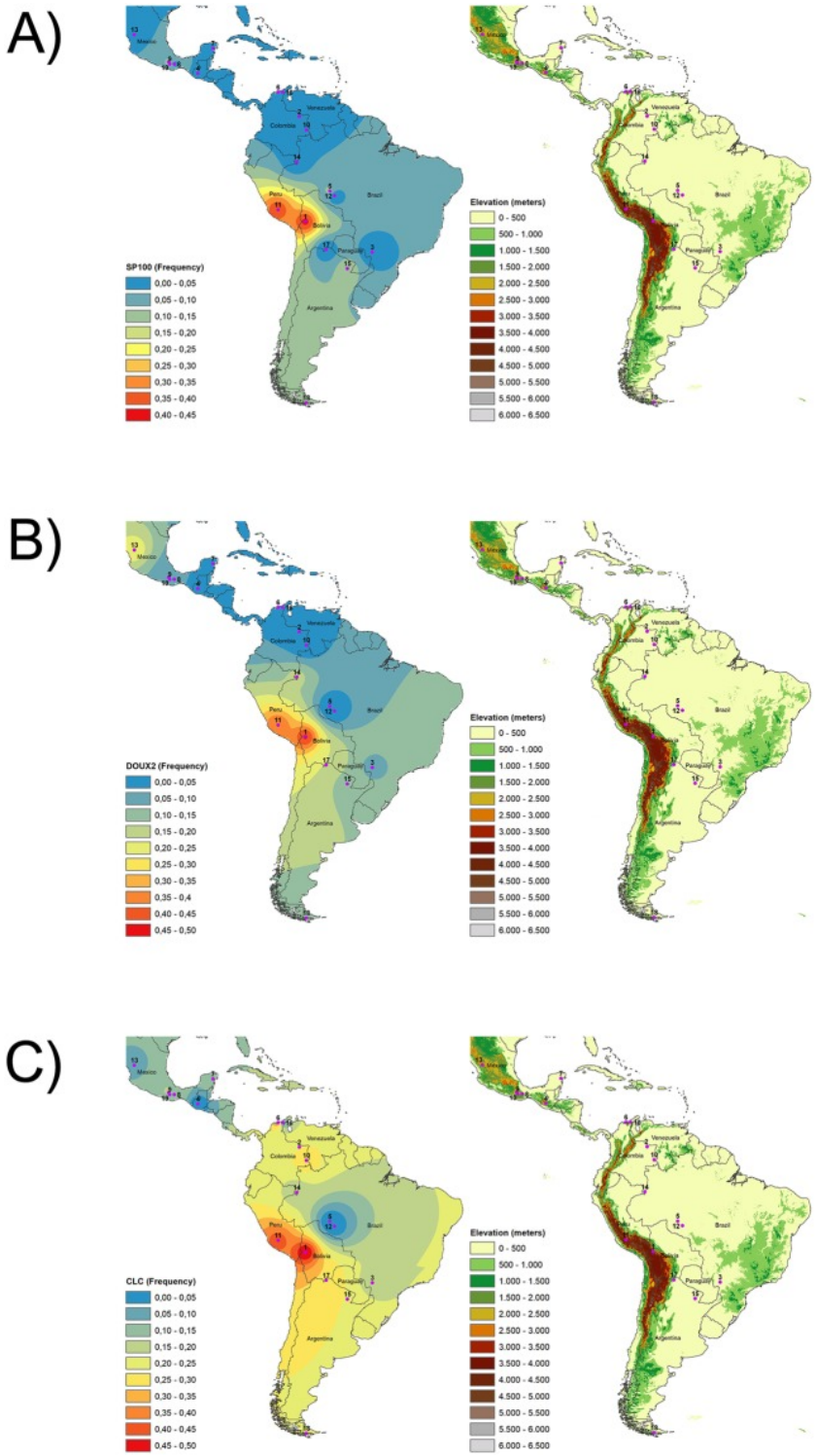

Figure 3.3 – **(a)** rs13411586_C (*SP100*). **(b)** rs269866_G (*DUOX2*) and **(c)** rs440191_A (*CLC*) allele frequency distributions according to altitude. Populations (n $\geq$ 3): 1. Aymara, 2. Guahibo, 3. Guarani, 4. Kaqchikel, 5. Karitiana, 6. Kogi, 7. Maya, 8. Mixe, 9. Mixtec, 10. Piapoco, 11. Quechua, 12. Surui,13. Tepehuano, 14. Ticuna, 15. Toba, 16. Wayuu, 17. Wichi, 18. Yaghan and 19. Zapotec.

Table 3.2 – Frequencies of the putatively selected alleles in the populational groups.

Population (n)	*DUOX2* G allele (rs269866)	*SP100* C allele (rs13411586)	*CLC* G allele (rs440191)
Mesoamerican Lowland (<2,500m.)			
Total (153)	0.068*	0.045*	0.128*
South American (Andean) Highland (≥4,000m.)			
Total (63)	0.420*	0.397*	0.452*
South American (Amazonian) Lowland (<2,500m.)			
Total (106)	0.048*	0.053*	0.142*

* Weighted average

Bootstrap simulations indicated that in all instances the 95% confidence interval of allele frequencies in lowlanders does not include the average values observed for populations living in high altitudes (>4000 m above sea level) (Supplementary Figure 3.5), suggesting that the differences found in allele frequencies between population groups might be caused by a non-random evolutionary process.

3.2.2.2 Effects of putatively selected alleles on gene expression

Homozygotes for the *DUOX2* putatively selected allele (rs269866 G) presented a slight increase in the expression of the *DUOX2* protein (Supplementary Figure 3.6). Multiple testing across tissues showed significant expression of this protein in thyroid (m-value = 1.0), lungs (m-value = 0.996) and aorta artery (0.996) (Supplementary Figure 3.7). Homozygotes for the rs13411586 (*SP100*) putatively selected allele (C) presented an increase in the expression of the SP100 protein in skeletal muscles (Supplementary Figure 3.8). Multiple testing across tissues showed significant expression of this protein in skeletal muscle (m-value = 1.0) and testis (m-value = 0.971) (Supplementary Figure 3.9). There is no information available about the *CLC* gene expression profile.

3.2.3 Discussion

We identified five loci under positive selection in Andean Native populations. Two of them were previously described: *AS3MT* was found to be under positive selection in Colla Andeans systematically exposed to arsenic water (EICHSTAEDT et al., 2015) while *TMEM38B* reduced the negative effects of polycythemia (elevated hematocrit or decreased plasma volume) at high altitudes (CRAWFORD et al., 2017). Three other genes, *SP100*, *DUOX2*, and *CLC* were identified for the first time in a high-altitude context in the present study. These genes are part of the *TP53* pathway, already indicated as a potential

candidate to be under natural selection in high altitude populations (EICHSTAEDT et al., 2014; JACOVAS et al., 2015).

SP100 is a single-copy gene in the human genome that produces several alternatively spliced Sp100 protein isoforms known as modulators of the p53 activity (BERSCHEMINSKI et al., 2016). We found three SNPs in the *SP100* gene with high and significant PBS values, as well as significant XP-EHH values when Andeans were compared to others. One of these SNPs, rs13411586, is differentially expressed in humans; our prediction showed that individuals homozygous for the putatively selected allele (C) have increased Sp100 production.

Interestingly, we also identified that the *SP100* gene is differentially expressed in skeletal muscles (Supplementary Figure 3.6). Studies have revealed that a member of the HIF pathway, HIF-1, plays an important role in the regulation of oxygen homeostasis, which includes the physiological skeletal and heart muscle adaptations in situations of oxygen reduction due to muscular effort (SEMENZA, 1999; VOGT; BILLETER; HOPPELER, 2003; LINDHOLM; RUNDQVIST, 2016) and ischemic cardiomyopathy, respectively (HERRER et al., 2015). Exposure to high altitude leads to reduced muscle mass and performance (*e.g.* lower work capacity and standing fatigue), except when one is evolutionarily adapted to it (MARCONI; MARZORATI; CERRETELLI, 2006; BRUTSAERT, 2008; COUNTER et al., 2017).

HIF-1 protects cell-survival during low oxygen supply, while p53 promotes genome cell-death under hypoxia. The reason for these apparently antagonistic roles can be in the difference of the oxygen quantity available; in a normal condition, both p53 and HIF-1 levels are low, but in mild hypoxia, the p53 level remains low, whereas the HIF-1 level increases, protecting cells still relatively healthy from destruction. In severe hypoxia, p53 accumulation promotes the repression or degradation of anti-apoptotic proteins like HIF-1, inducing apoptosis of the damaged cells (SCHMID; ZHOU; BRÜNE, 2004; OBACZ et al., 2013; ZHOU et al., 2015). Sp100 is known as a modulator of the p53 activity (BERSCHEMINSKI et al., 2016) and under tissue hypoxia due to ischemia, it is downregulated, leading to genomic instability (HERRER et al., 2015). The Andean population presents high allele C (rs13411586) frequency (Table 3.2), which in homozygosis increase Sp100 production according to our prediction test. Our result suggests an evolutionary solution to keep Sp100 at an adequate level in an environment with a constant low oxygen level. Furthermore, it is possible to speculate that there is an intricate balance in the level of expression of the *SP100*, *TP53* and *HIF-1* genes under hypoxia, considering both short (reversible physiological and metabolic adaptations) and long-term evolutionary adaptation scenarios.

DUOX2, expressed in epithelial cells of various tissues including nasal and lung, participates in the hydrogen peroxide (H_2O_2) pathway, which is required in the final steps

of thyroid hormones production. It is also involved in Reactive Oxygen Species (ROS), a byproduct of the normal oxygen metabolism even under normal physiologic conditions (DEVASAGAYAM et al., 2004). However, different stressor conditions can increase the ROS production, *i.e.* high-altitude exposure (hypoxia and UV exposure), and pathological conditions such as cancer (GUPTA et al., 2012). Salmeen, Park e Meyer (2010) provided evidence that *DUOX2* plays a role in a p53-dependent checkpoint mechanism for cell cycle entry.

In vitro and *in vivo* experiments showed that oxidative stress and generation of ROS caused by *DUOX2* overexpression, in both hypoxia and hyperoxia, contribute to inflammation, carcinogenesis and cell death (KIM et al., 2014; BAUTISTA-ORTEGA et al., 2014; FLETCHER et al., 2014; DIAS-FREITAS; METELO-COIMBRA; RONCON-ALBUQUERQUE, 2016; MIN et al., 2017; LIN et al., 2017; MACFIE et al., 2014). For instance, a functional study (KIM et al., 2014) showed that under hyperoxia conditions, mutant mice for *DUOX2* had significant lower acute lung injuries induced by hyperoxia. This finding pointed to the importance of these proteins in the response to changes of oxygen concentration in the environment. Another study (BAUTISTA-ORTEGA et al., 2014) found that chickens submitted to hypoxia (>3,000 m) had increased activity of DUOX/NOX proteins, indicating the physiological role of these enzymes in the process of adaptation to oxidative stress.

Our results on the expression of the *DUOX2* putatively selected allele G (high PBS values and significant XP-EHH value > 2; Table 3.1) also pointed to higher levels of protein expression in humans, mainly in the lungs and arteries. It is noteworthy that ROS contributes to inflammation in the vessel walls. (KIM; BYZOVA, 2014) demonstrated that ROS has an important role in angiogenesis, a process of new blood vessel growth. Angiogenesis is a key event in the physiological response to hypoxia and therefore might have a role in the adaptation to high altitude in long-term residents, especially in individuals with excessive erythropoiesis (like those found in the Chronic Mountain Sickness [CMS] phenotype), to compensate a plausible change in microcirculation (GE et al., 2011; BUROKER et al., 2012).

SNP rs440191 is located at the 3'UTR region of *CLC*, and the putatively selected allele G is in complete linkage disequilibrium with the *CLC* rs395892 G allele in the Mexican population (The 1000 Genomes Project Consortium; DELANEAU; MARCHINI, 2014). The latter is associated with eosinophil and basophil counts (1000 Genomes Project Consortium et al., 2015), while rs440191 has so far been investigated just in approaches assessing allergic susceptibilities (ASTLE et al., 2016). Gene expression queries did not show any significant eQTL related to this polymorphism, preventing any prediction of tissue-specific expression.

CLC (galectin-10) is still a poorly studied gene when compared to other members of

the functionally polyvalent galectin family. It is recognized as a lysophospholipase expressed in eosinophils and basophils, although some authors identified it just as an enzyme that interacts with lysophospholipases (ACKERMAN et al., 2002). The only functional study regarding this protein showed that hypoxia increases eosinophil accumulation and *CLC* production in humans, concomitant with a delay in constitutive apoptosis, antagonizing the normal pro-apoptotic effect of agents that normally induce eosinophil apoptosis (PORTER et al., 2017).

Regulation by the p53 transcription factor seems to be important in the galectin family genes' expression. For instance, the galectin-3 gene has a binding site for p53, and p53 increases the transcription of paralogue galectin-7 (POLYAK et al., 1997; RAIMOND et al., 1995; COOPER, 2002). Altered expression of galectin genes, including *CLC*, was implicated in cancer emergence and progression, highlighting the role of the galectins in cell proliferation via cell death programs (GOPALAN et al., 2016).

Investigations with galectin paralogues have shown that galectin-1 in the first term ovine gestation placenta prevented inflammatory processes that harm the fetus (IGLESIAS et al., 1998), while galectin-13, which has the highest homology to *CLC*, is a member of the group of the so-called "pregnancy-related proteins", due to its special immune functions at the feto-maternal interface (THAN et al., 2004; SU et al., 2018). These fundamental cell functions, already described for humans and other placental mammals, may indicate the path that connects our *CLC* findings and the selection pressure in the Andean hostile climate.

In conclusion, our results pointed to a complex adaptation that occurred in Andean natives, which involved the *CLC*, *SP100* and *DUOX2* genes, not previously correlated in contexts of long-time adaptation to high altitudes. We also reinforced the role of the *TP53* pathway at least for the adaptation to the Andean environmental stresses. Combined with other studies, and incorporating the present one, it is clear that Andeans have adapted to the Altiplano in different ways and using distinct molecular strategies than those of other natives living at high altitude.

3.2.4 Methods

3.2.4.1 Populations

We analyzed 213,987 SNPs determined with Illumina 610quad from 63 Native Americans living at extreme high altitude (≥4,000 m; 63% of the oxygen available at sea level; <http://www.altitude.org/air_pressure.php>) and 259 living at lowland areas (<2,500 m), data previously published by Reich et al. (2012). Highlanders included Aymara and Quechua Andeans, while lowlanders were represented by 25 populations from the Mesoamerican and South American lowlands. Details about these populations, sample

sizes and allelic frequencies are given in Suplementarry Table 3.4. Additional information, including ethical authorizations for evolutionary and anthropological studies, can be found in the primary publication58.

3.2.4.2 Population Branch Statistic (PBS) analysis

PBS determinations were performed between pair of populations, using Andean and Amazonian populations as sister groups and Mesoamericans as an outgroup. The analysis was carried out as described by Yi et al. (2010), with only the polymorphic SNPs in at least two of the populations being considered. From the genetic distances (FST) between the three population groups examined, PBS measures if there are alleles with extreme frequencies in the Andean group as compared to the other two. Under a scenario of genetic drift only, we expect that Andeans and Amazonians will be more similar genetically than both compared to Mesoamericans. If, however, there has been local adaptation, we should detect genes that have been targeted by selection in Andeans. PBS values were estimated for both individual SNPs and windows of 20 SNPs overlapped in five SNPs. The empiric distribution of PBS values, with a 99.5^{th} threshold, was used to determine signals of positive selection (more details in Amorim et al. (2017)).

3.2.4.3 Demographic simulations

To verify the significance of the observed positive selection signals we simulated different demographic models, according to reported historical population data and inferred effective population sizes. We adapted the models described by Valverde et al. (2015), to account for the divergence between Mesoamericans, Andeans and Amazonians. Assuming that the American continent was peopled beginning at 15,000 yBP, the Andes colonized by 12,000 yBP and the Amazon by 10,000 yBP, and based on Nes estimated by Valverde et al. (2015), we simulated the three demographic models proposed by them: (a) Constant Model: Ne of 7,000 individuals with constant size in all populations throughout history; (b) Bottleneck Model: Ne 8,000 in Mesoamerica, 4,000 in Andes and 2,000 in Amazon; and (c) Bottleneck + Expansion Model: model b with bottlenecks reducing the effective size of all populations by 50% in the last 10,000 years followed by a sharp expansion in the last 8,000 years. Simulations were performed in the MS program (HUDSON, 2002) with 10,000 replicates for each demographic scenario.

3.2.4.4 Linkage disequilibrium analysis

We also used three linkage disequilibrium-based methods: extended haplotype homozygosity (EHH) (SABETI et al., 2002), integrated haplotype score (iHS) (VOIGHT et al., 2006), and cross-population extended haplotype homozygosity (XP-EHH) (SABETI et al., 2007). These approaches adopt the same core principle, that an advantageous

allele under a hard sweep rise in frequency – carrying its neighbor alleles and therefore promoting homozygosity extension – quickly enough that recombination is not able to break down the haplotype. EHH statistics calculate the homozygosity rate from a core region (putative allele under selection) to the neutral scenery, *i.e.* the probability that any two randomly chosen chromosomes will be identical by descent, from the core region to a distance x. iHS evaluates the EHH considering both ancestral and derived alleles, and XP-EHH is used to calculate EHH/iHS between populations, therefore controlling for local variation. These tests are complementary; while iHS is better for detecting incomplete sweeps, XP-EHH has more power to detect sweeps near fixation (VITTI; GROSSMAN; SABETI, 2013). Both measurements and significance were calculated through the 'rehh' R package (GAUTIER; VITALIS, 2012).

3.2.4.5 Geographical analysis

To evaluate the variants spatial distribution, weighted inverse distance interpolation (IDW) was used to determine cell values using a weighted linear combination of a set of sample points. Weight is a function of the inverse distance (WATSON; PHILIP, 1985). The maps were made with the ArcGis 10.5 software and the cartographic base was georeferenced to the World Geodetic System (WGS84).

3.2.4.6 Bootstrap simulations

To verify whether the allele frequencies of the candidate variants under selection are significantly different among extreme high (>4,000 m) and lowland (<4,000 m) populations, we obtained the 95% confidence intervals of the average allele frequency of the lowland populations by means of 10,000 computer-assisted bootstrap simulations with replacement, considering a sample as having the same size and genotypic proportions observed in the real one. The average allele frequencies from high and lowland populations were obtained by weighing the observed frequencies according to their sample sizes.

3.2.4.7 Analysis of gene expression

We used the Genotype-Tissue Expression Portal (GTEx; <https://www.gtexportal.org/home/>) to evaluate possible associations between each of the candidate alleles with highest differentiation and gene expression across human tissues looking for evidence of quantitative trait loci (eQTLs). The m-value is the posterior probability that an eQTL effect exists in each tissue tested in the cross-tissue meta-analysis. The m-value ranges between 0 and 1 (m-values > 0.9 mean that the tissue is predicted to have an eQTL effect).

3.2.5 Supplementary Material

Table 3.3 – Significance of the PBS values for the extreme SNPs of each candidate gene, obtained for each simulated demographic model.

SNP (Gene)	Constant Model	Bottleneck+Expansion	Bottleneck
rs13411586 (SP100)	0.0001	0.0143	0.0074
rs10978240 (TMEM38B)	0.0017	0.0248	0.0147
rs1046778 (AS3MT)	0.0047	0.0315	0.0192
rs269866 (DUOX2)	0.0001	0.0132	0.0060
rs440191 (CLC)	0.0056	0.0321	0.0206

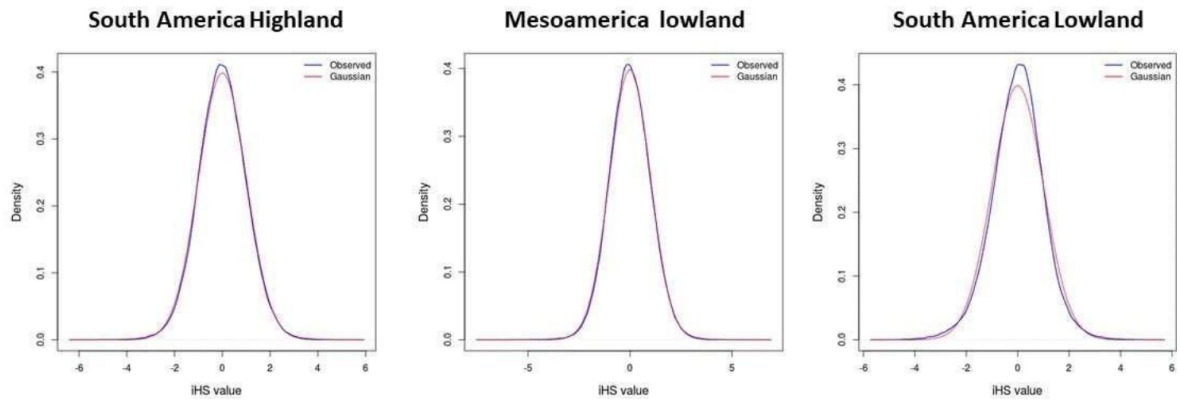

Figure 3.4 – iHS value distribution patterns for all three groups (South American highland, South America Lowland and Mesoamerica Lowland).

Table 3.4 – Allelic frequencies by Native American population analyzed in the present study.

Population (n)	**DUOX2 G allele (rs269866)**	**SP100 C allele (rs13411586)**	**CLC G allele (rs440191)**
Mesoamerican Lowland (< 2,500 m)			
Kaqchikel (13)	0	0.042	0
Maya (49)	0	0	0.138
Mixe (17)	0.059	0.029	0.147
Mixtec (5) 0.100 0.100 0.100			
Purepecha (1) 0 0 0			
Tepehuano (25)	0.240	0.020	0.080
Zapotec (43)	0.068	0.114	0.182
Total (153)	**0.068***	**0.045***	**0.128***
South American (Andean) Highland (≥ 4,000 m)			
Aymara (23)	0.457	0.413	0.500
Quechua (40)	0.400	0.388	0.425
Total (63)	**0.420***	**0.397***	**0.452***
South American (Amazonian) Lowland (< 2,500 m)			
Guahibo (6)	0	0	0.250
Guarani (6)	0.083	0	0.167
Jamamadi (1)	0	0.500	1
Kaingang (2)	0.500	0	0
Karitiana (13)	0	0.115	0
Kogi (4)	0	0	0.375
Maleku (3)	0	0	0
Palikur (3)	0	0.167	0.500
Parakana (1)	0	0.500	0
Piapoco (7)	0	0	0.286
Surui (24)	0	0	0
Teribe (3)	0	0.333	0
Ticuna (6)	0.167	0	0.167
Toba (4)	0.125	0.125	0.250
Waunana (3)	0	0.167	0.500
Wayuu (11)	0.056	0	0.056
Wichi (5)	0.200	0	0.300
Yaghan (4)	0.125	0.125	0.250
Total (106)	**0.048***	**0.053***	**0.142***

* Weighted average

	Mean (lowland)	Median (lowland)	95% c.i.	Mean (highland)
SP100 (rs13411586 C)	0.0656	0.0652	(0.0481; 0.0854)	0.3968
SP100 (rs9678342 C)	0.0637	0.0637	(0.0466; 0.0838)	0.3809
SP100 (rs7582700 T)	0.0965	0.0963	(0.0745; 0.1196)	0.3889
DUOX2 (rs269866 G)	0.0714	0.0714	(0.0528; 0.0916)	0.4206
CLC (rs440191 A)	0.1332	0.1335	(0.1071; 0.1599)	0.4524

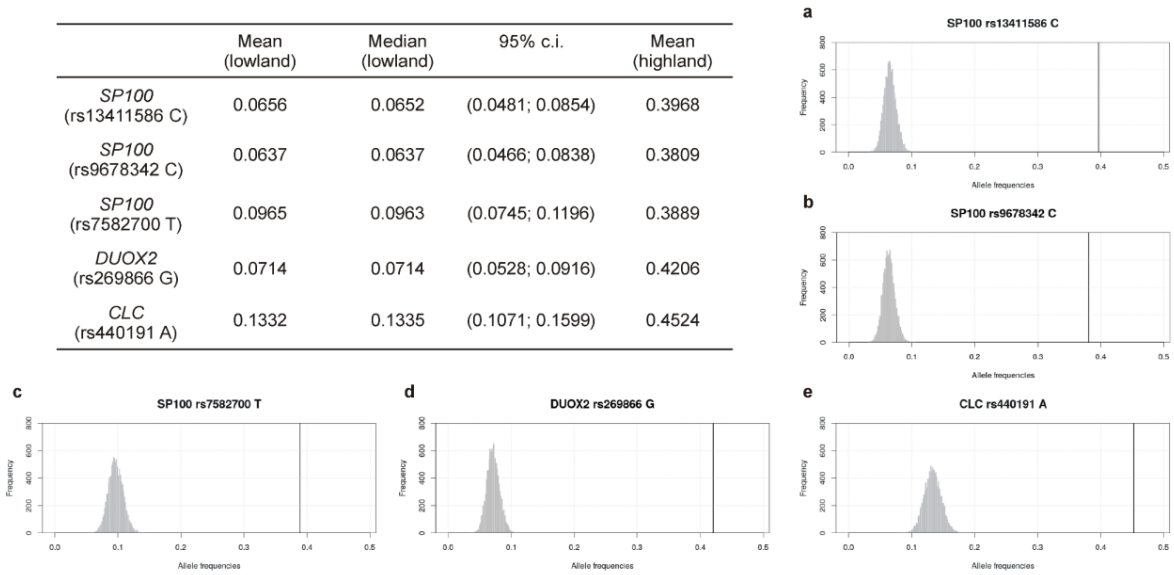

Figure 3.5 – Bootstrap simulations. (A) Table showing average values (mean and median) of the allele frequencies for each SNP in lowland populations, as well as, their 95% confidence intervals obtained by simulation, and the average allele frequency of the candidate variant in highland populations. (B–F). Distribution of allele frequencies obtained by 10,000 simulations for lowlanders considering all markers in putative selection. The corresponding average allele frequencies observed for highland populations are represented by black vertical lines.

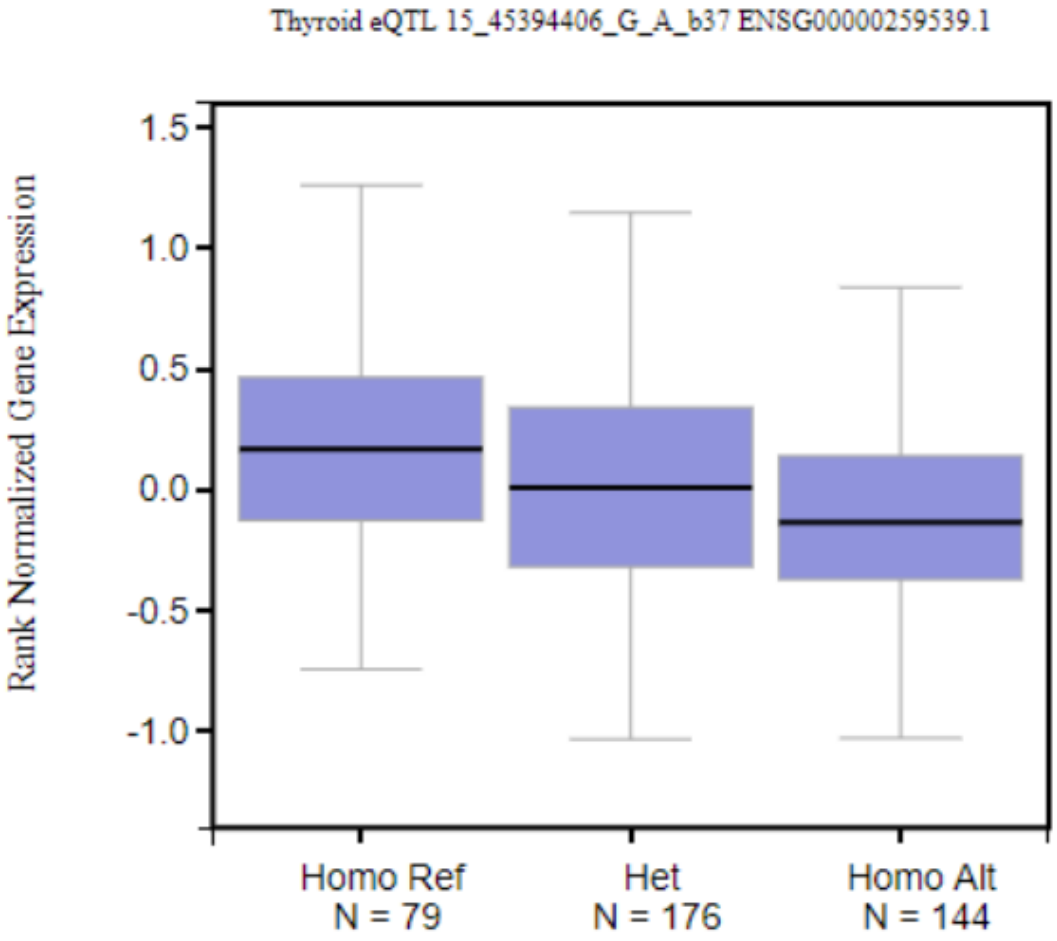

Figure 3.6 – Differential expression of the *DUOX2* putatively selected allele (rs269866 G) for a differential effect in gene expression in the thyroid tissue <(https://www.gtexportal.org/home/>, accessed 26/03/2018).

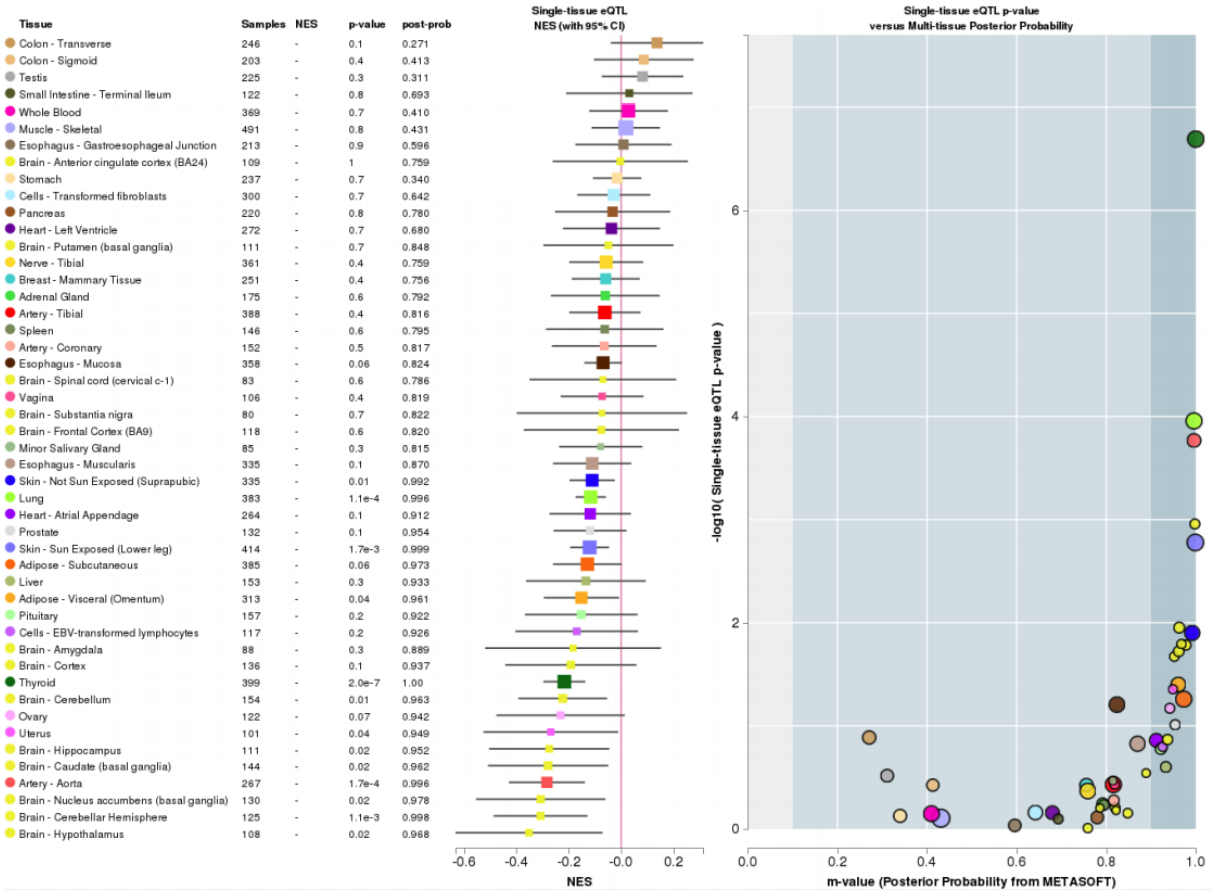

Figure 3.7 – Posterior probabilities of *DUOX2* putatively selected allele (rs269866G) for a differential effect in gene expression in multiple tissues (<https://www.gtexportal.org/home/>, accessed 26/03/2018).

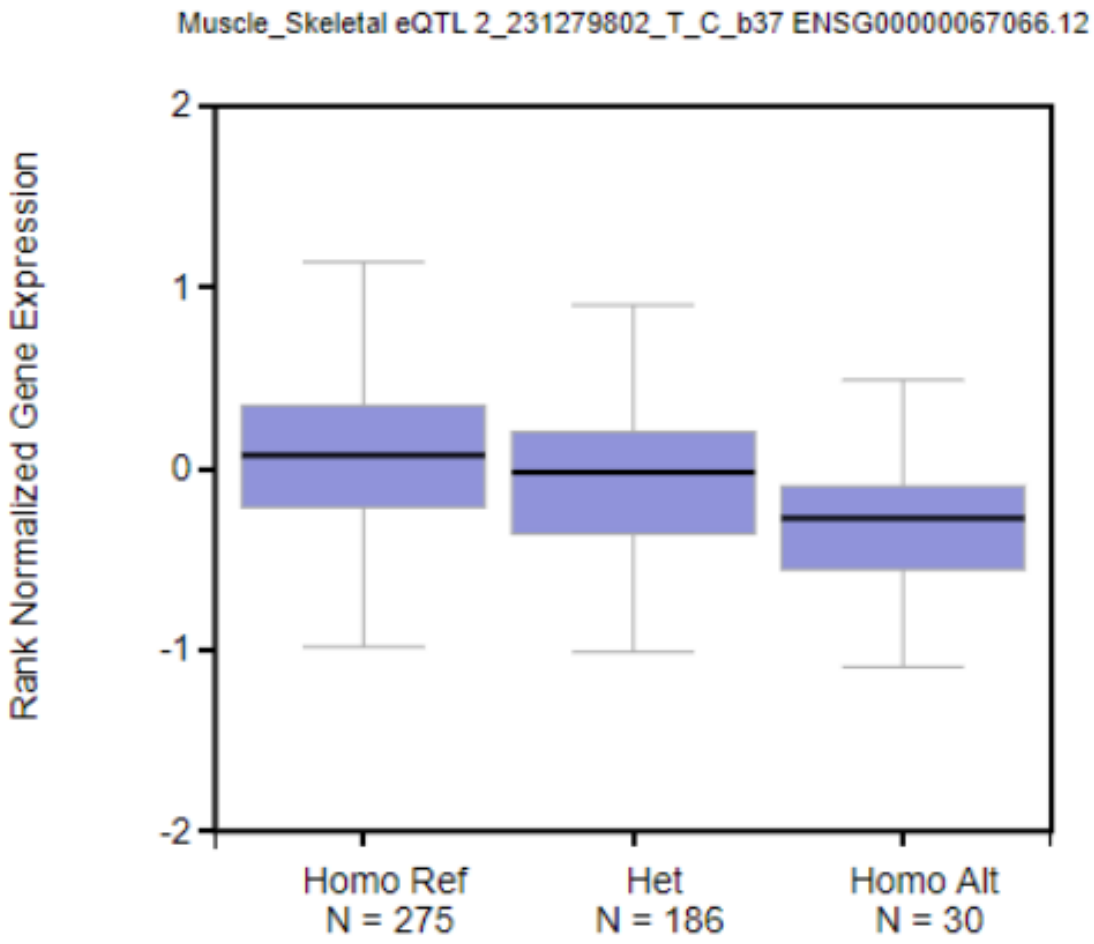

Figure 3.8 – Differential expression of the *SP100* putatively selected allele (rs13411586C) in the skeletal muscle (<https://www.gtexportal.org/home/>, accessed 26/03/2018).

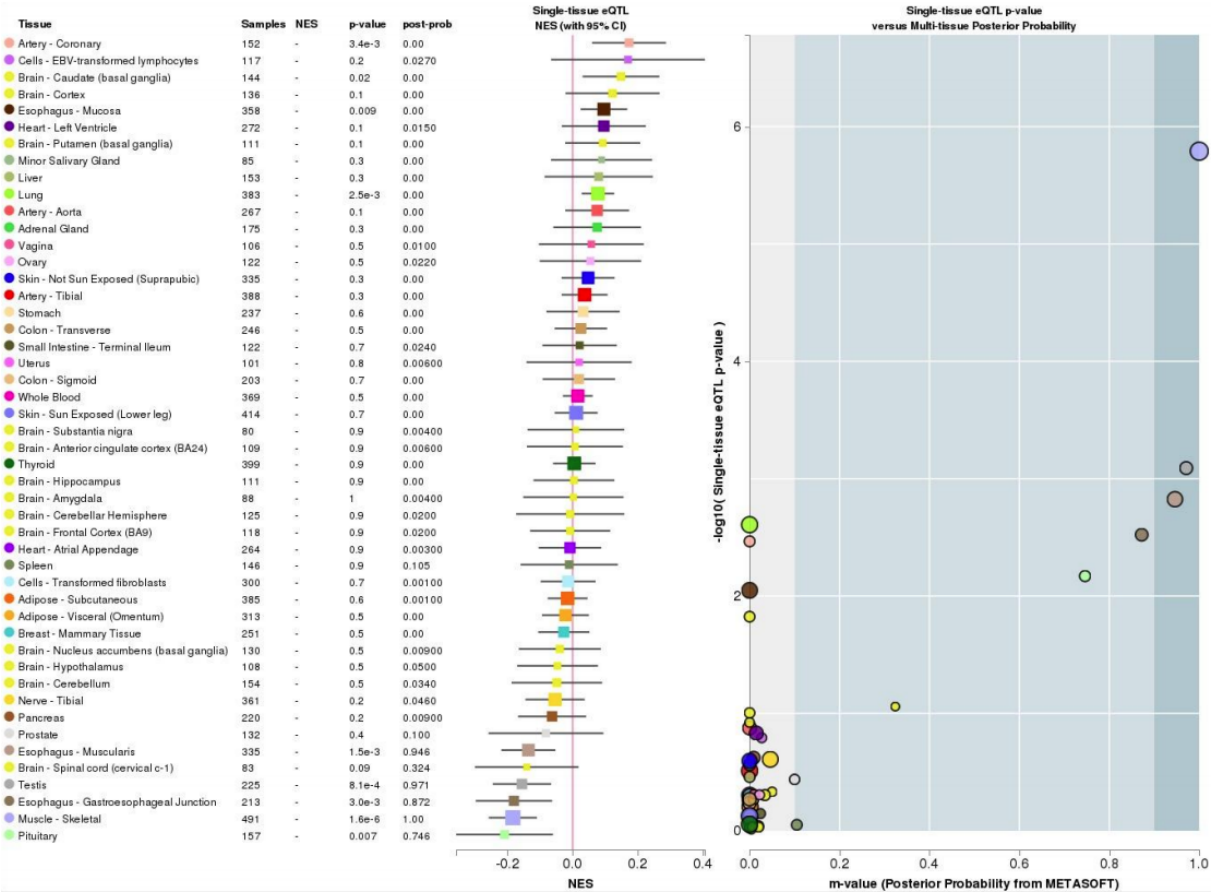

Figure 3.9 – Posterior probabilities of *SP100* putatively selected allele(rs13411586C) for a differential effect in gene expression in multiple tissues (<https://www.gtexportal.org/home/>, accessed 26/03/2018).

Capítulo 4

CONSIDERAÇÕES GERAIS

Neste trabalho, analisamos a influência da seleção natural em populações nativas americanas de duas ecorregiões principais: Floresta Amazônica e Andes.

Por meio de múltiplos testes de seleção positiva, análise de enriquecimento gênico, anotação dos genes e vias candidatas e metanálises, hipotetizamos que os genes e vias candidatas à seleção na Amazônia tiveram como pressão seletiva o estilo de vida caçador-coletor, sendo a procura de novidades, o período de escassez alimentícia e a resposta a de patógenos presentes neste ambiente os principais fatores seletivos. Neste contexto, alguns dos traços que outrora foram importantes para a subsistência destas populações podem hoje apresentar um efeito adverso dada a mudança do estilo de vida local, como a introdução de alimentos industrializados e o modo de vida sedentário. Por exemplo, as rotas para o sistema de recompensa do cérebro, que poderiam favorecer maior disposição à procura de novidades (exposição ao risco), podem hoje estar relacionados com os elevados índices de alcoolismo nestas populações. De forma similar, genes que possibilitaram a subsistência em períodos de escassez alimentícia dado à sazonalidade das florestas tropicais, podem estar relacionados com o aumento da obesidade e doenças relacionadas à síndrome metabólica. Em relação ao sistema imunológico, detectamos genes relacionados à seleção de linfócitos T, e genes relacionados à resposta sorológica ao agente transmissor da doença de Chagas, *Trypanosoma cruzi* (*e.g. PPP3CA* e *DYNC1I1*). Dada à pequena incidência da doença de Chagas nas populações amazônicas, quando comparado à populações andinas, é possível que os nativos amazônicos tenham desenvolvido um mecanismo de defesa ao protozoário *Trypanosoma cruzi*. Torna-se interessante, portanto, a realização de estudos funcionais com os genes aqui identificados para avaliar a taxa de infecção do protozoário.

Nas populações andinas, identificamos cinco genes candidatos principais – *DUOX2*, *SP100*, *CLC*, *TMEM38B*, *AS3MT* – pelo método PBS, corroborados por sua vez pelo método XP-EHH. Nossos resultados apontaram um mecanismo complexo de adaptação

nos nativos andinos, que envolveu os genes *CLC*, *SP100* e *DUOX2*, não encontrados em estudos prévios de adaptação ao ambiente andino. Considerando a diversidade de rotas evidenciadas relacionadas a adaptação exclusiva ao altiplano andino, inferimos que os nativos andinos se adaptaram a esse ambiente de maneiras diferentes e usando estratégias moleculares distintas das de outros povos que vivem em grandes altitudes.

Em conclusão, nosso trabalho identificou novos genes candidatos tanto na região da floresta amazônica quanto nas terras altas andinas, que muito provavelmente tiveram um papel essencial para adaptação local das populações nativas americanas na Amazônia e nos Andes. Os resultados desta tese abrem portas para outras pesquisas importantes, incluindo estudos funcionais, e levantam uma preocupação no que se refere à saúde pública indígena.

5 Resumo

O continente Americano foi povoado há aproximadamente 15.000 anos, e rapidamente os primeiros nativos americanos se dispersaram da América do Norte à América do Sul, passando por uma variedade de ecorregiões distintas, entre elas, a floresta tropical amazônica e as terras altas andinas. Ambas as regiões apresentam desafios para a subsistência humana. A floresta amazônica é úmida e apresenta baixa penetração solar, elevados índices de patógenos e agentes transmissores, bem como períodos de escassez de alimento humano, ao passo que as terras altas andinas são caracterizadas pela baixa concentração disponível de oxigênio, frio intenso e maior intensidade de radiação UV. Desta forma, as populações nativas nessas ecorregiões poderiam possuir variantes genéticas que favoreceram sua subsistência nestes ambientes, frente aos desafios mencionados. A fim de avaliar a influência da seleção natural nestes ambientes, utilizamos 285 indivíduos nativos sul americanos (222 terras baixas e 63 terras altas). Para detecção de seleção positiva utilizamos métodos baseados em haplótipos, como iHS e XP-EHH, ou baseados na diferenciação populacional via Fst, como PBS. Adicionalmente, utilizamos os resultados dos testes de seleção para aplicar testes de enriquecimento gênico, análise de eQTL in silico, anotação funcional dos genes candidatos, simulações demográficas, e realizamos uma metanálise dos resultados já publicados referentes às populações nativas de floresta tropical de diversos continentes. Como resultado, na região amazônica, identificamos genes e vias de sinalização candidatas relacionadas ao metabolismo energético, vias cardiovasculares, defesa imunológica e comportamento *Novelty seeking*. Na região andina, identificamos genes candidatos com funções essenciais ao metabolismo em situações de estresse promovida pela hipóxia, mostrando que os nativos andinos parecem apresentar vias alternativas de adaptação ao altiplano quando comparados com outras populações de altitude dos demais continentes. O presente trabalho apresenta dados inéditos relativos à adaptação dos nativos americanos a duas importantes ecorregiões no continente, evidenciando que rotas metabólicas antes importantes para a exploração e sobrevivência ao ambiente, hoje apresentam grande impacto no perfil epidemiológico dessas populações.

Palavras-chave: seleção natural, nativos americanos, adaptação local, genômica.

6 Abstract

The American continent was peopled approximately 15,000 years ago.he first Native Americans quickly spread out from North to South America, passing through various distinct ecoregions, including the Amazon rainforest and the Andean highlands. Both regions present challenges to human livelihood. The Amazon rainforest is humid with low solar penetration, high pathogenicity, and periods of scarcity of human food. On the other hand, Andean highlands are known for their low oxygen concentration, intense cold, and high UV radiation intensity. Therefore, native populations from such environments could have genetic variants that favored their subsistence in these enco.. To assess the influence of natural selection in these ecorregions, we used 285 Native South American individuals (222 from lowlands and 63 from highlands). To detect positive selection, we applied either haplotype-based methods, such as iHS and XP-EHH, or methods based on population differentiation via Fst, such as PBS. Furthermore, we used the results of the selection tests to apply gene enrichment tests, in silico eQTL analysis, functional annotation of the candidate genes, demographic simulations, and we carried out a metanalysis from already published data on native populations from tropical forests of different continents. As a result, in the Amazon region, we identified candidate genes and signaling pathways related to energy metabolism, cardiovascular pathways, immune defense, and novelty seeking behavior. In the Andean region, we have identified candidate genes with essential functions to metabolism in stress situations promoted by hypoxia, showing that Andean natives seem to have alternative ways for adapting to the altiplano when compared to other altitude populations from distinct continents. The present work reveals unprecedented data related to the adaptation of Native Americans to two leading ecoregions in the continent, showing that metabolic routes that were previously important to the environment exploration and survival, today play a big role in the epidemiological profile of these populations.

Keywords: natural selection, native americans, local adaptation, genomics.

REFERÊNCIAS BIBLIOGRÁFICAS

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, v. 526, n. 7571, p. 68–74, out. 2015. ISSN 1476-4687.

ACKERMAN, S. J. et al. Charcot-Leyden Crystal Protein (Galectin-10) Is Not a Dual Function Galectin with Lysophospholipase Activity but Binds a Lysophospholipase Inhibitor in a Novel Structural Fashion. *Journal of Biological Chemistry*, v. 277, n. 17, p. 14859–14868, abr. 2002. ISSN 0021-9258, 1083-351X. Disponível em: <http://www.jbc.org/lookup/doi/10.1074/jbc.M200221200>.

ACUÑA-ALONZO, V. et al. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. *Human Molecular Genetics*, v. 19, n. 14, p. 2877–2885, jul. 2010. ISSN 1460-2083.

AGUILAR, H. M. et al. Chagas disease in the Amazon Region. *Memórias do Instituto Oswaldo Cruz*, v. 1021, n. 1, p. 47–55, 2007.

ALBAKRI, A. Nutritional deficiency cardiomyopathy: A review and pooled analysis of pathophysiology, diagnosis and clinical management. *Research and Review Insights*, v. 3, n. 1, 2019. ISSN 25152637. Disponível em: <https://www.oatext.com/nutritional-deficiency-cardiomyopathy-a-review-and-pooled-analysis-of-pathophysiology-diagnosis-and-clinical-management.php>.

ALDENDERFER, M. Peopling the Tibetan Plateau: Insights from Archaeology. *High Altitude Medicine & Biology*, v. 12, n. 2, p. 141–147, jun. 2011. ISSN 1527-0297, 1557-8682. Disponível em: <http://www.liebertpub.com/doi/10.1089/ham.2010.1094>.

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. ISSN 1549-5469.

AMORIM, C. E. G. et al. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PloS One*, v. 10, n. 4, p. e0121557, 2015. ISSN 1932-6203.

AMORIM, C. E. G. et al. Genetic signature of natural selection in first Americans. *Proceedings of the National Academy of Sciences*, v. 114, n. 9, p. 2195–2199, fev. 2017. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/lookup/doi/10.1073/pnas.1620541114>.

ANTONELLI, A. et al. Amazonia is the primary source of Neotropical biodiversity. *Proceedings of the National Academy of Sciences*, v. 115, n. 23, p. 6034–6039, jun. 2018. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/lookup/doi/10.1073/pnas.1713819115>.

ARAÚJO, A. et al. Paleoparasitology of Chagas disease: a review. *Memórias do Instituto Oswaldo Cruz*, v. 104, n. suppl 1, p. 9–16, jul. 2009. ISSN 0074-0276. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762009000900004&lng=en&tlng=en>.

ARAÚJO, W. S. de et al. Nocturnal Mosquitoes of Pará State in the Brazilian Amazon: Species Composition, Habitat Segregation, and Seasonal Variation. *Journal of Medical Entomology*, p. tjaa103, jun. 2020. ISSN 0022-2585, 1938-2928. Disponível em: <https://academic.oup.com/jme/advance-article/doi/10.1093/jme/tjaa103/5850337>.

ARIAS, L. et al. High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. *American Journal of Physical Anthropology*, v. 165, n. 2, p. 238–255, 2018. ISSN 1096-8644.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, v. 25, n. 1, p. 25–29, maio 2000. ISSN 1061-4036.

ASTLE, W. J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, v. 167, n. 5, p. 1415–1429.e19, nov. 2016. ISSN 00928674. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0092867416314635>.

BAILEY, R. C. et al. Hunting and Gathering in Tropical Rain Forest: Is It Possible? *American Anthropologist*, v. 91, n. 1, p. 59–82, mar. 1989. ISSN 0002-7294, 1548-1433. Disponível em: <http://doi.wiley.com/10.1525/aa.1989.91.1.02a00040>.

BARGHI, N.; HERMISSON, J.; SCHLÖTTERER, C. Polygenic adaptation: a unifying framework to understand positive selection. *Nature Reviews. Genetics*, jun. 2020. ISSN 1471-0064.

BAUTISTA-ORTEGA, J. et al. Supplemental l-arginine and vitamins E and C preserve xanthine oxidase activity in the lung of broiler chickens grown under hypobaric hypoxia. *Poultry Science*, v. 93, n. 4, p. 979–988, abr. 2014. ISSN 00325791. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0032579119360900>.

BEALL, C. M. et al. Natural selection on EPAS1 (HIF2) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, v. 107, n. 25, p. 11459–11464, jun. 2010. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.1002443107>.

BECKER, N. S. A. et al. The role of GHR and IGF1 genes in the genetic determination of African pygmies' short stature. *European journal of human genetics: EJHG*, v. 21, n. 6, p. 653–658, jun. 2013. ISSN 1476-5438.

BEHR, A. A. et al. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics (Oxford, England)*, v. 32, n. 18, p. 2817–2823, set. 2016. ISSN 1367-4811.

BERGEY, C. M. et al. Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*, v. 115, n. 48, p. E11256–E11263, 2018. ISSN 1091-6490.

BERGSTRÖM, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, v. 367, n. 6484, p. eaay5012, mar. 2020. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.aay5012>.

BERSCHEMINSKI, J. et al. Sp100A is a tumor suppressor that activates p53-dependent transcription and counteracts E1A/E1B-55K-mediated transformation. *Oncogene*, v. 35, n. 24, p. 3178–3189, jun. 2016. ISSN 0950-9232, 1476-5594. Disponível em: <http://www.nature.com/articles/onc2015378>.

BHATIA, G. et al. Estimating and interpreting FST: the impact of rare variants. *Genome Research*, v. 23, n. 9, p. 1514–1521, set. 2013. ISSN 1549-5469.

BIGHAM, A. et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS genetics*, v. 6, n. 9, p. e1001116, set. 2010. ISSN 1553-7404.

BIGHAM, A. W.; LEE, F. S. Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes & Development*, v. 28, n. 20, p. 2189–2204, out. 2014. ISSN 0890-9369, 1549-5477. Disponível em: <http://genesdev.cshlp.org/lookup/doi/10.1101/gad.250167.114>.

BIGHAM, A. W. et al. Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Human Genomics*, v. 4, n. 2, p. 79–90, dez. 2009. ISSN 1479-7364.

BORDA, V. et al. *The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture*. [S.l.], 2020. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2020.01.30.916270>.

BRUTSAERT, T. D. Do high-altitude natives have enhanced exercise performance at altitude? *Applied Physiology, Nutrition, and Metabolism*, v. 33, n. 3, p. 582–592, jun. 2008. ISSN 1715-5312, 1715-5320. Disponível em: <http://www.nrcresearchpress.com/doi/10.1139/H08-009>.

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, 2019. ISSN 1362-4962.

BUROKER, N. E. et al. AKT3, ANGPTL4, eNOS3, and VEGFA associations with high altitude sickness in Han and Tibetan Chinese at the Qinghai-Tibetan Plateau. *International Journal of Hematology*, v. 96, n. 2, p. 200–213, ago. 2012. ISSN 0925-5710, 1865-3774. Disponível em: <http://link.springer.com/10.1007/s12185-012-1117-7>.

CANN, H. M. et al. A human genome diversity cell line panel. *Science (New York, N.Y.)*, v. 296, n. 5566, p. 261–262, abr. 2002. ISSN 1095-9203.

CAPELLI, J. d. C. S.; KOIFMAN, S. Avaliação do estado nutricional da comunidade indígena Parkatêjê, Bom Jesus do Tocantins, Pará, Brasil. *Cadernos de Saúde Pública*, v. 17, n. 2, p. 433–437, mar. 2001. ISSN 0102-311X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2001000200018&lng=pt&tlng=pt>.

CARDOSO, A. M.; MATTOS, I. E.; KOIFMAN, R. J. Prevalência de fatores de risco para doenças cardiovasculares na população Guaraní-Mbyá do Estado do Rio de Janeiro. *Cadernos de Saúde Pública*, v. 17, n. 2, p. 345–354, mar. 2001. ISSN 0102-311X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2001000200009&lng=pt&tlng=pt>.

CARVALHO, A. L. M.; OLIVEIRA, A. L. B. d. S.; GUIMARÃES, S. d. S. Caracterização epidemiológica das populações indígenas e do Subsistema de Saúde Indígena do Brasil: uma revisão integrativa da literatura. *Boletim Informativo Geum*, v. 5, p. 72–78, 2014.

CAVALLI-SFORZA, L. L. *African Pygmies*. [S.l.]: Academic Pr, 1986.

CHANG, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, v. 4, p. 7, 2015. ISSN 2047-217X.

CHEN, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, v. 14, p. 128, abr. 2013. ISSN 1471-2105.

CHICO, M. et al. Chagas Disease in Ecuador: Evidence for Disease Transmission in an Indigenous Population in the Amazon Region. *Memórias do Instituto Oswaldo Cruz*, v. 92, n. 3, p. 317–320, maio 1997. ISSN 0074-0276. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02761997000300002&lng=en&tlng=en>.

CLEMENT, C. et al. Origin and Domestication of Native Amazonian Crops. *Diversity*, v. 2, n. 1, p. 72–106, jan. 2010. ISSN 1424-2818. Disponível em: <http://www.mdpi.com/1424-2818/2/1/72>.

CLEMENT, C. R. et al. The domestication of Amazonia before European conquest. *Proceedings. Biological Sciences*, v. 282, n. 1812, p. 20150813, ago. 2015. ISSN 1471-2954.

CLIMER, S.; TEMPLETON, A. R.; ZHANG, W. Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis GWAS. *PLoS computational biology*, v. 10, n. 9, p. e1003766, set. 2014. ISSN 1553-7358.

CLIMER, S. et al. A custom correlation coefficient (CCC) approach for fast identification of multi-SNP association patterns in genome-wide SNPs data. *Genetic Epidemiology*, v. 38, n. 7, p. 610–621, nov. 2014. ISSN 1098-2272.

COIMBRA JR, C. E.; SANTOS, R. V. *Perfil Epidemiológico da População Indígena no Brasil: Considerações Gerais*. Porto Velho - Rondônia, 2001. 40 p. Disponível em: <http://www.cesir.unir.br/pdfs/doc3.pdf>.

COIMBRA JR, C. E. A. Human Settlements, Demographic Pattern, and Epidemiology in Lowland Amazonia: The Case of Chagas's Disease. *American Anthropologist*, v. 90, n. 1, p. 82–97, mar. 1988. ISSN 0002-7294, 1548-1433. Disponível em: <http://doi.wiley.com/10.1525/aa.1988.90.1.02a00060>.

COIMBRA JR, C. E. A.; SANTOS, R. V.; ESCOBAR, A. L. *Epidemiologia e saúde dos povos indígenas no Brasil*. Editora FIOCRUZ, 2003. ISBN 978-85-7541-261-9. Disponível em: <http://books.scielo.org/id/bsmtd>.

COLAGIURI, S.; MILLER, J. B. The 'carnivore connection'–evolutionary aspects of insulin resistance. *European Journal of Clinical Nutrition*, v. 56 Suppl 1, p. S30–35, mar. 2002. ISSN 0954-3007.

COOPER, D. Galectinomics: finding themes in complexity. *Biochimica et Biophysica Acta (BBA) - General Subjects*, v. 1572, n. 2-3, p. 209–231, set. 2002. ISSN 03044165. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304416502003100>.

CORDAIN, L. et al. Plant-animal subsistence ratios and macronutrient energy estimations in worldwide hunter-gatherer diets. *The American Journal of Clinical Nutrition*, v. 71, n. 3, p. 682–692, mar. 2000. ISSN 0002-9165.

COUNTER, S. A. et al. Assessment of the Brainstem-Mediated Stapedius Muscle Reflex in Andean Children Living at High Altitudes. *High Altitude Medicine & Biology*, v. 18, n. 1, p. 37–45, mar. 2017. ISSN 1557-8682. Disponível em: <http://www.liebertpub.com/doi/10.1089/ham.2016.0082>.

CRAWFORD, J. E. et al. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *The American Journal of Human Genetics*, v. 101, n. 5, p. 752–767, nov. 2017. ISSN 00029297. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0002929717303920>.

CROTEAU-CHONKA, D. C. et al. Expression Quantitative Trait Loci Information Improves Predictive Modeling of Disease Relevance of Non-Coding Genetic Variation. *PloS One*, v. 10, n. 10, p. e0140758, 2015. ISSN 1932-6203.

DARWIN, C. *The Origin of Species: By Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray, 1859. Disponível em: <http://darwin-online.org.uk/converted/pdf/1859_Origin_F373.pdf>.

DARWIN, C. *The Descent of Man, and Selection in Relation to Sex*. London: John Murray, 1871. Disponível em: <http://darwin-online.org.uk/converted/pdf/1871_Descent_F937.1.pdf>.

DARWIN, C.; WALLACE, A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Zoological Journal of the Linnean Society*, v. 3, n. 9, p. 45–62, ago. 1858. ISSN 19459475. Disponível em: <https://academic.oup.com/zoolinnean/article-lookup/doi/10.1111/j.1096-3642.1858.tb02500.x>.

DELANEAU, O.; MARCHINI, J.; ZAGURY, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods*, v. 9, n. 2, p. 179–181, fev. 2012. ISSN 1548-7091, 1548-7105. Disponível em: <http://www.nature.com/articles/nmeth.1785>.

DEVASAGAYAM, T. P. A. et al. Free radicals and antioxidants in human health: current status and future prospects. *The Journal of the Association of Physicians of India*, v. 52, p. 794–804, out. 2004. ISSN 0004-5772.

DIAS-FREITAS, F.; METELO-COIMBRA, C.; RONCON-ALBUQUERQUE, R. Molecular mechanisms underlying hyperoxia acute lung injury. *Respiratory Medicine*, v. 119, p. 23–28, out. 2016. ISSN 09546111. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0954611116301986>.

DILLEHAY, T. D. Probing deeper into first American studies. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, n. 4, p. 971–978, jan. 2009. ISSN 1091-6490.

EICHSTAEDT, C. A. et al. Positive selection of AS3MT to arsenic water in Andean populations. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, v. 780, p. 97–102, out. 2015. ISSN 00275107. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0027510715300269>.

EICHSTAEDT, C. A. et al. The Andean Adaptive Toolkit to Counteract High Altitude Maladaptation: Genome-Wide and Phenotypic Analysis of the Collas. *PLoS ONE*, v. 9, n. 3, p. e93314, mar. 2014. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0093314>.

ESPINOZA-NAVARRO, O. et al. Effects of Altitude on Anthropometric and Physiological Patterns in Aymara and Non-Aymara Population Between 18 and 65 Years in the Province of Parinacota Chile (3.700 masl). *International Journal of Morphology*, v. 29, n. 1, p. 34–40, mar. 2011. ISSN 0717-9502. Disponível em: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-95022011000100005&lng=en&nrm=iso&tlng=en>.

FAGUNDES, N. J. R. et al. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *American Journal of Human Genetics*, v. 82, n. 3, p. 583–592, mar. 2008. ISSN 1537-6605.

FAIRBANKS, D. J. Mendel and Darwin: untangling a persistent enigma. *Heredity*, v. 124, n. 2, p. 263–273, fev. 2020. ISSN 1365-2540. Number: 2 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41437-019-0289-9>.

FAN, S. et al. Going global by adapting local: A review of recent human adaptation. *Science (New York, N.Y.)*, v. 354, n. 6308, p. 54–59, 2016. ISSN 1095-9203.

FANG, Z. et al. *GSEApy: Gene Set Enrichment Analysis in Python*. Zenodo, 2020. Disponível em: <https://zenodo.org/record/3983639#.X2UIkXVKhhE>.

FEARNSIDE, P. M. *Destruição e Conservação da Floresta Amazônica*. 1. ed. Manaus: Editora do INPA, 2020. v. 1. Disponível em: <http://philip.inpa.gov.br/publ_livres/2019/Destruicao-v1/Destruicao_e_Conservacao_da_Floresta_Amazonica-Miolo-prova.pdf>.

FEHREN-SCHMITZ, L.; GEORGES, L. Ancient DNA reveals selection acting on genes associated with hypoxia response in pre-Columbian Peruvian Highlanders in the last 8500 years. *Scientific Reports*, v. 6, n. 1, p. 23485, mar. 2016. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep23485>.

FERREIRA, A. A. *A influência da ingestão de bebida alcoólica e transforno mentais comuns não psicóticos na pressão arterial dos indígenas Mura*. Tese (Tese) — Universidade de São Paulo, São Paulo, 2016. Disponível em: <https://www.teses.usp.br/teses/disponiveis/7/7139/tde-27042018-120745/publico/Alaidistania_Aparecida_Ferreira_Corrigida.pdf>.

FILHO, Z. A. d. S. et al. Hypertension prevalence among indigenous populations in Brazil: a systematic review with meta-analysis. *Revista da Escola de Enfermagem da USP*, v. 49, n. 6, p. 1012–1022, dez. 2015. ISSN 0080-6234. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342015000601012&lng=en&tlng=en>.

FLETCHER, N. M. et al. Nicotinamide Adenine Dinucleotide Phosphate Oxidase Is Differentially Regulated in Normal Myometrium Versus Leiomyoma. *Reproductive Sciences*, v. 21, n. 9, p. 1145–1152, set. 2014. ISSN 1933-7191, 1933-7205. Disponível em: <http://journals.sagepub.com/doi/10.1177/1933719114522552>.

FUMAGALLI, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science (New York, N.Y.)*, v. 349, n. 6254, p. 1343–1347, set. 2015. ISSN 1095-9203.

FUNAI, F. N. d. Í. *Terras Indígenas.* , [s.d.].

FUNASA. *Política Nacional de Atenção à Saúde dos Povos Indígenas*. Brasília, 2002. 40 p.

GAUTIER, M.; KLASSMANN, A.; VITALIS, R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, v. 17, n. 1, p. 78–90, jan. 2017. ISSN 1755-0998.

GAUTIER, M.; VITALIS, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, v. 28, n. 8, p. 1176–1177, abr. 2012. ISSN 1367-4803, 1460-2059. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts115>.

GE, R.-L. et al. B-type natriuretic peptide, vascular endothelial growth factor, endothelin-1, and nitric oxide synthase in chronic mountain sickness. *American Journal of Physiology-Heart and Circulatory Physiology*, v. 300, n. 4, p. H1427–H1433, abr. 2011. ISSN 0363-6135, 1522-1539. Disponível em: <https://www.physiology.org/doi/10.1152/ajpheart.00366.2010>.

GNECCHI-RUSCONE, G. A. et al. Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes-Amazonia Divide. *Molecular Biology and Evolution*, v. 36, n. 6, p. 1254–1269, 2019. ISSN 1537-1719.

GOPALAN, V. et al. The expression profiles of the galectin gene family in colorectal adenocarcinomas. *Human Pathology*, v. 53, p. 105–113, jul. 2016. ISSN 00468177. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0046817716000836>.

GOUY, A.; DAUB, J. T.; EXCOFFIER, L. Detecting gene subnetworks under selection in biological pathways. *Nucleic Acids Research*, v. 45, n. 16, p. e149, set. 2017. ISSN 1362-4962.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, v. 45, n. 6, p. 580–585, jun. 2013. ISSN 1546-1718.

GUERNIER, V.; HOCHBERG, M. E.; GUÉGAN, J.-F. Ecology drives the worldwide distribution of human diseases. *PLoS biology*, v. 2, n. 6, p. e141, jun. 2004. ISSN 1545-7885.

GUIMARÃES, L.; GRUBITS, S. Alcoolismo e violência em etnias indígenas: Uma visão crítica da situação Brasileira. *Psicologia & Sociedade*, v. 19, n. 1, p. 45–51, abr. 2007.

GÜNTHER, T.; COOP, G. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, v. 195, n. 1, p. 205–220, set. 2013. ISSN 0016-6731, 1943-2631. Disponível em: <http://www.genetics.org/lookup/doi/10.1534/genetics.113.152462>.

GUPTA, S. C. et al. Upsides and Downsides of Reactive Oxygen Species for Cancer: The Roles of Reactive Oxygen Species in Tumorigenesis, Prevention, and Therapy. *Antioxidants & Redox Signaling*, v. 16, n. 11, p. 1295–1322, jun. 2012. ISSN 1523-0864, 1557-7716. Disponível em: <http://www.liebertpub.com/doi/10.1089/ars.2011.4414>.

HAN, X. et al. Caloric Restriction Results in Phospholipid Depletion, Membrane Remodeling, and Triacylglycerol Accumulation in Murine Myocardium †. *Biochemistry*, v. 43, n. 49, p. 15584–15594, dez. 2004. ISSN 0006-2960, 1520-4995. Disponível em: <https://pubs.acs.org/doi/10.1021/bi048307o>.

HANCOCK, A. M.; RIENZO, A. D. Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. *Annual Review of Anthropology*, v. 37, n. 1, p. 197–217, out. 2008. ISSN 0084-6570, 1545-4290. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.anthro.37.081407.085141>.

HANCOCK, A. M. et al. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, v. 107, n. Supplement_2, p. 8924–8930, maio 2010. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.0914625107>.

HARRISON, G. F. et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. *Nature Ecology & Evolution*, v. 3, n. 8, p. 1253–1264, 2019. ISSN 2397-334X.

HART, T. B.; HART, J. A. The ecological basis of hunter-gatherer subsistence in African Rain Forests: The Mbuti of Eastern Zaire. v. 14, n. 1, 1986.

HASLAM, D. W.; JAMES, W. P. T. Obesity. *Lancet (London, England)*, v. 366, n. 9492, p. 1197–1209, out. 2005. ISSN 1474-547X.

HECKENBERGER, M.; NEVES, E. G. Amazonian Archaeology. *Annual Review of Anthropology*, v. 38, n. 1, p. 251–266, out. 2009. ISSN 0084-6570, 1545-4290. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev-anthro-091908-164310>.

HERRER, I. et al. Gene expression network analysis reveals new transcriptional regulators as novel factors in human ischemic cardiomyopathy. *BMC Medical Genomics*, v. 8, n. 1, p. 14, dez. 2015. ISSN 1755-8794. Disponível em: <http://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-015-0088-y>.

HIDRON, A. et al. Cardiac involvement with parasitic infections. *Clinical Microbiology Reviews*, v. 23, n. 2, p. 324–349, abr. 2010. ISSN 1098-6618.

HILL, W. D. et al. Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nature Communications*, v. 10, n. 1, p. 5741, 2019. ISSN 2041-1723.

HOFFECKER, J. F. et al. Beringia and the global dispersal of modern humans: Beringia and the Global Dispersal of Modern Humans. *Evolutionary Anthropology: Issues, News, and Reviews*, v. 25, n. 2, p. 64–78, mar. 2016. ISSN 10601538. Disponível em: <http://doi.wiley.com/10.1002/evan.21478>.

HSIEH, P. et al. Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations. *Molecular Biology and Evolution*, v. 34, n. 11, p. 2913–2926, nov. 2017. ISSN 1537-1719.

HUDSON, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, v. 18, n. 2, p. 337–338, fev. 2002. ISSN 1367-4803, 1460-2059. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.2.337>.

HUERTA-SÁNCHEZ, E. et al. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular Biology and Evolution*, v. 30, n. 8, p. 1877–1888, ago. 2013. ISSN 1537-1719, 0737-4038. Disponível em: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst089>.

HÜNEMEIER, T. et al. Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. *PloS One*, v. 7, n. 6, p. e38862, 2012. ISSN 1932-6203.

IGANSI, M. L.; ZATTI, C. A. GESTAÇÃO: CONHECENDO A REALIDADE DAS ALDEIAS INDÍGENAS NO BRASIL. *Brazilian Journal of Surgery and Clinical Research*, v. 23, n. 1, p. 48–52, 2018. Disponível em: <https://www.mastereditora.com.br/periodico/20180606_085304.pdf>.

IGLESIAS, M. M. et al. Galectin-1 from ovine placenta–amino-acid sequence, physicochemical properties and implications in T-cell death. *European Journal of Biochemistry*, v. 252, n. 3, p. 400–407, mar. 1998. ISSN 0014-2956.

IWEN, K. A.; OELKRUG, R.; BRABANT, G. Effects of thyroid hormones on thermogenesis and energy partitioning. *Journal of Molecular Endocrinology*, v. 60, n. 3, p. R157–R170, abr. 2018. ISSN 0952-5041, 1479-6813. Disponível em: <https://jme.bioscientifica.com/view/journals/jme/60/3/JME-17-0319.xml>.

JACOVAS, V. C. et al. Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Scientific Reports*, v. 8, n. 1, p. 12733, 2018. ISSN 2045-2322.

JACOVAS, V. C. et al. Genetic Variations in the TP53 Pathway in Native Americans Strongly Suggest Adaptation to the High Altitudes of the Andes. *PLOS ONE*, v. 10, n. 9, p. e0137823, set. 2015. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0137823>.

JARVIS, J. P. et al. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics*, v. 8, n. 4, p. e1002641, 2012. ISSN 1553-7404.

JEONG, C.; RIENZO, A. D. Adaptations to local environments in modern human populations. *Current Opinion in Genetics & Development*, v. 29, p. 1–8, dez. 2014. ISSN 0959437X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0959437X14000641>.

JOBLING, M. A. et al. *Human Evolutionary Genetics*. 2. ed. New York: Garland Science, 2014. ISBN 978-0-8153-4148-2.

JULIAN, C. G.; MOORE, L. G. Human Genetic Adaptation to High Altitude: Evidence from the Andes. *Genes*, v. 10, n. 2, 2019. ISSN 2073-4425.

KAHN, S. E.; HULL, R. L.; UTZSCHNEIDER, K. M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, v. 444, n. 7121, p. 840–846, dez. 2006. ISSN 1476-4687.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, v. 28, n. 1, p. 27–30, jan. 2000. ISSN 0305-1048.

KANEHISA, M. et al. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, v. 47, n. D1, p. D590–D595, 2019. ISSN 1362-4962.

KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, v. 8, n. 2, p. e1002375, 2012. ISSN 1553-7358.

KIM, M.-J. et al. Dual Oxidase 2 in Lung Epithelia Is Essential for Hyperoxia-Induced Acute Lung Injury in Mice. *Antioxidants & Redox Signaling*, v. 21, n. 13, p. 1803–1818, nov. 2014. ISSN 1523-0864, 1557-7716. Disponível em: <http://www.liebertpub.com/doi/10.1089/ars.2013.5677>.

KIM, Y.-W.; BYZOVA, T. V. Oxidative stress in angiogenesis and vascular disease. *Blood*, v. 123, n. 5, p. 625–631, jan. 2014. ISSN 0006-4971, 1528-0020. Disponível em: <https://ashpublications.org/blood/article/123/5/625/32854/Oxidative-stress-in-angiogenesis-and-vascular>.

KIMURA, M. The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, v. 63, n. 4, p. 1181–1188, ago. 1969. ISSN 0027-8424.

KLOPFENSTEIN, D. V. et al. GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, v. 8, n. 1, p. 10872, jul. 2018. ISSN 2045-2322.

KOSTER, J.; RAHMANN, S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, v. 28, n. 19, p. 2520–2522, out. 2012. ISSN 1367-4803, 1460-2059. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480>.

KREMER, R. et al. Vitamin D status and its relationship to body fat, final height, and peak bone mass in young women. *The Journal of Clinical Endocrinology and Metabolism*, v. 94, n. 1, p. 67–73, jan. 2009. ISSN 0021-972X.

KULESHOV, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, v. 44, n. W1, p. W90–97, 2016. ISSN 1362-4962.

LANGDON, E. J. O abuso de álcool entre os povos indígenas no Brasil: uma avaliação comparativa. *Tellus*, v. 5, n. 8/9, p. 103–124, 2005.

LEAL, J. L. d. S. *A AMAZÔNIA BRASILEIRA E O SEU CARÁTER TRANSNACIONAL: O APROVEITAMENTO DO SEU PATRIMÔNIO ESTRATÉGICO PARA GARANTIA DO DESENVOLVIMENTO*. Tese (Doutorado) — Universidade do Vale do Itajaí (UNIVALI), Itajaí - SC, abr. 2019. Disponível em: <https://www.univali.br/Lists/TrabalhosDoutorado/Attachments/259/TESE%20-%20JORGE%20LUIZ%20DOS%20SANTOS%20LEAL.pdf>.

LEE, C.-J. et al. Cholesterol effectively blocks entry of flavivirus. *Journal of Virology*, v. 82, n. 13, p. 6470–6480, jul. 2008. ISSN 1098-5514.

LEITE, M. S. et al. Crescimento físico e perfil nutricional da população indígena Xavánte de Sangradouro-Volta Grande, Mato Grosso, Brasil. *Cadernos de Saúde Pública*, v. 22, n. 2, p. 265–276, fev. 2006. ISSN 0102-311X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2006000200004&lng=pt&tlng=pt>.

LEWONTIN, R. C.; KRAKAUER, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, v. 74, n. 1, p. 175–195, maio 1973. ISSN 0016-6731.

LIAO, Y. et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, v. 47, n. W1, p. W199–W205, jul. 2019. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article/47/W1/W199/5494758>.

LIMA, C. M. d. A. et al. Iniciação sexual, gestação, parto e puerpério em comunidades indígenas do brasil: uma breve revisão integrativa. *Revisa de Saúde Pública de Mato Grosso do Sul*, v. 1, n. 1, p. 86–101, 2018. Disponível em: <https://revista.saude.ms.gov.br/index.php/rspms/article/view/11>.

LIN, S.-C. et al. High Immunoreactivity of DUOX2 Is Associated With Poor Response to Preoperative Chemoradiation Therapy and Worse Prognosis in Rectal Cancers. *Journal of Cancer*, v. 8, n. 14, p. 2756–2764, 2017. ISSN 1837-9664. Disponível em: <http://www.jcancer.org/v08p2756.htm>.

LINDHOLM, M. E.; RUNDQVIST, H. Skeletal muscle hypoxia-inducible factor-1 and exercise: Skeletal muscle hypoxia-inducible factor-1 and exercise. *Experimental Physiology*, v. 101, n. 1, p. 28–32, jan. 2016. ISSN 09580670. Disponível em: <http://doi.wiley.com/10.1113/EP085318>.

LLAMAS, B. et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances*, v. 2, n. 4, p. e1501385, abr. 2016. ISSN 2375-2548.

LOMBARDO, U. et al. Early Holocene crop cultivation and landscape modification in Amazonia. *Nature*, v. 581, n. 7807, p. 190–193, 2020. ISSN 1476-4687.

LOPEZ, M. et al. Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. *Current biology: CB*, v. 29, n. 17, p. 2926–2935.e4, 2019. ISSN 1879-0445.

LOSOS, J. B. et al. *The Princeton Guide to Evolution*. 1. ed. Princeton, New Jersey: Princeton University Press, 2014. ISBN 978-0-691-14977-6.

LOURENÇO, A. E. P. et al. Nutrition transition in Amazonia: obesity and socioeconomic change in the Suruí Indians from Brazil. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, v. 20, n. 5, p. 564–571, out. 2008. ISSN 1520-6300.

LU, D. et al. Ancestral Origins and Genetic History of Tibetan Highlanders. *The American Journal of Human Genetics*, v. 99, n. 3, p. 580–594, set. 2016. ISSN 00029297. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0002929716302737>.

LUCA, F.; PERRY, G. H.; RIENZO, A. D. Evolutionary adaptations to dietary changes. *Annual Review of Nutrition*, v. 30, p. 291–314, ago. 2010. ISSN 1545-4312.

MACARTHUR, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, v. 45, n. D1, p. D896–D901, 2017. ISSN 1362-4962.

MACFIE, T. S. et al. DUOX2 and DUOXA2 Form the Predominant Enzyme System Capable of Producing the Reactive Oxygen Species H2O2 in Active Ulcerative Colitis and are Modulated by 5-Aminosalicylic Acid:. *Inflammatory Bowel Diseases*, v. 20, n. 3, p. 514–524, mar. 2014. ISSN 1078-0998. Disponível em: <https://academic.oup.com/ibdjournal/article/20/3/514-524/4579005>.

MALÉCOT, G. *Les mathématiques de l'hérédité*. Paris: Masson & Cie, 1948.

MALLICK, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, v. 538, n. 7624, p. 201–206, out. 2016. ISSN 1476-4687.

MARCONI, C.; MARZORATI, M.; CERRETELLI, P. Work Capacity of Permanent Residents of High Altitude. *High Altitude Medicine & Biology*, v. 7, n. 2, p. 105–115, jun. 2006. ISSN 1527-0297, 1557-8682. Disponível em: <http://www.liebertpub.com/doi/10.1089/ham.2006.7.105>.

MARINHO, G. L. et al. Mortalidade infantil de indígenas e não indígenas nas microrregiões do Brasil. *Revista Brasileira de Enfermagem*, v. 72, n. 1, p. 57–63, fev. 2019. ISSN 1984-0446, 0034-7167. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71672019000100057&tlng=pt>.

MENDES, M. et al. The history behind the mosaic of the Americas. *Current Opinion in Genetics & Development*, v. 62, p. 72–77, jun. 2020. ISSN 0959437X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0959437X20301076>.

MIGLIANO, A. B.; VINICIUS, L.; LAHR, M. M. Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences of the United States of America*, v. 104, n. 51, p. 20216–20219, dez. 2007. ISSN 1091-6490.

MILLER, J. C.; COLAGIURI, S. The carnivore connection: dietary carbohydrate in the evolution of NIDDM. *Diabetologia*, v. 37, n. 12, p. 1280–1286, dez. 1994. ISSN 0012-186X.

MIN, H. J. et al. ROS-dependent HMGB1 secretion upregulates IL-8 in upper airway epithelial cells under hypoxic condition. *Mucosal Immunology*, v. 10, n. 3, p. 685–694, maio 2017. ISSN 1933-0219, 1935-3456. Disponível em: <http://www.nature.com/articles/mi201682>.

Ministério da Saúde. *Anais do Seminário sobre Alcoolismo e Vulnerabilidade às DST/AIDS entre os Povos Indígenas da Macrorregião Sul, Sudeste e Mato Grosso do Sul*. [S.l.: s.n.], 2001.

MISSAGGIA, B. O. et al. Adaptation and co-adaptation of skin pigmentation and vitamin D genes in native Americans. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, v. 184, n. 4, p. 1060–1077, dez. 2020. ISSN 1552-4876.

MOORE, L. G. Human Genetic Adaptation to High Altitude. *High Altitude Medicine & Biology*, v. 2, n. 2, p. 257–279, jun. 2001. ISSN 1527-0297, 1557-8682. Disponível em: <http://www.liebertpub.com/doi/10.1089/152702901750265341>.

MOORE, L. G. Human Genetic Adaptation to High Altitudes: Current Status and Future Prospects. *Quaternary International: The Journal of the International Union for Quaternary Research*, v. 461, p. 4–13, dez. 2017. ISSN 1040-6182.

MOREIRA, H. M. *A importância da Amazônia na definição da posição brasileira no regime internacional de mudanças climáticas*. 2009. Disponível em: <https://www.fclar.unesp.br/Home/Pesquisa/GruposdePesquisa/NPPA/C.E_Helena_MargaridoMoreiraHelena-LASA.pdf>.

MORENO-MAYAR, J. V. et al. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, v. 553, n. 7687, p. 203–207, jan. 2018. ISSN 0028-0836, 1476-4687. Disponível em: <http://www.nature.com/articles/nature25173>.

MORENO-MAYAR, J. V. et al. Early human dispersals within the Americas. *Science (New York, N.Y.)*, v. 362, n. 6419, 2018. ISSN 1095-9203.

MURDOCK, G. P. Ethnographic Atlas: A Summary. *Ethnology*, v. 6, n. 2, p. 109, abr. 1967. ISSN 00141828. Disponível em: <https://www.jstor.org/stable/3772751?origin=crossref>.

NIELSEN, R. et al. Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, v. 8, n. 11, p. 857–868, nov. 2007. ISSN 1471-0056, 1471-0064. Disponível em: <http://www.nature.com/articles/nrg2187>.

NOVEMBRE, J.; HAN, E. Human population structure and the adaptive response to pathogen-induced selection pressures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 367, n. 1590, p. 878–886, mar. 2012. ISSN 0962-8436, 1471-2970. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rstb.2011.0305>.

OBACZ, J. et al. Cross-talk between HIF and p53 as mediators of molecular responses to physiological and genotoxic stresses. *Molecular Cancer*, v. 12, n. 1, p. 93, 2013. ISSN 1476-4598. Disponível em: <http://molecular-cancer.biomedcentral.com/articles/10.1186/1476-4598-12-93>.

OHENJO, N. et al. Health of Indigenous people in Africa. *Lancet (London, England)*, v. 367, n. 9526, p. 1937–1946, jun. 2006. ISSN 1474-547X.

OLIVEIRA, W. C. d. *Caçadores Coletores na Amazônia: eles existem*. Tese (Tese) — Universidade de São Paulo, São Paulo, 2007. Disponível em: <https://teses.usp.br/teses/disponiveis/71/71131/tde-25022008-152739/publico/tdeWesley.pdf>.

PEDERSEN, M. W. et al. Postglacial viability and colonization in North America's ice-free corridor. *Nature*, v. 537, n. 7618, p. 45–49, 2016. ISSN 1476-4687.

PENG, Y. et al. Genetic Variations in Tibetan Populations and High-Altitude Adaptation at the Himalayas. *Molecular Biology and Evolution*, v. 28, n. 2, p. 1075–1081, fev. 2011. ISSN 0737-4038, 1537-1719. Disponível em: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msq290>.

PERRY, G. H.; DOMINY, N. J. Evolution of the human pygmy phenotype. *Trends in Ecology & Evolution*, v. 24, n. 4, p. 218–225, abr. 2009. ISSN 0169-5347.

POLYAK, K. et al. A model for p53-induced apoptosis. *Nature*, v. 389, n. 6648, p. 300–305, set. 1997. ISSN 0028-0836, 1476-4687. Disponível em: <http://www.nature.com/articles/38525>.

PORTER, L. M. et al. Hypoxia causes IL-8 secretion, Charcot Leyden crystal formation, and suppression of corticosteroid-induced apoptosis in human eosinophils. *Clinical & Experimental Allergy*, v. 47, n. 6, p. 770–784, jun. 2017. ISSN 09547894. Disponível em: <http://doi.wiley.com/10.1111/cea.12877>.

POSTH, C. et al. Reconstructing the Deep Population History of Central and South America. *Cell*, v. 175, n. 5, p. 1185–1197.e22, 2018. ISSN 1097-4172.

POTTER, B. A. et al. Current evidence allows multiple models for the peopling of the Americas. *Science Advances*, v. 4, n. 8, p. eaat5473, ago. 2018. ISSN 2375-2548. Disponível em: <https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aat5473>.

POUYET, F. et al. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, v. 7, 2018. ISSN 2050-084X.

PRITCHARD, J. K.; PICKRELL, J. K.; COOP, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology: CB*, v. 20, n. 4, p. R208–215, fev. 2010. ISSN 1879-0445.

PRITCHARD, J. K.; RIENZO, A. D. Adaptation – not by sweeps alone. *Nature Reviews Genetics*, v. 11, n. 10, p. 665–667, out. 2010. ISSN 1471-0056, 1471-0064. Disponível em: <http://www.nature.com/articles/nrg2880>.

PRUGNOLLE, F.; MANICA, A.; BALLOUX, F. Geography predicts neutral genetic diversity of human populations. *Current biology: CB*, v. 15, n. 5, p. R159–160, mar. 2005. ISSN 0960-9822.

RACIMO, F.; BERG, J. J.; PICKRELL, J. K. Detecting Polygenic Adaptation in Admixture Graphs. *Genetics*, v. 208, n. 4, p. 1565–1584, abr. 2018. ISSN 0016-6731, 1943-2631. Disponível em: <http://www.genetics.org/lookup/doi/10.1534/genetics.117.300489>.

RADEMAKER, K. et al. Paleoindian settlement of the high-altitude Peruvian Andes. *Science*, v. 346, n. 6208, p. 466–469, out. 2014. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.1258260>.

RAGHAVAN, M. et al. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science (New York, N.Y.)*, v. 349, n. 6250, p. aab3884, ago. 2015. ISSN 1095-9203.

RAIMOND, J. et al. The second intron of the human galectin-3 gene has a strong promoter activity down-regulated by p53. *FEBS Letters*, v. 363, n. 1-2, p. 165–169, abr. 1995. ISSN 00145793. Disponível em: <http://doi.wiley.com/10.1016/0014-5793%2895%2900310-6>.

RATNAM, J. et al. When is a 'forest' a savanna, and why does it matter?: When is a 'forest' a savanna. *Global Ecology and Biogeography*, v. 20, n. 5, p. 653–660, set. 2011. ISSN 1466822X. Disponível em: <http://doi.wiley.com/10.1111/j.1466-8238.2010.00634.x>.

RAVEL, C. et al. Polymorphisms in DLGH1 and LAMC1 in Mayer-Rokitansky-Kuster-Hauser syndrome. *Reproductive Biomedicine Online*, v. 24, n. 4, p. 462–465, abr. 2012. ISSN 1472-6491.

REECE, J. B. et al. *Campbell Biology*. 10. ed. Glenview,: Pearson, 2013.

REES, J. S.; CASTELLANO, S.; ANDRÉS, A. M. The Genomics of Human Local Adaptation. *Trends in genetics: TIG*, v. 36, n. 6, p. 415–428, jun. 2020. ISSN 0168-9525.

REFOYO-MARTÍNEZ, A. et al. *How robust are cross-population signatures of polygenic adaptation in humans?* [S.l.], 2020. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2020.07.13.200030>.

REICH, D. et al. Reconstructing Native American population history. *Nature*, v. 488, n. 7411, p. 370–374, ago. 2012. ISSN 0028-0836, 1476-4687. Disponível em: <http://www.nature.com/articles/nature11258>.

REYNOLDS, A. W. et al. Comparing signals of natural selection between three Indigenous North American populations. *Proceedings of the National Academy of Sciences of the United States of America*, v. 116, n. 19, p. 9312–9317, 2019. ISSN 1091-6490.

REYNOLDS, J.; WEIR, B. S.; COCKERHAM, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, v. 105, n. 3, p. 767–779, nov. 1983. ISSN 0016-6731.

ROSENBERG, N. A. et al. Genetic structure of human populations. *Science (New York, N.Y.)*, v. 298, n. 5602, p. 2381–2385, dez. 2002. ISSN 1095-9203.

SABETI, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, v. 419, n. 6909, p. 832–837, out. 2002. ISSN 0028-0836, 1476-4687. Disponível em: <http://www.nature.com/articles/nature01140>.

SABETI, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, v. 449, n. 7164, p. 913–918, out. 2007. ISSN 1476-4687.

SALAS, A. The natural selection that shapes our genomes. *Forensic Science International. Genetics*, v. 39, p. 57–60, 2019. ISSN 1878-0326.

SALMEEN, A.; PARK, B. O.; MEYER, T. The NADPH oxidases NOX4 and DUOX2 regulate cell cycle entry via a p53-dependent pathway. *Oncogene*, v. 29, n. 31, p. 4473–4484, ago. 2010. ISSN 0950-9232, 1476-5594. Disponível em: <http://www.nature.com/articles/onc2010200>.

SALVO, V. L. M. A. d. et al. Perfil metabólico e antropométrico dos Suyá: Parque Indígena do Xingu, Brasil Central. *Revista Brasileira de Epidemiologia*, v. 12, n. 3, p. 458–468, set. 2009. ISSN 1415-790X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2009000300014&lng=pt&tlng=pt>.

SAMSON, S. L.; GARBER, A. J. Metabolic syndrome. *Endocrinology and Metabolism Clinics of North America*, v. 43, n. 1, p. 1–23, mar. 2014. ISSN 1558-4410.

SANTIAGO, A. L. C. *PLANO DE INTERVENÇÃO PARA PREVENÇÃO DA GRAVIDEZ NA ADOLESCÊNCIA ENTRE OS 11 E 15 ANOS NO PBSI FORMOSO DO ARAGUAIA*. Tese (Monografia) — Universidade Federal de São Paulo, São Paulo, 2017. Disponível em: <https://ares.unasus.gov.br/acervo/html/ARES/12116/1/110560.pdf>.

SANTOS, R.; COIMBRA JR, C. E. A. *Saúde e povos indígenas. Rio de Janeiro: Ed. Fiocruz, 1994.* Rio de Janeiro: FIOCRUZ, 1994. Disponível em: <https://static.scielo.org/scielobooks/wqffx/pdf/santos-9788575412770.pdf>.

SCHEIB, C. L. et al. Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*, v. 360, n. 6392, p. 1024–1027, jun. 2018. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.aar6851>.

SCHEINFELDT, L. B. et al. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biology*, v. 13, n. 1, p. R1, 2012. ISSN 1465-6906. Disponível em: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-1-r1>.

SCHMID, T.; ZHOU, J.; BRÜNE, B. HIF-1 and p53: communication of transcription factors under hypoxia. *Journal of Cellular and Molecular Medicine*, v. 8, n. 4, p. 423–431, out. 2004. ISSN 1582-1838, 1582-4934. Disponível em: <http://doi.wiley.com/10.1111/j.1582-4934.2004.tb00467.x>.

SEMENZA, G. L. Regulation of Mammalian O $_2$ Homeostasis by Hypoxia-Inducible Factor 1. *Annual Review of Cell and Developmental Biology*, v. 15, n. 1, p. 551–578, nov. 1999. ISSN 1081-0706, 1530-8995. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.cellbio.15.1.551>.

SERAVALLE, G.; GRASSI, G. Obesity and hypertension. *Pharmacological Research*, v. 122, p. 1–7, ago. 2017. ISSN 1096-1186.

SHRINER, D. et al. Problems with Genome-Wide Association Studies. *Science*, v. 316, n. 5833, p. 1840c–1842c, jun. 2007. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.316.5833.1840c>.

SHRIVER, M. D. et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*, v. 1, n. 4, p. 274–286, maio 2004. ISSN 1473-9542.

SILVA, J. F.; OCARINO, N. M.; SERAKIDES, R. Thyroid hormones and female reproduction†. *Biology of Reproduction*, maio 2018. ISSN 0006-3363, 1529-7268. Disponível em: <https://academic.oup.com/biolreprod/advance-article/doi/10.1093/biolre/ioy115/4995900>.

SILVA, M. A. Castro e et al. Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proceedings of the National Academy of Sciences*, v. 117, n. 5, p. 2372–2377, fev. 2020. ISSN 0027-8424, 1091-6490. Disponível em: <http://www.pnas.org/lookup/doi/10.1073/pnas.1909075117>.

SIMONSON, T. S. et al. Adaptive genetic changes related to haemoglobin concentration in native high-altitude Tibetans: Tibetan adaptation and haemoglobin concentration at high altitude. *Experimental Physiology*, v. 100, n. 11, p. 1263–1268, nov. 2015. ISSN 09580670. Disponível em: <http://doi.wiley.com/10.1113/EP085035>.

SIMONSON, T. S. et al. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science*, v. 329, n. 5987, p. 72–75, jul. 2010. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.1189406>.

SKOGLUND, P. et al. Genetic evidence for two founding populations of the Americas. *Nature*, v. 525, n. 7567, p. 104–108, set. 2015. ISSN 1476-4687.

SMITH, J. M.; HAIGH, J. The hitch-hiking effect of a favourable gene. *Genetical Research*, v. 23, n. 1, p. 23–35, fev. 1974. ISSN 0016-6723, 1469-5073. Disponível em: <https://www.cambridge.org/core/product/identifier/S0016672300014634/type/journal_article>.

SOARES, L. P. *Perfil nutricional e alterações metabólicas na população adulta Xavante das reservas indígenas de São Marcos e Sangradouro - MT*. Tese (Tese) — Universidade de São Paulo, São Paulo, 2015. Disponível em: <https://www.teses.usp.br/teses/disponiveis/17/17139/tde-08122015-155437/publico/LuanaPaduaSoaresCorrig.pdf>.

SOARES, L. P. et al. Prevalence of metabolic syndrome in the Brazilian Xavante indigenous population. *Diabetology & Metabolic Syndrome*, v. 7, p. 105, 2015. ISSN 1758-5996.

SOHAIL, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, v. 8, 2019. ISSN 2050-084X.

SOUZA, J. G. de et al. Pre-Columbian earth-builders settled along the entire southern rim of the Amazon. *Nature Communications*, v. 9, n. 1, p. 1125, 2018. ISSN 2041-1723.

STELZER, G. et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, v. 54, p. 1.30.1–1.30.33, 2016. ISSN 1934-340X.

STEPHAN, W. Selective Sweeps. *Genetics*, v. 211, n. 1, p. 5–13, jan. 2019. ISSN 0016-6731, 1943-2631. Disponível em: <http://www.genetics.org/lookup/doi/10.1534/genetics.118.301319>.

SU, J. et al. Galectin-13, a different prototype galectin, does not bind β-galacto-sides and forms dimers via intermolecular disulfide bridges between Cys-136 and Cys-138. *Scientific Reports*, v. 8, n. 1, p. 980, dez. 2018. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/s41598-018-19465-0>.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, n. 43, p. 15545–15550, out. 2005. ISSN 0027-8424.

SUTTER, R. C. The Pre-Columbian Peopling and Population Dispersals of South America. *Journal of Archaeological Research*, mar. 2020. ISSN 1059-0161, 1573-7756. Disponível em: <http://link.springer.com/10.1007/s10814-020-09146-w>.

SUZUKI, Y. Statistical methods for detecting natural selection from genomic data. *Genes & Genetic Systems*, v. 85, n. 6, p. 359–376, 2010. ISSN 1880-5779, 1341-7568. Disponível em: <http://joi.jlc.jst.go.jp/JST.JSTAGE/ggs/85.359?from=CrossRef>.

SZPIECH, Z. A.; HERNANDEZ, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, v. 31, n. 10, p. 2824–2827, out. 2014. ISSN 1537-1719.

TAMM, E. et al. Beringian standstill and spread of Native American founders. *PloS One*, v. 2, n. 9, p. e829, set. 2007. ISSN 1932-6203.

TANG, K.; THORNTON, K. R.; STONEKING, M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS biology*, v. 5, n. 7, p. e171, jul. 2007. ISSN 1545-7885.

TARIM, Ö. Thyroid hormones and growth in health and disease. *Journal of Clinical Research in Pediatric Endocrinology*, v. 3, n. 2, p. 51–55, 2011. ISSN 1308-5735.

TEMPLETON, A. R. *Population Genetics and Microevolutionary Theory*. 1. ed. New Jersey: John Wiley & Sons, 2006. ISBN 978-0-471-40951-9.

THAN, N. G. et al. Functional analyses of placental protein 13/galectin-13: Functional analyses of PP13/galectin-13. *European Journal of Biochemistry*, v. 271, n. 6, p. 1065–1078, mar. 2004. ISSN 00142956. Disponível em: <http://doi.wiley.com/10.1111/j.1432-1033.2004.04004.x>.

The 1000 Genomes Project Consortium; DELANEAU, O.; MARCHINI, J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, v. 5, n. 1, p. 3934, set. 2014. ISSN 2041-1723. Disponível em: <http://www.nature.com/articles/ncomms4934>.

The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, v. 47, n. D1, p. D330–D338, 2019. ISSN 1362-4962.

TO, K. K. W.; HUANG, L. E. Suppression of hypoxia-inducible factor 1alpha (HIF-1alpha) transcriptional activity by the HIF prolyl hydroxylase EGLN1. *The Journal of Biological Chemistry*, v. 280, n. 45, p. 38102–38107, nov. 2005. ISSN 0021-9258.

TOVO-RODRIGUES, L. et al. Dopamine receptor D4 allele distribution in Amerindians: A reflection of past behavior differences? *American Journal of Physical Anthropology*, v. 143, n. 3, p. 458–464, nov. 2010. ISSN 00029483. Disponível em: <http://doi.wiley.com/10.1002/ajpa.21358>.

UEKI, M.; CORDELL, H. J. Improved statistics for genome-wide interaction analysis. *PLoS genetics*, v. 8, n. 4, p. e1002625, 2012. ISSN 1553-7404.

VALVERDE, G. et al. A Novel Candidate Region for Genetic Adaptation to High Altitude in Andean Populations. *PLOS ONE*, v. 10, n. 5, p. e0125444, maio 2015. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0125444>.

VERDU, P. African Pygmies. *Current biology: CB*, v. 26, n. 1, p. R12–14, jan. 2016. ISSN 1879-0445.

VISSCHER, P. M. et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, v. 101, n. 1, p. 5–22, jul. 2017. ISSN 00029297. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409>.

VITTI, J. J.; GROSSMAN, S. R.; SABETI, P. C. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, v. 47, n. 1, p. 97–120, nov. 2013. ISSN 0066-4197, 1545-2948. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev-genet-111212-133526>.

VOGT, M.; BILLETER, R.; HOPPELER, H. Einfluss von Hypoxie auf die muskuläre Leistungsfähigkeit: "Living low – Training high". *Therapeutische Umschau*, v. 60, n. 7, p. 419–424, jul. 2003. ISSN 0040-5930, 1664-2864. Disponível em: <https://econtent.hogrefe.com/doi/10.1024/0040-5930.60.7.419>.

VOIGHT, B. F. et al. A map of recent positive selection in the human genome. *PLoS biology*, v. 4, n. 3, p. e72, mar. 2006. ISSN 1545-7885.

WALKER, R. et al. Growth rates and life histories in twenty-two small-scale societies. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, v. 18, n. 3, p. 295–311, jun. 2006. ISSN 1042-0533.

WAN, X. et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, v. 87, n. 3, p. 325–340, set. 2010. ISSN 1537-6605.

WANG, J. et al. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, v. 41, n. Web Server issue, p. W77–83, jul. 2013. ISSN 1362-4962.

WANG, J.; LIAO, Y. *WebGestaltR: Gene Set Analysis Toolkit WebGestaltR*. 2020. Disponível em: <https://CRAN.R-project.org/package=WebGestaltR>.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, v. 38, n. 16, p. e164–e164, set. 2010. ISSN 0305-1048, 1362-4962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603>.

WATANABE, K. et al. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, v. 8, n. 1, p. 1826, 2017. ISSN 2041-1723.

WATLING, J. et al. Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. *PloS One*, v. 13, n. 7, p. e0199868, 2018. ISSN 1932-6203.

WATSON, D. F.; PHILIP, G. M. A refinement of inverse distance weighted interpolation. *Geo-processing*, v. 2, n. 4, p. 315–327, 1985. ISSN 0165-2273. Disponível em: <https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=8419722>.

WELCH, J. R. et al. Nutrition Transition, Socioeconomic Differentiation, and Gender Among Adult Xavante Indians, Brazilian Amazon. *Human Ecology*, v. 37, n. 1, p. 13–26, fev. 2009. ISSN 0300-7839, 1572-9915. Disponível em: <http://link.springer.com/10.1007/s10745-009-9216-7>.

WENG, L. et al. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC bioinformatics*, v. 12, p. 99, abr. 2011. ISSN 1471-2105.

WILCOXON, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, v. 1, n. 6, p. 80, dez. 1945. ISSN 00994987. Disponível em: <https://www.jstor.org/stable/10.2307/3001968?origin=crossref>.

WOOLLER, M. J. et al. A new terrestrial palaeoenvironmental record from the Bering Land Bridge and context for human dispersal. *Royal Society Open Science*, v. 5, n. 6, p. 180145, jun. 2018. ISSN 2054-5703, 2054-5703. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rsos.180145>.

WRIGHT, S. The genetical structure of populations. *Annals of Eugenics*, v. 15, n. 4, p. 323–354, 1949.

XU, S. et al. A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans. *Molecular Biology and Evolution*, v. 28, n. 2, p. 1003–1011, fev. 2011. ISSN 0737-4038, 1537-1719. Disponível em: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msq277>.

YATES, A. et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics (Oxford, England)*, v. 31, n. 1, p. 143–145, jan. 2015. ISSN 1367-4811.

YI, X. et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, v. 329, n. 5987, p. 75–78, jul. 2010. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.sciencemag.org/lookup/doi/10.1126/science.1190371>.

ZELEDÓN, R.; RABINOVICH, J. E. Chagas' disease: an ecological appraisal with special emphasis on its insect vectors. *Annual Review of Entomology*, v. 26, p. 101–133, 1981. ISSN 0066-4170.

ZHANG, B.; KIROV, S.; SNODDY, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, v. 33, n. Web Server issue, p. W741–748, jul. 2005. ISSN 1362-4962.

ZHOU, C.-H. et al. Modeling the interplay between the HIF-1 and p53 pathways in hypoxia. *Scientific Reports*, v. 5, n. 1, p. 13834, nov. 2015. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep13834>.