Marcos Araújo Castro e Silva

“Perspectiva genômica sobre a origem, história e diversidade dos povos indígenas da América do Sul: do povoamento inicial à colonização europeia”

“Genomic perspective on the origin, history, and diversity of indigenous peoples from South America: from the initial settlement to the European colonization”

Tábita Hünemeier
Orientadora

São Paulo
2021

Marcos Araújo Castro e Silva

"Perspectiva genômica sobre a origem, história e diversidade dos povos indígenas da América do Sul: do povoamento inicial à colonização europeia"

"Genomic perspective on the origin, history, and diversity of indigenous peoples from South America: from the initial settlement to the European colonization"

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

Orientadora: Tábita Hünemeier

São Paulo
2021

FICHA CATALOGRÁFICA

Castro e Silva, Marcos Araújo

Perspectiva genômica sobre a origem, história e diversidade dos povos indígenas da América do Sul: do povoamento inicial à colonização europeia.
237 páginas.

Tese (Doutorado) - Instituto de Biociências da Universidade de São Paulo. Departamento de Genética e Biologia Evolutiva.

Palavras-chave:
1. Genômica de populações;
2. Genética antropológica;
3. Nativos brasileiros;
4. Povoamento da América;
5. Diversidade genética.

I. Universidade de São Paulo. Instituto de Biociências. Departamento de Genética e Biologia Evolutiva.

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Profa. Dra. Tábita Hünemeier

Dedico aos meus professores
e aos povos indígenas do Brasil,
em especial aos Tupiniquim.

"Povo" só (r)existe no plural — povoS. Um povo é uma multiplicidade singular, que supõe outros povos, que habita uma terra pluralmente povoada de povos. Quanto em uma entrevista perguntaram ao escritor Daniel Munduruku se ele "enquanto índio etc.", ele cortou no ato: "não sou índio; sou Munduruku". Mas ser Munduruku significa saber que existem Kayabi, Kayapó, Matis, Guarani, Tupinambá, e que esses não são Munduruku, mas tampouco são Brancos. Quem inventou os "índios" como categoria genérica foram os grandes especialistas na generalidade, os Brancos, ou por outra, o Estado branco, colonial, imperial, republicano.

[...]

O povo tem a forma do Múltiplo. Forçados a se descobrirem "índios", os índios brasileiros descobriram que haviam sido unificados na generalidade por um poder transcendente, unificados para melhor serem des-multiplicados, homogeneizados, abrasileirados.

[...]

Os índios são os primeiros indígenas do Brasil. As terras que ocupam não são sua propriedade — não só porque os territórios indígenas são terras da "União", mas porque são eles que pertencem à terra e não o contrário. Pertencer à terra, em lugar de ser proprietário dela, é o que define o indígena.

(VIVEIROS DE CASTRO, 2017)

AGRADECIMENTOS

Agradeço à minha orientadora Profª. Drª. Tábita Hünemeier pela oportunidade de poder perseguir meus interesses e me dedicar a um tema de pesquisa que sempre me fascinou, no qual felizmente pude me sentir realizado, e pelo incentivo para sempre seguir em frente.

Agradeço à Drª. Kelly Nunes e ao Dr. Renan Barbosa Lemes pela amizade, pelos valiosos ensinamentos e pela generosidade em compartilhar o seu conhecimento comigo, principalmente na época mais difícil do início do doutorado.

Sou muito grato aos amigos do Laboratório de Genômica Populacional Humana (LGPH/USP), Tiago Ferraz, Cainã Couto e Gabrielle Rizzato por terem tornado mais fácil e agradável essa longa jornada do doutorado. Principalmente pela ajuda na adaptação a São Paulo e ao laboratório, pelo companheirismo, pelas brincadeiras e pelas muitas discussões sem fim ao longo desses anos.

Aos amigos do "Porão da Evolução" e do Departamento de Genética pelo encorajamento e pelos momentos agradáveis compartilhados, em especial ao André Fonseca, Lilian Kimura, Barbara Costa e Daniela Rossoni.

Aos colegas de "república" Jorge Arthuzzi, Fábio Sartorio e Flavio Segundo, pela amizade, pelas inúmeras conversas e momentos de descontração.

Aos amigos que embora distantes, se fizeram presentes no dia a dia com conselhos e escuta, dividindo o peso da jornada comigo, sobretudo à Karine Munck e ao Vinícius Carvalho.

Aos povos indígenas do Brasil pela generosidade e pelo exemplo de resistência e superação das adversidades com alegria e beleza.

À Profª. Drª. Maria Cátira Bortolini, Prof. Dr. David Comas, Prof. Dr. José Geraldo Mill, Prof. Dr. Alexandre da Costa Pereira, Prof. Dr. José Eduardo Krieger, Prof. Dr. Carlos Eduardo Guerra Amorim e Dr. Àlex Mas Sandoval, pela colaboração e apoio; e à Profª. Drª. Priscilla Zamberlan pela ajuda com a revisão da tese.

Ao Prof. Dr. Francisco Mauro Salzano (*in memoriam*) pelo pioneirismo no estudo da genética de populações humanas, principalmente dos povos indígenas brasileiros, abrindo caminhos para as novas gerações de geneticistas e pelo exemplo de amor à ciência.

Aos funcionários e técnicos do Departamento de Genética e Biologia Evolutiva e do Instituto de Biociências, especialmente à Erika Camargo pela disponibilidade e atenção em sempre ajudar nas questões burocráticas.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: 140155/2017-1) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP: 2015/26875-9 e 2018/013716) pelo apoio financeiro.

Finalmente, agradeço a minha família pelo incentivo, suporte e carinho que me impulsionaram até aqui, sem os quais nada disso seria possível. Ao meu pai pelos ensinamentos e pelo suporte. Ao meu irmão pelas muitas portas abertas, pela inspiração pra encontrar o meu caminho e principalmente pela amizade. Agradeço especialmente a minha mãe, pelo exemplo de vida e de luta que me ensinou a sonhar e por seu amor incondicional.

APOIO FINANCEIRO

Esta tese foi realizada com o apoio financeiro do projeto CNPq 140155/2017-1 e dos projetos FAPESP 2015/26875-9 e 2018/013716.

SUMÁRIO

LISTA DE PUBLICAÇÕES

- **Capítulo 1**:

CASTRO E SILVA, M. A. et al. Deep Genetic Affinity between Coastal Pacific and Amazonian Natives Evidenced by Australasian Ancestry. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 14, 6 abr. 2021.

- **Capítulo 2**:

CASTRO E SILVA, M. A. et al. Population histories and genomic diversity of South American natives. **Manuscrito submetido ao periódico Molecular Biology and Evolution (MBE) (em revisão)**.

- **Capítulo 3:**

CASTRO E SILVA, M. A. 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, 2020.

INTRODUÇÃO

Formação das populações americanas contemporâneas

A história humana nas Américas, segundo o modelo mais aceito pela comunidade científica atualmente, se inicia com a entrada dos primeiros grupos de *Homo sapiens* no continente durante o fim do último período glacial, há aproximadamente 16.000 AP[1]. Entretanto, é provável que a variabilidade genética dos povos americanos atuais tenha sido fortemente influenciada por eventos desencadeados pelo contato dos povos originários da atual América[2] com populações não nativas a partir do século 15. A invasão desse novo continente pelos povos europeus daria início então a alguns dos maiores eventos demográficos e migratórios da história humana.

À época da chegada dos europeus, estima-se que dezenas de milhões (entre 8,4 e 112,5 milhoẽs) de pessoas habitavam a América (G.; SANCHEZ-ALBORNOZ, 1976; THORNTON, 1987, 2005; DENEVAN, 1992), sendo que a maior densidade populacional ocorria nos Andes, com estimativas de 3 a 37 milhões de pessoas (DENEVAN, 1992), ao passo que no território do que viria a se tornar o Brasil, o número estimado era de aproximadamente 3 milhões de pessoas (IBGE, 2000), dos quais um terço deles teria habitado a região da costa do Atlântico. Este contingente populacional foi drasticamente reduzido em aproximadamente 90-95% a partir de 1492, como consequência de diferentes processos decorrentes da colonização europeia, com destaque para as epidemias devido a introdução de novas doenças, escravização, incitação de conflitos entre grupos[3] indígenas[4] rivais, guerras de conquista, deslocamento forçado de territórios, e destruição de habitats e disrupção dos meios de subsistência e conhecimentos tradicionais (THORNTON, 1987; STANNARD, 1993; MONTENEGRO; STEPHENS, 2006; UBELAKER, 2006). De tal forma que, ainda no século 18, as populações do litoral brasileiro

[1] Anos antes do presente.

[2] Antes de ser batizada de "América" pelos invasores europeus, os povos indígenas utilizaram vários nomes para designar o continente ou porções dele onde habitavam, como Abya Yala que na língua do povo Kuna ou Guna significa "Terra madura", "Terra viva" ou "Terra em florescimento" e Pindorama que nas línguas Tupí-Guaraní significa "Terra das palmeiras".

[3] Nesta tese, quando utilizamos o termo "grupo" estamos nos referindo a um conjunto de indivíduos de um mesmo grupo étnico (e.g. Parakanã, Xávante) ou população com uma delimitação espaço-temporal específica, aqui também chamado algumas vezes de "população", com o mesmo significado. Quando fazemos referência a outros níveis hierárquicos de agrupamento, o uso do termo "grupo" é acompanhado de alguma especificação, como por exemplo grupo linguístico (i.e. tronco ou família linguística).

[4] Por sua vez, os termos "indígena" (ou "indígena americano") e "nativo americano" são usados nesta tese como sinônimos para nos referirmos aos povos que já habitavam o continente americano antes da chegada dos europeus, assim como aos seus descendentes contemporâneos que fazem parte de comunidades indígenas ou se identificam como parte da herança cultural indígena, independente de critérios de genéticos e/ou fenotípicos de ancestralidade.

foram (equivocadamente) consideradas extintas e os grupos indígenas que resistiram a este processo se encontravam em sua maioria no interior do país, principalmente na região amazônica (DA CUNHA, 1992).

Concomitantemente, ocorreu também uma ampla miscigenação entre estes povos, anteriormente separados por milhares de anos de história evolutiva e que agora se encontravam reunidos no continente americano. As populações latino americanas atuais seriam desse modo majoritariamente formadas pela convergência de três grandes grupos continentais de ancestralidade, sendo o primeiro componente proveniente dos próprios nativos americanos e os outros dois introduzidos pela colonização das Américas pelos europeus, e por fim pelos africanos aprisionados e trazidos em enormes quantidades durante o tráfico transatlântico de escravos (SALZANO; BORTOLINI, 2005; ADHIKARI et al., 2016, 2017; ONGARO et al., 2019).

Num primeiro momento, a colonização da América foi predominantemente realizada por povos originários da península ibérica (i.e. Espanha e Portugal), iniciando a ocupação pelo Caribe e posteriormente avançando para a América Central e do Sul (MCALISTER, 1987; KAMEN, 2003; FERNÁNDEZ-ARMESTO, 2004). Outros povos europeus, sobretudo ingleses, franceses e holandeses, deram início ou intensificaram a colonização apenas no século 16. Durante o período colonial, até o século 19, um milhão de espanhóis e portugueses migraram para as suas colônias americanas, sendo aproximadamente metade de cada grupo (MCALISTER, 1987; KAMEN, 2003; FERNÁNDEZ-ARMESTO, 2004). Ao mesmo tempo, um milhão de ingleses, franceses e holandeses também migraram para as Américas (ALTMAN et al., 1991). Apenas nos séculos 19 e 20, o contingente de europeus que imigraram para as Américas totaliza mais de 50 milhões, sendo que 78% deles foram para a América do Norte e o restante (aproximadamente 11 milhões) foi para a América Latina, um terço dos quais vieram especificamente para o Brasil (ALVIM, 1998; SCHWARCZ; STARLING, 2015). Os imigrantes que se dirigiram a América Latina eram compostos de pouco mais de um terço de italianos (38%), quase um terço de espanhóis (28%), uma porção menor mas ainda significativa de portugueses (11%) e pequenas contribuições de franceses e alemães (3%) (ALVIM, 1998; SCHWARCZ; STARLING, 2015).

Neste contexto, a população brasileira foi formada por contribuições majoritárias de portugueses, italianos e espanhóis, respectivamente representando 38,05%, 28,04% e 12,02% da quantidade total estimada de 5.774.000 imigrantes europeus recebidos no período de 1500 a 1960 EC[5] (PENA, 2002), além de números menos significativos de imigrantes alemães (250.000) ou com origem no Leste Asiático e Oriente Médio, principalmente japoneses (229.000), sírios, libaneses e turcos (185.000) (PENA, 2002). Essa imigração é estruturada em dois períodos principais, no primeiro período de 1500 a 1808 EC virtualmente apenas portugueses migraram

[5] Era Comum.

para o Brasil, enquanto que de 1808 a 1960 EC a diversidade de origem dos imigrantes aumentou, particularmente após a abolição da escravatura em 1888 EC (PENA, 2002).

Os ibéricos foram os primeiros a iniciar o tráfico de africanos escravizados para as Américas, seguidos pelos ingleses, franceses e holandeses. Durante o período escravagista do Atlântico, africanos foram trazidos principalmente de algumas regiões específicas da África, as quais também possuíam destinos preferenciais nas Américas. Nesse sentido, particularmente a América portuguesa[6] recebeu em torno de 42% de todo o contingente populacional de aproximadamente 10 milhões de pessoas (CURTIN, 1972)[7] e os africanos trazidos eram originários majoritariamente de três regiões da África Subsaariana, as quais enumeradas de acordo com o volume do tráfico, segundo dados do banco *Slave Trade database*[8] (SEARING; ELTIS, 2001), são: (i) centro-oeste da África (Angola e Congo), (ii) sudeste da África (Moçambique) e (iii) oeste da África (Senegâmbia, Golfo do Benim, Golfo de Biafra, Costa do Marfim, Costa do Ouro e Serra Leoa); correspondendo a origem de 73,2%, 17,3% e 9,5% do número total de pessoas escravizadas, respectivamente. Estas também são as principais origens da diáspora africana para o restante do continente americano, sendo o centro-oeste africano, do mesmo modo, a origem predominante, e por sua vez as regiões oeste e sudeste africanas exibindo padrões contrastantes, com maiores contribuições para porção setentrional e meridional das Américas, respectivamente (ADHIKARI et al., 2017).

Este processo de miscigenação ocorreu de forma diferencial no tempo e no espaço, sendo bastante influenciado pelos contextos locais de densidade populacional indígena, de disponibilidade de recursos específicos de interesse dos europeus e do consequente volume de imigrantes, mas também dependente de outros fatores como a intensidade do emprego de mão-de-obra escrava africana e indígena em cada região específica, assim como dos contextos socioculturais que determinaram a frequência e a extensão com que a miscigenação ocorreu (ADHIKARI et al., 2017). Assim sendo, o resultado final é um mosaico em constante mudança formado por ancestralidades de diversas origens sub-continentais, as quais foram sendo introduzidas na América em momentos históricos distintos.

Além disso, ao longo do tempo ocorreram extensos movimentos populacionais macro e micro regionais no interior do continente e portanto os padrões de distribuição desse mosaico de ancestralidades se alteraram ainda mais até chegar a atual configuração (SALZANO; BORTOLINI, 2005; RUIZ-LINARES et al., 2014; MONTINARO et al., 2015; ADHIKARI et al., 2016, 2017; CHACÓN-DUQUE et al., 2018; ONGARO et al., 2019). Este processo foi também

[6] Regiões da América colonizadas pelos Portugueses, as quais em sua maioria deram origem ao Brasil.
[7] Do total de africanos escravizados, ~25% foram levados a colônias britânicas, ~15% colônias espanholas e ~14% para colônias francesas.
[8] O banco *Slave Trade database* reúne dados sobre 1,3 milhões de indivíduos.

influenciado por profundas diferenças em diversos aspectos socioculturais e históricos entre as colônias ibéricas e britânicas, que determinaram uma maior amplitude de miscigenação entre povos de diferentes ancestralidades nas colônias ibéricas, e a uma segregação mais pronunciada nas colônias britânicas, o que pode ainda ser observado em estudos sobre a diversidade genéticas de populações contemporâneas (MONTINARO et al., 2015; ADHIKARI et al., 2017; ONGARO et al., 2019).

Particularmente a miscigenação envolvendo indígenas ocorreu de forma preferencialmente local, de modo que hoje a ancestralidade indígena de populações miscigenadas remonta consistentemente aos grupos que ocupavam a mesma região no passado, permitindo assim revelar o padrão de estrutura e diversidade genética do período pré-contato[9] através do estudo destas populações, ao menos do período mais imediatamente anterior, padrão esse que foi ocultado pela chegada dos colonizadores e pelo extermínio da maior parte dos povos indígenas (WANG et al., 2008; MORENO-ESTRADA et al., 2014; KEHDY et al., 2015; MONTINARO et al., 2015; ADHIKARI et al., 2016; HARRIS et al., 2018).

Origem dos povos nativos americanos

Considerando esse quadro de dinâmicas populacionais intensas após o contato com os europeus, resta agora compreender onde se originaram, como se formaram e se dispersaram os diferentes povos nativos americanos. Nesse sentido, o povoamento inicial das Américas permanece um tópico de pesquisa ativa apesar de mais de um século de estudos e ainda que alguns dos principais aspectos desse processo tenham sido elucidados, persistem diversas dúvidas particularmente no que diz respeito a contextos mais específicos e locais, assim como em relação a dinâmicas populacionais internas ao continente, e também sobre a própria composição dos primeiros grupos humanos colonizadores e as rotas de entrada e dispersão[10] tomadas por elas. Atualmente há um consenso de que todos os povos nativos americanos têm origem em populações que antes ocupavam o leste asiático, o que é suportado por um vasto corpo de evidências provenientes de diversas áreas como a arqueologia (DILLEHAY, 2009; MELTZER, 2009; BRAJE et al., 2017; POTTER et al., 2017), a linguística (GREENBERG et al., 1986; NICHOLS, 2015), a análise de caracteres morfológicos (TURNER, 1985; NEVES; MEYER;

[9] Período anterior a chegada dos europeus às Américas, que tem início em 1492 EC com a chegada da frota de Cristóvão Colombo.

[10] O termo dispersão é usado nesta tese para fazer referência a movimentos populacionais de forma genérica, sem especificar se seria um movimento caracterizado pelo abandono do território original e motivado por fatores ambientais ou culturais, o que configura uma migração, ou se seria um movimento causado pelo crescimento populacional e difusão dessa população (i.e. difusão dêmica) pelo território configurando uma expansão (ou expansão dêmica) (SANJEK, 2003).

PUCCIARELLI, 1996; POWELL; NEVES, 1999) e a genética (REICH et al., 2012; LLAMAS et al., 2016; MORENO-MAYAR et al., 2018b; POSTH et al., 2018; SCHEIB et al., 2018). A hipótese de uma origem asiática dos indígenas americanos é bastante antiga (DE ACOSTA, 1589), tendo sido aventada inicialmente devido às semelhanças morfológicas evidentes entre os povos nativos americanos e asiáticos. Hoje sabemos que esta semelhança fenotípica se deve ao fato de que os nativos americanos se originaram no leste da Ásia e entraram nas Américas através da plataforma continental entre a Sibéria e o Alasca, exposta pelos níveis dos oceanos até 130 metros abaixo dos atuais (LAMBECK et al., 2014), durante o último máximo glacial[11] (UMG) entre ~24.000 a 17.000 AP (BROMLEY et al., 2016).

Para entender o povoamento do continente americano é necessário inicialmente contextualizar a dispersão dos primeiros humanos modernos no nordeste asiático, isto porque o povoamento das Américas através da Beríngia necessariamente ocorreu após a chegada e ocupação da Sibéria. Nessa perspectiva, as evidências arqueológicas mais antigas apontam para o início da ocupação do nordeste asiático ainda no período anterior ao UMG, com datações de ~31.600 AP na região do Rio Yana, próximo ao litoral do Oceano Ártico no nordeste da Rússia, (GRAF; BUVIT, 2017) e 24.000 AP em Mal'ta no centro-sul da Sibéria (RAGHAVAN et al., 2014). A Beríngia, por sua vez, teria passado a ser ocupada continuamente apenas entre ~15.000 e 13.600 AP (GRAF; BUVIT, 2017), o que estabelece um limite máximo para a entrada de humanos modernos na América, ou ao menos para a intensificação do processo de ocupação dos grupos que contribuíram para a composição genética dos nativos americanos. Nesse sentido, a Sibéria era então ocupada por uma população conhecida como Siberianos Antigos do Norte (ANS)[12], a qual estima-se que teria divergido dos eurasianos do oeste por volta de 39.000 AP, relativamente pouco tempo após a divergência entre os eurasianos do oeste e os asiáticos do leste (43.100 AP) (SIKORA et al., 2019). Os ANS apresentam uma afinidade genética tanto com populações contemporâneas do norte europeu quanto com nativos americanos, diferentemente de outros euroasiáticos, inclusive mais antigos, como os encontrados em Sunghir no oeste da Rússia (~34.000 AP) (SIKORA et al., 2017) e em Tianyuan no sudeste da China (39.565 AP) (FU et al., 2013), os quais apresentam similaridade genética mais específica com os euroasiáticos do oeste e asiáticos do leste, respectivamente. Os ANS não chegaram até os dias atuais como uma população geneticamente diferenciada, mas contribuíram geneticamente para os ancestrais de todos os nativos americanos através de um fluxo gênico entre eles e um grupo do Leste Asiático a aproximadamente 20.000-18.000 AP (SIKORA et al., 2019)[13]. Essa mistura teria levado a

[11] Faixa de tempo do último período glacial (o qual vai de 115.000 a 11.700 AP) na qual os glaciares continentais estavam na sua extensão máxima, com ápice a ~20.500 AP.

[12] *Ancient North Siberians*.

[13] Anteriormente esse fluxo gênico era inferido como uma contribuição de 40% da ancestralidade dos nativos americanos partindo de uma população chamada *Ancestral North Eurasians* (ANE), representada

formação de pelo menos duas linhagens distintas, a primeira delas são os chamados Paleo Siberianos Antigos[14], cujos descendentes hoje habitam o nordeste siberiano e a segunda linhagem seria a ancestral dos nativos americanos.

Ao contrário das populações americanas não-indígenas, as quais se constituíram como algumas das mais heterogêneas do mundo pelo seu histórico de ampla miscigenação, a análise da diversidade genética de populações indígenas americanas sempre revelou um nível mais baixo de diversidade em relação a outras regiões do globo. Tal padrão genético foi primeiramente detectado em análises de DNA mitocondrial e indica a presença de uma gargalo de garrafa[15] populacional antes ou durante o processo de entrada para as Américas, de modo que o tamanho da população formadora dos nativos americanos atuais é estimado entre algumas centenas até poucos milhares de indivíduos (WALLACE; GARRISON; KNOWLER, 1985; TORRONI et al., 1993; FAGUNDES; KANITZ; BONATTO, 2008; O'ROURKE; RAFF, 2010; LLAMAS et al., 2016; FAGUNDES et al., 2018; BERGSTRÖM et al., 2020).

Além disso, todos os nativos americanos estudados até hoje são descendentes de poucas linhagens fundadores: quatro patrilineares (os haplogrupos da região não recombinante do cromossomo Y: Q-M3, Q-CTS1780, C3-MPB373 e C3-P39/Z30536)[16] e nove matrilineares (os haplogrupos mitocondriais: A2, B2, C1b, C1c, C1d, D1 e os menos frequentes C4c, D4h3a e X2a). Ao mesmo tempo é possível concluir que estas linhagens se originaram no leste asiático, por integrarem um ramo irmão às linhagens asiáticas, e ainda pode-se inferir que passaram por um período de redução populacional prévia ou concomitante ao povoamento inicial, visto que algumas destas linhagens são completamente exclusivas às Américas, motivando a hipótese de que os primeiros humanos a migrarem para o continente teriam permanecido na região da Beríngia, ao menos parcialmente isolados de outras populações asiáticas durante um período do UMG, o que ficou conhecido como hipótese da permanência na Beríngia[17] (SZATHMARY, 1993; BONATTO; SALZANO, 1997; TAMM et al., 2007; WANG et al., 2007; KITCHEN; MIYAMOTO; MULLIGAN, 2008; MULLIGAN; KITCHEN; MIYAMOTO, 2008; BISSO-MACHADO et al., 2011; WATERS, 2019; BISSO-MACHADO; FAGUNDES, 2021). Estima-se que este período de isolamento tenha durado entre 4.600 (PINOTTI et al., 2019) e 15.000 AP (GRAF; BUVIT, 2017), porém o local onde eles permaneceram na Beríngia, ou mesmo fora dela, ainda não é um consenso

---

pelo indivíduo de Mal'ta (RAGHAVAN et al., 2014), posteriormente se demonstrou que essa linhagem pode ser modelada como descendente dos ANS (SIKORA et al., 2019).

[14] *Ancient Paleo-Siberians*.

[15] Uma redução populacional temporária.

[16] Q-M3, o haplogrupo mais frequente do cromossomo Y, além de três haplogrupos mais raros Q-CTS1780, C3-MPB373 e C3-P39/Z30536, estes últimos presentes apenas na América do Sul e do Norte, respectivamente. O haplogrupo C3-P39/Z30536 possivelmente representa uma dispersão posterior e portanto não seria uma linhagem fundadora (MORENO-MAYAR et al., 2018a; PINOTTI et al., 2019).

[17] *Beringian standstill hypothesis*.

(MORENO-MAYAR et al., 2018a) Outro aspecto ainda indeterminado é o motivo dessa permanência, que pode ter sido causado desde por barreiras ecológicas (TAMM et al., 2007) ou até mesmo pelo fato de a Beríngia ter atuado como um refúgio com condições bioclimáticas mais adequadas para a sobrevivência de seres humanos durante o UMG (SIKORA et al., 2019; RAE et al., 2020).

Um segundo evento de divergência populacional muito importante na história evolutiva dos nativos americanos teria ocorrido durante esse período de permanência na Beríngia ou próximo a ele, aproximadamente a 22.000-18.000 AP, quando a população dos chamados Beringianos Antigos (AB)[18] teria se separado dos ancestrais dos nativos americanos (ANA) (RAGHAVAN et al., 2015; MORENO-MAYAR et al., 2018a). Por sua vez, o ramo dos ANA teria se estruturado geneticamente entre 17.500 e 14.600 AP (MORENO-MAYAR et al., 2018b), dividindo a ancestralidade desta população nos componentes chamados de Nativos Americanos do Sul (SNA) e Nativos Americanos do Norte (NNA)[19] (REICH et al., 2012; RASMUSSEN et al., 2014; MORENO-MAYAR et al., 2018b; POSTH et al., 2018; SCHEIB et al., 2018). A determinação da região específica onde ocorrem esses eventos de estruturação e diversificação ainda é alvo de pesquisa e as evidências mais antigas apontam para o leste da Beríngia como a região candidata mais provável (POTTER et al., 2018; WATERS, 2019). Entretanto, evidências indicam que a divergência entre os AB e os ANA ocorreram no nordeste da Ásia e Sibéria (MORENO-MAYAR et al., 2018a), e que portanto a população beringiana já era estruturada geneticamente. Além disso, até mesmo a separação entre os NNA e os SNA pode ter ocorrido na Ásia, imediatamente após o período de permanência na Beríngia, de acordo com o modelo mais parcimonioso capaz de explicar um sinal de afinidade genética às populações da Bacia do Rio Amur, localizada no nordeste da China, presente em alguns AB (NING et al., 2020).

Todavia, durante o UMG a entrada por via terrestre nas Américas a partir da Beríngia estava bloqueada por geleiras continentais, as quais à época ocupavam todo o norte da América do Norte (MELTZER, 2009), o que não necessariamente impedia a entrada por rotas alternativas. Uma dessas possibilidades alternativas é a entrada pela chamada rota da costa do Pacífico[20], através da qual teria sido possível migrar para a América do Norte a partir de ~17.000 a 15.000 AP (FAGUNDES et al., 2008; PEREGO et al., 2009; MENOUNOS et al., 2017; DARVILL et al., 2018; LESNEK et al., 2018; DELSER et al., 2021), em oposição a hipótese de uma entrada mais tardia por um corredor livre de gelo formado ao longo das Montanhas Rochosas,

[18] *Ancient Beringians*.
[19] *Southern Native American* e *Northern Native American*, ou *Ancestral A* e *Ancestral B* (SCHEIB et al., 2018), respectivamente.
[20] *Pacific coastal route*.

que passava pela divisão entre duas geleiras continentais[21] e que permitia a passagem até o interior da América do Norte (PEREGO et al., 2009; POTTER et al., 2018). Entretanto esta rota pode não ter sido viável até 15.000-13.000 AP (HEINTZMAN et al., 2016; PEDERSEN et al., 2016; MARGOLD et al., 2019), sendo que o registro da presença humana mais antiga na região onde o corredor foi formado é datado em ~12.350 AP (DRIVER et al., 1996). No entanto, é importante ressaltar que estas rotas não são mutuamente exclusivas e que portanto ambas muito provavelmente foram utilizadas (POTTER et al., 2018), sendo assim, resta saber quais foram as contribuições e impactos de cada uma para a composição genética e cultural dos grupos nativos americanos.

Todos os nativos americanos, sejam antigos ou atuais, cujos dados genéticos foram analisados até o momento, derivam de um dos dois ramos, SNA ou NNA, dado que esses grupos divergiram entre 17.500 e 14.600 AP no nordeste da Ásia ou na Beríngia. Isto coloca um limite superior para a passagem pelas geleiras, visto que todos sul-americanos analisados até o momento apresentam exclusivamente ancestralidade SNA, de modo que a dispersão pela América do Sul deve ter ocorrido somente após a divergência entre os ramos (POSTH et al., 2018). Além disso, interessantemente alguns estudos indicam que ao menos parte das populações da América do Sul e Central são formadas por uma combinação das duas ancestralidades (SCHEIB et al., 2018). Exceções a esse padrão seriam as populações falantes de línguas Na-Dené (Chipewyan) e Esquimó-Aleuta (incluindo Paleo Inuítes e Thules, anteriormente conhecidos como Neo-Esquimós) para as quais seriam ainda necessários fluxos gênicos adicionais posteriores, provenientes de linhagens asiáticas distintas, para explicar a sua estrutura e diversidade genética (RASMUSSEN et al., 2010; REICH et al., 2012; MORENO-MAYAR et al., 2018a; FLEGONTOV et al., 2019). Por esse motivo, quando nos referirmos aos nativos americanos nesta tese, estaremos excluindo dessa definição os grupos Na-Dené e Esquimó-Aleuta (a menos que especificado de outra forma), e focando apenas nos grupos nativos americanos sub-árticos ou não-árticos.

Como dito anteriormente, apesar deste delineamento da história populacional inicial, algumas questões permanecem em aberto e portanto ainda é necessária uma maior resolução para o modelo de origem e dispersão dos nativos americanos. Uma das questões principais se refere a um excesso relativo de compartilhamento de alelos entre populações amazônicas brasileiras e populações australo-asiáticas (RAGHAVAN et al., 2015; SKOGLUND et al., 2015), que se expressa em maior grau quando se comparam algumas populações amazônicas com mesoamericanos da etnia Mixe, estes últimos são representantes de um dos ramos iniciais dos SNA, se posicionando como um grupo externo aos sul-americanos e portanto com divergência

[21] i.e. *Laurentide glacier* e *Cordilleran ice sheet*.

anterior a todos estes (REICH et al., 2012). Curiosamente, é inferido que os Mixe apresentam um sinal significativo de contribuição de uma população não amostrada[22], a qual teria divergido provavelmente durante o período de permanência na Beríngia a partir da população ancestral beringiana, que era estruturada geneticamente (MORENO-MAYAR et al., 2018b).

O excesso de ancestralidade australo-asiática por sua vez também foi modelado como a contribuição de uma população não amostrada (SKOGLUND et al., 2015), que recebeu o nome de população "Ypikuéra"[23] ou "Y", sugerindo uma história populacional mais complexa do que a definida até aquele momento, provavelmente envolvendo um fluxo populacional adicional a partir da Beríngia na formação da composição genética dos nativos americanos ou a existência de estruturação genética adicional das populações iniciais (dentre outras possibilidades bastante controversas e bem menos plausíveis, como migrações transpacíficas ou transatlânticas). Uma nova linha de evidência também citada anteriormente (NING et al., 2020), indica que a divergência entre os grupos AB, NNA e SNA pode ter ocorrido na Ásia, o que se comprovado teria favorecido o contato com diversos grupos do leste Asiático, tornando mais provável a introdução desta ancestralidade na composição genética dos grupos formadores dos nativos americanos. Contudo, interessantemente o sinal não pode ser detectado em nenhum indivíduo antigo da Sibéria ou da Beríngia, nem nos indivíduos de ancestralidade NNA (MORENO-MAYAR et al., 2018a; POSTH et al., 2018; SIKORA et al., 2019).

Ademais, como essa ancestralidade chegou até os dias atuais sendo detectada apenas em populações amazônicas ainda é uma questão que permanece sem resposta e portanto um dos enigmas mais intrigantes da história das migrações humanas. No passado a possibilidade de fluxos populacionais adicionais e distintos já havia sido aventada, sendo que uma das hipóteses mais relevantes se baseava na assim chamada morfologia craniana "paleoamericana", sobretudo de alguns indivíduos encontrados no sítio da Lapa do Santo, assim como de outras regiões como a Mesoamérica (NEVES; MEYER; PUCCIARELLI, 1996; POWELL; NEVES, 1999; GONZÁLEZ-JOSÉ et al., 2005), os quais supostamente representariam uma ocupação inicial do continente, composta por indivíduos com uma morfologia e ancestralidade distintas dos atuais grupos indígenas, ao passo que estes últimos teriam portanto substituído os primeiros colonizadores.

Por um lado esse modelo de dois componentes de ancestralidade distintos foi contestado por análises de morfologia craniofacial, cujos resultados apontavam para uma extensa diversidade morfológica, de modo que as morfologias craniofaciais paleoamericana e "mongolóide"[24] seriam apenas os extremos do espectro de variação, sendo assim a primeira

[22] Referência a população denominada de *unsampled population A*. Populações não amostradas são algumas vezes também chamadas de *ghost populations*.
[23] Termo Tupí cujo significado é "ancestral".
[24] Termo utilizado para designar a morfologia de povos da Ásia, Polinésia e também das Américas, de acordo com o modelo de divisão da espécie humana em três raças (Caucasianos, Mongolóides e

conservaria uma proporção maior de caracteres ancestrais e mais prevalentes nos grupos do povoamento inicial do Pleistoceno, enquanto que a segunda apresentaria um conjunto maior de fenótipos derivados como achatamento facial, os quais teriam evoluído e se dispersado a partir do Ártico durante o Holoceno[25] (GONZÁLEZ-JOSÉ et al., 2008; BORTOLINI et al., 2014). Por outro lado, uma análise recente da morfometria craniana das amostras de 89 populações antigas e contemporâneas do leste asiático (MATSUMURA et al., 2019) evidenciou uma separação desses indivíduos em dois grupos principais, a saber: (i) populações do nordeste e leste asiático, além de grupos iniciais de agricultores do sudeste asiático e do período mais recente; (ii) populações australo-papuanas[26] e grupos do Pleistoceno tardio e Holoceno inicial do sudeste asiático, estes últimos fortemente associados a contextos arqueológicos pré-agriculturalistas dos chamados hoabinhianos[27].

Para explicar esse padrão de variação foi também proposto um modelo de duas camadas (ou dois componentes), algo que já era indicado e suportado por análises anteriores[28]. Segundo o modelo durante a dispersão inicial pelo leste da Eurásia teria ocorrido uma divergência bastante antiga (até 65.000-50.000 AP) entre os grupos do nordeste (NEA)[29] e do sudeste (SEA)[30], sendo que estes últimos teriam migrado para o sudeste asiático, Sondalândia[31] e para o antigo continente do Sahul[32], ao passo que os primeiros teriam se dispersado para o nordeste asiático e posteriormente para as Américas (MATSUMURA et al., 2019). Interessantemente alguns grupos apresentam sinais de sobreposição ou troca entre os dois componentes, o que segundo Matsumura et al. indica a ocorrência de miscigenação, sendo que algumas amostras dos SEA, em especial alguns indivíduos das Ilhas Andamão, apresentam um sinal de afinidade com populações NEA (MATSUMURA et al., 2019), as quais provavelmente deram origem aos nativos americanos. Coincidentemente o grupo que produz a detecção mais significativa do sinal australo-asiático em populações nativas americanas é justamente o dos Onge das ilhas de

---

Negróides) proposto originalmente pela *Göttingen School of History*, entretanto o conceito de raças humanas se tornou obsoleto no sentido biológico e estes termos são hoje considerados pejorativos ou mesmo racistas (ACKERMANN et al., 2019).

[25] Período atual da escala de tempo geológica, iniciado a 11.700-11.650 AP.

[26] Populações contemporâneas nativas do sul da Ásia e de Pápua Nova Guiné.

[27] *Hòabìnhians.*

[28] Ver debates sobre os modelos de onda única (*Single Wave Model*) e ondas múltiplas (*Multiple Waves Model*) para a dispersão dos humanos anatomicamente modernos a partir da África (HUGO PAN-ASIAN SNP CONSORTIUM et al., 2009; STONEKING; DELFIN, 2010; RASMUSSEN et al., 2011; REYES-CENTENO et al., 2014; TASSI et al., 2015; BAE; DOUKA; PETRAGLIA, 2017).

[29] *Northeast asians*.

[30] *Southeast asians*.

[31] Região da plataforma continental do sudeste asiático, que une a Península da Malásia, as ilhas de Bornéu, Java e Sumatra, e arquipélagos da região, exposta pelos níveis até 130 metros mais baixos dos oceanos (LAMBECK et al., 2014) durante o último período glacial, sobretudo durante o UMG.

[32] Continente formado pela plataforma continental incluindo Austrália, Tasmânia e Nova Guiné, também exposto durante o último período glacial.

Andamão (RAGHAVAN et al., 2015; SKOGLUND et al., 2015; MORENO-MAYAR et al., 2018b; POSTH et al., 2018). Portanto, é plausível que este sinal de afinidade morfológica esteja remontando ao mesmo evento de miscigenação que teria introduzido a ancestralidade australo-asiática em alguns grupos nativos americanos e que desta forma essa ancestralidade estaria associada a uma fração do espectro de variação crânio-morfológica com caracteres ancestrais prevalentes, nos moldes do que ficou conhecida como morfologia paleoamericana (NEVES; MEYER; PUCCIARELLI, 1996; POWELL; NEVES, 1999; GONZÁLEZ-JOSÉ et al., 2005, 2008; BORTOLINI et al., 2014). Entretanto, a hipótese de associação entre a diversidade genética e morfológica foi testada em indivíduos nativos americanos e se demonstrou que aqueles identificados como portadores da morfologia paleoamericana não apresentam excesso significativo de compartilhamento de alelos com populações australo-asiáticas, com apenas uma exceção, e que portanto mais parcimoniosamente tais indivíduos poderiam ser considerados descendentes dos mesmos grupos ancestrais que o restante dos nativos americanos, antigos e contemporâneos, sem a necessidade da contribuição da ancestralidade de grupos adicionais (MORENO-MAYAR et al., 2018b; POSTH et al., 2018).

Cronologia e contexto do povoamento inicial

Uma hipótese muito duradoura sobre o povoamento inicial das Américas (HAYNES, 1964) preconizava que os primeiros americanos teriam sido caçadores de megafauna, abundante no continente até aquele momento, e se baseia na ampla distribuição geográfica de um conjunto específico de artefatos tecnológicos, particularmente de pontas de projéteis bastante características e uniformes, encontradas pela primeira vez em Clovis no Novo México, sudoeste dos Estados Unidos (FIGGINS; COOK, 1927). Segundo esse modelo, as pontas Clovis, como ficaram conhecidas, em conjunto com as chamadas pontas Rabo de Peixe[33], com distribuição na América Central e do Sul (LYNCH, 1978), seriam registros das primeiras ocupações humanas no continente, dado que ambas são aproximadamente contemporâneas e relativamente similares (SUTTER, 2021), e essa dispersão inicial teria ocorrido pelo corredor livre de gelo (POTTER et al., 2018).

Entretanto mais recentemente, esse modelo do povoamento inicial por parte de povos associados a cultura Clovis (chamado de modelo Clovis Primeiro)[34] (HAYNES, 1964) foi paulatinamente contestado por um corpo crescente de evidências, como por exemplo pelo período relativamente curto de existência da cultura, ao menos dos sítios seguramente datados,

[33] *Fishtail points*.
[34] *Clovis first model*.

aproximadamente entre 13.000 e 12.700 AP (GOEBEL; WATERS; O'ROURKE, 2008), muito posterior a faixa de tempo quando teoricamente se torna possível a entrada no continente pela Beríngia (~17.000-15.000 AP), sobretudo pela rota da costa do Pacífico (MENOUNOS et al., 2017; DARVILL et al., 2018; LESNEK et al., 2018). As evidências mais fortes vem principalmente da existência de sítios arqueológicos datados de período anterior e portanto definitivamente oriundos de ocupações pré-Clovis na América do Norte (e.g. Paisley Cave (GILBERT et al., 2008)) e do Sul (e.g. Monte Verde I e II (DILLEHAY et al., 2008, 2015)).

Soma-se isso a linha do tempo inferida a partir dos dados genéticos de nativos americanos atuais, a qual aponta que o clado formado pelos indivíduos SNA mais antigos, nomeadamente Anzick-1[35] (~12.800 AP), Spirit Cave (10.700 AP) e Lagoa Santa (10.400 AP), teria sofrido a primeira divergência entre ~14.900 e 13.200 AP, quando os ancestrais comuns de Anzick-1 e Spirit Cave se separaram dos ancestrais comuns de Lagoa Santa e da população mesoamericana contemporânea da etnia Mixe, ao passo que os ramos ancestrais de Lagoa Santa e Mixe teriam divergido entre ~14.800 e 12.800 (MORENO-MAYAR et al., 2018b). Ambas as estimativas de divergência são incompatíveis com o modelo Clovis Primeiro; no entanto, estão alinhadas com as datações de sítios pré-Clovis (DILLEHAY et al., 2008, 2015). Além disso, análises recentes indicam que a entrada na América do sul teria ocorrido entre ~15.500-14.600 AP e que o continente americano estaria totalmente colonizado a 13.200 AP (PRATES; POLITIS; PEREZ, 2020; DELSER et al., 2021).

Contudo é preciso fazer uma distinção entre os sítios pré-Clovis do período posterior ou final do UMG (pós-UMG), como os supracitados, e aqueles datados para o período anterior ao UMG (pré-UMG), dos quais alguns dos representantes mais importantes na América do Sul são: Pedra Furada no Piauí (GUIDON, 1986) e Pikimachay no Peru (MACNEISH et al., 1981) com datações entre ~50.000 e 30.000 AP, ou mesmo mais antigas. Ao contrário dos sítios pós-UMG que são mais amplamente aceitos pela comunidade científica, os sítios pré-UMG ainda são alvo de debate (SUTTER, 2021). Além disso, como discutido anteriormente, a ocupação por humanos modernos do nordeste Asiático e sobretudo da Sibéria necessariamente condiciona quando se torna possível a presença de humanos modernos nas Américas, o que ocorre apenas a partir de aproximadamente 32.000 AP (GRAF; BUVIT, 2017). Ao passo que no período posterior a chegada de grupos de humanos modernos ao nordeste asiático, se torna teoricamente possível que estes tenham chegado ao continente americano ainda no pré-UMG, considerando ainda que estes vivessem em grupos relativamente pequenos, seria plausível que os vestígios

[35] Anzick-1 é o esqueleto de uma criança do sexo masculino encontrada em Montana, no centro-sul dos Estados Unidos, em associação a artefatos presentes no local alegadamente pertencentes à cultura Clovis (LAHREN; BONNICHSEN, 1974; OWSLEY; HUNT, 2001), e portanto seria ao mesmo tempo um dos representantes mais antigos dos SNA e o único representante da cultura Clovis cujos dados genéticos estão disponíveis (RASMUSSEN et al., 2014).

arqueológicos deixados por eles sejam muito raros e que a contribuição para a composição genética dos nativos americanos seja nula ou insignificante (SUTTER, 2021).

Ancestralidades dos nativos da América do Sul

Considerando apenas o cenário mais provável do povoamento no pós-UMG[36], a dispersão dos nativos americanos da América do Norte até a América do Sul teria ocorrido de forma extremamente rápida. A luz das evidências atuais, os primeiros habitantes do continente sul-americano teriam chegado tão cedo quanto 15.000-14.000 AP (BODNER et al., 2012; RASMUSSEN et al., 2015; DILLEHAY et al., 2017) entretanto, uma estimativa baseada na distribuição de probabilidades de datações de radiocarbono de sítios arqueológicos coloca o limite cronológico mais provável para a primeira chegada em ~15.500 AP[37] (com limite superior de 16.600 AP e estimativa mais conservadora de 15.100 AP) (PRATES; POLITIS; PEREZ, 2020), o que também reforça uma cronologia pré-Clovis e pós-UMG do povoamento das Américas. Modelos demográficos de dispersão populacional indicam ainda, que considerando uma chegada a 13.800 AP dos primeiros habitantes do Brasil, partindo da abertura oeste da geleira *Cordilleran* (i.e. rota da costa do Pacífico) a 17.000 AP, a velocidade da dispersão necessária seria de 4,1 Km por ano, algo dentro do observado para grupos caçadores-coletores atuais (DELSER et al., 2021). De todo modo o continente já estava amplamente ocupado a ~13.200-12.000 AP (DELSER et al., 2021; SUTTER, 2021), usando como entrada o Istmo do Panamá, inicialmente através da costa do Pacífico (WANG et al., 2007; BODNER et al., 2012; DILLEHAY et al., 2017; LINDO et al., 2017; BRANDINI et al., 2018) e/ou da costa do Atlântico (GÓMEZ-CARBALLA et al., 2018). É preciso ponderar o fato de que quaisquer vestígios de ocupação inicial deixados nas plataformas continentais costeiras que se encontravam expostas durante o UMG, hoje provavelmente estão sob mais de uma centena de metros (~100-120 metros) de água dos Oceanos Pacífico e Atlântico, o que portanto dificulta enormemente o acesso aos vestígios do povoamento inicial da América do Sul e do continente americano como um todo, além de enviesar as inferências tanto sobre a data de entrada, quanto sobre as rotas migratórias utilizadas (LAMBECK et al., 2014; GOLDBERG; MYCHAJLIW; HADLY, 2016; SUTTER, 2021).

[36] Ou ao menos a intensificação do processo de povoamento, de modo que os primeiros grupos humanos a habitarem o continente não teriam contribuído para a composição genética dos nativos americanos antigos e contemporâneos.

[37] De modo geral é razoável a premissa de que durante o povoamento inicial de novos territórios são produzidos apenas registros altamente escassos e dispersos da presença humana e portanto bastante difíceis de serem detectados, sobretudo em decorrência da baixa densidade populacional característica dessa fase, consequentemente pode-se assumir que o início da ocupação ocorreu algum tempo antes da idade do registro mais antigo (PRATES; POLITIS; PEREZ, 2020).

No que se refere a ancestralidade, infere-se que os indígenas da América do Sul receberam a contribuição de quatro dispersões populacionais geneticamente distintas (POSTH et al., 2018). Os primeiros colonizadores da América do Sul seriam representantes do grupo SNA geneticamente próximos ao indivíduo Anzick-1 (RASMUSSEN et al., 2014). Essa primeira onda de povoamento inicial teria sido então substituída, a partir de aproximadamente 9.000 AP, por uma segunda dispersão de grupos SNA sem esta afinidade específica ao Anzick-1, evidenciando uma substituição dêmica ao menos parcial dos primeiros colonizadores (POSTH et al., 2018). Além destas duas dispersões principais, outras duas contribuições foram também identificadas para grupos específicos e mais restritos geograficamente. Isto posto, a terceira dispersão populacional seria portanto uma contribuição de grupos SNA geneticamente relacionados a indivíduos antigos das Ilhas do Canal da Califórnia[38] direcionadas especificamente para populações da porção mais central dos Andes, os quais teriam se difundido na região a aproximadamente 4.200 AP, sendo que o fluxo gênico que deu origem a esse padrão teria ocorrido portanto antes desta data (POSTH et al., 2018) e provavelmente está ligado à dispersão da agricultura a partir da Mesoamérica (SUTTER, 2021). Por fim, a quarta e última dispersão seria representada pela população Y, como discutido previamente, cuja contribuição pode ser detectada em populações viventes da Amazônia e planalto central brasileiro (Karitiana, Suruí e Xávante) (SKOGLUND et al., 2015), mas também como um sinal mais fraco em um indivíduo antigo de Lagoa Santa, em Minas Gerais, com datação de 10.000 AP (MORENO-MAYAR et al., 2018b).

Demografia e história populacional da América do Sul

A história demográfica do continente sul-americano compreende duas fases principais e distintas, de acordo com uma inferência feita pela análise conjunta da distribuição espaço-temporal de 5.464 datações de radiocarbono calibradas de 1.147 sítios arqueológicos da América do Sul, com datas distribuídas entre 13.000 e 2.000 AP (GOLDBERG; MYCHAJLIW; HADLY, 2016): (i) durante a primeira fase, no período compreendido entre 13.000 e ~5.500 AP, houve uma rápida expansão geográfica inicial com ocupação de grande parte do continente, seguida de uma fase de crescimento populacional dependente de densidade, de modo que no ínicio a população aumentou rapidamente até que a capacidade de suporte[39] foi atingida (a qual seria definida pelo ambiente e pelas estratégias de subsistências, nesse período

[38] *California Channel Islands*.

[39] É o tamanho ou densidade de uma população de uma dada espécie suportados por um ambiente específico, dada a disponibilidade de recursos do mesmo e as estratégias de subsistência adotadas, dentre outros fatores. Quando abaixo desse valor a população tende a aumentar, quando acima tende a diminuir.

predominantemente caçadoras-coletoras) e desse ponto em diante os tamanhos populacionais se mantiveram relativamente constantes entre 9.000 e ~5.500 AP; (ii) somente a partir de ~5.500 AP (até 2.000 AP), com a ampliação do sedentarismo e intensificação da produção de alimentos, se iniciou um período de crescimento populacional exponencial, ao menos em alguns centros culturais, sobretudo naqueles localizados nas porções central e setentrional dos Andes. Ainda segundo este modelo mais da metade do crescimento populacional teria ocorrido na fase de crescimento exponencial entre 5.500 e 2.000 AP, atingindo um total estimado de 615.000-1.000.000 de pessoas na América do Sul a 2.000 AP, considerando uma população inicial de 1.000 indivíduos (GOLDBERG; MYCHAJLIW; HADLY, 2016). Adicionalmente, esse padrão de crescimento logístico inicial (i.e. dependente de densidade) é corroborado por uma análise mais recente de um conjunto de dados com controle de qualidade rigoroso com mais 1.600 datações de radiocarbono do período inicial de ocupação da América do Sul (PRATES; POLITIS; PEREZ, 2020), indicando que o referido modelo se ajusta melhor aos dados do que um modelo de crescimento exponencial. Essa análise evidencia ainda que a estabilidade demográfica (i.e. limite da capacidade de suporte) é atingida a 11.000 AP e que ocorreram alguns picos temporários de crescimento acima da capacidade de suporte, apontando a existência de dinâmicas populacionais diferenciadas durante alguns períodos específicos.

Outros estudos baseados na análise conjunta de datações por radiocarbono e de dados moleculares de marcadores genéticos corroboram o padrão de aceleração do crescimento populacional no período de 7.000 a 5.000 AP na Patagônia (PEREZ et al., 2016), assim como evidenciam uma diferenciação espaço-temporal das taxas de crescimento populacional, de modo que, por exemplo, as regiões centro e sul dos Andes apresentam um crescimento precoce e mais acentuado a partir de 9.000 AP, enquanto que no noroeste da Patagônia o crescimento é mais lento e tardio, com início entre 7.500 e 7.000 AP (PEREZ; POSTILLONE; RINDEL, 2017). Tais observações também se alinham aos resultados da análise do banco curado de Prates et al., que apontam como pontos iniciais de adensamento populacional a parte central e meridional dos Andes, assim como a região dos Pampas, além de também indicarem que as maiores taxas de crescimento populacional no período inicial do povoamento (15.100-11.500 AP) ocorrem nas porções central e setentrional dos Andes (PRATES; POLITIS; PEREZ, 2020).

Revertendo um padrão de clima mais seco e maior variabilidade de precipitação do Holoceno médio (DEININGER et al., 2019), no fim do Holoceno médio (~6.000 AP) teve início um aumento consistente da precipitação nas florestas tropicais do Hemisfério Sul, uma tendência de longo prazo que se mantém até o presente (IRIARTE et al., 2017). Se por um lado o período mais seco do Holoceno médio é coetâneo a um decréscimo populacional (iniciado a ~8.600 AP) (RIRIS; ARROYO-KALIN, 2019), por outro lado o aumento da precipitação levou a uma expansão das florestas tropicais entre 5.000 e 1.000 AP, particularmente da porção sul da

floresta amazônica, o que inclusive possui implicações sobre a ocorrência da Expansão Tupí (IRIARTE et al., 2017), conforme será discutido posteriormente. Assim sendo, a aceleração da taxa de crescimento populacional é concomitante a uma transição para climas mais úmidos e também para um cenário de aumento do sedentarismo e do aumento da relevância da agricultura como estratégia de subsistência (GOLDBERG; MYCHAJLIW; HADLY, 2016), algo evidenciado pelas modificações da paisagem, que se tornam muito mais frequentes durante o Holoceno tardio, as quais eram realizadas intencionalmente ou não, mas de todo modo intensificaram a produção, o manejo e a coleta de alimentos (IRIARTE et al., 2020).

Ademais, é importante destacar que por volta de 9.000 e 8.000 AP as últimas espécies da megafauna foram extintas[40] (BORRERO, 2009; MARTÍNEZ et al., 2016), e foi aproximadamente nesse período também que teve início a domesticação de plantas e animais, a qual estava completa no caso de algumas espécies a aproximadamente 6.000 AP (LARSON et al., 2014) embora os primeiros sinais do uso de diversas espécies de plantas na dieta apareçam ainda no final do Pleistoceno e início do Holoceno (IRIARTE et al., 2020) e o cultivo de mandioca (*Manihot* sp.) a 10.400 AP e abóbora (*Cucurbita* sp.) a 10.300 AP (LOMBARDO et al., 2020). De todo modo, o consequente advento da produção de alimentos não conduziu diretamente a um incremento abrupto das taxas de crescimento populacional, o que teria ocorrido apenas com a transição da produção de alimentos como uma estratégia suplementar para uma consolidação de sistemas de agricultura intensiva em alguns centros culturais, o que por sua vez produziu um aumento significativo da capacidade de suporte desses ambientes específicos (PEREZ; POSTILLONE; RINDEL, 2017; SUTTER, 2021). Por fim, no caso específico da Amazônia, alguns estudos sugerem que este crescimento exponencial se sustentou até a chegada dos europeus (ARROYO-KALIN, 2018), porém algumas linhas de evidências vem apontando para uma desaceleração no período pré-contato, talvez novamente causada pelo tamanho populacional ter alcançado o limite da capacidade de suporte desses ambientes (ARROYO-KALIN; RIRIS, 2021). Uma análise recente de pólen fossilizado de 39 lagos amazônicos vai além e indica que o período mais intenso de desocupação das terras da região e intensificação do crescimento das florestas ocorreu entre 300 a 600 anos antes do contato com europeus, o que supostamente teria sido desencadeado por fatores como epidemias de doenças autóctones, mudanças climáticas e/ou sociais (BUSH et al., 2021). Todavia análises genéticas e arqueológicas apontam que o ponto mais intenso do processo de depopulação e mortalidade dos nativos americanos não ocorre no período pré-contato e provavelmente não imediatamente após a chegada dos europeus nas Américas, mas sim em um período mais tardio durante a colonização, com

[40] A extinção da megafauna provavelmente sofreu o maior impacto a partir de 12.000 AP (METCALF et al., 2016).

variações entre diferentes regiões subcontinentais (e.g. (BROWNING et al., 2018; JONES et al., 2021)).

No que concerne a história populacional, o primeiro marco temporal é representado pela divergência genética entre os grupos que povoaram as três principais regiões ecogeográficas do continente, ao menos da porção norte deste, são elas a costa do Pacífico, os Andes e a Amazônia, o que teria ocorrido por volta de ~12.000 AP (HARRIS et al., 2018), relativamente cedo na ocupação da América do Sul. Essa estruturação inicial entre ecorregiões distintas foi certamente modificada ao longo do tempo, sobretudo pela intensificação de movimentos populacionais, particularmente em três períodos principais: (i) no decorrer do Holoceno médio e tardio com as expansões demográficas e territoriais desencadeadas pela sedentarização e intensificação da produção de alimentos, com movimentos internos e fluxo de fora para dentro do continente (GOLDBERG; MYCHAJLIW; HADLY, 2016; POSTH et al., 2018; SUTTER, 2021); (ii) durante o período mais recente de desenvolvimento de civilizações agrícolas e socialmente hierarquizadas, as quais se mostraram bastante disruptivas em termos das dinâmicas populacionais, provocando a miscigenação, homogeneização e migração de diversos povos indígenas, como no casos dos impérios Tiwanaku, Wari, e Inca (HARRIS et al., 2018; BORDA et al., 2020); e (iii) a partir da invasão das Américas por colonizadores europeus e o consequente impacto causado na distribuição, densidade e dispersão de grupos indígenas (ADHIKARI et al., 2017; ONGARO et al., 2019).

A divisão entre estas três regiões, principalmente a divisão entre os Andes e a Amazônia, por um lado se mostrou um modelo útil para compreender diversos aspectos e padrões etnolinguísticos, arqueológicos, demográficos e genéticos, mas por outro lado limitou durante muitas décadas e gerou um viés na forma como os povos indígenas da América do Sul são estudados, particularmente no que se refere a diversidade genética (BARBIERI, 2020; FEHREN-SCHMITZ, 2020; SANTOS, 2020). Nessa perspectiva foi concebido um modelo evolutivo das populações nativas da América do Sul para o período pré-contato (TARAZONA-SANTOS et al., 2001), o qual postula um padrão antagônico de intensidade de forças evolutivas, sobretudo a deriva genética e o fluxo gênico, na história dos grupos de terras altas e dos grupos de terras baixas, grosso modo correspondendo a nativos da porção central dos Andes e nativos da Amazônia (e Platô central brasileiro), respectivamente. Essencialmente, esse modelo se baseia na observação de que os habitantes das terras altas apresentam uma diversidade genética intrapopulacional maior e evidências de um amplo fluxo gênico entre grupos com uma consequente baixa diferenciação genética entre eles, ao passo que os habitantes de terras baixas exibem uma baixa diversidade genética intrapopulacional, além de baixo fluxo gênico, de modo que existe uma maior diferenciação genética entre grupos (TARAZONA-SANTOS et al., 2001; FUSELLI et al., 2003; BARBIERI et al., 2014b; SANTOS, 2020).

Por sua vez, esses padrões seriam resultantes de histórias demográficas distintas entre as duas regiões, (i) de um lado os Andes sendo ocupado por grandes populações com produção intensiva de alimentos, e eventualmente dando origem a sociedades hierarquizadas e a um sistema interconectado de estradas, com uma relativa homogeneidade ambiental e sociocultural; (ii) por outro lado a Amazônia sendo habitada por grupos pequenos predominantemente caçadores-coletores, isolados, ocupando ambientes bastante heterogêneos e com ampla diversificação cultural interpopulacional (TARAZONA-SANTOS et al., 2001; FUSELLI et al., 2003; BARBIERI et al., 2014b; SANTOS, 2020). Entretanto, essas diferenças de complexidade sociocultural e demográfica vem sendo revistas de acordo com novas evidências de uma ocupação mais densa e da existência de sociedades mais complexas na Amazônia (HECKENBERGER et al., 2003; HECKENBERGER; NEVES, 2009; ROOSEVELT, 2013; DE SOUZA et al., 2019; BERESFORD-JONES; MURILLO, 2020). Além disso, estudos mais recentes baseados em dados de genotipagem de alta densidade inferem a ocorrência de fluxo gênico entre populações viventes dos Andes e da Amazônia (BARBIERI et al., 2014b, 2019; HARRIS et al., 2018; BARBIERI, 2020) e inclusive apontam um fluxo gênico assimétrico com origem em grupos andinos contribuindo com ~5% da ancestralidade de alguns grupos da Amazônia peruana (GNECCHI-RUSCONE et al., 2019a), evidenciando assim que ao menos um modelo de divisão abrupta entre os Andes e a Amazônia é equivocado.

Uma dimensão adicional dessa paisagem de diversidade dos povos indígenas é resultante da interação entre estes grupos humanos e a imensa diversidade ambiental do continente sul-americano, o que produziu uma série de adaptações aos diversos contextos ecogeográficos locais, mediadas tanto pela evolução biológica, quanto cultural, o que em última instância moldou a diversidade genética e cultural das populações indígenas antigas e atuais (FEHREN-SCHMITZ, 2020). Esse processo dinâmico proporcionou o surgimento de uma das maiores diversidades culturais do mundo (SALZANO et al., 1988), com a presença de diversas famílias linguísticas representadas por centenas de línguas faladas ainda hoje (DIXON et al., 1999; RODRIGUES, ARYON D. & CABRAL, ANA SUELLY A. C., 2012a). Além disso, os grupos humanos não apenas se adaptam aos seus habitats, como também os modificam, não necessariamente de forma intencional, de modo que estes passam a atender melhor às suas necessidades e ao seu modo de vida. Desta maneira tanto caçadores-coletores quanto horticultores e agriculturalistas estabelecem processos de construção de nicho, essenciais para a sua sobrevivência (HÜNEMEIER et al., 2012b; FLORES; LEVIS, 2021), o que inclui por exemplo a domesticação e o cultivo de espécies vegetais, assim como o manejo de espécies semi-domesticadas (LARSON et al., 2014; IRIARTE et al., 2020; FLORES; LEVIS, 2021), além da

formação das terras pretas da Amazônia[41] (i.e. terras pretas de índio) (MCMICHAEL et al., 2014), as quais são geralmente encontradas em terras mais altas e bem drenadas das chamadas terras firmes (terras não inundáveis e não ribeirinhas) (KERN et al., 2003), e que são tão antigas quanto 6.000 AP porém se tornam mais amplamente distribuídas pela Amazônia somente a ~2.500 AP (IRIARTE et al., 2020).

Contraditoriamente, algumas linhas de evidências apontam que os nativos amazônicos não modificaram de forma muito significativa os ambientes florestais, conforme indicado pela ausência de indícios de impacto da ação humana nos últimos 5.000 AP em grande parte das florestas das terras firmes das bacias hidrográficas do Médio Putumayo-Algodón no nordeste do Peru (PIPERNO; MCMICHAEL, 2021). Entretanto, deve-se salientar, que os povos amazônicos no período pré-contato praticavam o que ficou conhecido como policultura agroflorestal, a qual combina o cultivo de plantas domesticadas com o manejo de espécies semi-domesticadas em meio aos ambientes florestais, o que é historicamente bastante diferente das práticas de outras populações agriculturalistas do mundo, as quais geralmente envolviam uma ênfase na monocultura de uma ou poucas espécies cereais em ambientes homogêneos (ou artificialmente homogêneos), ou mesmo das práticas atuais de alguns grupos indígenas que incluem o corte e a queima da floresta (NEVES, 2013; IRIARTE et al., 2020), de modo que isso potencialmente explicaria o impacto reduzido causado pelos nativos amazônicos aos ambientes de floresta amazônica. Com efeito, diversas linhas de evidências indicam que os povos indígenas contribuíram para o enriquecimento e diversidade de plantas comestíveis e medicinais não apenas na Amazônia, mas também na Mata Atlântica. Alguns exemplos de plantas cuja distribuição e abundância foram modificadas pela ação humana são a castanha-do-pará (*Bertholletia excelsa*), o pequizeiro (*Caryocar* spp.), o cacaueiro (*Theobroma cacao*) e a araucária (*Araucaria angustifolia*) (FLORES; LEVIS, 2021).

Apesar dessa imensa diversidade ambiental e cultural, não há uma equivalência dos níveis de diversidade genética, os quais se apresentam extremamente baixos em comparação aos de populações dos demais continentes. Essa baixa diversidade genética se expressa entre outras formas como um baixo nível de heterozigosidade, o qual diminui num gradiente que se inicia no norte da América do norte e vai até o sul da América do Sul (WANG et al., 2007; REICH et al., 2012), ao passo que na América do Sul um segundo gradiente se direciona do oeste para o leste, ambos muito provavelmente remontando aos eventos iniciais de povoamento e aos gargalos de garrafa populacionais em série enfrentados por estes grupos, isto porque o efeito da deriva genética em grupos pequenos e isolados é maior, o que leva a um aumento da taxa de perda da diversidade genética, de modo que a Amazônia é muito possivelmente o lar das

[41] *Amazonian Dark Earths*.

populações viventes com os menores níveis de heterozigosidade do mundo (FEHREN-SCHMITZ, 2020; SANTOS, 2020). Ao contrário da diversidade genética, a diferenciação genética entre populações tende a aumentar de norte a sul nas Américas e de oeste a leste na América do Sul (WANG et al., 2007; O'ROURKE; RAFF, 2010; REICH et al., 2012; VERDU et al., 2014).

Dinâmicas populacionais e expansões dêmicas do Holoceno tardio

Apesar de estudos genéticos em escala continental das populações indígenas da América do Sul não terem encontrado uma relação clara entre a distribuição da variabilidade genética e cultural (HUNLEY et al., 2007; ROEWER et al., 2013; BISSO-MACHADO; FAGUNDES, 2021), é possível identificar alguns casos onde a cultura influenciou os padrões genéticos de forma significativa, especialmente em contextos regionais. Um bom exemplo dessa relação entre cultura e genética pode ser observada nos grupos das terras baixas, sobretudo nos Xávante do platô central brasileiro, que apresentam uma dinâmica populacional de eventos de fissão-fusão motivados por fatores culturais, que consiste na divisão e migração não-aleatória de grupos, os quais podem evoluir isoladamente ou se fundir novamente ao grupo original, ou ainda a outros grupos da mesma etnia (NEEL; SALZANO, 1967). Além disso, esses grupos tendem a ser altamente endogâmicos, fatores que em conjunto acarretam uma aceleração da taxa de evolução genotípica e fenotípica, conforme evidenciado pela rápida evolução da morfologia craniana dos próprios Xávante (HÜNEMEIER et al., 2012a).

Muitos padrões de estrutura genética hoje observados estão relacionados a fatores culturais subjacentes, sobretudo influenciados pelas estratégias de subsistência, sendo que o processo de transição e intensificação da produção de alimentos por meio da agricultura (acompanhada ou não do desenvolvimento de outras tecnologias como criação de animais, pesca, etc.) acarretou muitas vezes uma aceleração da taxa de crescimento populacional, com consequente aumento do tamanho populacional e difusão dêmica desses povos, um padrão bem documentado e recorrente na história humana (SOKAL; ODEN; WILSON, 1991; CORDAUX, 2004; WEN et al., 2004; DE FILIPPO et al., 2012; AMMERMAN; CAVALLI-SFORZA, 2014), o qual também é evidenciado pela distribuição das maiores famílias linguísticas do mundo (BELLWOOD; OTHERS, 2005).

Dessa forma, o panorama genético relativamente homogêneo das populações andinas seria o resultado não somente das taxas mais altas de fluxo gênico, como também de eventos de expansão dêmica de populações agriculturalistas no passado (BARBIERI et al., 2017, 2019), os quais se infere que teriam se originado na costa do Pacífico (STANISH, 2001), de maneira que ainda hoje grupos habitantes da região central dos Andes falantes de línguas Uro, e derivados de

grupos caçadores-coletores, são geneticamente diferenciados de grupos historicamente agriculturalistas e falantes de línguas Aymara e Quechua da mesma região (SANDOVAL et al., 2013).

Igualmente, esses eventos de expansão dêmica também ocorreram entre povos da Amazônia, associados a grupos com maior prevalência de estratégias de subsistência agriculturalistas, de tal modo que a estrutura genética de populações viventes coincide com as expectativas da ocorrência de expansões no passado. Como exemplo, os grupos agriculturalistas falantes de línguas Tupí apresentam um padrão de isolamento por distância indicativo de uma expansão populacional, ao passo que infere-se um padrão não-linear de dispersão para os grupos falantes de línguas Jê e predominantemente caçadores-coletores, um padrão que é discordante da expectativa da ocorrência de expansões dêmicas no passado (RAMALLO et al., 2013).

Ainda nesse sentido, as maiores famílias linguísticas da América do Sul possuem uma extensa distribuição geográfica no continente (DIXON et al., 1999) e desde muito tempo foi proposta uma correlação entre estas famílias e algumas das tradições[42] de cultura material mais importantes do Holoceno tardio, as quais também apresentam grande amplitude geográfica (NEVES, 2011), contudo, não há uma correspondência exata entre elas. De todo modo, as evidências suportam as hipóteses de que pelo menos algumas dessas tradições teriam se expandido de fato por meio de difusões dêmicas (NOELLI, 2008; GREGORIO DE SOUZA; ALCAINA MATEOS; MADELLA, 2020). Efetivamente, as quatro maiores famílias linguísticas da América do Sul, nomeadamente Arawak, Karib, Jê (parte do tronco Macro-Jê) e Tupí-Guaraní (parte do tronco Tupí), estariam relacionadas às tradições (GREGORIO DE SOUZA; ALCAINA MATEOS; MADELLA, 2020): Saladóide-Barrancóide (LATHRAP, 1970; BROCHADO, 1984), Inciso Ponteada (LATHRAP, 1970), Una (BROCHADO, 1984; NOELLI, 2005) e Tupiguarani (NOELLI, 2008; CORRÊA, 2014), respectivamente. Além disso, as famílias linguísticas possuem inerentemente um caráter expansivo e divergente, independente do mecanismo envolvido na sua expansão geográfica ser uma expansão dêmica ou uma difusão cultural[43]. De modo que, essa expansão geográfica implica necessariamente na migração de pelo menos alguns falantes da língua, os quais podem no limite levar os habitantes de outras regiões a um processo de substituição linguística, no qual a língua originalmente falada é substituída pela língua dos migrantes, seja por motivos culturais ou demográficos (HEGGARTY, 2020).

Dentre as quatro tradições citadas, as únicas a se expandirem para fora da Amazônia foram a Saladóide-Barrancóide e a Tupiguarani, a primeira atingiu até as ilhas de Porto Rico e

[42] Simplificadamente se refere a um conjunto de estilos e técnicas de produção cultural com ampla persistência temporal e amplitude espacial.

[43] Disseminação de traços culturais, práticas, tradições, hábitos, técnicas, idéias, estilos, religiões ou línguas entre indivíduos de um mesmo grupo cultural ou não.

Hispaniola no Caribe (KEEGAN, 1995), enquanto que a segunda se expandiu por um raio de mais de 4.000 quilômetros por boa parte das terras baixas amazônicas e não-amazônicas, incluindo o cerrado da região central brasileira, zonas de caatinga do Nordeste, uma porção significativa da costa do Atlântico, áreas de mata atlântica do Sudeste e do Sul, até as regiões dos pampas argentinos (NOELLI, 1998, 2008). A tradição Tupiguarani pode ainda ser divida em três sub-tradições principais, que seriam a Guaraní, Tupinambá da Amazônia e Tupinambá da Mata Atlântica, estando predominantemente circunscritas à bacia do Paraná, ao sudeste da Amazônia e ao litoral atlântico, respectivamente, e portanto ocorreram em regiões geográficas distintas (ALMEIDA; NEVES, 2015). Da mesma forma, línguas das famílias Arawak (DAVIS; GOODWIN, 1990) e Tupí-Guaraní (URBAN, 1992; RODRIGUES, ARYON D. & CABRAL, ANA SUELLY A. C., 2012b) se expandiram para as regiões correspondentes a das respectivas tradições associadas e eram faladas por povos nativos dessas regiões até o contato com europeus iniciado em 1492. De fato, há muito tempo se percebeu a similaridade linguística e cultural entre os Guaraní[44] e os Tupinambá[45] que permitiria a unificação destes num único grupo, o qual foi chamado de Tupí-Guaraní[46] (NOELLI, 1998, 2008). Desta forma, a família linguística com a distribuição geográfica mais ampla no território brasileiro é a Tupí-Guaraní, a qual integra o tronco linguístico Tupí em conjunto com outras nove famílias linguísticas restritas a populações amazônicas (URBAN, 1992; RODRIGUES, ARYON D. & CABRAL, ANA SUELLY A. C., 2012b).

Atualmente, é um consenso que toda a imensa diversidade das famílias linguísticas do tronco Tupí teve um centro de origem comum (NOELLI, 2008), ao passo que um conjunto de fortes evidências de análises linguísticas (e.g. (WALKER et al., 2012)), também suportados por dados arqueológicos (e.g. (MILLER, 2009)) e genéticos (RAMALLO et al., 2013; SANTOS et al., 2015) apontam como centro de diversidade e de dispersão dos povos falantes de línguas Tupí a região sudoeste da Amazônia, mais especificamente a região entre os rios Madeira e Guaporé (conhecida como Região Madeira Guaporé), localizada na bacia do rio Madeira. Em tempo, seguindo a mesma lógica o centro de dispersão dos Tupí-Guaraní por sua vez parece ter sido a região entre e próxima aos rios Xingu e Tocantins no sudeste amazônico, isto porque reúne a maior diversidade linguística e de cultura material do grupo, além de indícios consistentes de uma longa ocupação, sendo que a 1.000 AP a região já seria extensamente ocupada (ALMEIDA; NEVES, 2015). Entretanto, essa hipótese ainda não pode ser analisada e corroborada em termos da distribuição da diversidade genética.

[44] Grupos Tupí-Guaraní do Centro Oeste e Sul do Brasil.

[45] Grupos Tupí-Guaraní do litoral atlântico e da região amazônica. É também o nome de uma etnia específica do litoral brasileiro.

[46] Os nomes de famílias linguísticas ou das próprias línguas as vezes são utilizados para se referir aos seus respectivos grupos de falantes.

A dispersão dos Tupí-Guaraní a partir do centro de origem teria iniciado a aproximadamente 2.400 AP e muito rapidamente atingido tanto a bacia do Paraná a 2.200 AP quanto a costa do Atlântico a 1.800 AP, segundo a distribuição dos sítios arqueológicos mais antigos (NOELLI, 2008; MACARIO et al., 2009). Ao longo de mais de um século de pesquisa arqueológica, linguística, historiográfica e etnográfica, diversas rotas para a dispersão dos Tupí-Guaraní foram propostas, assim como os fatores climáticos, socioculturais e/ou demográficos envolvidos.

Nesse sentido, o antropólogo Alfred Métraux, baseado em dados históricos e de linguística histórica, propôs um modelo segundo o qual os Tupí-Guaraní teriam se utilizado das extensas conexões fluviais da bacia amazônica para se dispersar rapidamente para a região sul e a bacia do Paraná e a partir dessa região eles teriam ocupado o litoral atlântico, sendo que esse evento seria próximo temporalmente do contato com os europeus, tendo ocorrido apenas algumas centenas de anos antes (MÉTRAUX, 1927). É importante destacar que segundo Métraux o processo teria ocorrido sob a forma de migrações, o que significa dizer que as regiões originalmente ocupadas teriam sido abandonadas em busca de novos territórios. Esse modelo foi então mais tarde reinterpretado para incluir a ocorrência de mudanças climáticas como fator impulsionador da dispersão, as quais teriam provocado uma fragmentação da floresta e levado a uma busca por refúgios mais adequadas aos modos de vida e estratégias de subsistência desses grupos em regiões ao sul da Amazônia, e portanto assim desencadeado as migrações para a bacia do Paraná e assim por diante (MEGGERS, 1974, 1977, 1982; MEGGERS AND CLIFFORD, 1978).

Alternativamente o arqueólogo Donald Lathrap propôs o que ficou conhecido como o "modelo cardíaco" para a dispersão dos povos Tupí-Guaraní, o qual propunha uma expansão radial gradativa e mais profunda temporalmente, todavia também por meio das redes fluviais, a partir de um centro de dispersão amazônico, e se baseia em dados linguísticos, etnográficos e sobretudo arqueológicos (LATHRAP, 1970). Este modelo foi posteriormente aprimorado pelo arqueólogo brasileiro José Brochado, que buscando estabelecer relações entre os dados linguísticos e a evolução da cultura material, propôs que a diferenciação das línguas e tradições cerâmicas teria sido uma consequência da divergência e diferenciação dos Proto-Tupí, a qual por sua vez teria sido causada por um crescimento demográfico contínuo na região central amazônica causado pela produção de alimentos por meio da agricultura (BROCHADO; LATHRAP, 1982; BROCHADO, 1984). Dado que o caráter do evento proposto é o de uma expansão dêmica, isso implica que os territórios originalmente ocupados pelos Tupí-Guaraní não foram abandonados no processo, além disso esses grupos teriam se expandido em um movimento de pinça a partir de um centro de dispersão no sudeste amazônico, de modo que um dos ramos se dirigiu ao sul e a bacia do Paraná, e o outro em direção ao leste seguindo o curso

do Rio Amazonas, até a foz e depois em direção ao litoral atlântico, o qual teria sido ocupado em seguida até a região de Cananéia no litoral de São Paulo, respectivamente o ramo do sul daria origem então aos Guaraní e o ramo do leste aos Tupinambá (BROCHADO; LATHRAP, 1982; BROCHADO, 1984).

Em suma, esses dois modelos possuem premissas fundamentalmente distintas no que diz respeito a pontos centrais para a compreensão da dispersão dos povos Tupí-Guaraní, principalmente a forma com que os movimentos populacionais ocorreram, a profundidade temporal desses eventos e os fatores desencadeadores do processo. Entretanto, ao mesmo tempo eles concordam no entendimento de que tanto as línguas, quanto a cultura material foram dispersas por meio de um movimento populacional e não simplesmente pela difusão cultural. Um segundo ponto de concordância aqui citado é a visão de que estes deslocamentos teriam ocorrido por meio de rotas fluviais, o que vem sendo contestado por uma série de evidências que mostram uma ocupação e reocupação de terras firmes por parte dos Tupí-Guaraní em uma longa escala temporal, sobretudo na região do sudeste da Amazônia, entre os rios Tocantins e Xingu, indicando que rotas terrestres também tem de ser consideradas nos modelos que buscam explicar o fenômeno da expansão Tupí-Guaraní (ALMEIDA; NEVES, 2015).

É importante destacar também que essas regiões por onde os Tupí-Guaraní se expandiram não estavam desocupadas antes da chegada deles, sendo que o litoral atlântico, assim como algumas áreas ribeirinhas eram habitados por populações de pescadores-coletores construtoras de sambaquis[47] desde aproximadamente 8.000 AP (GASPAR et al., 2008). É possível que a ocupação seja ainda mais antiga, dado que muitos sambaquis podem ter sido formados sobre a plataforma continental, e atualmente estariam ocultos sob o oceano Atlântico (GASPAR et al., 2008). Estes grupos sambaquianos, como são conhecidos, teriam portanto ocupado o litoral brasileiro desde pelo menos 8.000 AP até a chegada de grupos Tupí-Guaraní e também de falantes de línguas Macro-Jê, não necessariamente nessa ordem, como evidenciado pela presença de cerâmicas das tradições Tupiguarani e Taquara/Itararé[48] nas camadas mais superiores de alguns sambaquis, respectivamente (GASPAR et al., 2008).

[47] Termo derivado do Tupi usado no Brasil para se referir aos "montes de conchas" ou "concheiros" amplamente distribuídos no litoral atlântico, assim como em lagunas e rios brasileiros. Os sambaqui são depósitos culturais com tamanho e estratigrafias variáveis, além de possuírem origens e funções diversas. Efetivamente existem concheiros em diversas partes do mundo, sendo chamados em inglês de *shell mounds* ou *middens*, entre outros termos.

[48] A tradição Taquara/Itararé é associada a grupos Jê do sul do Brasil, como os Xokleng e Kaingang (GASPAR et al., 2008).

Sub-representatividade de populações indígenas em estudos genéticos

Como foi demonstrado aqui a história dos povos americanos é extremamente profunda e complexa, se iniciando com a formação e entrada dos primeiros grupos nativos americanos ainda no Pleistoceno tardio a partir do nordeste asiático e da Beríngia, passando pela entrada no continente através da Beríngia, dispersão e diferenciação desses grupos por todo o continente, com alguns fluxos populacionais adicionais contribuindo especificamente para populações nativas do norte da América do Norte, prosseguindo com as dinâmicas e movimentos populacionais do Holoceno, os quais foram amplificados pela intensificação da produção de alimentos com a domesticação de plantas e animais, e o desenvolvimento da agricultura a partir do Holoceno médio, chegando ao desenvolvimento de sociedades agrícolas hierarquizadas que se mostraram bastante disruptivas em suas zonas de influência, sobretudo na Mesoamérica e na porção central dos Andes durante o Holoceno tardio, até por fim os eventos catastróficos causados e desencadeados pela invasão européia, que gerou uma reconfiguração gigantesca da distribuição, diversidade e no tamanho das populações indígenas.

Apesar desse panorama uma série de fatores inibiram e continuam prejudicando o estudo da diversidade genética de populações americanas, principalmente o fato de que historicamente populações não-europeias, sobretudo as indígenas, foram negligenciadas e continuam largamente sub-representadas no esforço global de estudo da diversidade genômica humana. Efetivamente, até o ano de 2009 os estudos de associação ampla do genoma[49] contavam com uma maioria absoluta de 96% (de um total de 1,7 milhões de amostras) de participantes de ascendência europeia (NEED; GOLDSTEIN, 2009). Esta situação não foi muito aprimorada até o ano de 2016, quando apenas 19% (de um total de 35 milhões de amostras) dos participantes possuíam ascendência não-europeia, ainda assim correspondendo predominantemente a indivíduos de origem asiática (POPEJOY; FULLERTON, 2016). No presente momento, segundo o GWAS Diversity Monitor[50] a proporção total de participantes europeus, asiáticos e africanos é de 88,64%, 7,11% e 0,33%, respectivamente, sendo o restante composto por afro-americanos e afro-caribenhos (0,85%), hispânicos ou latino-americanos (0,78%) e miscigenados ou outros (2,28%) (MILLS; RAHAL, 2020), demonstrando que uma

[49] *Genome-wide association studies* (GWAS), consiste basicamente na comparação das frequências de marcadores genéticos e caracteres fenotípicos entre indivíduos ou grupos, buscando identificar uma associação entre marcadores e fenótipos de interesse, como por exemplo fatores genéticos relacionados a doenças.

[50] Site consultado em 5 de Julho de 2021; https://gwasdiversitymonitor.com/.

posição absolutamente secundária é relegada às populações indígenas das Américas, da África e da Oceania.

A sub-representação dos nativos americanos nos painéis de dados genômicos globais, produz uma série de distorções, como por exemplo viés na aplicação de resultados de estudos de associação de marcadores genéticos a fenótipos, inclusive de fenótipos com relevância médica, o que se deve ao fato de que boa parte destes estudos foi conduzida em populações europeias, nas quais a frequência e a distribuição das variantes genéticas difere significativamente daquelas de populações indígenas de outras partes do globo. Exemplificadamente, devido a essas diferenças genéticas entre populações distintas, um marcador genético pode estar associado a um risco maior ou menor, de acordo com a ancestralidade do indivíduo ou até mesmo do segmento genômico no qual se localiza o marcador genético específico. Este viés de amostragem também significa que muitas associações relevantes entre caracteres fenotípicos e variantes genéticas estão sendo ignoradas (POPEJOY; FULLERTON, 2016).

Nesse sentido, a falta de representatividade de populações não-europeias pode levar a uma menor efetividade e até a risco na aplicação de certas drogas (BURKE, 2021). Soma-se a isto o fato de que as populações mais prejudicadas pela sub-representação tendem a ser aquelas originárias de regiões do globo mais periféricas e afetadas por fatores como baixo desenvolvimento econômico, problemas sociais e ambientais (POPEJOY; FULLERTON, 2016). Uma camada adicional do problema da baixa diversidade e falta de representatividade vem do fato de que a pesquisa em genômica é predominantemente realizada por institutos de pesquisa e universidades europeias e americanas, e os grupos de pesquisa também majoritariamente compostos por pessoas de ascendência europeia (BONHAM; GREEN, 2021), o que certamente produz ainda mais distorções e assimetrias no que diz respeito às abordagens utilizadas, aos objetivos das pesquisas e até ao benefício gerado para as populações estudadas. Desta forma, estas populações sub-representadas ainda têm o acesso bastante limitado aos avanços proporcionados por pesquisas em genômica, como serviços de testes genéticos, estimativas de risco de doenças e tratamentos médicos advindos destes estudos (BURKE, 2021).

Nesta perspectiva, a subamostragem de populações nativas das Américas, e sobretudo de amostras de indivíduos antigos, é particularmente problemática para o estudo da história destas populações, dado que peças essenciais do quebra-cabeças formado pelos eventos do passado certamente estão sendo completamente ignoradas por estarem ausentes nos dados disponíveis, um fato que na última década foi reiteradamente demonstrado pelas descobertas científicas resultantes do estudo de novas amostras (e.g. (REICH et al., 2012; SKOGLUND et al., 2015; POSTH et al., 2018)).

Esta tese portanto faz parte de um esforço para aumentar a representatividade de estudos sobre a diversidade genética das populações indígenas e não-indígenas brasileiras, essencialmente através (i) de uma amostragem mais ampla de povos nativos (e miscigenados) com distribuição por diferentes ecorregiões brasileiras; (ii) produção de dados genômicos a partir das melhores técnicas disponíveis, em quantidade e qualidade suficientes para permitir o estudo adequado destas populações; e (iii) finalmente pela análise da diversidade genética e produção de conhecimento acerca de aspectos históricos, demográficos e evolutivos.

OBJETIVOS

Nesta tese buscamos analisar os dados inéditos de populações indígenas brasileiras de diferentes ecorregiões (Amazônia, Cerrado e Mata Atlântica) combinados a um conjunto de dados públicos de populações indígenas contemporâneas de todo o continente americano (Mesoamérica, costa do Pacífico e Andes), incluindo também indivíduos antigos, de modo a permitir o estudo da formação da paisagem de diversidade genética nativa americana em diversas amplitudes espaço-temporais: (i) desde o povoamento inicial das Américas, e principalmente da América do Sul, particularmente relacionado a contribuição genética da população Y, detectada como um excesso de afinidade genética com grupos australo-asiáticos, a qual havia sido observada até o momento apenas em algumas populações amazônicas do Brasil; (ii) passando por eventos de dispersão iniciais e diferenciação genética dos principais grupos sul-americanos no Holoceno, assim como pela influência dos processos dinâmicos de interação entre a cultura, a genética e os ambientes na formação da paisagem de diversidade genética do continente, e também efeito das interações entre as populações de diferentes regiões, desde a costa do Pacífico até a do Atlântico; (iii) chegando por fim às dinâmicas populacionais do Holoceno médio e tardio, à transição das estratégias de sobrevivência para uma maior proeminência da produção de alimentos em algumas regiões e às expansões dêmicas, com foco principal nos Tupi e sobretudo nos Tupí-Guaraní.

Objetivos específicos

Capítulo 1

- Investigar a distribuição geográfica da ancestralidade australo-asiática no continente sul-americano.
- Analisar a variabilidade do padrão de detecção dessa ancestralidade nos níveis intra e interpopulacional.
- Produzir modelos de história populacional capazes de explicar a presença dessa ancestralidade em populações nativas americanas contemporâneas.
- Delimitar através desses modelos quais são as rotas mais prováveis de entrada dessa ancestralidade nas Américas, assim como a cronologia mais provável do evento.

Capítulo 2

- Explorar os padrões de distribuição da diversidade e estruturação genética na América do Sul e a influência tanto da diversidade etnolinguística quanto da distribuição geográfica sobre as mesmas.
- Avaliar a influência de outros fatores sobre a distribuição da diversidade genética, incluindo o contato entre diferentes grupos etnolinguísticos com a ocorrência de trocas culturais e a existência de continuidades genéticas locais.
- Investigar como os padrões de distribuição da diversidade genética se relacionam às diferenças demográficas e históricas entre os Andes e a Amazônia.
- Inferir e comparar o histórico demográfico dos diferentes grupos linguísticos durante o fim do Holoceno tardio, de modo a determinar se as populações indígenas apresentavam um tamanho populacional estável ou se existem evidências da ocorrência de expansões populacionais.
- Estudar os impactos da invasão e colonização europeia sobre o nível de diversidade genética e distribuição da variação genética nas populações da América do Sul.

Capítulo 3

- Estimar a proporção de contribuição das diferentes ancestralidades continentais para a comunidade Tupiniquim do litoral brasileiro e datar os períodos de intensificação da miscigenação, buscando estabelecer relações com eventos relevantes do registro histórico.
- Inferir o histórico demográfico dos Tupiniquim, buscando examinar os impactos do contato com europeus e averiguar a existência de um crescimento populacional durante o período da Expansão Tupí.
- Identificar e isolar a porção genômica de ancestralidade nativa americana dos Tupiniquim, e analisar os padrões de afinidade e parentesco com os outros grupos indígenas contemporâneos e antigos, sobretudo aqueles do Brasil.
- Testar as hipóteses alternativas da Expansão Tupí propostas por Métraux e Brochado, de modo a identificar quais são as rotas e cronologia mais provável do evento, além de estabelecer um modelo para a história populacional nos Tupí-Guaraní.

CAPÍTULO 1

Afinidade genética profunda entre nativos da costa do Pacífico e da Amazônia evidenciada pela ancestralidade australo-asiática

Manuscrito: CASTRO E SILVA, M. A. et al. Deep Genetic Affinity between Coastal Pacific and Amazonian Natives Evidenced by Australasian Ancestry. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 14, 6 abr. 2021.

Autores: Marcos Araújo Castro e Silva[a], Tiago Ferraz[a], Maria Cátira Bortolini[b], David Comas[c] e Tábita Hünemeier[a]

[a]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, 05508-090 São Paulo, SP, Brazil; [b]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil; [c]Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain

Resumo:

Diferentes modelos foram propostos para elucidar as origens das populações fundadoras da América, em conjunto com o número de ondas migratórias e rotas utilizadas por estes primeiros exploradores. O povoamento tanto ao longo da costa do Pacífico quanto pelo continente tem sido evidenciado por estudos genéticos e arqueológicos. Entretanto, o número de ondas migratórias e a origem dos imigrantes continuam sendo tópicos controversos. Aqui nós mostramos que o sinal genético australo-asiático está presente na região da costa do Pacífico, indicando uma distribuição muito mais extensa dentro da América do Sul e implicando um contato antigo entre habitantes do Pacífico e da Amazônia. Nós ainda demonstramos que a contribuição da população australo-asiática foi introduzida na América do Sul através da rota da costa do Pacífico antes da formação do ramo amazônico, provavelmente pela antiga população ancestral dos grupos da costa do Pacífico e Amazônia. Este estudo elucida as relações genéticas entre diferentes componentes ancestrais no povoamento inicial da América do Sul e propõe que a rota migratória utilizada pelos migrantes portadores da ancestralidade australo-asiática levou a ausência deste sinal em populações da América Central e do Norte.

Palavras-chave: povoamento da América do Sul | Australo-asiático | genética | Nativos americanos | rota da costa do Pacífico

Abstract:

Different models have been proposed to elucidate the origins of the founding populations of America, along with the number of migratory waves and routes used by these first explorers. Settlements, both along the Pacific coast and on land, have been evidenced in genetic and archeological studies. However, the number of migratory waves and the origin of immigrants are still controversial topics. Here, we show the Australasian genetic signal is present in the Pacific coast region, indicating a more widespread signal distribution within South America and implicating an ancient contact between Pacific and Amazonian dwellers. We demonstrate that the Australasian population contribution was introduced in South America through the Pacific coastal route before the formation of the Amazonian branch, likely in the ancient coastal Pacific/Amazonian population. In addition, we detected a significant amount of interpopulation and intrapopulation variation in this genetic signal in South America. This study elucidates the genetic relationships of different ancestral components in the initial settlement of South America and proposes that the migratory route used by migrants who carried the Australasian ancestry led to the absence of this signal in the populations of Central and North America.

Keywords: settlement of South America | Australasian | genetics | Native Americans | Pacific coastal route

Deep genetic affinity between coastal Pacific and Amazonian natives evidenced by Australasian ancestry

Marcos Araújo Castro e Silva[a,1], Tiago Ferraz[a,1], Maria Cátira Bortolini[b], David Comas[c], and Tábita Hünemeier[b,2]

[a]Departmento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, 05508-090 São Paulo, SP, Brazil; [b]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil; and [c]Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain

Edited by Elsa M. Redmond, American Museum of Natural History, New York, NY, and approved February 17, 2021 (received for review December 14, 2020)

Different models have been proposed to elucidate the origins of the founding populations of America, along with the number of migratory waves and routes used by these first explorers. Settlements, both along the Pacific coast and on land, have been evidenced in genetic and archeological studies. However, the number of migratory waves and the origin of immigrants are still controversial topics. Here, we show the Australasian genetic signal is present in the Pacific coast region, indicating a more widespread signal distribution within South America and implicating an ancient contact between Pacific and Amazonian dwellers. We demonstrate that the Australasian population contribution was introduced in South America through the Pacific coastal route before the formation of the Amazonian branch, likely in the ancient coastal Pacific/Amazonian population. In addition, we detected a significant amount of interpopulation and intrapopulation variation in this genetic signal in South America. This study elucidates the genetic relationships of different ancestral components in the initial settlement of South America and proposes that the migratory route used by migrants who carried the Australasian ancestry led to the absence of this signal in the populations of Central and North America.

settlement of South America | Australasian | genetics | Native Americans | Pacific coastal route

A signal of genetic affinity between present-day and ancient natives from South America and present-day indigenous groups of South Asia, Australia, and Melanesia has been previously reported (1–4). This Australasian–Native American connection persists as one of the most intriguing and poorly understood events in human history. The controversial Australasian population genetic component (i.e., "Ypikuéra population" or "Y population" component) was identified exclusively in the present-day Amazonian populations (2), suggesting at least two different founding waves leading to the formation of the people of this region. The first wave was inferred to be composed of direct descendants of the Beringian standstill population, and a second wave was formed by an admixed population of Beringian and southeast Asian ancestors that reached Beringia more recently. Both these populations would have settled and admixed in the Amazon region.

The contribution of an unsampled population to the autochthonous gene pool is thought to have led to the origin of the Australasian shared ancestry (2). In this sense, the Y population would be part of the first colonizing groups of the American continent. However, data from ancient South American samples indicated a weak Y signal around 10,000 yBP (3). This evidence indicates that, rather than a second wave entering South America from southeast Asia, the Y ancestry might be traced back to common ancestors of Native Americans, who lived in northeast Asia. Furthermore, a new line of evidence indicates that the first American clades split in East Asia, not in Beringia, which makes the gene flow of the Y ancestry from the ancestral East Asian groups even more likely (5). However, the paucity of the signal among present-day and ancient groups, along with the endemic and apparently random pattern of detection, has raised the possibility that it could be a false-positive detection, likely due to the strong genetic drift effects experienced by the Amazonian populations (and other indigenous South Americans). However, it might be the other way around, a scenario in which the signal went below the significance level in some populations, due to the high drift effects they experienced (i.e., false negatives).

We explored our dataset (*SI Appendix*, *Extended Methods*), which is currently the most comprehensive set of genomic data from South American populations (383 individuals; 438,443 markers), to shed light on this question. Ethical approval for sample collection was provided by the Brazilian National Ethics Commission (CONEP Resolutions 123 and 4599). CONEP also approved oral consent for the use of these samples in population history and human evolution studies. Individual and/or tribal informed oral consent was obtained from participants who were not able to read or write.

Our results showed that the Australasian genetic signal, previously described as exclusive to Amazonian groups, was also identified in the Pacific coastal population, pointing to a more widespread signal distribution within South America, and possibly implicating an ancient contact between Pacific and Amazonian dwellers. In addition, a significant amount of interpopulation and intrapopulation variation of this genetic signal was detected.

To test the existence of this excess allele sharing, we calculated the D(Mbuti, Australasian; Y, Z) statistic for every pair of Y and Z indigenous groups or individuals in our dataset (Dataset S1*A*), where "Australasian" is also iterated over the Australasian groups, namely Australian (and Australian.DG), Melanesian, Onge (i.e., ONG.SG), and Papuan (6–9). In the tests between groups, signal detection was reproduced in Karitiana and Suruí (Amazonia), but it was also observed in Chotuna (Mochica descendants from the Pacific coast), Guaraní Kaiowá (central west Brazil), and Xavánte (Central Brazilian Plateau) (Dataset S3). When we used the maximum unrelated set of individuals (Dataset S1*A*), the signal lost significance level in Karitiana, Suruí, and Guaraní Kaiowá (Dataset S3). However, the signal was still evident in the Pacific coast population and in the central Brazilian natives (Fig. 1 and Dataset S3).

We also aimed to detect whether some individuals would present a higher number of significant tests than others from the same population, which could indicate a heterogeneous genetic ancestry within the positive populations. Our analysis showed that, indeed, some individuals presented a higher number of tests

Author contributions: T.H. designed research; M.C.B. and D.C. contributed new reagents/analytic tools; M.A.C.e.S. and T.F. analyzed data; and M.A.C.e.S. and T.H. wrote the paper.

The authors declare no competing interest.

[1]M.A.C.e.S. and T.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: hunemeier@usp.br.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2025739118/-/DCSupplemental.

Published March 29, 2021.

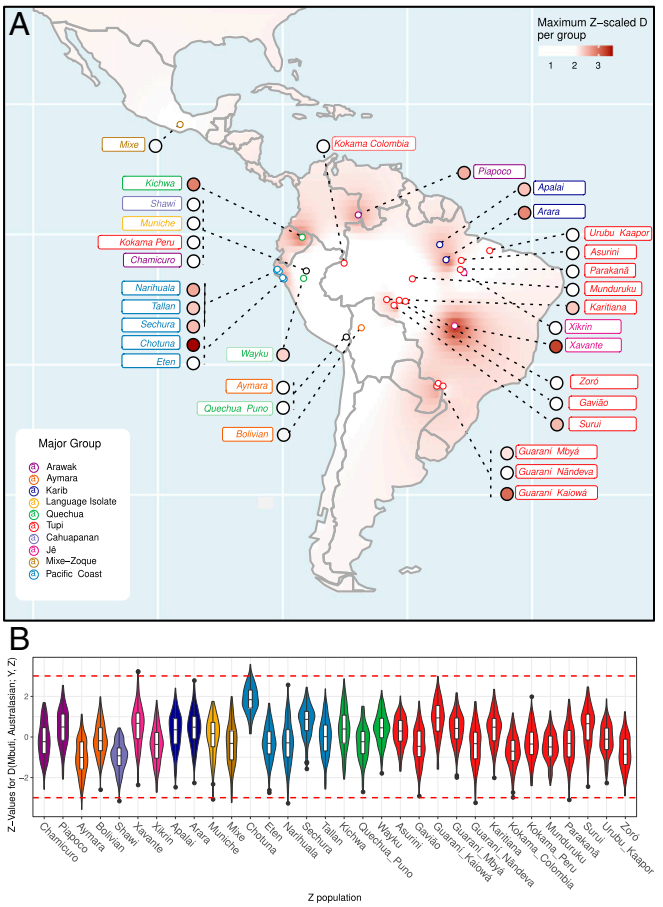


Fig. 1. Relative patterns of genetic affinity of Australasians among Native American groups. (*A*) Maximum Z values per population interpolated with the inverse distance weighting method. (*B*) Distribution of all estimated Z values (*y* axis) for each "Z" population (*x* axis) as violin and box plots. In *B*, the black dots represent outliers, and the red dashed lines indicate the Z-value thresholds of Z = −3 and Z = 3.

pointing to excess allele sharing, but also that some are more likely to present a significant deficit of this ancestry in comparison to the others (Fig. 2 and Dataset S4 *C* and *D*). From these results, it is evident that the loss of signal significance upon the shift from the complete set to the maximum unrelated set of samples (Dataset S3) was caused by the exclusion of specific individuals with higher levels of allele sharing with Australasians rather than by the removal of a bias caused by the relatedness among the tested samples in the first place.

This provides strong evidence that a significant variability of this signal exists not only at an interpopulation level but also between individuals from the same populations. These results suggest that the intrapopulation variability of this signal is not rare (Fig. 2) and is observed in several groups (Apalai, Guaraní Nãndeva, Karitiana, Munduruku, Parakanã, and Xavánte). Most significant tests detected this excess signal in Tupí-speaking individuals, but the signal was also detected in individuals from every major linguistic group (Fig. 2 and Dataset S4) and, at the same time, presented a widespread geographic distribution within South America (Fig. 1). Conversely, a considerable number of samples were inferred to have a deficit of allele sharing with Australasians (Fig. 2 and Dataset S4*D*). Strikingly, the individual PAR137 (Parakanã) presented an extremely high proportion of significant tests (31.64%), indicating a relative deficit. This individual is not an outlier neither in the principal component analysis of the Native American samples (Dataset S1 *B* and *C*), nor regarding its missingness rate (Dataset S1*A*), nor in a multidimensional scaling (MDS) of pairwise genetic distances between samples in the unrelated and unadmixed subset (Dataset S1*D*). Besides, the distribution of Y-population ancestry among present-day indigenous groups of South America showed no relationship with ethnolinguistic diversity or geographic location.

To further characterize the ancestry of Central and South American indigenous groups, we replicated a series of tests performed with qpWave by Skoglund et al. (2) to investigate the minimum number of ancestry streams necessary for the formation of these populations. Essentially, we selected four populations from each of the six global regions (sub-Saharan Africa, western Europe, East Asia, South Asia, Siberia/central Asia, and Oceania) as outgroups, and 14 indigenous groups with more than three unadmixed and unrelated individuals as test groups (*SI Appendix*, *Extended Methods*). These groups were tested in a few combinations, and the results are summarized in Dataset S5 (qpWave weights for the full dataset in Dataset S5*B*). These results reproduce the estimates obtained by Skoglund et al. (2) also indicating that at least two streams of migration are necessary to explain the present-day genetic diversity of Central and South American populations.

As the Chotuna group in the Pacific coast also exhibited excess allele sharing (Fig. 1 and Dataset S3) with the Australasians as estimated by D statistic (Mbuti; Australasians: Y, Z), we created admixture graph models based on the scaffold of Skoglund et al. (2) (Fig. 3*A*) with the addition of the Pacific coastal groups Sechura, Chotuna, and Narihuala. The best-fitted model showed that the Pacific coast is a mixed group of South American ancestry and a small non-American contribution associated with a sister branch of Onge (Fig. 3*C*), as also observed for Karitiana and Suruí. When the Xavánte were included in the analysis, the best-fitted model showed a direct contribution of the Australasian component in the Pacific coast, followed by a strong drift of

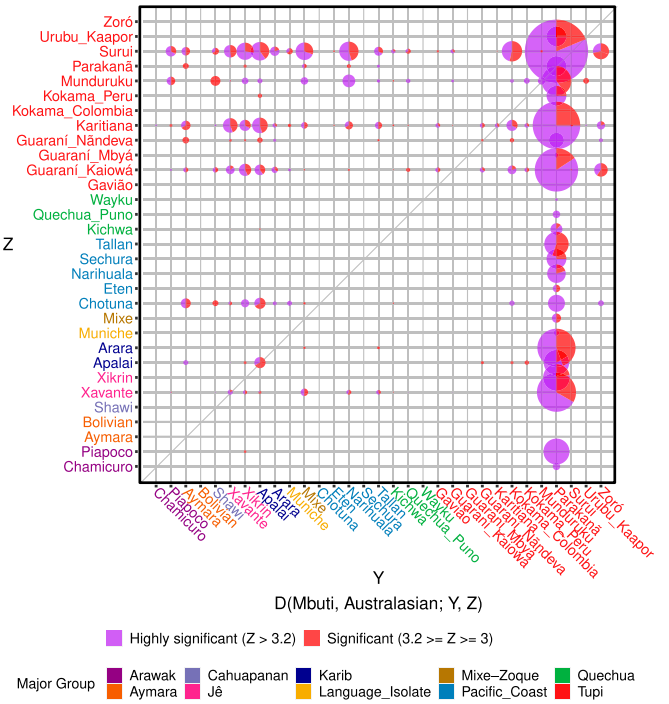


Fig. 2. Excess affinity of Native Americans with Australasians. The *y* axis indicates the group affiliation of the individual used at the Z position of the statistic (excess in allele sharing). The *x* axis represents the group affiliation of the individual at the Y position of the statistic (deficit in allele sharing). Estimates were clustered by groups, and the number of significant tests was weighted by the number of individuals in the comparison.

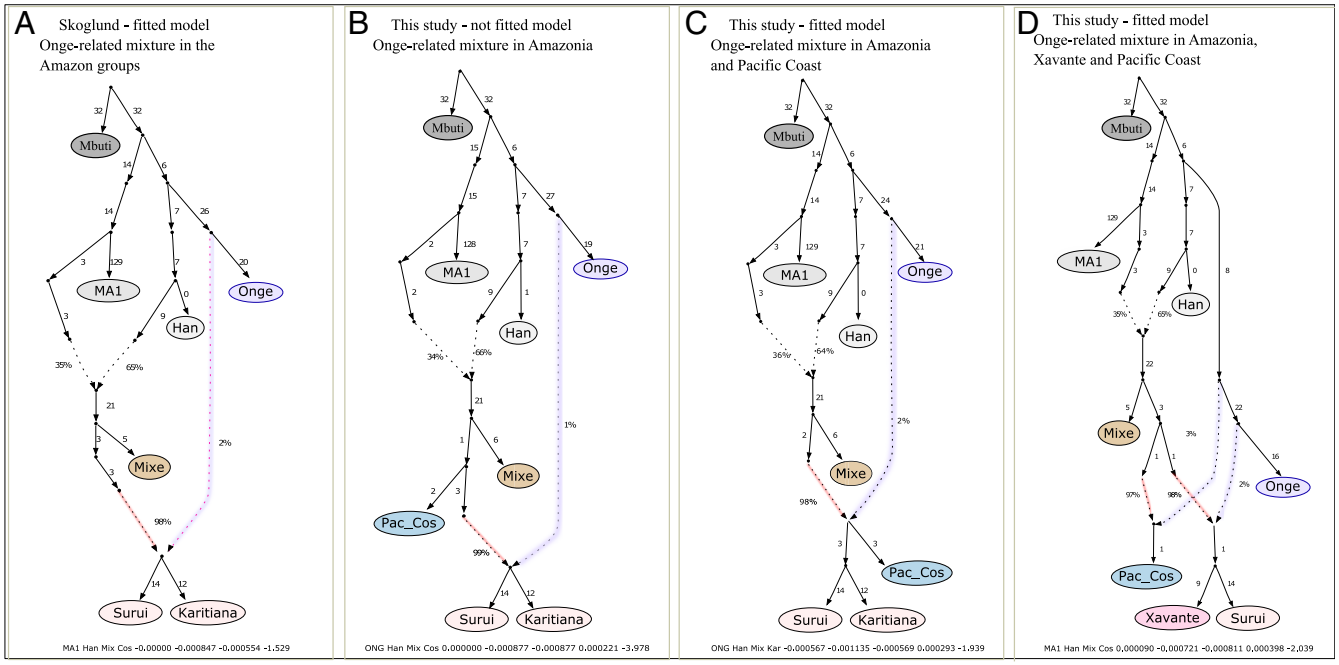


Fig. 3. Admixture graph modeling of the Y-population contribution to Amazonia and Pacific coast. (*A*) Previously published model proposed by Skoglund et al. (2). To investigate these genetic affinities, we first create (*B* and *C*) admixture graphs adding the Pacific coastal groups Sechura, Chotuna, and Narihuala (Pac_Cos) to the previous models, and (*D*) add Xavánte and the Pacific coast, followed by Suruí and Karitiana.

this signal, giving rise to Amazonian groups (Fig. 3*D*). Although Fig. 3*D* could indicate two independent events, the small genetic distances between the nodes in this model reinforced the single admixture event evidence. The Treemix (10) analysis also showed a pattern of diversification in which Pacific coastal and Andean groups diverged first (Dataset S6), followed by the eastern Andean slopes populations and then, finally, the Amazonians and other eastern South Americans. These findings suggest that the Y-population contribution was introduced before the formation of the Amazonian branch, likely in the ancestors of Pacific coastal/Amazonian populations.

Different migration routes to the South American region have been previously proposed and evidenced. Archeological and genetic data demonstrated that both routes, Pacific coastal and inland, were likely used by the first migrants (11). Our models point to an ancient genetic affinity between the Pacific coast and Amazonian populations that could be explained by the presence of Y ancestry in both geographic regions. In addition, this shared ancestry seems to precede the separation of the Pacific and Amazon branches, showing an entry through the west coast, followed by successive events of genetic drift in the Brazilian populations. This genetic evidence for the presence of Y ancestry on the South American Pacific coast indicates that this ancestry likely reached this region through the Pacific coastal route, and therefore could explain absence of this genetic component in the populations of North and Central America studied so far.

Data Availability. The newly genotyped datasets reported in this paper have been deposited in the European Genome-phenome Archive and are available for download under accession no. EGAS00001005022.

1. M. Raghavan *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
2. P. Skoglund *et al.*, Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
3. J. V. Moreno-Mayar *et al.*, Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
4. C. Posth *et al.*, Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197.e22 (2018).
5. C. Ning, D. Fernandes, P. Changmai, O. Flegontova, The genomic formation of First American ancestors in East and Northeast Asia. *bioRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.10.12.336628. Accessed 15 October 2020.
6. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
7. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
8. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
9. M. Mondal *et al.*, Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.* **48**, 1066–1070 (2016).
10. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
11. B. A. Potter *et al.*, Current evidence allows multiple models for the peopling of the Americas. *Sci. Adv.* **4**, eaat5473 (2018).

CAPÍTULO 2

Histórias populacionais e diversidade genômica dos nativos sul-americanos

Manuscrito: CASTRO E SILVA, M. A. et al. Population histories and genomic diversity of South American natives. **Manuscrito submetido ao periódico Molecular Biology and Evolution (MBE) (em revisão)**.

Autores: Marcos Araújo Castro e Silva[a], Tiago Ferraz[a], Cainã M. Couto-Silva, Renan B. Lemes[a], Kelly Nunes[a], Maíra R. Rodrigues[a], David Comas[b], and Tábita Hünemeier[a,2]

[a]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, 05508-090 São Paulo, SP, Brazil; [b]Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain

Resumo:

A América do Sul é o lar das mais diversas populações nativas contemporâneas e tem sido o foco de numerosas pesquisas multidisciplinares por mais de um século. Entretanto, o padrão de dispersão, a sub-estruturação genética e a complexidade demográfica dentro da América do Sul ainda é pouco compreendida. Aqui nós reportamos o mais abrangente estudo de grupos indígenas sul-americanos sob a luz de dados genômicos e linguísticos. Estes dados revelam um claro padrão de estrutura genética entre sul-americanos de terras baixas, apresentando pelo menos quatro agrupamentos genéticos primários nas regiões amazônica e de savana (Cerrado). Nossos resultados não suportam a existência de uma barreira genética abrupta entre os Andes e a Amazônia, indicando a existência de uma clina de isolamento por distância, a qual se alinha a um eixo de oeste para leste. Nós encontramos evidências de trocas culturais no oeste amazônico que muito provavelmente levaram a um processo de substituição linguística no período pré-contato. Adicionalmente, nós mostramos que a subestruturação genética dos sul-americanos contemporâneos recapitula ancestralidades locais antigas. Nossas inferências demográficas apontam para uma maior resistência dos sul-americanos do oeste aos colapsos populacionais causados pela invasão europeia e indicam a existência de expansões dêmicas no período pré-contato tanto na América do Sul quanto na Mesoamérica.

Palavras-chave: Povoamento da América do Sul | Divisão Andes-Amazônia | Genética | Nativos Americanos

Abstract:

South America is home to the most diverse present-day native populations and has been the focus of numerous multidisciplinary studies for more than a century. However, the dispersion pattern, genetic substructure, and demographic complexity within South America are still poorly understood. Here, we report the most comprehensive study of South American indigenous groups in light of the genomic and linguistic data. These data reveal a clear pattern of genetic structure among the South American lowlanders, presenting at least four primary genetic clusters in the Amazonian and savanna regions. Our results do not support a hard genetic division between the Andes and the Amazonia, indicating the existence of an isolation-by-distance cline, which aligns with a west to east axis. We found genetic evidence of cultural exchanges in the western Amazonia that most likely led to language replacement in the pre-contact times. In addition, we showed that present-day South American substructures recapitulated ancient local ancestries. Our demographic inferences point to a higher resilience of western South Americans to the population collapses caused by the European invasion and point out the existence of pre-contact demic expansions in both South America and Mesoamerica.

Keywords: Settlement of South America | Andes-Amazonia divide | Genetics | Native Americans

Population histories and genomic diversity of South American natives

Marcos Araújo Castro e Silva[a], Tiago Ferraz[a], Cainã M. Couto-Silva, Renan B. Lemes[a], Kelly Nunes[a], David Comas[b] and Tábita Hünemeier[a,2]

[a]*Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil;* [c]*Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain*

[2]Corresponding author: E-mail: hunemeier@usp.br

Introduction

The earliest hunter-gatherers entered the Amazon region *circa* 13,000 years before the present (BP), and the first signs of villages emerged at 8,000 BP. Archaeogenetic studies showed that the settlement of all contemporaneous native Amazonians was the result of a second migration wave from Beringia that arrived in South America at approximately 9,000 BP, replacing the first ancient wave that initially settled in this area (Posth et al. 2018). The genetic diversity of this region was most likely shaped by a continuous and dynamic interaction between the autochthonous population and the heterogeneous ecological features of this environment mediated by both cultural and biological evolution.

Amazon harbors the largest rainforest (also known as Amazonia) and the most voluminous river in the world. However, only approximately 2% of the entire ecosystem is suitable for agricultural practice (fertile floodplains), which may have contributed to the agglomeration and movement of people along the rivers and nearby banks, as well as to the emergence of trading networks along them. Although it has been suggested in the past that Amazonia was mainly populated by small, dispersed and disconnected communities (Tarazona-Santos et al. 2001), certain historical records indicate that these floodplains were densely populated and exploited when the first Europeans arrived (Carvajal 2011). Consequently, complex indigenous societies settled in Amazonia, creating a rich network of exchanges between different cultures and populations and housing one of the most diverse linguistic landscapes in the world.

Studies on human dispersals in this region are limited. The origin of the linguistic and genetic diversity of the people and the roles of the environmental and cultural practices in maintaining this diversity are still unknown. Here, we dissected the population and demographic history of the South American natives by analyzing genomic data from populations belonging to 50 different groups from eastern and western Amazon, tropical savanna (central Brazilian plateau), the Andes, the Pacific Coast, and Mexico.

Results and Discussion

Patterns of genetic ancestry in South America

First, using ADMIXTURE we estimated the proportion of African and European components present in the admixed individuals (we use a threshold of more than 1% inferred non-Native American ancestry to identify those with post-contact admixture) and the proportions per individual vary from 0 to 22% (mean 1%) and 0 to 44% (mean 6%) for the African and European ancestry components (Figure S1), respectively. Additionally we demonstrated that

these component proportions are not significantly correlated with the geographic location of these indigenous groups, represented by latitude and longitude, as inferred with a linear regression model (Figure S2). Then, we began to unravel the relationship among the current unadmixed (i.e. non-admixed) indigenous groups (Figure S3 and Figure S4) and examined the influence of geographic and cultural factors on the distribution and structure of genetic diversity. We identified nine major clusters with moderate to high similarity of genetic profiles: (i) northern Mexico, (ii) southern Mexico, (iii) northern Andes, (iv) southern Andes, (v) Pacific Coast and central Andes, (vi) western Amazonia, including the eastern Andean slopes, (vii) Karib and Tupí speakers from central and eastern Amazonia and the Atlantic Coast, (viii) Jê speakers from eastern Amazonia and the central Brazilian plateau, and (ix) Guaraní from central-west Brazil (also Tupí speakers) (Figure 1, Figure S5, and Figure S6).

The division between western and eastern South America is not patterned as an abrupt genetic division between the Andes and the Amazonia. As a result, different ancestry levels are shared between the Andean and Pacific coastal populations and the western Amazonian populations, especially those located in the eastern Andean slopes, corroborating previous genetic studies (Harris et al. 2018; Barbieri et al. 2019; Gnecchi-Ruscone et al. 2019). The western Amazonian groups situated more closely to the Andes exhibit a mixed profile that includes both eastern and western South American components (Figure 1, Figure S5, and Figure S6). The distribution of the putative genetic ancestry components suggests that the Andean lineages were introduced in the western Amazonia via gene flow. However, the opposite does not seem to have happened, as our finding showed that Andeans would have had much greater population sizes in comparison with Amazonians (see Population dynamics and demography section). This scenario could have caused the asymmetrical contribution of ancestry from Andeans to western Amazonians. These broad ancestry patterns also indicate, along with evidence from the principal component analysis (PCA) (Figure 2C, D and Figure S7), the influence of geographic distribution, and environmental and cultural diversity on the genetic similarity of South American groups. Notably, within the current Brazilian territory, we identified four primary genomic divisions in the Amazonian and savanna regions. When only the unadmixed and unrelated individuals were analyzed, the most prominent genetic structure separated Jê-speakers from the other Native Americans (Figure S8 and Figure S9). The second group to be differentiated is the southern branch of Tupí speakers, Guaraní. The remaining non-Jê/non-Guaraní component is then subdivided into a predominant component in western South America and Mesoamerica and another predominant component in eastern South America.

The historical relationships among populations inferred with Treemix (Pickrell and Pritchard 2012) broadly recovers the same clusters observed in the previous analyses, namely the Pacific

Coast, southern Andes, and the Kichwa group, remains isolated as the sole representative of a northern Andean cluster in this analysis, another division between western and eastern Amazonia, and the remaining Brazilian samples clustered with the latter (Figure 2). The eastern South American/Brazilian cluster is subdivided into two main regions: north (with Apalai and Wajãpi) and south of the Amazon River, which is further subdivided into central Amazonia (i.e., Munduruku) and Madeira-Guaraporé region (Gavião, Karitiana, Suruí, and Zoró), and Xavante and Urubu-Kaapor with no further clustering.

By analyzing the population substructure in the Amazon region, we found evidence that rivers might have acted as physical barriers or at least deterrents to gene flow. As some of these genetic divisions are drawn along the main river courses, we found a cluster delimited by the Xingu River in the west and by the Araguaia and Tocantins rivers in the east (Asurini, Parakanã, and Xikrin), including Arara, located in the Xingu River near the west margin (Figure 2).

Additionally, the cluster of Guaraní groups is differentiated from the other groups (Figure 2) located in central-west Brazil. This result also suggests that geographic distance and environmental diversity play important roles in shaping genetic similarities among groups. Genetic clusters are circumscribed to specific ecogeographical regions and are not necessarily related to cultural diversity. However, the effects of geographic distribution, environmental and cultural diversities are difficult to disentangle, as groups speaking the same language and sharing a common culture tend to be in proximity and settle in the same kinds of environments.

This relationship between genetic variation and geography is also evidenced by the PCA, where at least one of the principal components is correlated at some degree with longitude (Figure 2C: PC4 ~ Longitude: R^2 = 0.3962 and p-value = 2.2e–16; and Figure 2D: PC1 ~ Longitude: R^2 = 0.5905 and p-value = 2.2e–16), although some groups are clearly outliers (Xavante, Xikrin, Karitiana and Suruí). In addition, we also computed matrices of genetic (estimated as 1 – Outgroup F_3) and geographic distances (great circle distances) between all pairs of individuals and applied a linear regression model in the form 'lm(genetic_distance ~ geographic_distance)' to the subset of unadmixed and unrelated indigenous individuals, resulting in an estimated R^2 of 0.515 (p = < 2.2e–16).

Finally, we also inferred putative gene flow events using Treemix by fitting an increasing number of gene flow events, until a model likelihood plateau was reached with six events (color coded arrows in Figure 2 and Figure S10). It is noteworthy that most inferred gene flow events occur between eastern South Americans, especially among Tupi-speakers: (i) Wajãpi to Asurini; (ii) Guaraní Kaiowá to Guaraní Mbyá; (iii) the inferred most recent common ancestor (MRCA) of

Tupí-Mondé (Surui, Gavião, and Zoró) to Guaraní Mbyá; (iv) Zoró to the inferred MRCA of northeastern Amazonians located in the southern margin of the Amazon River (Arara, Parakanã, Xikrin, and Asurini); and (v) Xavante to Xikrin (both Jê-Speakers). We also detected one gene flow event from Quechua of southern Peruvian Andes to a group of Bolivians (with Aymara ancestry).

Pre-Columbian genetics and linguistic interchange

The Tupí linguistic group currently has the largest number of speakers among South American lowlanders. In addition to this large proportion of speakers, the group's wide geographical distribution draws attention, with populations ranging from the Atlantic coast (Tupiniquim) to the northwest boundary between the Amazon forest and the Andes (Kokama), totaling more than 5,000 km in distance between these two far ends.

Interestingly, the Kokama groups (from Colombia and Peru) presented genetic ancestry profiles much more similar to populations from western Amazonia, especially those from the Loreto region located close to the eastern Andean slopes (i.e., Chamicuro, Muniche, Shawi, and Wayku), rather than to other Tupí-speaking groups, as shown by ADMIXTURE, Treemix, and PCA (Figure 1 and Figure 2). Additionally, admixture graph models showed a good fit for the 1-way model or single origin model for the Peruvian Kokama, placing them as a sister branch of the Chamicuro, an Arawak-speaking group (Figure 3). Furthermore, the Colombian Kokama can be modeled as a basal group to the Peruvian Kokama and Chamicuro (Figure 3A and Figure S11A). Other evidence comes from the multi-dimensional scaling (MDS) of pairwise genetic distances (1 – Outgroup F_3), in which both Kokama groups are clustered with western Amazonians (Figure 4A), rather than with the other Tupí, and also from F_4-statistics, we detected an excess affinity among both Kokama groups (Figure 4B and Figure S12), as well as between them and the Arawak speakers, especially strong when comparing Peruvian Kokama and Chamicuro (median Z-value < −3 in Figure 4B, left top panel). This genetic similarity could theoretically be the consequence of the recent population influx from rural areas into large urban centers in the Andean region, which resulted in the homogenization of the local genetic diversity (Tarazona-Santos et al. 2001). While this could partially explain this pattern in the Peruvian Kokama, it cannot explain why the Colombian Kokama also presents genetic profiles very similar to those of the Arawak-speaking groups and the other western South American populations, even though they are located hundreds of kilometers away.

This striking result aligns with a previous hypothesis about the origins of present-day Kokama. According to this interpretation, the Kokama were non-Tupí-speaking people who went through

a process of language replacement due to contact with speakers of Tupí-Guaraní (Noelli 2008; Michael 2014), likely during the period of the Tupí Expansion event around 3,000–2,000 BP (Castro e Silva et al. 2020). In that regard, a combination of linguistic, historical, and ethnographic analysis (Michael 2014) pointed in favor of a pre-Columbian origin of the Kokama group. According to this time frame, this language replacement would not be an aftermath of the contact with Europeans and the consequent disruption of indigenous societies. Therefore, our results on the marked genetic similarity between Kokama and the Arawak speakers added to the latter's unique cultural characteristics (e.g., exogamy, high mobility, and expansiveness of the Arawakan identity; (Hornborg 2005)) supports the Kokama as former Arawak-speakers (or closely related groups) that adopted a Tupí-Guaraní language, likely due to the Tupí expansion.

At the other end of the Tupí expansion, on the border among Argentina, Brazil, and Paraguay, are the Guaraní peoples, the most numerous speakers of Tupí today, who are divided into three ethnic groups: Kaiowá, Mbyá and Ñandeva. Previous studies have shown that the pre-Columbian admixture with Gran Chaco, Mesoamerican, or other related sources probably contributed to the formation of this southern Tupí branch (Reich et al. 2012; Gnecchi-Ruscone et al. 2019; Castro e Silva et al. 2020).

Accordingly, we used Admixture Graphs (AG) modeling to infer the population history of the Guaraní group, and several models presented a good fit to the genetic data, considering all possible combinations of single or mixture models. On the one hand, the single origin of Guaraní Kaiowá (Figure 3B) as a sister group of Xavante (Jê-speakers) indicates a closer relationship between the ancestors of the two groups in comparison with the Amazonians, here represented by Suruí and Karitiana. On the other hand, the Guaraní Ñandeva presented several models with good fit (Figure 3C and Figure S11B–I). Among these possibilities, Ñandeva can be fitted as an admixed population resulting from a major Tupí-Guaraní ancestral component and a minor component related to early branches of South American groups (e.g., Mixe, Andes, Pacific Coast, eastern Andean slopes groups, or basal South Americans) (Figure 3C and Figure S11B–I). We were not able to model Guaraní Mbyá into a well-fitted model. The lowest maximum Z of the mixture model was 3.141 (Figure 3D) in a pattern very similar to the population history models inferred for Guaraní Ñandeva.

It is likely that we were unable to fit the Guaraní Mbyá into a model, as well as to distinguish a clear best fitted model for Guaraní Ñandeva, due to significant contributions from Gran Chaco lineages or other closely related groups not represented in our dataset. This admixed origin of Guaraní was first inferred by Reich et al. (Reich et al. 2012) as a mixture of ancestral lineages of sister branches of the present-day Wichi (Gran Chaco) and Suruí (Amazonia); later, this

inference of an admixture event was also reproduced by others (Gnecchi-Ruscone et al. 2019). Additionally, we have previously detected an ancestry contribution from a Mesoamerican source into the Guaraní branch (Castro e Silva et al. 2020), which likely represents the same event, as we also lacked samples from the Gran Chaco and other adjacent regions for that specific analysis.

Genetic affinities amongst present-day South American natives

We tested the existence of excess allele sharing among particular Native American groups and calculated the F_4(Mbuti, X, Y, Z) for every combination of X, Y, and Z Native American test groups (Figure 4B and Figure S12). Using this approach, we detected continental-wide excess affinity patterns of interest: (i) an excess affinity between Kokama and Arawak speakers (Figure 4B - left top panel); (ii) an ubiquitous high affinity between the groups from the Madeira-Guaporé region (Amazonia - Rondonia state) (Figure 4B - right top panel); (iii) a pattern of high affinity among the Guaraní groups (Kaiowá, Mbyá, and Ñandeva) (Figure 4B - left middle panel); (iv) Jê speakers present high affinities with one another (Figure 4B - right middle panel); (v) Finally, groups from the Pacific coast and the Andes regions present high and almost exclusive intra-regional affinity (Figure 4B - bottom panels).

The genetic affinities among the groups were also examined through the identification of genomic segments using identity by descent (IBD). By selecting only those segments present in the genomic regions of Native American local ancestry, we enabled the detection of shared ancestry and uncovered information about the time and intensity of these past events. Any haplotype in the genome undergoes an exponential reduction in length over time owing to the randomness of the recombination process. For this reason, shorter segments of IBD correspond to older events and are likely to be more widely and ubiquitously shared; in the case of present-day Native American populations, these would have been mostly formed during a series of population bottlenecks in the ancient past. Considering this, we adapted the approach used by Barbieri et al. (Barbieri et al. 2019), restricting this analysis to connections between groups that share a total average of at least 5 cM of IBD. As only IBD segments of Native American ancestry were analyzed, we focused exclusively on the population histories of the indigenous peoples of the Americas, thereby eliminating the confounding effect of the IBD introduced by admixture events with populations from Africa and Europe that took place in the post-Columbian period. We also removed related individuals to avoid biased estimates of IBD and the selected segments were grouped into three categories to focus on different time periods (Baharian et al. 2016; Harris et al. 2018): (i) Pre-Columbian period $\sim <$ 1500 CE (segments $\leq$ 9 cM) (Figure 5); (ii)

Colonial period ~ 1500–1850 CE (9 cM < segments ≤ 22 cM) (Figure S13A), and (iii) the recent period ~ 1850–Present CE (22 cM > segments) (Figure S13B).

The results showed a widespread network of IBD sharing formed in the Pre-Columbian period, with a much higher number and intensity of connections among groups within a few main areas: the Pacific coast; the southern, central, and northern Andes; northern and southern Mexico; and especially central and eastern Amazonia, along with other Brazilian regions with less intense connections among these main areas (Figure 5). This observation also goes against a hard Andes-Amazonia divide model (Pearce et al. 2020), as western Amazonians show a similar connectivity with both the Andeans and the central and eastern Amazonians, presenting themselves as intermediaries or transitional groups. In addition, for IBD originating in the colonial period (Figure S13A) as well as in the more recent period (Figure S13B), IBD sharing is even more restricted to these main areas with very few inferred long-distance connections. These observations also strengthen the evidence on how the genetic diversity of the groups from these areas was distinctively shaped, as they faced different geographic, climatic, ecological, and cultural forces in their recent and ancient histories.

Present-day South American substructure recapitulates deep local ancestries

To assess which ancient individuals from these different regions contributed to the genetic formation of specific present-day groups, our approach aimed at detecting the existence of excess allele sharing between present-day Native American groups and ancient individuals, in relation to other present-day indigenous groups in the Americas. To accomplish this, we computed F_4(Mbuti, X; Y, Z), where X is any ancient sample, and Y and Z are any present-day Native American group. The complete set of statistics was then summarized as the number of highly significant tests (Z > 4) grouped by pairs of X ancient individuals and Z present-day groups summed over all the Y present-day groups (Figure S14). While the specific pattern of affinity changes between different sample ages and locations, our findings demonstrated a general excess affinity between ancient and present-day groups from the same or adjacent regions. This finding suggests at least some level of genetic continuity within large continental areas. The broad patterns indicate an excess affinity between ancient samples from Bolivia, Chile, and Peru with present-day southern Andean populations (Aymara and Quechua), and in the case of some Argentinian and Peruvian ancient samples to present-day Pacific coastal groups (Figure S14). On the other hand, ancient samples from eastern South America (Brazil), Central America (Belize), and the Caribbean (Bahamas) tend to exhibit higher affinity to present-day groups from Brazil, more specifically to speakers of Tupí, Jê, and Karib languages (Figure S14).

We also analysed the global pattern of genetic affinities among present-day Native Americans and ancient individuals from across the whole American continent and Siberia. For this we estimated the genetic distances (1 – Outgroup F_3) between all pairs of individuals and applied an MDS to the matrix of pairwise genetic distances. The analysis identifies the most distinctive groups (Figure 6A; Dimension 1), with ancient Siberians and Ancient Beringians (AB) separated from the Northern Native Americans (NNA) and Southern Native Americans (SNA) (Reich et al. 2012; Rasmussen et al. 2014; Raghavan et al. 2015; Moreno-Mayar et al. 2018; Posth et al. 2018), with all present-day Native Americans clustered with the latter. Ancient individuals from the San Nicolas Island are outliers in the MDS (Figure 6A; Dimension 2). With a closer look at the SNA, the same cline of genetic differentiation detected in previous analyses is observable. However this time the gradient broadly goes from (Figure 6B; top right corner) ancient samples from North, and then Central and South America, followed by present-day groups from Mesoamerica (Mixe), Pacific Coast, Andes, Amazonia and ending with Jê-speakers from the central Brazilian plateau and Guaraní groups from central-west Brazil (Figure 6B; bottom left corner). Accordingly, if only the values of present-day individuals obtained in this MDS are considered and linear regression models are applied between the first and second dimensions of the MDS with the Longitude or Latitude of each individual, the models 'lm(Dimension 1 ~ Longitude)' and 'lm(Dimension 2 ~ Longitude)' present estimated R^2 of 0.4882 ($p = 3.195e^{-14}$) and of 0.5274 ($p = 1.041e^{-15}$), while the models 'lm(Dimension 1 ~ Latitude)' and 'lm(Dimension 2 ~ Latitude)' exhibit estimated R^2 of 0.2830 ($p = 6.855e^{-08}$) and of 0.2904 ($p = 4.348e^{-08}$), respectively. Once again evidencing the relationship between genetic variation and geography, especially with longitudinal distribution.

Population dynamics and demography

Contact with the Europeans and the colonization of the American continent led to massive depopulation of the indigenous peoples caused by the introduction of new diseases (e.g., smallpox), enslavement, warfare, disruption of subsistence strategies, and forced displacement from territories, among other processes (Montenegro and Stephens 2006; Adhikari et al. 2017). To evaluate the impact of these population bottlenecks in the genetic diversity of indigenous populations of the Americas, inferring when they occurred and how strong they were, we applied ASCEND (Tournebize et al. 2020) to all populations with more than five unrelated individuals (Figure 7B). Some populations were also analyzed in clusters of speakers of languages from the same linguistic families (to reach the minimum sample size required for the analysis). For each population and cluster, the Founder Intensity (FI) (an estimate of the genetic drift strengh caused by a bottleneck) and Founder Age (FA) (inferred date of the bottleneck)

were estimated along with the associated 95% confidence intervals (CIs) (Tournebize et al. 2020).

Most groups produced FI estimates concentrated around 5–10% and FA estimates below 500 BP or approximately 18 generations before present (gBP; considering 28 years per generation) (Figure 7B and Figure S15). The median FA and median FI for the groups were 8 gBP (or ~224 BP) and 5.3%, respectively. The highest inferred FI estimates for groups were obtained for a few Pacific coastal groups (Narihuala: 15.4%; Tallan: 12.5%) and a few groups speaking Jê (Xikrin: 15.3%; Xavante: 11.2%), Tupí-Guaraní (Guaraní_Kaiowá: 10.7%), Uto-Aztecan (Pima: 8.8%), Quechua (Huancas: 8.2%), and Mixe-Zoque (Mixe: 7.3%) languages. The lowest FI estimates were obtained for some Mexican (Zapotec, 4.1%; Yaquis, 4.2%; Maya, 4.4%) and Southern Andean (Bolivians: 3.2% and Quechua_Cusco2: 3.7%) groups. Most of the inferred population bottlenecks occurred in post-contact times, which aligns with that found in historical records.

In addition, the eastern South Americans present significantly higher levels of population inbreeding coefficient F_{ROH} (Figure 8 and Figure S16), obtained from the runs of homozygosity (ROH), in comparison to that observed for the western South Americans, which means that the genetic diversity estimated for lowlander populations is significantly lower (p-value = 2×10^{-12}). This suggests that populations close to the Amazonia were subjected, on average, to extreme genetic drift processes, such as serial bottlenecks followed by subsequent founder events, which would have considerably reduced their genetic variability. When the longitude coordinates are considered, we observe that the farther east the population is located, the greater its F_{ROH} value is (Spearman correlation coefficient $r = -0.58$; p-value < 0.001) (Figure 8), showing that the genetic variability increases according to the proximity to the Andes region, which is consistent with an isolation by distance model. Likewise, it indicates that Amazonia might be the region with the lowest genetic variability in the Americas, and thus, probably the lowest in the world. We also analyzed the distribution of IBD within populations and Homozygosity-by-Descent (HBD), which demonstrated a distinctive pattern between these regions, with ubiquitously lower IBD and HBD levels in the highlands, Pacific Coast and Mesoamerica, and the opposite pattern on the eastern South American lowlands (Figure S17), as expected.

Finally, to estimate the population sizes and evaluate how they have changed over time in the ancestors of the present-day groups from these diverse regions, we also leveraged the IBD segments by applying IBDNe (S.R. Browning et al. 2018) to infer the effective population size histories. This analysis was conducted by sub-setting the IBD segments into the major ethnolinguistic/geographic regions and selecting only those with at least 20 samples (to minimize the effects of low sample sizes in the analysis), therefore keeping on the dataset the

following clusters: Pacific Coast, Quechua, Tupí, Jê, Maya, Oto-Manguean, and Uto-Aztecan. These subsets were then used to infer the Native American ancestry-specific N_e history of these groups by including only the IBD segments identified in the genomic regions of Native American local ancestry as inferred by RFMix (Maples et al. 2013).

As expected, the historical N_e of eastern South Americans is smaller (median, minimum and maximum of 9,280, 867, and 90,200, respectively) than that of western South Americans, which shows the highest N_e (median, minimum and maximum of 28,000, 7,380, and 143,000, respectively) in the last 100 gBP, while Mesoamerica and northern Mexico have similar values to western South America (median, minimum and maximum of 26,500, 805, and 32,700, respectively) (Figure 7A and Figure S18). The effective population size estimates also evidence the effect of contact with Europeans and colonization on the Native American peoples, particularly on the already lower genetic diversity of eastern South Americans. Conversely, the western South American genetic diversity seems to have been more resilient to this process, although it also faces an N_e reduction of an order of magnitude, with a slight recent recovery.

Looking at the inferred effective population size histories of the major ethnolinguistic and geographic groups represented here (those with more than 20 individuals), it is noticeable that most groups still preserve an N_e of at least $\sim 10^3$ (Mayan, Oto-Manguean, Andes, Pacific Coast, Quechua); however, some of them present a steep decline in N_e going even below $\sim 10^2$ (Uto-Aztecan, Jê and Tupí) (Figure 7A and Figure S18). Notably, an additional bottleneck is inferred for the Tupí, starting at ~37 gBP (~1036 BP) before the contact with Europeans and going up until ~15 gBP (~420 BP), though these inferences should be taken cautiously, this population decline goes in line with a recent report based on Amazonian fossil pollen analysis (Bush et al. 2021), which shows that landscapes started to be abandoned and forests to regrow around 300 to 600 years prior to the European arrival, which is hypothesize to might have been caused by pre-European diseases, societal and/or climate change, and most surprisingly it was also inferred almost no substantial forest increase during the post-contact population collapse period. These results also show substantial population expansions of some of these groups in the pre-contact period, especially between ~75 and ~20 gBP. In Mesoamerica, the speakers of Oto-Manguean languages expanded from an N_e of 10^4 to almost 10^6. In South America, especially Tupí-speakers and the Andean populations presented an increase of at least two orders of magnitude (Figure 7A and Figure S18). The Tupí-speakers exhibit an N_e increase that goes from ~75 to ~37 gBP (approximately between 2100 and 1000 BP), therefore partially overlapping with the hypothesized period of the Tupí Expansion which would have started between 3000 and 2000 BP (Noelli 2008). This result suggests that Tupí-speaking groups would have gone

through population growth during this time, which is one of the proposed drivers of their territorial expansion.

Conclusions

South American genetic variation is related to linguistic and environmental diversity, which is more pronounced in local contexts and within the same ethnolinguistic groups. We found at least four primary clusters of genetic similarity in the Amazonian and savanna regions, partially mirroring the main linguistic diversity. Also, no hard genetic division between the Andes and the Amazonia was noticed. Furthermore, genetic variation and homozygosity level are correlated with longitude, supporting isolation by distance model, possibly tracing back to an initial settlement from the Pacific coast. We also described an extensive ancient genetic interchange among the eastern lowland populations with reduced allele sharing between the eastern and western Amazonian lowlanders, and genetic evidence of cultural exchanges in the pre-contact times, leading to language replacement. In addition, the present-day Native American diversity recapitulates ancient local ancestries. Finally, demographic analyses indicate that Western South Americans were less affected by the process of European colonization and show that the population size of some South American and Mesoamerican groups varied greatly over the past 2-3 thousand years.

Materials and Methods

Dataset assembly

Here we used the dataset from Castro e Silva et al. (Castro e Silva et al. 2021), which combined new and publicly available datasets, genotyped with the Axiom Human Origins array - Affymetrix/Thermo Fisher (Patterson et al. 2012) or whole-genome sequenced, as described below. The dataset includes Brazilian populations from the Amazonian Rainforest, from Southwestern Brazil (near the Paraguayan border), the central Brazilian plateau, and also from the Brazilian Atlantic Coast (Patterson et al. 2012; Skoglund et al. 2015; Castro e Silva et al. 2020). To increase the scope of the study and to enable the examination of how the human genetic diversity is patterned across the Amazon-Andes divide and also more broadly on the American continent, populations from Mexico, Colombia, Ecuador and Peru were also included (Lazaridis et al. 2014; Mallick et al. 2016; Barbieri et al. 2019). The dataset was also merged with the 1240K_HO dataset assembly (v42.4), to include publicly available data for other Native American groups and potential unadmixed individuals (with no or negligible signal of contributions from Non-Native American sources) from other present-day populations from the American continent, as well as the publicly available ancient samples from the Americas. The complete dataset contains 383 individuals from 50 different present-day indigenous groups (Dataset S1), although it contains more than one set of samples for some of these ethnolinguistic groups (e.g. Quechua_Cusco, Quechua_Cusco2, Quechua_Cusco3, Quechua.DG, and Quechua_Puno). Please refer to Dataset S1, Dataset S2 and Dataset S3 for more information on the test samples (i.e. Native Americans), on the reference panel of samples and on the ancient samples (i.e. aDNA samples from ancient Native Americans) used in this study, respectively.

Data curation

Before any analysis was performed, some genome-wide summary statistics were estimated for the complete dataset. Considering a threshold of missingness rate per individual of 10% no sample was removed from the dataset. While an insignificant number of SNPs (74 SNPs) present a missingness rate above 5%, though not exceeding 7%, which were removed. The dataset contains a proportion of 10.96% of rare alleles (SNPs with a minor allele frequency lower than 1%). Following this initial evaluation, the complete dataset was pruned to remove markers with a pairwise correlation above 20% ($r^2 > 0.2$ inside a sliding window of 50 Kb size and step size of 10 Kb) thus producing an LD-pruned dataset. Next, a supervised clustering analysis was performed with ADMIXTURE (Alexander et al. 2009) on a subset of the LD-pruned dataset (keeping all of the samples from the American continent, Sub-Saharan Africa, and western

Europe), in order to assess whether the samples from Native American communities and other populations from the American continent have genetic contributions from non-Native American sources. Finally, the LD-pruned dataset was also assessed for the presence of related individuals, for this we used PLINK 1.9 (Chang et al. 2015) pairwise IBD estimation method to obtain the proportion of shared IBD between all pairs of individuals (i.e. PI_HAT = P(IBD=2) + 0.5*P(IBD=1)), these estimates were used as input to the PRIMUS (Staples et al. 2013) in order to identify the maximum set of unrelated individuals, considering the 1st degree of relatedness (PI-HAT > 0.375) as the threshold. The complete dataset could then be filtered to remove admixed (< 99% inferred non-Native American ancestry) and related individuals (PI-HAT > 0.375). Thus allowing the selection of the subset of unadmixed individuals (with 150 individuals from 33 indigenous groups), the subset of unrelated individuals (with 312 individuals from 58 indigenous groups), and the subset unadmixed and unrelated individuals (with 87 individuals also distributed in 33 groups) (Dataset S1).

Global ancestry and population structure

Initially, Principal Component Analysis (PCA) was applied with SNPRelate R/Bioconductor (Zheng et al. 2012) to the LD-pruned subset of unadmixed and unrelated Native Americans and one additional PCA excluding outliers (Xavante, Xikrin, Karitiana, and Suruí), to examine the broad patterns of ancestry and genetic differentiation. Next, we applied ADMIXTURE (Alexander et al. 2009) to the unrelated and to the unadmixed-unrelated LD-pruned datasets in order to estimate the individual global ancestry components and to investigate the population structure and the patterns of shared ancestry of the South American indigenous groups. In order to do this, we executed one unsupervised ADMIXTURE analysis for each dataset, with the number of putative genetic components K ranging from 2 to 10. To evaluate the estimated ancestry components we used the distribution of cross-validation errors and likelihoods (provided by ADMIXTURE) for each value of K, and we used PONG (Behr et al. 2016) to evaluate multimodality and plot the estimates. The proportions of individual ancestry components were also plotted on maps using the mean group values of these components and the geographical coordinates for each group, highlighting their linguistic affiliations. To do this some R packages were used as described below, “ggmap” to obtain the map, data for the main rivers of the American continent was provide by “mapdata” (wider blue lines), and data for the main rivers of Brazil (thin blue lines; which were not added to all map figures to avoid overplotting) was obtained from the website of the Laboratório de Pesquisas em Geografia Física (LAPEGE; http://www.uel.br/laboratorios/lapege/pages/base-de-dados-br.php), “scatterpie” to plot the pie charts on the map, and “ggrepel” to create labels for each group,

Patterns of shared ancestry

As a first step we analyzed how Identity by descent (IBD) blocks are shared between the entire set of pairs of Native American groups and how these connections were distributed along the geographical space, following the same approach used by Barbieri and collaborators (Barbieri et al. 2019). In order to do so the subset of Native American groups data was filtered to remove markers and samples with more than 5% of missingness, monomorphic SNPs were also pruned. Then, this dataset was phased with BEAGLE v.5.1 (B.L. Browning et al. 2018), the IBD segments were identified with Refined IBD (Browning and Browning 2013) and a Local Ancestry Inference was performed with RFMix (Maples et al. 2013), with a window size of 0.2 cM, with 5 as the minimum number of reference haplotypes per tree node, and the unadmixed Native Americans, Sub-Saharan Africans, and western Europeans, as the reference panel of populations. Next, the short gaps in the IBD segments were removed with the program merge-ibd-segments.17Jan20.102.jar, using default parameters (i.e. 0.6 as maximum gap length and 1 as maximum number of discordant homozygotes), then the IBD segments were classified and separated in different subsets according to their local ancestry.

We only analysed IBD segments of Native American ancestry, identified in the maximum unrelated subset of individuals and to secure reliable results we selected blocks with more than 2 cM and with an estimated LOD score above 3 (Browning and Browning 2013), furthermore only pairs of populations with more than 5 cM of IBD on average were considered as population pairwise sharing. The IBD blocks were then classified into 3 length categories essentially corresponding to 3 distinct time periods (Baharian et al. 2016; Harris et al. 2018): (i) Pre-Columbian period ~ < 1500 CE (segments <= 9 cM) ; (ii) colonial period ~ 1500–1850 CE (9 cM < segments <= 22 cM); and recent period ~ 1850–Present CE (22 cM > segments). The average number of IBD segments and the average length of shared IBD were then calculated for each length category independently, with in-house R scripts, by dividing these averages, one at a time, by the product of the sample sizes of the 2 populations being compared. The map of pairwise connections was produced with the "ggmap" R package.

Next, to assess the patterns of allele sharing between indigenous groups and to test prior hypotheses regarding the formation of some specific groups, we first estimated the Outgroup F_3(Y, Z; Mbuti), as well as the formal test F_4(Mbuti, Test; Y, Z). Both statistics (F_3 and F_4) were calculated for every pair of Y and Z indigenous groups, and "Test" was iterated over every single indigenous group or individual in the F_4 statistics computation. Additionally, a matrix of Outgroup F_3(Y, Z; Mbuti) calculated for all Y and Z pairs of groups and pairs of individuals, was converted to genetic distances (Genetic distance = 1 – Outgroup F_3 estimate). A

multidimensional scaling analysis (MDS) was then applied to the matrix of pairwise genetic distances with the "stats" R package.

Patterns of genetic continuity

To shed light on the local patterns of genetic continuity in the South American continent, first we wanted to investigate the relative patterns of excess affinity between ancient individuals and each pair of present-day Native American groups, by computing F_4(Mbuti, X; Y, Z), where X is any ancient sample, and Y and Z is any pair of present-day Native American groups. First we filtered the highly significant tests (Z > 4) and then the selected set was summarized as the number of significant tests, grouping by pairs of X ancient individuals and Z present-day groups, and summing over all the Y present-day groups. We also estimated the pairwise genetic distance (1 – Outgroup F_3 estimate) matrix between all present-day (Native Americans) and ancient (from the American continent and Siberia) individuals and applied an MDS to this matrix using the "stats" R package.

Genetic diversity and geography

Furthermore, we wanted to investigate if and how much the genetic diversity of indigenous populations was influenced by their geographic distances. We used the 1 – Outgroup F_3 as the measure of genetic distance and the geographical distances were calculated based on the coordinates of every pair of groups as great circle distances with the R package "geosphere". We then applied a linear regression model (R "stats" package) to the matrices of genetic and geographical distance of all 87 indigenous individuals from the unadmixed-unrelated dataset, with and without comparisons of samples from the same place (i.e. geographical distance = 0, equivalent to removing intra-population comparisons). Finally, we looked at the relationship between continental ancestry components (i.e. Native American, African and European) and Latitude/Longitude, and also between the estimated Principal Components and the later, by fitting linear regression models.

Population histories and admixture events in present-day indigenous groups

First, we aimed to produce an outline of the population history of the Native American groups here represented, therefore providing subsidies, along with other lines of evidence from archeology and linguistics, for a framework to model how these groups relate to each other and for how their population history unfolded. This was initially explored with Treemix program (Pickrell and Pritchard 2012), which uses an unsupervised method for estimating the Maximum

Likelihood tree based on population pairwise allelic covariances and allows for putative gene flow events to be adjusted between branches of the tree.

Second, we used the qpGraph (v.6450) from Admixtools (Patterson et al. 2012) to model the population history of the present-day indigenous groups, by compiling several F statistics to infer the best fit between the genetic variation observed and the model. In order to do so we computed the models with the unrelated and unadmixed set of Native American individuals, using only groups/individuals that presented a coverage of at least 200k SNPs. The default settings for qpGraph were used, except for the parameters "outpop: NULL" and the "all SNPs: YES". The models were constructed using a scaffold tree composed by groups representing the major possible contributions for the tested groups (Mesoamericans, Andes, Amazonian - Tupí-speaking groups, Shawi - Muniche - Lowland Peruvians, and Jê-speaking groups) and we modelled the test groups in all possible placements along the scaffold branches.

Next, we tested both the one-way model assuming a single origin of the tested group and the two-way model assuming a mixed origin. We compared the maximum Z score of all computed admixture graphs, and selected the best fitted models (i.e. lowest maximum |Z|) as candidates to represent the population history. When both models (single and admixed origin) presented a good fit, we gave preference to the single-origin model (one-way), as it is the more parsimonious scenario. However, if among the multiple models there was no one-way with a good fit, we used the criteria of the lowest maximum Z score among all two-way models and the number of outliers as a proxy to decide which is the best candidate model to explain the genetic patterns observed. In search of the best fit, we tested one group at a time (Kokama_Peru, Kokama_Colombia, Kaiowá, Ñandeva and Mbyá), placing the test group in all terminal positions.

Demographic inferences

We then used the ASCEND software (Tournebize et al. 2020) to infer the age and intensity of the bottleneck events on the Native American groups. First we selected the Native Americans from the complete dataset, then we pruned markers and samples with more than 5% of missing data (no samples were removed due to this criterion). Next we selected groups with more than 5 unrelated samples (keeping samples with evidence of admixture with other continental ancestries), to ensure the minimum required sample size (N >= 5) and to avoid the confounding effect of consanguinity. Basically, ASCEND infers the age and intensity of founder events (or bottlenecks in this case) as parameters of a model based on the empirical curve of exponential decay of allele sharing correlation between pairs of individuals inside the same population as a function of the genetic distance. The cross-population correlation with an outgroup is subtracted from this intra-population correlation to exclude the effect of ancestral allele sharing.

A random sample of 15 individuals from the complete dataset was used as the outgroup population as in Tournebize et al. (Tournebize et al. 2020).

We also analyzed the IBD segments identified as previously described in the section "Patterns of shared ancestry", first by selecting segments shared between individuals from the same population (i.e. intra-population IBD) and present in genomic regions of Native American ancestry as inferred with RFMix (Maples et al. 2013), and second by selecting segments shared within the same individual (i.e. homozygosity-by-descent - HBD). Then, these segments were binned into 5 length categories (1–2, 2–4, 4–8, 8–16, >16), which are informative about events that happened in different time frames (longer segments were formed more recently, and vice-versa). Then the average length and number of HBD was calculated for major groups as well as for populations, and finally the average length of IBD shared in the intra-population level was also estimated. The data was grouped by the segment length categories and the averages were obtained for each of them.

Additionally, we used IBDNe (25) to estimate N_e history also based on the inferred IBD segments (section "Patterns of shared ancestry"), but first we selected those located inside the blocks of Native American ancestry identified through a Local Ancestry Inference conducted with RFMix (Maples et al. 2013). The IBDNe was then applied to the IBD segments of Native American ancestry to infer the ancestry-specific N_e history, using the default parameters, with exception to the parameter "filtersamples=false", which was used due to the small sample sizes per group, and for this reason we did not excluded the related samples from the analysis and therefore the estimates obtained for recent generations are expected to be particularly biased by the presence of relatedness between samples. In order to assess this bias we also performed a second set of analyses in which we used the parameter "filtersamples=true", hence removing related samples, and it is presented in the Supplementary Materials.

Finally, we combined the unadmixed set of Native Americans, with a dataset containing information from 952 individuals from worldwide populations from Human Genome Diversity Project (HGDP) (Bergström et al. 2020) and from Simons Genome Diversity Project (SGDP) (Mallick et al. 2016), resulting in a merger of 1,102 individuals. Then a quality control was performed by removing: (1) autosomal triallelic markers, (2) SNPs within the 2Mb of the extremities of all chromosome arms, (3) loci with extreme deviations from Hardy-Weinberg proportions (p-value $\leq 10^{-8}$), and (4) SNPs with more than 10% of missing values. The resulting set of markers is composed of 251,940 autosomal SNPs.

Next, the ROH identification was performed with PLINK v1.9 (Chang et al. 2015), using a sliding window of 50 SNPs, with a maximum of five missing calls and one heterozygous genotype per window, a proportion of 5% of overlapping windows in an homozygous segment, with at least one SNP each 50kb, a maximum gap of 100kb between consecutive SNPs, and a minimum ROH

length of 500kb. The ROH were used to estimate the individual inbreeding coefficient from ROH, as proposed by McQuillan *et al.* (McQuillan et al. 2008), in which the estimate corresponds to the genomic proportion composed by ROH, i.e. the total ROH length divided by the total genomic region covered by the SNPs. Then, population F_{ROH} averages were estimated from the individual values. Last, the F_{ROH} estimates were compared to the geographic position of these groups, by assessing the correlation between the F_{ROH} estimates for each group with their Longitude and Latitude.

References

Adhikari K, Chacón-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, Ruiz-Linares A. 2017. The Genetic Diversity of the Americas. *Annu. Rev. Genomics Hum. Genet.* 18:277–296.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–1664.

Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, Bustamante CD, Kenny EE, Williams SM, Aldrich MC, et al. 2016. The Great Migration and African-American Genomic Diversity. *PLoS Genet.* 12:e1006059.

Barbieri C, Barquera R, Arias L, Sandoval JR, Acosta O, Zurita C, Aguilar-Campos A, Tito-Álvarez AM, Serrano-Osuna R, Gray RD, et al. 2019. The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol. Biol. Evol.* 36:2698–2713.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32:2817–2823.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* [Internet] 367. Available from: http://dx.doi.org/10.1126/science.aay5012

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194:459–471.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103:338–348.

Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, Laurie CC. 2018. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* 14:e1007385.

Bush MB, Nascimento MN, Åkesson CM, Cárdenes-Sandí GM, Maezumi SY, Behling H, Correa-Metrio A, Church W, Huisman SN, Kelly T, et al. 2021. Widespread reforestation before European influence on Amazonia. Science 372:484–487.

Carvajal FG de. 2011. Relación de Descubrimiento del río de las amazonas, edición y notas de Nieves Pinillos Iglesias, realizada para Babelia.

Castro e Silva MA, Ferraz T, Bortolini MC, Comas D, Hünemeier T. 2021. Deep genetic affinity between coastal Pacific and Amazonian natives evidenced by Australasian ancestry. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 118. Available from: http://dx.doi.org/10.1073/pnas.2025739118

Castro e Silva MA, Nunes K, Lemes RB, Mas-Sandoval À, Amorim CEG, Krieger JE, Mill JG, Salzano FM, Bortolini MC, da Costa Pereira A, et al. 2020. Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proceedings of the National Academy of Sciences* 117:2372–2377.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Gnecchi-Ruscone GA, Sarno S, De Fanti S, Gianvincenzo L, Giuliani C, Boattini A, Bortolini E, Di

Corcia T, Sanchez Mellado C, Dàvila Francia TJ, et al. 2019. Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide. *Mol. Biol. Evol.* 36:1254–1269.

Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U. S. A.* 115:E6526–E6535.

Hornborg A. 2005. Ethnogenesis, Regional Integration, and Ecology in Prehistoric Amazonia. *Current Anthropology* [Internet] 46:589–620. Available from: http://dx.doi.org/10.1086/431530

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. *Nature* 536:419–424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93:278–288.

McQuillan R, Leutenegger A-L, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, et al. 2008. Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83:359–372.

Michael L. 2014. On the Pre-Columbian Origin of Proto-Omagua-Kokama. *Journal of Language Contact* 7:309–344.

Montenegro RA, Stephens C. 2006. Indigenous health in Latin America and the Caribbean. *Lancet* 367:1859–1869.

Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, et al. 2018. Early human dispersals within the Americas. *Science* [Internet] 362. Available from: http://dx.doi.org/10.1126/science.aav2621

Noelli FS. 2008. The Tupi Expansion. *The Handbook of South American Archaeology* [Internet]:659–670. Available from: http://dx.doi.org/10.1007/978-0-387-74907-5_33

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192:1065–1093.

Pearce AJ, Beresford-Jones DG, Heggarty P. 2020. Rethinking the Andes–Amazonia Divide: A cross-disciplinary exploration. UCL Press

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8:e1002967.

Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E, et al. 2018. Reconstructing the Deep Population History of Central and South America. *Cell* 175:1185–1197.e22.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. 2015. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349:aab3884.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al. 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* 506:225–229.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing Native American population history. *Nature* 488:370–374.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. 2015. Genetic evidence for two founding populations of the Americas. *Nature* 525:104–108.

Staples J, Nickerson DA, Below JE. 2013. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet. Epidemiol.* 37:136–141.

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SD, Santos FR. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. *Am. J. Hum. Genet.* 68:1485–1496.

Tournebize R, Chu G, Moorjani P. 2020. Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. *bioRxiv* [Internet]. Available from: https://www.biorxiv.org/content/10.1101/2020.09.07.286450v1.abstract

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328.

Acknowledgments

We are grateful for the continuous support of Professor Francisco M. Salzano (in memoriam) in the development of this project. We also thank all the native communities who participated in the study.

Funding: M.A.C.e.S., T.F., C.M.C.S and R.B.L.. were supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; 2018/013716; 2016/12371-1; 17/14916-8, 2020/10136-0; 2015/26875-9) and K.N. was funded by NIH (R01 GM075091).

Author contributions: M.A.C.S., T.F., C.M.C.S, R.B.L. and K.N.. performed the analyses and M.A.C.S. and T.H. wrote the paper and incorporated inputs from other authors.

Competing interests: The authors declare no conflicts of interest.

Data and material availability: The datasets reported in this paper have been deposited in the European Genome-phenome Archive and are available for download under the accession numbers EGAD00010002061 and EGAD00010001803.

Tables

Table 1 - Description of samples and major group labels used throughout the text. Including continental region of origin, number of individuals sampled from each major group, and also the label and number of individuals from each indigenous group, along with each respective reference. See Dataset S1 for a complete description of the samples. Major group labels are based on affiliations to common linguistic stocks, or geographic and genetic proximity (in cases where no detailed linguistic information was available).

Continental Region	Major group	N	Description
Eastern South America	Jê	45	Jê speakers from northeastern Amazonia and central Brazilian plateau: Xavante (N = 27) (Castro e Silva et al. 2021); Xavante (N = 11) (Skoglund et al. 2015); Xikrin (N = 7) (Castro e Silva et al. 2021)
	Karib	8	Karib speakers from northeastern Amazonia: Apalai (N = 4) (Skoglund et al. 2015); Arara (N = 4) (Skoglund et al. 2015)
	Tupí	65	Tupí speakers from Amazonia, central-west Brazil and Brazilian Atlantic coast: Asurini (N = 1) (Castro e Silva et al. 2021); Gavião (N = 2) (Castro e Silva et al. 2020); Guaraní Kaiowá (N = 10) (Skoglund et al. 2015); Guaraní Mbyá (N = 4) (Castro e Silva et al. 2020); Guaraní Nãndeva (N = 7) (Skoglund et al. 2015); Karitiana (N = 13) (HGDP (Patterson et al. 2012)); Karitiana (N = 4) (Skoglund et al. 2015); Munduruku (N = 2) (Castro e Silva et al. 2021); Parakanã (N = 3) (Castro e Silva et al. 2020); Surui (N = 8) (HGDP (Patterson et al. 2012)); Surui (N = 4) (Skoglund et al. 2015); Tupiniquim (N = 1) (Castro e Silva et al. 2020); Urubu Kaapor (N = 3) (Skoglund et al. 2015); Wajãpi (N = 2) (Castro e Silva et al. 2020); Zoró (N = 1) (Skoglund et al. 2015)
Mesoamerica and northern Mexico	Mayan	21	Mayan speakers from the Mexican Yucatán Peninsula: Maya (N = 21) (HGDP (Patterson et al. 2012))
	Mixe-Zoque	12	Mixe from southern Mexico: Mixe (N = 10) (Lazaridis et al. 2014); Mixe.DG (N = 2) (Skoglund et al. 2015)
	Oto-Manguean	24	Mixtec and Zapotec from southern Mexico: Mixtec (N = 10) (Lazaridis et al. 2014); Mixtec.DG (N = 2) (SGDP (Mallick et al. 2016)); Zapotec (N = 10) (Lazaridis et al. 2014); Zapotec.DG (N = 2) (SGDP (Mallick et al. 2016))
	Uto-Aztecan	22	Pima and Yaquis from northern Mexico: Pima (N = 14) (HGDP (Patterson et al. 2012)); Yaquis (N = 8) (Barbieri et al. 2019)
Western South America	Andes	31	Descendants of indigenous groups from central Peruvian Andes: Huancas (N = 5) (Barbieri et al. 2019); LaJalca (N = 11) (Barbieri et al. 2019); Luya (N = 11) (Barbieri et al. 2019); UtcubambaSouth (N = 4) (Barbieri et al. 2019)
	Arawak	6	Arawak speakers from central Peru (eastern Andean slopes) and southern Colombia: Chamicuro (N = 1)

			(Barbieri et al. 2019); Piapoco (N = 5) (HGDP (Patterson et al. 2012))
	Aymara	9	Aymara from southern Peruvian Andes and Aymara descendants from northern Bolivian Andes (Lake Titicaca and adjacent regions): Aymara (N = 2) (Barbieri et al. 2019); Bolivian (N = 7) (Lazaridis et al. 2014)
	Cahuapanan	1	Shawi from central Peru (eastern Andean slopes): Shawi (N = 1) (Barbieri et al. 2019)
	Language Isolate	10	Speakers of language isolates from southwestern Colombia and central Peru (eastern Andean slopes): Cofan (N = 4) (Barbieri et al. 2019); Kamentsa (N = 4) (Barbieri et al. 2019); Muniche (N = 2) (Barbieri et al. 2019)
	Pacific Coast	58	Descendants of indigenous groups from the Pacific Coast of central and northern Peru: Cao (N = 10) (Barbieri et al. 2019); Chotuna (N = 4) (Barbieri et al. 2019); Chulucanas (N = 8) (Barbieri et al. 2019); Eten (N = 5) (Barbieri et al. 2019); Narihuala (N = 5) (Barbieri et al. 2019); Olmos (N = 4) (Barbieri et al. 2019); Paran (N = 3) (Barbieri et al. 2019); Sechura (N = 3) (Barbieri et al. 2019); Tallan (N = 8) (Barbieri et al. 2019); Tumbes (N = 8) (Barbieri et al. 2019)
	Quechua	61	Quechua speakers from southern Colombia, northern Ecuador, central Peru (western Amazonia), southern Peruvian Andes and Lake Titicaca region: Inga (N = 13) (Barbieri et al. 2019); Kichwa (N = 17) (Barbieri et al. 2019); Quechua Cusco (N = 3) (Barbieri et al. 2019); Quechua Cusco2 (N = 7) (Barbieri et al. 2019); Quechua Cusco3 (N = 5) (Lazaridis et al. 2014); Quechua Puno (N = 5) (Barbieri et al. 2019); Quechua.DG (N = 1) (SGDP (Mallick et al. 2016)); Wayku (N = 10) (Barbieri et al. 2019)
	Tupí	10	Tupí speakers from southern Colombia and central Peru (eastern Andean slopes): Kokama Colombia (N = 7) (Barbieri et al. 2019); Kokama Peru (N = 3) (Barbieri et al. 2019)

Figures

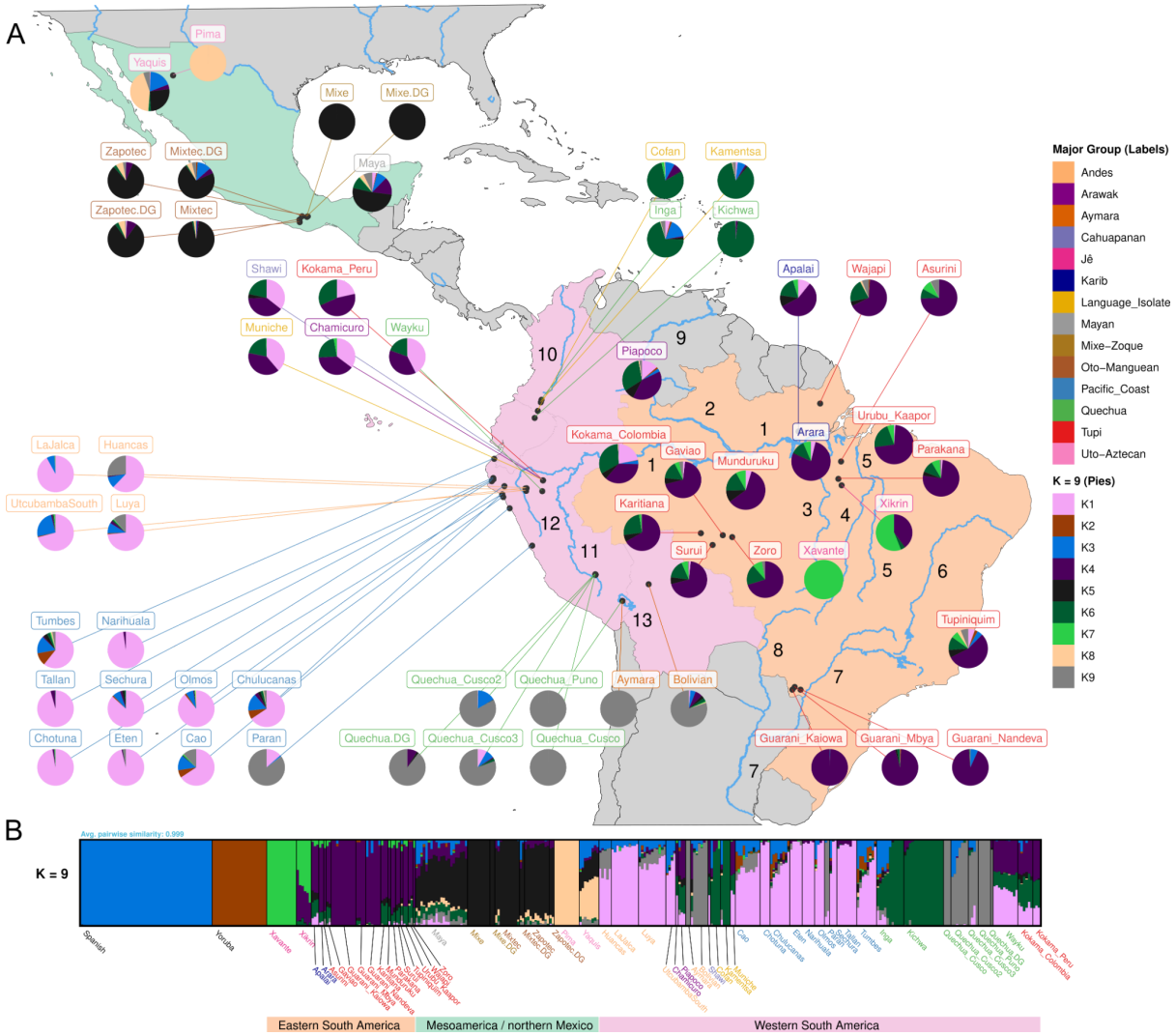


Figure 1 - Genetic structure of the Native Americans. An unsupervised admixture analysis with the number of putative ancestry components (K) ranging from 2 to 10 was applied to the LD-pruned set of unrelated Native Americans and the results with K=9 are shown here. **(A)** A partial map of the American continent with mean putative ancestry component estimates per group plotted in their approximate sampling locations. **(B)** Bar plot of the individual ancestry component estimates created using PONG (Behr et al. 2016). In **(A)** and **(B)**, the name tags and the putative ancestry components are color coded as indicated by the legends on the right. Finally, the three main continental regions are indicated by color shade in **(A)** and colored bar at the bottom in **(B)**: Mesoamerica and northern Mexico in light green, western South America in pink and eastern South American in beige. Some of the main South American rivers (and lakes) are indicated by numbers: (1) Amazon, (2) Negro, (3) Xingu, (4) Araguaia, (5) Tocantins, (6) São Francisco, (7) Paraná, (8) Paraguay, (9) Orinoco, (10) Magdalena, (11) Ucayali, (12) Marañón, and (13) Lake Titicaca.

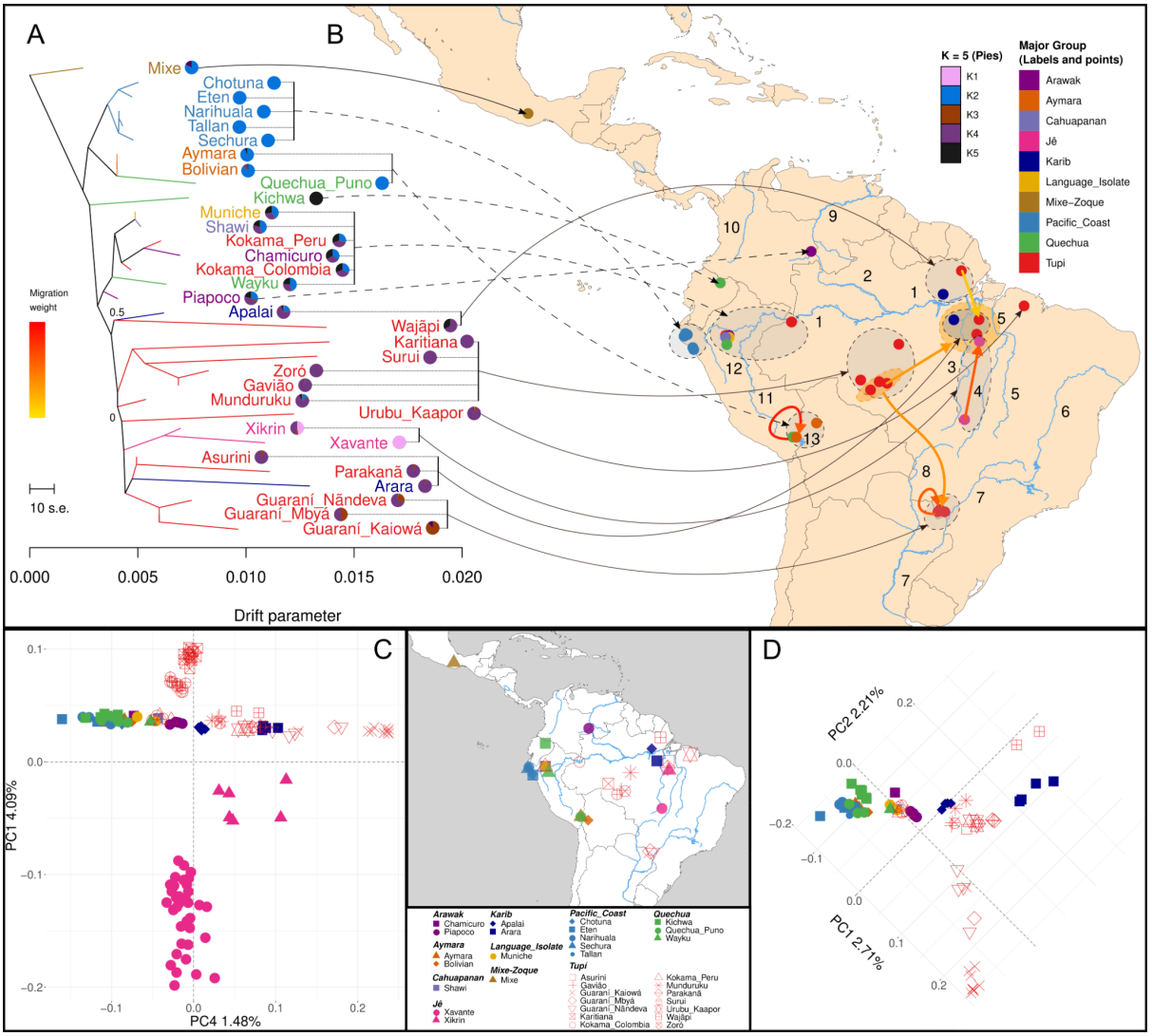


Figure 2 - Population diversification patterns reflect geographic distribution. Using the LD-pruned subset of unadmixed and unrelated Native Americans, **(A)** a Maximum Likelihood (ML) tree was estimated based on pairwise population covariance using Treemix (Pickrell and Pritchard 2012) and gene flow events were progressively modeled between the branches of the ML tree with the poorest fit. The model likelihood reaches a plateau at six gene flow events; therefore, we additionally present these gene flow events (exclusively in **B**). Using the same dataset we also performed an unsupervised admixture analysis and we present the results with K = 5 as pie charts at the right side of each group in the ML tree. **(B)** Group geographic locations are indicated as points on a map, which along with group labels on the ML tree **(A)** are color coded to indicate affiliation to the major groups. Finally, we also link the groups on the ML tree **(A)** to their geographic locations on the map **(B)** (black arrows; to enhance readability solid and dashed lines are used for eastern and western South American groups, respectively), as well as gene flow events (color coded arrows) inferred in the six gene flows model, indicating their direction (arrow heads) and intensity (color coded). In **(B)** Some of the main South American rivers (and lakes) are indicated by numbers: (1) Amazon, (2) Negro, (3) Xingu, (4) Araguaia, (5) Tocantins, (6) São Francisco, (7) Paraná, (8) Paraguay, (9) Orinoco, (10) Magdalena, (11) Ucayali, (12) Marañón, and (13) Lake Titicaca. PCAs were also applied to **(C)** this dataset, and we show PC1 and PC4, because it captures the longitudinal cline, and to **(D)** a subset excluding the outliers (namely: Xavante, Xikrin, Karitiana, and Suruí), where PC1 and PC2 are shown. The groups and major group affiliations are coded as shapes and colors, respectively, as indicated in the legend and map at the bottom center.

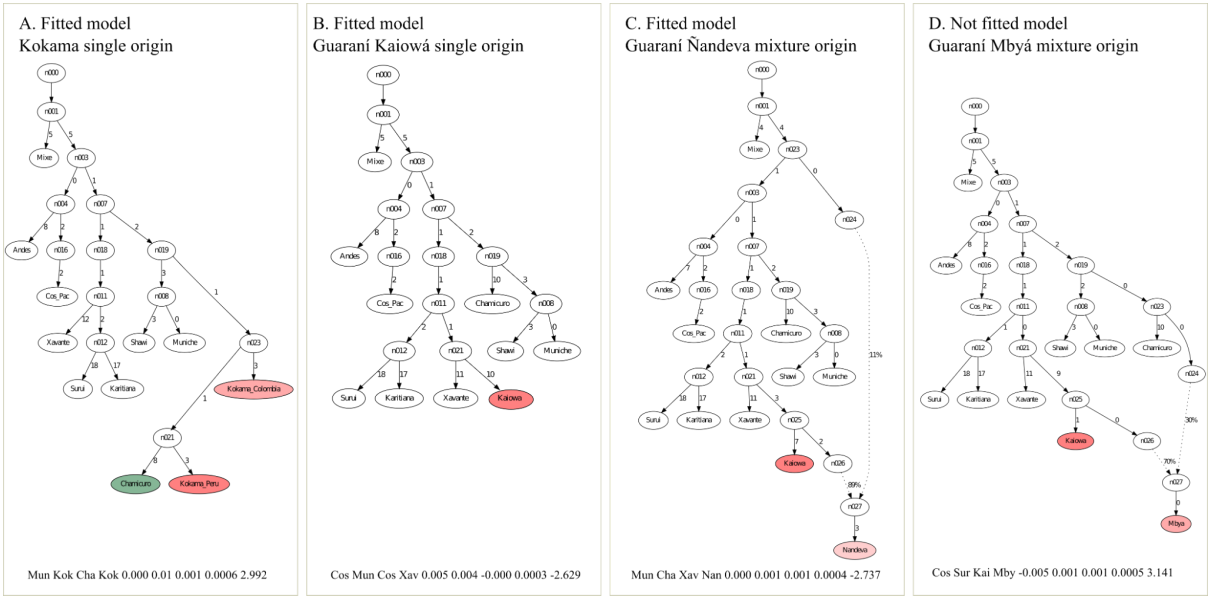


Figure 3 - Population history of Native South Americans. We leveraged the unadmixed and unrelated subset of Native American groups to model the possible ancestral contributions to the Kokama and Guaraní ethnolinguistic groups. **(A)** Best fitted model for the Kokama group (from Peru and Colombia) - single origin. **(B)** Best fitted model for the Guaraní Kaiowá group - single origin. **(C)** Best fitted model for Guaraní Nãndeva group - mixed origin. **(D)** Best fitted model for Guaraní Mbyá group - mixed origin.

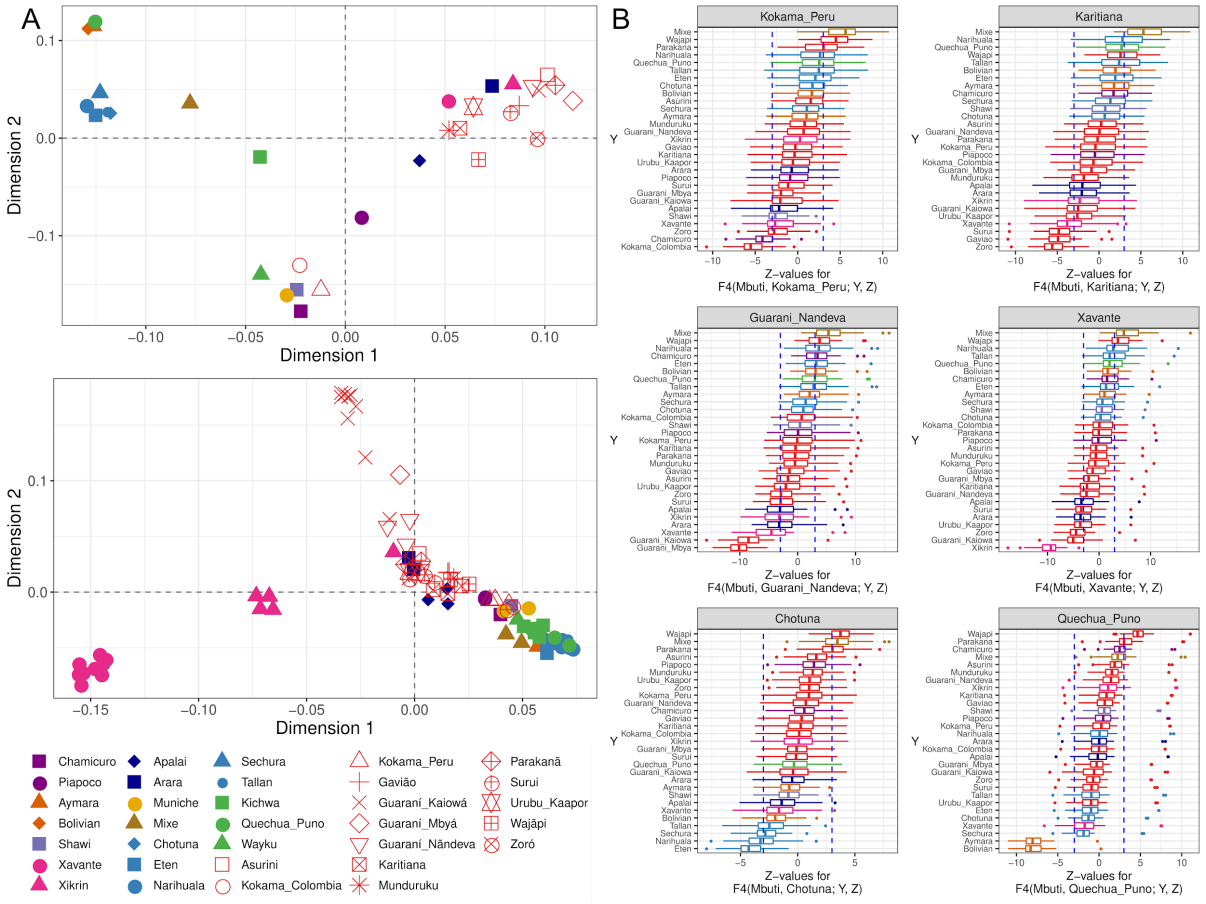


Figure 4 - Genetic affinities of present-day Native Americans. (A) Using the subset of unadmixed and unrelated Native Americans we estimated genetic distances as 1 – outgroup F_3(Mbuti; Y, Z), where Y and Z are any indigenous group or individual. MDS was then applied to the matrices of pairwise genetic distances. (**A Top**) MDS of the population-wise genetic distances matrix. (**A bottom**) MDS of the individual-wise genetic distances matrix. The legend at the right bottom shows the symbol and color used for each present-day group. **(B)** We also estimated F_4(Mbuti, X; Y, Z), where X, Y, and Z are any group. Here, we present the estimates for a subset of tests of particular interest; for the complete set of tests, see Figure S8. In each panel, one X and Y test groups are shown at the top stripe and in the y-axis, respectively, while the x-axis presents the Z-values for each estimate obtained by the comparison with every Z test group, in the form of boxplots, which are sorted by the median Z-value. Therefore, Z-values below −3 indicate an excess affinity between the Y test group (*y*-axis) and the Z test group (top stripe). The complete sets of F_3 and F_4 statistics are presented in Dataset S5A-B and Dataset S6A, respectively.

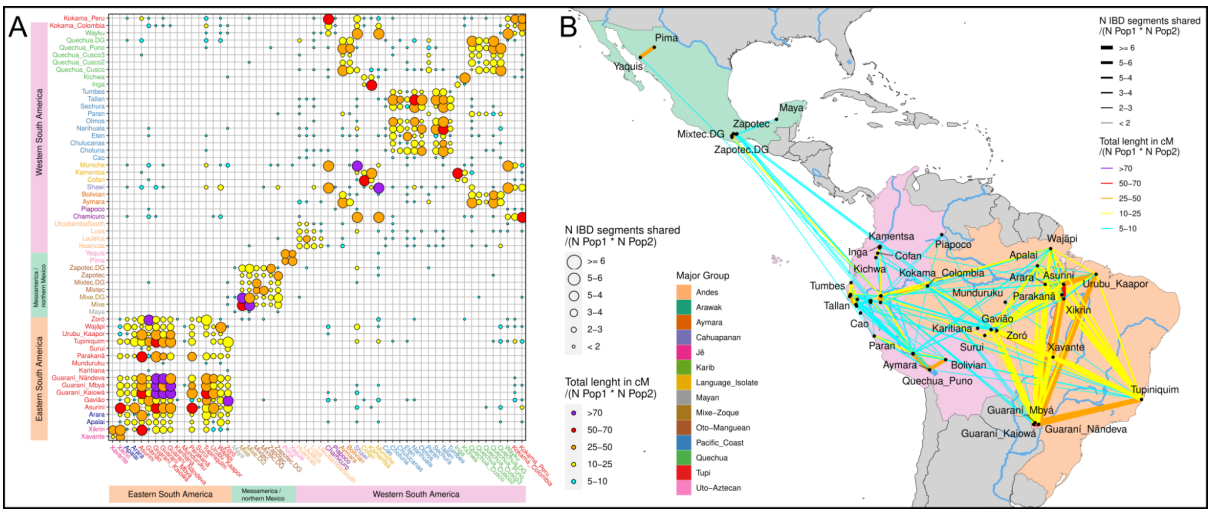


Figure 5 - Network of Pre-Columbian IBD sharing among present-day Native American groups. The IBD genomic segments were identified based on the phased data subset of unrelated Native Americans, then these segments were filtered to select only those inferred to be in genomic regions of Native American local ancestry. Segments shorter than 2 cM were removed and pairwise connections with less than 5 cM shared on average were also not considered. Here, we present the results obtained using IBD segments with at most 9 cM of length, which correspond to those originated in the Pre-Columbian period (approximately before 1500 CE). The average number of IBD segments (color) and the average length of IBD in cM (size), are shown as a matrix (**A**) and as a network on a map (**B**). The classification of populations into major groups is also color coded, as indicated in the legend at the center (axes labels in **A**). The three main continental regions are indicated by a set of colored bars at the left and at the bottom of **A**, matching the same colors used in the map regions in **B**. The intrapopulation IBD is shown in the diagonal in **A**. Some group labels are shown in **B** for reference. For the patterns IBD sharing in the colonial and recent periods, see Figure S13. The complete set of IBD segments inferred are presented in Dataset S4.

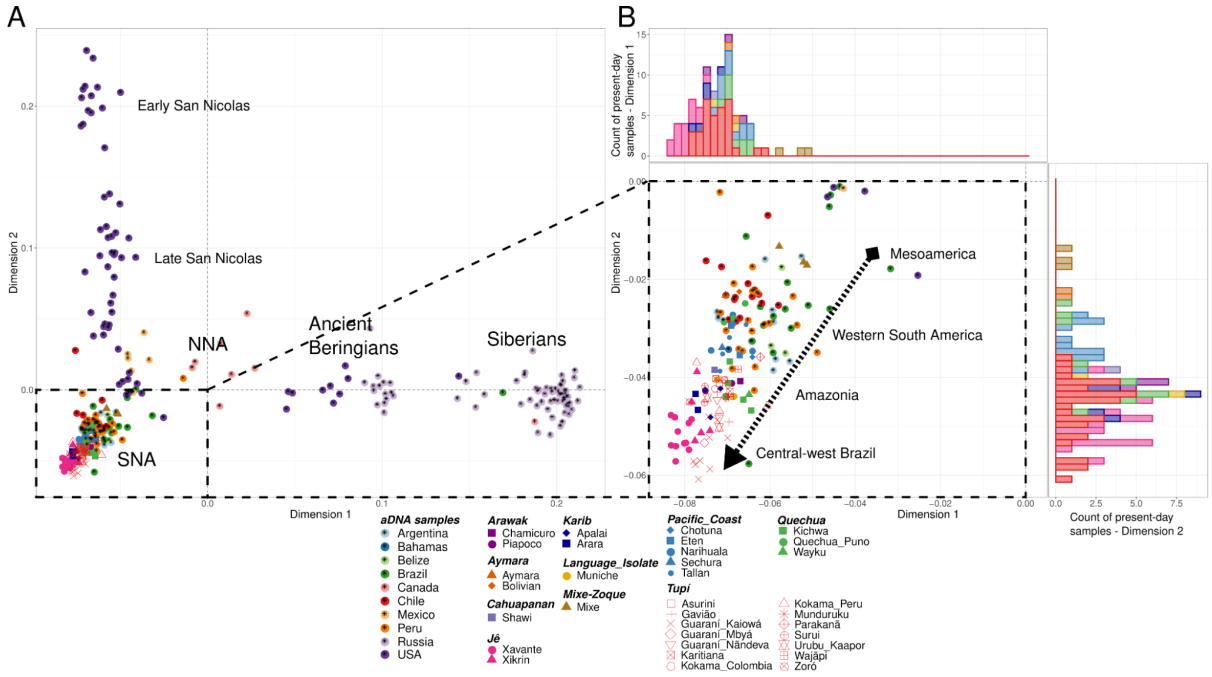


Figure 6 - Using a dataset combining the subset unadmixed and unrelated present-day Native Americans and ancient DNA samples from across the whole American continent and Siberia, we estimated the pairwise genetic distances as 1 – outgroup F_3(Mbuti; Y, Z), where Y and Z are any present day and ancient individuals. The matrix of pairwise distances between individuals was then used to perform an MDS. **(A)** Complete MDS plot, highlighting some of the most differentiated groups. **(B)** Inset showing the Southern Native Americans (SNA) with more detail. To enhance pattern visualization, the top and right bar plots in **(B)** exhibit only the count of present-day samples from each major group (color code) at each position. The legend at the bottom shows the symbol and color used for each present-day group or for the country of origin of each aDNA sample. The complete set of statistics is presented in Dataset S5C.

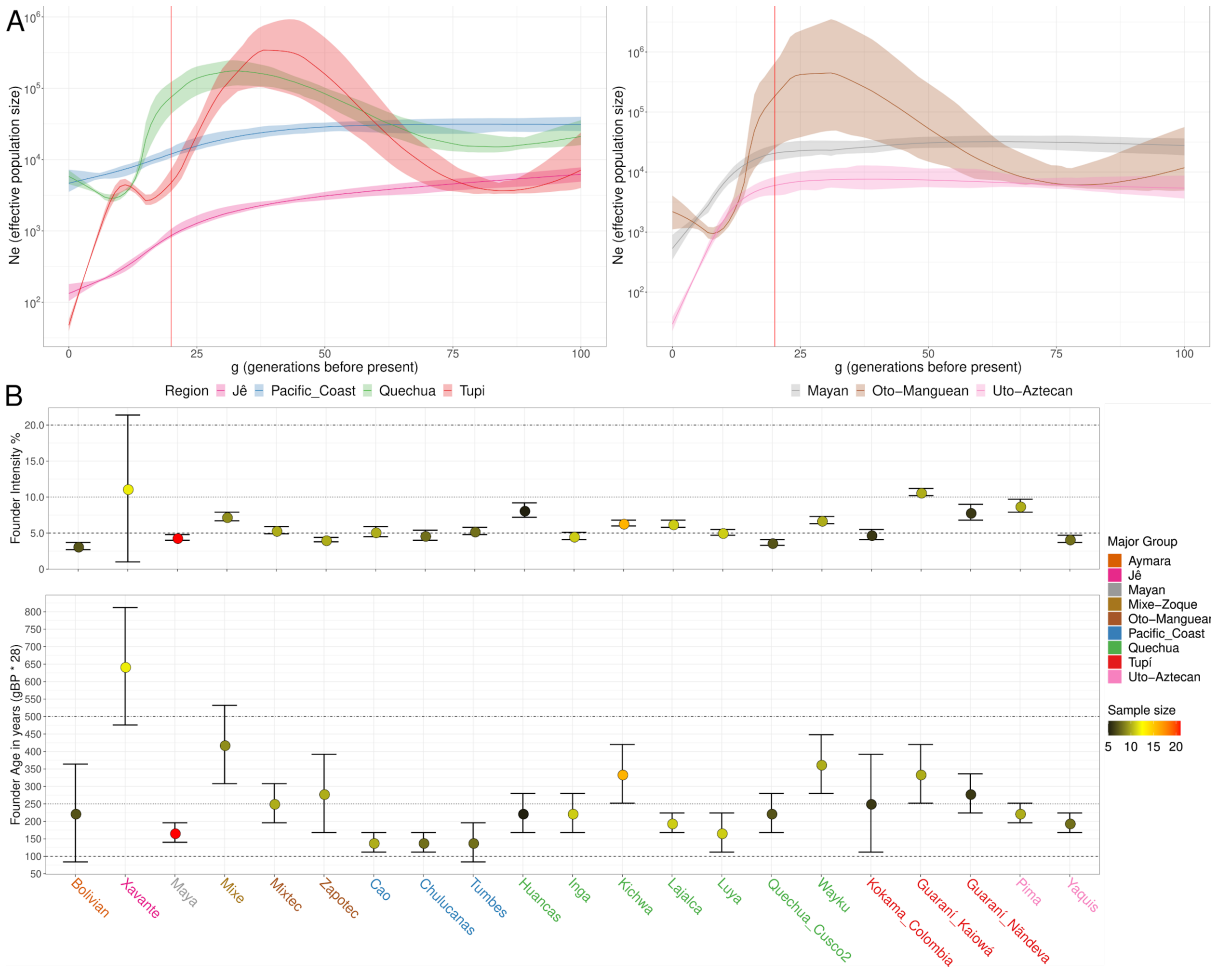


Figure 7 - Native American effective population size (N_e) histories and the post-Contact population collapse. (A) The IBD genomic segments were identified with the phased data set of Native American groups, followed by a selection of the segments inferred to be in genomic regions of Native American ancestry. The complete set of IBD segments was separated into subsets of major groups from South America **(A left)** and Mesoamerica/Northern Mexico **(A right)**, and then each set was used to infer the N_e history of each specific major group. The ancestry-specific Ne values are coded in the y axis (log scale) and indicated by a line for each generation before present (gBP) depicted in the x-axis. The shaded areas show a 95% bootstrap confidence interval for each major group. The vertical red line indicates 20 gBP (approximately 1500 CE) and therefore the time of the first contacts with Europeans. Here, we show the results of IBDNe using the parameter filtersamples = "false", alternatively the results produced with the parameter filtersamples = "true" are shown in Figure S18. **(B)** We also applied the ASCEND method to every Native American group with more than 5 unrelated samples (and also to some clusters of groups in order to reach the minimum sample size of 5, see Figure S15). In **B,** the top panels depict the founder intensity (FI), and the bottom panels show the mean estimate of the founder age (FA) for each indigenous group. For each group, the estimated FI and FA are shown along with their associated 95% confidence interval. The sample size is color coded on the points and the affiliations with major groups are indicated in the group label IDs at the x-axis, both indicated in the legend. In the top panels, the y-axis indicates the FI percentage and in the bottom panel, the y-axis shows the estimated FA calculated as: 'x' generation before present (gBP) * 28 years per generation = 'y' years before present (BP).

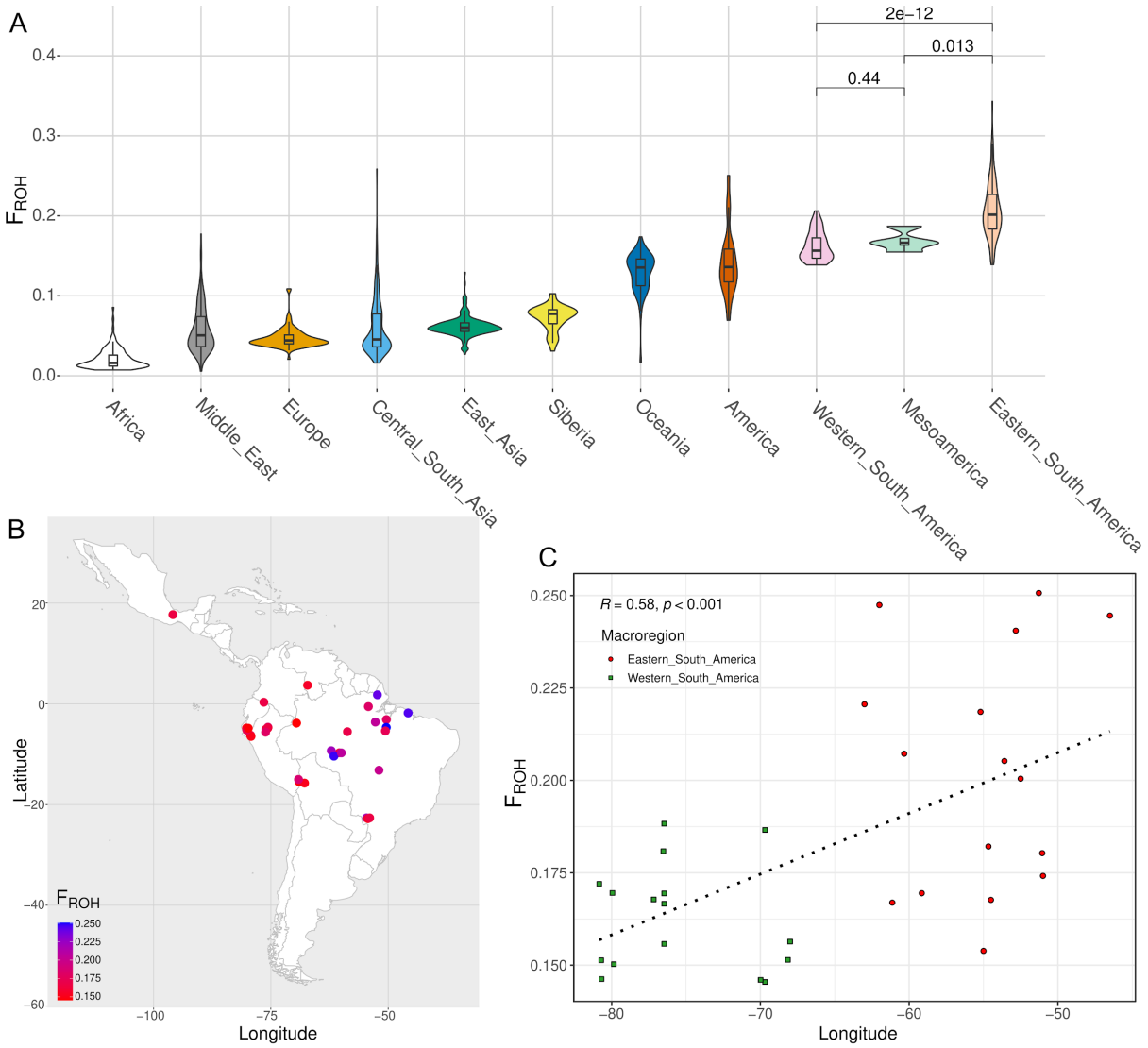


Figure 8 - Distribution of inbreeding coefficient from ROH in Native Americans. (A) The distribution of F_{ROH} was obtained averaging the individual estimates from a combined set of the unadmixed Native Americans along with HGDP and SGDP databases (Africa, Middle East, Europe, Central South Asia, East Asia, Siberia, and America). The p-values were obtained from a nonparametric Wilcoxon rank-sum test. **(B)** Population average estimates of F_{ROH} were plotted according to the corresponding geographic location. **(C)** Correlation of F_{ROH} values according to the longitude of each population. The dotted line was estimated by linear regression. The Spearman correlation coefficient and its corresponding p-value are also presented.

CAPÍTULO 3

Inferências genômicas sobre as origens e dispersão dos nativos da costa brasileira

Manuscrito: CASTRO E SILVA, M. A. 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, 2020.

Autores: Marcos Araújo Castro e Silva[a], Kelly Nunes[a], Renan Barbosa Lemes[a], Àlex Mas-Sandoval[b,c], Carlos Eduardo Guerra Amorim[d], Jose Eduardo Krieger[e], José Geraldo Mill[f], Francisco Mauro Salzano[b,1], Maria Cátira Bortolini[b], Alexandre da Costa Pereira[e], David Comas[c], and Tábita Hünemeier[a,2]

[a]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil 05508-090; [b]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil 91501-970; [c]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [d]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095; [e]Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil 05403-000; and [f]Departamento de Fisiologia, Universidade Federal do Espírito Santo, Espírito Santo, Brazil 29040-090

Resumo:

No século 15 aproximadamente 900.000 nativos americanos, a maioria deles falantes de Tupí, viviam na costa brasileira. Ao fim do século 18, as populações nativas da costa foram declaradas extintas. Os Tupí chegaram a costa leste depois de partir da bacia amazônica a aproximadamente 2.000 anos antes dos presente; contudo, não há um consenso sobre como esta migração ocorreu: primeiramente em direção ao norte amazônico e depois diretamente para a costa atlântica ou avançando para o sul pelo interior do continente e então migrando para a costa. Aqui nós exploramos os dados genômicos de um dos últimos representantes putativos da linhagem dos Tupí da costa, uma comunidade pequena e miscigenada de autodeclarados Tupiniquim, assim como dados de uma população Guaraní Mbyá nativa do sul do Brasil e outras três populações nativas da região amazônica. Nós demonstramos que a ancestralidade nativa americana dos Tupiniquim não é relacionada às outras populações nativas

brasileiras estudadas até aqui e portanto podem ser considerados os únicos representantes vivos do extinto ramo Tupí que ocupava a costa atlântica brasileira. Além disso, nossos dados evidenciam uma migração direta da Amazônia para a costa do Nordeste no período pré-contato, dando origem às populações Tupí da costa, ao passo que uma migração distinta se dirigiu ao sul originando os povos Guaraní do Brasil e Paraguai. Este estudo elucida as dinâmicas populacionais e a diversificação dos nativos brasileiros no nível genômico, o que se tornou possível devido ao resgate de dados da população brasileira costeira original através dos genomas de indivíduos miscigenados.

Palavras-chave: Nativos americanos | povoamento da América do Sul | falantes de Tupí | Brasileiros | genética

Abstract:

In the 15th century, ~900,000 Native Americans, mostly Tupí speakers, lived on the Brazilian coast. By the end of the 18th century, the coastal native populations were declared extinct. The Tupí arrived on the east coast after leaving the Amazonian basin ~2,000 y before present; however, there is no consensus on how this migration occurred: toward the northern Amazon and then directly to the Atlantic coast, or heading south into the continent and then migrating to the coast. Here we leveraged genomic data from one of the last remaining putative representatives of the Tupí coastal branch, a small, admixed, self-reported Tupiniquim community, as well as data of a Guaraní Mbyá native population from Southern Brazil and of three other native populations from the Amazonian region. We demonstrated that the Tupiniquim Native American ancestry is not related to any extant Brazilian Native American population already studied, and thus they could be considered the only living representatives of the extinct Tupí branch that used to settle the Atlantic Coast of Brazil. Furthermore, these data show evidence of a direct migration from Amazon to the Northeast Coast in pre-Columbian time, giving rise to the Tupí Coastal populations, and a single distinct migration southward that originated the Guaraní people from Brazil and Paraguay. This study elucidates the population dynamics and diversification of the Brazilian natives at a genomic level, which was made possible by recovering data from the Brazilian coastal population through the genomes of mestizo individuals.

Keywords: Native Americans | peopling of South America | Tupí speakers | Brazilian | genetics

Genomic insight into the origins and dispersal of the Brazilian coastal natives

Marcos Araújo Castro e Silva[a], Kelly Nunes[a], Renan Barbosa Lemes[a], Àlex Mas-Sandoval[b,c], Carlos Eduardo Guerra Amorim[d], Jose Eduardo Krieger[e], José Geraldo Mill[f], Francisco Mauro Salzano[b,1], Maria Cátira Bortolini[b], Alexandre da Costa Pereira[e], David Comas[c], and Tábita Hünemeier[a,2]

[a]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil 05508-090; [b]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil 91501-970; [c]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [d]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095; [e]Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil 05403-000; and [f]Departamento de Fisiologia, Universidade Federal do Espírito Santo, Espírito Santo, Brazil 29040-090

Edited by Anne C. Stone, Arizona State University, Tempe, AZ, and approved December 12, 2019 (received for review May 25, 2019)

In the 15th century, ∼900,000 Native Americans, mostly Tupí speakers, lived on the Brazilian coast. By the end of the 18th century, the coastal native populations were declared extinct. The Tupí arrived on the east coast after leaving the Amazonian basin ∼2,000 y before present; however, there is no consensus on how this migration occurred: toward the northern Amazon and then directly to the Atlantic coast, or heading south into the continent and then migrating to the coast. Here we leveraged genomic data from one of the last remaining putative representatives of the Tupí coastal branch, a small, admixed, self-reported Tupiniquim community, as well as data of a Guaraní Mbyá native population from Southern Brazil and of three other native populations from the Amazonian region. We demonstrated that the Tupiniquim Native American ancestry is not related to any extant Brazilian Native American population already studied, and thus they could be considered the only living representatives of the extinct Tupí branch that used to settle the Atlantic Coast of Brazil. Furthermore, these data show evidence of a direct migration from Amazon to the Northeast Coast in pre-Columbian time, giving rise to the Tupí Coastal populations, and a single distinct migration southward that originated the Guaraní people from Brazil and Paraguay. This study elucidates the population dynamics and diversification of the Brazilian natives at a genomic level, which was made possible by recovering data from the Brazilian coastal population through the genomes of mestizo individuals.

Native Americans | peopling of South America | Tupí speakers | Brazilian natives | genetics

In the 15th century, the Brazilian coast was densely populated by Native American populations. At that time, a total of 3 million indigenous individuals lived in the territory currently corresponding to Brazil, with about a third inhabiting its coast (1). The conquest of the Brazilian territory by the Portuguese (circa 1500) led to a rapid decline of the coastal native populations, culminating in their extinction by the end of the 18th century (2). This massive depopulation completely changed the distribution of the Native American populations within Brazil, delimiting their territory to the Amazon region and the inland. At present there are just two small admixed communities self-reported as coastal Tupí (Tupiniquim and Tupinambá) living in Brazil; however, they do not speak any indigenous language.

When the Portuguese first arrived in South America, the Tupiniquim and Tupinambá, both originally Tupí speakers, were the dominant groups in the Brazilian Atlantic Coast (2). It is not clear how the Tupí speakers arrived on the east coast after they left the Amazonian basin. The origins of the Proto-Tupí (Amazonian, southern, and coastal Tupí ancestrals) dates back to possibly 5,000 y before present (YBP) in the Northwest Amazon (ref. 3 and references therein). More than 2,000 YBP, different Tupí populations expanded from this region over 4,000 km eastward and southward, respectively peopling the Atlantic coast and the western Brazilian inland. They expanded to most of the South American lowlands during the late Holocene epoch, becoming one of the most populous and diverse linguistic families (with >35 languages still spoken). The Tupí expansion is comparable in importance to the Bantu expansion in Africa; however, relatively little is known about the event. There is no consensus in the literature regarding linguistic expansion models for the Tupí family (4, 5). Genetic studies based on uniparental markers are consistent with linguistic data indicating that northwestern Amazon was the center of diversification of the Tupí (3, 6), but they do not define any clear route of expansion, mainly due to lack of data from coastal populations. The causes of expansion are also unknown, and could have involved ecological adaptation or cultural issues (7). The Tupí-Guaraní branch (which includes coastal and southern Tupí groups) has assumed an expansionist character over the last 2,000 to 3,000 y, populating the Brazilian southwest, northeast, and entire coast, distinguishing them from the other Tupí speakers. On the basis primarily of archeological and linguistic evidence (2, 8), two main broad and contrasting hypotheses

Significance

The indigenous populations of the Brazilian coast were decimated by European conquerors and declared extinct by the 18th century. The disappearance of these populations created a gap in the understanding of South American settlement. The present study rescues the genome of an extinct coastal lineage of the Tupí branch through the examination of a small, admixed, self-reported Native American community. Our results suggest that genetic lineages representative of the Tupí peoples who inhabited the coast survived in this specific extant population. We also show the relationships among Coastal, Amazonian, and ancient Brazilian populations and elucidate the putative migratory routes used by Amazonian peoples between the Amazon and the Atlantic coast ∼2,000 y ago.

Author contributions: T.H. designed research; M.A.C.e.S., D.C., and T.H. performed research; A.M.-S., J.E.K., J.G.M., F.M.S., M.C.B., A.d.C.P., and T.H. contributed new reagents/analytic tools; J.G.M. and F.M.S. collected the biological data; M.A.C.e.S., K.N., and R.B.L. analyzed data; and M.A.C.e.S. and T.H. wrote the paper with contributions from C.E.G.A., M.C.B., A.d.C.P., and D.C.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: The newly genotyped datasets reported in this paper have been deposited in the European Genome-phenome Archive and are available for download under accession no. EGAS00001004036.

[1]Deceased September 28, 2018.

[2]To whom correspondence may be addressed. Email: hunemeier@usp.br.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1909075117/-/DCSupplemental.

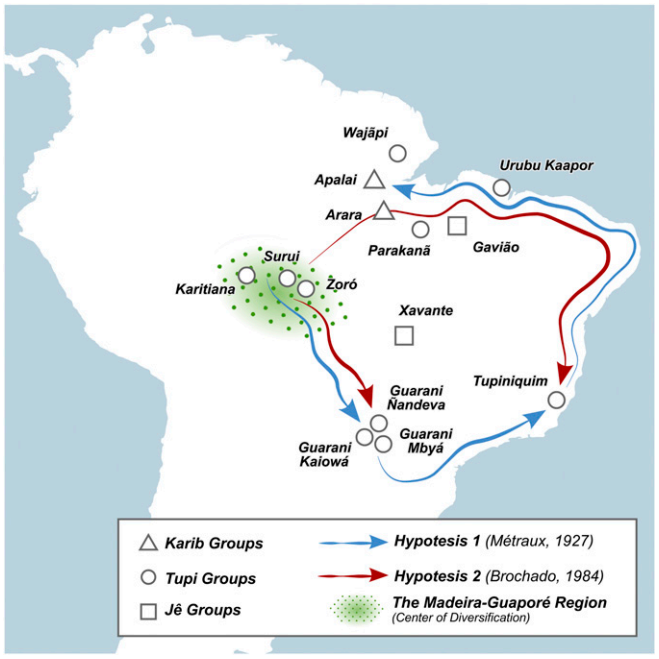


Fig. 1. Tupí Expansion hypotheses. Two main contrasting broad hypotheses can be recognized from literature (2, 8), which try to explain the Tupí Expansion. In hypothesis 1, the coastal Tupí would have derived from Guaraní populations in the south, which would have arrived there expanding southward from the Amazon Basin, here represented by the blue arrow. Conversely, hypothesis 2 postulates that the coastal Tupí and the Guaraní populations would have been originated in two separated expansions, with the former expanding eastward along the coast from the Amazon River mouth, the latter southward from the Amazon, here indicated by the two red arrows.

regarding the settlement of the Brazilian coast by the Tupí groups can be distinguished in the literature (Fig. 1). The first proposes that the Tupí from the Brazilian coast reached this region after coming from southwest Brazil, deriving from the same Tupí-Guaraní branch of Guaraní populations (9, 10) (blue arrow in Fig. 1). This hypothesis (10, 11) is based on archaeological data, linguistic analysis, and paleoenvironmental data, and associates the Tupí expansion with forest reductions that would have occurred during the Holocene. In this context, changes in vegetation would have forced these nonceramicist, preagriculturalist populations to seek new subsistence niches. Although these forest refuges were located both to the south and the east, linguistic data suggest that the most likely migration route to the Atlantic coast would have been through Brazil's western border, and then to the east shore. The alternative hypothesis assumes that one branch of Tupí moved first eastward, reaching the coast, and then southward along the coast, originating the coastal Tupí, whereas the other branch went southward, originating the Guaraní people (12) (red arrows in Fig. 1). According to this interpretation, the Proto-Tupí were already agriculturalists and ceramists, and the reason for their expansion was likely the demographic pressure caused by a continuous increase in population, which forced them to disperse in search of new lands to cultivate. This proposition (12) is motivated by the independent mode and evolution of Guaraní and Tupinambá potteries from the Amazonian Polychrome Tradition of Proto-Tupí speakers (characterized by the use of red and black paint on a white engobe). Tupinambá pottery is only found in the northeast Amazon and along the Brazilian coast to the Tropic of Capricorn, while Guarani pottery has been found from southern Amazon to northern Argentina, Paraguay, and southern Brazil.

To reconstruct the history of the Tupí, we generated genomic data for the last remaining putative representatives of the Tupí coastal branch, a small admixed self-reported community of Tupiniquim people; for a Guaraní native population from Southern Brazil; and for three native populations from the Amazonian region. We investigated their genetic origins and demonstrated that the Tupiniquim Native American ancestry is not related to any extant Brazilian Native American population for which genetic data have been generated to date. Therefore, we infer that the Tupiniquim are the only living representatives of this extinct Tupí branch that was settled along the Brazilian Atlantic Coast at the arrival of the Europeans. Leveraging genomic information of the coastal Tupí branch retrieved from these admixed individuals, we elucidated the pre-Columbian dispersion of the Tupí-stock from the Amazon to Southern Brazil and to the coast, finding evidence of two migrations: there was a direct migration from the Amazon to the coast, which originated the Tupí coastal populations, and a single distinct migration to south that originated the Guaraní people from Brazil and Paraguay. We further showed the existence of genetic continuity within Brazil when comparing ancient and modern individuals. The intensity of this continuity changed when the linguistic groups split and became structured, around 6,000 YBP, producing specific patterns of shared ancestry.

Results and Discussion

Overview of the Data. We generated data for more than 600,000 SNPs in 102 Native Americans from Brazil (47 Tupiniquim, 48 Guaraní Mbyá, 2 Wajãpi, 3 Parakanã and 2 Gavião; *SI Appendix*, Table S1). The following public datasets were also used in the analysis: 48 Native Americans (13), Human Genome Diversity Project dataset 11 (http://www.cephb.fr/hgdp/), 1000 Genomes Project (https://www.internationalgenome.org/), Anzick-1 (14), and 15 ancient DNA samples from Brazil (15) (*SI Appendix*, Table S2). *SI Appendix*, Fig. S1 shows the geographic position of the analyzed samples.

Postcontact History. We performed Global Ancestry Inferences of the Tupiniquim and Guaraní Mbyá with ADMIXTURE (16). The Tupiniquim community exhibits a greater proportion of European and African (25.9% and 22.54%, respectively) admixture (*SI Appendix*, Fig. S2*A* and Dataset S1) in comparison with the Guaraní Mbyá (15.62% and 7.1%) (*SI Appendix*, Fig. S2*B* and Dataset S1). The Tupiniquim presented higher Native American ancestry (51.55%) when compared with the general Brazilian population (~7%; refs. 17 and 18). The Wajãpi, Parakanã, and Gavião populations presented no evidence of admixture with Africans and Europeans (*SI Appendix*, Figs. S3–S5).

To establish a timeline for these admixture events, we analyzed the decay of linkage disequilibrium between markers with Rolloff (19). Using the Guaraní Mbyá in addition to the Iberian and Yoruba populations from the 1000 Genomes Project as Tupiniquim parental populations, the last intense gene flow between the Native American and the European components was dated to seven generations ago (*SI Appendix*, Fig. S6*A* and Dataset S3), and between the Native American and African components to 5.5 generations ago (*SI Appendix*, Fig. S6*B* and Dataset S3), which is also approximately the same date estimated for the African and European components (*SI Appendix*, Fig. S6*C* and Dataset S3). We also performed an analysis of time and admixture dynamics inferred from TRACTS (20, 21). The results suggest that the admixture process in the Tupiniquim population was complex and continuous, involving two pulses of admixture followed by a continuous migratory flow (*SI Appendix*, Fig. S7 and Dataset S4). The results indicated the beginning of the process of admixture with Europeans ~11.2 generations ago and with Africans ~8.3 generations ago. This initial admixing was followed by a second major pulse that started ~5.2 generations ago and with a continuous flow of Africans and Europeans for subsequent generations. Eleven generations ago (the time of the first European pulse) coincides with the height of the Brazilian Gold Cycle (1690–1750;

ref. 22), a period in which the Brazilian population increased from 300,000 to almost 3 million (1), forcing King João V of Portugal to restrict free Portuguese access to Brazilian lands. The exploitation of gold led to the enslavement of the indigenous people by the Portuguese to work in the mines (23), which decimated a great part of the coastal and central Brazilian native populations. Interestingly, also approximately eight generations ago (1807; time of the first African pulse), the Portuguese Royal family moved the court to Rio de Janeiro to escape the invasion of the Kingdom of Portugal by Napoleon Bonaparte (24), which quickly intensified colonization of the Brazilian coast and promoted rapid population growth. The transfer of the Portuguese Court to Brazil also intensified the slave trade, and between 1806 and 1830 alone, more than 850,000 Africans were forcibly brought to Brazil (1). Approximately five generations ago (1888; the time of the second pulse of admixture with Africans and Europeans), slavery was abolished in Brazil (24), which resulted in increases in the African-derived populations in all regions. During the same period, a massive migration of ∼1.5 million Italians to the Brazilian southeastern region was encouraged by the government (1) to replace slave labor.

We also analyzed the distribution of runs of homozygosity (ROH) in the Tupiniquim in comparison with modern Native Americans (newly genotyped Guaraní Mbyá, Wajãpi, Parakanã, and Gavião; ref. 13), Africans, Europeans, and East Asians (HGDP). The distribution of ROH reflects demographic processes and mating patterns occurring throughout the population's history, since the longer tracts represent recent events, such as inbreeding, while shorter segments were formed by older demographic processes (e.g., bottlenecks and founder effects). Our results (Fig. 2 and *SI Appendix*, Fig. S8) showed that Amazonian and non-Amazonian Native Americans present, on average, larger amounts of short/intermediate ROH (0.5 to 8 Mb), while Tupiniquim exhibits an ROH pattern similar to that observed for Mesoamericans,

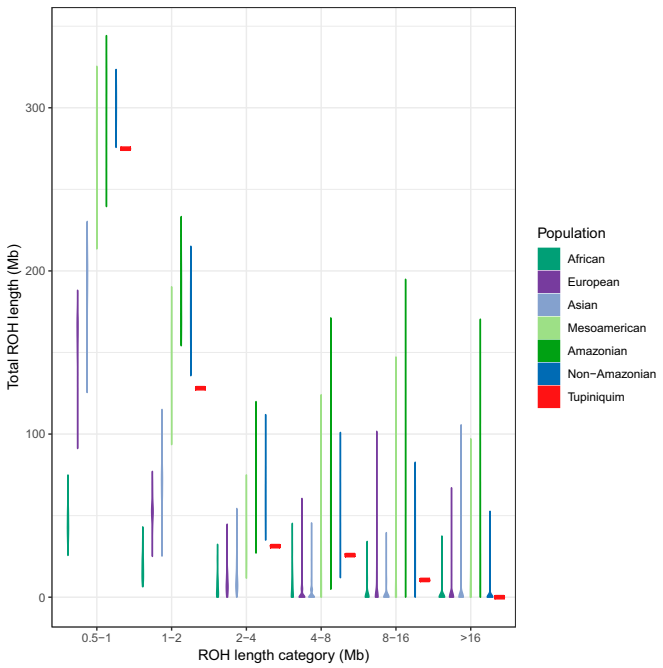


Fig. 2. ROH distribution of Native American, Asian, European, and African individuals. ROH identification was performed using PLINK v1.9 software (35) on a set of 395,840 SNP markers obtained from 609 individuals including the Tupiniquim, other modern Native Americans [newly genotyped and public (8)], and African, European, and Asian populations (Human Genome Diversity Project). The average total ROH lengths obtained per population are presented, binned by the ROH length.

with a higher amount of genetic diversity. This could be a result of the greater effective population size of the Tupiniquim population during the time of the Conquest, estimated at ∼90,000 individuals living along the Brazilian coast (1).

To refine the estimation of the effective population size (N_e), we used a method recently introduced by Browning et al. (25), based on identity by descent (IBD) and local ancestry, to estimate the effective population size histories of the analyzed samples. The ancestry-specific (African, European, and Native American), historical N_e estimates are somewhat similar to and within the variation observed for Caribbean and Central American admixed populations (25), although the bottleneck seems to have been stronger in the Tupiniquim, with a minimum N_e of $\sim 10^2$ (Fig. 3), while the same estimate for most of the former populations is at least an order of magnitude higher. Interestingly, the minimum value of N_e in the Tupiniquim was caused by a bottleneck 7 generations ago, which coincides with the estimated date for the admixture between the Native American and European components (*SI Appendix*, Fig. S6*A* and Dataset S3). There are few historical data available regarding the Tupiniquim census; however, the last count was 55 individuals in 1876 (26), which is ∼7 generations ago, so the genetic data seem to recover the collapse time of this population.

Precontact History. Individuals from the Tupiniquim population showed large variance in proportions of admixture (*SI Appendix*, Fig. S2*C* and Dataset S1), with one of them being almost entirely of Native American descent, whereas others showed less than 5% Native American ancestry. We leveraged this feature in each individual to look in more detail at the Native American ancestry component of the admixed genome and make inferences about deeper timescales in two ways: 1) by performing a Local Ancestry Inference on every Tupiniquim and masking the non-Native American markers as missing data, after which the individuals presented no or negligible evidence of non-Native American ancestry (*SI Appendix*, Fig. S9), and selecting individuals with more than 80% estimated Native American ancestry; and 2) by using exclusively the individual with the highest proportion of Native American ancestry (94.06%) and performing genotyping with high-density SNPs array Axiom Human Origins (19); yielding 1) ∼70,000 and 2) ∼600,000 SNPs overlapping those of the reference panels used (*SI Appendix*).

Principal component analysis (17) of the Native American populations clustered the Tupiniquim with Amazonian Tupí-Guaraní populations (the Parakanã, Urubu Kaapor, and Wajãpi), as well as with Karib speakers (the Apalai and Arara; *SI Appendix*, Fig. S32). However, principal component analysis of the Tupí populations placed the Tupiniquim next to the Parakanã and Urubu Kaapor (*SI Appendix*, Fig. S32), providing evidence for closer genetic relationships between these groups.

We then used the F_3-statistics, as implemented by AdmixTools (19), to investigate the Tupiniquim Native American ancestry component and its relationship with other modern Native American populations. We calculated the F_3-statistics in the form F_3 (Tupiniquim, Y, Z) for every pair (Y and Z) of modern Native American populations, and found no evidence of admixture between Tupiniquim and other Native American populations (*SI Appendix*, Fig. S10 *A* and *B*), as no significant negative F_3 estimate was observed. Furthermore, using Treemix (27), Maximum Likelihood trees were inferred for all Native American, and separately for all Tupí, populations. Then we allowed the algorithm to fit up to 5 gene flow events between branches of the trees (*SI Appendix*); no gene flow was detected from any Native American population toward the Tupiniquim branch (*SI Appendix*, Fig. S11 *A*–*D*).

Considering then the absence of admixture, we investigated the patterns of ancestral allele sharing among these groups. To assess this, the outgroup-F_3 was calculated in the form F_3 (Mbuti Pygmy; Y, Z) for every pair (Y and Z) of modern Native American populations, and the estimated F_3 values were plotted

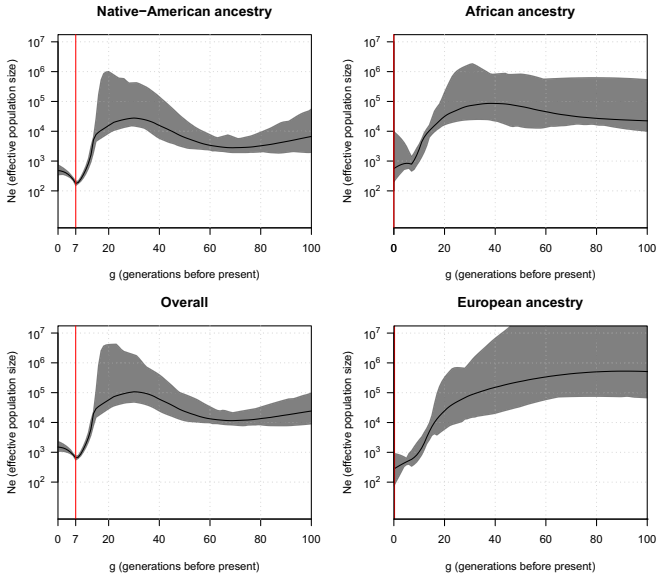


Fig. 3. Ancestry-specific effective population size (Ne) history estimated for the Tupiniquim. Related Tupiniquim samples (k < 0.0625) were removed from the data, which was then phased with Beagle v.5 (32). On the basis of the phased data, IBD segments were estimated with RefinedIBD (33) and Local Ancestry Inference with RFMix (34). Using IBDNe (25) ancestry specific and overall Ne were estimated. The ancestry-specific Ne values are coded in the *y* axis and indicated by the line for each generation before present depicted in the *x* axis. The gray areas show a 95% bootstrap confidence interval. Results for Native American, African, and European ancestry are shown in different panels, along with the Ne estimates obtained using all IBD segments (Overall). The red line indicates the generation with the minimum estimated value of Ne.

as heat map points on the map of the American continent (*SI Appendix*, Fig. S12). The Tupiniquim share more alleles with South Native Americans; however, they display no linguistic, geographic, or any other specific pattern of allele sharing among the latter, and they are not genetically close to the Guaraní Mbyá, who are currently settled near them (*SI Appendix*, Fig. S12*A*). In contrast, Guaraní Mbyá are related to the other Guaraní groups (*SI Appendix*, Fig. S12*B*). This pattern of geographic-genomic relationships was also present in some populations located next to the Madeira-Guaporé region in the Amazon basin (*SI Appendix*, Figs. S13 and S14).

We also performed the outgroup-F_3 analysis to investigate the relationships between modern-day Native American populations and ancient individuals from various periods of time located in current Brazilian territory; namely, Lapa do Santo (9,600 YBP) and 3 Sambaquis (Brazilian coastal and fluvial shell mounds, which are cultural deposits with diverse sizes and stratigraphies, mainly composed of shells: Laranjal [6,700 YBP], Moraes [5,800 YBP], and Jabuticabeira [2,000-2,100 YBP]), along with the Anzick-1 (12,900 to 12,700 YBP; a Clovis Culture-associated sample). In each comparison, ancient Brazilians were more closely related to the modern Brazilian native populations than to the modern Mesoamerican natives (*SI Appendix*, Figs. S15 and S16). More specifically, the most ancient individuals (Lagoa Santa and Laranjal) seemed to be more broadly related to the modern Brazilian natives (*SI Appendix*, Figs. S15 *A* and *B* and S16 *B* and *C*), indicating that the genetic similarity between modern populations and these ancient populations is independent of linguistic affiliation or geographic location. Most interestingly, the same was observed for the Anzick-1 individual (*SI Appendix*, Fig. S16*A*), implying some level of a distinctive contribution of Anzick-1–related lineages to modern South Americans (at least for the Brazilian populations represented here) when compared with Mesoamericans (Maya and Pima). Beginning around 5,800 YBP (Moraes), these patterns of shared ancestry between paleo and modern individuals become progressively more distinctive (i.e., specific to some populations, such as Xavánte [Jê-speaker population] and Arara [Karib-speaker population; *SI Appendix*, Figs. S15 *C* and *D* and S16 *D* and *E*]), replicating the long-standing continuity inside South American regions described by Posth et al. (15). Here we also detected some level of genetic continuity inside South America between the modern and the most ancient South Americans (Lagoa Santa and Laranjal), as well as with the Anzick-1, given that they share significantly more alleles than are shared between modern Mesoamericans and ancient Americans (*SI Appendix*, Figs. S17–S20). These most recent paleo individuals, Moraes and Jabuticabeira, and particularly the latter, presented differential high affinity with some populations (mainly Jê and Karib groups), but showed less close relationships with the Tupí groups (*SI Appendix*, Figs. S15 *C* and *D* and Fig. S16 *D* and *E*).

The scenario that emerges is one of increasing differentiation between Native American populations since the initial settlement of South America. Posth et al. (15) provide evidence for the existence of different migrations to this continent and subsequent replacement of the initial populations to a large extent. We add to this model, specifically in the case of eastern South America, the idea of the effects of demographic movements that occurred after the linguistic split (i.e., Tupí-Jê split), which involved several fission-fusion events (nonrandom migration processes that affect the structure of hunter-gatherer populations; ref. 28 and references therein) that genetically differentiated modern native populations from each other over time (28); this is more pronounced in recent samples (i.e., Jabuticabeira ∼2,000 YBP).

Furthermore, we examined the relationship between ancient Brazilian samples and modern populations to see if we could detect any patterns of specific shared ancestry among them. With this purpose, we calculated F_4 (Mbuti Pygmy, aDNA; Y, Z), with aDNA iterating over all groups of samples according to archeological sites and Y and Z over all modern populations. For comparisons involving Pima and Maya, virtually any modern native South American population exhibited higher levels of allele sharing with all ancient samples, including the Anzick-1 (*SI Appendix*, Figs. S19 and S20). A similar pattern was also present in the outgroup-F3 results (*SI Appendix*, Figs. S15 and S16). In addition, analysis of variance (Dataset S5) and Tukey's honestly significant difference test demonstrated that these differences were significant (*SI Appendix*, Figs. S17 and S18). In general, Xavánte (Jê-speaker population) is the population that shares the most alleles with ancient samples, whereas the Mesoamericans (Maya and Pima), along with Wajãpi, Guaraní Mbyá, and Parakanã, are the populations that share the least alleles with ancient samples (*SI Appendix*, Figs. S19 and S20). This may indicate distinctive demographic processes acting in these southern Native American populations, such as higher genetic drift or a more complex demographic history involving differential gene flow among the populations.

The Tupí Expansion. Our results thus far suggest that the Tupiniquim Native American ancestry is of Tupí origin, and they therefore may be used as proxies for the Tupí populations extinct from the Brazilian coast in investigations into the process of their expansion toward the coast. Thus, we tried to shed light on this question of the Tupí expansion, using two approaches.

First, we used an unsupervised approach, in which no prior expectations about the Tupí population history would be assumed. In this sense, we tried to produce trees depicting the evolutionary relationship between all Tupí populations using three methods: pairwise F_{ST} and pairwise F_2 Neighbor Joining Trees and Treemix (27). Using qpGraph (19), we tested these trees (*SI Appendix*, Figs. S21–S24), and only two of them showed a good fit to the data

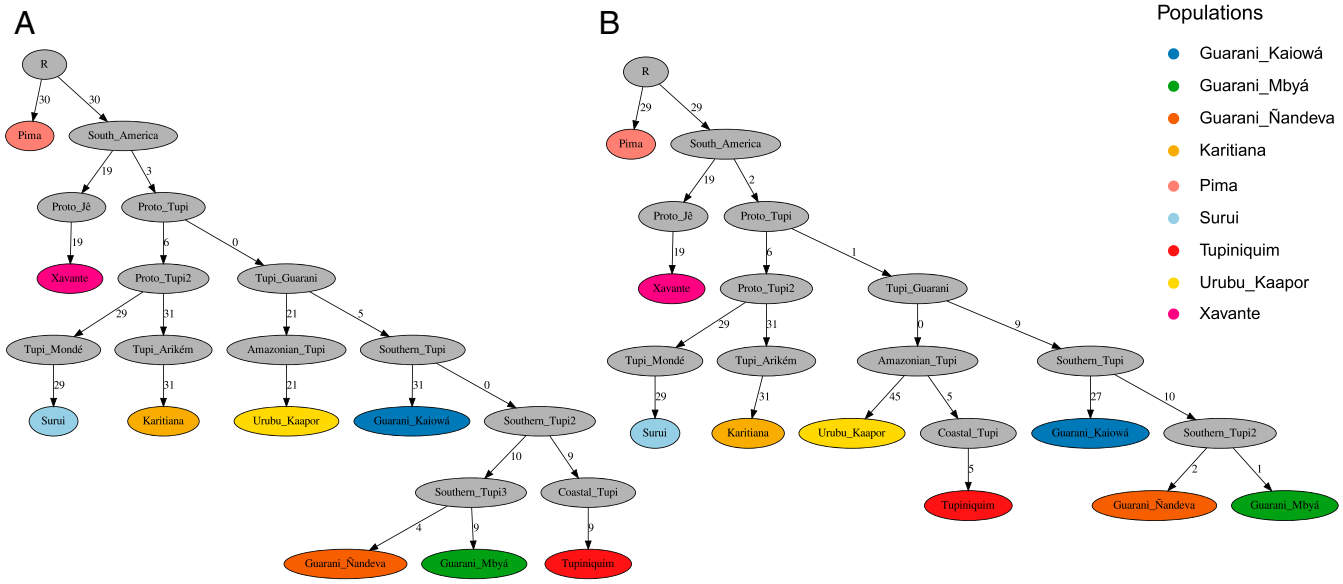


Fig. 4. Modeling Tupí expansion. The two main concurrent Tupí expansion hypotheses were modeled and assessed with qpGraph (19) to test the fit between all expected and observed F-statistics. A good fit to the data are indicated by absence of |Z| > 3 values. (*A*) Example of Tupí expansion hypothesis 1 model, with maximum |Z| equal to 6.087. (*B*) Example of Tupí expansion hypothesis 2 model, where the maximum |Z| is 2.763. Gray circles represent internal nodes of the tree for which there are no data, while colored circles stand for the modern Native American populations, in the same color scheme of other figures. The branch lengths are presented as units of FST, multiplied by 1,000.

(*SI Appendix*, Fig. S24 *B* and *C*). Both trees have very similar topologies, differing only in the relationships inside the Tupí-Guaraní, also presenting very small branches (values of 1 or less; *SI Appendix*, Fig. S24 *B* and *C*) inside this group, excepting the Guaraní populations, which are likely a monophyletic group (also evident in all other produced trees; *SI Appendix*, Figs. S21 and S22). This pattern of star-like radiation of the Tupí-Guaraní suggests that the common ancestor populations had relatively large effective population sizes and/or that the expansion happened in a short time. Thus, a polytomy appears in the root of Tupí-Guaraní populations, obscuring their relationships with one another, and with the Guaraní cluster being the only more easily discernible group. Interestingly, evidence for an excess of Native Mesoamerican-related ancestry was detected in the Guaraní, as a gene flow event from a Mesoamerican population to the Guaraní cluster was inferred with Treemix (27) (*SI Appendix*, Fig. S11 *A* and *C*). This pattern has been previously described between Andean and Native Mexican populations (29), but here we showed evidence of Mesoamerican gene introgression in a lowland population. This could indicate that the barrier between Andeans and non-Andeans was not so strict in the past and that the division observed in modern populations is likely related to both the establishment of agriculture in the highlands and strong drift in the lowlands.

Our second approach was to model population history scenarios (also with qpGraph) (19) to test the two main broad Tupí Expansion hypotheses (2, 8): hypothesis 1, the Tupí reaching the coast through a single expansion direction initially going through the south and then moving upward, deriving from the Guaraní people (9–11) (Fig. 1, blue arrow); and hypothesis 2, in which the Tupí occupied the coast after originally expanding eastward from the mouth of the Amazon and the Guaraní spread from the Amazon to the Paraná basin (12) (Fig. 1, red arrows). Essentially, we tried to differentiate between two hypotheses: 1) that the Tupiniquim would have reached the coast from the south and would, therefore, be genetically closer to the Guaraní populations, as they would share an exclusive most recent common ancestor with them; 2) that they would be more closely related to Tupí-Guaraní populations from the north (e.g., the Parakanã), based on the same rationale. Hence, we used Pima as an outgroup for all South Native Americans, and Xavante (Jê-speaker population) as an outgroup for all Tupi. Several models were produced for each hypothesis, including different sets of populations and switching the Tupiniquim position (*SI Appendix*, Figs. S25–S28). Models of the second hypothesis consistently presented a better fit to the data in comparison with those from the first hypothesis, as inferred with qpGraph (19) (Fig. 4 *A* and *B* and *SI Appendix*, Figs. S25–S28). According to hypothesis 2, the Atlantic coast was peopled by Amazonian Tupí-speakers that probably reached this region ∼2,000 YBP through a route along the northeast coast of Brazil (Fig. 1).

In this context, our results support the notion that expansion was caused by a search for new lands to cultivate by incipient Amazonian agriculturalists. Pottery found in the south Amazon (Guaraní Tradition) and the east Amazon and coast (Tupinambá Tradition) are separated from each other by as much as 4,000 km and present features that reveal the distinct evolution of these two groups after they left the Amazon basin.

Conclusion. Our study rescued part of the Native American history that had been concealed by European colonization. First, we recovered genomes from extinct coastal populations through the genomes of admixed people historically related to the Tupiniquim. Then, using this information of the Coastal Tupí populations, combined with data from natives of other regions, we managed to retrace how the occupation of the Brazilian territory by the Tupí occurred before the arrival of the Europeans. Notably, we reveal how the Atlantic coast was occupied by Amazonian peoples through a migratory wave from the northwest of the Amazon, and we further show that the Guaraní peoples of southern Brazil and Paraguay came from a separate migration but share a common ancestor. We also detected a subsequent migratory wave coming from Mesoamerica that may have influenced the formation of the southern Tupí groups (Guaraní branch). In addition, we found genomic evidence of the collapse of the coastal population, with an extreme bottleneck effect on the admixed Tupiniquim population. Last, when comparing modern and ancient individuals, we see that originating 6,000 y ago, there is genetic structure between populations that is most likely

generated by the strong drift events caused by the language diversification in South America.

Materials and Methods

To investigate the admixture events, we applied Rolloff (19) and TRACTS software (20, 21). Using AdmixTools (19) we computed F_3, outgroup-F_3, D-statistics, and F_4, clustering the Tupiniquim as a population and treating them as separate individuals in the calculation in some of the analyses. For these analyses, datasets v [47 Tupiniquim (masked) + 48 Guaraní Mbyá + 48 Native Americans (13) + 7 newly genotyped Native Americans + HGDP], vi [1 Tupiniquim (ID: 2004) + 4 Guaraní Mbyá (IDs: 3001, 3036, 3038, 3051) + 48 Native Americans (13) + 7 newly genotyped Native Americans + HGDP], ix [47 Tupiniquim (masked) + 48 Guaraní Mbyá +48 Native Americans (13) + 7 newly genotyped Native Americans + HGDP +15 Ancient DNA samples (15)], and x [1 Tupiniquim (ID: 2004) + 4 Guaraní Mbyá (IDs: 3001, 3036, 3038, 3051) + 48 Native Americans (13) + 7 newly genotyped Native Americans + HGDP +15 Ancient DNA samples (15) + Anzick-1 Clovis Culture associated ancient DNA (14); *SI Appendix*, Table S3] were used. We calculated F_{ST} and F_2 for all pairs of populations (*SI Appendix*, Table S3: datasets v and vi), to shed light on the relations between these populations and to pinpoint where the Tupiniquim fit within the Native American groups. Matrices containing pairwise genetic distance values were produced using R scripts (https://github.com/BenjaminPeter/cph_course/blob/master/scripts/analysis.R) and plotted as Neighbor-Joining trees using R packages *ape* and *ggtree* (30, 31) to provide models for the history of population splits between these populations. We also used Treemix (27) to estimate the Maximum Likelihood tree and fit putative admixture events. For a subset of populations that included all Tupí (*SI Appendix*, Table S3: datasets v and vi), we tested the fit between empirical data and the pairwise F_{ST} and F_2 NJ trees, along with the Maximum Likelihood trees produced with Treemix, using AdmixTools (19). Finally, we tried to explicitly model the two main Tupí Expansion hypotheses (2, 8), producing several models for each hypothesis with different populations, repositioning the Tupiniquim in the trees (again using datasets v and vi; *SI Appendix*, Table S3). Model fit was assessed by the differences between estimated and expected F-statistics values. Models with $|Z| < 3$ for all (or almost all) differences were considered to present a good fit to the data.

Ancestry-specific Effective Population Size (N_e) history was reconstructed for both the Tupiniquim and the Guaraní Mbyá (*SI Appendix*, Table S3; datasets vii [47 Tupiniquim (unmasked) + 48 Guaraní Mbyá + Sub-Saharan Africans, Europeans, and East Asians (1000 Genomes Project)] and xi [48 Guaraní Mbyá + Peruvians from Lima, Sub-Saharan Africans and Europeans (1000 Genomes Project)], respectively). First phasing was done with Beagle v.5 (32), and IBD segment estimation with RefinedIBD (33) and Local Ancestry Inference implemented with RFMix (34). Finally, IBDNe (25) was used to estimate ancestry-specific N_e from the estimated IBD segments and the ancestry blocks identified through the Local Ancestry Inference. ROH were identified using PLINK v1.9 (35) with a minimum length of 500 Kb, using a sliding window of 50 SNPs, a maximum gap of 100 Kb between consecutive SNPs, a proportion of 5% overlapping windows forming homozygous segments, and an SNP density of at least one per 50 Kb. A complete description of sampling, genotyping strategies, dataset assembly, quality control procedures, and methods is included in the *SI Appendix*.

Ethical approval for sample collection was provided by the Brazilian National Ethics Commission (CONEP Resolution no. 123 and 4599). CONEP also approved the oral consent procedure and the use of these samples in studies of population history and human evolution. Individual and/or tribal informed oral consents were obtained from participants who were not able to read or write. All sampling was coordinated by coauthors of this study (F.M.S. and J.G.M.) and their collaborators, in a manner consistent with the Helsinki Declaration and Brazilian laws and regulations applicable at the time of sampling. Logistical support for the sample collection was provided by the Fundação Nacional do Índio. The results of this study were discussed with the participating communities. A description of the sampling and genotyping strategies, along with the dataset assembly and quality control procedures is included in the *SI Appendix*.

Our dataset has been deposited at the European Genome-phenome Archive, which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAS00001004036. The informed consent associated with these samples is restricted to population history/evolutionary analyses. The data will be available to researchers who sign the Data Access Agreement with the Data Access Committee on the European Genome-phenome Archive website.

ACKNOWLEDGMENTS. We thank Rui Sérgio Sereni Murrieta and André Menezes Strauss for their helpful comments on the historical and archeological data. We are also grateful to Regina Cália Mingroni Netto and Lilian Kimura for laboratory assistance and technical support. Finally, we would like to thank all the native communities who participated in the study without whom this work would not have been possible. M.A.C.e.S was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2018/013716; 2015/26875-9) and K.N was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (PNPD/1645581); NIH (R01 GM075091).

1. IBGE, Instituto Brasileiro de Geografia e Estatística. https://brasil500anos.ibge.gov.br/estatisticas-do-povoamento.html. Accessed 1 December 2019.
2. M. C. da Cunha, *História dos índios no Brasil* (Editora Companhia das Letras, 1992).
3. V. Ramallo *et al*., Demographic expansions in South America: Enlightening a complex scenario with genetic and linguistic data. *Am. J. Phys. Anthropol.* **150**, 453–463 (2013).
4. R. S. Walker, S. Wichmann, T. Mailund, C. J. Atkisson, Cultural phylogenetics of the Tupi language family in lowland South America. *PLoS One* **7**, e35025 (2012).
5. A. V. Galucio *et al*., Genealogical relations and lexical distances within the Tupian linguistic family. *Bol. Mus. Para. Emílio Goeldi Ciênc. Hum.* **10**, 229–274 (2015).
6. E. J. M. dos Santos, A. L. S. da Silva, P. D. Ewerton, L. Y. Takeshita, M. H. T. Maia, Origins and demographic dynamics of Tupí expansion: A genetic tale. *Bol. Mus. Para. Emílio Goeldi Ciênc. Hum.* **10**, 217–228 (2015).
7. H. Silverman, W. Isbell, *Handbook of South American Archaeology* (Springer Science & Business Media, 2008).
8. F. S. Noelli, "The Tupi expansion" in *The Handbook of South American Archaeology* (Springer, 2008), 659–670.
9. A. Métraux, Migrations historiques des Tupi-Guarani. *J. Soc. Am.* **19**, 1–45 (1927).
10. B. J. Meggers, C. Evans, A reconstituição da pré-história amazônica: Algumas considerações teóricas (1973).
11. B. J. Meggers, C. Evans, "Lowland South America and the Antilles" in *Ancient Native Americans*, J. D. Jennings, ed. (W. H. Freeman and Company, CA: San Francisco, 1978), pp 543–591.
12. J. P. Brochado, "An ecological model of the spread of pottery and agriculture into Eastern South America," PhD dissertation, University of Illinois at Urbana-Champaign (1984).
13. P. Skoglund *et al*., Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
14. M. Rasmussen *et al*., The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
15. C. Posth *et al*., Reconstructing the deep population history of central and south America. *Cell* **175**, 1185–1197.e22 (2018).
16. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
17. A. Ruiz-Linares *et al*., Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* **10**, e1004572 (2014).
18. F. S. G. Kehdy *et al*.; Brazilian EPIGEN Project Consortium, Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8696–8701 (2015).
19. N. Patterson *et al*., Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
20. S. Gravel, Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
21. S. Gravel *et al*.; 1000 Genomes Project, Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet.* **9**, e1004023 (2013).
22. T. E. Skidmore, Brazil: Five Centuries of Change. *OUP Catalogue* (2009). https://ideas.repec.org/b/oxp/obooks/9780195374551.html. Accessed 9 August 2019.
23. H. Langfur, The return of the bandeira: Economic calamity, historical memory, and armed expeditions to the sertao in minas gerais, Brazil, 1750-1808. *Americas* **61**, 429–461 (2005).
24. H. M. Starling, L. M. Schwarcz, *Brazil: A Biography* (Penguin, UK, 2018).
25. S. R. Browning *et al*., Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* **14**, e1007385 (2018).
26. B. Ricardo, F. Ricardo, *Povos indígenas no Brasil, 2011/2016* (Instituto Socioambiental, 2017).
27. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
28. F. M. Salzano, The fission-fusion concept. *Curr. Anthropol.* **50**, 959 (2009).
29. G. A. Gnecchi-Ruscone *et al*., Dissecting the pre-columbian genomic ancestry of native Americans along the andes-amazonia divide. *Mol. Biol. Evol.* **36**, 1254–1269 (2019).
30. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
31. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: Anrpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
32. B. L. Browning, Y. Zhou, S. R. Browning, A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
33. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
34. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
35. C. C. Chang *et al*., Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

DISCUSSÃO

Desafios e abordagens para o estudo da genética de populações nativas americanas

A ampla miscigenação iniciada a partir da chegada dos europeus nas Américas e que transcorreu durante e após o processo de colonização é uma questão que sempre deve ser avaliada no estudo da genética de populações americanas. Isto porque, a miscigenação é um fator que influencia em maior ou menor grau quaisquer inferências sobre a ancestralidade e história populacional de grupos indígenas e não-indígenas das Américas, sendo que a intensidade dessa influência varia de acordo com uma gama de fatores socioculturais e históricos que determinaram a amplitude e frequência desses eventos de miscigenação em cada região e população (e.g. (RUIZ-LINARES et al., 2014; ADHIKARI et al., 2016, 2017; CHACÓN-DUQUE et al., 2018)), o que também foi inferido e constatado nas amostras dos grupos indígenas aqui analisadas. Uma exceção são os grupos amazônicos que se apresentam significativamente menos miscigenados que o restante e de fato, na maior parte, não apresenta sinais de miscigenação (Figura suplementar 1 e conjunto de dados suplementares 1 do Capítulo 2; Figura suplementar 2 e conjunto de dados suplementares 1 do Capítulo 3). Por esse motivo seu efeito sobre os métodos de inferência genéticos deve ser sempre levado em consideração e, se possível, identificado e controlado através de alguma metodologia (BOLNICK et al., 2016).

Concomitante ao processo de miscigenação, a chegada dos europeus às Américas também alterou a paisagem de diversidade genética e cultural de uma outra forma, ainda mais drástica, através do extermínio massivo e em alguns casos completo de povos nativos americanos (THORNTON, 1987; STANNARD, 1993; MONTENEGRO; STEPHENS, 2006; UBELAKER, 2006). Em especial, algumas regiões como o leste da América do Sul foram mais afetadas por esse declínio populacional, conforme demonstrado em nossas análises sobre o histórico de tamanho efetivo populacional (Figura 7A e figura suplementar 18 do Capítulo 2). Soma-se a isto o fato de que mesmo no período pré-contato populações destas regiões já apresentavam níveis baixos de diversidade genética em comparação com outras populações do mundo, em grande parte devido às dinâmicas e ao histórico de povoamento das Américas (WANG et al., 2007; REICH et al., 2012), além do fato de serem grupos endogâmicos, de modo que os os menores níveis de diversidade genética são hoje encontrados nesta porção do continente, sobretudo na Amazônia (1000 GENOMES PROJECT CONSORTIUM et al., 2015; CEBALLOS et al., 2018; BERGSTRÖM et al., 2020), como claramente evidenciado pelas análises da diversidade genética da América do Sul apresentados no Capítulo 2, onde encontramos uma

correlação entre a longitude e o coeficiente de endogamia, o qual aumenta do oeste para o leste. Desta forma, os padrões de estrutura e afinidade genética subjacentes aos povos nativos americanos, os quais pode-se inferir que eram mais sutis devido à baixa diversidade genética, foram ainda mais ofuscados pelos eventos desencadeados após o contato com europeus.

Adicionalmente, grande parte das metodologias desenvolvidas para o estudo genômico da espécie humana encontra alguns obstáculos para a aplicação em populações indígenas e miscigenadas. Isto ocorre porque estes métodos foram predominantemente desenvolvidos tendo como base bancos de dados majoritariamente compostos por populações europeias, as quais possuem polimorfismos[51], frequências alélicas[52] e padrões de desequilíbrio de ligação[53] distintos das demais populações (PATTERSON et al., 2012; 1000 GENOMES PROJECT CONSORTIUM et al., 2015; BERGSTRÖM et al., 2020). Além disso, a própria diversidade genética de populações indígenas foi consideravelmente menos estudada através do sequenciamento completo de genomas, de modo que certamente uma parte significativa dos polimorfismos exclusivos de nativos americanos ainda não foi identificada (BOLNICK et al., 2016). Um exemplo mais extremo desse problema são os próprios arranjos de SNPs[54] comerciais, os quais de modo geral são desenhados com base em polimorfismos identificados em populações europeias e asiáticas e portanto não necessariamente informativos ou mesmo presentes em populações de outras regiões do mundo (PATTERSON et al., 2012). Como consequência, tais arranjos podem não detectar e subestimar a diversidade genética presente em populações não-europeias, resultando assim no viés de determinação[55], o que interfere e prejudica a aplicação de diferentes tipos de análises, como por exemplo inferências sobre a estrutura genética de uma população (PATTERSON et al., 2012).

Além disso, historicamente as populações indígenas e miscigenadas das Américas têm sido menos estudadas do ponto de vista genético[56], levando a uma escassez relativa desses estudos e a uma subamostragem de populações indígenas contemporâneas, principalmente de comunidades indígenas dos Estados Unidos[57], dificultando o estudo da história populacional e

[51] Um *locus* (i.e. posição do genoma) é dito polimórfico quando apresenta mais de uma forma possível (i.e. mais de um alelo). Entretanto, o termo polimorfismo é tradicionalmente usado apenas para *loci* com pelo menos dois alelos com mais de 1% de frequência na população.

[52] Frequência dos alelos de um *locus* em uma população.

[53] Se refere ao comprimento do segmento cromossômico herdado sem recombinação de um ancestral comum, o que causa uma associação não aleatória de alelos em diferentes *loci* de uma dada população.

[54] *Single-nucleotide polymorphism*, são polimorfismos envolvendo um único nucleotídeo (e.g. transição de Adenina [A] para Guanina [G]).

[55] *Ascertainment bias*.

[56] Apesar dos enormes esforços de pioneiros da genética de populações nativas americanas como o geneticista Francisco Mauro Salzano, professor emérito da Universidade Federal do Rio Grande do Sul.

[57] Um histórico de problemas éticos da pesquisa genômica nos Estados Unidos gerou uma grande desconfiança por parte das comunidades indígenas locais, as quais por esse motivo se negam ou se mostram bastante receosas em participar de estudos genéticos (BOLNICK et al., 2016).

evolutiva dos grupos humanos desta região (POPEJOY; FULLERTON, 2016; HINDORFF et al., 2018; MILLS; RAHAL, 2019). Dessa forma, o estudo do povoamento das Américas até pouco antes da última década estava largamente restrito somente à análise de evidências arqueológicas, linguísticas e de marcadores uniparentais em populações humanas contemporâneas, o que reduzia consideravelmente o poder de elucidar em maior resolução alguns aspectos e dinâmicas destes eventos do passado (BOLNICK et al., 2016; BISSO-MACHADO; FAGUNDES, 2019). Apenas recentemente o desenvolvimento contínuo de plataformas de sequenciamento de nova geração e arranjos de SNPs permitiram o estudo de centenas de milhares a milhões de marcadores moleculares a um custo acessível, promovendo uma maior democratização ao acesso à tecnologia, a qual possibilita a detecção de padrões de estrutura e compartilhamento genético muito mais sutis do que os marcadores uniparentais (SKOGLUND; REICH, 2016).

Sobre esse ponto de vista, até recentemente, mesmo os estudos genéticos de marcadores autossômicos se focaram predominantemente em utilizar métodos de inferência de ancestralidade global de modo a estimar as proporções de miscigenação e descrever as populações americanas em termos de seus principais componentes de ancestralidade continental (RUIZ-LINARES et al., 2014; SALZANO; SANS, 2014). Entretanto nos últimos anos, o desenvolvimento de metodologias de análise que exploram não apenas as frequências alélicas de forma independente, mas também os padrões de desequilíbrio de ligação, somados à crescente disponibilidades de dados genômicos de alta densidade, obtidos por meio de genotipagens em arranjos de SNPs (polimorfismo de base única) ou sequenciamentos de genoma completo, vem inclusive convertendo o estudo do genoma de indivíduos miscigenados num ativo valioso na análise da história e evolução de populações. Nesse sentido vários estudos se voltaram a uma análise conjunta e aprofundada de populações indígenas e miscigenadas, onde se demonstrou que as populações miscigenadas atuaram como repositórios de linhagens genéticas indígenas (KEHDY et al., 2015; MONTINARO et al., 2015; SKOGLUND et al., 2015; BROWNING et al., 2018; CHACÓN-DUQUE et al., 2018; BARBIERI et al., 2019; GNECCHI-RUSCONE et al., 2019a; GOUVEIA et al., 2019; MAS-SANDOVAL et al., 2019; ONGARO et al., 2019), o que também foi corroborado pelos resultados aqui apresentados, sobretudo pela análise dos Tupiniquim e Guaraní Mbyá apresentada no Capítulo 3.

Ainda nesse ponto de vista, as dificuldades no estudo das populações indígenas em particular vêm sendo também aos poucos superadas, primeiro como consequência de um crescente interesse pelo estudo genômico da história populacional, demográfica e evolutiva das populações indígenas americanas (e.g. (REICH et al., 2012; RAGHAVAN et al., 2015; SKOGLUND et al., 2015; MORENO-MAYAR et al., 2018b; POSTH et al., 2018; SCHEIB et al., 2018)), segundo pelo desenvolvimento e fortalecimento de grupos de pesquisa locais, como o nosso próprio

grupo do Laboratório de Genômica Populacional Humana (LGPH/USP), os quais têm diversificado o escopo das questões investigadas, sejam elas de cunho histórico, antropológico ou demográfico, focando por vezes também em contextos mais específicos e locais (e.g. (RAFF; TACKNEY; O'ROURKE, 2010; HÜNEMEIER et al., 2012a, 2012b; MORENO-ESTRADA et al., 2013, 2014; RAMALLO et al., 2013; BARBIERI et al., 2014a; FAGUNDES et al., 2018)).

Como discutido, em geral arranjos de SNPs comerciais estão sujeitos ao viés de determinação, contudo de modo a lidar especificamente com este problema foi desenvolvido o arranjo *Axiom Human Origins* (PATTERSON et al., 2012), o qual possibilita a condução de estudos de história populacional com maior resolução em diferentes grupos humanos, ao mesmo tempo em que minimiza o efeito do viés de determinação. Isso é possível porque esse arranjo utiliza uma série de painéis de SNPs identificados pelo sequenciamento de 12 indivíduos de ancestralidade conhecida do HGDP-CEPH (*Human Genome Diversity Panel - Centre d'Etude du Polymorphisme Humain*) provenientes de diversas regiões do globo, incluindo um indivíduo da etnia Karitiana da região amazônica brasileira, desta forma garantindo que polimorfismos de populações de diferentes regiões do globo estão representados no painel. As amostras utilizadas para as análises desta tese foram em sua maioria genotipadas nesse arranjo ou provenientes do sequenciamento completo do genoma, com exceção de alguns dados públicos de amostras de aDNA[58], os quais foram genotipados por meio de uma técnica que consiste no enriquecimento de um painel de SNPs sobre fragmentos amplificados de aDNA[59], além de amostras dos Tupiniquim e Guaraní Mbyá do Capítulo 3 que foram genotipadas no arranjo *Axiom InCor BB*, com ~70.000 SNPs compartilhados com o arranjo *Axiom Human Origins*.

Nos últimos anos uma série de avanços nas estratégias de genotipagem e seleção de marcadores, o desenvolvimento de novas tecnologias de sequenciamento e principalmente o advento das tecnologias de sequenciamento de DNA de alta produtividade reduziram sistematicamente o custo do sequenciamento por genoma, o que por sua vez está lentamente democratizando o acesso a este tipo de tecnologia por parte de grupos de pesquisa fora da Europa e dos Estados Unidos. Adicionalmente, a identificação das melhores fontes e a evolução dos métodos de extração, purificação e sequenciamento (ou genotipagem) de aDNA são desenvolvimentos científicos extremamente disruptivos em termos do conhecimento gerado e vem viabilizando uma produção crescente de genomas com boa qualidade e cobertura provenientes de indivíduos antigos, incluindo os nativos americanos (WILLERSLEV; MELTZER, 2021). Soma-se a isso um esforço de amostragem tanto de populações viventes, quanto de indivíduos antigos, de modo que nos últimos anos a representatividade dos diferentes grupos

[58] DNA antigo.
[59] *1240k SNP capture (FU et al., 2015)*.

americanos tem sido otimizada, reduzindo-se levemente o hiato em relação às populações europeias (SKOGLUND; REICH, 2016; WILLERSLEV; MELTZER, 2021).

Com efeito, o primeiro estudo publicado sobre DNA antigo humano do continente americano ainda é muito recente e data apenas de 2010, o qual trata de um indivíduo Paleoesquimó (Saqqaq) da Groenlândia (RASMUSSEN et al., 2010), seguido pelo genoma do Anzick-1 encontrado em Montana nos Estados Unidos e publicado em 2014 (RASMUSSEN et al., 2014). Desde então uma quantidade crescente de trabalhos com DNA antigo de nativos americanos têm sido publicados (e.g. (RAGHAVAN et al., 2015; RASMUSSEN et al., 2015), culminando em 2018 com a publicação de três trabalhos fundamentais que produziram dados genômicos para 49 (POSTH et al., 2018), 15 (MORENO-MAYAR et al., 2018b) e 91 amostras (SCHEIB et al., 2018) de aDNA, respectivamente. Essas publicações representam uma revolução na capacidade de investigação acerca da história dos povos Nativos Americanos, a qual tem crescido significativamente e vem possibilitando um progresso na elucidação de processos históricos e evolutivos macro e micro-regionais do continente americano (WILLERSLEV; MELTZER, 2021).

Por esse motivo, recentemente a análise genética de indivíduos antigos passou a ser o foco dos estudos da história evolutiva e das migrações humanas. Entretanto, mesmo com o aumento relativo da disponibilidade de dados, o estudo genômico da história populacional dos indígenas americanos, ainda precisa de outros métodos capazes de acessar a informação sobretudo do Holoceno tardio e do período mais próximo ao contato com europeus, principalmente devido à escassez relativa de esqueletos humanos preservados e disponíveis para serem analisados em regiões tropicais do globo[60]. Desta forma, o estudo de populações humanas viventes também nos permite recuperar diversos aspectos do passado através da informação armazenada nos seus genomas, especialmente quando combinados aos dados de indivíduos antigos, de modo a compor um panorama espaço-temporal mais completo das dinâmicas e história populacionais. Como discutido anteriormente é possível analisar a informação genômica de populações viventes indígenas e não-indígenas para resgatar as linhagens genéticas nativas americanas, inclusive em escalas de tempo mais profundas, para tanto o primeiro passo consiste em selecionar apenas esse conjunto de linhagens de interesse e isolá-las daquelas provenientes de outras ancestralidades (SCHRAIBER; AKEY, 2015).

Com este objetivo algumas metodologias podem ser utilizadas, dentre as quais duas abordagens foram aplicadas às amostras e aos dados desta tese, a saber: (i) a primeira consiste em estimar as proporções de ancestralidade global dos indivíduos e selecionar apenas aqueles indivíduos sem sinais, ou com sinais negligenciáveis de miscigenação (i.e. inferência de

[60] Devido a dificuldade de preservação de materiais biológicos em regiões de clima úmido e quente.

ancestralidade não-nativa americana); (ii) a segunda compreende a realização de uma inferência de ancestralidade local[61], de modo a identificar e selecionar apenas os segmentos genômicos de ancestralidade nativa americana (i.e. segmentos genômicos com ancestrais comuns mais recentes compartilhados com grupos indígenas) ou definir como dados faltantes todas as posições genômicas para as quais foi inferida uma ancestralidade não-nativa americana[62]. Ambas abordagens foram utilizadas em conjunto ou separadamente nas etapas de filtragem e controle de qualidade dos dados utilizados nesta tese, com algumas variações dos critérios de seleção de acordo com os objetivos específicos de cada análise, de toda forma pretendendo minimizar os efeitos da miscigenação nos resultados obtidos e ao mesmo tempo maximizar a informação disponível, sempre buscando conservar o maior número possível de amostras e marcadores à disposição. Além disso, também foram removidos os indivíduos aparentados dos conjuntos de dados, de modo a reduzir possíveis vieses, através do cálculo do coeficiente de parentesco entre pares de indivíduos e seleção do maior conjunto possível de indivíduos não aparentados, por meio de uma função desenvolvida (na linguagem de programação R) para este propósito (Capítulo 3) ou com o uso do programa PRIMUS (STAPLES; NICKERSON; BELOW, 2013) (Capítulos 1 e 2).

Adicionalmente, de acordo com o que foi aqui exposto, esta tese cumpre antes de tudo um papel importante de aumentar a representatividade das populações indígenas brasileiras em estudos sobre a diversidade genômica humana atual. Esse padrão fica claro quando observamos a distribuição das amostras de populações contemporâneas publicadas e incluídas no banco *Allen Ancient DNA Resource* (AADR)[63] (círculos coloridos na Figura 1A). É possível observar que as amostras se concentram predominantemente na Europa e Ásia, seguidas por números bem menores na África, nas Américas e por fim na Oceania, estas três últimas apresentando elevada sub-representação e uma amostragem bastante desigualmente distribuída (Figura 1A). Nota-se que a maioria das amostras genotipadas/sequenciadas tanto no período de 2012 a 2014, e de 2016 a 2020 se concentravam predominantemente na África, Ásia e principalmente na Europa (Figura 1A-B). Em 2015 foi publicado o projeto 1000 Genomas da fase 3 (1000 GENOMES PROJECT CONSORTIUM et al., 2015), sendo que os dados de 2015 da Figura 1 são em sua maioria desse projeto, o que reduziu em parte a assimetria de amostragem que existe entre os continentes. Entretanto os dados do continente americano incluídos nesse projeto são provenientes de populações miscigenadas, desta forma não contribuindo para

[61] *Local ancestry inference* ou *local ancestry deconvolution*.

[62] *Local ancestry masking*.

[63] É um banco de dados público, curado e organizado pelo grupo do pesquisador David Reich da *Harvard Medical School* e que conta com 11.686 indivíduos antigos e contemporâneos (no conjunto 1240K+HO) em sua versão 44.3 de 20 de Janeiro de 2021 (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data).

aumentar a representatividade dos povos nativos americanos. Os dados inéditos publicados no âmbito desta tese (Capítulos 1 e 3) em conjunto com os dados de outras publicações que analisaram populações nativas americanas e também incluídos nas análises desta tese (ver quadrados coloridos na Figura 1A) vem dessa forma contribuir para sanar, ao menos em parte, essa lacuna que havia não apenas na amostragem mas principalmente na compreensão da variabilidade genética de populações indígenas sul americanas e brasileiras. Além de lançar bases para que mais estudos sejam realizados dentro da perspectiva genômica sobre a história populacional, demográfica e evolutiva dessas populações.

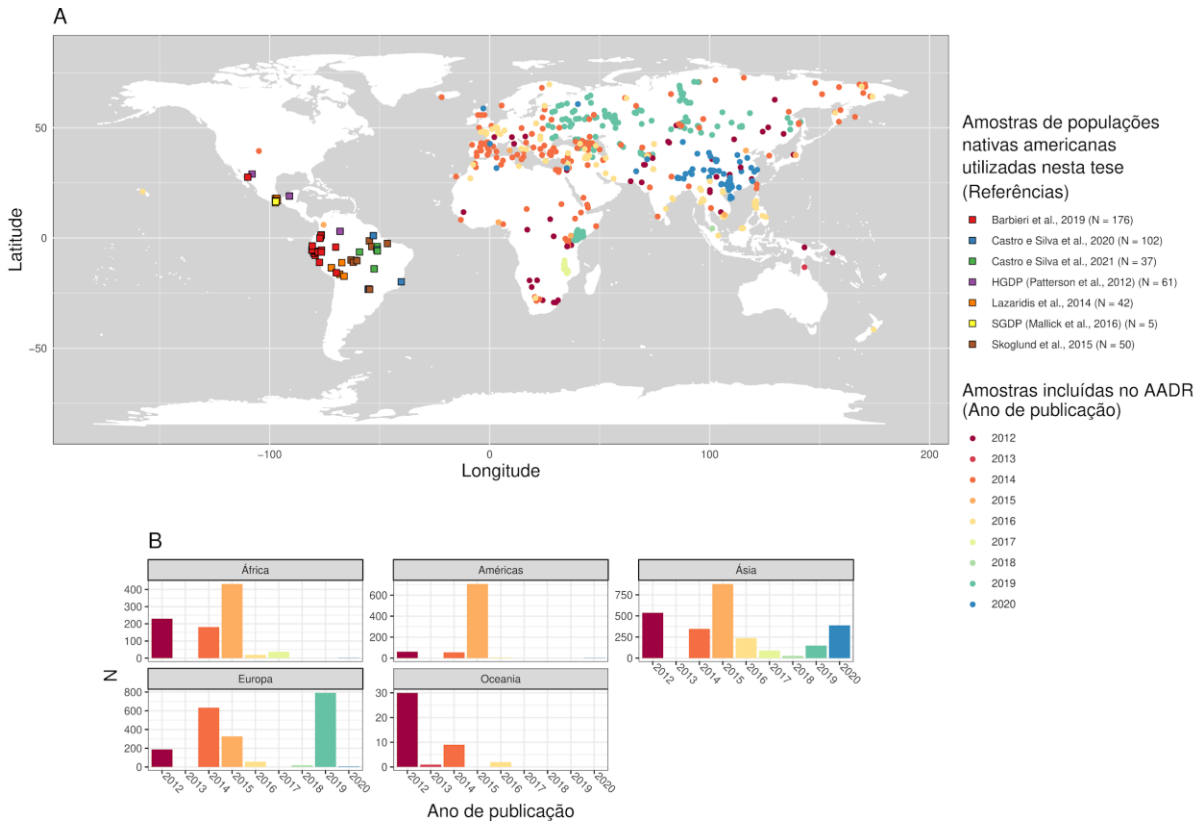


Figura 1 - Sub-representatividade das populações indígenas americanas nos painéis globais de diversidade genômica. Distribuição geográfica **(A)** e contagem do número de amostras por ano de publicação **(B)**, de um total de 6458 indivíduos contemporâneos incluídos no conjunto de dados do *Allen Ancient DNA Resource* (AADR), em sua versão 44.3 de 20 de Janeiro de 2021 (https://reichdata.hms.harvard.edu/pub/datasets/amh_repo/curated_releases/). Adicionalmente, são plotados apenas no mapa **(A)** os dados de grupos nativos americanos utilizados nesta tese: (i) tanto aqueles gerados e publicados nos artigos que compõem esta tese; (ii) quanto os públicos, alguns dos quais não incluídos no AADR. As amostras que compõem o AADR e as amostras de nativos americanos usadas nesta tese são indicadas por pontos em formato de círculo e quadrado, respectivamente, coloridos de acordo com a legenda ao lado direito dos gráficos. O número de amostras oriundos de cada artigo é mostrado entre parênteses à direita das referências, totalizando 383 amostras genotipadas na plataforma *Axiom Human Origins* e 95 na *Axiom InCor BB*. As escalas do eixo Y dos histogramas variam para facilitar a visualização. Algumas amostras não possuem metadados sobre as coordenadas geográficas e portanto não aparecem no mapa, mas são contabilizadas nos histogramas.

Impactos do contato com os europeus e da colonização

Primeiramente, é importante destacar que modelos demográficos e de histórias populacionais, como os apresentados nesta tese, são por definição representações simplificadas da realidade e portanto não são capazes de abranger toda a diversidade de evidências sobre o passado, sejam elas de origem arqueológica, histórica, etnográfica, linguística, entre outras disponíveis, tampouco são capazes de refletir a riqueza de histórias e tradições orais dos povos nativos americanos. Nesse sentido, hoje há um amplo consenso de que estudos multidisciplinares que integram os conhecimentos das diversas áreas são imprescindíveis para uma compreensão correta e mais completa do passado. Com essas ressalvas, passamos a discussão dos resultados.

Nos Capítulos 1 a 3 desta tese analisamos dezenas de grupos nativos americanos distribuídos por toda a América do Sul, desde a costa do Pacífico até a costa do Atlântico, além de algumas populações do norte e sul do México (pontos em formato de quadrado Figura 1) e a primeira vista o padrão mais evidente encontrado é o do impacto causado pelo contato com europeus e pelo processo de colonização. Como discutido, a probabilidade de miscigenação e consequentemente as estimativas de ancestralidade de origem europeia e africana nas comunidades indígenas variam bastante de acordo com uma série de fatores históricos e culturais desde a distribuição e densidade pré-contato dos grupos indígenas, até ao processo de colonização de cada região determinado pela disponibilidade de recursos e condições ambientais específicos e ainda pelo emprego ou não de mão-de-obra escrava (ADHIKARI et al., 2017), assim como pela própria cultura dos colonizadores que resultaram em maior segregação no caso dos britânicos ou maior miscigenação no caso dos ibéricos (MONTINARO et al., 2015; ADHIKARI et al., 2017; ONGARO et al., 2019).

No Capítulo 2, a proporção máxima e mínima de contribuição europeia para as comunidades indígenas é respectivamente de 24,95% no grupo do sul da província peruana de Utcubamba e menor que 1% em 22 grupos do total de 58, com média de 6,21% (Figura 2; Conjunto de dados suplementares 1 do Capítulo 2). Já as contribuições africanas inferidas são muito menores, com proporção máxima de 11,63% da comunidade de Tumbes no extremo norte do litoral pacífico peruano, média de 1,15% e com contribuição menor do que 1% em 41 grupos dos 58. Desta forma, 20 comunidades não apresentam ou apresentam sinal insignificante de miscigenação (i.e. ancestralidade nativa americana inferida é superior a 99%), as quais estão majoritariamente localizadas na Amazônia, corroborando a hipótese de que essa região atuou como um refúgio, ao menos parcial, para os efeitos da colonização europeia. Apesar dos níveis diferenciais de miscigenação observados não foi identificada nenhuma correlação

entre as coordenadas geográficas (i.e. latitude e longitude) e as proporções de ancestralidade nativa americana, europeia e africana inferidas (Figura suplementar 2 do Capítulo 2). Enquanto que nos dados do litoral atlântico brasileiro das comunidades Tupiniquim e Guaraní-Mbyá[64] estudados no Capítulo 3, a proporção de miscigenação inferida é muito maior. Os Tupiniquim apresentam uma proporção média de ancestralidade europeia de 25,9% (4,08%-82,39%) e africana de 22,54% (1,86%-78,01%), enquanto que os Mbyá apresentam ancestralidades europeia e africana média de 15,62% (0-84,82%) e 7,1% (0-40,53%), respectivamente (Figura 2; Figura suplementar 2 e conjunto de dados suplementares 1 do Capítulo 3).

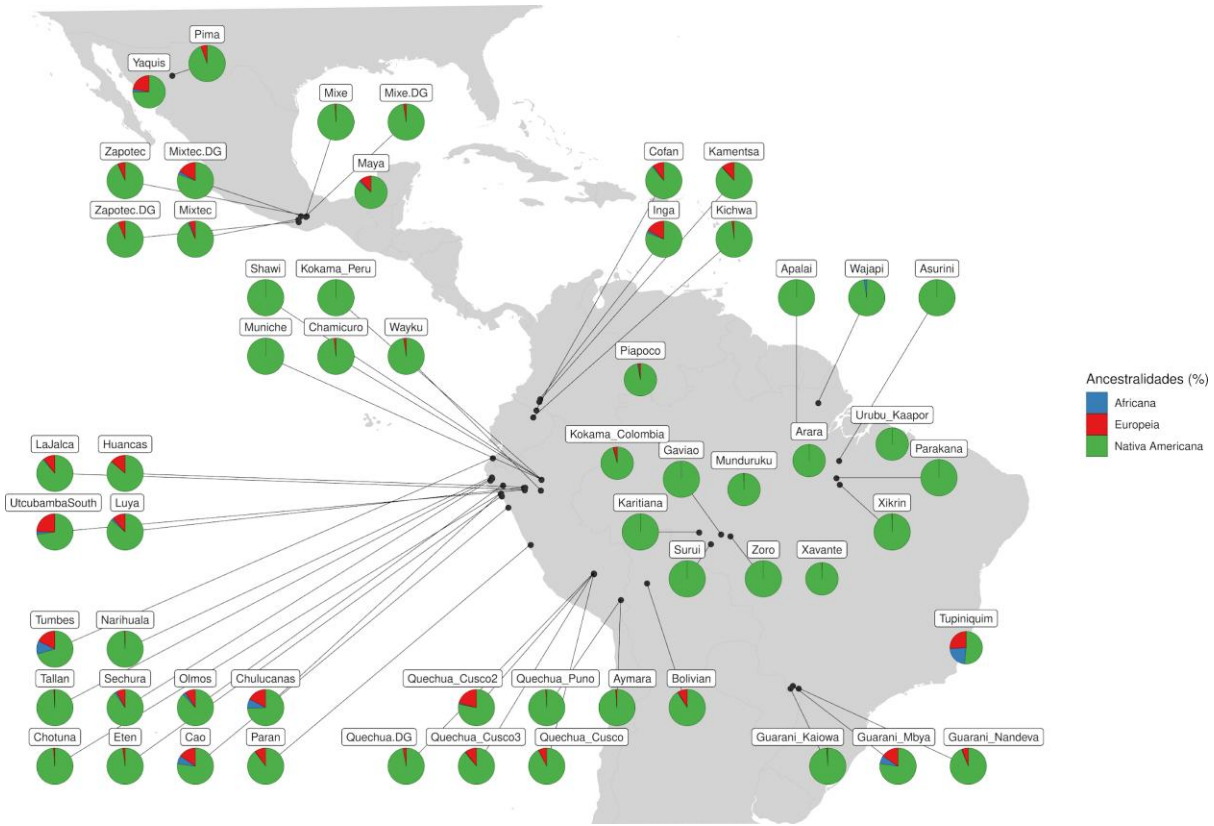


Figura 2 - Distribuição geográfica dos componentes continentais de ancestralidade. Os perfis de ancestralidade global (africana, europeia e nativa americana) obtidos para os Tupiniquim e Guaraní Mbyá no Capítulo 3 (Figura suplementar 1) e para o restante das comunidades indígenas no Capítulo 2 (Figura suplementar 2) foram plotados em suas localizações geográficas aproximadas.

Efetivamente, a observação de maiores proporções de miscigenação no litoral brasileiro se alinha às expectativas do registro histórico, visto que o processo de colonização impactou com intensidade muito maior o litoral e regiões adjacentes do que regiões mais ao interior do continente e da Amazônia, pois o litoral foi historicamente mais densamente povoado por

[64] É importante notar que os Guaraní Mbyá foram reassentados na década de 1960 para a sua localização atual em Aracruz no Espírito Santo, mas são representados nos mapas dessa tese na localização aproximada de seus territórios originários, próximo aos outros Guaraní em Mato Grosso do Sul, como discutido no material suplementar do Capítulo 3.

europeus e consequentemente também o destino final da maioria dos africanos escravizados, um padrão de ocupação do território que se mantém até os dias atuais (SKIDMORE, 2009; SCHWARCZ; STARLING, 2015). Nossas análises indicam ainda que houve três períodos de intensificação da miscigenação os quais coincidem com eventos e processos históricos que levaram à ampliação e interiorização da colonização, além do adensamento populacional. São eles o Ciclo do Ouro, a vinda da corte portuguesa para o Brasil e a abolição da escravatura[65] (Figura suplementar 7 e conjunto de dados suplementares 4 do Capítulo 3), evidenciando assim o quanto esses eventos foram disruptivos na história das populações indígenas e não indígenas do Brasil e consequentemente na composição genética contemporânea das mesmas, certamente indicando o impacto que eventos similares tiveram em outras regiões das Américas.

As consequências da colonização também podem ser observadas nas inferências do histórico do tamanho efetivo (Ne) dos grupos nativos americanos (Figura 7A e figura suplementar 18 do Capítulo 2). No período pós-contato, iniciado a aproximadamente a 18-20 GP[66] (~1500 EC), é observada uma drástica diminuição populacional, com redução média do Ne no período de 90,38%[67] (ou de 90,08% após a remoção dos indivíduos aparentados)[68]. As maiores reduções do Ne são inferidas para os grupos do México, com declínio de 99,49% (99,11%) nos Uto-Astecas, 99,34% (99,34%) nos Oto-Mangueanos e 97,31% (97,31%) nos Maias, sendo apenas equiparadas aos 98,93% nos Tupí (95,49%; da Amazônia, Centro-Oeste brasileiro e costa do Atlântico brasileira), seguidos um pouco mais distante pelos Quechua (dos Andes) com 95,60% (98,39%). Por fim, infere-se uma redução sensivelmente menor nos Jê (Planalto central brasileiro e sudeste da Amazônia) de 83,01% (97,51%; embora igualmente alta na estimativa com filtro de aparentados) e bastante menor na costa do Pacífico de 59,03% (43,45%; ainda menor quando se removem os aparentados). Esses resultados indicam que as consequências do processo de colonização foram diferencialmente distribuídos ao longo do continente e mais extremas nas regiões da Mesoamérica (e norte do México), e também no leste da América do Sul, com um impacto menor mas ainda assim imenso sobre as populações do oeste da América do Sul.

[65] Também deve ser levada em consideração a migração massiva de europeus incentivada pelo governo brasileiro que ocorreu após a abolição da escravatura (i.e. política de branqueamento racial (SCHWARCZ, 1993)).

[66] Gerações antes do presente.

[67] Diferença entre o Ne máximo e mínimo para o período, dividida pelo máximo, para o conjunto de populações de cada grupo linguístico.

[68] Apresentamos os resultados da análise conduzida pelo método IBDNe (BROWNING et al., 2018) no Capítulo 2, utilizando o parâmetro "*filtersamples = false*" (no texto) ou "*filtersamples = true*" (entre parênteses). Esse parâmetro determina se indivíduos aparentados devem ser removidos ou não. Na análise com remoção dos aparentados espera-se um maior o efeito do pequeno tamanho amostral, porém com uma redução do viés causado pela presença de indivíduos aparentados, especialmente sobre as estimativas de Ne das gerações mais recentes, no caso da manutenção dos indivíduos aparentados espera-se o efeito contrário.

Entretanto, a análise do Ne histórico a partir dos dados dos 47 Tupiniquim do Capítulo 3 mostra uma diminuição de 99,72% do Ne específico do componente nativo americano de ancestralidade no período (Figura 3 do Capítulo 3), de forma que esta é a maior redução do tamanho efetivo populacional inferida a partir dos nossos dados para a América como um todo. Adicionalmente, a data média estimada para o ponto mais baixo da redução da diversidade genética (e populacional) é de 8 GP (~1820 EC) para o conjunto completo de populações analisadas no Capítulo 2 (Figura 7B do Capítulo 2), e o mínimo valor de Ne inferido para os Tupiniquim analisados no Capítulo 3 ocorre há 7 gerações (Figura 3 do Capítulo 3), mostrando que o processo de declínio das populações nativas americanas provavelmente se acentuou mais tardiamente durante a colonização e não imediatamente após a chegada dos europeus, algo corroborado por evidências sólidas de estudos genéticos e também arqueológicos (e.g. (BROWNING et al., 2018; JONES et al., 2021)). No entanto, o mínimo N_e estimado é um ponto de inflexão, a partir do qual a diversidade genética passa a aumentar com o tempo, o que portanto sugere que em média as populações indígenas vem se recuperando desde o século 19. Entretanto, alguns estudos recentes apontam uma estabilização (ARROYO-KALIN; RIRIS, 2021) ou mesmo declínio da população 300 a 600 anos antes da chegada dos europeus (BUSH et al., 2021), o que teria sido causado pelo tamanho e/ou densidade da população ter ultrapassado a capacidade de suporte, ou mesmo pela ocorrência de epidemias de doenças autóctones, mudanças climáticas e sociais, como conflitos e outros tipos de eventos socialmente disruptivos.

Além disso, as análises compartilhamento intrapopulacional de segmentos em IBD[69] mostram um aumento significativo do comprimento médio total de IBD na categoria de blocos com mais de 16 cM[70] em praticamente todos os grupos analisados[71] (Figura suplementar 17D do Capítulo 2), indicando que no período mais recente ocorreu um aumento da frequência de casamento consanguíneos[72]. Isso, parcimoniosamente, seria uma consequência da fragmentação das populações e redução do número de indivíduos por população, processo desencadeado pela chegada dos europeus e portanto não relacionado a quaisquer mudanças culturais que possam ter levado à adoção de um comportamento reprodutivo de casamentos preferenciais, visto que a mudança ocorre de forma ubíqua no continente.

[69] *Identity-by-descent*, um locus ou segmento genômico é dito idêntico por descendência entre dois indivíduos quando ambos foram herdados de um mesmo cromossomo, e portanto de um mesmo indivíduo, na história recente.

[70] Centimorgan, unidade de medida de ligação genética, a qual define a distância entre dois *loci* no genoma para a qual são esperados em média 1% de cromossomos recombinantes a cada geração.

[71] Devido a aleatoriedade do processo de recombinação genética, qualquer haplótipo no genoma tem seu comprimento reduzido exponencialmente ao longo das gerações, de modo que segmentos maiores foram formados mais recentemente e vice-versa.

[72] Casamento entre indivíduos aparentados.

Nesse sentido chama a atenção ainda a proporção de segmentos em HBD[73] com mais de 16 cM em alguns grupos (Figura suplementar 17C do Capítulo 2), sobretudo nos nos Arara (Karib), mas também nos Urubu Kaapor e nos Suruí (ambos Tupí), que apresentam um aumento bastante expressivo do compartilhamento de HBD da penúltima para a última categoria de comprimento dos segmentos (de 8-16 para >16 cM), indicando um grande aumento da proporção de casamentos consanguíneos na história recente. A análise dos ROHs[74] do Capítulo 3 também apontou um aumento da taxa de casamentos consanguíneos nesse período, como inferido pelos segmentos com mais de 16 cM, especialmente nos grupos amazônicos (Figura 2 do Capítulo 3). Ademais, a IBD interpopulacional[75] é bastante reduzida no período colonial e recente da história em comparação ao período pré-contato (comparar Figura 5 e Figura suplementar 13 do Capítulo 2), sugerindo por consequência uma diminuição dos casamentos entre indivíduos de populações diferentes e possivelmente aumento da endogamia[76], sendo que alguns grupos apresentam IBD em segmentos gerados no período recente, mas estão circunscritos aos mesmos grupos linguísticos e/ou regiões geográficas, com clara divisão entre grupos da Mesoamérica (e norte do México), do leste e oeste da América do Sul. No período mais recente apenas os Guaraní no Centro-Oeste brasileiro e os Gavião e Zoró da região amazônica do Madeira-Guaporé apresentam IBD interpopulacional significativo (Figura suplementar 13B do Capítulo 2), sugerindo um fluxo gênico entre esses grupos na história recente.

Mapeamento da história da população Y no espaço e no tempo

No Capítulo 1 foi demonstrado que a distribuição do excesso de afinidade genética com populações australo-asiáticas entre grupos nativos americanos contemporâneos é muito mais extensa do que se sabia anteriormente (RAGHAVAN et al., 2015; SKOGLUND et al., 2015), sendo que a sua presença foi inferida não apenas na Amazônia (Karitiana e Suruí), no platô central brasileiro (Xávante) e no sul da região Centro-Oeste do Brasil (Guaraní Kaiowá), mas também na costa do Pacífico (Chotuna) (Figura 4 - ponto 41; Figura 1A do Capítulo 1). Nossos resultados sugerem também que em muitos grupos o sinal de ancestralidade pode estar um pouco abaixo do limite de significância estatística, dada a grande variação das estimativas de afinidade genética encontrada quando diferentes grupos são comparados (Figura 1B do Capítulo 1).

[73] *Homozygosity-by-descent*, por sua vez um locus ou segmento genômico é dito homozigoto por descendência caso um segmento de ambos cromossomos homólogos tenha sido herdado de um mesmo cromossomo na história recente.
[74] *Runs of Homozygosity*, são segmentos longos do genoma em homozigose, identificados através da análise de marcadores moleculares.
[75] Identidade por descendência entre indivíduos de populações diferentes.
[76] Casamento entre indivíduos aparentados, ou casamento consanguíneo.

Nesse sentido ainda, a inferência de uma grande variação intrapopulacional do sinal reforça a noção de que a presença desta ancestralidade pode estar sendo mascarada em alguns grupos, visto que ela é encontrada em maior proporção apenas em alguns indivíduos (Figura 2 do Capítulo 1), o que faz com que o sinal seja diluído nas comparações envolvendo a população inteira. Este efeito que pode ser observado nos resultados obtidos com o conjunto onde foi feita a seleção do maior número possível de indivíduos não aparentados, no qual o sinal perde a significância estatística nos grupos Guaraní Kaiowá, Karitiana e Suruí (Conjunto de dados suplementares 3 do Capítulo 1). Essa possibilidade de que o sinal possa estar sendo ocultado também é congruente com a baixa proporção de contribuição dessa ancestralidade australo-asiática, cujas inferências ficam na faixa de 1 a 3% (Figura 3 do Capítulo 1 e (SKOGLUND et al., 2015)), embora teoricamente poderiam ser ainda menores em outros grupos, aumentando a dificuldade de detecção. Desta forma, se torna ainda mais evidente que as comparações específicas de populações a serem testadas, bem como a composição das próprias populações, que de modo geral são delimitadas pelo pertencimento a grupos etnolinguísticos e proximidade geográfica, são fatores cruciais para a capacidade de detecção do sinal subjacente. Portanto, é possível que esses elementos, somados à sub amostragem da diversidade de grupos etnolinguísticos existentes e ao pequeno número de indivíduos amostrados por grupo, sejam responsáveis pelo padrão irregular e aparentemente aleatório de detecção encontrado até o momento.

A ausência dessa afinidade genética na maioria das populações nativas americanas poderia ainda ser a consequência de alguns outros fatores. Os grupos portadores desta ancestralidade australo-asiática podem ter se miscigenado múltiplas vezes com outros de ancestralidade inteiramente nativa americana (mais especificamente SNA), além de certamente terem sofrido forte efeito da deriva genética, intensificado em diversos momentos, sobretudo em decorrência da série de gargalos populacionais e efeitos fundadores enfrentados durante o povoamento inicial (WANG et al., 2007; REICH et al., 2012; FAGUNDES et al., 2018), assim como também das dinâmicas populacionais transcorridas ao longo do Holoceno, como por exemplo os eventos de fissão-fusão populacionais (NEEL; SALZANO, 1967) e dispersões populacionais (e.g. Capítulo 3), os quais podem ter contribuído também para o aumento do fluxo gênico e consequentemente para uma homogeneização genética. Esse conjunto de forças evolutivas pode assim ter contribuído para que o sinal tenha se diluído na maior parte dos grupos onde estava originalmente presente, ao mesmo tempo em que pode ter sido preservado e se mantido em uma proporção mais significativa em alguns grupos mais isolados, ou que por algum outro motivo estavam menos propensos ao fluxo gênico. Nesse sentido, o maior nível de isolamento dos grupos do leste da América do Sul poderiam em tese ter levado a manutenção de uma proporção mais alta dessa ancestralidade em alguns desses grupos. As expectativas dessa

hipótese se alinham em boa medida ao padrão de detecção do sinal na América do Sul, onde a afinidade é encontrada apenas em algumas das populações mais isoladas e diferenciadas geneticamente (como demonstrado sobretudo pelos resultados do Capítulo 2), os Karitiana, Suruí e Xávante, e os Guaraní Kaiowá provavelmente com menor grau de isolamento e uma população maior (Figura 1 e conjunto de dados suplementares 3 do Capítulo 1). Por outro lado, é possível também que o extenso efeito da deriva genética enfrentado por alguns grupos tenha contribuído para obscurecer o sinal de alguma forma, por meio da fixação de alelos e perda da diversidade genética, ainda mais no contexto da baixa diversidade genética dos nativos americanos. De todo modo, é preciso ponderar que os padrões genéticos observados nos nossos resultados se referem especificamente ao período recente do Holoceno tardio, o que não é o suficiente para explicar a suposta preservação do sinal durante o Holoceno inicial e médio.

Adicionalmente, aqui nós demonstramos que esse sinal de afinidade com populações australo-asiáticas também está presente nos Chotuna, localizados na costa do Oceano Pacífico peruana, para os quais não há evidências de que tenham permanecido particularmente mais isolados do que os outros grupos da mesma região. A presença do sinal na costa do Pacífico indica claramente que o sinal também estava presente no outro lado da cordilheira dos Andes e por conseguinte condiciona alguns cenários possíveis para a introdução do sinal na América do Sul. São eles:

(i) A ancestralidade australo-asiática teria sido introduzida durante ou pouco tempo depois do povoamento inicial da América do Sul (hipótese apresentada na Figura 4 - setas laranja e setas rosa no mapa da América do Sul), seja por meio dos primeiros migrantes que constituiriam uma população geneticamente estruturada ou por um processo de povoamento contínuo, com variação da ancestralidade dos migrantes ao longo do período de povoamento.

> (i.a) No caso de uma população formadora geneticamente estruturada, a ancestralidade australo-asiática teria provavelmente sido introduzida em um número maior de populações, mas o sinal teria sido diluído de acordo com as dinâmicas e movimentos populacionais posteriores, e teria ainda necessariamente ocorrido em uma data de entrada anterior a primeira divergência entre os grupos ancestrais que deram origem às populações da costa do Pacífico, Andes e Amazônia, que é datada em ~12.000 AP (Figura 4 - ponto 27) (HARRIS et al., 2018).
>
> (i.b) Por outro lado, se o povoamento ocorreu de forma contínua, a ancestralidade australo-asiática potencialmente também poderia ter se dispersado pelas mesmas regiões que os primeiros migrantes, mas muito provavelmente teria contribuído para um número menor de regiões onde a densidade populacional dos grupos portadores dessa ancestralidade foi maior. Entretanto, até o momento o sinal foi detectado apenas em um indivíduo antigo com ~10.000 AP de Lagoa Santa (Figura 4 -

ponto 30) (MORENO-MAYAR et al., 2018b) e o fato de que ele não é detectado em nenhum outro indivíduo antigo, nem mesmo aqueles encontrados no mesmo sítio ou em região próxima, ou com idade similar, ainda não pôde ser explicado. De qualquer forma, a dispersão da ancestralidade australo-asiática deve ter ocorrido necessariamente em algum momento entre o início do povoamento da América do Sul e 10.000 AP, de modo a explicar a presença do sinal em Lagoa Santa nesta data.

(ii) A ancestralidade australo-asiática não estaria presente na população ancestral comum (nem em um povoamento inicial contínuo) que contribuiu com a maior parte da ancestralidade das populações antigas e contemporâneas da costa do Pacífico, Andes e Amazônia e portanto teria uma origem em uma outra dispersão para o continente, em momento posterior ou anterior a ~12.000 AP (i.e. data inferida de divergência entre os referidos grupos populacionais). Esse tipo de cenário provavelmente envolveria pelo menos dois eventos independentes de miscigenação, um na costa do Pacífico e outro na Amazônia.

(ii.a) No caso de uma dispersão anterior à chegada dos ancestrais dos nativos americanos, esta população inicial possivelmente seria composta por populações pequenas e esparsas, que dessa maneira teriam contribuído para a ancestralidade de apenas alguns poucos grupos nativos americanos, provavelmente restritos à América do Sul, no entanto essa hipótese não pode ser testada propriamente pela ausência de dados genéticos de indivíduos do período pré-Clovis (WILLERSLEV; MELTZER, 2021).

(ii.b) Ao passo que, no cenário de uma dispersão posterior, os novos migrantes possivelmente também teriam contribuído apenas para a ancestralidade dos grupos indígenas de algumas regiões. Presumivelmente deveria se encontrar indivíduos antigos que representassem essa dispersão populacional adicional, o que ainda não aconteceu apesar de uma quantidade significativa de indivíduos antigos analisados, com a exceção de um indivíduo que viveu a ~10.000 AP de Lagoa Santa (MORENO-MAYAR et al., 2018b; WILLERSLEV; MELTZER, 2021), da mesma forma a dispersão necessariamente teria ocorrido após o povoamento inicial e antes de 10.000 AP.

Além disso, camadas adicionais de complexidade são introduzidas a qualquer dos cenários apresentados acima. Primeiramente, porque as dinâmicas populacionais e eventos de mistura (e.g. com grupos de ancestralidade inteiramente SNA) possivelmente teriam diluído, apagado ou tornado o sinal insignificante na maior parte dos grupos. Tais eventos sabidamente foram mais frequentes e intensos nos Andes, onde de fato o sinal não foi detectado, possivelmente devido ao histórico da região de maior fluxo gênico, dispersões populacionais (motivadas pela expansão dos impérios Tiwanaku, Wari e Inca (HARRIS et al., 2018; NAKATSUKA et al., 2020) ou por migrações motivadas por instabilidade climática (FEHREN-SCHMITZ et al., 2014) e com uma consequente homogeneização genética. Ao mesmo tempo, essa

ancestralidade poderia ter se conservado em proporções mais elevadas e estatisticamente significativas em algumas populações mais isoladas, como argumentado anteriormente.

Em segundo lugar, o contato com os europeus e as suas drásticas consequências, principalmente a vertiginosa redução do tamanho populacional, da diversidade genética e da distribuição geográfica dos diversos grupos nativos americanos (THORNTON, 1987; STANNARD, 1993; MONTENEGRO; STEPHENS, 2006; UBELAKER, 2006; ADHIKARI et al., 2016, 2017; ONGARO et al., 2019), poderia ajudar a explicar em parte a ausência do sinal, não apenas nas populações sul americanas como também naquelas da América Central e do Norte. Soma-se a isso o fato de que populações nativas americanas dos Estados Unidos estão largamente ausentes dos painéis de dados genômicos, o que impossibilita não apenas investigar a extensão desse sinal na América do Norte, como também uma melhor compreensão da história populacional da entrada dessa ancestralidade no continente como um todo.

Esse conjunto de elementos nos ajuda a compor um quadro mais detalhado sobre as possibilidades de quando e como a ancestralidade australo-asiática chegou e se dispersou na América do Sul. No entanto, ainda persistem algumas questões fundamentais sobre a identidade do grupo ancestral causador da introdução dessa ancestralidade e por onde ele teria se dispersado pelo interior do continente. Em 2015, Skoglund et al. propuseram um modelo para responder a pergunta sobre quem seria a população responsável pela introdução do excesso de afinidade de alguns nativos americanos com populações australo-asiáticas, sobretudo com os Onge das Ilhas de Andamão. Segundo esse modelo, uma população ainda não amostrada (População Y) seria formada pela mistura de um ramo irmão dos Onge e um ramo irmão dos grupos Pima e Mixe (SNA), a qual teria através de uma segunda mistura, desta vez com um ramo SNA mais próximo apenas aos Mixe, dado origem a alguns grupos nativos brasileiros (Karitiana, Suruí e Xávante). Diferentes frações de ancestralidade provenientes dos grupos envolvidos nesses dois eventos de mistura supracitados possuem suporte estatístico, porém todas as combinações com bom ajuste constringem a contribuição da ancestralidade australo-asiática a 1-2% do genoma dos grupos nativos brasileiros contemporâneos onde ela foi detectada (SKOGLUND et al., 2015). Todavia, a própria existência de uma população putativa nos moldes do modelo proposto para a População Y ainda permanece uma questão em aberto, dado que o sinal australo-asiático foi detectado apenas em um indivíduo antigo com ~10.000 AP de Minas Gerais, o qual apresenta uma fração de ancestralidade na mesma faixa (~3%) que as populações contemporâneas (MORENO-MAYAR et al., 2018b).

Nossos resultados, nessa perspectiva, contribuem não apenas com mais informações sobre linha do tempo da chegada dessa ancestralidade ao continente, como também fornecem elementos para distinguir qual a rota de entrada mais provável, particularmente no que diz respeito à América do Sul. De acordo com nossos modelos de história populacional, segundo o

cenário mais ajustado aos dados genéticos a ancestralidade australo-asiática teria chegado através de uma dispersão pela costa do Pacífico, visto que os grupos costeiros divergem primeiro após o evento de mistura dos ramos nativo americano e australo-asiático (Figura 3C do Capítulo 1), ou apresentam concomitantemente menores estimativas de deriva genética no seu componente australo-asiático e maior fração de contribuição do mesmo (Figura 3D do Capítulo 1). É possível que tenha ocorrido a entrada e dispersão tanto pela costa oeste, quanto pelo interior do continente (Figura 3D do Capítulo 1); no entanto, no modelo que permite essa interpretação a diferenciação genética entre os ramos nativos americanos formadores do grupo costeiro e amazônico é a mínima (1 F_{ST} multiplicado por 1000, em cada ramo), indicando que poderiam fazer parte de um mesmo grupo estruturado geneticamente e com frações diferentes de ancestralidade australo-asiática.

Portanto, nossos resultados indicam que a dispersão teria acontecido a partir de uma ocupação inicial pelo litoral do Oceano Pacífico, embora não seja possível precisar onde e como exatamente se deu essa dispersão partindo do litoral, faz sentido que tenha ocorrido inicialmente mais ao norte do continente onde a cordilheira dos Andes possui trechos de menor altitude. Essa dispersão pode ter ocorrido tanto durante o povoamento inicial ou em outro momento, de acordo com os cenários possíveis descritos anteriormente. De todo modo, a inferência de que a ancestralidade australo-asiática se dispersou pela costa oeste corrobora um corpo crescente de evidências que aponta a viabilidade da rota da costa do Pacífico como via de a entrada no continente sul-americano e também nas Américas, ainda que a ancestralidade australo-asiática tenha se dispersado por ela em momento posterior ao povoamento inicial.

Até recentemente havia um debate em torno da questão, sendo favorecida a hipótese de que a entrada inicial na América do Norte teria acontecido através de um corredor formado ao fim do período glacial, próximo às Montanhas Rochosas, entre as geleiras *Cordilleran* e *Laurentide* que ocupavam toda a porção norte do continente durante o UMG desde aproximadamente 23.000 AP (POTTER et al., 2018; WILLERSLEV; MELTZER, 2021). Entretanto, hoje se sabe que o corredor não estaria aberto antes de pelo menos 15.000 a 14.000 AP (MARGOLD et al., 2019) e ainda que provavelmente não seria viável como passagem para grupos humanos antes de 13.000 AP, devido a restrições ecológicas que não permitiriam a sobrevivência (HEINTZMAN et al., 2016; PEDERSEN et al., 2016). Portanto essa linha do tempo não é compatível com a existência de sítios ao sul das geleiras no mínimo tão antigos quanto ~14.500 AP, alguns inclusive localizados na América do Sul (e.g. Monte Verde I e II) (DILLEHAY et al., 2008, 2015; GILBERT et al., 2008; WATERS et al., 2011, 2018; JENKINS et al., 2012; WILLIAMS et al., 2018). Sendo assim os primeiros nativos americanos teriam muito provavelmente atravessado as geleiras através da costa pacífica do Alasca e do Canadá (hipótese apresentada na Figura 4), uma rota que teria se tornado possível a partir de ~17.000-15.000

com o degelo da borda oeste da geleira *Cordilleran* (MENOUNOS et al., 2017; DARVILL et al., 2018; LESNEK et al., 2018), o que consequentemente se alinha às datações dos sítios pré-Clovis supracitados. Além disso, após a chegada ao sul das geleiras no norte dos Estados Unidos, os primeiros americanos teriam ocupado e se dispersado ao longo da costa do Pacífico, avançando relativamente rápido até a América do Sul (BODNER et al., 2012), chegando ao continente tão cedo quanto ~15.500 AP de acordo com estimativas obtidas pela distribuição de probabilidades de datações de radiocarbono (PRATES; POLITIS; PEREZ, 2020).

Essa ocupação predominantemente litorânea pode ser explicada pelo fato de que esses grupos seriam culturalmente adaptados à vida à beira-mar, logo apresentariam uma série de estratégias de subsistência dependentes dos recursos naturais marítimos obtidos sobretudo a partir da pesca e coleta. Nessa linha de argumentação é postulada a chamada hipótese da rodovia de *kelp*[77], segundo a qual os ambientes costeiros altamente produtivos das florestas de algas *kelp* (ricos em mariscos, peixes, mamíferos marinhos, pássaros e algas), presentes desde a costa do Pacífico asiática até a porção mais meridional da América do Sul, teriam fornecido um habitat contínuo[78] adequado às estratégias de sobrevivência de grupos pescadores-coletores que portanto poderiam se dispersar por toda a costa, sendo necessárias apenas pequenas adaptações graduais (ERLANDSON et al., 2007). Portanto, essas adaptações à vida na costa poderiam ajudar a explicar a ausência do sinal australo-asiático em populações da América Central e do Norte, considerando que esses primeiros grupos, portadores dessa ancestralidade, fossem pescadores-coletores e tivessem se concentrado apenas na região da costa do Pacífico, migrando rapidamente para a América do Sul e contribuindo pouco ou nada para a composição genética de grupos da América Central e do Norte. É preciso ainda ponderar o fato de que possíveis evidências dessas dispersões humanas na porção mais costeira da plataforma continental, que seriam consequentemente as mais antigas, estariam hoje ocultas pelo aumento do nível dos oceanos com o fim do período glacial (LAMBECK et al., 2014), impedindo portanto que a hipótese seja testada de maneira razoável e definitiva.

Ainda considerando os níveis mais baixos dos oceanos, um estudo recente evidencia a existência de uma grande quantidade de ilhas no mar de Bering entre 30.000 e 8.000 AP, as quais conjuntamente receberam o nome de arquipélago transitório de Bering[79] (DOBSON; SPADA; GALASSI, 2021) (Figura 4 - ponto 13). Interessantemente, no contexto do povoamento inicial das Américas esse arquipélago teria fornecido uma maior abundância de recursos marinhos, disponibilidade de ilhas que podem ter atuado como pontos de parada seguros na

[77] *Kelp Highway Hypothesis*.
[78] Com exceção da porção mais tropical onde se encontram recifes de corais e mangues, de todo modo também extremamente ricos em recursos (ERLANDSON et al., 2007).
[79] *The Bering Transitory Archipelago*.

travessia pela costa e a possibilidade de deslocamentos marítimos por águas mais facilmente navegáveis e protegidas. Ainda mais relevante é o fato de que a formação dessas ilhas a medida que o nível oceânico variou fornece um mecanismo bastante plausível para explicar o longo isolamento da população beringiana do fluxo gênico de povos do leste asiático (i.e. a hipótese de permanência na Beríngia), além da própria estruturação genética dos beringianos, visto que podem ter permanecido parcialmente isolados em ilhas diferentes (DOBSON; SPADA; GALASSI, 2021). Tomando esses fatores em conjunto, uma nova hipótese sobre o povoamento inicial das Américas foi proposta por Dobson et al.[80], segundo a qual o arquipélago teria fornecido não apenas uma rota, com ilhas localizadas relativamente próximas umas às outras, e os recursos necessários para a dispersão inicial e povoamento das Américas, como também teria selecionado e moldado os modos de vida e habilidades exigidas para que isso ocorresse (e.g. pesca, caça, canoagem, sobrevivência em ambientes costeiros/marítimos). Por fim, o aumento do nível oceânico, particularmente a partir de ~16.000 AP teria ainda simultaneamente aumentando o isolamento entre grupos beringianos e desses com o leste da Ásia à medida que a densidade de ilhas diminuía, além de ter forçado a dispersão do oeste para o leste, visto que o arquipélago foi progressivamente inundado também nesse sentido, de modo que a migração ao longo da costa pode ter sido desencadeada ou mesmo impelida por esse processo (DOBSON; SPADA; GALASSI, 2021). Desta forma, a existência do arquipélago transitório de Bering reforça o modelo de uma dispersão contínua (ou estruturada temporalmente) pela costa do Pacífico, fornecendo todos os elementos necessários para explicar os diversos aspectos desse processo, como a permanência na Beríngia e a provável estruturação genética entre grupos portadores e não-portadores de ancestralidade australo-asiática; entretanto, a maioria dos registros arqueológicos deixados por essa possível dispersão estariam hoje no fundo do mar de Bering.

As dificuldades aqui expostas poderiam ao menos em parte serem superadas com uma amostragem mais representativa da diversidade de grupos etnolinguísticos nativos americanos e também mais bem distribuída geograficamente, no melhor dos casos incluindo sítios subaquáticos da costa do Pacífico. Esses dados poderiam ser analisados buscando inferir e delimitar grupos nativos americanos mais geneticamente homogêneos (não necessariamente correspondendo aos grupos étnicos ou geográficos) para serem usados como grupos teste nas estatísticas, desse modo minimizando os efeitos da grande variação intrapopulacional do sinal australo-asiático identificados (Figura 2 do Capítulo 1), o que certamente contribuiria para uma melhor delimitação da contribuição dessa ancestralidade no continente americano e consequentemente da sua história populacional.

[80] *Stepping-Stones hypothesis*.

Além disso, uma melhor amostragem e estudo do leste e nordeste asiáticos poderiam trazer algumas respostas necessárias para compreender o surgimento e o padrão de dispersão dessa ancestralidade na Ásia e sua entrada nas Américas, como evidenciado pela descoberta recente da afinidade entre a população contemporânea dos Onge (i.e. sinal australo-asiático) e dois indivíduos antigos (La368 com 7950-7795 AP e Ma911 com 4415-4160 AP) do sudoeste asiático continental (Laos e Malásia) (Figura 4 - pontos 10 e 11) (MCCOLL et al., 2018), pertencentes ao grupo de caçadores-coletores conhecidos como hoabinhianos, o qual teria se formado (dados arqueológicos) a pelo menos ~44.000 AP (Figura 4 - ponto 2) (JI et al., 2016), aproximadamente ~21.000 AP após a chegada dos humanos modernos no sudeste asiático (~65.000 AP) (Figura 4 - pontos 1) (DEMETER et al., 2017; WESTAWAY et al., 2017). Os hoabinhianos analisados (La368 e Ma911) podem ser modelados como um grupo irmão dos Onge contemporâneos, e compartilham componentes de ancestralidade não apenas com o próprios Onge, mas também com os Jehai (Malásia peninsular), sendo ambos os grupos supostamente descendentes dos hoabinhianos (MCCOLL et al., 2018). Ademais um indivíduo Jōmon do Japão pode ser modelado como uma mistura entre as ancestralidades hoabinhiana e do leste asiático, indicando o quão ao norte teria chegado a contribuição genética dos hoabinhianos (Figura 4 - seta azul pontilhada partindo do ramo dos hoabinhianos e chegando à ilha do Japão). Em contrapartida, o sinal de afinidade australo-asiático não é observado nos grupos mais recentes do sudeste asiático, algo que segundo Mccoll et al. seria explicado pela expansão dêmica de grupos agriculturalistas do leste asiático para a região sudeste iniciada a partir de ~4.000 AP, os quais teriam substituído parcialmente e se miscigenado aos hoabinhianos *(MCCOLL et al., 2018)*.

Essas evidências portanto sugerem que grupos relacionados aos hoabinhianos podem estar envolvidos nos eventos de fluxo gênico que resultaram no sinal de afinidade genética entre populações nativas americanas e australo-asiáticas. Isso é também sugerido por análises morfométricas de populações (antigas e contemporâneas) do leste asiático, onde podem ser identificados dois agrupamentos principais, com prevalência no nordeste (NEA) e sudeste (SEA) da Ásia, muito provavelmente oriundos de uma divergência no oeste da Eurásia e anterior a ocupação do leste da Eurásia (tão antiga quanto ~65.000-50.000 AP) (MATSUMURA et al., 2019). Os SEA teriam se dispersado pelo sudeste da Ásia, Sondalândia e Sahul, onde formariam os grupos antigos pré-agriculturalistas hoabinhianos e os australo-papuanos contemporâneos, enquanto que os NEA teriam avançado pelo nordeste da Ásia e provavelmente entrado nas Américas pela Beríngia, dando origem aos nativos americanos, povos do nordeste e centro-leste da Ásia, e eventualmente também a grupos agriculturalistas que posteriormente teriam

expandido do norte substituindo e miscigenando com os hoabinhianos do sudeste asiático (MATSUMURA et al., 2019)[81].

A conexão entre as evidências genéticas e morfológicas em relação a origem e dispersão da ancestralidade australo-asiática nas Américas vem do fato de que os Onge portadores de ancestralidade hoabinhiana (MCCOLL et al., 2018) e também causadores da maior intensidade de detecção do sinal australo-asiático nos nativos americanos (Conjunto de dados suplementares 3 do Capítulo 1 e (SKOGLUND et al., 2015)), apresentam um sinal de afinidade morfológica com populações do nordeste asiático, o que segundo Matsumura et al. indica a existência de um evento de miscigenação (entre grupos SEA e NEA) (MATSUMURA et al., 2019), e em última instância aponta que um ramo próximo aos Onge pode ter sido o responsável pela introdução da ancestralidade australo-asiática em alguns nativos americanos. Soma-se a isto uma análise demográfica de modelos genético-espaciais que buscou testar o papel do clima e do relevo na composição genética dos grupos humanos (DELSER et al., 2021)[82], onde de acordo com os modelos mais bem ajustados aos dados dos genomas antigos de caçadores coletores disponíveis, após a primeira divergência fora da África um grupo se dispersou para o oeste das montanhas Zagros[83] e deram origem aos Caçadores-Coletores Ocidentais[84], enquanto que um segundo grupo se dirigiu para leste dando origem aos asiáticos. O grupo dos asiáticos, por sua vez, teria divergido novamente devido à barreira promovida pelos Himalaias e deserto de Gobi, com ramo ao sul e outro ao norte, os quais teriam se reunido novamente no leste asiático (DELSER et al., 2021). Segundo esse modelo, houve contato e oportunidades para fluxo gênico entre populações do nordeste e sudeste asiático, o que em última análise poderia resultar na presença de uma fração de ancestralidade australo-asiática nos grupos formadores dos nativos americanos.

Essa hipótese se torna ainda mais provável à luz de evidências recentes que apontam para uma divergência ainda na Ásia não apenas dos ramos AB e ANA (Figura 4 - ponto 14), como também dos NNA e SNA (NING et al., 2020) (Figura 4 - pontos 16), reforçando dessa forma que houve muitas oportunidades para a interação e fluxo gênico entre os grupos formadores dos nativos americanos e outras populações do leste asiático ainda fora da Beríngia. Tais interações possibilitariam que a introgressão tivesse ocorrido apenas em alguns SNA (Figura 4 - seta rosa pontilhada e laranja na Sibéria), explicando assim a ausência do sinal nos NNA e AB. Desta forma, investigar as relações entre grupos nativos da América e do leste da Ásia com diversas

[81] Entretanto, existe a possibilidade de que os SEA tenham se dispersado primeiramente pela Sibéria antes de migrar para o sudeste asiáticos, e portanto os NEA teriam se originado na Sibéria pela adaptação ao clima frio da região durante o UMG e posteriormente alguns grupos agriculturalistas descendentes dos NEA teriam se expandido também para o sudeste asiático (MATSUMURA et al., 2019).

[82] *Climate-Informed Spatial Genetic Modelling* (CISGeM).

[83] Cordilheira de montanhas que cruza o Irã, norte do Iraque e sudeste da Turquia.

[84] *Western Hunter-Gatherers*.

profundidades temporais, no caso dos últimos com foco especial nos hoabinhianos e grupos mais relacionados, parece ser um caminho bastante promissor para elucidar a origem e dispersão inicial da ancestralidade australo-asiática para as Américas.

Fatores influenciadores da diversidade genética

No Capítulo 2 nós demonstramos que a ancestralidade dos grupos atuais remonta ao menos parcialmente a ancestralidade local de indivíduos antigos, como indicado pelo padrão de excesso relativo de compartilhamento de alelos entre populações nativas contemporâneas e indivíduos antigos da mesma região ou de regiões próximas, em comparação com populações modernas de outras regiões (Figura suplementar 14 do Capítulo 2). Adicionalmente, a análise da distância genética entre indivíduos antigos e populações modernas (MDS[85] dos valores de 1 – *outgroup* F_3) evidencia a existência de um gradiente de diferenciação genética ao longo do espaço e do tempo, recapitulando os eventos de divergência e as rotas de dispersão, iniciando a partir do ponto de entrada do povoamento do continente com as amostras antigas da Sibéria, passando pelos AB (Beríngia), pelo ramo dos NNA (América do Norte) e finalmente chegando aos SNA (Figura 6 do Capítulo 2), grupo no qual se incluem todas as populações contemporâneas analisadas. A existência dessa variação genética clinal é corroborada pela inferência de uma correlação moderada das duas primeiras dimensões do MDS das distâncias genéticas entre grupos modernos com a latitude e principalmente com a longitude, algo que também é apontado pela inferência de uma correlação entre pelo menos um dos componentes principais de variação genética (PCA)[86] dos grupos nativos modernos e a longitude (Figura 2C-D do Capítulo 2). Do mesmo modo, uma correlação entre valores de distância genética (1 – *Outgroup* F_3) e geográfica entre pares de indivíduos modernos também é inferida (Capítulo 2). Tomados em conjunto, esses resultados indicam que a distribuição atual da variação genética foi influenciada significativamente pelo processo de povoamento inicial das Américas, assim como pela dispersão inicial dentro da América do Sul que muito provavelmente se deslocaram primordialmente do oeste para o leste.

Efetivamente a existência de um gradiente de variação genética de norte a sul e de oeste para leste já havia sido observada anteriormente, contudo o padrão identificado era de uma redução da diversidade genética e um aumento da diferenciação genética entre grupos a partir do ponto setentrional de entrada no continente no Alasca até as regiões mais meridionais da América do Sul, inclusive com o mesmo padrão observado de oeste para leste da América do Sul

[85] *Multidimensional scaling*.
[86] *Principal component analysis*.

(WANG et al., 2007; O'ROURKE; RAFF, 2010; REICH et al., 2012; VERDU et al., 2014; FEHREN-SCHMITZ, 2020; SANTOS, 2020). Interessantemente esse padrão também é encontrado nos nossos resultados, onde demonstramos que existe uma diferença significativa (Teste U de Mann-Whitney) entre a diversidade genética (como medido pelo coeficiente de endogamia F_{ROH}) entre grupos do leste e oeste da América do Sul, assim como entre os grupos do leste da América do Sul e da Mesoamérica (Figura 8A do Capítulo 2), além disso é inferida uma correlação entre o coeficiente de endogamia e a longitude, a qual aumenta do oeste para o leste (Figura 8A do Capítulo 2).

Desta forma, por um lado a correlação entre as distâncias geográficas e genéticas entre populações indica um padrão de isolamento por distância[87], por outro lado, a redução da diversidade genética do oeste para o leste da América do Sul sugere que este gradiente de variação genética pode ser resultante dos gargalos populacionais em série que ocorreram ao longo das sucessivas dispersões dentro do continente. Em vista disso, esses resultados são similares ao padrão global de variação genética das populações humanas, como demonstrado pela correlação entre as distâncias genéticas (F_{ST}) e geográficas[88] de todos os pares de populações do CEPH-HGDP (RAMACHANDRAN et al., 2005) e diminuição da heterozigosidade a partir da África, ponto de origem e centro de dispersão dos humanos modernos para o restante do planeta (PRUGNOLLE; MANICA; BALLOUX, 2005), observações que em conjunto indicam que a distribuição da variabilidade genética atual foi influenciada tanto pelo efeito do isolamento por distância quanto pela ocorrência de gargalos populacionais em série na história das dispersões humanas. Portanto essa conclusão também é plausível em relação aos padrões aqui identificados para as populações nativas americanas. Todavia, em nosso caso, essas observações de padrões genético-geográficos devem ser tomadas com cuidado, visto que apesar de dispormos do maior conjunto de dados genéticos autossômicos de populações contemporâneas da América do Sul até o momento, as amostras ainda são bastante dispersas, desigualmente distribuídas e escassas no território, de modo que no futuro uma distribuição mais uniforme e representativa de todas as regiões pode indicar outros padrões, especialmente no que se refere a contextos subcontinentais e locais. Contraditoriamente, um estudo recente do maior conjunto de dados já reunido de marcadores genéticos uniparentais, não encontrou correlação significativa entre distâncias genéticas e geográficas (BISSO-MACHADO; FAGUNDES, 2021), o que talvez possa ser explicado pela menor resolução da informação genética contida nesses

[87] Declínio da similaridade genética em função do aumento da distância geográfica, causado pelo fato de que a probabilidade do acasalamento entre indivíduos diminui com o aumento da distância, de modo que a partir de certo ponto o efeito do fluxo gênico é superado pela deriva genética resultando em diferenciação genética.

[88] *Great-circle distances*: distâncias geográficas que levam em consideração a curvatura da terra. Além disso, no cálculo das distâncias foram usados pontos de referência para evitar obstáculos geográficos.

marcadores quando comparados com centenas de milhares de marcadores autossômicos, poderia ainda ser explicado pela amostragem do referido trabalho cobrir toda a América de forma mais igualmente distribuída no território, o que poderia reduzir o sinal de um padrão mais específico da América do Sul ou mesmo reduzir a correlação inferida entre as distâncias por capturar melhor movimentos secundários ocorridos após o povoamento inicial.

Um indício adicional que corrobora a existência desse eixo de variação genética oeste-leste vem do padrão de diversificação populacional (como inferido pelo Treemix (PICKRELL; PRITCHARD, 2012)), que se inicia com a divergência dos grupos da costa do Pacífico, seguidos pelos andinos, passando pelos grupos do oeste, centro e leste da Amazônia, até os Guaraní no sul da região Centro-Oeste brasileira, exatamente nesta ordem (Figura 2A-B do Capítulo 2). Esse padrão de diversificação reforça ainda três pontos importantes discutidos anteriormente: (i) que o povoamento da América do Sul provavelmente se iniciou pela costa do Pacífico; (ii) que o gradiente de variação genética longitudinal portanto remonta a dispersão inicial, de forma que ele teria sido formado no mínimo em parte como consequência de gargalos populacionais em série ao longo da dispersão para o leste do continente; (iii) que grupos mais próximos geograficamente são mais similares geneticamente, algo que também pode ter sido influenciado pelo efeito do isolamento por distância. Além disso, a diferenciação genética entre grupos aumenta de oeste para leste como indicado pelos maiores comprimentos de ramo (i.e. maior efeito de deriva genética) dos grupos dessa região, algo também já observado no passado (WANG et al., 2007; REICH et al., 2012), entretanto esse último padrão pode ter sido amplificado por meio de dinâmicas populacionais específicas da região leste, as quais intensificaram a deriva genética (e.g. eventos de fissão-fusão populacionais (NEEL; SALZANO, 1967)).

Além da geografia a diversidade genética também é influenciada pela diversidade cultural e ambiental de forma dinâmica e interdependente, contudo é bastante difícil isolar os efeitos de cada um desses fatores visto que há uma relação de influência mútua entre eles, de modo que grupos próximos geograficamente geralmente falam línguas das mesmas famílias, possuem práticas culturais e estratégias de subsistência compartilhadas, e também tendem a ocupar os mesmos tipos de ambientes. A Figura 3 apresenta uma matriz de distância genética (F_{ST}) estimada entre todos os pares de grupos nativos americanos com indivíduos não-miscigenados e não-aparentados, como dito populações dos mesmos grupos linguísticos ocupam predominantemente os mesmos biomas e ecorregiões (ver barras codificadas por cores a esquerda do mapa de calor), a análise de agrupamento hierárquico apresentada pelos dendrogramas indica que diferentes populações dentro dos mesmos grandes grupos linguísticos não necessariamente apresentam perfis de distâncias genéticas similares, mas tendem a ser reunidas nos mesmos grupos, assim como as populações das diferentes regiões continentais são majoritariamente agrupadas em concordância com sua localização, entretanto esses

dendrogramas não devem ser tomados como representações da história populacional desses grupos, visto que diferentes fatores influenciam a diferenciação genética entre cada um dos pares de grupos além da relação de parentesco entre eles, como por exemplo o histórico do tamanho efetivo populacional e o efeito da deriva genética, o contato e a miscigenação com outros grupos e assim por diante.

Buscando avaliar a existência de uma estrutura genética significativa entre grupos étnicos (i.e. populações) e entre grupos linguísticos, como delimitados por diferentes classificações linguísticas (JOSEPH; RUHLEN, 2007; EBERHARD, D. M., SIMONS, G. F., & FENNIG, C. D., 2020; HAMMARSTRÖM, 2021), assim como entre ecorregiões (BAILEY; HOGG, 1986) e biomas (COSTA et al., 2018), e concomitantemente quantificar a proporção da variação genética explicada por cada uma dessas categorias, foi utilizada a AMOVA (EXCOFFIER; SMOUSE; QUATTRO, 1992). Nesse sentido, análises independentes foram realizadas para cada classificação linguística e ecogeográfica das populações (Tabela 1). Foi inferida uma estrutura genética significativa para todos os níveis hierárquicos das análises, com exceção da variação entre amostras dentro das populações e entre biomas (Tabela 1E). Portanto os resultados suportam a existência de acasalamento aleatório entre indivíduos das mesmas populações, ao mesmo tempo é inferida uma diferenciação genética entre populações e entre grupos linguísticos assim como entre ecorregiões (Tabela 1). Considerando apenas a variação genética relativa, representada pela porcentagem da variação total que não inclui a variação dentro dos indivíduos (que compõe ~93% do total), podemos observar que a menor parte da variação é atribuída a diferenciação entre grupos linguísticos (13,39-16,22%; Tabela 1A-C) e pouco menos ainda a ecorregiões (9,06%; Tabela 1D) e biomas (2,71%; Tabela 1E), ao passo que a variação genética relativa entre populações dentro dos grupos é maior, entre 35,59-38,87% na análise de grupos linguísticos (Tabela 1A-C) e de 45,39% e 49% nas análises das ecorregiões e biomas (Tabela 1D-E), respectivamente. Por fim, apesar de se estimar que uma grande parte da variação relativa se encontra entre indivíduos dentro das populações (45,54-48,29%), não é possível rejeitar a hipótese de acasalamento aleatório entre indivíduos da mesma população. Essas observações indicam assim que populações são homogêneas geneticamente, porém diferenciadas entre si, um padrão esperado para populações relativamente pequenas e isoladas, que sofreram efeito significativo de deriva genética e apresentam baixo fluxo gênico. Além disso, a diversidade linguística e ambiental parece explicar uma parte menor, embora significativa, da variação genética.

Nas análises sumarizadas na Tabela 1 foram testadas as mesmas classificações em grupos linguísticos e ecorregiões analisadas por Bisso-Machado et al. (BISSO-MACHADO;

FAGUNDES, 2021)[89], onde foram feitas AMOVAs de marcadores uniparentais, as quais indicaram que a maior parte da variação genética encontra-se dentro das populações, com uma proporção menor, mas também considerável entre populações dentro dos grupos (linguísticos ou ecorregionais) e por fim com a menor parte da variação explicada pelas diferenças entre grupos linguísticos ou ecorregiões. As análises dos marcadores uniparentais indicaram ainda que a estruturação genética inferida a partir dos marcadores mitocondriais é sempre superior à inferida para marcadores do cromossomo Y, apontando para a existência de fluxo gênico assimétrico com maior contribuição de homens e possivelmente predominância da matrilocalidade. Entretanto, não podemos comparar diretamente a estruturação genética inferida a partir da nossa análise dos dados autossômicos e essa análise de dados uniparentais, visto que não é possível estimar a variação genética dentro dos indivíduos a partir de marcadores haplóides, caso dos marcadores uniparentais. De qualquer forma o padrão e a proporção de distribuição da variação nos diferentes níveis hierárquicos é bastante similar, com a maior parte concentrada dentro das populações (variação dentro e entre amostras na Tabela 1), uma porção também grande entre populações e uma fração menor entre grupos linguísticos ou ecorregiões.

Além da dificuldade de se separar os efeitos das variáveis culturais e ambientais sobre a variação genética, para se estudar a relação entre a diversidade genética, cultural e ambiental, bem como as interações e trocas culturais entre grupos de diferentes regiões americanas, existem ainda limitações acerca do conhecimento sobre a cultura e os ambientes onde esses povos originalmente viviam. Devido a natureza fragmentária, incompleta e dispersa dos registros se faz necessária a integração de dados genéticos, arqueológicos, geológicos, paleobotânicos, históricos, etnográficos, entre outros, sobretudo no que se refere a indivíduos antigos e a populações contemporâneas de regiões muito impactadas pela colonização o que ocasionou muitas vezes a perda do contato com a sua herança cultural e/ou com seus territórios.

[89] Com exceção da classificação em biomas (COSTA et al., 2018) que foi exclusivamente testada nesta tese.

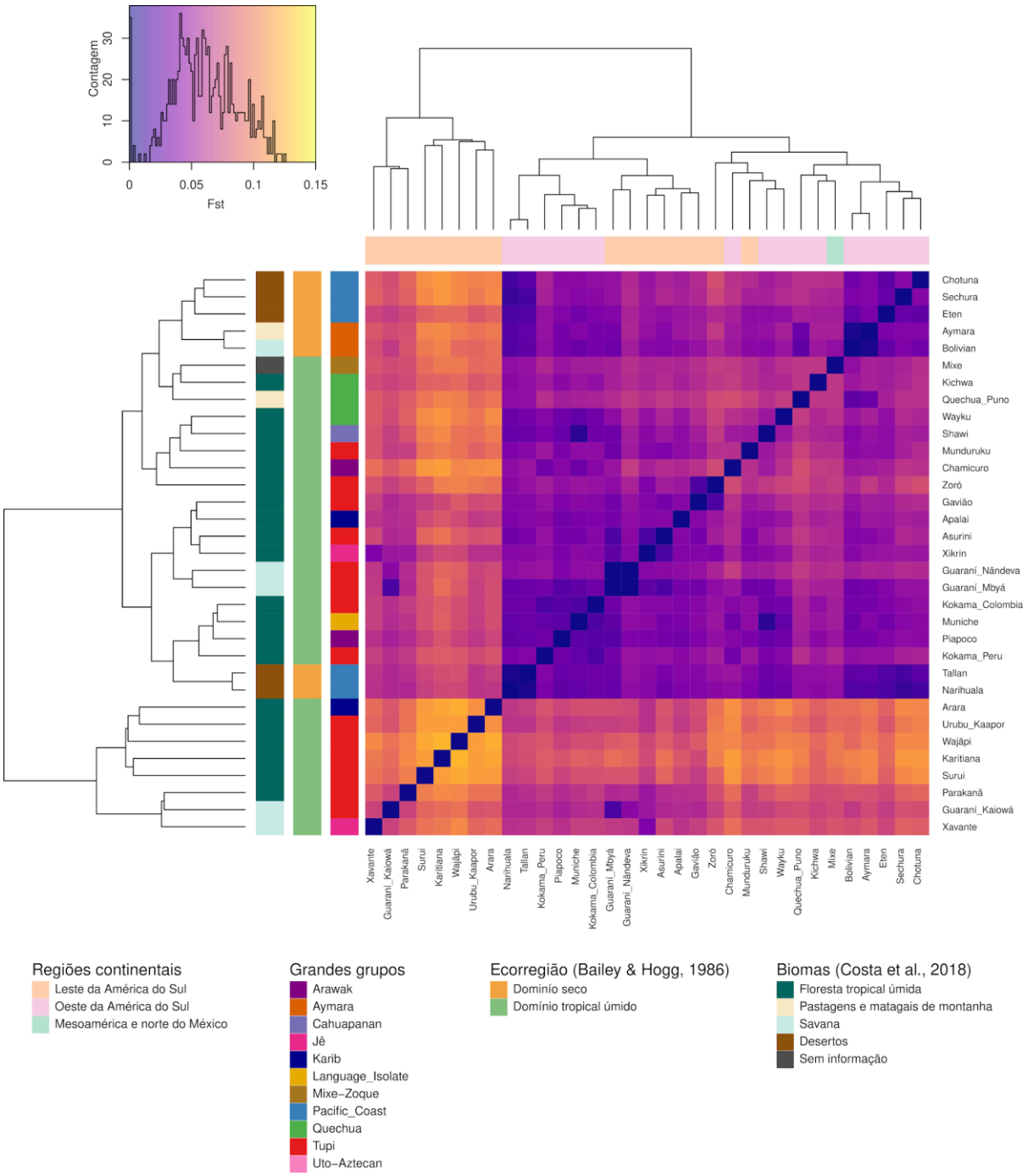


Figura 3 - Relação entre distância genética, geografia, cultura e ambiente. Uma matriz simétrica de F_{ST} par a par foi estimada com o pacote 'SNPRelate' (ZHENG et al., 2012) para todos os grupos no conjunto de dados de indivíduos não-miscigenados e não-aparentados e plotado na forma de mapa de calor com agrupamento hierárquico de linhas e colunas realizado pelo método padrão da função 'hclust' do pacote 'stats' (i.e. *complete linkage*). Os valores de F_{ST} são codificados de acordo com a legenda no canto superior esquerdo e a distribuição dos valores é mostrada na forma de um histograma sobre a legenda (eixo x: F_{ST} e eixo y: contagem). Adicionalmente o agrupamento hierárquico é apresentado na forma de dendogramas a esquerda e acima do gráfico. Por fim, as regiões continentais e grandes grupos (etnolinguísticos) como definidos nesta tese, assim como classificações em ecorregiões (BAILEY; HOGG, 1986) e biomas (COSTA et al., 2018), são codificados por cores na parte esquerda e superior do gráfico, como indicado na legenda abaixo do gráfico.

Tabela 1 - Resultados das análises de AMOVA. Análises independentes de AMOVA (EXCOFFIER; SMOUSE; QUATTRO, 1992) foram conduzidas através do pacote 'poppr' (KAMVAR; TABIMA; GRÜNWALD, 2014; KAMVAR; BROOKS; GRÜNWALD, 2015), sob o conjunto de indivíduos não-miscigenados e não-aparentados, removendo SNPs com dados faltantes e com correlação par a par maior do que 1% (selecionando 10.236 SNPs), testando diferentes classificações etnolinguísticas (**A-C**) e ecogeográficas (**D-E**) para as populações incluídas nesta tese.

Agrupamentos	Variação (%)	p-valor[1]	Variação relativa (%)[2]
A. Grandes grupos			
Variação entre grandes grupos	1,0811	0.004*	0,1622
Variação entre populações dentro dos grandes grupos	2,4328	0.001*	0,3651
Variação entre amostras dentro das populações	3,1501	0,148	0,4727
Variação dentro das amostras[3]	93,3360	0.002*	
Variação total	100,00		
B. Greenberg and Ruhlen et al., 2007			
Variação entre grupos linguísticos	1,1754	0.001*	0,1751
Variação entre populações dentro dos grupos linguísticos	2,3892	0.001*	0,3559
Variação entre amostras dentro das populações	3,1485	0,155	0,4690
Variação dentro das amostras	93,2869	0.002*	
Variação total	100,00		
C. Eberhard et al., 2020 (Ethnologue)[4]			
Variação entre grupos linguísticos	0,8843	0.006*	0,1339
Variação entre populações dentro dos grupos linguísticos	2,5666	0.001*	0,3887
Variação entre amostras dentro das populações	3,1522	0,148	0,4774
Variação dentro das amostras	93,3969	0.002*	
Variação total	100,00		
D. Bailey & Hogg, 1986 (Ecorregiões)			
Variação entre ecorregiões	0,6253	0.018*	0,0906
Variação entre populações dentro das ecorregiões	3,1318	0.001*	0,4539
Variação entre amostras dentro das populações	3,1422	0,173	0,4554
Variação dentro das amostras	93,1007	0.002*	
Variação total	100,00		
E. Costa et al., 2018 (Biomas)			
Variação entre biomas	0,1773	0,159	0,0271
Variação entre populações dentro dos biomas	3,2011	0.001*	0,4900
Variação entre amostras dentro das populações	3,1546	0,174	0,4829
Variação dentro das amostras	93,4670	0.005*	
Variação total	100,00		

*Valores estatisticamente significativos (p-valor < 0.05); [1]Significância estimada com o teste de randomização feito por meio do pacote 'ade4'; [2] Variação relativa = Variação de um dado nível hierárquico / (Variação total - Variação dentro das amostras); [3]Genótipos são separados em haplótipos e usados para calcular a variação dentro dos indivíduos (equivalente a configuração padrão do programa Arlequin); [4]A classificação linguística de Eberhard et al., 2020 (*Ethnologue*) apresenta as mesmas divisões que Hammarström et al., 2021 (*Glottolog*).

Por sua vez, é preciso também considerar que a diversidade cultural nem sempre está diretamente relacionada à diversidade genética, isto porque os mecanismos de herança e as taxas de evolução são distintos, além de interações e trocas culturais nem sempre resultarem

em fluxo gênico (FEHREN-SCHMITZ, 2020). Em vista disso, nós identificamos um caso bastante emblemático desse fenômeno onde há uma discordância entre os padrões de similaridade genética e cultural, no qual o grupo dos Kokama falantes de uma língua Tupí-Guaraní do oeste Amazônico se apresenta geneticamente mais próximo a grupos das encostas andinas orientais e sobretudo aos Arawak, visto que podem inclusive ser modelados com bom ajuste estatístico como grupo irmão dos Chamicuro (falantes de Arawak) (Figuras 1-4 e Figuras Suplementares 11-12 do Capítulo 2). Como discutido no Capítulo 2 nossos resultados indicam que os Kokama sofreram um processo de substituição linguística, de forma que eles muito provavelmente eram originalmente falantes de línguas Arawak que posteriormente adotaram uma língua Tupí-Guaraní, corroborando uma hipótese precedente sobre a origem do grupo, a qual apontava para uma cronologia pré-contato desse evento da substituição linguística (MICHAEL, 2014). Essa cronologia é suportada pela observação de que tanto os Kokama do Peru quanto os da Colômbia apresentam um perfil genético bastante similar, de modo que seria muito improvável que populações tão distantes geograficamente tivessem sofrido uma homogeneização genética com os mesmos grupos como consequência do contato com europeus e/ou dos processos de êxodo rural e adensamento populacional nos perímetros urbanos mais recentemente (TARAZONA-SANTOS et al., 2001; SANTOS, 2020).

Esse evento de substituição é muito relevante por indicar o quanto as características culturais dos diferentes grupos podem afetar a probabilidade de que ocorram interações entre diferentes povos, inclusive trocas culturais bastante consideráveis, sem que haja necessariamente em contrapartida qualquer vestígio genético das mesmas. Uma vez que características como comportamentos reprodutivos e mobilidade podem variar bastante entre grupos, o conhecimento acerca desses aspectos culturais pode assim guiar desde estratégias de amostragem mais adequadas e efetivas para estudar cada grupo etnolinguístico específico, até ajustar as hipóteses e expectativas sobre como provavelmente ocorreram as dinâmicas e histórias populacionais. Sob esse ponto de vista, os Arawak parecem reunir um conjunto de características que os tornam especialmente propensos a esse tipo de trocas culturais, como exemplificados pela prática da exogamia, pela natureza territorialmente expansiva dos Arawak evidenciada pelas extensas redes de comércio que estabeleceram nos rios amazônicos, as quais foram impulsionadas pela grande mobilidade, amplificada pelo uso de canoas para o deslocamento fluvial (HORNBORG, 2005; HILL, 2009).

Além da divisão Andes-Amazônia: demografia e estrutura genética

Dentre os fatores culturais influenciadores da diversidade genética as estratégias de subsistência e construção de nicho exercem um papel fundamental, visto que são um dos elementos principais da determinação da capacidade de suporte de uma dada população e logo também da taxa de crescimento populacional da mesma, por consequência influenciam se esta tende a uma estabilização populacional ao longo do tempo e do espaço ou a uma expansão territorial com o passar do tempo, em conjunto com outros fatores, como a coevolução gene-cultura (HÜNEMEIER et al., 2012b). Assim sendo, a transição dos grupos caçadores-coletores para o modo de vida agriculturalista levou num primeiro momento a uma aceleração da taxa de crescimento populacional (GOLDBERG; MYCHAJLIW; HADLY, 2016) e consequentemente a um aumento no movimento de populações humanas (LOOG et al., 2017; DELSER et al., 2021), o que em alguns casos se traduziu em eventos de expansão dêmica responsáveis por reconfigurar consideravelmente a paisagem de diversidade genética e cultural, não apenas da América do Sul como também de todo o globo (SOKAL; ODEN; WILSON, 1991; CORDAUX, 2004; WEN et al., 2004; DE FILIPPO et al., 2012; AMMERMAN; CAVALLI-SFORZA, 2014).

Dessa forma, os centros culturais onde primeiro se intensificaram a sedentarização e a produção de alimentos teriam passado por um processo de crescimento populacional, expansão dêmica, homogeneização genética e cultural (GOLDBERG; MYCHAJLIW; HADLY, 2016; PEREZ; POSTILLONE; RINDEL, 2017; SUTTER, 2021), eventualmente levando ao surgimento de sociedades agrícolas altamente hierarquizadas em alguns desses centros (HARRIS et al., 2018; BORDA et al., 2020). Esse processo ocorreu na América do Sul sobretudo nas regiões centrais e setentrionais andinas, onde culminou com a formação dos impérios Tiwanaku, Wari e Inca (HARRIS et al., 2018; BORDA et al., 2020), os quais acentuaram ainda mais o processo de homogeneização como consequência da construção de um intrincado sistema de estradas que permitia não apenas o fluxo de produtos alimentícios e culturais como também de pessoas, facilitando assim o contato entre diferentes grupos e proporcionando oportunidades para o fluxo gênico.

Por outro lado, a Amazônia seria habitada na maior parte por populações pequenas e isoladas de caçadores-coletores, entretanto hoje se sabe que a domesticação de plantas e o desenvolvimento da produção de alimentos também ocorreu na região amazônica e deu origem a estratégias de subsistência híbridas, como a policultura agroflorestal que envolvia tanto o cultivo de plantas domesticadas, manejo de plantas semi-domesticadas e coleta de alimentos nos ambientes florestais (NEVES, 2013; IRIARTE et al., 2020). Essa produção de alimentos apesar de não apresentar o mesmo grau de intensificação daquela observada nos Andes,

também teria levado a expansões dêmicas de alguns grupos, como os Arawak e os Tupí-Guaraní, conforme apontado por evidências arqueológicas e linguísticas (LATHRAP, 1970; BROCHADO, 1984; NOELLI, 2008; CORRÊA, 2014; GREGORIO DE SOUZA; ALCAINA MATEOS; MADELLA, 2020) e também indicado por nossos dados que evidenciam a ocorrência de aumento populacional durante o Holoceno tardio .

Nessa linha, nossos resultados do Capítulo 3 indicando uma maior similaridade dos Tupiniquim da costa atlântica brasileira com populações Tupí amazônicas e sobretudo com grupos de falantes de línguas Tupí-Guaraní, inclusive apresentando um bom ajuste estatístico quando modelados como um grupo irmão desses últimos (e.g. Urubu Ka'apor) (Figura 4 e Figuras suplementares 3-5;11;12-14;21-28;32 do Capítulo 3). Fundamentalmente, essas observações demonstram que os Tupiniquim e outros Tupí da costa tiveram origem em ancestrais Tupí amazônicos e que a dispersão das línguas e cultura material dos Tupí-Guaraní no passado seria explicada pela dispersão dos povos falantes dessas línguas e não apenas pela difusão cultural delas. Com efeito, modelos de história populacional foram gerados de acordo as hipóteses de Expansão Tupí propostas por Métraux (MÉTRAUX, 1927) e Brochado (BROCHADO, 1984), e o ajuste estatístico dos modelos foi estimado, consistentemente demonstrando que modelos estruturados de acordo com a hipótese de Brochado possuem bom ajuste e/ou explicam consideravelmente melhor os dados das populações amostradas (Figura 4 e figuras suplementares 25-28 do Capítulo 3).

Resumidamente, segundo a hipótese de Brochado os Tupí-Guaraní teriam se expandido a partir da Amazônia devido ao crescimento populacional proporcionado pela produção de alimentos a partir da agricultura (BROCHADO; LATHRAP, 1982; BROCHADO, 1984), se expandindo a partir de um centro de origem provavelmente localizado no sudeste da Amazônia (entre os rios Xingu e Tocantins) (ALMEIDA; NEVES, 2015), de onde um ramo foi em direção ao sul e a bacia do Paraná formando os Guaraní e outro ramo acompanhando inicialmente o rio Amazonas se dirigiu ao litoral atlântico, o qual povoou até a região de Cananéia, dando origem aos Tupí-Guaraní da costa (Figura 4 - pontos 36-38 e setas verdes; Figura 1 do Capítulo 3). Por outro lado a hipótese de Métraux propunha uma migração como explicação para a dispersão dos Tupí-Guaraní (Figura 1 do Capítulo 3), que teria sido motivada por mudanças climáticas de acordo com interpretações posteriores, segundo as quais uma redução e fragmentação das florestas a aproximadamente 2.000 AP teria impelido a dispersão (MEGGERS, 1974, 1977, 1982; MEGGERS AND CLIFFORD, 1978). Entretanto uma análise mais recente de dados paleoecológicos e paleoclimáticos evidencia um cenário diferente (IRIARTE et al., 2017), no qual o clima mais úmido do Holoceno tardio teria produzido uma expansão das florestas no hemisfério sul da América do Sul durante o período de 3.000 a 2.000 AP, particularmente a expansão das florestas ribeirinhas teria assim criado uma oportunidade ecológica para a

expansão dos Tupí-Guaraní, por fornecer as condições ambientais necessárias para a produção de alimentos por meio da policultura florestal a que eles estavam adaptados e possivelmente também contribuído para fomentar o ethos[90] expansivo do grupo. Sendo assim, estas novas evidências são compatíveis com o modelo de Brochado, inclusive cronologicamente, ao mesmo tempo que evidenciam a existência de fatores adicionais que teriam facilitado ou possivelmente desencadeado o processo de expansão.

Desse modo, o fato da hipótese de Brochado ser suportada pelos dados genômicos e de que esses dados indicam uma maior proximidade dos Tupiniquim com os Tupí-Guaraní amazônicos dentre todos os nativos brasileiros, em conjunto com as demais evidências (e.g. arqueológicas, linguísticas, paleoclimáticas e paleoecológicas) corrobora fortemente que a Expansão Tupí (NOELLI, 2008; CORRÊA, 2014) foi uma expansão dêmica e dessa forma soma-se ao corpo crescente de evidências que apontam para a ocorrência de expansões dêmicas com origem na Amazônia (GREGORIO DE SOUZA; ALCAINA MATEOS; MADELLA, 2020). Esses dados vão de encontro a algumas premissas do modelo de uma divisão genética e cultural entre os Andes e a Amazônia, o qual assumia que as populações amazônicas seriam predominantemente pequenas e isoladas (TARAZONA-SANTOS et al., 2001; FUSELLI et al., 2003; BARBIERI et al., 2014b; SANTOS, 2020).

O modelo da divisão Andes/Amazônia tinha como pressuposto a existência de forças evolutivas contrastantes entre as duas regiões, dado que por um lado os Andes seriam historicamente habitados por populações grandes e interconectadas, ao passo que na Amazônia, como dito antes, as populações seriam pequenas e isolados (TARAZONA-SANTOS et al., 2001; FUSELLI et al., 2003; BARBIERI et al., 2014b; SANTOS, 2020). De fato no Capítulo 2 nós inferimos um N_e mediano durante as últimas 100 gerações de 28.000 para os grupos do oeste sul-americano e de 9.280 para os do leste (Figura 7A and Figura suplementar 18 do Capítulo 2), suportando o histórico de um maior tamanho populacional nos Andes, ao menos nesse período do Holoceno tardio. Essas diferenças demográficas seriam as responsáveis por uma maior diversidade genética nos Andes e oeste da América do Sul e uma menor diversidade na Amazônia e leste do continente (TARAZONA-SANTOS et al., 2001; FUSELLI et al., 2003; BARBIERI et al., 2014b; SANTOS, 2020), um padrão que também é corroborado pelos nossos resultados do Capítulo 2, que indicam uma diferença significativa no nível de diversidade entre leste e oeste da América do Sul, como inferido pelo Teste U de Mann-Whitney dos coeficientes de endogamia F_{ROH} dos grupos das duas regiões (Figura 8A e figura suplementar 16 do Capítulo 2), adicionalmente são observados níveis muito superiores de IBD e HBD no leste do continente em comparação com o oeste (Figura suplementar 17 do Capítulo 2). Apesar disso, nós

[90] Conjunto de traços comportamentais e hábitos característicos de um determinado grupo.

demonstramos que o F_{ROH} é correlacionado a longitude (Figura 8C do Capítulo 2), o que sugere uma transição gradual dos padrões demográficos e das zonas de influência entre grupos andinos e amazônicos, e dessa forma contrapõe uma separação abrupta entre eles. Uma segunda expectativa do modelo de divisão Andes/Amazônia seria que a diferenciação genética entre grupos amazônicos é maior do que aquela observada entre os andinos, devido a um menor fluxo gênico entre os primeiros. Efetivamente os maiores valores de distância genética são estimados para grupos amazônicos (ver valores de F_{ST} na Figura 3 desta discussão, e comprimentos de ramo da Figura 2 e da Figura suplementar 10 do Capítulo 2), indicando uma maior diferenciação genética dos mesmos. Contudo AMOVAs independentes (do mesmo conjunto de dados usados para obter os resultados da Tabela 1) estimam que a fração de variação entre populações no leste da América do Sul (2,88%; p = 0.001)[91] é menor do que a observada entre populações do oeste (3,44%; p = 0.001)[92], contrariando assim as expectativas do modelo, porém a análise compara diversos grupos do leste e oeste do continente e não apenas populações andinas e amazônicas, além disso a porcentagem de variação estimada é proporcional à variabilidade total que diverge entre as duas regiões, a qual é significativamente maior no oeste como demonstrado no Capítulo 2 (Figura 8A e figura suplementar 16 do Capítulo 2).

Em adição ao já discutido para a distribuição da variabilidade genética, uma série de outras evidências também argumentam contrariamente à existência de uma divisão abrupta entre populações dos Andes e da Amazônia, e também contrapõe parcialmente o modelo proposto segundo o qual os históricos demográficos e evolutivos antagônicos entre as duas regiões teriam levado a formação de padrões genéticos contrastantes, como por exemplo a premissa de que populações amazônicas seriam historicamente pequenas e isoladas, o que teria resultado em uma menor variação genética intrapopulacional e maior diferenciação genética entre populações. As principais evidências são:

(i) Primeiramente, a ocorrência de crescimento populacional e expansões dêmicas no Holoceno tardio evidenciada pelos resultados do Capítulo 3 sobre os Tupí-Guaraní da costa atlântica e do Centro-Oeste e Sul do Brasil, além do histórico de tamanho efetivo populacional de diversos grupos linguísticos, estimados no Capítulo 2, que evidenciam a ocorrência de crescimentos populacionais significativos nas últimas 100 gerações, como exemplificado pelos Tupí que apresentam um crescimento populacional de 98,92%[93] (N_e máximo e mínimo de 343.000 e

[91] O restante da variação se distribui em 3,32% (p = 0.226) entre as amostras e 93,8% (p = 0.016) dentro das amostras.

[92] 3,41% (p = 0.29) entre as amostras e 93,15% (p = 0.032) dentro das amostras.

[93] Esses são os resultados obtidos pela análise conduzida com IBDNe (BROWNING et al., 2018) utilizando o parâmetro "*filtersamples = false*", o qual determina se indivíduos aparentados devem ser removidos ou não. Nesta análise buscamos reduzir o efeito do pequeno tamanho amostral, entretanto é esperado que a

3.690 respectivamente) no período entre 100 e 20 GP (desconsiderando o período pós-contato) (Figura 7A do Capítulo 2). Os quais demonstram que a Amazônia não necessariamente era habitada apenas por populações pequenas e isoladas como era descrito até recentemente, o que somado a outros estudos recentes indica uma maior densidade populacional e complexidade das sociedades que habitavam a região (HECKENBERGER et al., 2003; HECKENBERGER; NEVES, 2009; ROOSEVELT, 2013; DE SOUZA et al., 2019; BERESFORD-JONES; MURILLO, 2020), e inclusive demonstram a ocorrência de expansões para fora da Amazônia, provavelmente desencadeadas por mudanças climáticas e pelo aumento da taxa de crescimento populacional proporcionado pela produção de alimentos durante o Holoceno tardio.

(ii) Em segundo lugar, a distribuição dos componentes putativos de ancestralidade indica o compartilhamento de ancestralidade entre populações da costa do Pacífico, Andes e Amazônia ocidental (Figura 1-2 e figuras suplementares 5-6 e 8-9 do Capítulo 2), corroborando observações anteriores no mesmo sentido (HARRIS et al., 2018; BARBIERI et al., 2019; GNECCHI-RUSCONE et al., 2019b), evidenciando portanto a permeabilidade ao fluxo gênico entre as duas regiões.

(iii) Adicionalmente, um dos componentes principais de variação se alinha e é correlacionado a longitude, indicando uma variação contínua entre o oeste e leste da América do Sul, sem indícios de uma separação entre as duas regiões (Figura 2C-D do Capítulo 2). O MDS das distâncias genéticas também suporta as mesmas interpretações que os resultados de PCA (Figura 4A e 6 do Capítulo 2).

(iv) Por fim, grupos do leste e oeste da América do Sul compartilham segmentos de IBD formados no período pré-contato (Figura 5 do Capítulo 2), algo também demonstrado anteriormente (BARBIERI et al., 2019), indicando que mesmo antes da chegada dos europeus, e do processo de disrupção das sociedades indígenas, ocorreram dispersões populacionais entre as duas regiões ou possivelmente também a existência de fluxo gênico entre elas, apesar do compartilhamento de segmentos ser maior entre grupos da mesma região.

Tomados em conjunto, esses resultados demonstram que a distribuição da diversidade genética no continente sul-americano é mais complexa do que poderia ser explicada por um modelo de divisão entre os Andes e Amazônia, ou mesmo entre o oeste e o leste do continente, de modo que é possível identificar a existência 3 a 4 grupos principais com maior similaridade genética em cada região (Figura 1 e Figura suplementar 9 do Capítulo 2). No oeste basicamente se dividem as porções norte, sul e central dos Andes, com as populações da costa pacífica se agrupando com os últimos, no leste é possível identificar o grupo dos Guaraní, dos Jê, e um

presença de indivíduos aparentados introduza um viés especialmente estimativas de N_e das gerações mais recentes, por isso apresentamos também o resultado da análise com o parâmetro "*filtersamples = true*" (Figura suplementar 18 do Capítulo 2), no qual o aumento populacional estimado para os Tupí no período é de 72,77% (Ne máximo e mínimo de 29.200 e 7.950 respectivamente).

terceiro componente maximizado nas populações do sudeste amazônico (Tupí e Karib), por fim os grupos do oeste amazônico incluindo aqueles das encostas orientais andinas (Tupí, Arawak, Quechua, Cahuapanos e isolados linguísticos) parecem ser constituídos por uma mistura entre ancestralidades do leste e oeste (Figura 1 e Figura suplementar 9 do Capítulo 2). O padrão de estrutura genética identificado é essencialmente o mesmo inferido por outros estudos recentes (BARBIERI et al., 2019; GNECCHI-RUSCONE et al., 2019a).

Como discutido no Capítulo 3 é interessante notar que os Guaraní se configuram como um clado[94] com considerável diferenciação genética em relação aos demais nativos brasileiros e inclusive aos outros Tupí (e.g. Figura 2A do Capítulo 2; Figura 4 e figura suplementar 11 do Capítulo 3), além disso apresentam grande afinidade genética mútua (Figura 4B e figura suplementar 12 do Capítulo 2) e significativo compartilhamento de segmentos de IBD em todas as categorias de comprimento (Figura 5 e figura suplementar 13 do Capítulo 2), um conjunto de evidências que indicam uma história populacional independente e algum grau de isolamento em relação aos outros grupos Tupí desde a divergência, o que também corrobora a hipótese de que foram formado por um ramo distinto dos demais Tupí-Guaraní, o qual teria se expandido em direção ao Sul segundo Brochado (BROCHADO, 1984). A provável mistura com grupos do Chaco[95] (Figura 3B-D e figura suplementar 11), certamente, também contribuiu para a maior diferenciação genética dos Guaraní em relação aos outros Tupí.

Por outro lado, os Tupí e Karib amazônicos não se apresentam diferenciados geneticamente (e.g. Figuras 1, 2 e 4 e figuras suplementares 5-9 do Capítulo 2; Figura suplementar 3-6 do Capítulo 3), sugerindo a ocorrência de fluxo gênico entre esses grupos apesar das diferenças linguísticas e culturais, e consequentemente a existência de mecanismos capazes de manter a identidade ou continuidade cultural frente ao contato e fluxo de migrantes entre diferentes grupos étnicos. Entretanto, existe a possibilidade de que não haja resolução o suficiente para identificar a estrutura genética existente entre os Tupí e Karib amazônicos, por isso no futuro a possível disponibilidade de genomas completos de indivíduos desses grupos pode ajudar a esclarecer essa questão.

Por fim, entre os nativos brasileiros, os grupos Jê de modo geral apresentam a maior afinidade genética com os indivíduos antigos do Brasil, sobretudo com os do Holoceno inicial (Figura suplementar 14 do Capítulo 2; Figuras suplementares 15-20 do Capítulo 3), além de apresentarem um componente de ancestralidade exclusivo (Figuras 1, 2 e 4 e figuras suplementares 5-9 do Capítulo 2), o que sugere que eles são formados por um ramo diferente dos outros nativos brasileiros, possivelmente mais basal, e apresentando uma maior

[94] Também chamado de grupo monofilético, essencialmente é um grupo no qual estão incluídos um ancestral comum e todos os seus descendentes, caso contrário um grupo é chamado de parafilético.

[95] Região quente e semi árida da bacia do Rio da Prata incluindo territórios da Bolívia, Paraguai, Argentina e dos estados brasileiros do Mato Grosso e Mato Grosso do Sul.

contribuição genética (i.e. continuidade genética) dos grupos antigos do leste da América do Sul. Dessa forma o estudo genético dos grupos Jê é promissor e pode revelar aspectos importantes sobre o povoamento do leste da América do Sul.

O padrão de IBD inferido revela ainda que as populações do leste do continente sul-americano compartilham um maior número e comprimento total de segmentos, o que mais parcimoniosamente deve ser explicado pela ocorrência de expansões dêmicas mais recentes nessa região, ao passo que no oeste, seguindo a mesma lógica, as expansões teriam ocorrido num passado mais distante, uma hipótese que se alinha à aceleração do crescimento populacional mais precoce entre as populações andinas (GOLDBERG; MYCHAJLIW; HADLY, 2016) e por consequência um crescimento populacional mais tardio na Amazônia, como indicado pela própria estimativa de início da Expansão Tupí no Holoceno tardio entre 3.000 a 2.000 AP (NOELLI, 2008; CORRÊA, 2014).

Em conclusão, o estudo das populações americanas, nativas e miscigenadas, vem revelando progressivamente a imensa diversidade de linhagens ancestrais que tiveram as Américas como ponto de encontro, algumas delas separadas a dezenas de milhares de anos como as europeias e africanas, trazidas com a colonização e o tráfico de escravo do atlântico (Figura 4 - pontos 39 e 40), respectivamente, e outras separadas a menos tempo, embora ainda no final do Pleistoceno durante o povoamento do nordeste asiático, Sibéria, Beríngia e entrada nas Américas, como por exemplo as NNA e SNA (Figura 4 - pontos 16). Além de outras populações antigas não amostradas até o momento que contribuíram com linhagens que divergiram a ainda mais tempo, como as linhagens australo-asiáticas, possivelmente formadas durante o povoamento do sudeste asiático (Figura 4 - pontos 10 e 11). Por fim, dinâmicas e dispersões populacionais, sobretudo no Holoceno médio e tardio, provocadas tanto por mudanças climáticas quanto pela transição progressiva a sedentarização e a intensificação da produção de alimentos, reconfiguram consideravelmente os padrões de distribuição da diversidade genética nativa americana através de movimentos internos inter e intra-continentais, sendo a Expansão Tupí provavelmente o evento mais impactante para o leste da América do Sul (Figura 4 - pontos 36-38).

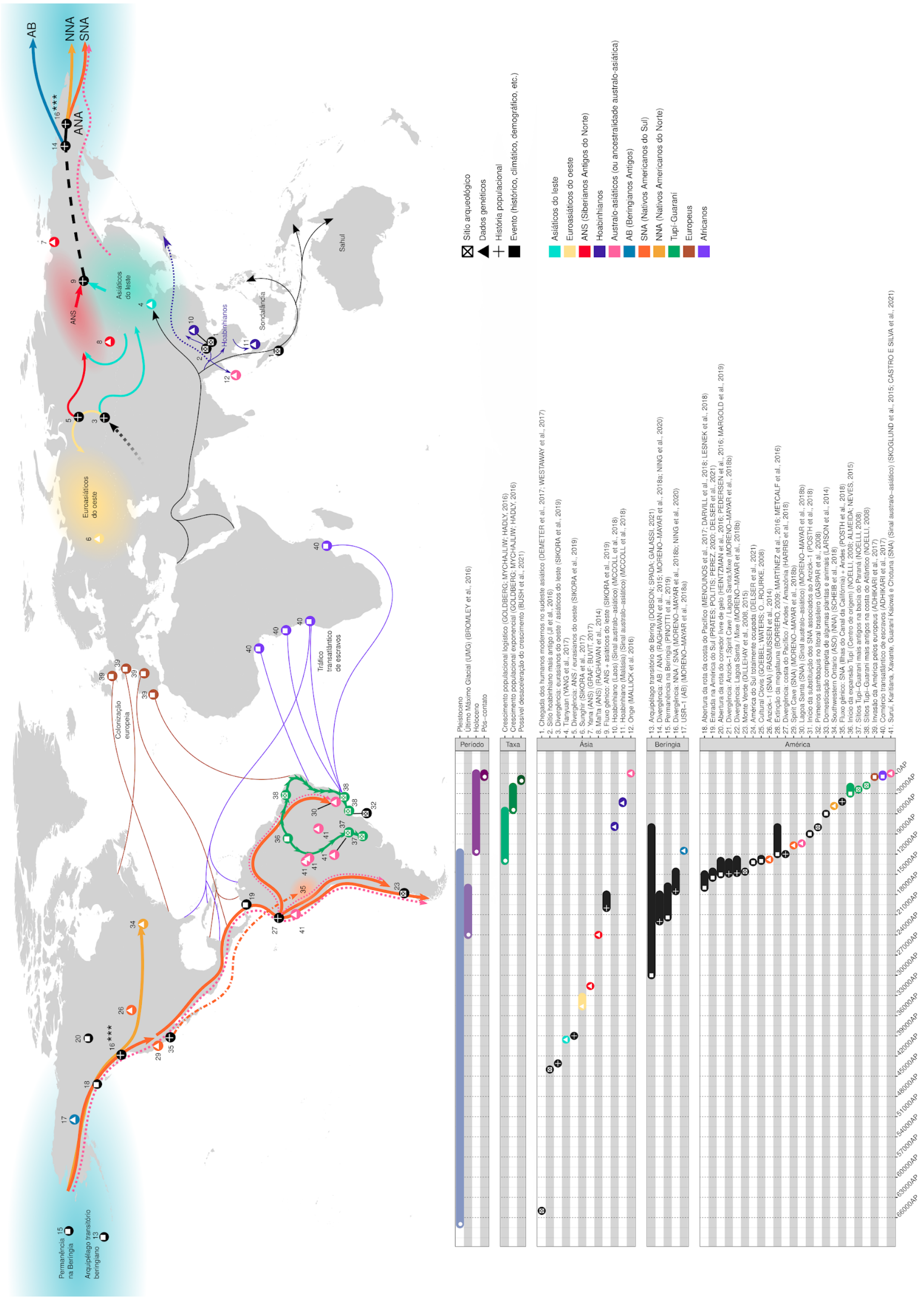

AB
NNA
SNA
ANA
16 ***
ANS
Asiáticos do leste
Euroasiáticos do oeste
Hoabinhianos
Sondalândia
Sahul
Colonização europeia
Tráfico transatlântico de escravos
Permanência na Beríngia
Arquipélago transitório beringiano
Período
Taxa
Ásia
Beríngia
América
Pleistoceno
Último Máximo Glacial (UMG) (BROMLEY et al., 2016)
Holoceno
Pós–contato
Crescimento populacional logístico (GOLDBERG; MYCHAJLIW; HADLY, 2016)
Crescimento populacional exponencial (GOLDBERG; MYCHAJLIW; HADLY, 2016)
Possível desaceleração do crescimento (BUSH et al., 2021)
1. Chegada dos humanos modernos no sudeste asiático (DEMETER et al., 2017; WESTAWAY et al., 2017)
2. Sítio hoabinhiano mais antigo (JI et al., 2016)
3. Divergência: eurasianos do oeste / asiáticos do leste (SIKORA et al., 2019)
4. Tianyuan (YANG et al., 2017)
5. Divergência: ANS / eurasianos do oeste (SIKORA et al., 2019)
6. Sunghir (SIKORA et al., 2017)
7. Yana (ANS) (GRAF; BUVIT, 2017)
8. Mal'ta (ANS) (RAGHAVAN et al., 2014)
9. Fluxo gênico: ANS + asiáticos do leste (SIKORA et al., 2019)
10. Hoabinhiano (Laos) (Sinal australo–asiático) (MCCOLL et al., 2018)
11. Hoabinhiano (Malásia) (Sinal australo–asiático) (MCCOLL et al., 2018)
12. Onge (MALLICK et al. 2016)
13. Arquipélago transitório de Bering (DOBSON; SPADA; GALASSI, 2021)
14. Divergência: AB / ANA (RAGHAVAN et al., 2015; MORENO–MAYAR et al., 2018a; NING et al., 2020)
15. Permanência na Beríngia (PINOTTI et al., 2019)
16. Divergência: NNA / SNA (MORENO–MAYAR et al., 2018b; NING et al., 2020)
17. USR–1 (AB) (MORENO–MAYAR et al., 2018a)
18. Abertura da rota da costa do Pacífico (MENOUNOS et al., 2017; DARVILL et al., 2018; LESNEK et al., 2018)
19. Entrada na América do Sul (PRATES; POLITIS; PEREZ, 2020; DELSER et al., 2021)
20. Abertura da rota do corredor livre de gelo (HEINTZMAN et al., 2016; PEDERSEN et al., 2016; MARGOLD et al., 2019)
21. Divergência: Anzick–1;Spirit Cave / Lagoa Santa;Mixe (MORENO–MAYAR et al., 2018b)
22. Divergência: Lagoa Santa / Mixe (MORENO–MAYAR et al., 2018b)
23. Monte Verde (DILLEHAY et al., 2008, 2015)
24. América do Sul totalmente ocupada (DELSER et al., 2021)
25. Cultural Clovis (GOEBEL; WATERS; O...ROURKE, 2008)
26. Anzick–1 (SNA) (RASMUSSEN et al., 2014)
28. Extinção da megafauna (BORRERO, 2009; MARTÍNEZ et al., 2016; METCALF et al., 2016)
27. Divergência: costa do Pacífico / Andes / Amazônia (HARRIS et al., 2018)
29. Spirit Cave (SNA) (MORENO–MAYAR et al., 2018b)
30. Lagoa Santa (SNA) (Sinal australo–asiático) (MORENO–MAYAR et al., 2018b)
31. Início da substituição dos SNA associados ao Anzick–1 (POSTH et al., 2018)
32. Primeiros sambaquis no litoral brasileiro (GASPAR et al., 2008)
33. Domesticação completa de algumas plantas e animais (LARSON et al., 2014)
34. Southwestern Ontario (ASO) (NNA) (SCHEIB et al., 2018)
35. Fluxo gênico: SNA (Ilhas do Canal da Califórnia) + Andes (POSTH et al., 2018)
36. Início da expansão Tupí (Centro de origem) (NOELLI, 2008; ALMEIDA; NEVES, 2015)
37. Sítios Tupi–Guarani mais antigos na bacia do Paraná (NOELLI, 2008)
38. Sítios Tupi–Guarani mais antigos na costa do Atlântico (NOELLI, 2008)
39. Invasão da América pelos europeus (ADHIKARI et al., 2017)
40. Comércio transatlântico de escravos (ADHIKARI et al., 2017)
41. Suruí, Karitiana, Xavante, Guaraní Kaiowá e Chotuna (SNA) (Sinal australo–asiático) (SKOGLUND et al., 2015; CASTRO E SILVA et al., 2021)
66000AP
63000AP
60000AP
57000AP
54000AP
51000AP
48000AP
45000AP
42000AP
39000AP
36000AP
33000AP
30000AP
27000AP
24000AP
21000AP
18000AP
15000AP
12000AP
9000AP
6000AP
3000AP
0AP
Sítio arqueológico
Dados genéticos
História populacional
Evento (histórico, climático, demográfico, etc.)
Asiáticos do leste
Euroasiáticos do oeste
ANS (Siberianos Antigos do Norte)
Hoabinhianos
Australo-asiáticos (ou ancestralidade australo-asiática)
AB (Beringianos Antigos)
SNA (Nativos Americanos do Sul)
NNA (Nativos Americanos do Norte)
Tupí-Guarani
Europeus
Africanos

Figura 4 - Sumário da história populacional e demográfica das Américas. Na linha do tempo apresentada na parte inferior da figura são listados os principais marcos da história das populações americanas (nativas e miscigenadas), ordenados de acordo com a data de início, bem como numerados e descritos, além de separados por continente (Painéis Ásia, Beríngia e América), sendo ainda classificados em (codificado nas formas dos pontos - legenda à direita): sítios arqueológicos, indivíduos antigos dos quais se obteve aDNA (dados genéticos), eventos da história populacional cujas datas foram inferidas a partir de análises de dados genéticos (divergências e fluxo gênico) e por fim eventos históricos, climáticos, demográficos entre outros, que possuem maior relevância. Além disso, as diferentes ancestralidades são identificadas pelas cores mostradas na legenda à direita. São apresentados também os períodos geológicos (Painel Período) e a taxa de crescimento populacional (Painel Taxa) entre o fim Pleistoceno tardio e o Holoceno tardio, inferida para as populações da América do Sul. No mapa mostrado na parte superior são representadas as localizações aproximadas de alguns dos principais marcos (pontos), em conjunto com as rotas (ou hipóteses) mais prováveis de dispersão (setas), entretanto essas não devem ser interpretadas de forma literal, pois também apenas indicam aproximadamente a direção desses movimentos. No caso dos imigrantes europeus e africanos são mostrados somente alguns dos principais pontos de origem e destino. ***A separação entre os ramos NNA e SNA provavelmente aconteceu ao sul das geleiras da América do Norte (WILLERSLEV; MELTZER, 2021), entretanto novas evidências apontam que o evento teria ocorrido mais parcimoniosamente ainda na Ásia (NING et al., 2020). Os eventos aqui listados são descritos e discutidos em maior profundidade na introdução da tese.

CONCLUSÕES

- As amostras inéditas publicadas nos artigos que compõem essa tese somam-se a um esforço recente para reduzir a sub-representação de populações nativas americanas nos painéis globais de dados genômicos, contribuindo principalmente para preencher uma lacuna de dados genômicos sobre a diversidade das populações do leste da América do Sul.

- As comunidades dos Tupiniquim e Guaraní Mbyá localizadas em Aracruz (ES) na costa atlântica são compostas por indivíduos com um grau de miscigenação superior ao observado em outras populações nativas da América do Sul. Essa maior proporção de miscigenação no litoral brasileiro aponta um impacto mais drástico do contato com os europeus e do processo de colonização para a região, algo esperado pela densidade populacional ter sido historicamente muito maior no litoral e regiões adjacentes. Entretanto, não há correlação entre o grau de miscigenação e a localização geográfica, ao menos no nível continental.

- O histórico de tamanho efetivo populacional do período pós-contato dos diferentes grupos indígenas americanos revela um maior impacto da colonização na porção leste da América do Sul e no México, apontando uma maior redução proporcional da diversidade nas populações do litoral atlântico brasileiro (Tupiniquim) e México, seguidas pelos Tupí (incluindo os Guaraní e Tupí Amazônicos), pelos Quechua, pelos Jê e por fim pelas comunidades da costa do Pacífico. O mínimo de diversidade genética é estimado há 7-8 gerações e demonstra que o processo de extermínio dos povos indígenas provavelmente se intensificou ao longo do tempo desde o contato.

- O compartilhamento de segmentos de IBD intrapopulacional (e HBD) indica um aumento da frequência de casamentos consanguíneos no período pós-contato, muito provavelmente causado pela redução e fragmentação das populações indígenas. Ao passo que o compartilhamento de IBD entre populações é reduzido no mesmo período, sugerindo um aumento da proporção de casamentos dentro dos grupos (i.e. intrapopulacionais).

- O excesso de afinidade genética com populações australo-asiáticas foi encontrado nos Guaraní Kaiowá do sul da região Centro-Oeste brasileira e nos Chotuna do litoral norte do Peru, indicando que a extensão da presença dessa ancestralidade é muito superior à anteriormente inferida, que estava restrita à Amazônia (Karitiana e Suruí) e ao planalto central brasileiro (Xávante). A pequena contribuição da ancestralidade australo-asiática para os grupos onde foi detectada (entre 1 e 3%) e a grande variabilidade do sinal entre populações e indivíduos sugerem ainda que a significância estatística pode estar sendo perdida por meio da diluição do sinal e sua presença, ocultada em alguns grupos.

- Outros fatores que podem contribuir para explicar a ausência do sinal australo-asiático na maioria das populações nativas americanas são:
    - Miscigenação com grupos de ancestralidade inteiramente nativa americana;

    - Efeito extremo da deriva genética com grande perda da diversidade;
    - Dinâmicas populacionais e expansões dêmicas que contribuíram para Homogeneização genética a nível continental e sub-continental;
    - Extinção de diversos grupos e drástica redução da diversidade genética como consequência da invasão e colonização europeia;
    - Sub-amostragem, sobretudo na América do Norte.

- A ancestralidade australo-asiática teria sido introduzida, mais parcimoniosamente, durante o povoamento inicial da América do Sul, o qual teria sido feito por uma população bastante estruturada geneticamente e tem como limite máximo de chegada ~12.000 AP, a data de divergência inferida entre os grupos da costa do Pacífico, Andes e Amazônia (HARRIS et al., 2018). Isto porque hipóteses alternativas requerem mais de um evento de miscigenação independente com grupos amazônicos e da costa do pacífico, resultando em proporções muito similares de ancestralidade australo-asiática. Além disso, nossos modelos de história populacional corroboram a hipótese de dispersão inicial pela rota da costa do Pacífico, o que também poderia ajudar a explicar a ausência do sinal australo-asiático na América Central e do Norte, principalmente considerando que estes primeiros grupos humanos adotariam estratégias de subsistência predominantemente pescadoras-coletoras e portanto adaptadas à vida costeira.

- A ancestralidade das comunidades indígenas atuais apresentam evidências de uma continuidade genética parcial em relação aos grupos que ocupavam as mesmas regiões ou regiões adjacentes no passado. Adicionalmente, o padrão de afinidade genética entre grupos contemporâneos e antigos indica a existência de um gradiente de variação genética que remonta aos eventos de povoamento inicial, com diferenciação no sentido norte-sul e oeste-leste. Indicando portanto que a distribuição da variação genética nativa americana foi significativamente influenciada pelo povoamento inicial das Américas.

- A correlação entre distâncias genéticas e geográficas, somadas a redução da diversidade genética do oeste para o leste da América do sul, como demonstrada pela correlação entre o coeficiente de endogamia e a longitude e pela diferença significativa entre o coeficiente de endogamia médio de grupos das duas regiões, evidencia que a variação genética clinal observada foi formada como consequência tanto do isolamento por distância quanto dos gargalos populacionais em série, provavelmente durante o povoamento inicial. Soma-se ainda a essas evidências o padrão inferido de divergência dos grupos ao longo da história populacional que se inicia no oeste e continua progressivamente em direção ao leste. Finalmente, são inferidas maiores distâncias genéticas entre grupos do leste da América do Sul.

- As diversidades linguística e ambiental são capazes de explicar apenas uma parte da variação genética e não é possível rejeitar a hipótese de que os acasalamentos são aleatórios entre indivíduos dos mesmos grupos, entretanto existe variação genética significativa entre os grupos, um padrão esperado para populações pequenas e isoladas, onde o efeito da deriva genética é maior e do fluxo gênico menor.

- Um conjunto robusto de evidências indica que os Kokama localizados no oeste amazônico e falantes de uma língua da família Tupí-Guaraní passaram por um processo de substituição linguística e são muito provavelmente descendentes de povos falantes de Arawak ou proximamente relacionados, visto que possuem uma afinidade genética muito maior com grupos do oeste amazônico e com os Arawak, em especial com os Chamicuro do Peru, do que com os Tupí ou mesmo com os outros Tupí-Guaraní. A grande similaridade do perfil de afinidades genéticas dos Kokama do Peru e dos Kokama da Colômbia corrobora uma cronologia pré-contato para o evento de substituição linguística, isto porque estes encontram-se a centenas de quilômetros de distância um do outro, o que torna improvável que movimentos e dinâmicas populacionais desencadeadas pela colonização ou eventos recentes de êxodo rural e adensamento em centros urbanos tenham gerado independentemente esses perfis genéticos similares, como consequência de uma possível homogeneização genética das populações da região.

- Os Tupiniquim do litoral atlântico localizados na região Sudeste brasileira (ES) são mais similares geneticamente a grupos Tupí amazônicos do que a outros nativos brasileiros, incluindo os próprios Guaraní (Tupí-Guaraní do Centro-Oeste e Sul brasileiros). Os Tupiniquim ainda apresentam bom ajuste como grupo irmão dos Tupí-Guaraní amazônicos em modelos de história populacional. Esse resultado possui algumas implicações importantes para a história dos Tupí-Guaraní:
    - Demonstra que a Expansão Tupí foi um processo de expansão dêmica, visto que os Tupiniquim, descendentes do ramo costeiro da expansão, são geneticamente mais próximos aos Tupí Guaraní amazônicos como seria esperado, o que ao mesmo tempo se contrapõe a possibilidade de que as línguas Tupí-Guaraní tenham se expandido apenas sob a forma de uma difusão cultural, caso em que seria esperada uma maior similaridade com outros grupos indígenas, provavelmente com os Jê;
    - Rejeita a hipótese proposta por Métraux (MÉTRAUX, 1927) e corrobora a hipótese proposta por Brochado (BROCHADO, 1984), segundo a qual os Tupí-Guaraní se expandiram para fora da amazônia devido a um aumento da taxa de crescimento populacional proporcionado pela agricultura no interior da Amazônia, ainda segundo esta hipótese os Tupí-Guaraní da costa (i.e. os Tupinambá) e os do sul (i.e. os Guaraní) foram formados por ramos distintos que teriam divergido no início da Expansão Tupí, entre 3.000 e 2.000 AP.

- Ao passo que os Guaraní do sul da região Centro-Oeste brasileira (MS), sobretudo os Mbyá e os Ñandeva, muito provavelmente foram formados por um evento de miscigenação entre o ramo sul da Expansão Tupí e populações do Chaco, e se apresentam como um grupo monofilético em relação aos outros Tupí e nativos brasileiros. Além disso, os Guaraní são mais relacionados aos Jê do que os outros Tupí, visto que os Guaraní Kaiowá podem ser modelados como grupo irmão dos Xávante, e os outros Guaraní como uma mistura de um ramo irmão dos Kaiowá e de uma outra linhagem mais basal, a qual ao que tudo indica seria proveniente de um grupo do Chaco.

- Os padrões de diversidade genética encontrados corroboram a existência de diferenças

entre o leste e oeste da América do Sul, indicando um histórico de maior tamanho efetivo e um menor coeficiente de endogamia no oeste em comparação com o leste, enquanto que o número e comprimento total dos segmentos de IBD (HBD, intra e interpopulacional) apresentam o padrão oposto, ou seja maiores níveis no leste e menores no oeste. Além disso, a distância genética entre grupos também é maior no leste, apesar da estimativa aparentemente contraditória de uma maior porcentagem da variação genética entre grupos do oeste.

- Em contrapartida, um modelo de divisão abrupta entre os Andes e a Amazônia é refutado por diversas linhas de evidência, entre elas: a ocorrência de expansões dêmicas com centro de dispersão na Amazônia, o compartilhamento de componentes de ancestralidade e de segmentos de IBD formados no período pré-contato entre grupos da costa do Pacífico, Andes e Amazônia, assim como a correlação do gradiente de variação genética e do nível de diversidade genética com a longitude.

- Por fim, foram identificados 4 grupos principais de similaridade genética entre os nativos das terras baixas do leste sul-americano: os Guaraní, os Jê, um grupo incluindo toda porção central e leste da Amazônia, e último grupo localizado no oeste amazônico e que se apresenta como uma mistura entre os componentes do oeste sul-americano e do centro-leste amazônico. A análise de estrutura genética ainda identificou pelo menos 3 grupos no oeste da América do Sul, sendo separados basicamente na porção norte dos Andes, incluindo o Equador e Bolívia, e porções norte e sul dos Andes peruanos, corroborando resultados anteriormente observados.

RESUMO

A origem dos nativos americanos remonta a povos do nordeste asiático que teriam chegado a Beríngia durante o Último Máximo Glacial e após o seu fim teriam cruzado as geleiras do norte da América do Norte dando início ao povoamento do continente. Com exceção de populações do ártico, o restante dos nativos americanos foram formados por uma população ancestral comum, entretanto algumas comunidades indígenas atuais e um indivíduo antigo do Brasil apresentam um componente de ancestralidade adicional, que exibe maior afinidade genética com populações do sul da Ásia, Melanésia e Austrália e foi modelado como a contribuição de uma população não-amostrada (População Y). Durante o povoamento inicial os primeiros americanos passaram por um processo de adaptação à imensa diversidade ambiental das Américas, mediado tanto pela evolução biológica quanto cultural, assim como pela construção de nichos, resultando em uma das maiores diversidades culturais do mundo, com aproximadamente 350 grupos étnicos e mais de 180 línguas nativas apenas na Amazônia. Posteriormente, durante o Holoceno médio e tardio alguns fatores como a intensificação da produção de alimentos e mudanças climáticas produziram dinâmicas populacionais que reconfiguraram significativamente a paisagem genética, entre esses eventos a Expansão Tupí é provavelmente o mais relevante para o leste da América do Sul. Além disso, a invasão e colonização europeia, a partir do fim do século 15, levou a um extermínio massivo dos indígenas, particularmente do litoral brasileiro que era predominantemente ocupado pelos Tupí. Aqui nós analisamos os dados genômicos inéditos de 139 nativos americanos contemporâneos de 9 grupos étnicos do Brasil, genotipados nas plataformas *Axiom Human Origins - Affymetrix* (49) e *Axiom InCor BB - Affymetrix* (95) (5 indivíduos presentes em ambas) e combinados a bancos de dados públicos de comunidades indígenas contemporâneas e de indivíduos antigos. No Capítulo 1, nós demonstramos que o sinal de afinidade genética australo-asiática apresenta grande variação intra e interpopulacional, e está muito mais extensamente distribuído na América do Sul do que anteriormente demonstrado, estando presente mais ao sul da região Centro-Oeste brasileira nos Guaraní Kaiowá e também do outro lado da Cordilheira dos Andes nos Chotuna do litoral norte do Peru. Adicionalmente, os modelos de história populacional corroboram que o povoamento inicial das Américas ocorreu pela rota da costa do Pacífico. Enquanto que, no Capítulo 2 nós encontramos um padrão de estrutura genética parcialmente relacionado a diversidade linguística entre os indígenas do leste e oeste da América do Sul, com pelo menos 3 divisões primárias em cada região e um grupo transicional no oeste amazônico, incluindo populações das encostas orientais andinas. Contudo, tanto a variação genética quanto o nível de homozigose são correlacionados à longitude, além da distância genética que é correlacionada a geográfica, sugerindo portanto um efeito conjunto de isolamento por distância e de gargalos populacionais em série, possivelmente remontando a dispersão inicial a partir da costa do Pacífico. Interessantemente, a subestruturação genética de populações indígenas contemporâneas recapitula a ancestralidade subcontinental de indivíduos antigos. Nós ainda encontramos evidências de expansões dêmicas no Holoceno tardio e de uma maior resistência ao colapso populacional no pós-contato entre os grupos do oeste sul-americano. Por fim, diversas linhas de evidências apontam que os Kokama do oeste amazônico passaram por um processo de substituição linguística. Ao passo que, no Capítulo 3 nós mostramos que os Tupiniquim do Espírito Santo no litoral brasileiro, são descendentes do antigo ramo Tupí da costa, além disso nós também datamos os períodos de intensificação da miscigenação com povos europeus e africanos, os quais remontam a alguns eventos históricos particularmente relevantes para a demografia brasileira. Por último, modelos baseados em hipóteses alternativas da Expansão Tupí corroboram a interpretação do arqueólogo brasileiro José P. Brochado, segundo a qual os Tupí da costa (i.e. Tupinambá) e os Guaraní teriam divergido e se expandido a partir da Amazônia por rotas independentes, os primeiros seguindo o curso Rio Amazonas em direção ao leste e depois através do litoral atlântico e os últimos indo diretamente para o sul ao longo dos rios até a Bacia do Paraná.

ABSTRACT

The origin of the native americans traces back to northeastern asian peoples which would have reached Beringia during the Last Glacial Maximum and after its end they would have crossed the northern North America ice sheets initiating the settlement of the continent. Excepting the arctic populations, the remaining native americans were formed by a common ancestral population, however some present-day indigenous communities and one ancient individual from Brazil exhibit an additional ancestry component, which presents higher genetic affinity with populations from southern Asia, Melanesia and Australia and has been modelled as a contribution from an unsampled population (Population Y). During the initial settlement which the first americans underwent a process of adaptation to the vast environmental diversity of the Americas, mediated by both biological and cultural evolution, as well as by niche construction, resulting in one of the greatest cultural diversities of the world, with approximately 350 ethnic groups and over 180 native languages in the Amazon alone. Later, during the mid and late Holocene some factors such as the intensification of food production and climate change produced population dynamics that significantly reconfigured the genetic landscape, among these events the Tupí Expansion is probably the most relevant for eastern South America. Furthermore, the European invasion and colonization, starting in the end of the 15th century, led to a massive extermination of the indigenous people, particularly on the Brazilian coast, which was predominantly occupied by the Tupi. Here we analyze unpublished genomic data from 139 contemporary Native Americans from 9 ethnic groups of Brazil, genotyped on the Axiom Human Origins - Affymetrix (49) and Axiom InCor BB - Affymetrix (95) platforms (5 individuals present in both) and merged to public databases of contemporary indigenous communities and ancient individuals. In Chapter 1, we demonstrate that the Australasian genetic affinity signal has high intra- and interpopulation variation, and is much more widely distributed in South America than previously demonstrated, being present further south in the Brazilian midwest region in the Guaraní Kaiowá and also across the Andes in the Chotuna of the northern peruvian coast. Additionally, the population history models corroborate that the initial settlement of the Americas occurred along the Pacific coast route. Whereas, in Chapter 2 we find a pattern of genetic structure partially related to linguistic diversity among the indigenous peoples of eastern and western South America, with at least 3 primary divisions in each region and a transitional group in the western Amazon, including populations from the Andean eastern slopes. However, both genetic variation and homozygosity level are correlated with longitude, in addition to genetic distance, which is correlated with geographic distance, thus suggesting a joint effect of isolation by distance and serial population bottlenecks, possibly dating back to the initial dispersion from Pacific coast. Interestingly, the genetic substructure of contemporary indigenous populations recapitulates the subcontinental ancestry of ancient individuals. We also find evidence of demic expansions in the late Holocene and a greater resistance to post-contact population collapse among western South American groups. Finally, several lines of evidence indicate that the Kokama of the western Amazon underwent a process of linguistic substitution. While in Chapter 3 we show that the Tupiniquim from Espírito Santo, on the Brazilian coast, are descendants of the ancient coastal Tupi branch, furthermore we also date the periods of admixture intensification with European and African peoples, which trace back to some historical events particularly relevant to Brazilian demography. Finally, models based on alternative Tupí expansion hypotheses corroborate the interpretation of the Brazilian archaeologist José P. Brochado, according to which the coastal Tupí (i.e. Tupinambá) and the Guaraní would have diverged and expanded from the Amazon by independent routes, the first following the Amazon River course towards the east and then through the Atlantic coast and the latter going directly south along the rivers to the Paraná Basin.

APROVAÇÃO ÉTICA E PLANO DE GESTÃO DOS DADOS

A aprovação ética para a coleta das amostras apresentadas nesta tese foi fornecida pela Comissão Nacional de Ética em Pesquisa (CONEP Resolution no. 123 and 4599). A CONEP também aprovou o consentimento oral e o uso dessas amostras em estudos de história populacional e evolução humana. Consentimentos orais individuais e/ou tribais foram obtidos dos participantes que não eram capazes de ler ou escrever. Todas as amostras foram coletadas pelos pesquisadores Francisco Mauro Salzano e José Geraldo Mill e seus colaboradores, de forma consistente com a Declaração de Helsinki e com as leis e regulações vigentes à época da amostragem. Suporte Logístico para a coleta das amostras foi dada pela Fundação Nacional do Índio.

Os dados genômicos inéditos gerados no âmbito deste projeto e incluídos nas análises dos manuscritos publicados (CASTRO E SILVA et al., 2020, 2021) foram integralmente depositados no repositório *European Genome-phenome Archive* (EGA) e encontram-se disponíveis sob os números de acesso EGAS00001004036 e EGAS00001005022. Colocamos abaixo os links para acesso, assim como uma breve descrição dos datasets depositados em relação ao número de indivíduos, grupos de origem e plataforma de genotipagem:

1. (CASTRO E SILVA et al., 2020):

https://ega-archive.org/studies/EGAS00001004036/

   a. EGAD00010001801: 47 Tupiniquim (with local ancestry masking of non-Native American ancestry) (Axiom InCor BB - Affymetrix)
   b. EGAD00010001802: 47 Tupiniquim e 48 Guaraní Mbyá (Axiom InCor BB - Affymetrix)
   c. EGAD00010001803: 1 Tupiniquim, 4 Guaraní Mbyá, 2 Wajãpi, 3 Parakanã e 2 Gavião (Axiom Human Origins - Affymetrix)

2. (CASTRO E SILVA et al., 2021):

https://ega-archive.org/studies/EGAS00001005022/

   a. EGAD00010002061: 1 Asurini, 2 Munduruku, 7 Xikrin e 27 Xávante (Axiom Human Origins - Affymetrix)

Os dados podem ser requisitados através do sistema do EGA, o qual permite o contato entre os solicitantes e um comitê de acesso aos dados (*Data Access Committee* - DAC). Tal comitê julga a adequação dos pedidos de acesso em relação a um conjunto de requisitos sobre o grupo de pesquisa e objetivos do projeto, e por fim, caso defira a solicitação o DAC também requisita a assinatura de um acordo de acesso aos dados (*Data Access Agreement* - DAA).

REFERÊNCIAS BIBLIOGRÁFICAS

1000 GENOMES PROJECT CONSORTIUM et al. A Global Reference for Human Genetic Variation. **Nature**, v. 526, n. 7571, p. 68–74, 1 out. 2015.

ACKERMANN, R. et al. AAPA Statement on Race and Racism. **American Association of Physical Anthropologists**, v. 27, 2019.

ADHIKARI, K. et al. Admixture in Latin America. **Current opinion in genetics & development**, v. 41, p. 106–114, dez. 2016.

ADHIKARI, K. et al. The Genetic Diversity of the Americas. **Annual review of genomics and human genetics**, v. 18, p. 277–296, 31 ago. 2017.

ALMEIDA, F. O. de; NEVES, E. G. Evidências arqueológicas para a origem dos tupi-guarani no leste da amazônia. **Mana**, 2015. . Disponível em: <http://dx.doi.org/10.1590/0104-93132015v21n3p499>.

ALTMAN, I. et al. **"To Make America": European Emigration in the Early Modern Period**. [s.l.] University of California Press, 1991.

ALVIM, Z. Imigrantes: a vida privada dos pobres do campo. In: **República: Da Belle Époque à era do rádio**. [s.l: s.n.]p. 215–287.

AMMERMAN, A. J.; CAVALLI-SFORZA, L. L. **The Neolithic Transition and the Genetics of Populations in Europe**. [s.l.] Princeton University Press, 2014.

ARROYO-KALIN, M. Human Niche Construction and Population Growth in Pre-Columbian Amazonia. **Archaeology International**, v. 20, p. 122–136, 2018. . Acesso em: 14 jun. 2021.

ARROYO-KALIN, M.; RIRIS, P. Did Pre-Columbian Populations of the Amazonian Biome Reach Carrying Capacity during the Late Holocene? **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 376, n. 1816, p. 20190715, 18 jan. 2021.

BAE, C. J.; DOUKA, K.; PETRAGLIA, M. D. On the Origin of Modern Humans: Asian Perspectives. **Science**, v. 358, n. 6368, 8 dez. 2017. Disponível em: <http://dx.doi.org/10.1126/science.aai9067>.

BAILEY, R. G.; HOGG, H. C. A World Ecoregions Map for Resource Reporting. **Environmental conservation**, v. 13, n. 3, p. 195–202, 1986. Acesso em: 13 jul. 2021.

BARBIERI, C. et al. Between Andes and Amazon: The genetic profile of the Arawak-speaking Yanesha. **American journal of physical anthropology**, v. 155, n. 4, p. 600–609, 2014a.

BARBIERI, C. et al. Between Andes and Amazon: The Genetic Profile of the Arawak-Speaking Yanesha. **American journal of physical anthropology**, v. 155, n. 4, p. 600–609, dez. 2014b.

BARBIERI, C. et al. Enclaves of Genetic Diversity Resisted Inca Impacts on Population History. **Scientific reports**, v. 7, n. 1, p. 17411, 12 dez. 2017.

BARBIERI, C. et al. The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. **Molecular biology and evolution**, v. 36, n. 12, p. 2698–2713, 1 dez. 2019.

BARBIERI, C. Genetic exchanges in the highland/lowland transitional environments of South America. In: **Rethinking the Andes--Amazonia Divide: a cross-disciplinary exploration**. [s.l.] UCL Press, 2020. p. 152–163.

BELLWOOD, P.; OTHERS. Examining the farming/language dispersal hypothesis in the East Asian context. **The peopling of East Asia: Putting together archaeology, linguistics and genetics**, v. 1, p. 17–30, 2005.

BERESFORD-JONES, D. G.; MURILLO, E. M. Archaeology. In: PEARCE, A. J.; BERESFORD-JONES, D. G.; HEGGARTY, P. (Ed.). **Rethinking the Andes–Amazonia Divide: A cross-disciplinary exploration**. [s.l.] UCL Press, 2020. p. 21–34.

BERGSTRÖM, A. et al. Insights into Human Genetic Variation and Population History from 929 Diverse Genomes. **Science**, v. 367, n. 6484, 20 mar. 2020. Disponível em: <http://dx.doi.org/10.1126/science.aay5012>.

BISSO-MACHADO, R. et al. Distribution of Y-Chromosome Q Lineages in Native Americans. **American journal of human biology: the official journal of the Human Biology Council**, v. 23, n. 4, p. 563–566, jul. 2011.

BISSO-MACHADO, R.; FAGUNDES, N. J. R. Homo sapiens dispersal and the peopling of the Americas. **A Companion to Anthropological Genetics**, p. 165–185, 2019.

BISSO-MACHADO, R.; FAGUNDES, N. J. R. Uniparental Genetic Markers in Native Americans: A Summary of All Available Data from Ancient and Contemporary Populations. **American journal of physical anthropology**, n. ajpa.24357, 28 jun. 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/ajpa.24357>.

BODNER, M. et al. Rapid Coastal Spread of First Americans: Novel Insights from South America's Southern Cone Mitochondrial Genomes. **Genome research**, v. 22, n. 5, p. 811–820, maio 2012.

BOLNICK, D. A. et al. Native American Genomics and Population Histories. 21 out. 2016. Disponível em: <https://www.annualreviews.org/doi/abs/10.1146/annurev-anthro-102215-100036>. Acesso em: 23 jun. 2021.

BONATTO, S. L.; SALZANO, F. M. Diversity and Age of the Four Major mtDNA Haplogroups, and Their Implications for the Peopling of the New World. **American journal of human genetics**, v. 61, n. 6, p. 1413–1423, dez. 1997.

BONHAM, V. L.; GREEN, E. D. The Genomics Workforce Must Become More Diverse: A Strategic Imperative. **American journal of human genetics**, v. 108, n. 1, p. 3–7, 7 jan. 2021.

BORDA, V. et al. The Genetic Structure and Adaptation of Andean Highlanders and Amazonians Are Influenced by the Interplay between Geography and Culture. **Proceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 51, p. 32557–32565, 22 dez. 2020.

BORRERO, L. A. The elusive evidence: the archeological record of the South American extinct megafauna. In: HAYNES, G. (Ed.). **American Megafaunal Extinctions at the End of the Pleistocene.** Springer, 2009. pp 145-168.

BORTOLINI, M. C. et al. Reconciling pre-Columbian settlement hypotheses requires integrative, multidisciplinary, and model-bound approaches. **Proceedings of the National Academy of Sciences of the United States of America**, 14 jan. 2014. .

BRAJE, T. J. et al. Finding the First Americans. **Science**, v. 358, n. 6363, p. 592–594, 3 nov. 2017.

BRANDINI, S. et al. The Paleo-Indian Entry into South America According to Mitogenomes. **Molecular biology and evolution**, v. 35, n. 2, p. 299–311, 1 fev. 2018.

BROCHADO, J. P. **An Ecological Model of the Spread of Pottery and Agriculture Into Eastern South America**. [s.l.] University of Illinois at Urbana-Champaign, 1984.

BROCHADO, J. P.; LATHRAP, D. Chronologies in the New World: Amazonia. **Illinois: University of Illinois, Urbana-Champaign**, 1982.

BROMLEY, G. R. M. et al. A cosmogenic 10 Be chronology for the local last glacial maximum and termination in the Cordillera Oriental, southern Peruvian Andes: Implications for the tropical role in global climate. **Quaternary science reviews**, v. 148, p. 54–67, 15 set. 2016.

BROWNING, S. R. et al. Ancestry-Specific Recent Effective Population Size in the Americas. **PLoS genetics**, v. 14, n. 5, p. e1007385, maio 2018.

BURKE, W. Utility and Diversity: Challenges for Genomic Medicine. **Annual review of genomics and human genetics**, 1 abr. 2021. Disponível em: <http://dx.doi.org/10.1146/annurev-genom-120220-082640>.

BUSH, M. B. et al. Widespread Reforestation before European Influence on Amazonia. **Science**, v. 372, n. 6541, p. 484–487, 30 abr. 2021.

CASTRO E SILVA, M. A. 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, 2020.

CASTRO E SILVA, M. A. et al. Deep Genetic Affinity between Coastal Pacific and Amazonian Natives Evidenced by Australasian Ancestry. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 14, 6 abr. 2021. Disponível em: <http://dx.doi.org/10.1073/pnas.2025739118>.

CEBALLOS, F. C. et al. Runs of Homozygosity: Windows into Population History and Trait Architecture. **Nature reviews. Genetics**, v. 19, n. 4, p. 220–234, abr. 2018.

CHACÓN-DUQUE, J.-C. et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. **Nature Communications**, 2018. . Disponível em: <http://dx.doi.org/10.1038/s41467-018-07748-z>.

CORDAUX, R. Genetic Evidence for the Demic Diffusion of Agriculture to India. **Science**, 2004. . Disponível em: <http://dx.doi.org/10.1126/science.1095819>.

CORRÊA, Â. A. **Pindorama de mboîa e îakaré: continuidade e mudança na trajetória das populações Tupi**. 2014. Universidade de São Paulo, 2014. Disponível em: <http://seer.pucgoias.edu.br/index.php/habitus/article/viewFile/4257/2447>.

COSTA, G. C. et al. Biome Stability in South America over the Last 30 Kyr: Inferences from Long-term Vegetation Dynamics and Habitat Modelling. **Global ecology and biogeography: a journal of macroecology**, v. 27, n. 3, p. 285–297, mar. 2018.

CURTIN, P. D. **The Atlantic Slave Trade: A Census**. [s.l.] Univ of Wisconsin Press, 1972.

DA CUNHA, M. C. (ed.). **História dos índios no Brasil**. [s.l.] Companhia das Letras, 1992.

DARVILL, C. M. et al. Retreat of the Western Cordilleran Ice Sheet Margin during the Last Deglaciation. **Geophysical research letters**, v. 45, n. 18, p. 9710–9720, 28 set. 2018.

DAVIS, D. D.; GOODWIN, R. C. Island Carib Origins: Evidence and Nonevidence. **American antiquity**, v. 55, n. 1, p. 37–48, 1990.

DE ACOSTA, J. **Natural and moral history of the indies**. [s.l: s.n.] 1589.

DE FILIPPO, C. et al. Bringing Together Linguistic and Genetic Evidence to Test the Bantu Expansion. **Proceedings. Biological sciences / The Royal Society**, v. 279, n. 1741, p. 3256–3263, 22 ago. 2012.

DEININGER, M. et al. Late Quaternary Variations in the South American Monsoon System as Inferred by Speleothems—New Perspectives Using the SISAL Database. **Quaternary**, v. 2, n. 1, p. 6, 28 jan. 2019. . Acesso em: 2 ago. 2021.

DELSER, P. M. et al. **Climate and mountains shaped human ancestral genetic lineages**. 13 jul. 2021. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.07.13.452067v1>. Acesso em: 16 jul. 2021.

DEMETER, F. et al. Early Modern Humans from Tam Pà Ling, Laos: Fossil Review and Perspectives.

Current anthropology, v. 58, n. S17, p. S527–S538, 1 dez. 2017.

DENEVAN, W. M. **The Native Population of the Americas in 1492**. [s.l.] Univ of Wisconsin Press, 1992.

DE SOUZA, J. G. et al. Climate Change and Cultural Resilience in Late Pre-Columbian Amazonia. **Nature ecology & evolution**, v. 3, n. 7, p. 1007–1017, jul. 2019.

DILLEHAY, T. D. et al. Monte Verde: Seaweed, Food, Medicine, and the Peopling of South America. **Science**, v. 320, n. 5877, p. 784–786, 9 maio 2008.

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, 27 jan. 2009.

DILLEHAY, T. D. et al. New Archaeological Evidence for an Early Human Presence at Monte Verde, Chile. **PloS one**, v. 10, n. 11, p. e0141923, 18 nov. 2015.

DILLEHAY, T. D. et al. Simple Technologies and Diverse Food Strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru. **Science advances**, v. 3, n. 5, p. e1602778, maio 2017.

DIXON, R. M. W. et al. **The Amazonian Languages**. [s.l.] Cambridge University Press, 1999.

DOBSON, J. E.; SPADA, G.; GALASSI, G. The Bering Transitory Archipelago: Stepping Stones for the First Americans. **Comptes rendus: Geoscience**, v. 353, n. 1, p. 55–65, 28 abr. 2021.

DRIVER, J. C. et al. Stratigraphy, Radiocarbon Dating, and Culture History of Charlie Lake Cave, British Columbia. **Arctic**, v. 49, n. 3, p. 265–277, 1996.

EBERHARD, D. M., SIMONS, G. F., & FENNIG, C. D. (ed.). **Ethnologue. Languages of the World**. [s.l.] SIL International, 2020.

ERLANDSON, J. M. et al. The Kelp Highway Hypothesis: Marine Ecology, the Coastal Migration Theory, and the Peopling of the Americas. **The Journal of Island and Coastal Archaeology**, v. 2, n. 2, p. 161–174, 30 out. 2007.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. **Genetics**, v. 131, n. 2, p. 479–491, jun. 1992.

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.

FAGUNDES, N. J. R. et al. How Strong Was the Bottleneck Associated to the Peopling of the Americas? New Insights from Multilocus Sequence Data. **Genetics and molecular biology**, v. 41, n. 1 suppl 1, p. 206–214, 2018.

FAGUNDES, N. J. R.; KANITZ, R.; BONATTO, S. L. A Reevaluation of the Native American mtDNA Genome Diversity and Its Bearing on the Models of Early Colonization of Beringia. **PloS one**, v. 3, n. 9, p. e3157, 17 set. 2008.

FEHREN-SCHMITZ, L. et al. Climate Change Underlies Global Demographic, Genetic, and Cultural Transitions in Pre-Columbian Southern Peru. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 26, p. 9443–9448, 1 jul. 2014.

FEHREN-SCHMITZ, L. Genetics. In: PEARCE, A. J.; BERESFORD-JONES, D. G.; HEGGARTY, P. (Ed.). **Rethinking the Andes–Amazonia Divide: A cross-disciplinary exploration**. [s.l.] UCL Press, 2020. p. 48–57.

FERNÁNDEZ-ARMESTO, F. **The Americas: The History of a Hemisphere**. [s.l.] Orion Publishing Group, 2004.

FIGGINS, J. D.; COOK, H. J. **The antiquity of man in America**. [s.l.] American Museum of Natural History, 1927.

FLEGONTOV, P. et al. Palaeo-Eskimo Genetic Ancestry and the Peopling of Chukotka and North America. **Nature**, v. 570, n. 7760, p. 236–240, jun. 2019.

FLORES, B. M.; LEVIS, C. Human-Food Feedback in Tropical Forests. **Science**, v. 372, n. 6547, p. 1146–1147, 11 jun. 2021.

FU, Q. et al. DNA Analysis of an Early Modern Human from Tianyuan Cave, China. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 6, p. 2223–2227, 5 fev. 2013.

FU, Q. et al. An Early Modern Human from Romania with a Recent Neanderthal Ancestor. **Nature**, v. 524, n. 7564, p. 216–219, 13 ago. 2015.

FUSELLI, S. et al. Mitochondrial DNA Diversity in South America and the Genetic History of Andean Highlanders. **Molecular biology and evolution**, v. 20, n. 10, p. 1682–1691, out. 2003.

GASPAR, M. D. et al. Sambaqui (Shell Mound) Societies of Coastal Brazil. In: SILVERMAN, H.; ISBELL, W. H. (Ed.). **The Handbook of South American Archaeology**. New York, NY: Springer New York, 2008. p. 319–335.

GILBERT, M. T. P. et al. DNA from Pre-Clovis Human Coprolites in Oregon, North America. **Science**, v. 320, n. 5877, p. 786–789, 9 maio 2008.

G., M.; SANCHEZ-ALBORNOZ, N. **The Population of Latin America: A History Population and Development Review**, 1976. . Disponível em: <http://dx.doi.org/10.2307/1971673>.

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, 20 mar. 2019a. . Acesso em: 15 set. 2020.

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, 1 jun. 2019b. . Acesso em: 9 ago. 2019.

GOEBEL, T.; WATERS, M. R.; O'ROURKE, D. H. The Late Pleistocene Dispersal of Modern Humans in the Americas. **Science**, v. 319, n. 5869, p. 1497–1502, 14 mar. 2008.

GOLDBERG, A.; MYCHAJLIW, A. M.; HADLY, E. A. Post-Invasion Demography of Prehistoric Humans in South America. **Nature**, v. 532, n. 7598, p. 232–235, 14 abr. 2016.

GÓMEZ-CARBALLA, A. et al. The Peopling of South America and the Trans-Andean Gene Flow of the First Settlers. **Genome research**, v. 28, n. 6, p. 767–779, jun. 2018.

GONZÁLEZ-JOSÉ, R. et al. Late Pleistocene/Holocene Craniofacial Morphology in Mesoamerican Paleoindians: Implications for the Peopling of the New World. **American journal of physical anthropology**, v. 128, n. 4, p. 772–780, dez. 2005.

GONZÁLEZ-JOSÉ, R. et al. The Peopling of America: Craniofacial Shape Variation on a Continental Scale and Its Interpretation from an Interdisciplinary View. **American journal of physical anthropology**, v. 137, n. 2, p. 175–187, out. 2008.

GOUVEIA, M. H. et al. **Origins, admixture dynamics and homogenization of the African gene pool in the Americas**. 28 maio 2019. . Disponível em: <https://www.biorxiv.org/content/10.1101/652701v1.abstract>. Acesso em: 22 out. 2019.

GRAF, K. E.; BUVIT, I. Human Dispersal from Siberia to Beringia: Assessing a Beringian Standstill in Light of the Archaeological Evidence. **Current anthropology**, v. 58, n. S17, p. S583–S603, 1 dez. 2017.

GREENBERG, J. H. et al. The Settlement of the Americas: A Comparison of the Linguistic, Dental, and Genetic Evidence [and Comments and Reply]. **Current anthropology**, v. 27, n. 5, p. 477–497, 1 dez. 1986.

GREGORIO DE SOUZA, J.; ALCAINA MATEOS, J.; MADELLA, M. Archaeological Expansions in Tropical South America during the Late Holocene: Assessing the Role of Demic Diffusion. **PloS one**, v. 15, n. 4, p. e0232367, 27 abr. 2020.

GUIDON, N. Las unidades culturales de São Raimundo Nonato—sudeste del Estado de Piauí— Brazil. In: BRYAN, A. L. (Ed.). **New evidence for the Pleistocene peopling of the Americas**. [s.l.] Center for the Study of Early Man, University of Maine Orono, ME, 1986. p. 157–171.

HAMMARSTRÖM, H. **glottolog/glottolog: Glottolog database 4.4**, 14 maio 2021. . Disponível em: <https://zenodo.org/record/4761960>. Acesso em: 13 jul. 2021.

HARRIS, D. N. et al. Evolutionary Genomic Dynamics of Peruvians Before, During, and after the Inca Empire. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 28, p. E6526–E6535, 10 jul. 2018.

HAYNES, C. V., Jr. Fluted Projectile Points: Their Age and Dispersion: Stratigraphically Controlled Radiocarbon Dating Provides New Evidence on Peopling of the New World. **Science**, v. 145, n. 3639, p. 1408–1413, 25 set. 1964.

HECKENBERGER, M. J. et al. Amazonia 1492: Pristine Forest or Cultural Parkland? **Science**, v. 301, n. 5640, p. 1710–1714, 19 set. 2003.

HECKENBERGER, M.; NEVES, E. G. Amazonian Archaeology. 11 set. 2009. Disponível em: <https://www.annualreviews.org/doi/abs/10.1146/annurev-anthro-091908-164310>. Acesso em: 29 jul. 2021.

HEGGARTY, P. Linguistics. In: **Rethinking the Andes--Amazonia Divide: a cross-disciplinary exploration**. [s.l.] UCL Press, 2020. p. 35–47.

HEINTZMAN, P. D. et al. Bison Phylogeography Constrains Dispersal and Viability of the Ice Free Corridor in Western Canada. **Proceedings of the National Academy of Sciences of the United States of America**, v. 113, n. 29, p. 8057–8063, 19 jul. 2016.

HILL, J. History, power, and identity: Amazonian perspectives. **Acta Historica Universitatis Klaipedensis XIX, Studia Anthropologica**, v. 3, p. 25–47, 2009.

HINDORFF, L. A. et al. Prioritizing Diversity in Human Genomics Research. **Nature reviews. Genetics**, v. 19, n. 3, p. 175–185, mar. 2018.

HORNBORG, A. Ethnogenesis, Regional Integration, and Ecology in Prehistoric Amazonia. **Current Anthropology**, 2005. . Disponível em: <http://dx.doi.org/10.1086/431530>.

HUGO PAN-ASIAN SNP CONSORTIUM et al. Mapping Human Genetic Diversity in Asia. **Science**, v. 326, n. 5959, p. 1541–1545, 11 dez. 2009.

HÜNEMEIER, T. et al. Cultural Diversification Promotes Rapid Phenotypic Evolution in Xavánte Indians. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 1, p. 73–77, 3 jan. 2012a.

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, 21 jun. 2012b.

HUNLEY, K. L. et al. A Formal Test of Linguistic and Genetic Coevolution in Native Central and South America. **American journal of physical anthropology**, v. 132, n. 4, p. 622–631, abr. 2007.

IBGE. **500 anos de povoamento.** Rio de Janeiro. Instituto Brasileiro de Geografia e Estatística, , 2000. . Disponível em: <https://brasil500anos.ibge.gov.br/estatisticas-do-povoamento.html>.

IRIARTE, J. et al. Out of Amazonia: Late-Holocene climate change and the Tupi–Guarani trans-continental expansion. **Holocene**, v. 27, n. 7, p. 967–975, 1 jul. 2017.

IRIARTE, J. et al. The origins of Amazonian landscapes: Plant cultivation, domestication and the spread of food production in tropical South America. **Quaternary science reviews**, v. 248, p. 106582, 15 nov. 2020.

JENKINS, D. L. et al. Clovis Age Western Stemmed Projectile Points and Human Coprolites at the Paisley Caves. **Science**, v. 337, n. 6091, p. 223–228, 13 jul. 2012.

JI, X. et al. The oldest Hoabinhian technocomplex in Asia (43.5 ka) at Xiaodong rockshelter, Yunnan Province, southwest China. **Quaternary International**, 2016. . Disponível em: <http://dx.doi.org/10.1016/j.quaint.2015.09.080>.

JONES, T. L. et al. Historic and bioarchaeological evidence supports late onset of post-Columbian epidemics in Native California. **Proceedings of the National Academy of Sciences**, 2021. . Disponível em: <http://dx.doi.org/10.1073/pnas.2024802118>.

JOSEPH, G.; RUHLEN, M. **An Amerind Etymological Dictionary. Department of Anthropological Sciences.** Stanford: Stanford University, , 2007. .

KAMEN, H. **Spain's Road to Empire: The Making of a World Power, 1492-1763**. [s.l.] Penguin UK, 2003.

KAMVAR, Z. N.; BROOKS, J. C.; GRÜNWALD, N. J. Novel R Tools for Analysis of Genome-Wide Population Genetic Data with Emphasis on Clonality. **Frontiers in genetics**, v. 6, p. 208, 10 jun. 2015.

KAMVAR, Z. N.; TABIMA, J. F.; GRÜNWALD, N. J. Poppr: An R Package for Genetic Analysis of Populations with Clonal, Partially Clonal, And/or Sexual Reproduction. **PeerJ**, v. 2, p. e281, 4 mar. 2014.

KEEGAN, W. F. Modeling dispersal in the prehistoric West Indies. **World archaeology**, v. 26, n. 3, p. 400–420, 1 fev. 1995.

KEHDY, F. S. G. et al. Origin and Dynamics of Admixture in Brazilians and Its Effect on the Pattern of Deleterious Mutations. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 28, p. 8696–8701, 14 jul. 2015.

KERN, D. C. et al. Distribution of Amazonian Dark Earths in the Brazilian Amazon. In: LEHMANN, J. et al. (Ed.). **Amazonian Dark Earths: Origin Properties Management**. Dordrecht: Springer Netherlands, 2003. p. 51–75.

KITCHEN, A.; MIYAMOTO, M. M.; MULLIGAN, C. J. A Three-Stage Colonization Model for the Peopling of the Americas. **PloS one**, v. 3, n. 2, p. e1596, 13 fev. 2008.

LAHREN, L.; BONNICHSEN, R. Bone Foreshafts from a Clovis Burial in Southwestern Montana. **Science**, v. 186, n. 4159, p. 147–150, 11 out. 1974.

LAMBECK, K. et al. Sea level and global ice volumes from the Last Glacial Maximum to the Holocene. **Proceedings of the National Academy of Sciences**, 2014. . Disponível em: <http://dx.doi.org/10.1073/pnas.1411762111>.

LARSON, G. et al. Current Perspectives and the Future of Domestication Studies. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 17, p. 6139–6146, 29 abr. 2014.

LATHRAP, D. W. **The Upper Amazon**. London: Thames & Hudson, 1970. 256 p.

LESNEK, A. J. et al. Deglaciation of the Pacific Coastal Corridor Directly Preceded the Human Colonization of the Americas. **Science advances**, v. 4, n. 5, p. eaar5040, maio 2018.

LINDO, J. et al. Ancient Individuals from the North American Northwest Coast Reveal 10,000 Years of

Regional Genetic Continuity. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 16, p. 4093–4098, 18 abr. 2017.

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.

LOMBARDO, U. et al. Early Holocene Crop Cultivation and Landscape Modification in Amazonia. **Nature**, v. 581, n. 7807, p. 190–193, maio 2020.

LOOG, L. et al. Estimating Mobility Using Sparse Data: Application to Human Genetic Variation. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 46, p. 12213–12218, 14 nov. 2017.

LYNCH, T. F. The South American Paleo-Indians. **Ancient Native Americans**, p. 455–489, 1978.

MACARIO, K. D. et al. The Long-Term Tupiguarani Occupation in Southeastern Brazil. **Radiocarbon**, v. 51, n. 3, p. 937–946, 2009. . Acesso em: 17 jun. 2021.

MACNEISH, R. S. et al. **Prehistory of the Ayacucho Basin, Peru: Volume II: Excavations and Chronology**. [s.l.] University of Michigan Press, 1981.

MARGOLD, M. et al. Beryllium-10 dating of the Foothills Erratics Train in Alberta, Canada, indicates detachment of the Laurentide Ice Sheet from the Rocky Mountains at\ 15 ka. **Quaternary Research**, v. 92, n. 2, p. 469–482, 2019.

MARTÍNEZ, G. et al. Subsistence strategies in Argentina during the late Pleistocene and early Holocene. **Quaternary science reviews**, v. 144, p. 51–65, 15 jul. 2016.

MAS-SANDOVAL, A. et al. Reconstructed Lost Native American Populations from Eastern Brazil Are Shaped by Differential Jê/Tupi Ancestry. **Genome biology and evolution**, v. 11, n. 9, p. 2593–2604, 1 set. 2019.

MATSUMURA, H. et al. Craniometrics Reveal "Two Layers" of Prehistoric Human Dispersal in Eastern Eurasia. **Scientific reports**, v. 9, n. 1, p. 1–12, 5 fev. 2019. . Acesso em: 6 jul. 2021.

MCALISTER, L. N. Spain and Portugal in the New World, 1492-1700. **The American Historical Review**, 1987. Disponível em: <http://dx.doi.org/10.2307/1862976>.

MCCOLL, H. et al. The Prehistoric Peopling of Southeast Asia. **Science**, v. 361, n. 6397, p. 88–92, 6 jul. 2018.

MCMICHAEL, C. H. et al. Predicting Pre-Columbian Anthropogenic Soils in Amazonia. **Proceedings. Biological sciences / The Royal Society**, v. 281, n. 1777, p. 20132475, 22 fev. 2014.

MEGGERS AND CLIFFORD, B. Lowland South America and the Antilles. In: JENNINGS, J. D. (Ed.). **Ancient Native Americans**. CA: San Francisco: W. H. Freeman and Company, 1978. p. 543–591.

MEGGERS, B. J. **A reconstituição da pré-história amazônica: algumas considerações teóricas**. [s.l.] Instituto de Geografia da Universidade de São Paulo, 1974.

MEGGERS, B. J. Vegetational fluctuation and prehistoric cultural adaptation in Amazonia: Some tentative correlations. **World archaeology**, v. 8, n. 3, p. 287–303, 1 fev. 1977.

MEGGERS, B. J. Archeological and ethnographic evidence compatible with the model of forest fragmentation. 1982. Disponível em: <https://agris.fao.org/agris-search/search.do?recordID=US19840108828>.

MELTZER, D. J. **First Peoples in a New World: Colonizing Ice Age America**. [s.l.] University of California Press, 2009.

MENOUNOS, B. et al. Cordilleran Ice Sheet Mass Loss Preceded Climate Reversals near the Pleistocene Termination. **Science**, v. 358, n. 6364, p. 781–784, 10 nov. 2017.

METCALF, J. L. et al. Synergistic Roles of Climate Warming and Human Occupation in Patagonian Megafaunal Extinctions during the Last Deglaciation. **Science advances**, v. 2, n. 6, p. e1501682, jun. 2016.

MÉTRAUX, A. Migrations historiques des Tupi-Guarani. **Journal de la Société des Américanistes**, 1927. . Disponível em: <http://dx.doi.org/10.3406/jsa.1927.3618>.

MICHAEL, L. On the Pre-Columbian Origin of Proto-Omagua-Kokama. **Journal of Language Contact**, v. 7, n. 2, p. 309–344, 14 maio 2014. Acesso em: 18 set. 2020.

MILLER, E. T. A Cultura Cerâmica do Tronco Tupí no alto Ji-Paraná, Rondônia, Brasil: algumas reflexões teóricas, hipotéticas e conclusivas. **Revista Brasileira de Linguística Antropológica**, 2009. Disponível em: <http://periodicos.unb.br/index.php/ling/article/download/12288/10774>.

MILLS, M. C.; RAHAL, C. A Scientometric Review of Genome-Wide Association Studies. **Communications biology**, v. 2, p. 9, 7 jan. 2019.

MILLS, M. C.; RAHAL, C. The GWAS Diversity Monitor Tracks Diversity by Disease in Real Time. **Nature genetics**, v. 52, n. 3, p. 242–243, mar. 2020.

MONTENEGRO, R. A.; STEPHENS, C. Indigenous Health in Latin America and the Caribbean. **The Lancet**, v. 367, n. 9525, p. 1859–1869, 3 jun. 2006.

MONTINARO, F. et al. Unravelling the Hidden Ancestry of American Admixed Populations. **Nature communications**, v. 6, p. 6596, 24 mar. 2015.

MORENO-ESTRADA, A. et al. Reconstructing the Population Genetic History of the Caribbean. **PLoS genetics**, v. 9, n. 11, p. e1003925, nov. 2013.

MORENO-ESTRADA, A. et al. Human Genetics. The Genetics of Mexico Recapitulates Native American Substructure and Affects Biomedical Traits. **Science**, v. 344, n. 6189, p. 1280–1285, 13 jun. 2014.

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, 11 jan. 2018a.

MORENO-MAYAR, J. V. et al. Early Human Dispersals within the Americas. **Science**, v. 362, n. 6419, 7 dez. 2018b. Disponível em: <http://dx.doi.org/10.1126/science.aav2621>.

MULLIGAN, C. J.; KITCHEN, A.; MIYAMOTO, M. M. Updated Three-Stage Model for the Peopling of the Americas. **PloS one**, v. 3, n. 9, p. e3199, 17 set. 2008.

NAKATSUKA, N. et al. A Paleogenomic Reconstruction of the Deep Population History of the Andes. **Cell**, v. 181, n. 5, p. 1131–1145.e21, 28 maio 2020.

NEED, A. C.; GOLDSTEIN, D. B. Next Generation Disparities in Human Genomics: Concerns and Remedies. **Trends in genetics: TIG**, v. 25, n. 11, p. 489–494, nov. 2009.

NEEL, J. V.; SALZANO, F. M. Further Studies on the Xavante Indians. X. Some Hypotheses-Generalizations Resulting from These Studies. **American journal of human genetics**, v. 19, n. 4, p. 554–574, jul. 1967.

NEVES, E. G. Archaeological cultures and past identities in the pre-colonial Central Amazon. **Ethnicity in ancient Amazonian: reconstructing past identities from Archaeology, Linguistic and Etnohistory. Boulder: University Press of Colorado**, p. 1–27, 2011.

NEVES, E. G. Was Agriculture a Key Productive Activity in Pre-Colonial Amazonia? The Stable Productive Basis for Social Equality in the Central Amazon. In: BRONDÍZIO, E. S.; MORAN, E. F. (Ed.). **Human-Environment Interactions: Current and Future Directions**. Dordrecht: Springer Netherlands, 2013. p. 371–388.

NEVES, W. A.; MEYER, D.; PUCCIARELLI, H. M. Early skeletal remains and the peopling of the Americas. **Revista de antropologia social / Departamento de Antropologia Social, Facultad de Ciencias Politicas y Sociologia, Universidad Complutense de Madrid**, v. 39, n. 2, p. 121–139, 1996.

NICHOLS, J. **How America Was Colonized: Linguistic EvidenceMobility and Ancient Society in Asia and the Americas**, 2015. Disponível em: <http://dx.doi.org/10.1007/978-3-319-15138-0_9>.

NING, C. et al. The genomic formation of First American ancestors in East and Northeast Asia. **bioRxiv**, 2020. Disponível em: <https://www.biorxiv.org/content/10.1101/2020.10.12.336628v1.abstract>.

NOELLI, F. S. The Tupi: explaining origin and expansions in terms of archaeology and of historical linguistics. **Antiquity**, v. 72, n. 277, p. 648–663, set. 1998. . Acesso em: 16 jun. 2021.

NOELLI, F. S. Rethinking Stereotypes and the History of Research on Jê Populations in South Brazil: An Interdisciplinary Point of View. In: FUNARI, P. P.; ZARANKIN, A.; STOVEL, E. (Ed.). **Global Archaeological Theory: Contextual Voices and Contemporary Thoughts**. Boston, MA: Springer US, 2005. p. 167–190.

NOELLI, F. S. The Tupi Expansion. **The Handbook of South American Archaeology**, 2008. Disponível em: <http://dx.doi.org/10.1007/978-0-387-74907-5_33>.

ONGARO, L. et al. The Genomic Impact of European Colonization of the Americas. **Current biology: CB**, v. 29, n. 23, p. 3974–3986.e4, 2 dez. 2019.

O'ROURKE, D. H.; RAFF, J. A. The Human Genetic History of the Americas: The Final Frontier. **Current biology: CB**, v. 20, n. 4, p. R202–7, 23 fev. 2010.

OWSLEY, D. W.; HUNT, D. R. Clovis and Early Archaic Crania from the Anzick Site (24PA506), Park County, Montana. **Plains anthropologist**, v. 46, n. 176, p. 115–124, 1 maio 2001.

PATTERSON, N. et al. Ancient Admixture in Human History. **Genetics**, 2012. . Disponível em: <http://dx.doi.org/10.1534/genetics.112.145037>.

PEDERSEN, M. W. et al. Postglacial Viability and Colonization in North America's Ice-Free Corridor. **Nature**, v. 537, n. 7618, p. 45–49, 1 set. 2016.

PENA, S. D. J. Homo brasilis: aspectos genéticos, lingüísticos, históricos e socioantropológicos da formação do povo brasileiro. In: **Homo Brasilis: aspectos genéticos, lingüísticos, históricos e socioantropológicos da formação do povo brasileiro**. [s.l: s.n.]p. 192–192.

PEREGO, U. A. et al. Distinctive Paleo-Indian Migration Routes from Beringia Marked by Two Rare mtDNA Haplogroups. **Current biology: CB**, v. 19, n. 1, p. 1–8, 13 jan. 2009.

PEREZ, S. I. et al. Peopling time, spatial occupation and demography of Late Pleistocene-Holocene human population from Patagonia. **Quaternary international: the journal of the International Union for Quaternary Research**, v. 30, p. 1e10, 2016.

PEREZ, S. I.; POSTILLONE, M. B.; RINDEL, D. Domestication and Human Demographic History in South America. **American journal of physical anthropology**, v. 163, n. 1, p. 44–52, maio 2017.

PICKRELL, J. K.; PRITCHARD, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. **PLoS genetics**, v. 8, n. 11, p. e1002967, 15 nov. 2012.

PINOTTI, T. et al. Y Chromosome Sequences Reveal a Short Beringian Standstill, Rapid Expansion, and Early Population Structure of Native American Founders. **Current biology: CB**, v. 29, n. 1, p. 149–157.e3, 7 jan. 2019.

PIPERNO, D. R.; MCMICHAEL, C. H. A 5,000-year vegetation and fire history for tierra firme forests in the Medio Putumayo-Algodón watersheds, northeastern Peru. **Proceedings of the National Academy of Sciences of the United States of America**, 2021. Disponível em: <https://www.pnas.org/content/early/2021/06/04/2022213118.short>.

POPEJOY, A. B.; FULLERTON, S. M. Genomics is failing on diversity. **Nature**, 13 out. 2016. .

POSTH, C. et al. Reconstructing the Deep Population History of Central and South America. **Cell**, v. 175, n. 5, p. 1185–1197.e22, 15 nov. 2018.

POTTER, B. A. et al. Early colonization of Beringia and Northern North America: Chronology, routes, and adaptive strategies. **Quaternary international: the journal of the International Union for Quaternary Research**, v. 444, p. 36–55, 20 jul. 2017.

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.

POWELL, J. F.; NEVES, W. A. Craniofacial Morphology of the First Americans: Pattern and Process in the Peopling of the New World. **American journal of physical anthropology**, v. Suppl 29, p. 153–188, 1999.

PRATES, L.; POLITIS, G. G.; PEREZ, S. I. Rapid Radiation of Humans in South America after the Last Glacial Maximum: A Radiocarbon-Based Study. **PloS one**, v. 15, n. 7, p. e0236023, 22 jul. 2020.

PRUGNOLLE, F.; MANICA, A.; BALLOUX, F. Geography Predicts Neutral Genetic Diversity of Human Populations. **Current biology: CB**, v. 15, n. 5, p. R159–60, 8 mar. 2005.

RAE, J. W. B. et al. Overturning Circulation, Nutrient Limitation, and Warming in the Glacial North Pacific. **Science advances**, v. 6, n. 50, dez. 2020. Disponível em: <http://dx.doi.org/10.1126/sciadv.abd1654>.

RAFF, J.; TACKNEY, J.; O'ROURKE, D. H. South from Alaska: A Pilot aDNA Study of Genetic History on the Alaska Peninsula and the Eastern Aleutians. **Human biology**, v. 82, n. 5-6, p. 677–693, dez. 2010.

RAGHAVAN, M. et al. Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans. **Nature**, v. 505, n. 7481, p. 87–91, 2 jan. 2014.

RAGHAVAN, M. et al. POPULATION GENETICS. Genomic Evidence for the Pleistocene and Recent Population History of Native Americans. **Science**, v. 349, n. 6250, p. aab3884, 21 ago. 2015.

RAMACHANDRAN, S. et al. Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 44, p. 15942–15947, 1 nov. 2005.

RAMALLO, V. et al. Demographic expansions in South America: Enlightening a complex scenario with genetic and linguistic data. **American Journal of Physical Anthropology**, 2013. Disponível em: <http://dx.doi.org/10.1002/ajpa.22219>.

RASMUSSEN, M. et al. Ancient Human Genome Sequence of an Extinct Palaeo-Eskimo. **Nature**, v. 463, n. 7282, p. 757–762, 11 fev. 2010.

RASMUSSEN, M. et al. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. **Science**, v. 334, n. 6052, p. 94–98, 7 out. 2011.

RASMUSSEN, M. et al. The Genome of a Late Pleistocene Human from a Clovis Burial Site in Western Montana. **Nature**, v. 506, n. 7487, p. 225–229, 13 fev. 2014.

RASMUSSEN, M. et al. The Ancestry and Affiliations of Kennewick Man. **Nature**, v. 523, n. 7561, p. 455–458, 23 jul. 2015.

REICH, D. et al. Reconstructing Native American Population History. **Nature**, v. 488, n. 7411, p. 370–374, 16 ago. 2012.

REYES-CENTENO, H. et al. Genomic and Cranial Phenotype Data Support Multiple Modern Human Dispersals from Africa and a Southern Route into Asia. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 20, p. 7248–7253, 20 maio 2014.

RIRIS, P.; ARROYO-KALIN, M. Widespread Population Decline in South America Correlates with Mid-Holocene Climate Change. **Scientific reports**, v. 9, n. 1, p. 6850, 9 maio 2019.

RODRIGUES, ARYON D. & CABRAL, ANA SUELLY A. C. The indigenous languages of South America: a comprehensive guide. In: CAMPBELL, L.; GRONDONA, V. (Ed.). **The Indigenous Languages of South America: A Comprehensive Guide**. [s.l.] Walter de Gruyter, 2012a. p. 495–574.

RODRIGUES, ARYON D. & CABRAL, ANA SUELLY A. C. Tupían. In: CAMPBELL, L.; GRONDONA, V. (Ed.). **The Indigenous Languages of South America: A Comprehensive Guide**. Berlin: Walter de Gruyter Mouton, 2012b. p. 495–574.

ROEWER, L. et al. Continent-Wide Decoupling of Y-Chromosomal Genetic Variation from Language and Geography in Native South Americans. **PLoS genetics**, v. 9, n. 4, p. e1003460, abr. 2013.

ROOSEVELT, A. C. The Amazon and the Anthropocene: 13,000 years of human influence in a tropical rainforest. **Anthropocene**, v. 4, p. 69–87, 1 dez. 2013.

RUIZ-LINARES, A. et al. Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. **PLoS genetics**, v. 10, n. 9, p. e1004572, set. 2014.

SALZANO, F. M. et al. **South American Indians: A Case Study in Evolution**. [s.l.] Clarendon Press, 1988.

SALZANO, F. M.; BORTOLINI, M. C. **The Evolution and Genetics of Latin American Populations**. [s.l.] Cambridge University Press, 2005.

SALZANO, F. M.; SANS, M. Interethnic Admixture and the Evolution of Latin American Populations. **Genetics and molecular biology**, v. 37, n. 1 Suppl, p. 151–170, mar. 2014.

SANDOVAL, J. R. et al. The Genetic History of Indigenous Populations of the Peruvian and Bolivian Altiplano: The Legacy of the Uros. **PloS one**, v. 8, n. 9, p. e73006, 11 set. 2013.

SANJEK, R. Rethinking Migration, Ancient to Future. **Global Networks-A Journal Of Transnational Affairs**, v. 3, n. 3, p. 315–336, jul. 2003.

SANTOS, E. J. M. dos et al. Origins and Demographic Dynamics of Tupí Expansion: A Genetic Tale. **Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas**, v. 10, n. 2, p. 217–228, 2015. . Acesso em: 8 ago. 2019.

SANTOS, F. R. Genetic diversity patterns in the Andes and Amazonia. In: PEARCE, A. J.; BERESFORD-JONES, D. G.; HEGGARTY, P. (Ed.). **Rethinking the Andes–Amazonia Divide: A cross-disciplinary exploration**. [s.l.] UCL Press, 2020. p. 143–151.

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, 1 jun. 2018.

SCHRAIBER, J. G.; AKEY, J. M. Methods and Models for Unravelling Human Evolutionary History. **Nature reviews. Genetics**, v. 16, n. 12, p. 727–740, dez. 2015.

SCHWARCZ, L. M. O espetáculo das raças: cientistas, instituições e questão racial no Brasil. **São Paulo: Companhia das Letras**, p. 99–133, 1993.

SCHWARCZ, L. M.; STARLING, H. M. **Brasil: uma biografia: Com novo pós-escrito**. [s.l.] Editora Companhia das Letras, 2015.

SEARING, J. F.; ELTIS, D. The Trans-Atlantic Slave Trade: A Database on CD-ROM. **The American Historical Review**, 2001. Disponível em: <http://dx.doi.org/10.2307/2692332>.

SIKORA, M. et al. Ancient Genomes Show Social and Reproductive Behavior of Early Upper Paleolithic Foragers. **Science**, v. 358, n. 6363, p. 659–662, 3 nov. 2017.

SIKORA, M. et al. The Population History of Northeastern Siberia since the Pleistocene. **Nature**, v. 570, n. 7760, p. 182–188, jun. 2019.

SKIDMORE, T. E. Brazil: Five Centuries of Change. **OUP Catalogue**, 2009. Disponível em: <https://ideas.repec.org/b/oxp/obooks/9780195374551.html>. Acesso em: 9 ago. 2019.

SKOGLUND, P. et al. Genetic Evidence for Two Founding Populations of the Americas. **Nature**, v. 525, n. 7567, p. 104–108, 3 set. 2015.

SKOGLUND, P.; REICH, D. A Genomic View of the Peopling of the Americas. **Current opinion in genetics & development**, v. 41, p. 27–35, dez. 2016.

SOKAL, R. R.; ODEN, N. L.; WILSON, C. **Genetic evidence for the spread of agriculture in Europe by demic diffusionNature**, 1991. . Disponível em: <http://dx.doi.org/10.1038/351143a0>.

STANISH, C. The Origin of State Societies in South America. **Annual review of anthropology**, v. 30, n. 1, p. 41–64, out. 2001.

STANNARD, D. E. **American Holocaust: The Conquest of the New World**. [s.l.] Oxford University Press, USA, 1993.

STAPLES, J.; NICKERSON, D. A.; BELOW, J. E. Utilizing Graph Theory to Select the Largest Set of Unrelated Individuals for Genetic Analysis. **Genetic epidemiology**, v. 37, n. 2, p. 136–141, fev. 2013.

STONEKING, M.; DELFIN, F. The Human Genetic History of East Asia: Weaving a Complex Tapestry. **Current biology: CB**, v. 20, n. 4, p. R188–93, 23 fev. 2010.

SUTTER, R. C. The Pre-Columbian Peopling and Population Dispersals of South America. **Journal of Archaeological Research**, v. 29, n. 1, p. 93–151, 1 mar. 2021.

SZATHMARY, E. J. mtDNA and the Peopling of the Americas. **American journal of human genetics**, v. 53, n. 4, p. 793–799, out. 1993.

TAMM, E. et al. Beringian Standstill and Spread of Native American Founders. **PloS one**, v. 2, n. 9, p. e829, 5 set. 2007.

TARAZONA-SANTOS, E. et al. Genetic Differentiation in South Amerindians Is Related to Environmental and Cultural Diversity: Evidence from the Y Chromosome. **American journal of human genetics**, v. 68, n. 6, p. 1485–1496, jun. 2001.

TASSI, F. et al. Early Modern Human Dispersal from Africa: Genomic Evidence for Multiple Waves of Migration. **Investigative genetics**, v. 6, p. 13, 6 nov. 2015.

THORNTON, R. **American Indian Holocaust and Survival: A Population History Since 1492**. [s.l.] University of Oklahoma Press, 1987.

THORNTON, R. Native American Demographic and Tribal Survival into the Twenty-first Century. **American studies** , v. 46, n. 3/4, p. 23–38, 2005.

TORRONI, A. et al. Asian Affinities and Continental Radiation of the Four Founding Native American mtDNAs. **American journal of human genetics**, v. 53, n. 3, p. 563–590, set. 1993.

TURNER, C. G. I. I. The dental search for Ntive American origins. **Out of Asia: peopling the Americas and the Pacific**, p. 31–78, 1985.

UBELAKER, D. H. Population size, contact to nadir. **Handbook of North American Indians**, v. 3, p. 694–701, 2006.

URBAN, G. A história da cultura brasileira segundo as línguas nativas. In: DA CUNHA, M. C. (Ed.). **História dos índios no Brasil**. São Paulo: Companhia das Letras, 1992.

VERDU, P. et al. Patterns of Admixture and Population Structure in Native Populations of Northwest North America. **PLoS genetics**, v. 10, n. 8, p. e1004530, ago. 2014.

VIVEIROS DE CASTRO, E. Os involuntários da pátria: elogio do subdesenvolvimento. **Caderno de leituras**, n. 65, p. 1–9, 2017.

WALKER, R. S. et al. Cultural Phylogenetics of the Tupi Language Family in Lowland South America. **PloS one**, v. 7, n. 4, p. e35025, 10 abr. 2012.

WALLACE, D. C.; GARRISON, K.; KNOWLER, W. C. Dramatic Founder Effects in Amerindian Mitochondrial DNAs. **American journal of physical anthropology**, v. 68, n. 2, p. 149–155, out. 1985.

WANG, S. et al. Genetic Variation and Population Structure in Native Americans. **PLoS genetics**, v. 3, n. 11, p. e185, nov. 2007.

WANG, S. et al. Geographic Patterns of Genome Admixture in Latin American Mestizos. **PLoS genetics**, v. 4, n. 3, p. e1000037, 21 mar. 2008.

WATERS, M. R. et al. Pre-Clovis Mastodon Hunting 13,800 Years Ago at the Manis Site, Washington. **Science**, v. 334, n. 6054, p. 351–353, 21 out. 2011.

WATERS, M. R. et al. Pre-Clovis Projectile Points at the Debra L. Friedkin Site, Texas—Implications for the Late Pleistocene Peopling of the Americas. **Science Advances**, v. 4, n. 10, p. eaat4505, 1 out. 2018. . Acesso em: 1 jul. 2021.

WATERS, M. R. Late Pleistocene Exploration and Settlement of the Americas by Modern Humans. **Science**, v. 365, n. 6449, 12 jul. 2019. Disponível em: <http://dx.doi.org/10.1126/science.aat5447>.

WEN, B. et al. Genetic Evidence Supports Demic Diffusion of Han Culture. **Nature**, v. 431, n. 7006, p. 302–305, 16 set. 2004.

WESTAWAY, K. E. et al. An Early Modern Human Presence in Sumatra 73,000-63,000 Years Ago. **Nature**, v. 548, n. 7667, p. 322–325, 17 ago. 2017.

WILLERSLEV, E.; MELTZER, D. J. Peopling of the Americas as Inferred from Ancient Genomics. **Nature**, v. 594, n. 7863, p. 356–364, jun. 2021.

WILLIAMS, T. J. et al. Evidence of an Early Projectile Point Technology in North America at the Gault Site, Texas, USA. **Science advances**, v. 4, n. 7, p. eaar5954, jul. 2018.

ZHENG, X. et al. A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. **Bioinformatics** , v. 28, n. 24, p. 3326–3328, 15 dez. 2012.

ANEXOS

Material suplementar do Capítulo 1

PNAS
www.pnas.org

Supplementary Information for

Deep genetic affinity between Coastal Pacific and Amazonian natives evidenced by Australasian ancestry

Marcos Araújo Castro e Silva[a,1], Tiago Ferraz[a,1], Maria Cátira Bortolini[b], David Comas[c], and Tábita Hünemeier[b,2]

Tábita Hünemeier
Email: hunemeier@usp.br

[a]*Departmento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil;* [b]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil;* [c]*Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain*

This PDF file includes:

Supplementary text
Legends for Datasets S1 to S6
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S6

Summary

Supplementary Information Text

Extended Methods

Data overview

Newly generated data for 37 Brazilian natives from 4 indigenous communities, namely Asurini, Munduruku (both Tupí-speaker groups), Xavánte, and Xikrin (both Jê-speaker groups), were genotyped in the Axiom Human Origins array - Affymetrix/Thermo Fisher (1). These populations are settled in the Amazonian rainforest (Asurini, Munduruku, and Xikrin), or in the Brazilian central plateau tropical savanna (Xavánte).

Ethical approval for sample collection was provided by the Brazilian National Ethics Commission (CONEP Resolution no. 123 and 4599). CONEP also approved oral consent for the use of these samples in population history and human evolution studies. Individual and/or tribal informed oral consent was obtained from participants who were not able to read or write. All sampling was coordinated by Francisco Mauro Salzano (*in memoriam*) and their collaborators, in a manner consistent with the Declaration of Helsinki and Brazilian laws and regulations applicable at the time of sampling. Logistical support for the sample collection was provided by the Fundação Nacional do Índio. The results of this study were presented to the participating communities.

Dataset assembly and quality control

These data were merged with publicly available datasets (Axiom Human Origins array - Affymetrix/Thermo Fisher (1) genotyped or whole-genome sequenced) of populations from Brazil (1–3) and other countries in South America (Colombia, Ecuador, and Peru) and Mexico (4–6). Finally, we also combined the 1240K_HO dataset assembly (v42.4) and the merging procedure was conducted using PLINK 1.9 (7), sharing single nucleotide polymorphisms (SNPs) across merging datasets. The resulting dataset contained 383 individuals from 58 indigenous groups (Dataset S1A), along with the 67 world-wide reference populations (Dataset S2) and a total of 438,443 SNPs. Next, we removed markers with more than a 5% absence rate, and no sample was removed with a 10% absence rate criteria. We also excluded markers with a pairwise correlation above 20% ($r^2 > 0.2$ inside a sliding window of 50 kb size and step size of 10 kb), obtaining a subset of

127,931 markers and applied an unsupervised ADMIXTURE (8) analysis with K = 3 on the subset of samples from the American continent, Sub-Saharan Africa, and Western Europe. We then estimated the pairwise IBD with PLINK 1.9 (plink --file mydata --genome), which uses the which uses the method-of-moments to calculate the probability of sharing 0 (ZO), 1 (Z1), or 2 (Z2) alleles identical by descent between any given pair of individuals over all the loci, and the total proportion of IBD is estimated between a pair of individuals as PI_HAT = Z2+0.5*Z1. We then used a threshold of PI-HAT < 0.375 to identify the maximum unrelated dataset with PRIMUS (9). Finally, we filtered the data to remove admixed (< 99% inferred non-Native American ancestry; 150 unadmixed samples) and selected the maximum unrelated set of individuals (PI_HAT < 0.375; 312 unrelated samples). The subset of unrelated and unadmixed Native American samples includes 87 individuals. Metadata for every Native American sample (test samples) and every reference population sample are summarized in Dataset S1A and Dataset S2, respectively. The complete set of SNPs and the subset of unrelated and unadmixed Native American samples was used for all analyses, unless otherwise specified.

Exploratory data analysis

Initially, Principal Component Analyses were applied with SNPRelate R/Bioconductor (10) to the LD pruned dataset, obtained as above mentioned, to check data quality, inconsistencies introduced by the merging process and most importantly if any Native American sample was an outlier in relation to the other American indigenous groups. We also applied the ancestry estimates obtained with the unsupervised ADMIXTURE analysis, as previously described, to visualize and evaluate the influence of the proportions genetic ancestry created by the recent post-Contact 3-way admixture between (Native Americans, European conquistadors, and enslaved Sub-Saharan Africans). Next, we performed a PCA on the subset of unadmixed and unrelated Native Americans, in order to examine the broad patterns of ancestry and genetic differentiation, as well as to ensure the absence of outliers in our data set.

Multidimensional scaling of genetic distances

Next, to assess the patterns of allele sharing between individuals, we estimated the *Outgroup* F_3(Y, Z; Mbuti), calculating for every pair of Y and Z indigenous groups. Additionally, a matrix of *Outgroup* F_3(Y, Z; Mbuti) calculated for all Y and Z pairs of individuals, was converted to genetic distances (Genetic distance = 1 - *Outgroup* F_3 estimate). A multidimensional scaling analysis (MDS) was then applied to the matrix of pairwise genetic distances with the "stats" R package.

D-statistics

First, we examined the presence of an excess allele sharing between all Native American groups in the unadmixed and unrelated dataset and present-day indigenous Papuans, Australians, Melanesians, and Andamanese, which was considered to be a signal of the ancestry contribution from the so-called "Population Y" (2). To accomplish this we used Admixtools (1) to estimate D-statistics (Mbuti, Austro-Melanesian; Y, Z) we defined "Australasian" as any Australasian group present in our dataset (i.e., Australian, Melanesian, Onge, or Papuan), and Y and Z as any modern Native American group or individual (e.g., Mixe, Karitiana, or Xavánte). Therefore we estimated the D-statistic for all pairs of Native American individuals and groups. The D-statistic as well as the standard error is estimated by qpDstat program from Admixtools (3), using a weighted Block Jackknife procedure, in which the genome is divided into blocks of 5 cM (default parameter) then multiple runs are executed deleting one block at a time, which allows the estimation of the statistic mean and standard error.

qpWave

Second, we used qpWave (1) to infer how many admixture flows from outside the American continent would be necessary to produce the genetic diversity of present-day Central (represented only by Mixe in this analysis) and South American indigenous groups. The qpWave software infers that if a given set f4(W, X; Y, Z) statistics are consistent with rank 0, 1, or 2 (or more), the test populations (W and X) derive from 1, 2, and 3 (or more) streams of ancestry from the outgroup populations (Y and Z), respectively. To do so, a set of tests in the form f4($test_1$, $test_2$; $outgroup_1$, $outgroup_2$) was

performed, following the original design used by Skoglund et al. (2). As test populations, we analyzed 14 indigenous groups with a minimum of 3 individuals (unadmixed and unrelated) and as outgroups, 4 populations from 6 world regions: Sub-Saharan Africa (Mbuti, Yoruba, Bantu-SouthAfrica, and Bantu-Kenya), Western Europe (Sardinian, French, Orcadian, and Spanish), East Asia (Han, Japanese, Miao, and Uygur), South Asia (Onge [ONG.SG], Sindhi, Cambodian, and Dai), Siberia/Central Asia (Mongola, Yakut, Oroqen, and Hezhen), Oceania (Papuan, Melanesian, and Australian) (Dataset S6A-B). Next, a series of tests were performed by dropping one of the above-mentioned regions at each time, followed by a series of tests performed by dropping one of the test groups (Native Americans) at each time, and finally, two sets of tests were performed by keeping Africa or Siberia/Central Asia plus one of the other regions at each time (Dataset S6A). Furthermore, we also tried to produce a more refined overview of the presence of these deeply divergent ancestries and to evaluate the extent of the variation between contemporary Native Americans; to accomplish this, qpWave was applied to every pair and trio of the test groups, including all worldwide regions mentioned above as outgroups, and the data is summarized in Dataset S6C-D.

qpgraph

Finally, we aimed to assess the population history models to investigate how this deeply diverged shared ancestry between present-day indigenous Australasians and South Native American groups emerged, especially now in the light of the new evidence (D-statistics) pointing to the existence of such affinity, not only amongst Amazonians (Karitiana and Suruí) but also in the Pacific coastal population (Chotuna) and populations from other Brazilian regions (Xavánte from the central Brazilian plateau). In this sense, we applied the models proposed and published by Skoglund et al. (2) as a scaffold admixture graph to model and test these additional groups (i.e., Chotuna and Xavánte). These groups were included by computing all possible positions in the admixture graph scaffold independently. To test the existence of this genetic affinity, we added the Pacific coastal groups Chotuna, Narihuala, and Sechura to the above-mentioned models; next, we also included Xavánte. We also tried another approach in which we started by first adding Xavánte and the Pacific coastal groups to the scaffold, and only then adding Suruí and

Karitiana. Finally, we compared the worst estimated Z-value of all computed models, selecting the candidates with the best fit to the data to represent the population history (i.e., tree topology and admixture events).

Treemix

We also aimed to produce an outline of the population history of the Native American groups here represented, by using an alternative method, distinct from the F-statistics (1, 11) framework. This was done with Treemix (12), which implements an unsupervised method of estimating a Maximum Likelihood tree of the population pairwise allelic covariances and allows the inference of putative gene flow between branches of the tree. First, we inferred the ML tree and then allowed the model to fit a growing number of gene flow events until a plateau of the model likelihood was reached.

Datasets (S1 to S6)

Dataset S1. Metadata and exploratory analysis of test samples.

This dataset includes (**A**) metadata for each Native American individual used in our analyses. The information includes original group name (as used in the data source study), group name (as used in this study), individual ID, major ethnolinguistic group affiliation, country of origin, data source study, data source method (e.g., Axiom Human Origins array or Shotgun sequencing), geographic coordinates, inclusion in the maximum unrelated dataset (True or False), and presence of non-Native American admixture (True or False). This dataset also presents estimates produced by an unsupervised ADMIXTURE (8) analysis on the subset of Native Americans with K = 3. The colors used to represent each individual throughout the study are also included. Additionally, it contains data for the PCAs performed with **(B)** the complete set of Native Americans and (**C**) the subset of unadmixed and unrelated samples. Finally (**D**) a multidimensional scaling analysis was performed on a matrix of the genetic distances (1 - Outgroup F_3) between all pairs of samples in the unrelated and unadmixed subset.

Dataset S2. Metadata for reference samples.

This dataset includes metadata for each individual from a reference population used in our analyses. The information includes group name, individual ID, country of origin, data source study, macro-region of origin, continent of origin, and data source method (e.g., Axiom Human Origins array or Shotgun sequencing).

Dataset S3. Significant Z-values for the D(Mbuti, Australasian; Y, Z) statistic.

The D-statistics (Mbuti, Australasian; Y, Z) (1) for every combination of an Australasian group (i.e., Australian, Australian.DG, Melanesian, Onge, or Papuan) and a pair of Y and Z American indigenous groups, with one set including related samples and another excluding related samples (i.e., maximum unrelated dataset). The complete set of D statistics are accessible in Dataset S4A-B.

Dataset S4. Estimates of D-statistics (Mbuti, Australasian; Y, Z) for every pair of Y and Z indigenous groups and individuals.

This dataset presents the D-statistics D(Mbuti, Australasian; Y, Z) (1) for every combination of an Australasian group (i.e., Australian, Melanesian, Onge, or Papuan) and a pair of Y and Z American indigenous groups, with (**A**) one set including related samples and (**B**) another excluding related samples (i.e., maximum unrelated dataset tab), and also (**C**) for every pair of Y and Z individuals in our dataset. It includes information on D-statistic, Z-Value, number of ABBA and BABA positions, and the total number of shared SNPs across the tested populations. Finally, (**D**) a summary of the number of significant tests per ID when they are at the Y and Z positions of the statistic is provided.

Dataset S5. Number of ancestry streams consistent with the Central and South American genetic diversity.

The consistency of 1 to 4 admixture flows between Central and South American indigenous groups (test populations - Dataset S1A) and other global-wide populations (reference populations - Dataset S2) was tested with qpWave (1). (**A**) A series of tests were performed using different combinations of reference populations as described in the first column. The remaining columns exhibit the *p*-value for each number of tests (i.e., 1 to 4), and significant values are marked with an asterisk. We also present **(B)** the group qpWave weights for the Full dataset analysis, along with a summary of the results of another series of qpWave analysis testing all **(C)** pairs and **(D)** trios of Native American groups.

Dataset S6. Maximum likelihood tree and gene flow events.

A maximum likelihood tree based on the population pairwise allelic covariances matrix was obtained with Treemix (12)) and an increasing number of gene flow events were adjusted by the model (up until 8 events). Here we present (**A**) the likelihood for each of these models, the covariance matrices (**B**) for the ML tree with no gene flow events and (**C**) for the model with 6 gene flow events, which is the first model to reach the likelihood plateau. Finally we (**D**) present the ML tree with no gene flow events and (**E**) the model with 6 gene flow events.

SI References

1. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

2. P. Skoglund, *et al.*, Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).

3. M. A. Castro e Silva, *et al.*, Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proceedings of the National Academy of Sciences* **117**, 2372–2377 (2020).

4. I. Lazaridis, *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

5. S. Mallick, *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

6. C. Barbieri, *et al.*, The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol. Biol. Evol.* **36**, 2698–2713 (2019).

7. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

8. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

9. J. Staples, D. A. Nickerson, J. E. Below, Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet. Epidemiol.* **37**, 136–141 (2013).

10. X. Zheng, *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

11. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

12. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

Material suplementar do Capítulo 2

Supplementary Materials for:

Population histories and genomic diversity of South American natives

Marcos Araújo Castro e Silva[a], Tiago Ferraz[a], Cainã M. Couto-Silva, Renan B. Lemes[a], Kelly Nunes[a], David Comas[b] and Tábita Hünemeier[a,2]

[a]*Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil;* [c]*Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain*

[2]Corresponding author: E-mail: hunemeier@usp.br

Supplementary Figures

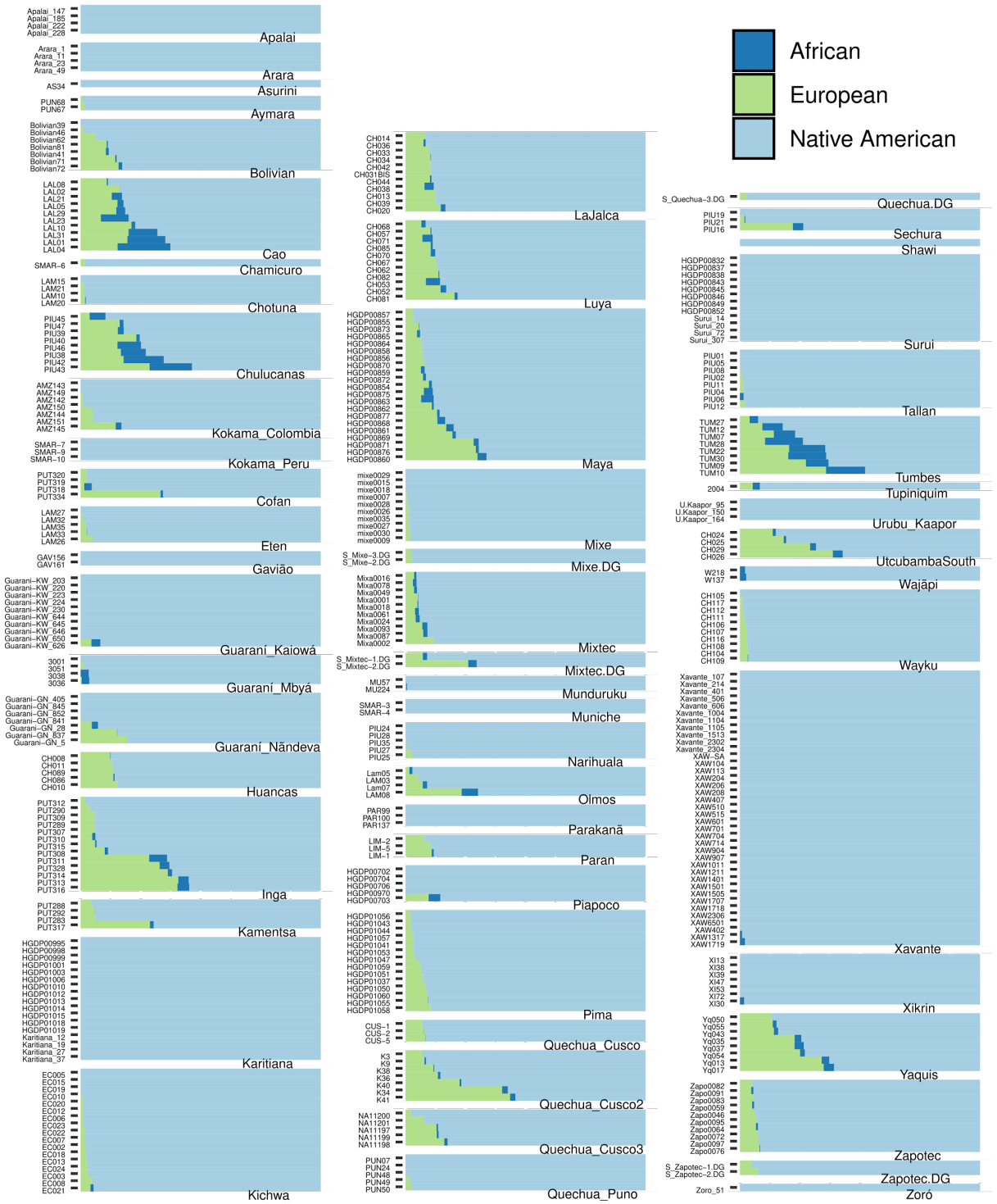


Figure S1 - Non-Native American admixture profile of American indigenous groups. An unsupervised ADMIXTURE (Alexander et al. 2009) analysis on the subset of Native Americans with K = 3 was performed in order to estimate their proportions of African and European admixture. The African populations used as proxies of the parentals were Bantu from Kenya, Bantu from South Africa, Biaka, Mandenka, Mbuti and Yoruba from HGDP, additionally the European populations used as parentals were Basque, French, Italian, Orcadian, Sardinian, Tuscan from HGDP, and Basque, Southern Italian, Sicilian, Spanish, Northern Spanish from Lazaridis et al. (Lazaridis et al. 2014; Lazaridis et al. 2016). The three panels exhibit the individual proportions of African (dark blue), European (green) and Native American

(light blue) ancestries, individual labels are placed on the left side of each bar and group labels are placed at the bottom right of each group. A total of 150 unadmixed individuals can be identified and they are listed on the Dataset S1.

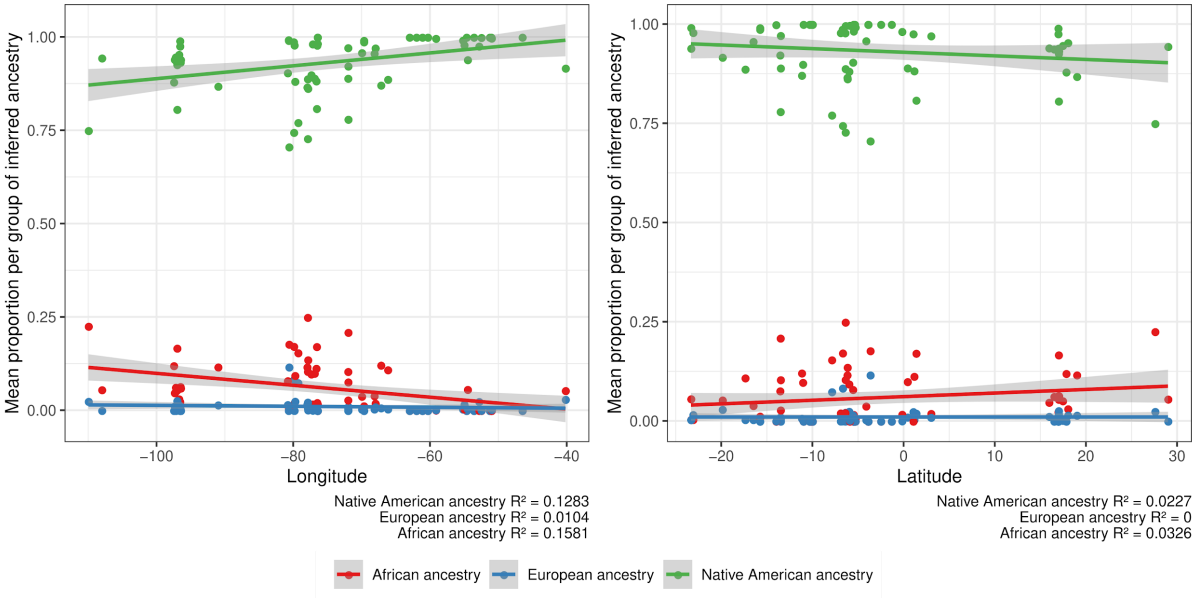


Figure S2 - Admixture proportions components in relation to geography. Here we use the continental ancestry components in the analysis presented in Figure S1, to test if the ancestry proportions were correlated with the geographic position (longitude and latitude) of each group. The two panels show the mean proportion of each inferred ancestry (see bottom legend) for each group in the form of color coded points as functions of their longitude (left) and latitude (right), additionally a linear regression model is fitted for each ancestry component along with their 95% confidence interval presented as shaded areas. The coefficients of determination (R^2) of each model are shown at the bottom right of both panels.

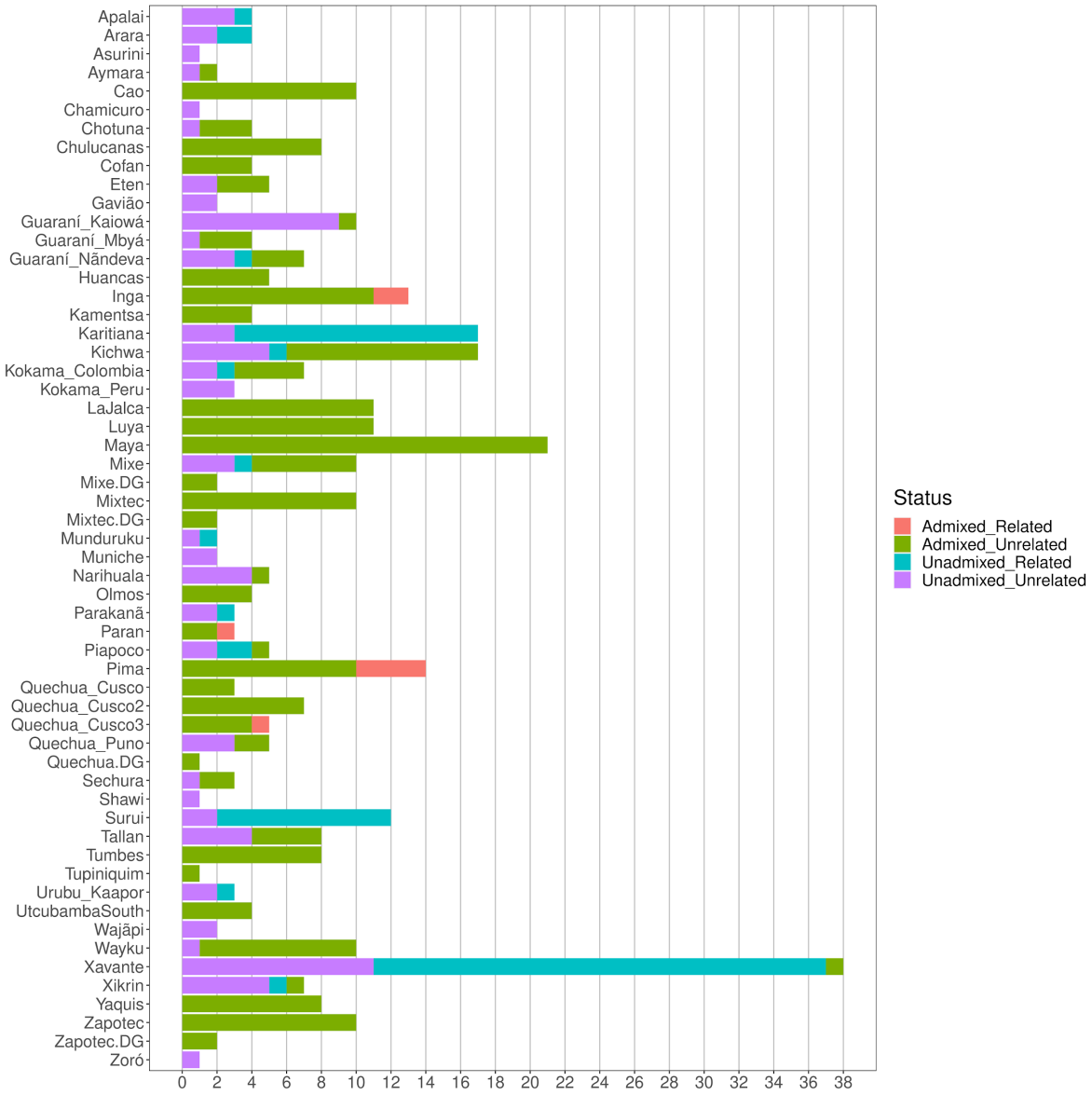


Figure S3 - Assessing non-Native American admixture and relatedness in the set of Native American groups. Ids are classified into the categories unadmixed and unrelated, respectively when they present more than 99% of inferred Native American ancestry in ADMIXTURE (Alexander et al. 2009) analysis and are selected in the maximum unrelated (or independent) set of ids with a PI-HAT < 0.375 (1st degree relatedness) with PRIMUS (Staples et al. 2013). Wajãpi individuals were included despite having a non-negligible contribution from non-Native american ancestors (~ 3%; see Figure S1 and Dataset S1).

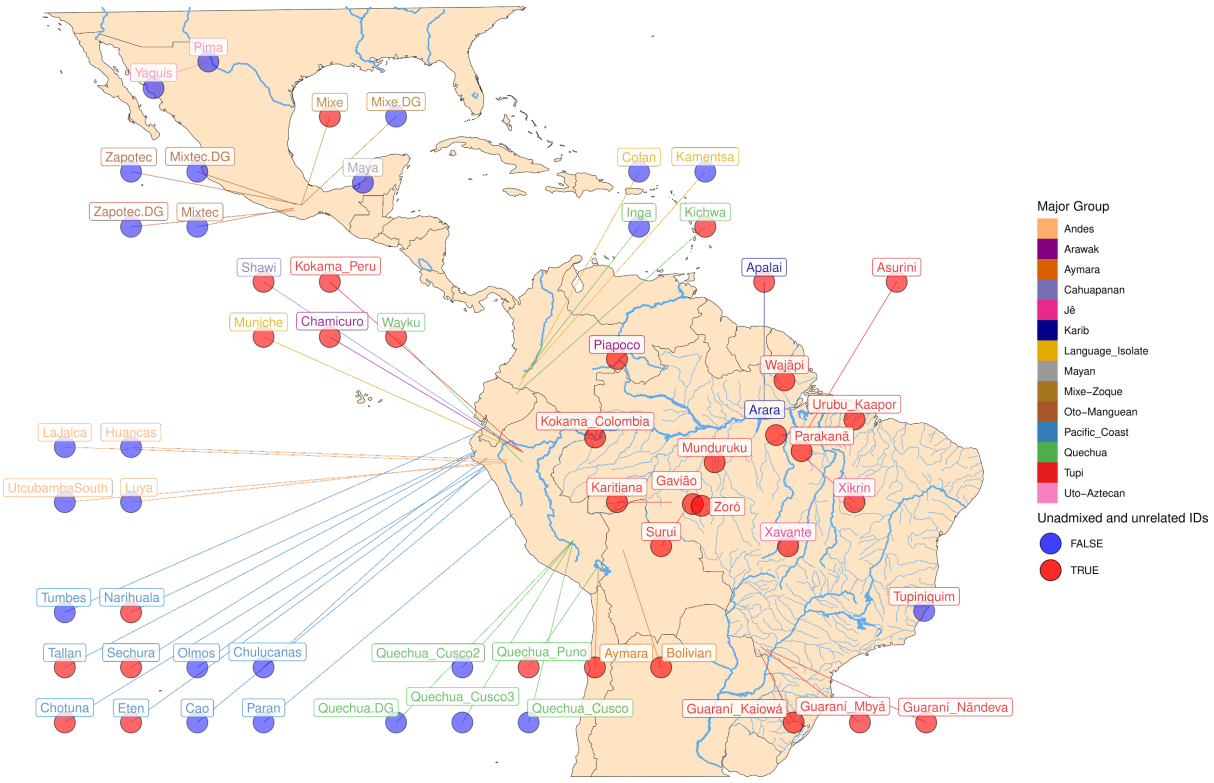


Figure S4 - Map of indigenous groups from the American continent. Labels and circles indicate the group names and their approximate location. Affiliation to the major groups used throughout this paper is color coded in the labels and indicated on the legend at the right side of the plot. Finally, if a given group contains at least one individual unadmixed and unrelated to any other individual in the dataset, the circle is colored red and when this is not the case, blue.

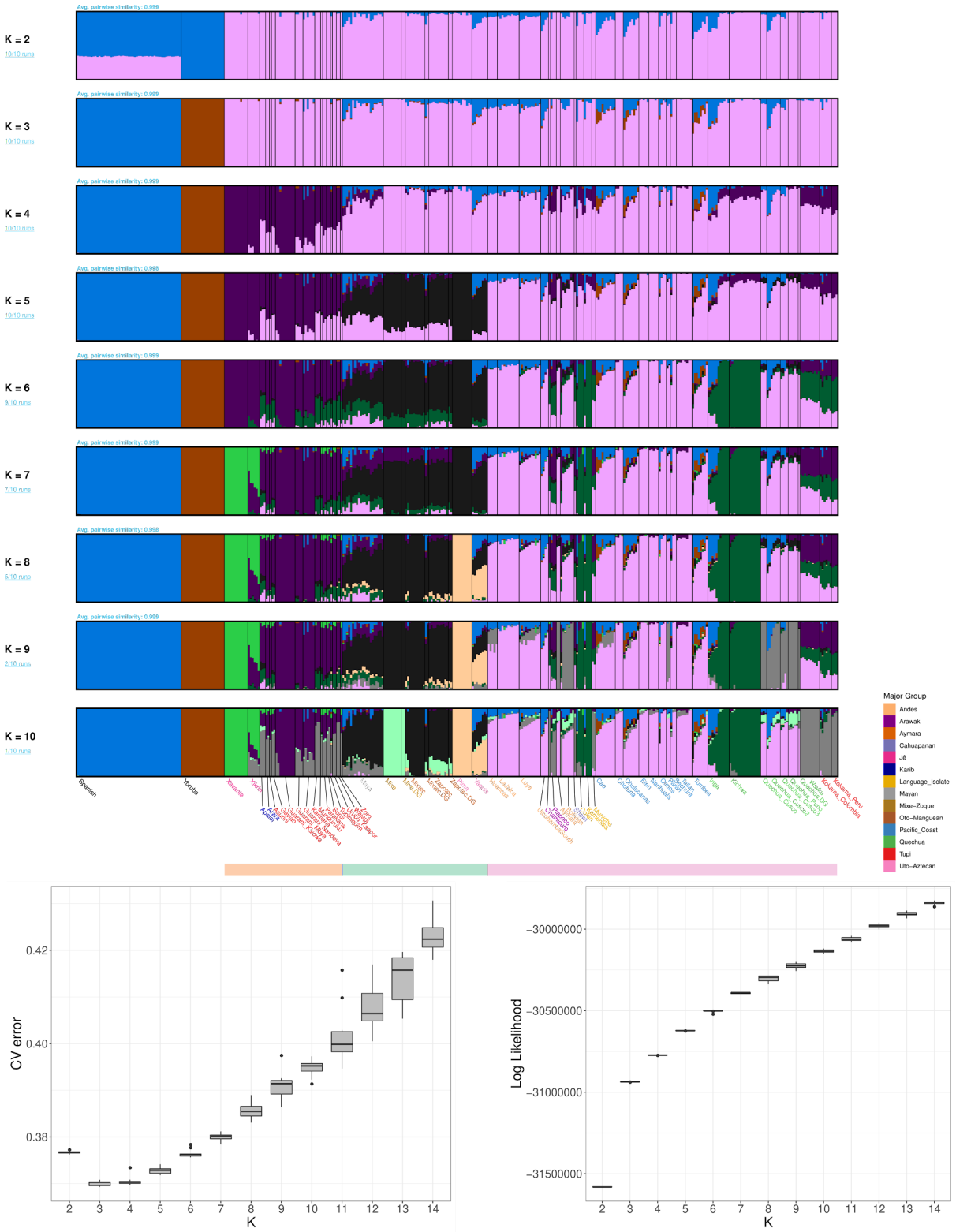


Figure S5 - Genetic structure and European and African admixture in Native Americans. Individual values of the putative ancestry components were estimated with an unsupervised ADMIXTURE (Alexander et al. 2009) analysis of the complete set of Native Americans with K values from 2 to 10 (Top panels) and the barplots of the estimates were produced with PONG (Behr et al. 2016). The number of putative ancestry components tested increases from top to bottom, group labels are given at the bottom of the last barplot and they are color coded to indicate their affiliation to major groups, as shown in the legend at the bottom right. The three main continental regions are indicated by the colored bar at the bottom: Mesoamerica in light green, western South America in pink and eastern South American in beige.

The cross-validation error (bottom left) and likelihood (bottom right) of each iteration of the algorithm were also estimated by ADMIXTURE.

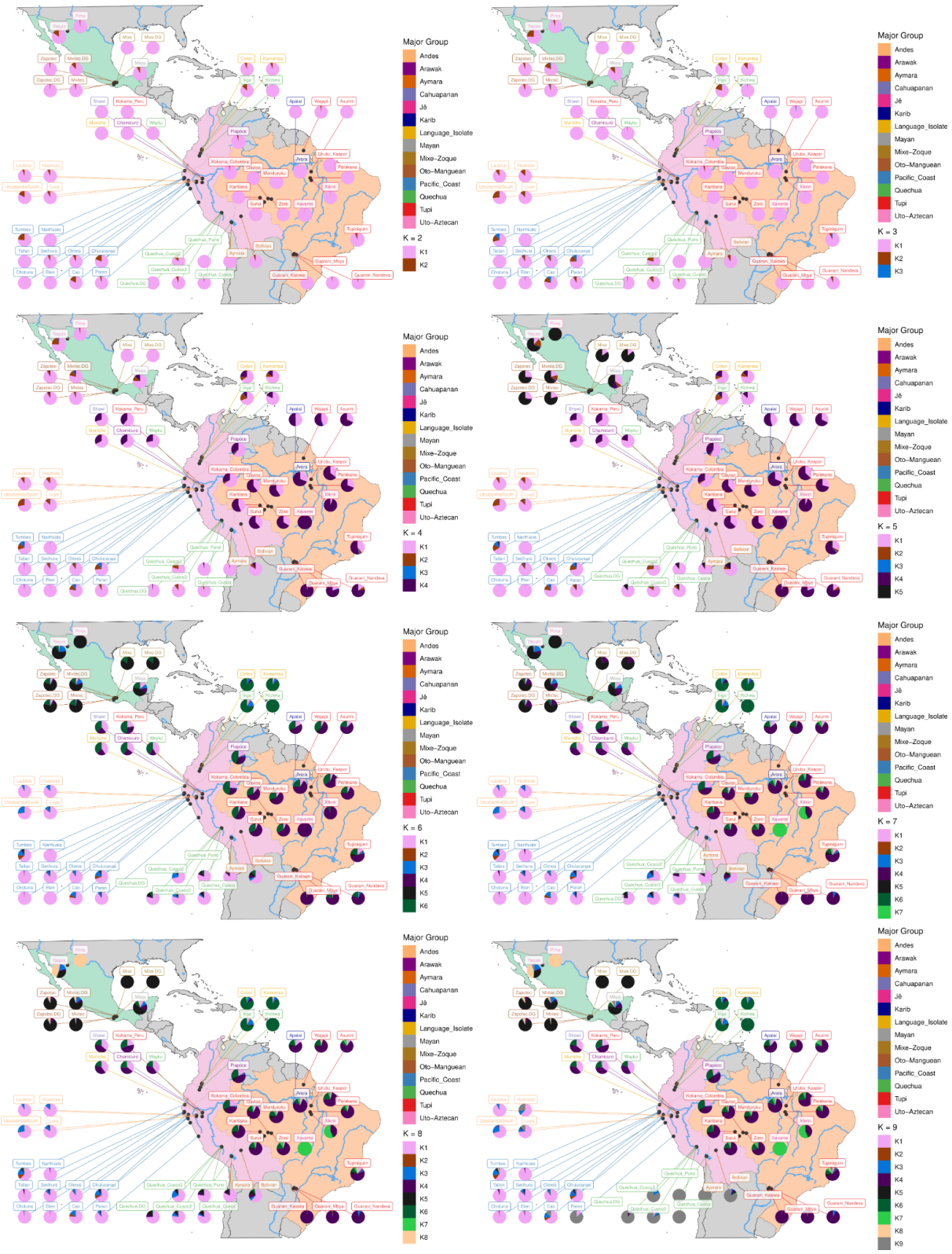


Figure S6 - Mean ancestry components of the complete set of Native American groups. Population mean values of the putative ancestry components were estimated with an unsupervised ADMIXTURE analysis with K values from 2 to 10 of the complete set of Native Americans (same estimates used in Figure S4) and plotted as pie charts a map of Central and South America, based on the approximate sampling location of each group. The linguistic affiliations as well as the putative ancestry components (K), are color coded as indicated in the legend at the right.

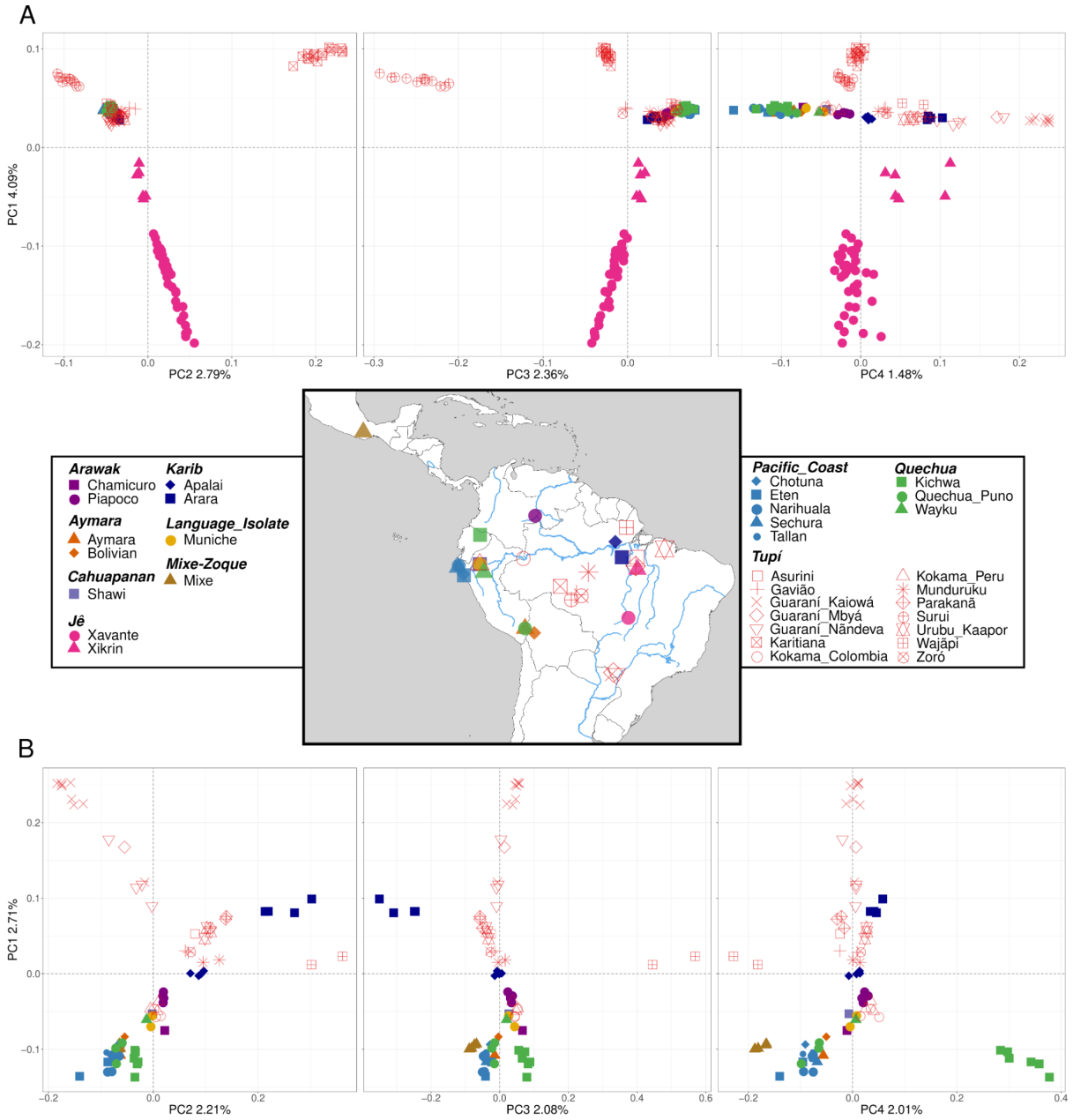


Figure S7 - Broad patterns of shared ancestry among Native Americans. A PCA was applied to **(A)** the LD-pruned set of unadmixed and unrelated Native Americans and to **(B)** a subset excluding the most divergent groups, namely: Xavante, Xikrin, Karitiana, and Suruí. The plot exhibits the combinations of the first PC (x-axis) with the second to fourth PCs (y-axis), from left to right: PC1 and PC2; PC1 and PC3; PC1 and PC4. The groups and major groups affiliations are coded as shapes and colors, respectively, as indicated in the legend at the center of the plot.

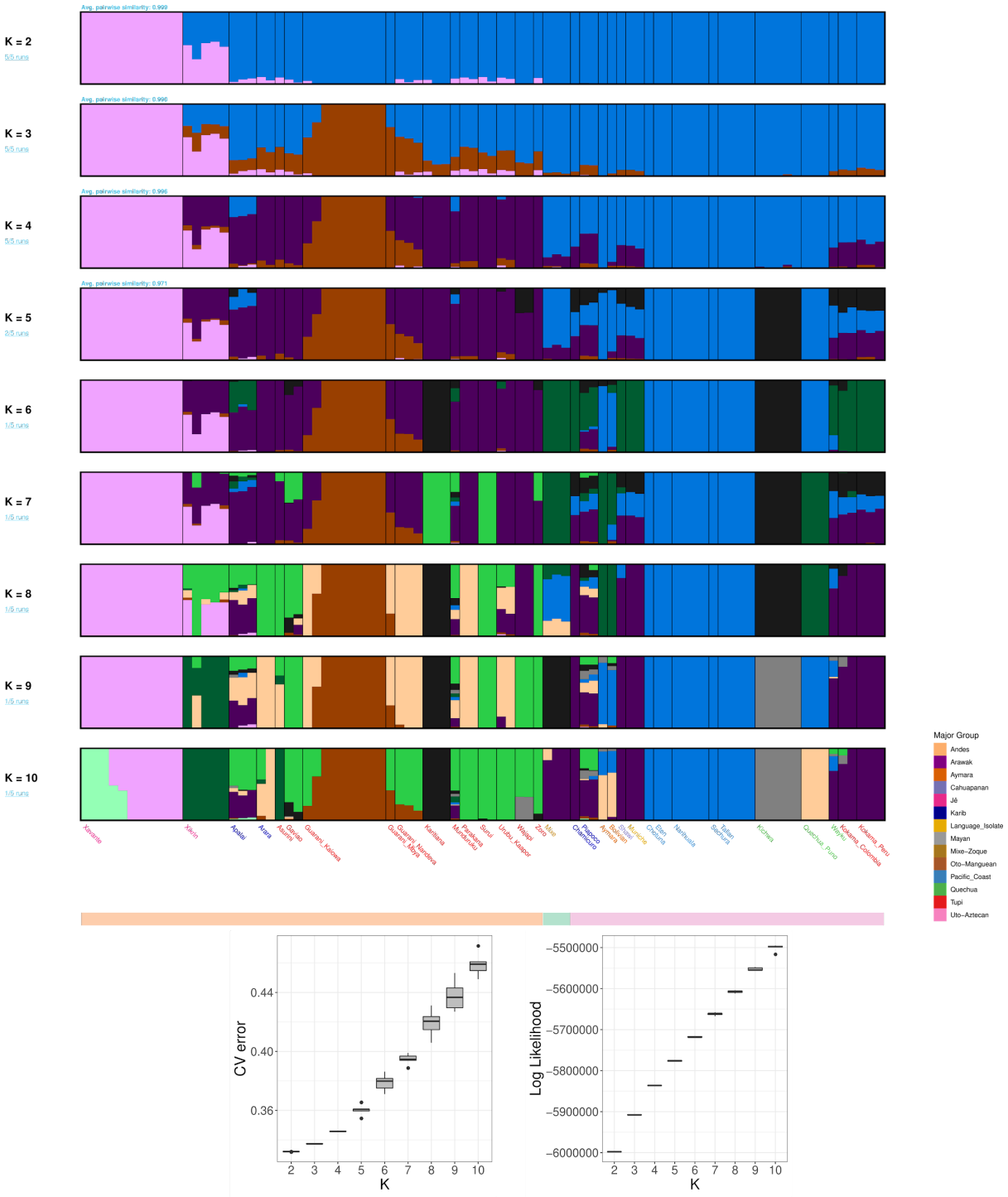


Figure S8 - Genetic structure and patterns of shared ancestry in unadmixed and unrelated Native Americans. Individual values of the putative ancestry components were estimated with an unsupervised ADMIXTURE (Alexander et al. 2009) analysis of the subset of unadmixed and unrelated Native Americans with K values from 2 to 10 (Top panels) and the barplots of the estimates were produced with PONG (Behr et al. 2016). The number of putative ancestry components tested increases from top to bottom, group labels are given at the bottom of the last barplot and they are color coded to indicate their affiliation to major groups, as shown in the legend at the bottom right. The three main continental regions are indicated by the colored bar at the bottom: Mesoamerica in light green, western South America in pink and eastern South American in beige. The cross-validation error (bottom left) and likelihood (bottom right) of each iteration of the algorithm were also estimated by ADMIXTURE.

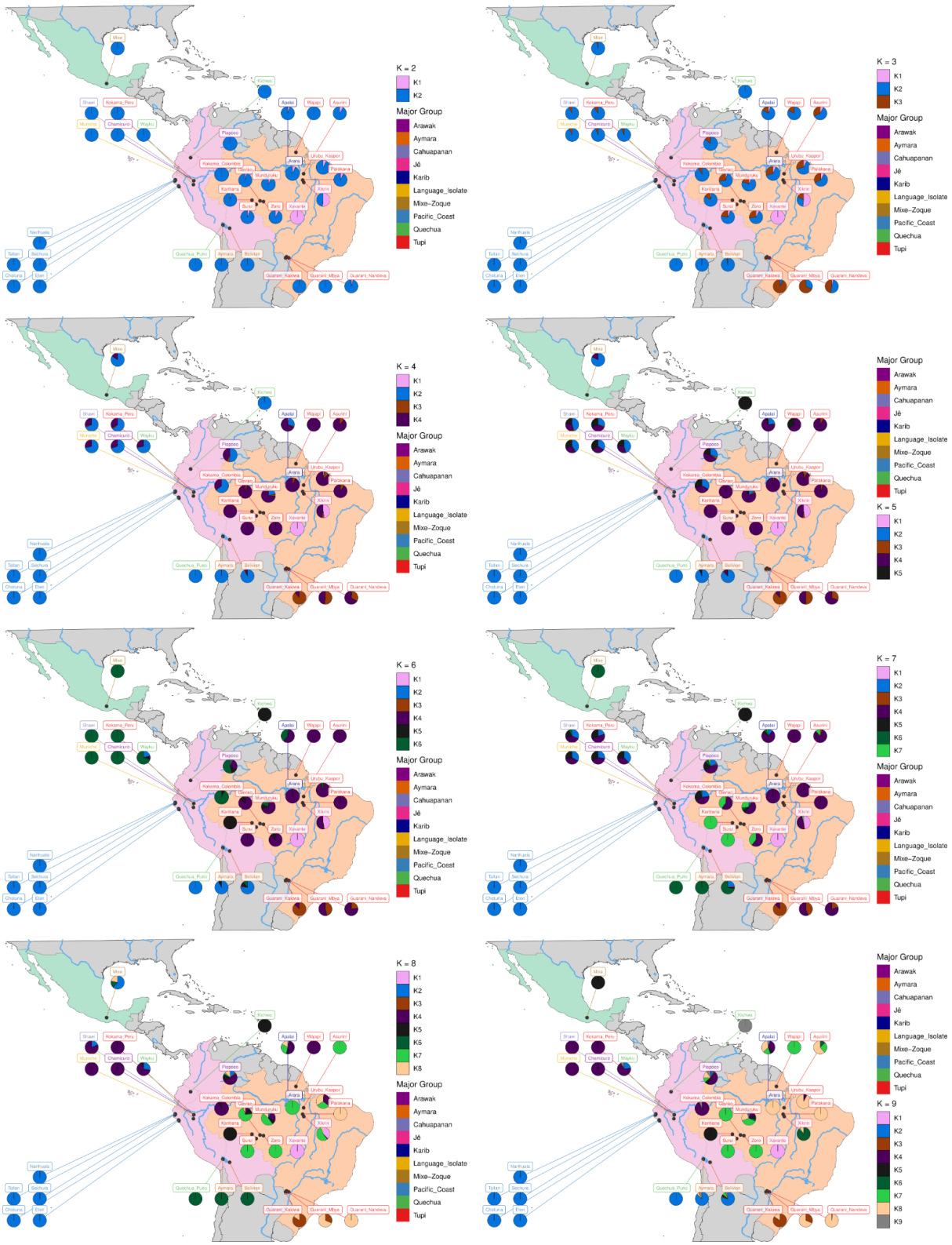


Figure S9 - Mean ancestry components of the subset of unadmixed and unrelated Native American groups. Population mean values of the putative ancestry components were estimated with an unsupervised ADMIXTURE analysis with K values from 2 to 10 of the subset of unadmixed and unrelated Native Americans (same estimates used in Figure S5) and plotted as pie charts a map of Central and South America, based on the approximate sampling location of each group. The linguistic affiliations as well as the ancestry components (K), are color coded as indicated in the legend at the right.

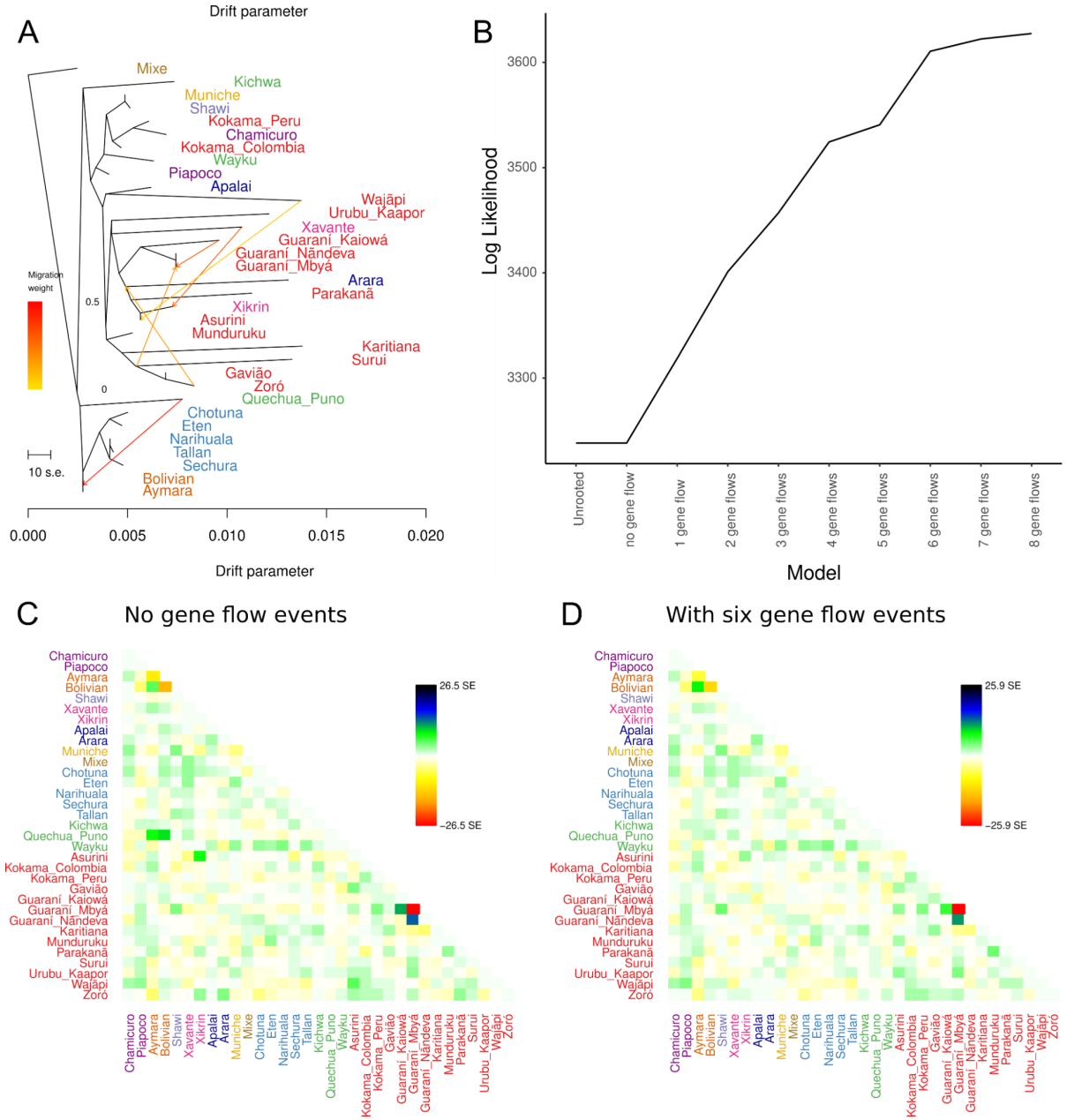


Figure S10 - Maximum likelihood tree modelling. Using Treemix (Pickrell and Pritchard 2012) Maximum Likelihood (ML) trees were estimated and gene flow events were modeled between poorly fitted branches. **(A)** ML tree with six gene flow events. **(B)** Plot showing the likelihood (y-axis) for each model (x-axis). **(C)** Residue matrix for the no gene flow model. **(D)** Residue matrix for the six gene flow events model.

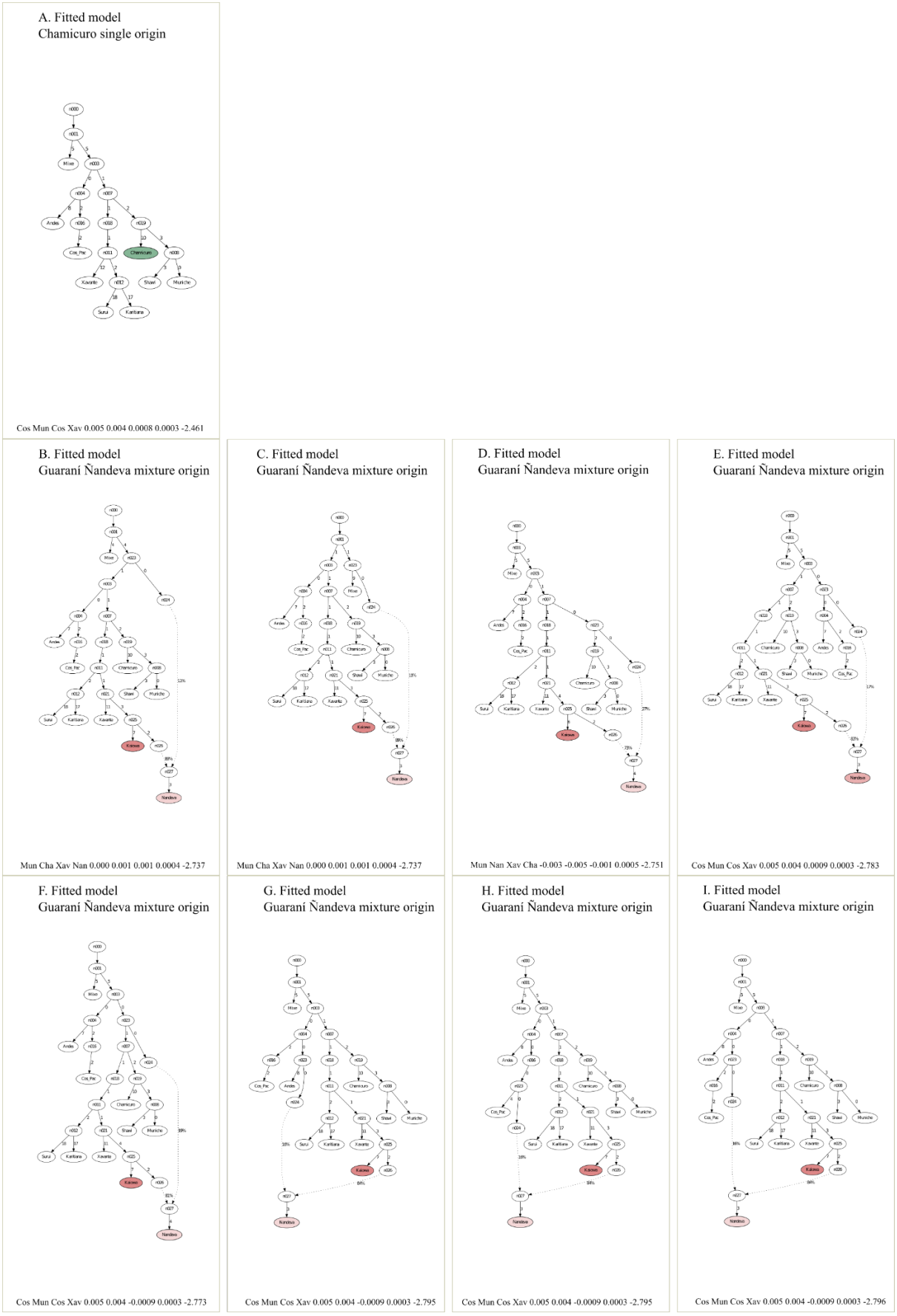


Figure S11 - Complementary set of admixture graph models. (A) Graphic representation of the best fitted model for a single origin of Chamicuro - Arawak speakers - this model was used as a scaffold tree to

build up all the other models. **(B-I)** Models with good fit to the data for a mixture origin of the Guaraní Ñandeva. Note that **B** is the same model presented in FigureFigure 3C.

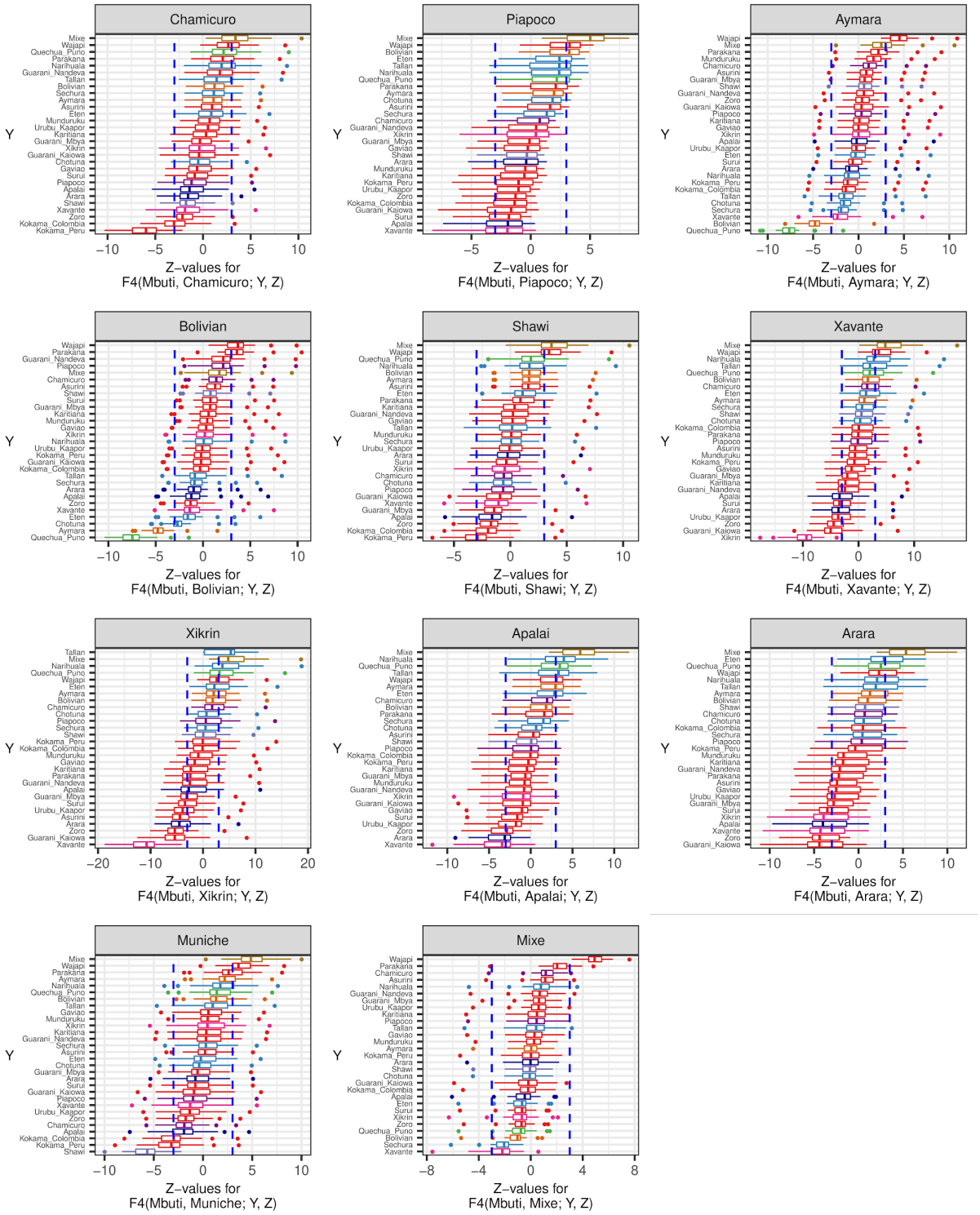


Figure S12 - Genetic affinities among present-day indigenous groups. To examine the patterns of allele sharing we estimated F_4(Mbuti, X; Y, Z) for every combination of X, Y and Z present-day Native American groups. Each panel shows the combinations of X (top stripe) and Z (y-axis) test groups, and the Z-values (x-axis) obtained by the comparison with every Z test group are presented in the form of boxplots. Panels ordered by major groups. The complete set of statistics is presented in Dataset S6A. (Start)

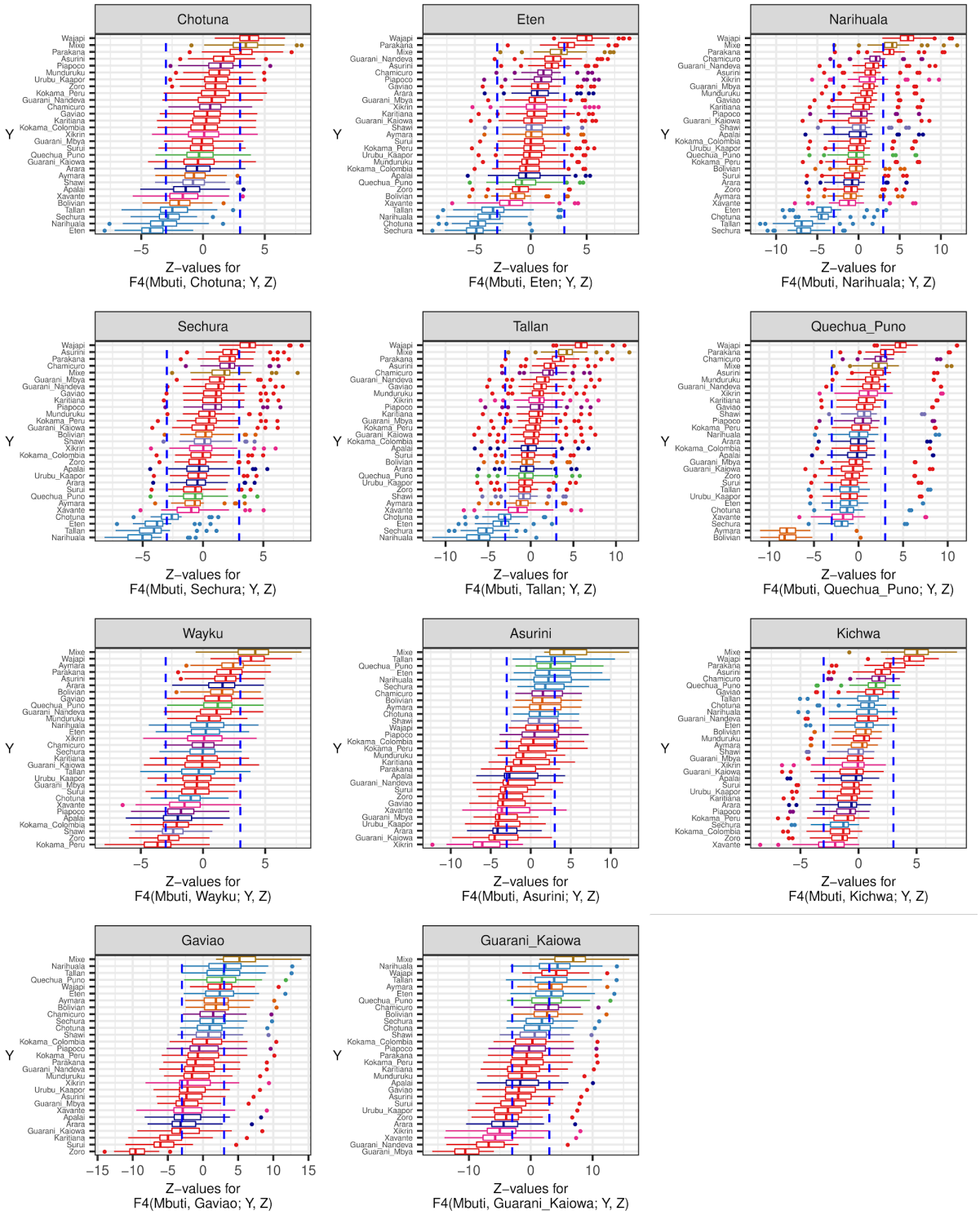


Figure S12 (continued)

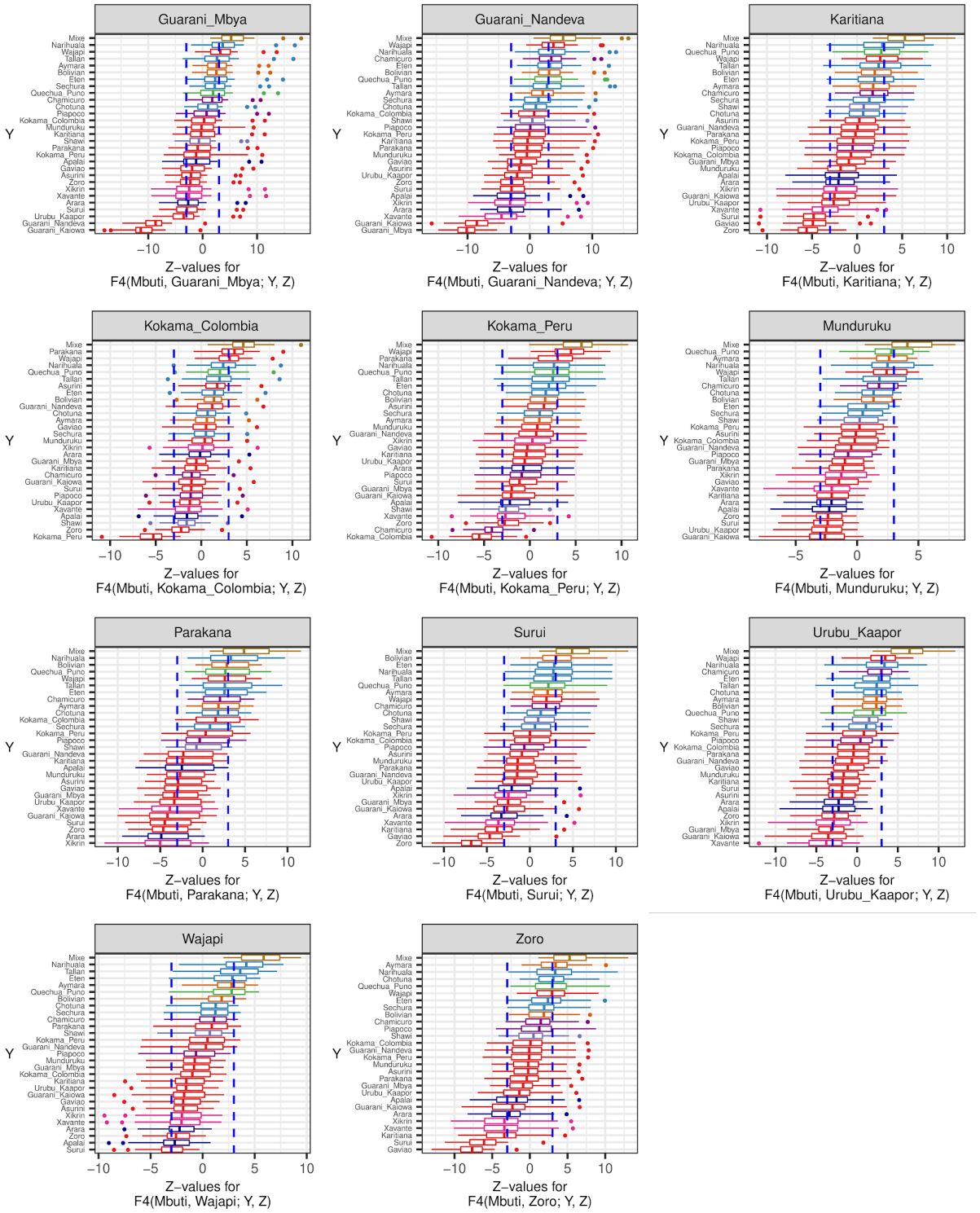


Figure S12 (continued)

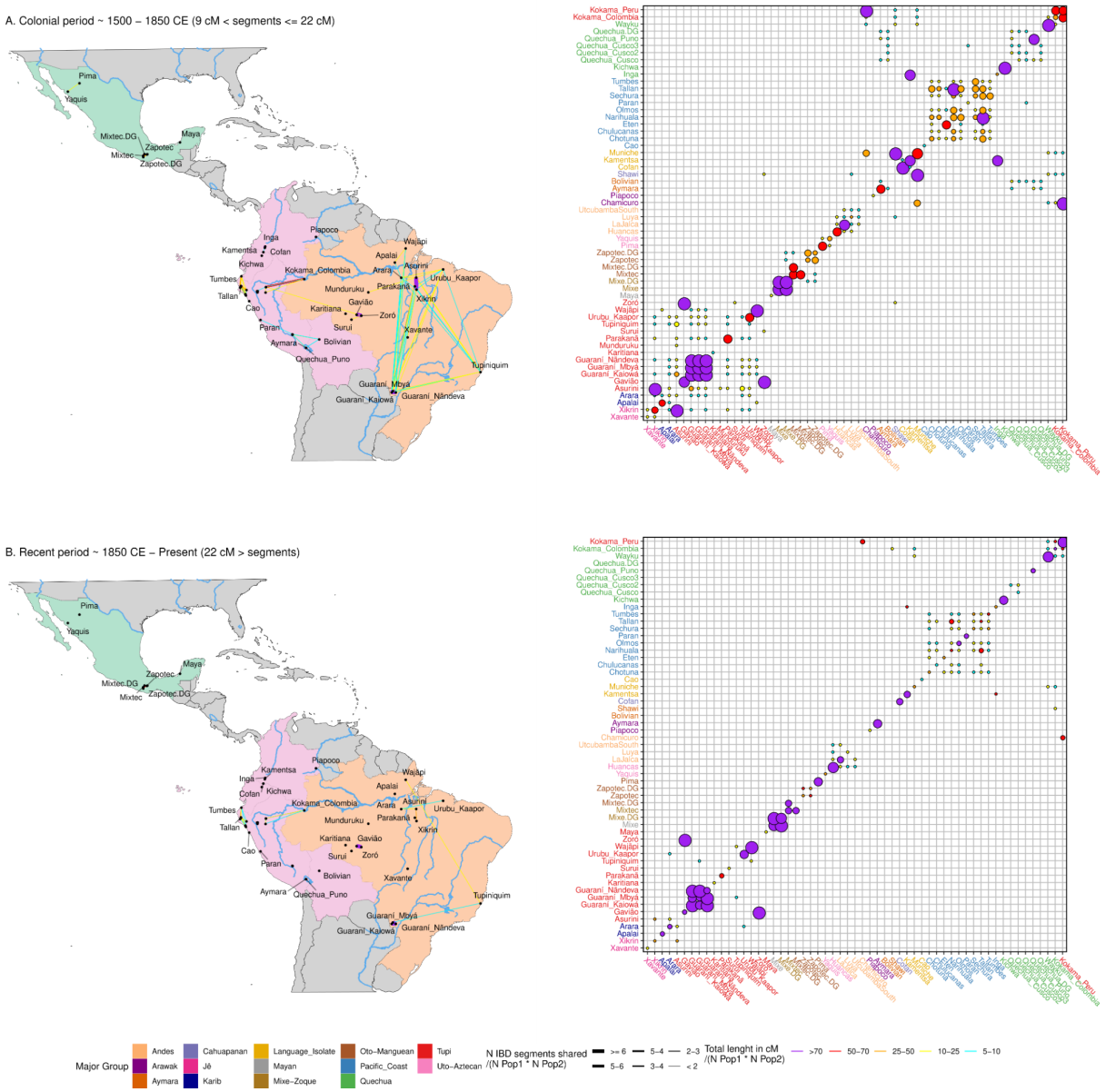


Figure S13 - **Levels of genetic connections among Native American groups.** Here we present two networks and matrices created by estimating the average length of total IBD sharing and the average number of IBD segments, based on the same subsets of IBD genomic segments with length **(A)** between 9 and 22 cM and **(B)** with more than 22 cM. In this sense, the IBD genomic segments were identified based on the phased data subset of unrelated Native Americans, then they were filtered to select only those inferred to be in genomic regions of Native American local ancestry. We also removed segments shorter than 2 cM and pairwise connections with less than 5 cM shared on average were also not considered. Each map and matrix exhibits the average number of IBD segments (color) and the average length of IBD in cM (size), as indicated in the legends at the bottom. The three main continental regions are indicated by the colors used in each map. The complete set of IBD segments inferred are presented in Dataset S4.

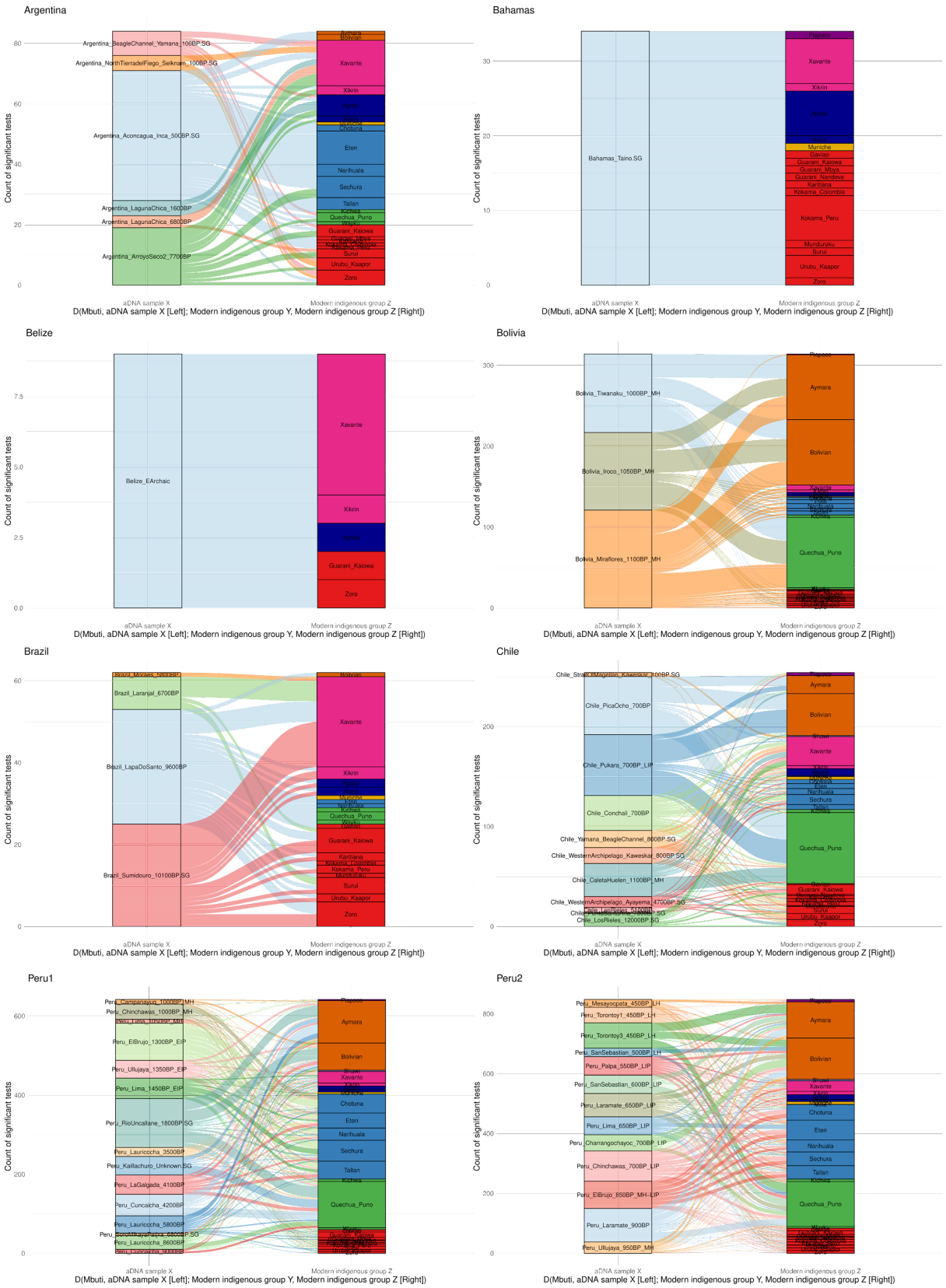


Figure S14 - Genetic affinities between present-day indigenous groups and ancient individuals (aDNA). To examine the patterns of allele sharing we estimated F_4(Mbuti, X; Y, Z) for every combination of X ancient individuals, and Y and Z present-day Native American groups. Ancient individuals from different countries are shown in separate panels, as indicated in their top left. Each panel shows the number of highly significant F_4 statistics (i.e. Z-value > 4) in the y-axis for each pair of X aDNA sample (left) and Z

present-day indigenous group (right), iterated over all Y groups. Metadata for the ancient samples used are included in Dataset S3 and the complete set of statistics is present in Dataset S6B.

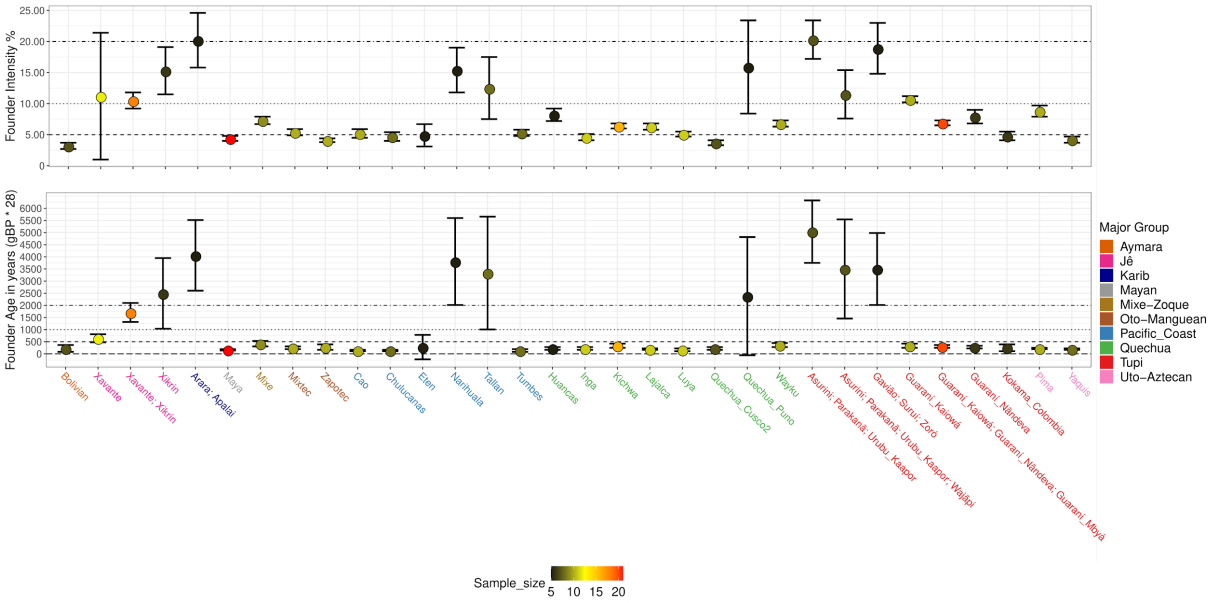


Figure S15 - Population bottlenecks in the history of Central and South American natives. The ASCEND method was applied to Native American groups with more than 5 unrelated samples and also to some clusters of groups (in order to reach the minimum sample size of 5). Here are presented the estimates of the complete set of clusters and groups **(A)** (the subset containing only the estimates for groups is presented in Figure 8). The top panels depict the founder intensity (FI) and the bottom panels exhibit the mean estimate for the founder age (FA) for each indigenous group or cluster of groups. For each group the estimated FI and FA are shown, along with their associated 95% confidence interval. The sample size is color coded on the points and the affiliations with major groups are indicated in the label IDs at the x-axis, both indicated in the legend. In the top panels the y-axis indicates the FI percentage and in the bottom panel the y-axis exhibits the estimated FA calculated as: 'x' generation before present (gBP) * 28 years per generation = 'y' years before present (BP).

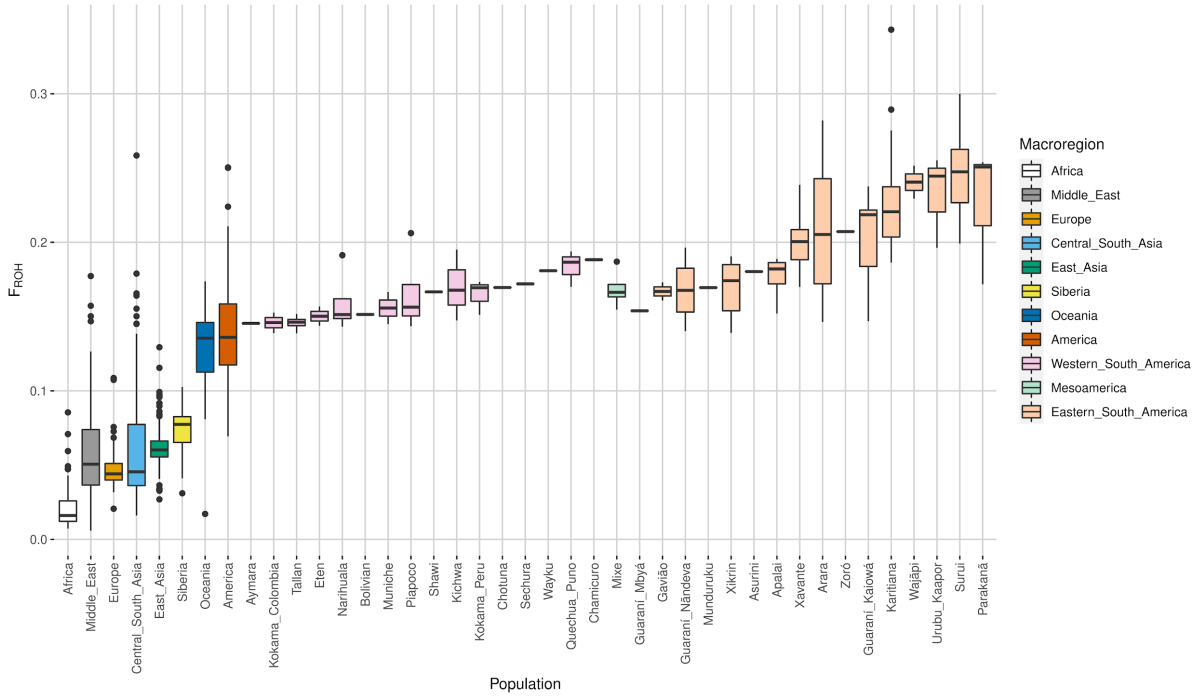


Figure S16 - Distribution of inbreeding coefficient per population. The distribution of F_{ROH} was obtained averaging the individual estimates from HGDP and SGDP databases (Africa, Middle East, Europe, Central South Asia, East Asia, Siberia, and America) and from each unadmixed Native American population independently.

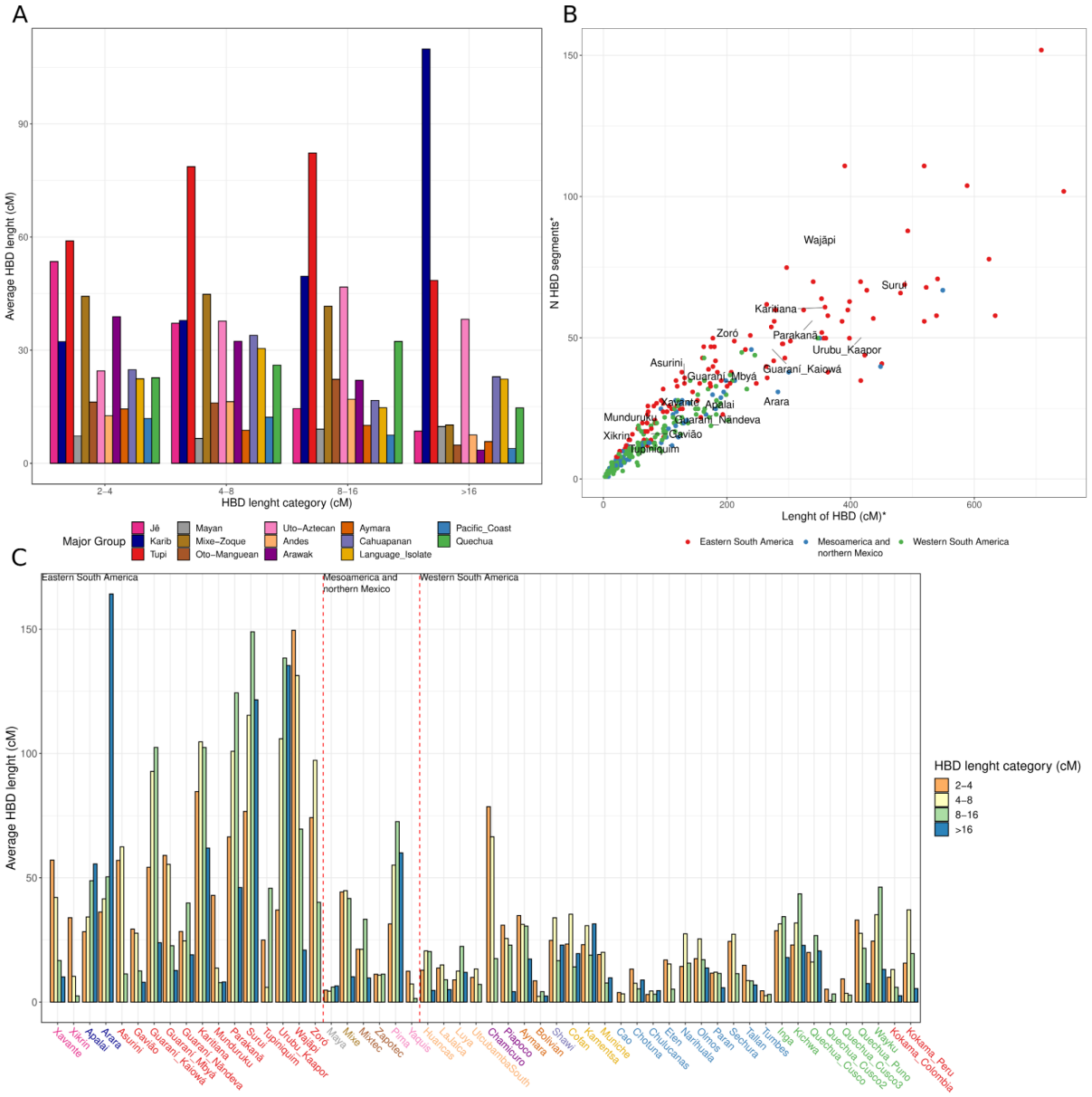


Figure S17 - Patterns of intrapopulation sharing of identity-by-descent (IBD) and homozygosity-by-descent (HBD) segments. The IBD and HBD genomic segments were identified with Refined IBD (S.R. Browning et al. 2018), based on the phased subset of Native American populations. Only IBD segments located in regions of local Native American ancestry as inferred with RFMix (Maples et al. 2013) were used. **A)** The average HBD total length (y-axis) was estimated for each major group (color coded) and binned into 5 segment length categories (x-axis). **B)** Number of HBD segments (y-axis) plotted against the total length of HBD in cM (x-axis). *Absolute values for all individuals and average values for eastern South American groups are represented by filled circles and labels, respectively. Color code indicates whether the group or individual is from Central America, eastern South America or western South America. **C)** The average HBD total length (y-axis) was estimated for each group (x-axis) and binned into 5 segment length categories (color coded). **D)** The average IBD length shared between individuals from the same groups (y-axis) was estimated for each group (x-axis; some groups are absent, as they have only 1 individual or none intra-population shared IBD) and binned into 4 segment length categories (color coded; the category 1–2 cM is absent). In **C** and **D**, populations are sorted by geographic location, from left to right: eastern South America (Xavante to Wajãpi), Mesoamerica (Maya to Yaquis), and western South America (Piapoco to Kokama Peru), as indicated by the red dashed lines and the top labels.

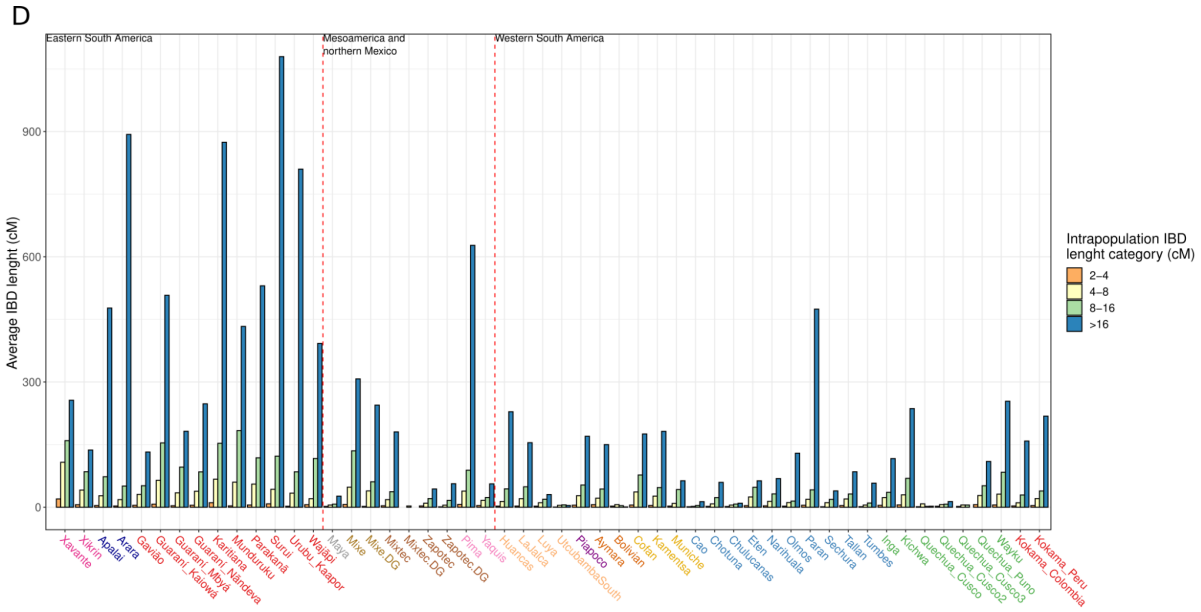


Figure S17 (continued)

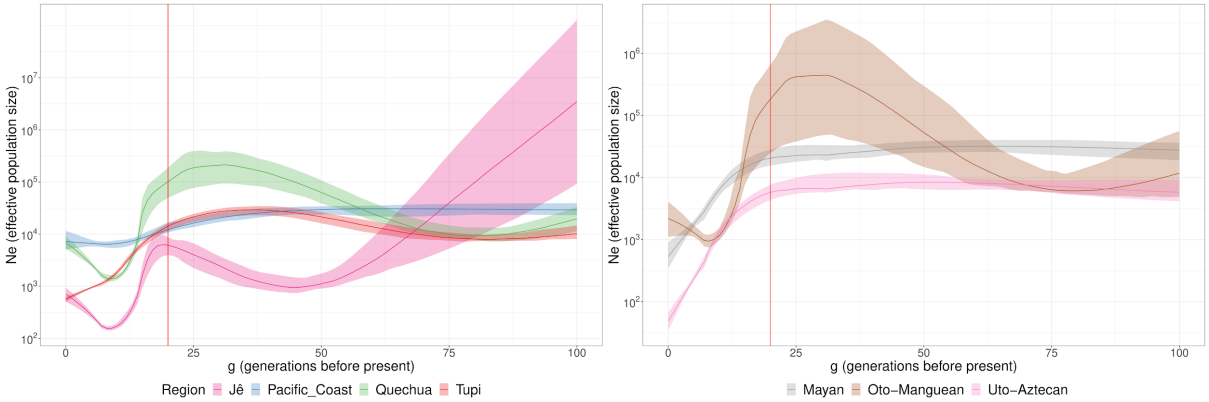


Figure S18 - Native American effective population size (N_e) histories. The IBD genomic segments were identified with the phased data subset of Native American groups, followed by a selection of the segments inferred to be in genomic regions of Native American ancestry. The complete set of IBD segments was separated into subsets of major groups from South America **(left)** and Mesoamerica **(right)**, and then each set was used to infer the N_e history of each specific major group. The ancestry-specific Ne values are coded in the y axis (log scale) and indicated by the line for each generation before present (gBP) depicted in the x axis. The shaded areas show a 95% bootstrap confidence interval for each major group. The vertical red line indicates 20 gBP (approximately 1500 C.E.) and therefore the time of the first contacts with Europeans. Here we show the results of IBDNe using the parameter filtersampes = "true", alternatively the results produced with the parameter filtersampes = "false" are shown in Figure 8.

Supplementary Data Files

Dataset S1 - Metadata for test samples. This dataset presents metadata for each present-day Native American individual used in our analyses, the information includes for example: original group name (as used in the data source study), group name (as used in this study), individual ID, major ethnolinguistic group affiliation, country of origin, data source study, data source method (e.g. Axiom Human Origins array or Shotgun sequencing), geographic coordinates, inclusion on the maximum unrelated dataset (True or False) and presence of non-Native American admixture (True or False). This dataset also presents estimates produced by an unsupervised ADMIXTURE (Alexander et al. 2009) analysis on the subset of Native Americans with K = 3. These estimates are the same presented in Figure S1. Finally the colors used to represent each individual throughout the study is also included.

Dataset S2 - Metadata for reference samples. Here we present metadata for each individual from a reference population used in our analyses, the information includes: group name, individual ID, country of origin, data source study, macro region of origin, continent of origin, data source method (e.g. Axiom Human Origins array or Shotgun sequencing).

Dataset S3 - Metadata for ancient samples (aDNA). This dataset contains metadata for each ancient individual sample used in our analyses, as obtained from 1240K+HO (v42.4; https://reichdata.hms.harvard.edu/pub/datasets/amh_repo/curated_releases/index_v42.4.html) curated dataset, including information on: individual IDs, study publication, representative contact, date mean in BP, date CIs, group label, locality, country, geographic coordinates, data source method (e.g. 1240K or Shotgun sequencing), coverage, library type and quality assessment.

Dataset S4. Inferred IBD segments of Native American ancestry. This dataset presents the complete set of IBD segments inferred from genomic regions of Native American local ancestry based on the subset of unrelated present-day Native Americans, and includes information about: group identification, individual identification, haplotype index, chromosome, starting and ending genomic positions, and length in cM for each IBD segment.

Dataset S5. Estimates of outgroup F_3(Y, Z; Mbuti). This dataset includes the F_3-statistics for every combination of Y and Z indigenous groups **(A)** or individuals **(B)**, in the unadmixed and unrelated subset of present-day Native Americans, as well as the F_3-statistics for every pair of Y and Z present-day and/or ancient individuals **(C)**. The datasets include information on F_3-statistic, standard error (SE), Z-value, and the total number of shared SNPs across the tested populations.

Dataset S6. Estimates of F_4(Mbuti, X; Y, Z). This dataset presents **(A)** the F_4-statistics F_4(Mbuti, X; Y, Z) for every combination of X, Y and Z present-day American indigenous groups; and also contains **(B)** the F_4-statistics F_4(Mbuti, X; Y, Z) for every combination of X ancient individuals, with pairs of Y and Z present-day American indigenous groups. Present-day samples include only those present in the unadmixed and unrelated subset. The datasets present information on F_4-statistic, Z-Value, number of ABBA and BABA positions, and the total number of shared SNPs across the tested populations.

Material suplementar do Capítulo 3

PNAS
www.pnas.org

SI Appendix for

Genomic insight into the origins and dispersal of the Brazilian Coastal Natives

Marcos Araújo Castro e Silva[a], Kelly Nunes[a], Renan Barbosa Lemes[a], Àlex Mas-Sandoval[b,c], Carlos Eduardo Guerra Amorim[d], Jose E. Krieger[f], José Geraldo Mill[e], Francisco M. Salzano[b,*], Maria Cátira Bortolini[b], Alexandre Pereira[f], David Comas[c] and Tábita Hünemeier[a,1]

[a]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil; [b]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil; [c]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain; [d]University of California, Los Angeles, United States of America, [e]Departamento de Fisiologia, Universidade Federal do Espírito Santo, Espírito Santo, Brazil. [f]Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil

Tábita Hünemeier
Email: hunemeier@usp.br

www.pnas.org/cgi/doi/10.1073/pnas.1909075117

Summary

SI Appendix Text

Dataset Assembly

In the present study, we analyzed 47 Tupiniquim and 48 Guaraní Mbyá new samples from Aracruz, state of Espírito Santo, Brazil (Table S1). We genotyped the samples in the Axiom InCor BB (Affymetrix) array with 842.019 SNPs. This customized array has a relatively low proportion of overlapped SNPs with current commercial arrays (e.g., *Axiom Human Origins* (1)) with approximately 70K shared SNPs across platforms. For this reason, we assembled a series of datasets in order to execute different analyses and investigate various aspects of the data (Table S3). We also included some newly-genotyped data from Native American populations, which are 2 Wajãpi, 3 Parakanã, and 2 Gavião, provided by Prof. Francisco Mauro Salzano (Table S1).

After an estimation of the admixture proportions of putative ancestral components (i.e. African, European and Native American) in the Tupiniquim and Guaraní Mbyá (Figure S2a-b; Dataset S1), as will be described later, we selected 1 Tupiniquim and 4 Guaraní, with 94,06%, 99,99%, 99,05%, 98,75% and 99,99% of estimated Native American proportions, respectively. We selected these samples in order to exploit a higher density of SNPs by genotyping them in the *Axiom Human Origins* array - Affymetrix/Thermo Fisher (1) (Table S1).

The following public datasets (Table S2) were combined into different sets with the new data (Table S1), producing our working datasets (Table S3): 48 Native Americans (2), Human Genome Diversity Project dataset 11 (http://www.cephb.fr/hgdp/), 1000 Genomes Project (http://www.internationalgenome.org/), Anzick-1 (3) and 15 Ancient DNA samples from Brazil (4).

The merging strategy consisted of maintaining only the markers present in all data used to assemble each dataset, and we performed all merging steps with PLINK v1.9 (5). Data curation processes were performed only on the final datasets listed in Table S3. All commercial arrays, in general, are prone to ascertainment bias, however here we have used data genotyped on the *Axiom Human Origins* array, which was explicitly designed to deal with this problem and to enable the implementation of population studies. Addressing this problem was made possible through the selection of SNPs on individuals with known ancestry, identified from the resequencing of 12 individuals of the HGDP-CEPH (Human Genome Diversity Panel) from populations from all over the globe (including one individual from a Brazilian Native American population, the Karitiana). These SNPs were posteriorly validated as polymorphic in 44 samples from the HapMap and 952 HGDP-CEPH. For this reason, it is possible to identify SNPs on the principal ethnic groups (Africans, Europeans, Asians, Melanesians, and Native Americans) through this array.

The geographical locations of all populations (new and public data) used in this work are depicted in Figure S1, and the numbers of individuals and SNPs reported in

Table S3 are those obtained before quality control steps, which are described in the next section. Throughout the text, we will describe the analyses in which each dataset was used.

Quality Control (QC) and Exploratory Data Analysis (EDA)

Before any of the analyses were conducted, the data were submitted to the following QC steps, ensuring higher reliability to the results. First, the individuals previously reported for being admixed (2) and related in our sample were excluded. Also, sites with more than 10% of missingness were removed from the data. We developed a function in R, which could select the maximum number of unrelated individuals, bearing the highest possible proportion of a given ancestry. This function needs as inputs: (1) estimates of kinship parameter (k) between all pairs of individuals (as estimated by IBD - Identity by descent - using PLINK method of moments) and (2) admixture proportions for each individual (as generated by ADMIXTURE - (6). It was applied to the Tupiniquim and Guaraní Mbyá, selecting the maximum unrelated dataset ($k < 0.1$) with the maximum contribution of Native American ancestry. For the other Native American populations, we just selected the maximum unrelated set of individuals ($k < 0.1$). Beyond that, we withdraw all Native Americans with a proportion of estimated Native American ancestry less than 95% from data, for the Guaraní Mbyá the threshold used was 98% of estimated Native American ancestry. The global ancestry inferences used to select the Tupiniquim and Guaraní Mbyá come from the analysis presented in Figure S2, while those used to select the remaining Native Americans were produced in the analysis displayed in Figure S5 (with K=3).

Given that these populations are geographically close, we also estimated k for all pairs of Tupiniquim and Guaraní Mbyá individuals. The maximum estimated k was 0.066, which corresponds to the expected for a 1st cousin. To examine this more carefully, we tried to determine the ancestry of the shared ancestors between Tupiniquim and Mbyá individuals. To this end, we selected pairs of individuals from both populations with k estimates higher than 0.0625 (first cousins), corresponding to 10 individuals (4 Tupiniquim and 6 Guaraní Mbyá). Then we phased their data with Beagle v.5 (7), along with Peruvians from Lima, sub-Saharan-African and West-European populations (1000 Genomes Project) as the Native American, African and European parentals, respectively. A Local Ancestry Inference was performed with RFMix (8) and IBD segment estimation with RefinedIBD (9). Then IBD segments were classified into the ancestries mentioned above, with the method developed by Browning *et al*. (10).

Selecting only the IBD segments larger than 10Mb (created by recent inbreeding) shared between Tupiniquim and Mbyá individuals, we found that 85.97% and 14.03% of the total number of these segments are of Native American and European ancestries, respectively, which are not significantly different from 3:1. The same is true considering the proportions of the total length of segments classified in these ancestries, 75.82%, and 24.18% respectively. These individuals related across

populations are also related inside their populations (Dataset S2), in other words, there are two clusters or families, one in each population, with individuals related to the other. Considering the proportions 3:1 of IBD ancestries, the simplest scenario would be the existence of a common couple of grandparents between these two clusters, responsible for creating long IBD segments being one completely Native American and one admixed with 50% of Native American and 50% of European contributions. However, it is uncertain whether these common ancestors were part of the Tupiniquim or Guaraní Mbyá or even another source. Nonetheless, this result suggests that a single migration event could explain the kinship observed between the populations, which would only be possible after the Guaraní Mbyá were resettled from Southern Brazil to their current location near the Tupiniquim in the 1960 decade. In this sense, only unrelated individuals were used (maximum estimated k is 0.018) and our analysis of the pre-contact time looking into the relatedness among Native American populations would not be biased by this recently introduced kinship between some Tupiniquim and Guaraní Mbyá individuals.

However, it is essential to note that some analyses (as well as the EDA) were also conducted for all Tupiniquim individuals, considering each individual separately, to study the effect of admixture with African and European components on the data. When regarding the Tupiniquim as a population, we used only unrelated samples based in the QC steps.

After QC an initial set of EDA was performed, in order to access the general quality of the data, its characteristics and sources of variation, more specifically this was done through Principal Component Analysis (PCA), as implemented by SNPRelate R/Bioconductor package (11). PCA of dataset iv (Figure S29a) show the 47 Tupiniquim individuals (in red) scattered between the three major clusters represented by the African, European and Native American populations, evidencing different proportions of admixture. The Tupiniquim with more than 80% of Native American estimated ancestry are marked by their IDs (1214, 2004, 4002, 4005 and 4019) placed near their respective dots in the plot, and are closer to the Native Americans, in comparison to the other Tupiniquim. Whereas, in the PCA of dataset vi (Figure S29c) we have only 1 Tupiniquim (~ 94% Native American ancestry and genotyped in the *Axiom Human Origins* - Affymetrix/Thermo Fisher - (1)) which is also evidently grouped with the Native American cluster (2004).

Genetic structure and post-contact admixture

As explained in the first section of the SI Appendix (Dataset Assembly), we had access to two newly-genotyped populations from the Espírito Santo state, which are the Tupiniquim and the Guaraní Mbyá. Both are populations of Tupí-Guaraní Language speakers, part of one major South American linguistic stock, the Tupí. The present-day Tupiniquim form an admixed community settled in Aracruz (ES) and, currently, they speak Portuguese and do not speak any native language. Nevertheless, they self-report as Tupiniquim and were culturally identified as such by the Fundação

Nacional do Índio (FUNAI) in 1975, though lacking clear evidence of genetic continuity with the extinct pre-contact Tupiniquim people.

In this sense, as an initial approach, we wanted to estimate the proportions of admixture with other ancestral continental components. Nevertheless, distinctively for this task, we have done an LD-based pruning of markers above an r^2 threshold of 0.2, inside a sliding window of 50 Kb size and a step size of 10 Kb. Then a Global Ancestry Inference was conducted with ADMIXTURE software (6) in its supervised form with K=3, fixing the Native American, African and European populations as the ancestral components, based on datasets i and ii (Table S3). A set of 100 independent runs were performed and posteriorly concatenated with Pophelper/R package (v.2.2.0; (12)) and used as an input to CLUMPP (Jakobsson & Rosenberg, 2007) to obtain a consensus. Considering the lack of suitable parental Native American populations with a large proportion of overlapping SNPs with the Tupiniquim genotype data, we decided to use Guaraní Mbyá as such, since they were genotyped in the same array. Thus the Guaraní Mbyá ancestry component proportions were also examined in the same way described for the Tupiniquim, to make possible the selection of individuals with a high proportion of Native American ancestry (> 90%) to be used as the Tupiniquim parentals. The results show that the Tupiniquim have lower levels of Native American ancestry (51.55%; Figure S2a) in comparison to the Guaraní Mbyá (77,27%; Figure S2b), though there is substantial variance (standard deviation of ± 24,59% and 26,94%, respectively) in admixture proportions of both populations (Figure S2c-d; Figure S29a; Dataset S1) and that there are individuals with a high proportion of Native American ancestry. This high ancestry proportion was later exploited to enable the study of the Tupiniquim Native American component and to recover some aspects of its pre-contact history.

Additionally, we aimed to estimate time since the admixture events of the continental ancestry components: (a) Native American and European, (b) Native American and African and (c) European and African. To this end, Rolloff (1) was used. In general lines, when an admixture event happens, chunks of DNA sequence in Linkage Disequilibrium (LD) are introduced and with time these chunks are shortened through the action of the recombination process, based on this rate of exponential decay Rolloff estimates the time since the admixture event took place. In practice, it is calculated as a correlation between a signed LD statistic for pairs of SNPs and a weight that represents the allele frequency differentiation in their ancestral populations.

Therefore, Tupiniquim were used as targets, and as parentals we used all pairs of combinations of Guaraní Mbyá and one African population of dataset vii and all pairs of combinations of Guaraní Mbyá and one European population of the same dataset, to estimate time since admixture. We also estimated time since admixture between the African and European components, by using Tupiniquim as targets and all pairs of combinations between one African and one European population of dataset vii.

However, it is important to notice that the method adjusts a single exponential

distribution to the empirical decay rate of admixture LD, assuming, therefore, a single pulse model of admixture. We expect that in a scenario of continuous migration, the recovered date should fall within the period during which gene flow occurred. The admixture event between Native American and European components has a mean estimate of 7.05 generations ago (Figure S6a; Dataset S3), with very little variance of estimates regardless of the parental populations pair used, the same is observed for the Native American and African component (Figure S6b; Dataset S3), although an earlier mean date is estimated, 5.51 generations ago. Slightly earlier estimates in comparison to the Native American and African are recovered for the European and African components (with mean 5.66; Figure S6c; Dataset S3), which might indicate these components were already admixed before, hence the African component could have been introduced in the Tupiniquim via mestizo individuals.

To further investigate how the admixture dynamics the model-based approach implemented by TRACTS software (25, 26) was applied. First, we identified regions with less than 0.9 inferred posterior probability (RFMix; (8)) for a given ancestry. These regions were designated as "undefined ancestry" and were excluded from the analysis. The selected data was then used to test three demographic models of increasing complexity: (a) Single-pulse admixture model: with a single admixture pulse between Native Americans and Europeans, followed by a single admixture pulse with Africans; (b) Discrete double pulse admixture model: in which a second pulse is included to the model, with additional input from the European and African component. In these first two models, the pulses are discrete. (c) Double-pulse admixture model: with a continuous migratory flow, similar to model 2, but with continuous pulses rather than discrete, allowing continuous migratory flow by following generations. Each model was tested three times to verify the convergence pattern, and we fit each model with 1000 bootstrapping attempts. The analysis of time and admixture dynamics inferred from TRACTS indicates the "Double pulse admixture model with continuous migratory flow" as the best-fit model (Dataset S4; Figure S7). This result suggests that the admixture process in Tupiniquim people was complex and continuous. The admixture events are estimated to have happened at ~11.2 generations ago with contact with Europeans and at ~8.3 generations ago with Africans. This initial admixing was followed by a second pulse that started ~5.2 generations ago and with a continuous flow of Africans and Europeans for subsequent generations (Dataset S4). It is important to point out that the continuous admixture process in Brazil means that individuals of admixed origin (European and African) may be admixing with the Tupiniquim, but this most recent dating will not be detected by methods that consider only the length of the ancestry tracts.

Leveraging the Tupiniquim Native American ancestry component

The Tupiniquim individuals with a higher proportion of Native American ancestry (Figure S2c-d; Figure S29a; Dataset S1) were used to investigate the relationship of the Tupiniquim with other modern and ancient Native Americans. To

do this, we conducted two approaches, and the first one was to use only an almost non-admixed individual (~94% of Native American Ancestry) to represent the Tupiniquim. Furthermore, this sample was genotyped in the *Axiom Human Origins* platform (1), making it possible to use a much higher number of markers. In the second approach we performed a local ancestry inference on all the Tupiniquim, which was implemented with RFMix (8), because in admixed individuals, chunks of the genome from different ancestries can be identified, especially in populations were the admixture event was reasonably recent, and therefore the chunks of Native American ancestry can be marked and selected. Thus, based on the previous Global Ancestry Inference, as described before, we could select a group of 25 mostly non-admixed (> 90% Native American ancestry) Guaraní Mbyá, to be used as the Native American parental. African (excepting ASW and ACB populations) and European populations from the *1000 Genomes Project* were also used as parentals, randomly sampling 25 individuals from each one, to match the number of Guaraní Mbyá, which is required because the Local Ancestry Inference as executed by RFMix (8) could be biased by the sample sizes (thus producing the dataset iii).

The data was then phased with ShapeIT (*Segmented Haplotype Estimation and Imputation tool* v.2; (13), which uses Hidden Markov Model (HMM) to estimate the haplotypic phase from genotype data. As previously stated, Local Ancestry Inference was then implemented by RFMix (8), with 100 iterations of the Expectation-Maximization algorithm, a window size of 0.2 cM and with five as the minimum number of reference haplotypes per tree node. In this method, the genome is divided into windows of SNPs, and through a Random Forest algorithm, the posterior probability of each window being more ancestrally related to each population in a reference panel is calculated.

For the subsequent analyses, only the sites with more than 99% of posterior probability of being more ancestrally related to Native Americans were selected, masking the remaining genotype data present in the Tupiniquim (i.e. setting as missing data), with a custom-made function in R. We assigned both alleles as missing data for loci in which the inferred local ancestry was heterozygous, i.e. a locus presented one inferred ancestry for one allele (e.g. Native American ancestry) and another ancestry for the other allele (e.g. African ancestry). When considering all 47 Tupiniquim individuals, from a total of ~600K SNPs, an average of ~45% of the sites were removed due to heterozygous local ancestry. However when we consider only the 5 Tupiniquim individuals with more than 80% inferred Native American ancestry, which were selected to represent the Tupiniquim population, an average of ~29% of the sites were set as missing data, while an average of ~32% of the sites were removed in total. Using ADMIXTURE (6) in its supervised form with K=3 the Tupiniquim presents considerable West-Eurasian-related and sub-Saharan-African-related ancestry proportions before masking (Figure S9a), but almost none afterward (Figure S9b).

To investigate the effects of the admixture and the number of individuals and

markers used, we used three datasets, the first containing all Tupiniquim without any treatment, the second containing 5 Tupiniquim with more than 80% inferred Native American ancestry and masked data, and the last containing only the Tupiniquim with 94.06% inferred Native American ancestry (and genotyped in *Axiom Human Origins* (1)). PCAs of these datasets show the effect of the masking and the selection of individuals, Figure S29a and Figure S29b show data before (dataset iv) and after masking the data (dataset v), respectively, in Figure S29b the selection of Tupiniquim individuals with high Native American ancestry can also be observed. Whereas in Figure S29c (dataset vi) there is only one Tupiniquim (ID 2004) and it can be seen that there is almost no effect of using a much higher number of markers, from ~ 70K (datasets iv and v) to ~ 570K (dataset vi), as only minimal differences are observed between PCAs (Figure S29). The same is true for PCAs of these datasets containing only the Native Americans (Figure S32a-d).

Relatedness between modern Native American populations

In order to examine the allele sharing between Native Americans, we used AdmixTools (1) to compute F_3, D-statistics, and F_4, treating the Tupiniquim as a population and as separate individuals in some calculations. In order to investigate and isolate the effect of post-contact admixture in the Tupiniquim datasets iv and v were used (~70K SNPs), making it possible to compare the estimates for the unmasked and masked Tupiniquim genotype data. We also computed all sets of F-statistics for dataset vi, hence using a much larger number of markers (~570K SNPs).

In order to investigate the effect of admixture with other continental ancestral components in the F-statistics estimation, we calculated F_3(Mbuti Pygmy; Tupiniquim, Z), for every pair of Tupiniquim individuals and Z modern Native American populations. The results demonstrate the effect of masking the non-Native American ancestry in the Tupiniquim (Figure S30a and Figure S30b, before and after masking, respectively). Before the masking procedure, F_3 values clearly correlates ($r^2 = 0.8939$) with the Native American ancestry proportion of individuals (Figure S31a) and after the masking F_3 estimates are virtually identical for every individual (Figure S30b and Figure S31b), with the exception of individuals with very high proportion of missing data (top of Figure S30b and right of Figure S31b). As we have only used individuals with more than 80% of inferred Native American ancestry to represent the Tupiniquim population, the analyses are very unlikely biased by the proportion of missing data.

F-statistics are based in estimates of genetic drift between pairs of populations, so we used the three population statistic F_3 as implemented by AdmixTools (14), to test if the Tupiniquim are admixed, more specifically to test if there is evidence of admixture between the Tupiniquim Native American ancestry component with other modern Native American populations. To accomplish this we calculated the F_3-statistics in the form F_3(Tupiniquim, Y, Z), for every pair of Y and Z modern Native American

populations and we do not find any evidence of admixture with other Native American groups, as would be indicated by an estimated significative negative value for the statistic (Figure S10).

In order to corroborate or contradict this finding, Treemix (14) was used to estimate the Maximum Likelihood tree and additionally to fit admixture events between branches, thus producing in the last case admixture graph models. Usually, the tree is seen as a null model, in which there is no gene flow, in opposition to 1 or more events of gene flow, which could potentially improve the model fit of the empirical data. The tree is based in population pairwise allelic covariance, and the gene flow events are fitted between pairs of populations that have the worst fit between the observed and the expected covariance based on the fitted tree model. In practice, we added gene flow events, one at a time, up to 5. To infer these trees, all available Native American and Tupí populations from datasets v and vi were used (Figure S11a-d). In these models not a single gene flow event to the Tupiniquim branch is inferred, consequently supporting what was observed before in the three population statistic (Figure S10). Though a gene flow event is fitted leaving the base of the Tupiniquim branch to the Tupí Mondé (Zoró and Suruí) cluster node when using the Tupí populations from dataset vi and fitting five events (Figure S11d).

We next aimed to investigate the patterns of allele sharing among groups. Hence outgroup-F_3 was used to measure the amount of shared drift (i.e., allele sharing) between these populations. We calculated F_3(Mbuti Pygmy; Y, Z), for every pair of Y and Z modern Native American populations. Fixing one population in the Y position, and iterating populations in Z, the amount of shared genetic drift to every other population can be estimated. The results were plotted as heatmap points based on their geographical coordinates on a map (https://github.com/pontussk/point_heatmap/blob/master/heatmap_Pontus_colors.R). Patterns of geographically-genomic relationships are observed for the Guaraní populations (Figure S12b; Figure S13-Figure S14) and the Amazonian Madeira Guarporé Region populations (Figure S13-Figure S14), on the other hand, the Tupiniquim does not show any distinctive pattern of allele sharing (Figure S12a).

We also assessed global patterns of ancestry in the Native American and Tupí populations from datasets v and vi, by performing a PCA (SNPRelate R/Bioconductor package (11)). PCA from both datasets (v and vi) are mostly indistinguishable (Figure S32a-d), with exception of the observed patterns of the Tupiniquim, which differ more, most probably due to the inclusion of 5 Tupiniquim individuals with masked non-Native American ancestries in dataset v (Figure S32a-b) and only one Tupiniquim individual with some (6%) non-Native American ancestry in dataset vi (Figure S32c-d). This result shows the effect of using fewer genetic markers (~600K (dataset vi) and ~70K (dataset v)) does not significantly influence the PCA, although the use fewer individuals (i.e. one in dataset vi) and the presence of little non-Native American ancestry (6%) seems to do so. For this reason, results obtained for dataset v

are more likely to represent the distribution of the Tupiniquim global patterns of ancestry. These results indicate a closer relationship between the Tupiniquim and some Amazonian Tupí Populations, namely Parakanã, Urubu Kaapor (Tupí-Guaraní) and Zoró (Tupí-Mondé) (Figure S32b), but also some similarity with Karib-speaker populations (Apalai and Arara) from the Amazon (Figure S32b).

Additionally, we designed a set of global ancestry inferences using ADMIXTURE (6), to examine the patterns of shared ancestry and genetic structure across the Native American populations, including the Tupiniquim. An independent inference was performed for datasets iv, v and vi, each with K values between 2 and 8. This inference was done with ten independent runs for each value of K and obtaining the consensus for all runs and Ks with Pophelper/R package (v.2.2.0; (12)) and CLUMPP (15). With ADMIXTURE (6), the cross-validation error for each K was also estimated. Separate plots were produced for each analysis (Figure S3, Figure S4, and Figure S5), and the value of K that minimizes the cross-validation error is 6, 5 and 3 for dataset iv, v and vi, respectively.

Long-standing patterns of ancestry

Furthermore, we wanted to examine the relationship between ancient Brazilian samples (4) and modern populations, to see if we could detect any patterns of distinctive shared ancestry among them. To accomplish this the F-statistics was also used to investigate the patterns of allele sharing between ancient samples from Brazil (15 Ancient DNA samples from Posth et al. (4)), the Anzick-1 (Clovis Culture associated ancient DNA sample - (3)) and modern Native Americans (using datasets ix and x). It should be noticed, however, that all analyses involving ancient DNA samples were conducted with the same parameters as in Posth *et al*. (4). More precisely "inbreed: YES" parameter was used for all F_3-statistics, therefore using pseudo-haploid data and ''f4mode: YES'' when using qpDstat (thus applying F_4-statistic rather than D-statistic).

We calculated F_3(Mbuti Pygmy; aDNA, Z) for each 'aDNA' archeological site (e.g., Jabuticabeira) separately, where Z is any modern Native American population. The comparisons were also plotted as heatmap points on a map (https://github.com/pontussk/point_heatmap/blob/master/heatmap_Pontus_colors.R).

Broader patterns of allele sharing are observed for older populations, becoming increasingly restricted in more recent populations (Figure S15a-d and Figure S16). We then tested the significance of these differences between estimated F_3 for each comparison in two ways: (i) calculating the statistic F_3(Mbuti Pygmy; Y, Z) for all pairs of Native American individuals in Y modern populations and Z ancient sites, generating multiple estimates of F_3 for each comparison, and using ANOVA and Tukey HSD (Honestly Significant Difference) to test the significance of the differences; (ii) making a more fine-grained examination, comparing all pairs of modern populations and inferring which modern population shares more alleles with

each of the ancient samples, we used F_4(Mbuti Pygmy, aDNA; Y, Z), again "aDNA" is an archeological site and Y and Z are any modern Native American population.

Significative differences between estimated F_3 values are detected for comparisons of each ancient site with all modern populations as demonstrated by ANOVA (p-value < 0.001; Dataset S5), i.e., for every ancient site there is at least one modern population that shares significantly more or fewer alleles with it. Both Tukey HSD (Honestly Significant Difference) tests (Figure S17-Figure S18) and F_4-statistics (Figure S19-Figure S20) demonstrate that Xavante and Mesoamericans (Pima and Maya) consistently display higher and lower levels of allele sharing with ancient samples, respectively. When using dataset x, the Tupiniquim consistently shows low values of outgroup-F_3 (Figure S18 and Figure S20), most probably due to admixture with African and European populations.

Taken together these results indicate some level of genetic continuity inside Brazilian territory, with older sites (Lapa do Santo and Laranjal) sharing alleles with most modern populations and more recent sites exhibiting a more distinctive pattern of shared ancestry. Posth *et al*. (4) detected evidence for this continuity in comparisons of ancient individuals, with the highest estimated age of ~5,800 YBP, with modern populations inside South America, a result replicated here, as seen for Moraes and Jabuticabeira (Figure S17-Figure S20). We detected higher allele sharing between the oldest Native South Americans (i.e., Lapa do Santo and Laranjal) and the modern Native South Americans in comparison to the allele sharing between the former group and the modern Native Mesoamericans. This pattern was not observed by Posth et al. (4). Interestingly modern native South-Americans also share significantly more alleles with Anzick-1, than modern Mesoamericans do.

Tupiniquim population history and the Tupí Expansion hypotheses

In order to recover the underlying history of the Tupiniquim Native American lineages and shed light on the process of Tupí expansion to the coast, basically two approaches were developed: (i) to build population trees using models of genetic similarity and/or dissimilarity and clustering algorithms; (ii) to explicitly model the 2 main hypotheses for the Tupí Expansion (16) (Figure 1).

Since our objective was to recover the Tupiniquim population history, we decided to study only the Tupí, as it would be a simpler model. Therefore we produced trees using all Tupí populations for datasets v and vi and tested their fit to the data using the F-statistics framework. For this end, here we used F_{ST} and F_2 pairwise distances Neighbor-Joining trees and Treemix (14) ML trees. The Maximum Likelihood trees produced by Treemix presented good fit to the data, as indicated by no significative (> 3 SE) differences between observed and expected covariance in the inferred trees. Therefore we used these trees without any gene flow events for datasets v and vi.

To test the different hypotheses about the trees, we used the qpGraph algorithm

(14), by modeling and assessing the models. We assessed how well the F-statistics estimates calculated with our observed data fitted the expected estimates as predicted by the models, the usual criterion adopted for a good fit is a threshold of $|Z| < 3$ for all differences between predicted and empirical estimated statistics.

Regarding all trees produced for the Tupí populations (Figure S21-Figure S24), only two (Neighbor-Joining Tree based on pairwise F_2 values and Treemix) based on the dataset vi presented a good fit to the data (Figure S22b-c; Figure S24b-c). These two trees show a very similar topology and in a more careful analysis it is possible to identify that there is very little genetic drift separating the Tupi-Guaraní populations, with the exception of the Guaraní populations, which seems to be a more consistent and likely a monophyletic group (Figure S24b-c; also observed in all other produced trees). These results argue in favor of the Guaraní having had a population history separated from the other Tupi-Guaraní and diverging from them relatively soon in the context of the Tupí Expansion, possibly while they were still inside the Amazon. The other Tupi-Guaraní seem to have gone through rather rapid radiation, maybe involving a relatively large Ne, producing this pattern of star-like radiation. This seems to complicate the establishment of the underlying relations between these Tupi-Guaraní populations.

To provide more subsidies for these results, as previously described, the second approach was to model the two main hypotheses for the Tupí Expansion (16, 17) (Figure 1) namely: the expansion initially going southwards and then reaching the southeast region of present-day Brazil (Blue arrow in Figure 1), with the coastal Tupí deriving from the Guaraní populations cluster; or the alternative, the expansion first following the course of the Amazon river and then spreading eastwards across the coast, through the northeast region, until finally reaching the southeast region, in this case, with the Guaraní expanding in a separate wave from the Amazon to the south (Red arrows in Figure 1). These hypotheses would be differentiated by observing a close relationship between the Guaraní populations or the north Tupí populations (e.g. Urubu Kaapor or Parakanã) of Brazil with the Tupiniquim. Ideally, models would have a differential fit, and the most likely hypothesis identified.

Assessment of the different models produced using dataset vi for both hypotheses argues in favor of a hypothesis 2 type scenario, as hypothesis 2 models present a good fit to the empirical data ($|Z| < 3$), in comparison to hypothesis 1 models (Figure 4a-b). To test the consistency of the results different models were produced using dataset vi with distinct sets of populations and replacing the Tupiniquim population, even though not all hypotheses 2 type models present good fit to the data ($|Z| < 3$), they clearly adjust much better to the data (Figure S25-Figure S26 and Dataset S6). Specifically, the inclusion of Zoró and specially Wajãpi populations significantly impair the models fit (Figure S26c-e), which suggests they have a more complex population history, maybe involving admixture events. However, even when including all Tupí populations, the hypotheses 2 models have a much better fit (Figure S25e and Figure S26e). Assessment of models based on dataset v also points to

hypothesis 2 being more likely (Figure S27-Figure S28). Although the hypothesis 2 models based on dataset v do not present a good fit to data (i.e. all estimated $|Z| > 3$), they exhibit very few outlier F-statistics and a reasonable maximum estimated $|Z|$ around 3.5. On the other hand, the hypothesis 1 models present a much worse fit to data with many outliers and maximum $|Z|$ of 7.5 for all models (Figure S27-Figure S28 and Dataset S6).

Demographic inferences

Regarding the fact that present-day Tupiniquim are admixed with other continental ancestral components, we wanted to examine their demographic history as well as that of their parental populations. This analysis was made possible through a recently introduced method (10), which exploits two properties of the genome, the first is that populations with fairly recent admixture events, usually have relatively long chunks of a given ancestry, because recombination has not had the time to reduce their length. The second feature is the relation between the number and length of IBD segments in a population and the probability of coalescence, and therefore, the effective population size (Ne). This relation can be expressed in an equation that allows for the Ne history estimation (i.e. Ne in generations before present).

Based on both these properties, the IBD segments can be classified based on their ancestry and used to estimate the ancestry specific Ne history and also the overall Ne history, in this last case using all estimated IBD segments. In practice the data was phased by Beagle v.5 (7), followed by the IBD segment estimation with RefinedIBD (9), which is specific to heterogeneous data and the Local Ancestry Inference with RFMix (8), running 10 iterations of the Expectation–maximization algorithm, a window size of 0.2 cM and with 5 as the minimum number of reference haplotypes per tree node. The ancestry specific and overall effective population size was then estimated, applying both the estimated IBD segments and the Local Ancestry Inference to IBDNe (10).

The Ne history estimation was performed for the Tupiniquim (with dataset vii) using the sub-Saharan-African and West-European populations from the *1000 Genomes Project*, and Guaraní Mbyá as the African, European and Native American parentals, respectively. Whereas, for the Guaraní Mbyá we used non-admixed (with more 90% estimated Native American ancestry) Peruvians from Lima (PEL) population from 1000 Genomes Project Consortium (2015) as the Native American parentals (dataset xi). Only unrelated ($k < 0.0625$) Tupiniquim and Guaraní Mbyá were used in the estimations. In general the Ne estimates for Native American ancestry are somewhat similar, and a substantial post-colonization bottleneck is observed for both populations, with minimum Native American ancestry specific Ne around 7 generations ago (175 YBP considering a generation of 25 years), Ne=178 for the Tupiniquim and Ne=108 for the Guaraní Mbyá (Figure 3 and Figure S33). Still, all these estimates need to be taken cautiously, given the small sample size used (26 Tupiniquim and 22 Guaraní), particularly for the African and European ancestries,

though for the Tupiniquim the 95% significance bootstrap intervals are reasonably restricted (Figure 3). The African and European ancestry specific estimates for the Guaraní Mbyá are based on very few IBD segments due to its limited amount of admixture (~20% in total; Figure S2b) and small sample size, for this reason, we could not recover good estimates.

Another way of studying the demography of a population is through the analysis of the distribution of Runs of Homozygosity (ROH) lengths, which reflects its demographic history. The longer ROH are the result of recent events that occurred few generations ago with little time for the recombination process to break up the segments (i.e., recent inbreeding), whereas the shorter tracts represent much earlier events, such as bottlenecks and founder effect (18, 19). Considering the out-of-Africa model of human origins, for example, we expect an increment of shorter ROH and consequently a reduction of the genetic diversity according to the walking distance from Africa (18, 19).

Here, we analyzed (dataset viii) the Tupiniquim ROH distribution (considering only those five individuals with more than 80% of Native American ancestry) alongside with that of the Guaraní Mbyá, African, and European populations from the *1000 Genomes Project* and the Anzick-1 (Clovis Culture associated ancient DNA sample - (3)), as a proxy for the ancestral Native American population. We excluded data from SNPs with deviations from Hardy-Weinberg proportions ($P \geq 10^{-8}$) and minor allele frequency MAF = 0, resulting in a final merged dataset of 417,477 SNPs. The ROH identification was performed with PLINK v1.9 (5), considering a minimum length of 500kb, a sliding window with 50 SNPs, a maximum gap between consecutive SNPs of 100kb; a maximum of one heterozygous genotype and 25 missing calls allowed per window; a proportion of windows that overlap to form a homozygous segment of 5%, and a density of at least one SNP per 50kb.

As shown in Figure S34a, the Native American Clovis, Guaraní Mbyá, and Tupiniquim individuals showed the highest levels of short ROH (<2Mb), with the Tupiniquim presenting slightly lower homozygosity levels. This is probably caused by the admixture process that affects mainly short ROH (<2Mb) in two different ways: creating lower estimates of average ROH length in admixed populations than that expected by the proportional contribution of parental populations, probably due to the increase in genetic variation resulting from the admixture process (20), and creating high estimates of the variance of ROH lengths when compared with non-admixed populations (21). It can be observed, for example, in Figure S34b, in which the Tupiniquim individuals with more significant Native American contribution have homozygosity levels similar to the observed for Guaraní Mbyá and Clovis (populations with less genomic diversity). As already observed by (22), the ancient Clovis genome has an apparent excess of intermediate ROH (2-8Mb).

Additionally, a second ROH analysis was performed, using dataset vi, making a more detailed comparison between the Tupiniquim and the modern Native American populations, grouping them by Amazonian and non-Amazonian South Native

Americans, alongside Mesoamericans, Asians, Africans and Europeans from the *Human Genome Diversity Project*. Again we excluded all data from SNPs: (i) within 2Mb of the extremities of all chromosomal arms; (ii) with high deviations from Hardy-Weinberg proportions ($p \geq 10\text{-}8$); with missing genotypes proportion above 10%; and (3) with minor allele frequency MAF = 0. The data cleaning process resulted in a final merged dataset with 395,840 SNPs. We used the software PLINK v1.9 (5) in order to identify ROH in our sample. The parameters considered were: minimum length of 500kb; sliding window with 50 SNPs; maximum gap between consecutive SNPs of 100kb; a maximum of one heterozygous genotype; 5 missing calls allowed per window; proportion of windows that overlap to form an homozygous segment of 5%; and density of at least one SNP per 50kb. The Tupiniquim average total ROH length is very much like the one from the Mesoamerican populations, except large ROH (> 8 Mb) in which the former show less ROH (Figure 2 and Figure S8). Overall, Amazonian and Non-Amazonian South Native Americans show a similar pattern to each other. Conversely, both have more short and intermediate ROH (0.5-8Mb) than the Tupiniquim (Figure 2 and Figure S8).

Supplementary Figures (S1 to S34)

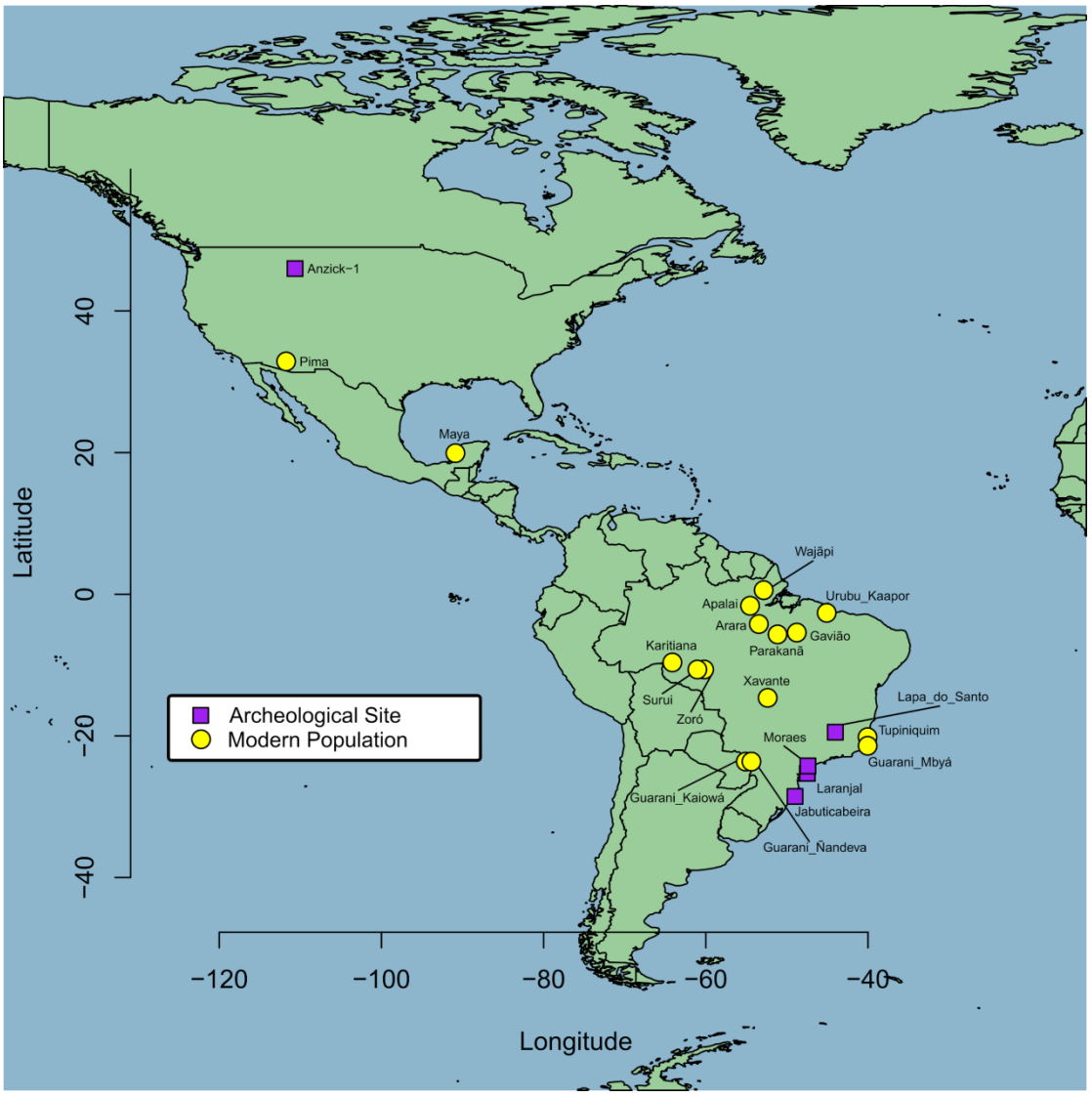


Figure S1. The geographic position of all samples used. Coordinates for populations newly-genotyped and obtained from public data are displayed over the map of the Americas. Modern populations and archeological sites are indicated by yellow circles and purple squares, respectively.

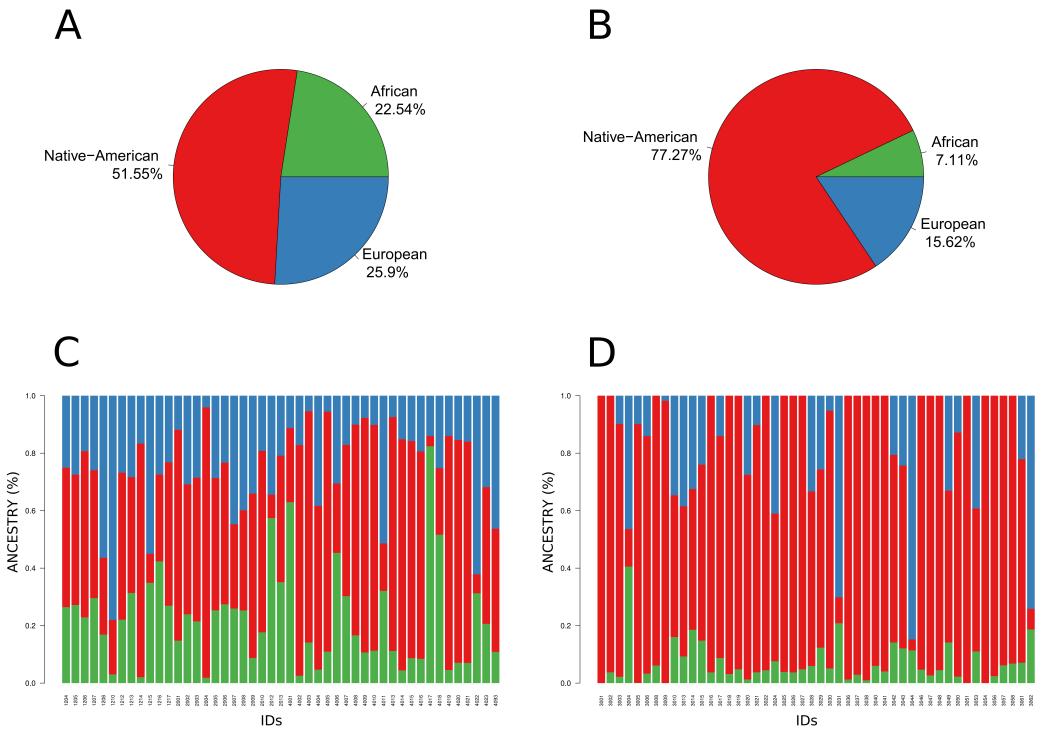


Figure S2. Global Ancestry Inference profile. Estimates were obtained for the Guaraní Mbyá and Tupiniquim using datasets i and ii, with a supervised analysis with K=3 using ADMIXTURE (6) to determine the three continental ancestral components ancestry proportions. The Tupiniquim and Guaraní Mbyá estimated ancestry proportions were plotted as pie charts (**A** and **B**) and as bar plots with samples in the X-axis and percentages of ancestry in the Y-axis (**C** and **D**). Ancestry proportions are color-coded as red, green and blue for Native American, African and European ancestries, respectively. Individual ancestry proportions are presented in Dataset S1.

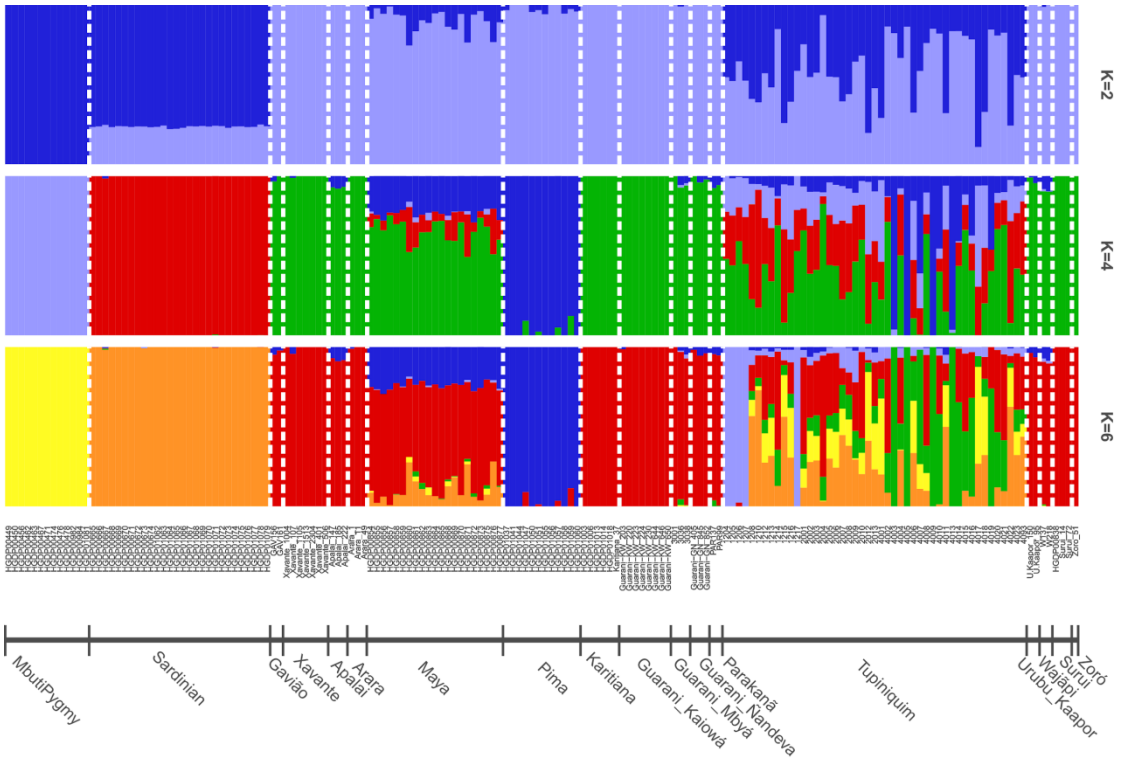


Figure S3. Shared ancestry and genetic structure in Native Americans (dataset iv). An unsupervised analysis was performed with ADMIXTURE (6) using dataset iv, with the number of putative ancestry components (K) varying from 2 to 8. Each vertical bar represents one individual (IDs below bars), and colors represent ancestry assignments given the number of K. Three values of K (2, 4 and 6) were plotted and the value of K that minimizes the cross-validation error is 6. The Horizontal axis beneath the panels indicates population assignments.

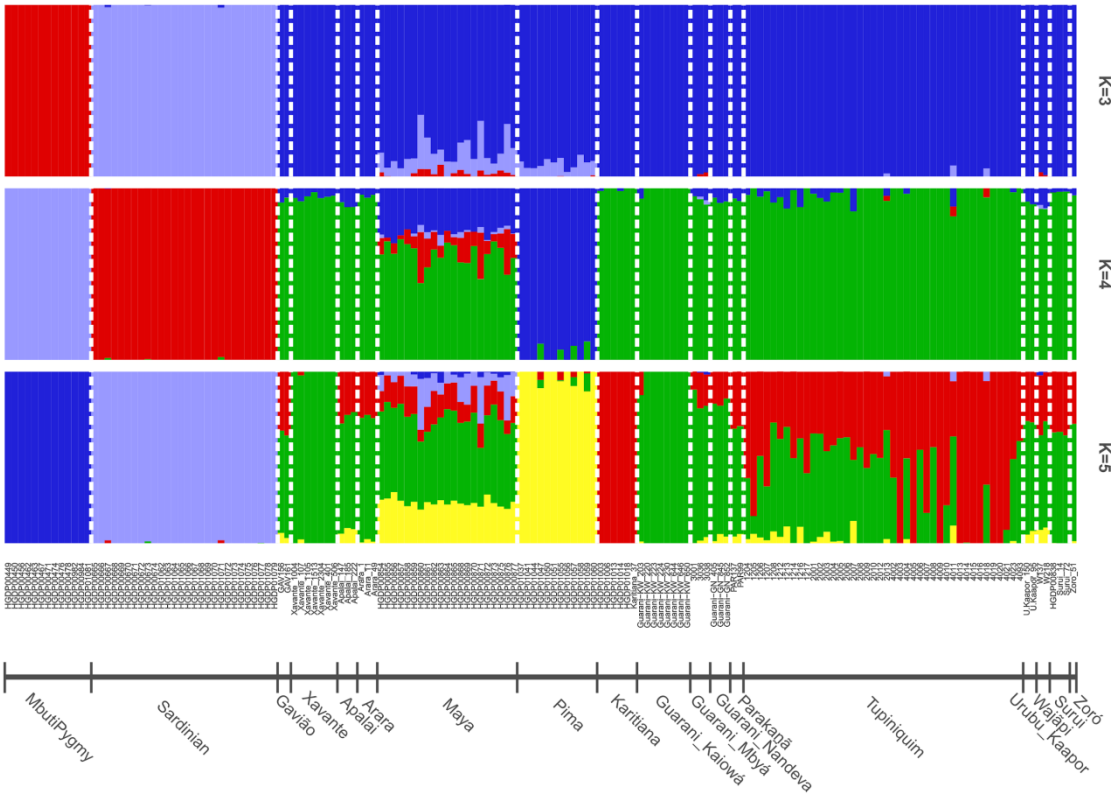


Figure S4. Shared ancestry and genetic structure in Native Americans (dataset v). An unsupervised analysis was performed with ADMIXTURE (6) using dataset v, with the number of putative ancestry components (K) varying from 2 to 8. Each vertical bar represents one individual (IDs below bars), and colors represent ancestry assignments given the number of K. Three values of K (3, 4 and 5) were plotted and the value of K that minimizes the cross-validation error is 5. The Horizontal axis beneath the panels indicates population assignments.

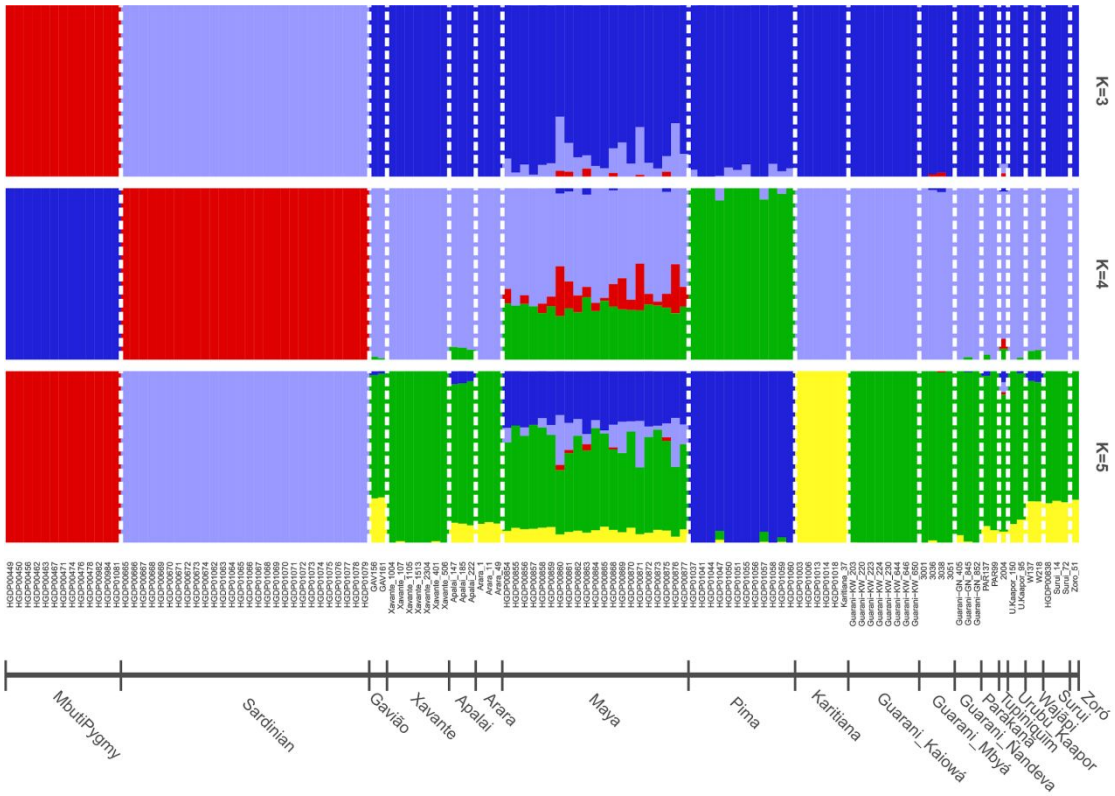


Figure S5. Shared ancestry and genetic structure in Native Americans (dataset vi). An unsupervised analysis was performed with ADMIXTURE (6) using dataset vi, with the number of putative ancestry components (K) varying from 2 to 8. Each vertical bar represents one individual (IDs below bars), and colors represent ancestry assignments given the number of K. Three values of K (3, 4 and 5) were plotted and the value of K that minimizes the cross-validation error is 3. The Horizontal axis beneath the panels indicates population assignments.

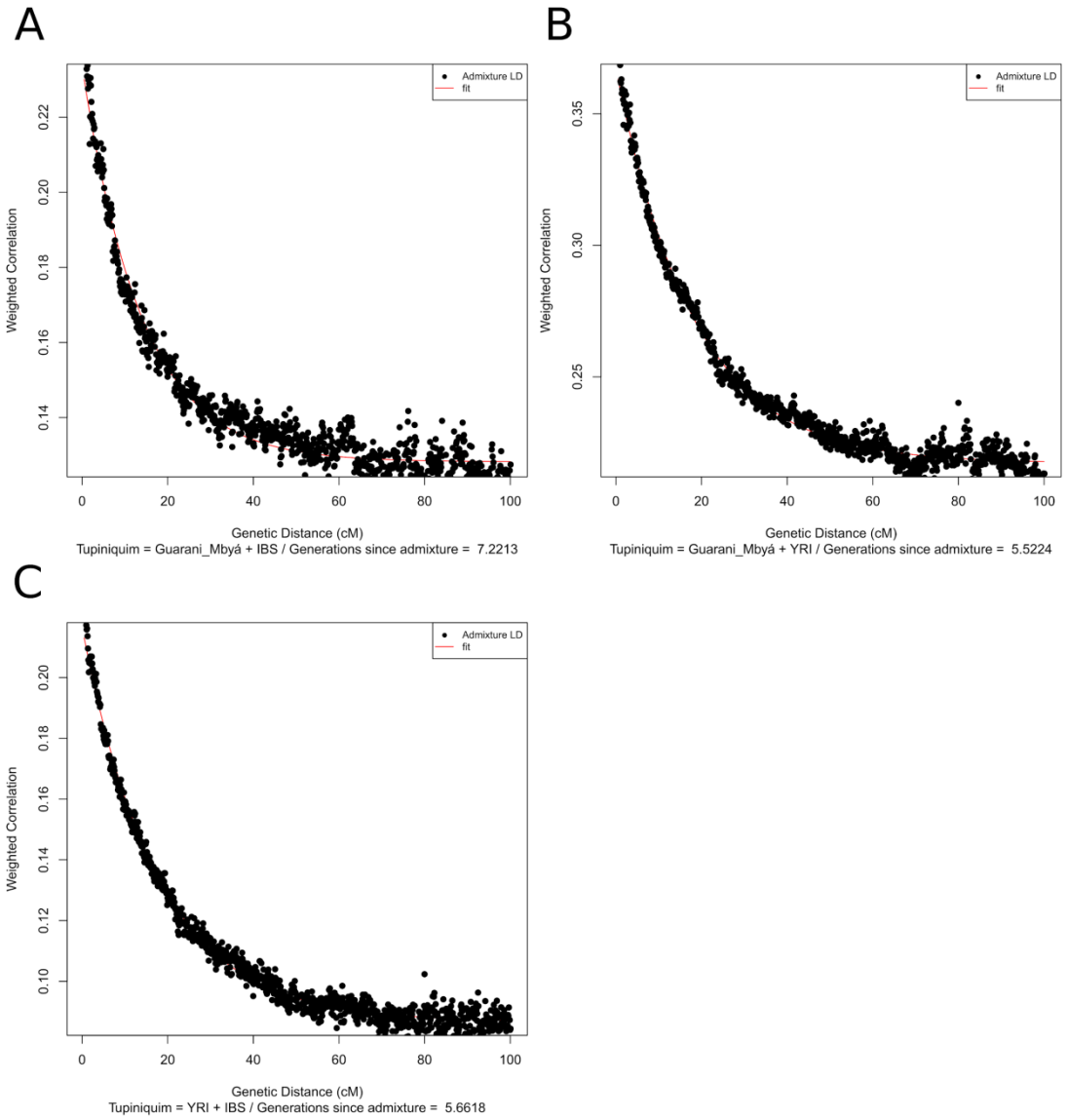


Figure S6. Dating admixture events. Inference of the time since the admixture events was performed with Rolloff (1) using dataset vii. The Y-axis represents an LD statistics weighted by the difference in their allele frequency in the ancestral populations (Admixture LD), and the X-axis indicates the genetic distance (cM). The black dots indicate the estimated Admixture LD and the red line an exponential distribution fitted to the data, which allows the estimation of an admixture date. The estimated admixture dates between Native American and European components (Parentals: Guaraní Mbyá and Iberians) is 7.2213 generations ago (**A**), Native American and African components (Parentals: Guaraní Mbyá and Yoruba) is 5.5224 generations ago (**B**) and African and European components (Parentals: Yoruba and Iberians) is 5.6618 (**C**).

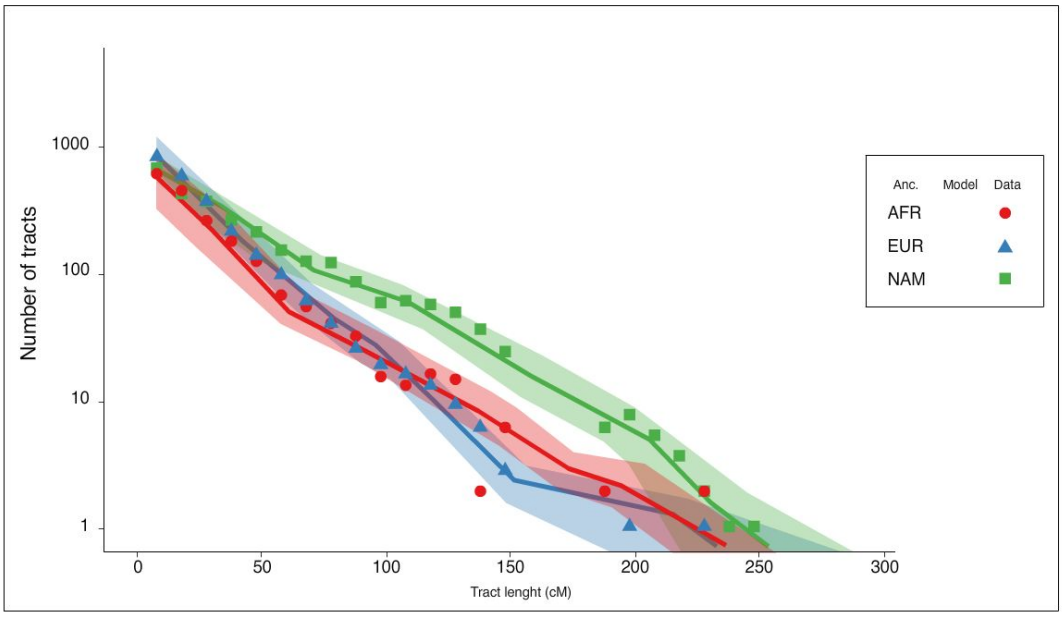


Figure S7. Distribution of tracts and fit of the best model (Model 3). The tracts (or segments) of Native American ancestry inferred by RFMix (8) with a posterior probability above 0.9 were applied to the model-based approach implemented in the TRACTS software (23, 24), in order to test three demographic models: (Model 1) Single-pulse admixture model, (Model 2) Discrete double pulse admixture model and (Model 3) Double-pulse admixture model with continuous migratory flow. For a description of the models, see the SI Appendix text. The analysis indicates Model 3 as the best-fit model, and therefore, we show here the distribution of inferred tracts (dots) and the Model 3 fit (lines). Shaded areas represent 95% Confidence Intervals based on 1000 bootstrap iterations. Colors (and shapes) are coded to represent the ancestry assignment of the tracts distributions and the associated fitted model, along with each 95% CIs. The inferred parameters for all models are also presented in Dataset S4.

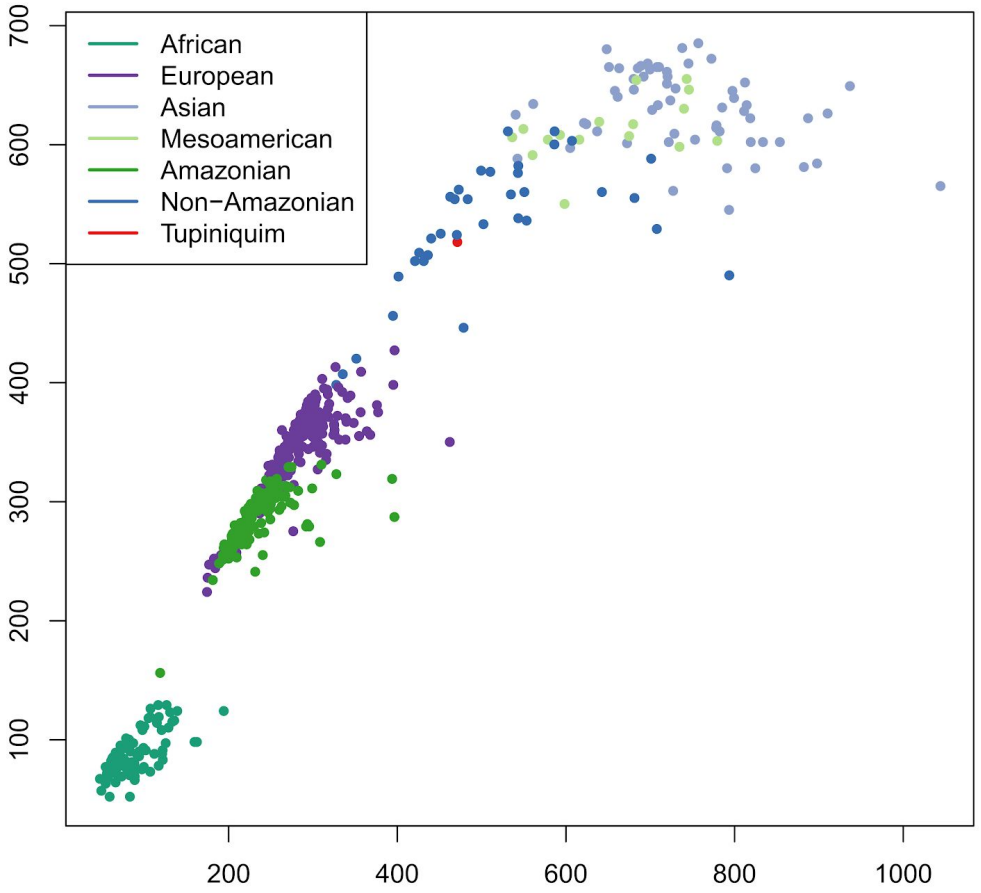


Figure S8. Individual patterns of ROH in Native American, Asian, European, and African populations. The ROH identification was performed with PLINK v1.9 (5), considering ROH above 0.5Mb and taking into account a set of 395,840 SNP markers from a subset of 609 individuals from dataset vi (Table S3). Each point represents the number, and the total length of ROH of each individual identified according to the population they belong.

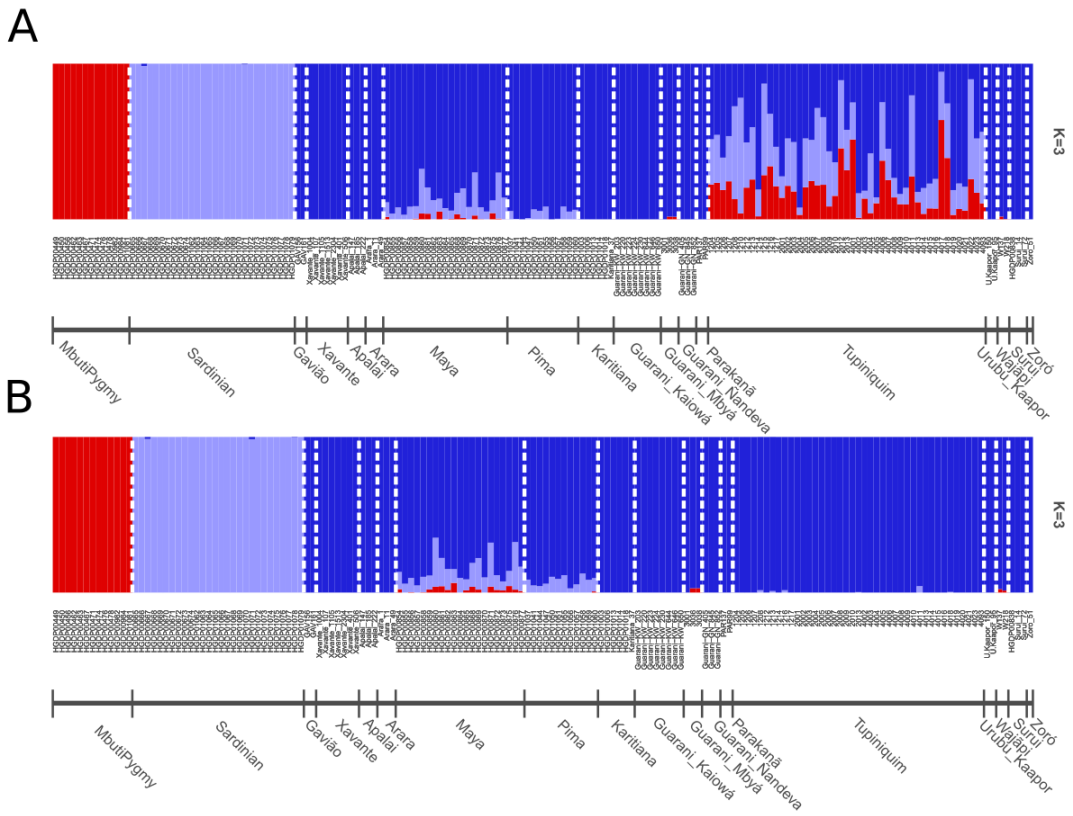


Figure S9. Pre and Post data masking Global Ancestry Inferences. A unsupervised analysis with K=3 using ADMIXTURE (6) to estimate the proportions of Native American, west-Eurasian and sub-Saharan African components in the Tupiniquim before (**A**) (dataset iv) and after (**B**) (dataset v) the Local Ancestry masking of markers with a posterior probability smaller than 0.99 of being more ancestrally related to Native Americans as estimated by RFMix (Maples et al., 2013). Populations are indicated by name tags in the axis below each plot.

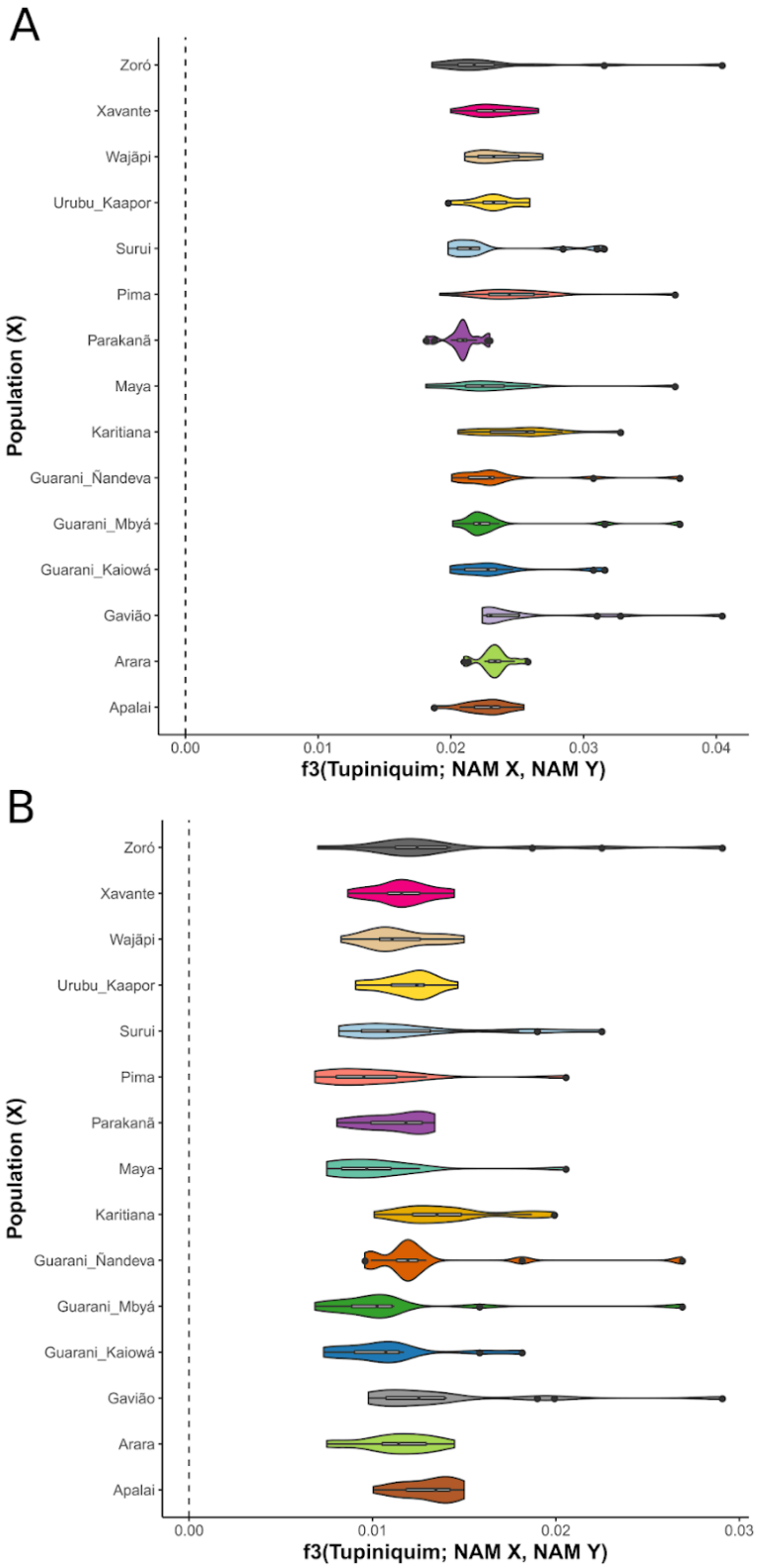


Figure S10. No evidence of admixture in the Tupiniquim Native American component. F-Statistic was estimated in the form F_3(Tupiniquim; Y, Z) with Admixtools (1), where Y and Z are modern Native American populations. Estimated F_3 for the comparison between each population indicated at the Y-Axis, and all of the others are plotted in each vertical line, while the X-axis indicates the F_3 values. Evidence of admixture would be indicated by an estimated negative value of the statistic. **A)** Estimates obtained with dataset v. **B)** Estimates obtained with dataset vi.

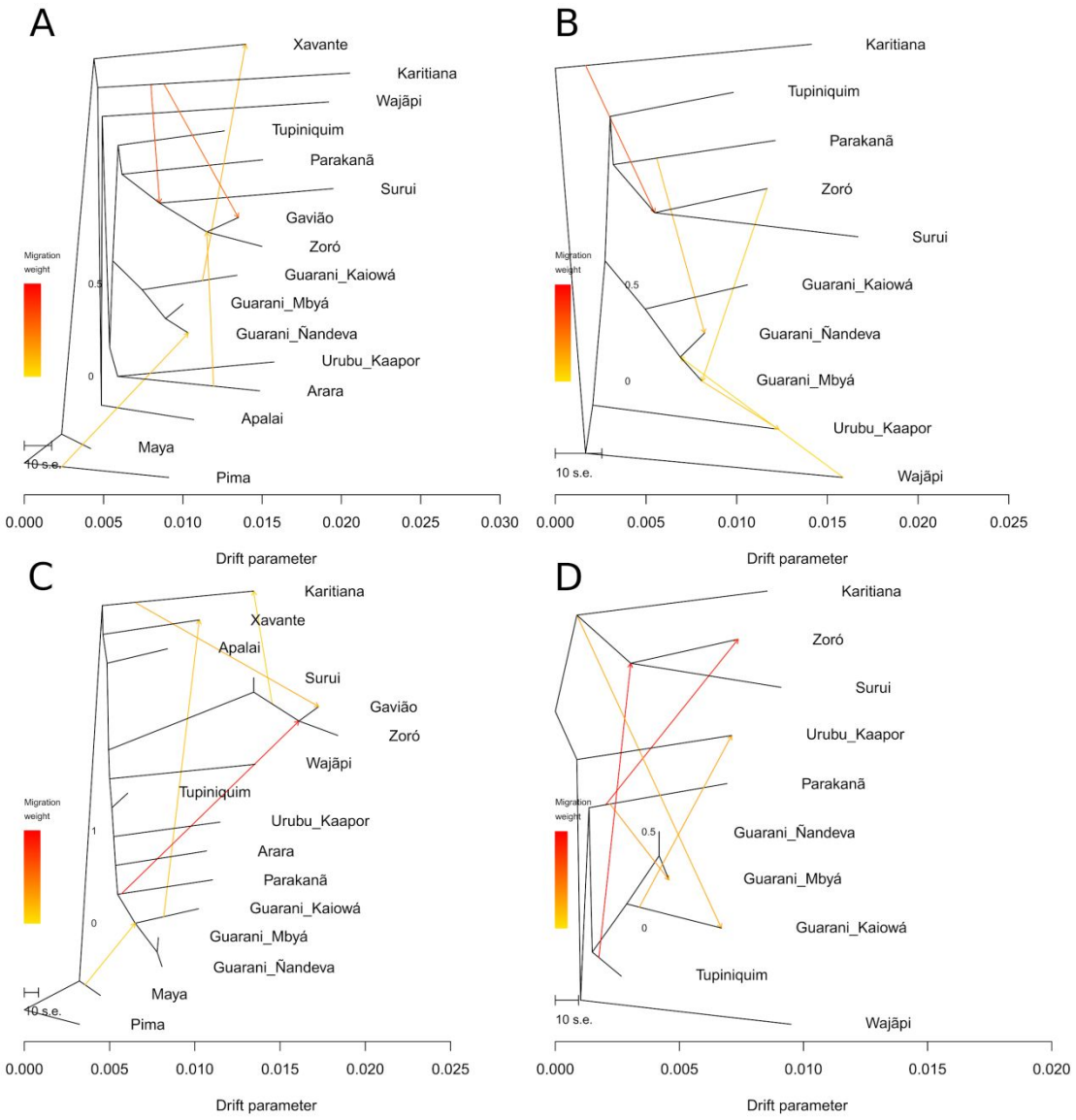


Figure S11. **No gene flow events to the Tupiniquim branch.** Maximum Likelihood trees based on population pairwise allelic covariance were inferred with Treemix (Pickrell & Pritchard, 2012) and admixture graph models were produced, allowing gene flow events between branches of the ML trees. Fitting up to five events detected none gene flow to the Tupiniquim branch. Here are depicted the admixture graphs with five gene flow events of each of the following analyses: **A)** Native American populations from dataset v. **B)** Tupí populations from dataset v. **C)** Native American populations from dataset vi. **D)** Tupí populations from dataset vi.

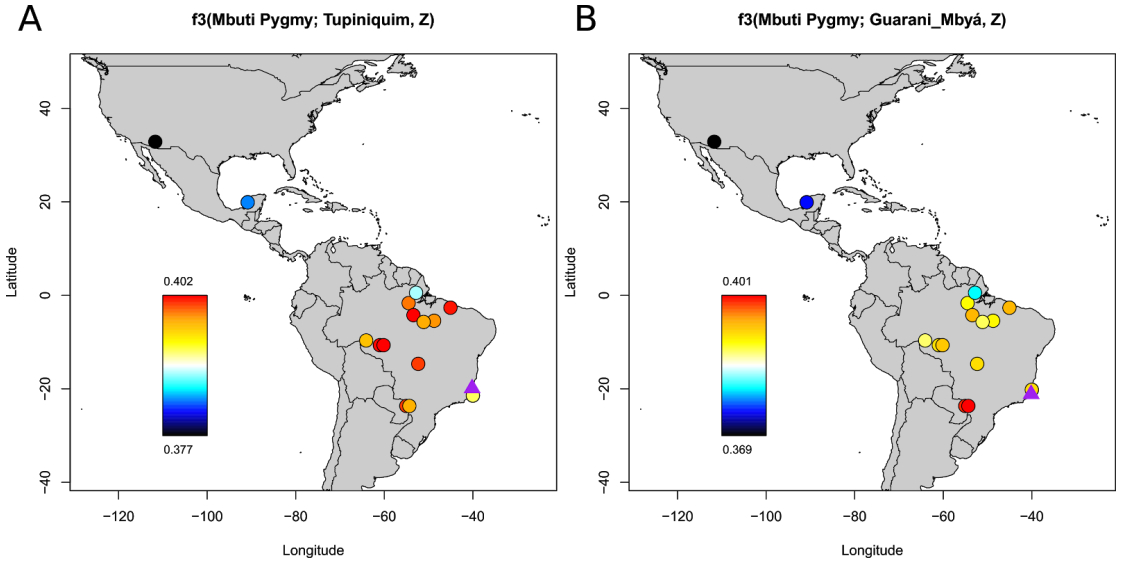


Figure S12. Patterns of allele sharing between modern Native American populations. The three population F-statistic (19) was calculated in the form F_3(Mbuti Pygmy; Y, Z) for all pairs of Y and Z modern Native American populations. The purple triangle represents the population fixed in Y, while the circles are the populations iterated over in Z. The heatmap depicts the spectrum between the maximum and minimum F_3 estimated for the set of comparisons in the panel. The estimates of each pair of Y and Z populations are color coded in the circles, with warmer and cold colors indicating more and less allele sharing, respectively. **A)** Y = Tupiniquim. **B)** Y = Guaraní Mybá.

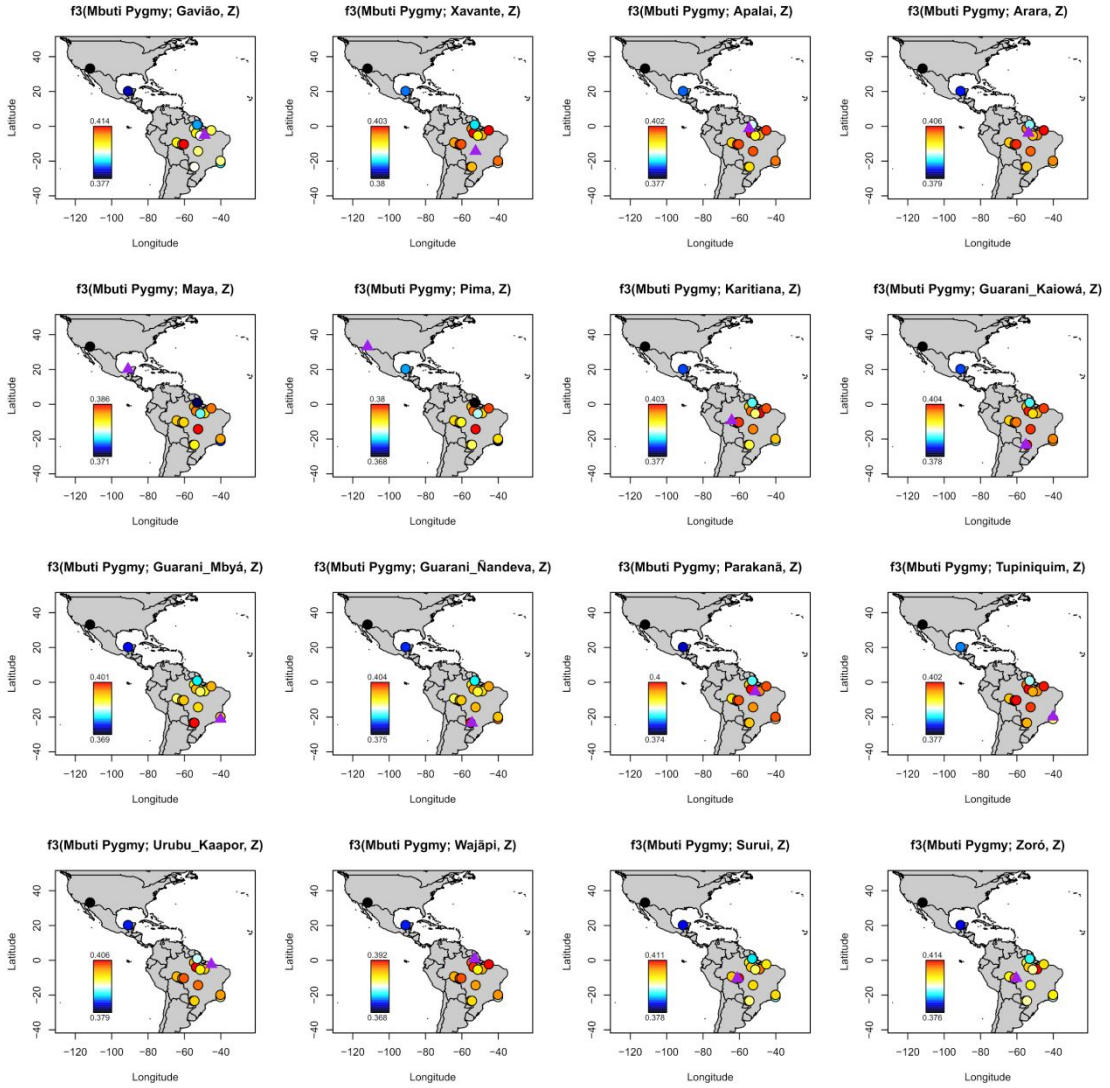


Figure S13. Patterns of allele sharing among Native Americans (dataset v). Outgroup-F_3 was estimated with Admixtools (1) using dataset v in the form F_3(Mbuti Pygmy; Y, Z), for every pair of Y and Z modern Native American populations. The purple triangle represents the population fixed in Y, while the circles are the modern populations iterated over in Z. The y and x-axes indicate Latitude and Longitude, respectively. In each panel, one population is fixed in Y position, and the legends present a scale of outgroup-F_3 values (i.e. allele sharing) between that population and each population at Z. In the set of comparisons of each panel, red and black indicates maximum and minimum allele sharing, respectively.

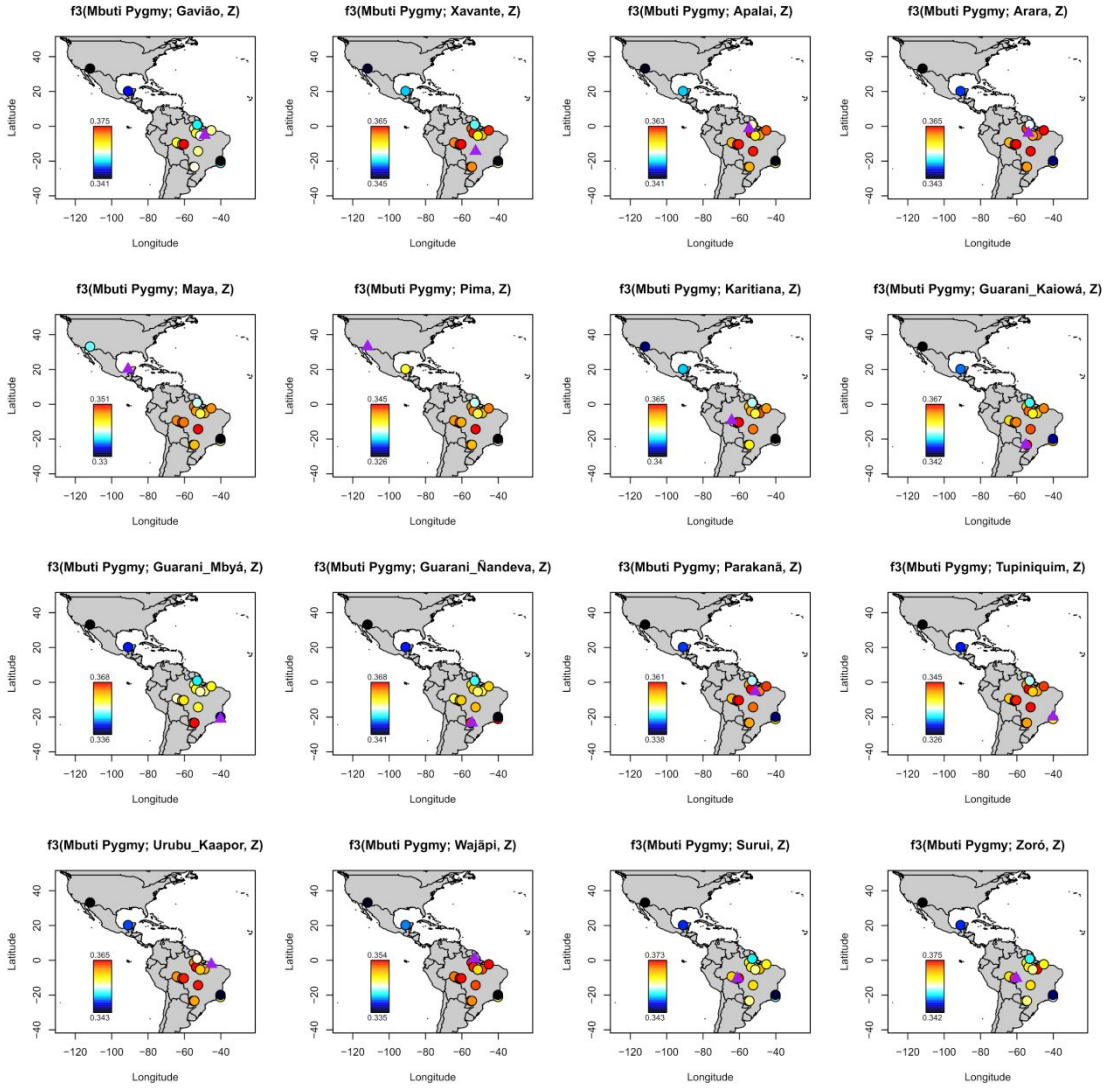


Figure S14. Patterns of allele sharing among Native Americans (dataset vi). Outgroup-F_3 was estimated with Admixtools (1) using dataset vi in the form F_3(Mbuti Pygmy; Y, Z), for every pair of Y and Z modern Native American populations. The purple triangle represents the population fixed in Y, while the circles are the modern populations iterated over in Z. The y and x-axes indicate Latitude and Longitude, respectively. In each panel, one population is fixed in the Y position, and the legends present a scale of outgroup-F_3 values (i.e., allele sharing) between that population and each population at Z. In the set of comparisons of each panel, red and black indicates maximum and minimum allele sharing, respectively.

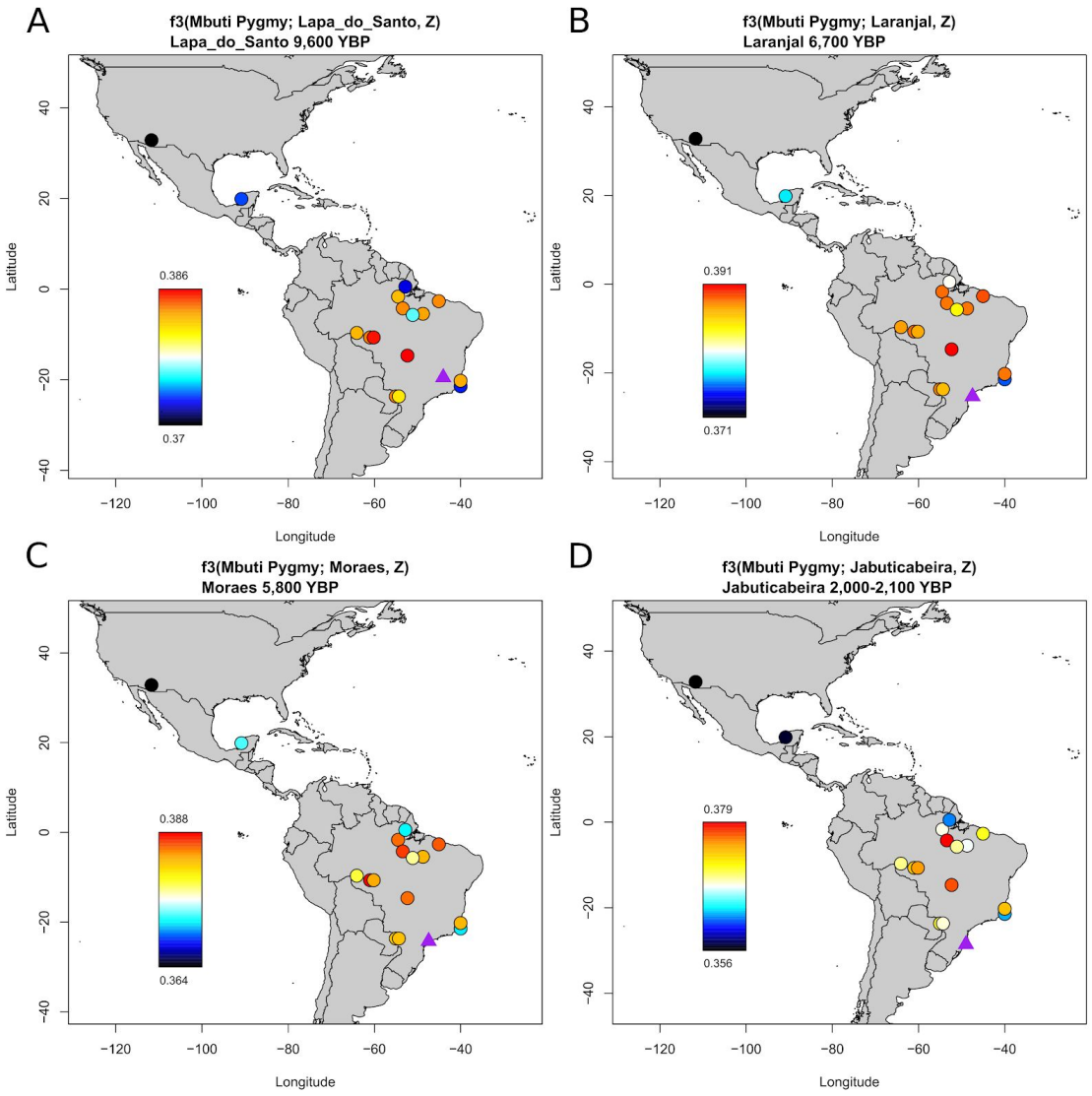


Figure S15. Patterns of allele sharing between modern and ancient Native Americans (dataset vi). The three population F-statistic (1) was calculated in the form F_3(Mbuti Pygmy; Y, Z) using dataset vi, for every Y archeological site (samples clustered by archeological site) and Z modern population. The purple triangle represents the site fixed in Y, while the circles are the modern populations iterated over in Z. The heatmap depicts the spectrum between the maximum and minimum F_3 estimated for the set of comparisons in the panel. The estimates of each pair of Y and Z populations are color-coded in the circles, with warm and cool colors indicating more and less allele sharing, respectively. **A)** Y = Lapa do Santo. **B)** Y = Laranjal. **C)** Y = Moraes. **D)** Y = Jabuticabeira.

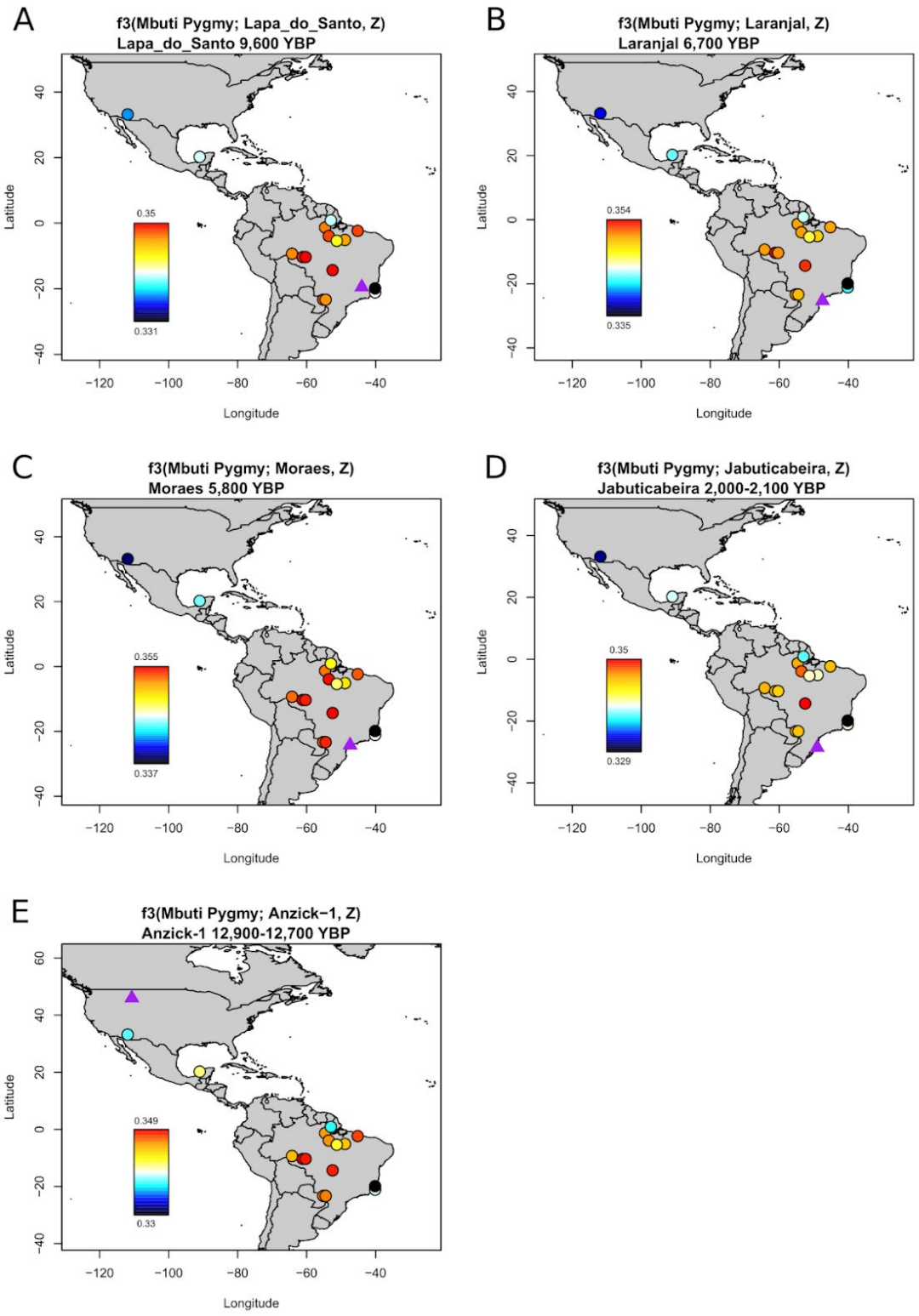


Figure S16. Patterns of allele sharing between modern and ancient Native Americans (dataset vi). The three population F-statistic (1) was calculated in the form F_3(Mbuti Pygmy; Y, Z) using dataset vi, for every Y archeological site (samples clustered by archeological site) and Z modern population. The purple triangle represents the site fixed in Y, while the circles are the modern populations iterated over in Z. The heatmap depicts the spectrum between the maximum and minimum F_3 estimated for the set of comparisons in the panel. The estimates of each pair of Y and Z populations are color-coded in the circles, with warm and cool colors

indicating more and less allele sharing, respectively. **A)** Y = Lapa do Santo. **B)** Y = Laranjal. **C)** Y = Moraes. **D)** Y = Jabuticabeira. **E)** Y = Anzick-1.

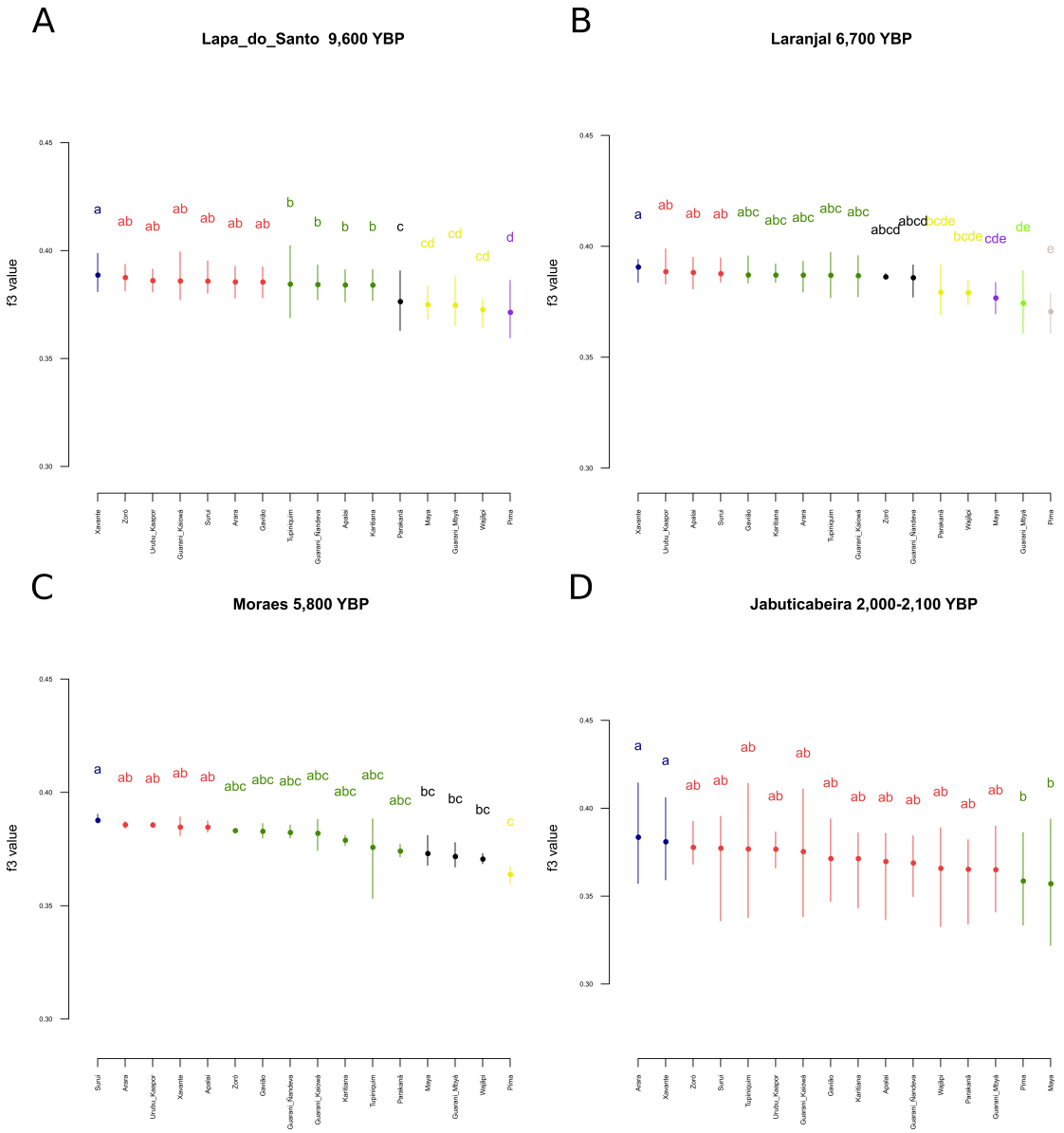


Figure S17. Significance of differences in allele sharing with ancient Native Americans (dataset v). Applying F-statistic to dataset v, in the same form as in Figure 4, i.e., F_3(Mbuti Pygmy; Y, Z), again for every Y archeological site (samples clustered by archeological site) and Z modern individual, instead of clustering individuals in populations as before. Therefore multiple estimates of F_3 are produced for each comparison and using these estimates the significance of the differences between averages can be tested. ANOVA and Tukey HSD (Honestly Significant Difference) test (R-scripts) were performed for each set of comparisons fixing one archeological site in the Y position. Significative differences between estimated F_3 values are detected for all comparisons (ANOVA - p-value < 0.001; Dataset S5). The X-axis presents the population from which individuals were drawn from to calculate the F_3 values, which are indicated in the Y-axis. Each panel shows the results for the comparisons when each of the following populations are fixed in Z position: **A)** Lapa do Santo. **B)** Laranjal. **C)** Moraes. **D)** Jabuticabeira. Pairs of means that are not significantly different according to a Tukey HSD test receive the same letter and groups of means that have received the same letters are painted with the same color, meaning that they share the same relationship with all other means.

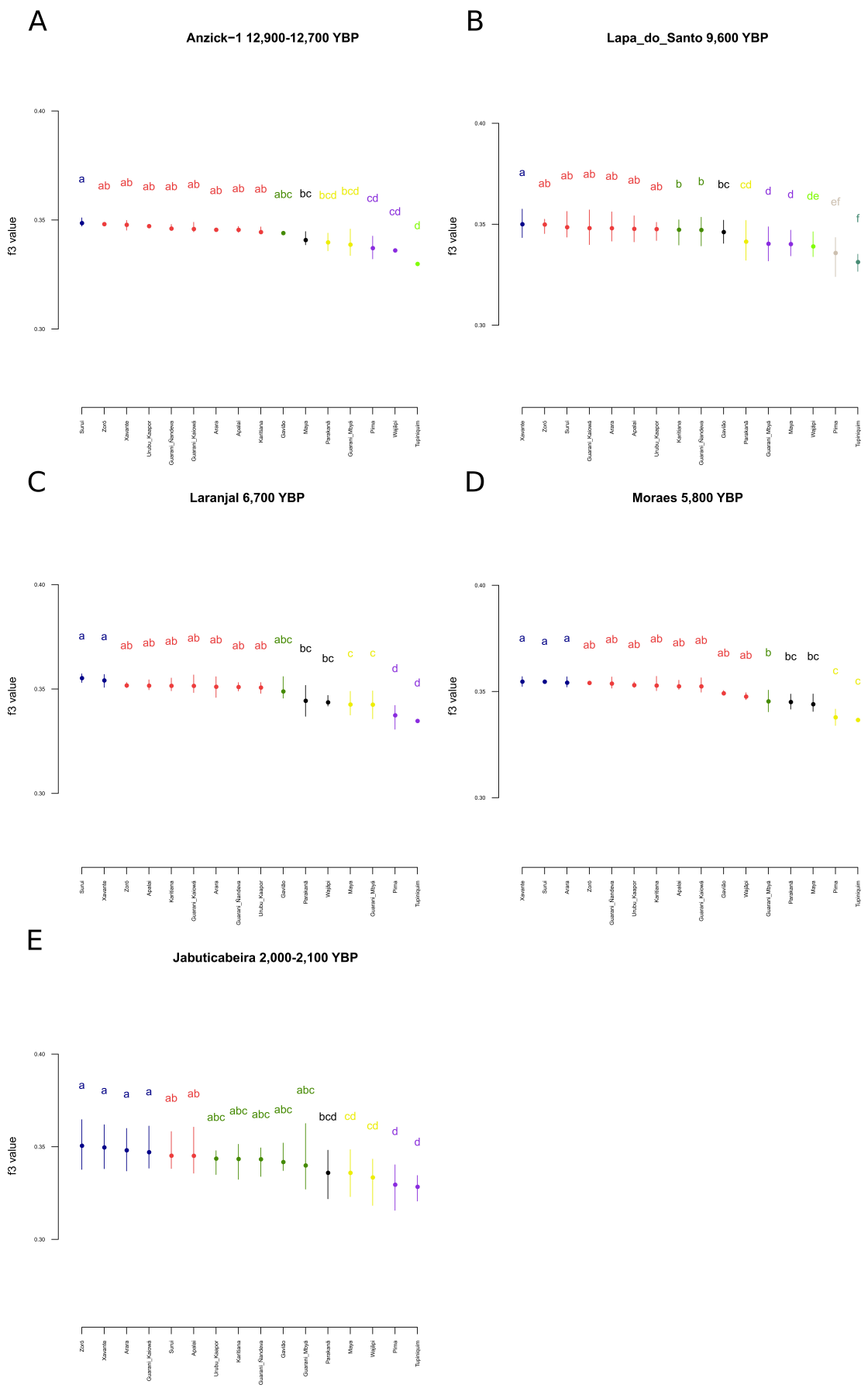

A
Anzick−1 12,900-12,700 YBP
B
Lapa_do_Santo 9,600 YBP
C
Laranjal 6,700 YBP
D
Moraes 5,800 YBP
E
Jabuticabeira 2,000-2,100 YBP
f3 value

Figure S18. Significance of differences in allele sharing with ancient Native Americans (dataset vi). Applying F-statistic to dataset vi, in the same form as in Figure S18, i.e., F_3(Mbuti Pygmy; Y, Z), again for every Y archeological site (samples clustered by archeological site) and Z modern individual, instead of clustering individuals in populations as before. Therefore multiple estimates of F_3 are produced for each comparison and using these estimates the significance of the differences between averages can be tested. ANOVA and Tukey HSD (Honestly Significant Difference) test (R-scripts) were performed for each set of comparisons fixing one archeological site in the Y position. Significative differences between estimated F_3 values are detected for all comparisons (ANOVA - p-value < 0.001; Dataset S5). The X-axis presents the population from which individuals were drawn from to calculate the F_3 values, which are indicated in the Y-axis. Each panel shows the results for the comparisons when each of the following populations are fixed in Z position: **A)** Anzick-1. **B)** Lapa do Santo. **C)** Laranjal. **D)** Moraes. **E)** Jabuticabeira. Pairs of means that are not significantly different according to a Tukey HSD test receive the same letter and groups of means that have received the same letters are painted with the same color, meaning that they share the same relationship with all other means.

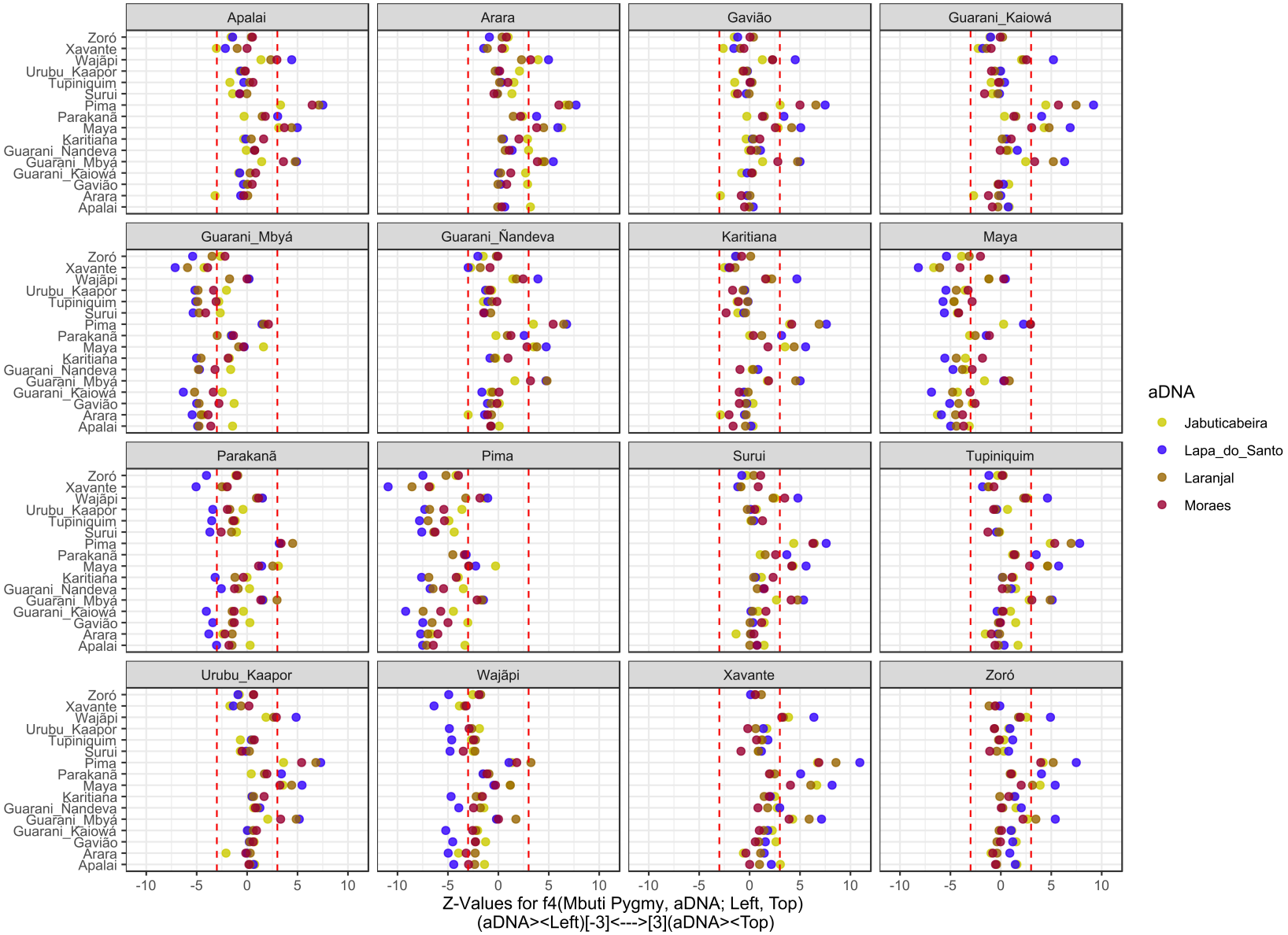

Apalai
Arara
Gavião
Guarani_Kaiowá
Guarani_Mbyá
Guarani_Ñandeva
Karitiana
Maya
Parakanã
Pima
Surui
Tupiniquim
Urubu_Kaapor
Wajãpi
Xavante
Zoró
Zoró
Xavante
Wajãpi
Urubu_Kaapor
Tupiniquim
Surui
Pima
Parakanã
Maya
Karitiana
Guarani_Nandeva
Guarani_Mbyá
Guarani_Kaiowá
Gavião
Arara
Apalai
-10
-5
0
5
10
Z-Values for f4(Mbuti Pygmy, aDNA; Left, Top)
(aDNA><Left)[-3]<--->[3](aDNA><Top)
aDNA
Jabuticabeira
Lapa_do_Santo
Laranjal
Moraes

Figure S19. Distinctive patterns allele sharing with ancient Native Americans (dataset v). F_4-Statistics was estimated with Admixtools (1) using dataset v in the form F_4(Mbuti Pygmy, aDNA; Left, Top); where 'aDNA' are ancient samples clustered by archeological site, while 'Left' and 'Top' are any modern Native American populations. In each panel one Native American population is fixed at 'Top' position, populations at the 'Left' position are represented in the Y-axis, and the 'aDNA' archeological site used in each calculation is labeled by color, as coded in the legend. The red dashed lines indicate the Z-value thresholds of Z = -3 and Z= 3, in each panel. Significantly negative estimated Z-values (Z < -3) indicate the 'Left' population shares more alleles with the referred archeological site (color) than the 'Top' population do, significantly positive values (Z > 3) indicate the opposite.

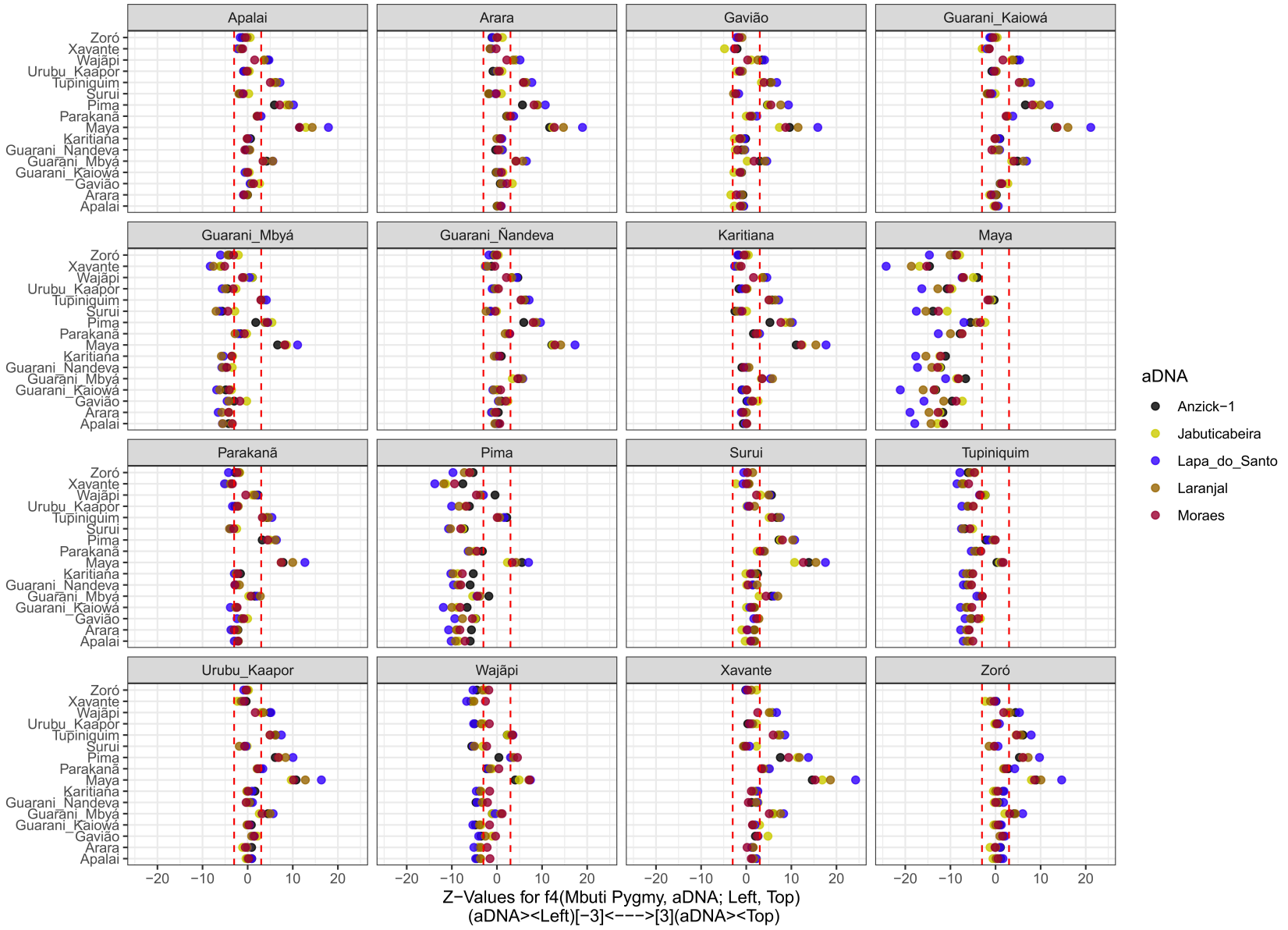

Apalai
Arara
Gavião
Guarani_Kaiowá
Guarani_Mbyá
Guarani_Ñandeva
Karitiana
Maya
Parakanã
Pima
Surui
Tupiniquim
Urubu_Kaapor
Wajãpi
Xavante
Zoró
−20
−10
0
10
20
Z−Values for f4(Mbuti Pygmy, aDNA; Left, Top)
(aDNA><Left)[−3]<−−−>[3](aDNA><Top)
aDNA
Anzick−1
Jabuticabeira
Lapa_do_Santo
Laranjal
Moraes

Figure S20. Distinctive patterns allele sharing with ancient Native Americans (dataset vi). F_4-Statistics was estimated with Admixtools (1) using dataset vi in the form F_4(Mbuti Pygmy, aDNA; Left, Top); where 'aDNA' are ancient samples clustered by archeological site, while 'Left' and 'Top' are any modern Native American populations. In each panel one Native American population is fixed at 'Top' position, populations at the 'Left' position are represented in the Y-axis, and the 'aDNA' archeological site used in each calculation is labeled by color, as coded in the legend. The red dashed lines indicate the Z-value thresholds of Z = -3 and Z= 3, in each panel. Significantly negative estimated Z-values (Z < -3) indicate the 'Left' population shares more alleles with the referred archeological site (color) than the 'Top' population do, significantly positive values (Z > 3) indicate the opposite.

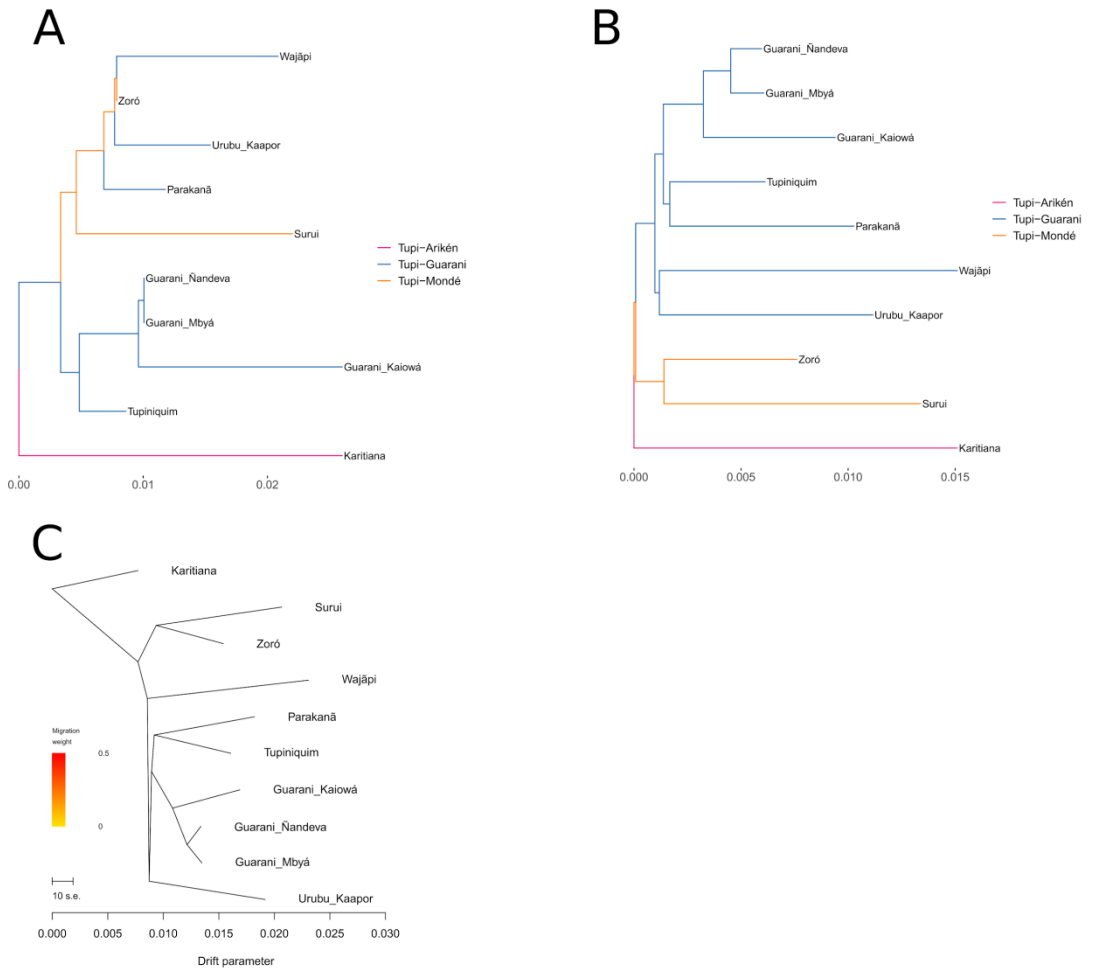


Figure S21. Tupí population phylogenies (dataset v). Pairwise F_{ST} and F_2 values were estimated for all pairs of Native American populations in dataset v using R scripts (https://github.com/BenjaminPeter/cph_course/blob/master/scripts/analysis.R) and plotted as Neighbor-Joining trees using the R packages *ape* and *ggtree* ((25); (26)). Branch lengths represent F_{ST} and $F_{2,}$ which are indicated in the X-axis. Based on population pairwise allelic covariance, Treemix (Pickrell & Pritchard, 2012) estimates a Maximum Likelihood tree. **A)** F_{ST} NJ tree. **B)** F_2 NJ tree. **C)** Pairwise allelic covariance Maximum Likelihood tree.

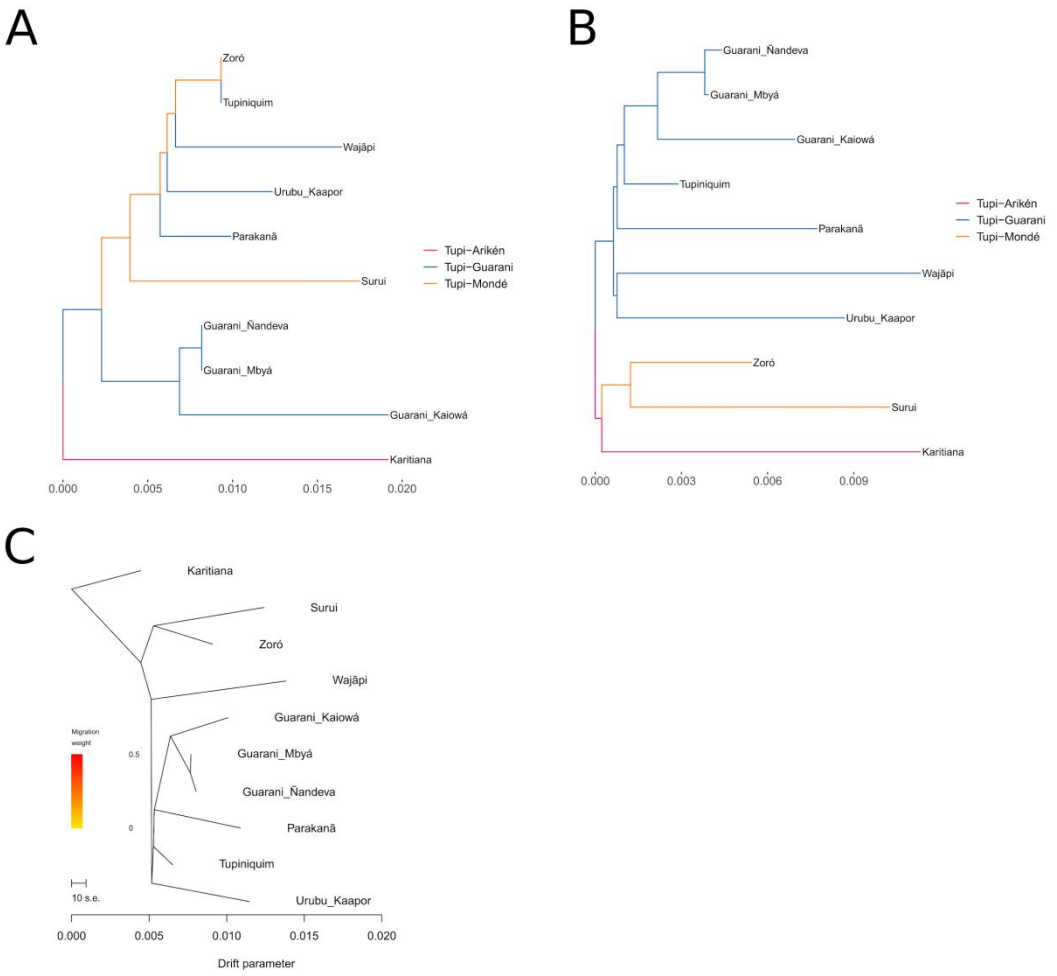


Figure S22. Tupí population phylogenies (dataset vi). Pairwise F_{ST} and F_2 values were estimated for all pairs of Native American populations in dataset vi using R scripts (https://github.com/BenjaminPeter/cph_course/blob/master/scripts/analysis.R) and plotted as Neighbor-Joining trees using the R packages *ape* and *ggtree* ((25); (26)). Branch lengths represent F_{ST} and $F_{2,}$ which are indicated in the X-axis. Based on population pairwise allelic covariance, Treemix (14) estimates a Maximum Likelihood tree. **A)** F_{ST} NJ tree. **B)** F_2 NJ tree. **C)** Pairwise allelic covariance Maximum Likelihood tree.

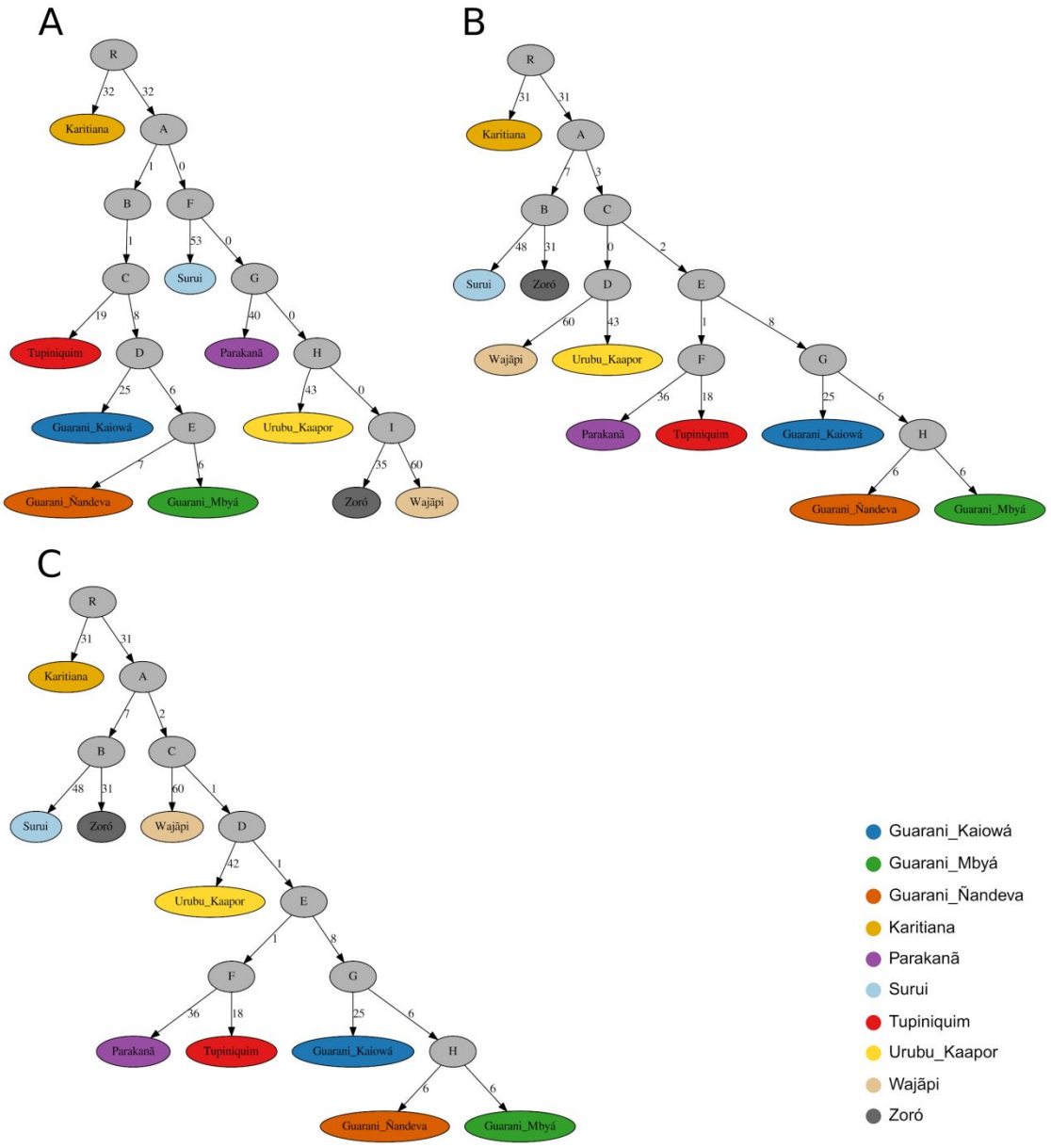


Figure S23. Assessment of the Tupí trees (dataset v). The fit between F-statistics estimated for the different Tupí trees and the expected estimates based on empirical data of dataset v was examined with *qpGraph* AdmixTools (1). Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. The maximum |Z| values estimated are 6.688, 3.682 and 3.334, for the models based on the (**A**) F_{ST} NJ tree, (**B**) the F_2 NJ tree and (**C**) the Pairwise allelic covariance Maximum Likelihood tree, respectively. The branch lengths are presented as units of F_{ST}, multiplied by 1000.

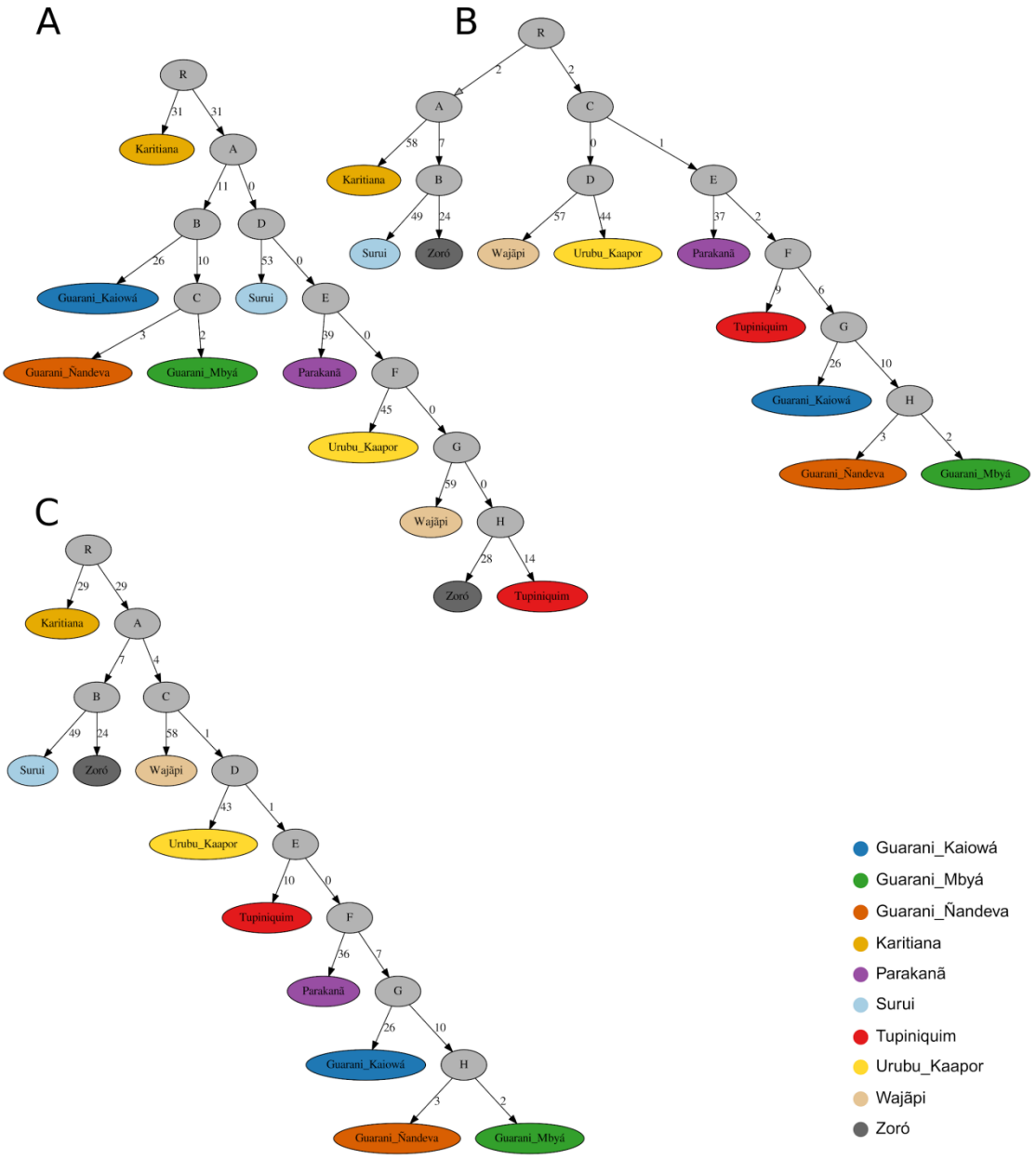


Figure S24. Assessment of the Tupí trees (dataset vi). The fit between F-statistics estimated for the different Tupí trees and the expected estimates based on empirical data of dataset vi was examined with *qpGraph* AdmixTools (1). Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. The maximum $|Z|$ values estimated are 7.034, 2.778 and 2.964, for the models based on the (**A**) F_{ST} NJ tree, (**B**) the F_2 NJ tree and (**C**) the Pairwise allelic covariance Maximum Likelihood tree, respectively. The branch lengths are presented as units of F_{ST}, multiplied by 1000.

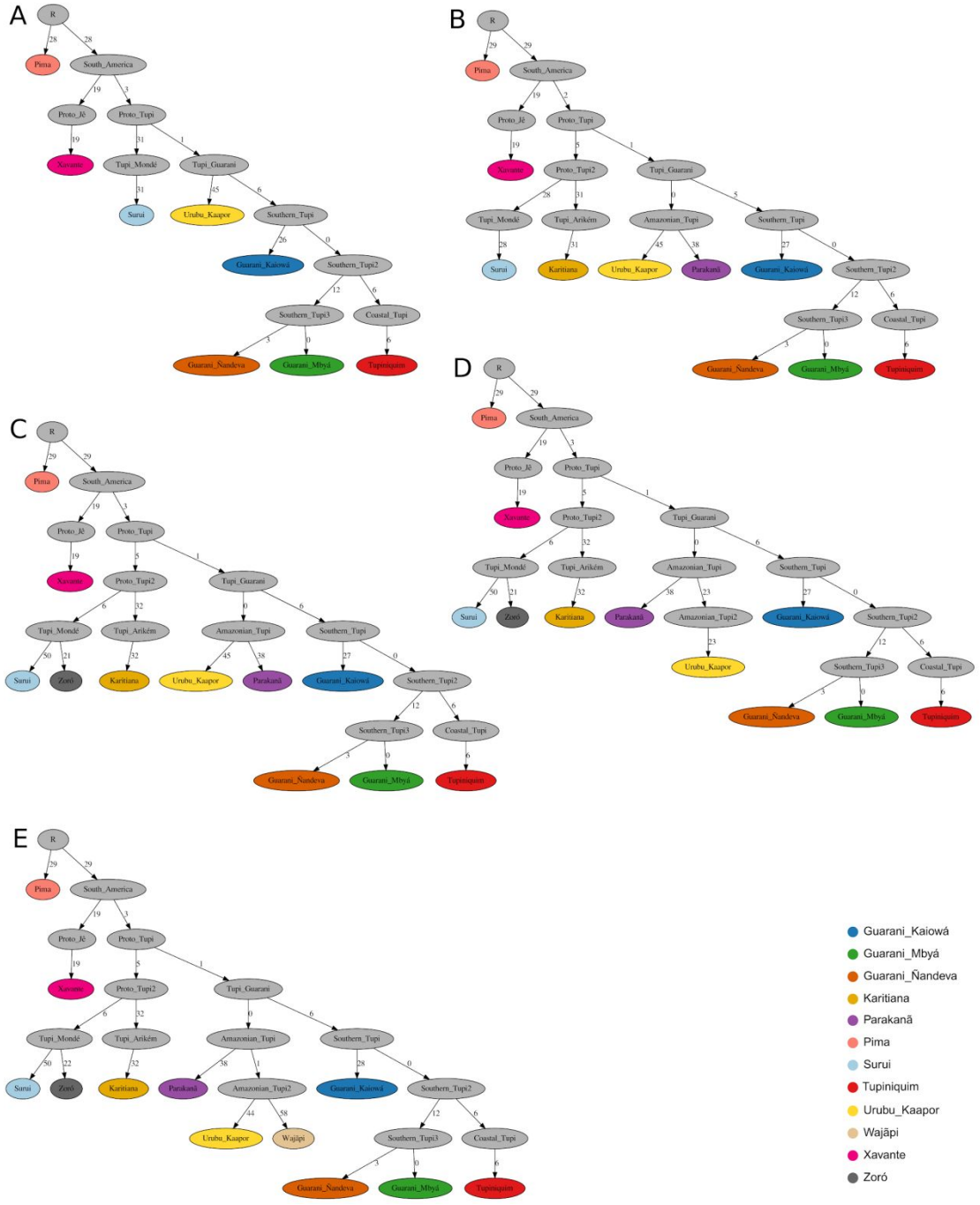


Figure S25. Alternative models of Tupí expansion hypothesis 1 (dataset vi). Several alternative models for hypothesis 1 of the Tupí expansion hypotheses were produced and assessed with *qpGraph* AdmixTools (1). The branch lengths are presented as units of F_{ST}, multiplied by 1000. Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. A summary of the maximum |Z| value estimated for each model is given in Dataset S6.

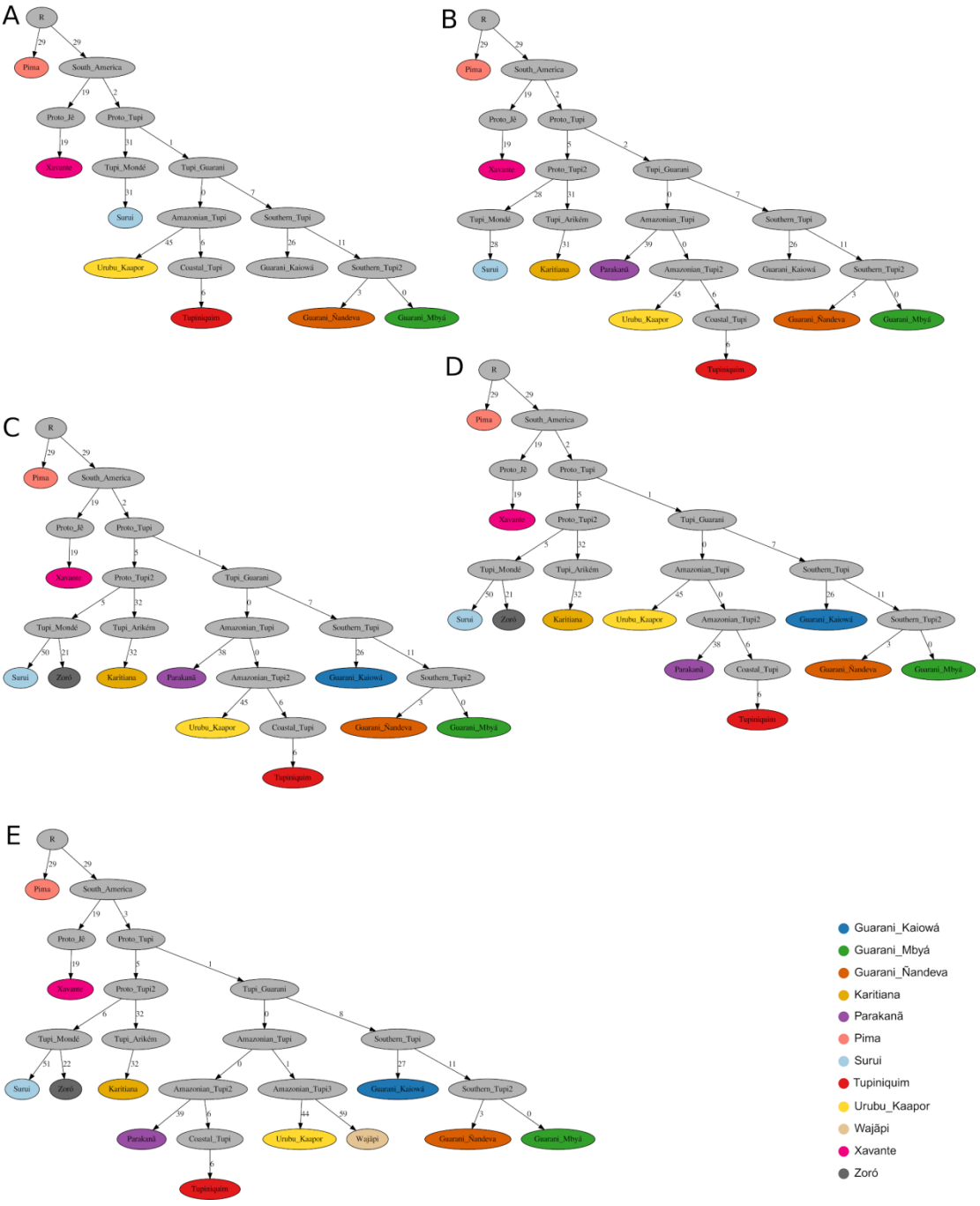


Figure S26. Alternative models of Tupí expansion hypothesis 2 (dataset vi). Several alternative models for hypothesis 2 of the Tupí expansion hypotheses were produced and assessed with *qpGraph* AdmixTools (1). The branch lengths are presented as units of F_{ST}, multiplied by 1000. Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. A summary of the maximum |Z| value estimated for each model is given in Dataset S6.

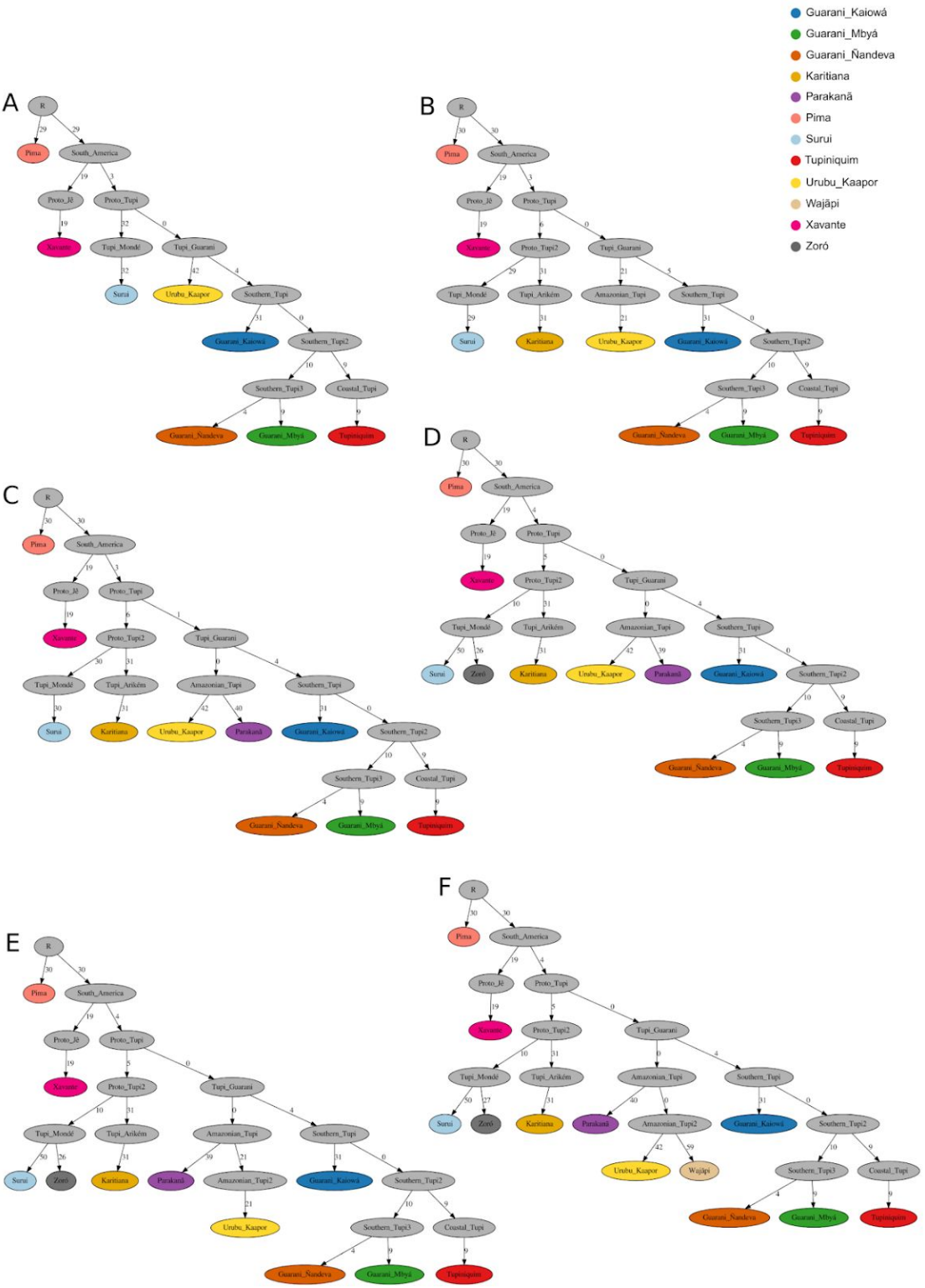


Figure S27. Alternative models of Tupí expansion hypothesis 1 (dataset v). Several alternative models for hypothesis 1 of the Tupí expansion hypotheses were produced and assessed with *qpGraph* AdmixTools (1). The branch lengths are presented as units of F_{ST}, multiplied by 1000. Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. A summary of the maximum |Z| value estimated for each model is given in Dataset S6.

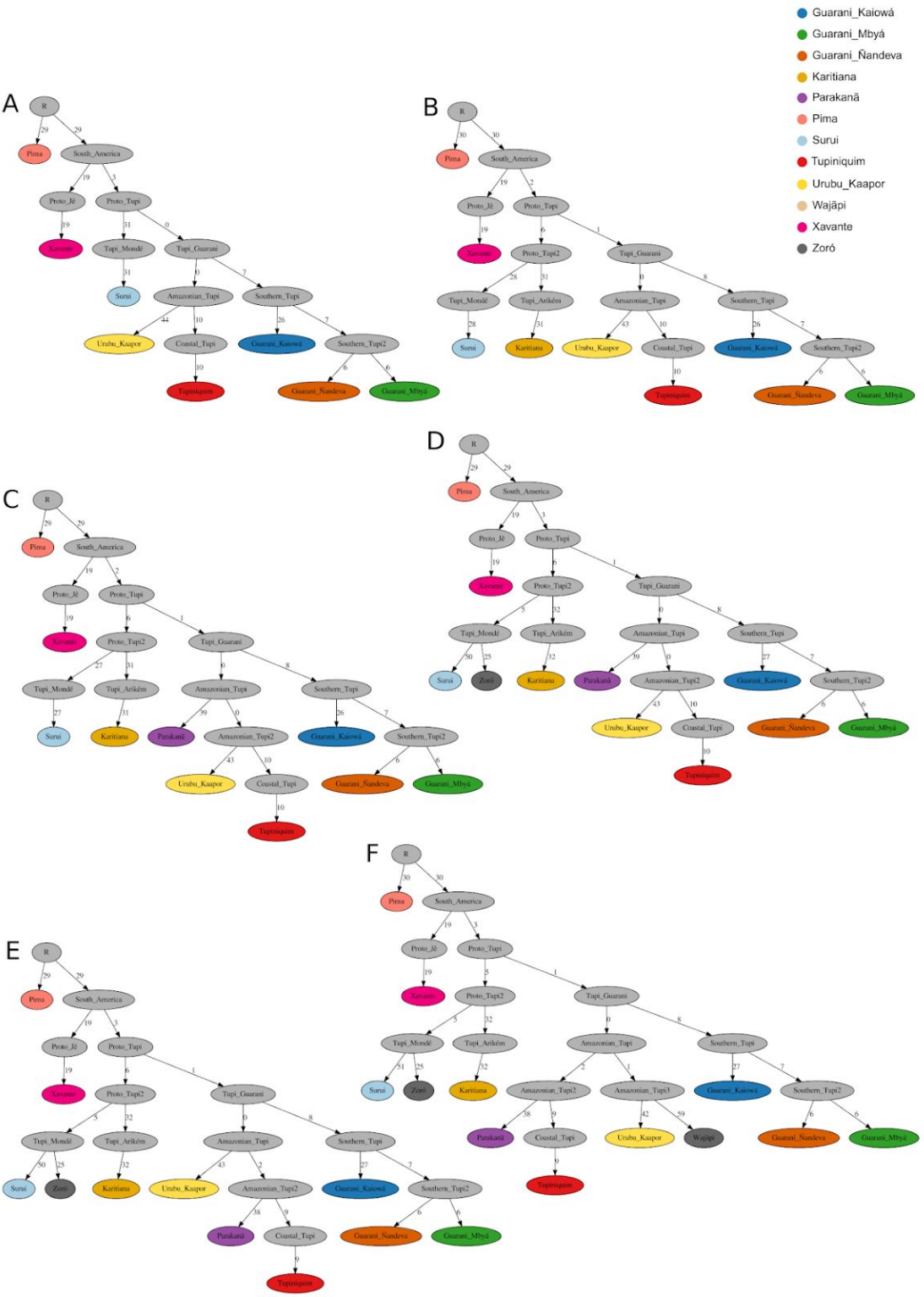


Figure S28. Alternative models of Tupí expansion hypothesis 2 (dataset v). Several alternative models for hypothesis 2 of the Tupí expansion hypotheses were produced and assessed with *qpGraph* AdmixTools (1). The branch lengths are presented as units of F_{ST}, multiplied by 1000. Conventionally, a good fit is indicated by a threshold of $|Z| < 3$ for all differences between expected and observed estimates. A summary of the maximum |Z| value estimated for each model is given in Dataset S6.

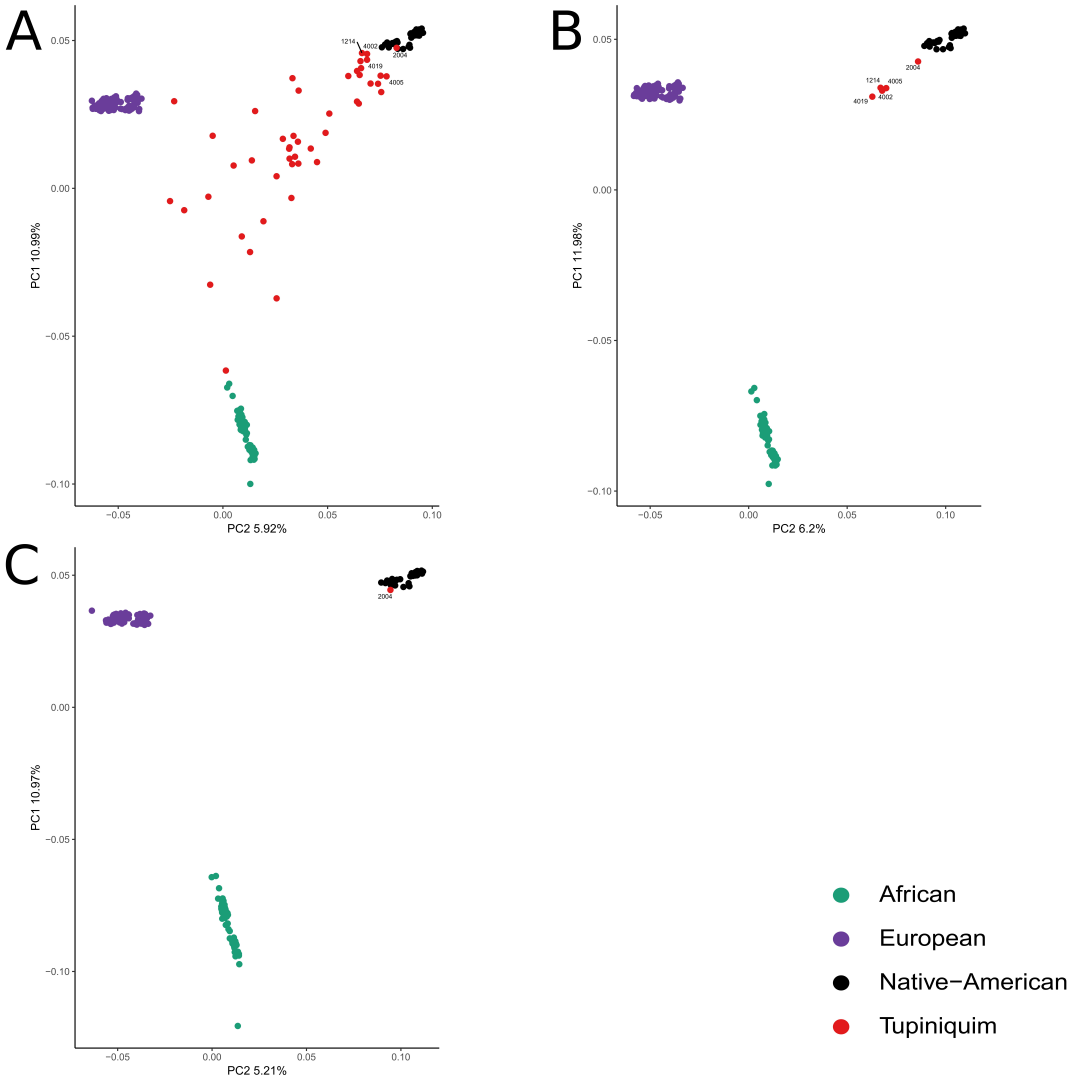


Figure S29. Global patterns of ancestry of the ancestral continental components in the Tupiniquim population. PCA was implemented with SNPRelate R/Bioconductor package (11), in all plots, the y and x-axes represent the first and second principal components (PCs) respectively. Samples are color-coded as in the legend. PCA plot of **A)** Dataset iv. **B)** Dataset v. **C)** Dataset vi. In **A** and **B** ID tags mark Tupiniquim individuals with high estimated Native American ancestry (1214, 2004, 4002, 4005 and 4019), and in **C** they mark the Tupiniquim with ~ 94% Native American ancestry (2004).

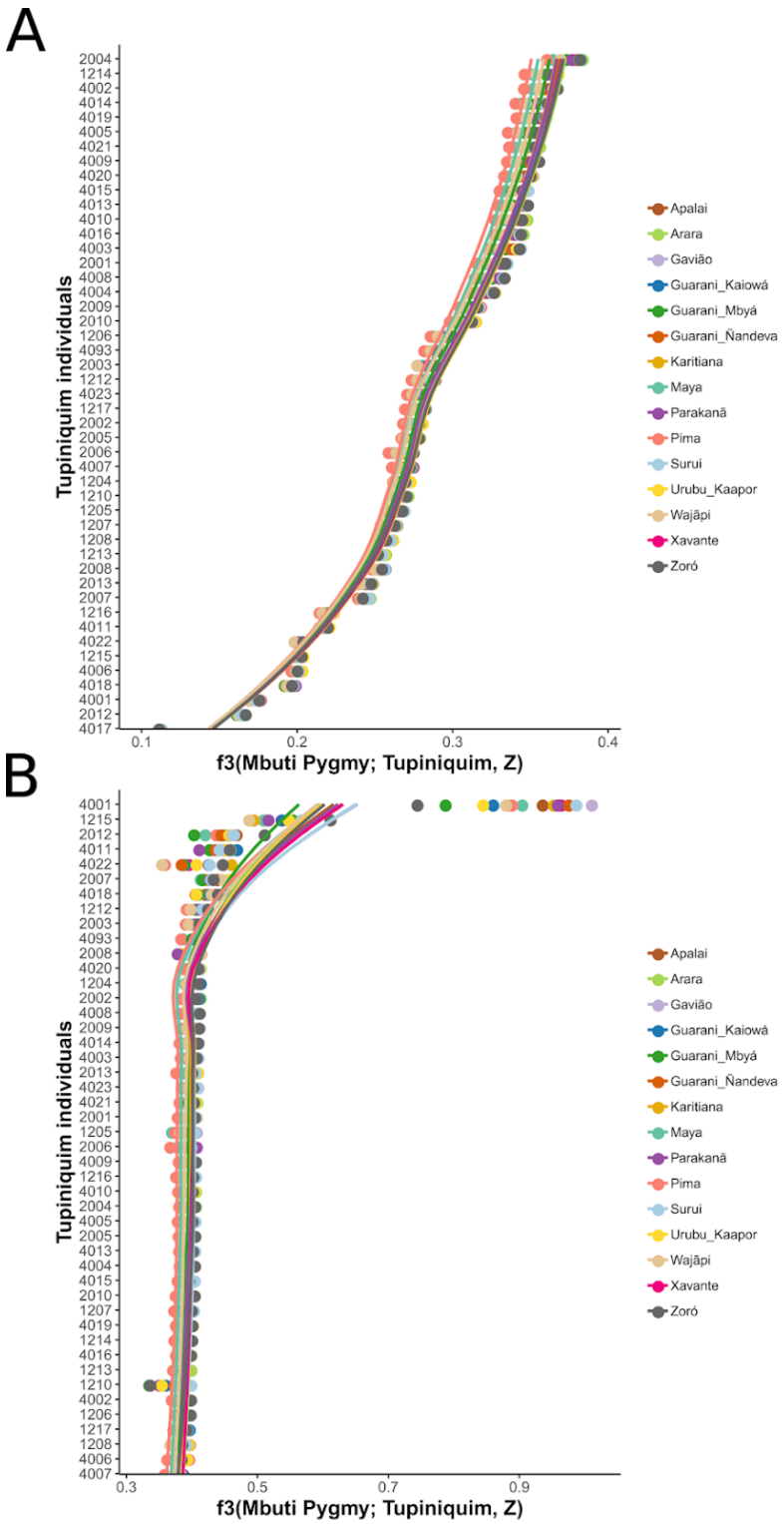


Figure S30. Allele sharing between Tupiniquim individuals and modern Native Americans. F-Statistic was estimated in the form F_3(Mbuti Pygmy; Tupiniquim, Z) with Admixtools (1), for every Tupiniquim individual and Z modern Native American population. The X-axis represents estimated F_3 values for every comparison between a Tupiniquim individual at the Y-axis and a modern population, the later labeled by the color. Circles indicate point estimates of F_3 statistic and smoothing lines were added with Loess (Local Polynomial Regression Fitting) method. **A)** Estimates obtained before masking the data (dataset iv). **B)** Estimated F_3 values after masking the data (dataset v).

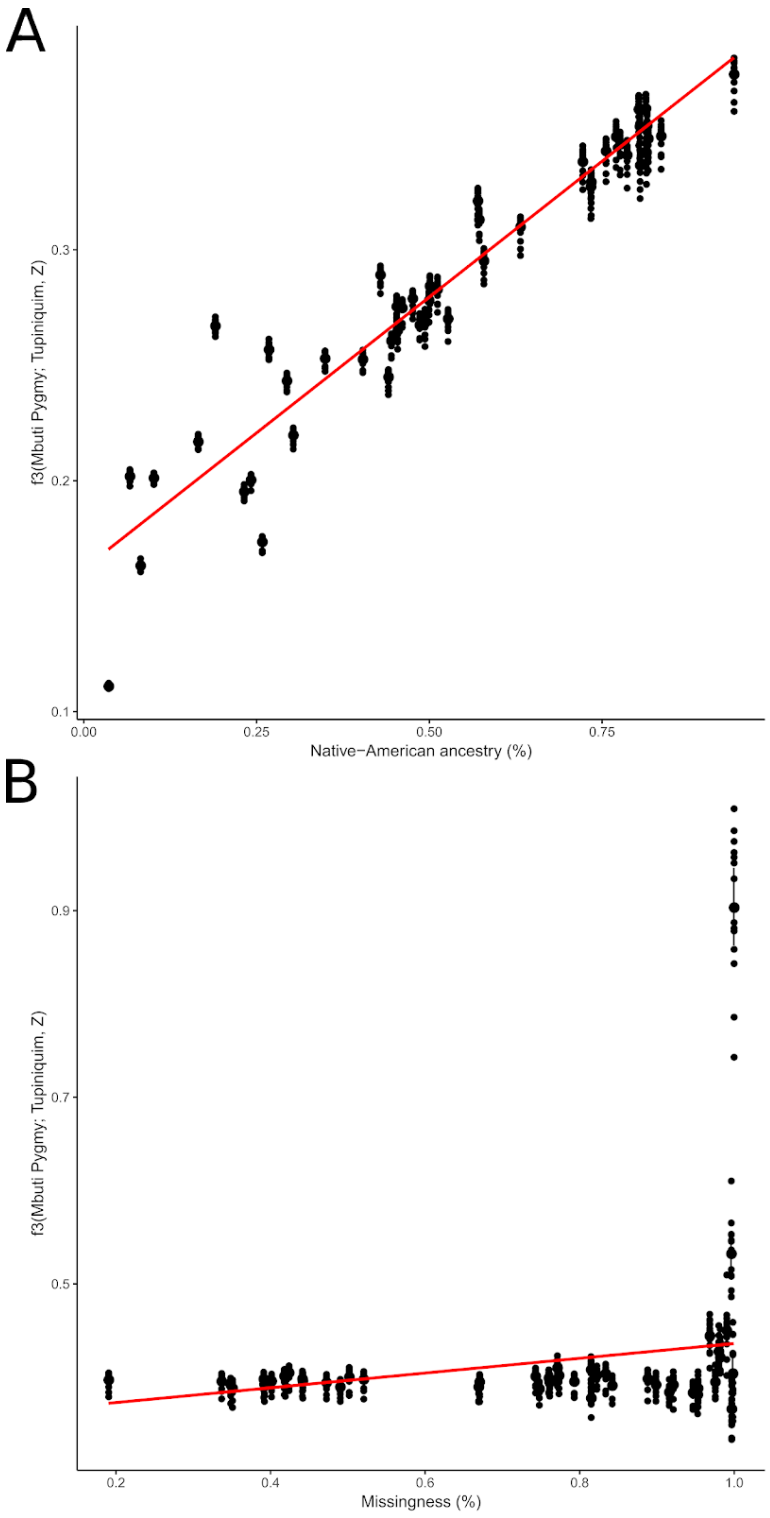


Figure S31. Interference of admixture and missingness in F-Statistics. As shown in Figure S28, estimates were obtained using the three population test in the form F_3(Mbuti Pygmy; Tupiniquim, Z) with Admixtools (1), for every Tupiniquim individual and Z modern Native American population, using datasets iv and v. Here the estimated F_3 values (Y-axis) are plotted against the proportion of Native American ancestry (Y-axis) for dataset iv (**A**) and also against the proportion of missing data (Y-axis) for dataset v (**B**). The red line is a linear regression fitted to the data, r^2 for **A** is 0.8939 and for **B** is 0.05731.

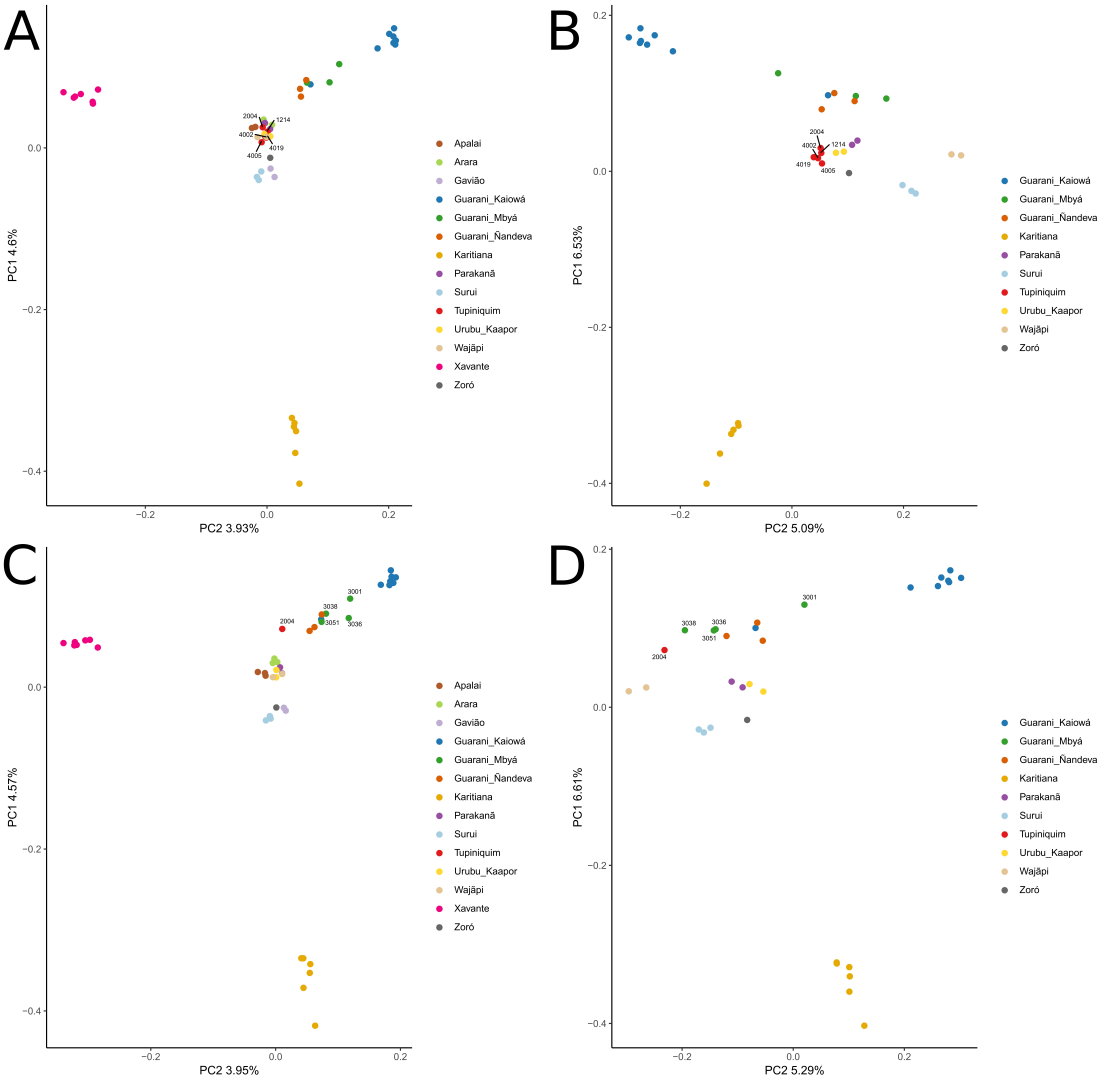


Figure S32. Global patterns of ancestry of the Native American and Tupí populations. PCA was realized through SNPRelate R/Bioconductor package (11), in all plots, the y and x-axes represent the first and second principal components (PCs) respectively. Samples are color-coded and for each plot, there is a legend on the right side. PCA plot of **A)** Native American populations from Dataset v; **B)** Tupí populations from Dataset v; **C)** Native American populations from Dataset vi; and **D)** Tupí populations from Dataset vi.

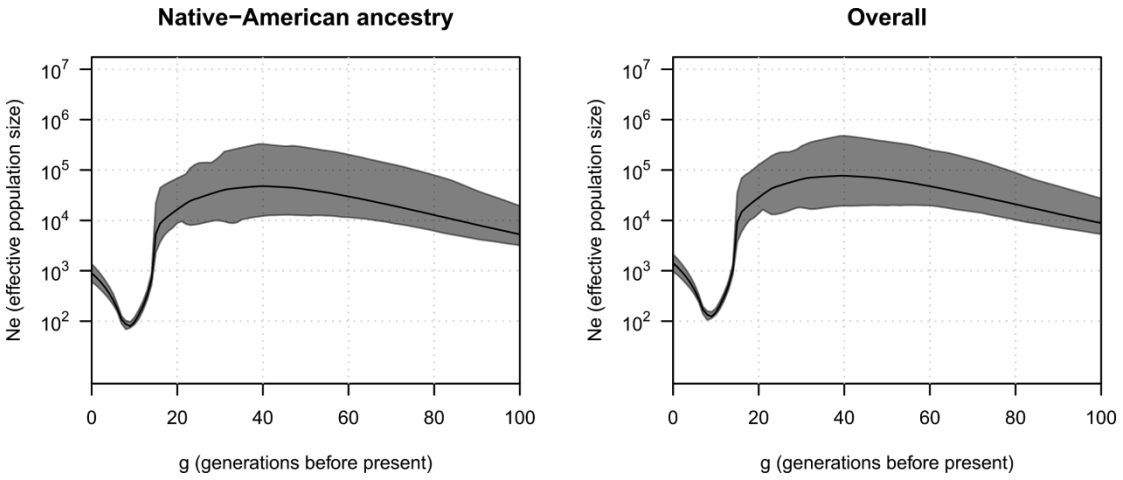
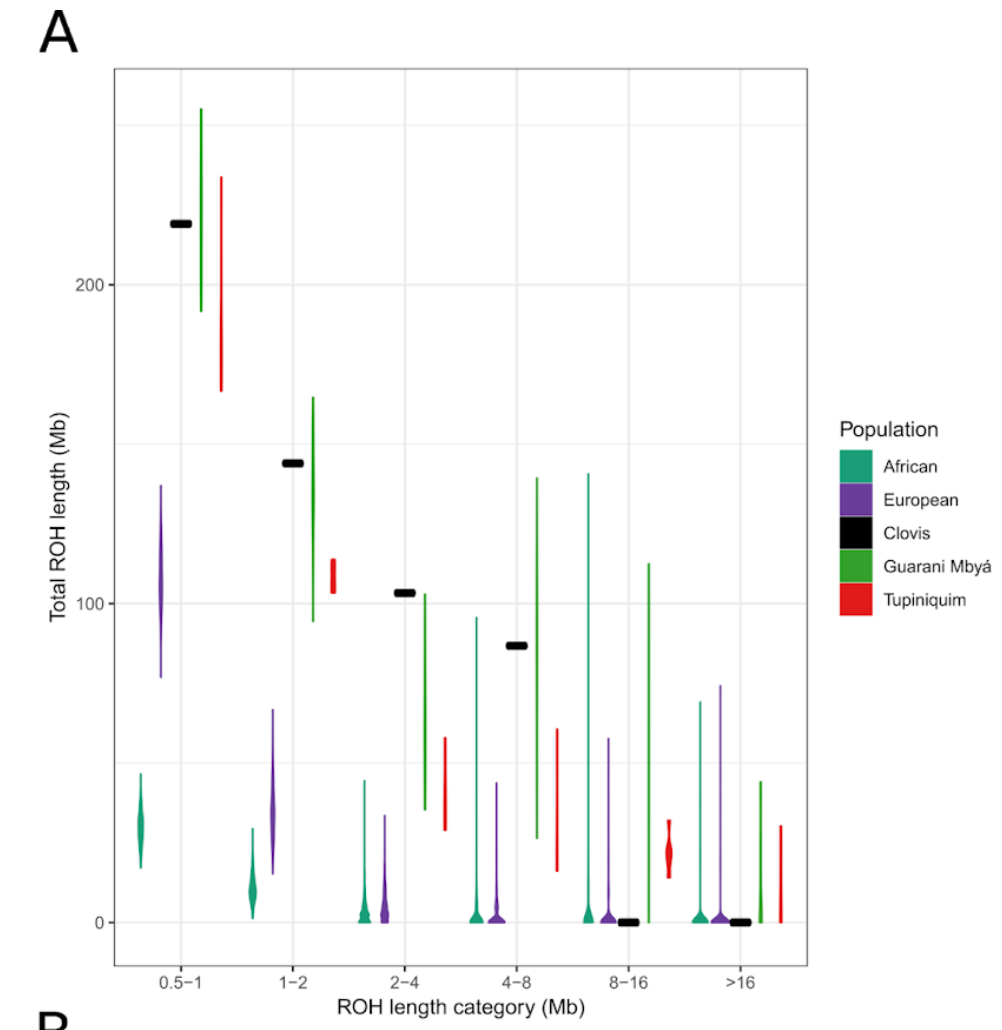


Figure S33. Ancestry-specific effective population size (Ne) history estimated for the Guaraní Mbyá. Related Guaraní Mbyá samples (k < 0.0625) were removed from dataset xi, which was then phased with Beagle v.5 (7). Based on the phased data, IBD segments were estimated with RefinedIBD (9) and Local Ancestry Inference with RFMix (8). Using IBDNe (10) Native American ancestry specific and overall Ne were estimated. Given the small amount of admixture (~ 20% in total; Figure S2b) and sample size African and European ancestry specific estimates could not be recovered. The ancestry-specific Ne values are coded in the Y-axis and indicated by the line for each generation before the present depicted in the X-axis. The grey areas show a 95% bootstrap confidence interval. Results for the Native American ancestry specific along with the Ne estimates obtained using all IBD segments (Overall) are shown in different panels.


A
Total ROH length (Mb)
200
100
0
0.5–1
1–2
2–4
4–8
8–16
>16
ROH length category (Mb)
Population
African
European
Clovis
Guarani Mbyá
Tupiniquim

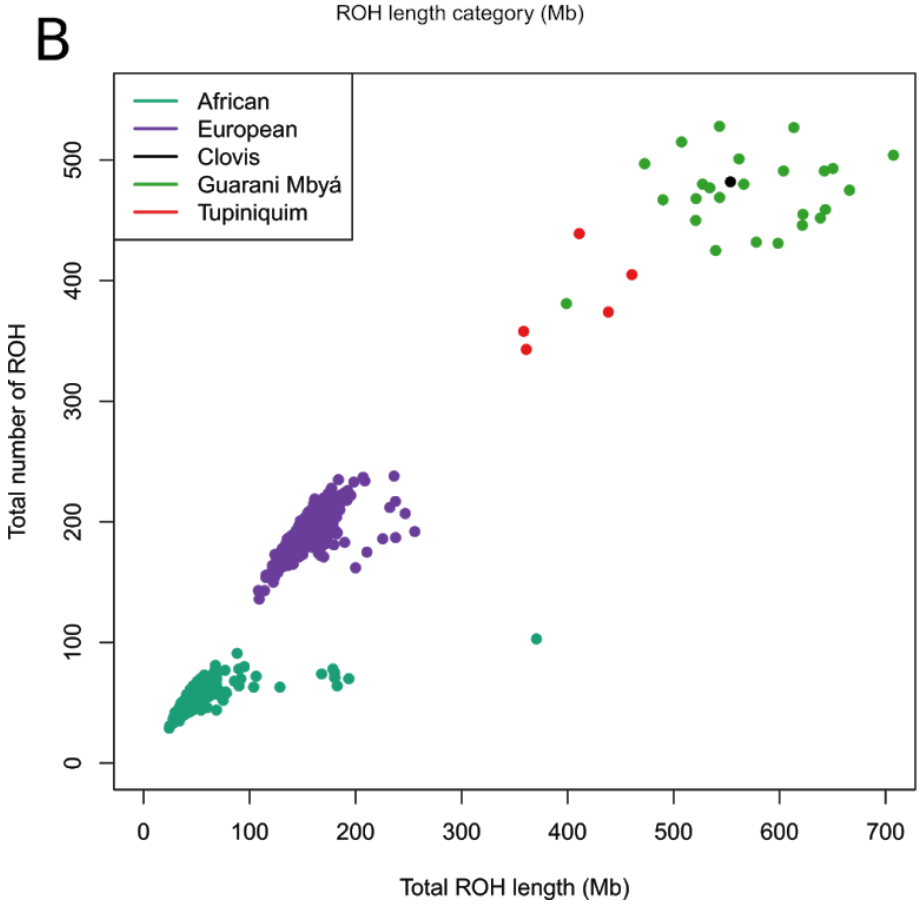

B
African
European
Clovis
Guarani Mbyá
Tupiniquim
Total number of ROH
0
100
200
300
400
500
0
100
200
300
400
500
600
700
Total ROH length (Mb)

Figure S34. ROH distribution in African, European, and Native American populations. The ROH identification was performed using the software PLINK v1.9 (5). **A)** Violin plots representing average total ROH lengths obtained per population, binned by the ROH length category. **B)** Scatterplot showing total length and the number of ROH of each individual according to population.

Supplementary Tables (S1 to S3)

Table S1 - Summary of the unpublished data used in this work.

	Datasets	Platform	SNPs
i	47 Tupiniquim / 48 Guaraní Mbyá	Axiom InCor BB (Affymetrix)	842019
ii	1 Tupiniquim / 4 Guaraní Mbyá / 2 Wajãpi 3 Parakanã / 2 Gavião	Axiom Human Origins (Affymetrix)	632293

Table S2 - Summary of the public data used in this work.

	Datasets	Platform
i	*Human Genome Diversity Project* dataset 11	Axiom Human Origins (Affymetrix)
ii	1000 Genomes Project	Whole-genome sequencing
iii	Skoglund *et al*., 2015	Axiom Human Origins (Affymetrix)
iv	Rasmussen *et al*., 2014	Whole-genome sequencing
v	Posth *et al*., 2018	'1240k SNP capture' (27) (Ancient DNA)

Table S3 - Summary of assembled datasets. In the first column, datasets are briefly described, followed in column 2 and 3 by the number of individuals and SNPs present in them, respectively.

nº	Datasets	N	SNPs
i	48 Guaraní Mbyá + Sub-Saharan Africans and Europeans (HGDP) + 48 Native Americans (2)	442	62607
ii	47 Tupiniquim (unmasked) + Sub-Saharan Africans and Europeans (HGDP) + 48 Native Americans (2)	441	62607
iii	47 Tupiniquim (unmasked) + 25 Guaraní Mbyá (> 90% Native American Ancestry) + 25 Sub-Saharan Africans and 25 Europeans (*1000 Genomes Project*)	122	676624
iv	47 Tupiniquim (unmasked) + 48 Guaraní Mbyá + 48 Native Americans (2) + 7 newly-genotyped Native Americans + HGDP	1101	71983
v	47 Tupiniquim (masked) + 48 Guaraní Mbyá + 48 Native Americans (2) + 7 newly-genotyped Native Americans + HGDP	1101	70464
vi	1 Tupiniquim (ID: 2004) + 4 Guaraní Mbyá (IDs: 3001, 3036, 3038, 3051) + 48 Native Americans (2) + 7 newly-genotyped Native Americans + HGDP	1006	572277
vii	47 Tupiniquim (unmasked) + 48 Guaraní Mbyá + Sub-Saharan Africans, Europeans and East Asians (*1000 Genomes Project*)	1079	676624
viii	47 Tupiniquim (unmasked) + 25 Guaraní Mbyá (> 90% Native American Ancestry) + Sub-Saharan Africans and Europeans (*1000 Genomes Project*) + Anzick-1 (Clovis Culture associated ancient DNA (3))	1080	436606
ix	47 Tupiniquim (masked) + 48 Guaraní Mbyá + 48 Native Americans (2) + 7 newly-genotyped Native Americans + HGDP + 15 Ancient DNA samples (4)	1150	70231
x	1 Tupiniquim (ID: 2004) + 4 Guaraní Mbyá (IDs: 3001, 3036, 3038, 3051) + 48 Native Americans (2) + 7 newly-genotyped Native Americans + HGDP + 15 Ancient DNA samples (4) + Anzick-1 Clovis Culture associated ancient DNA (3)	1059	356201

xi	48 Guaraní Mbyá + Peruvians from Lima (PEL), Sub-Saharan Africans and Europeans (*1000 Genomes Project*)	102	678050

Datasets (S1 to S6)

Dataset S1. Tupiniquim and Guaraní Mbyá Global ancestry proportions for each individual. The estimates presented here are the same used to produce Figure S2. Therefore they were obtained for Guaraní Mbyá and Tupiniquim individuals using datasets i and ii, applying ADMIXTURE (6) with a supervised analysis to estimate African, European, and Native American continental ancestry proportions. Tupiniquim and Guaraní Mbyá individual ancestry proportions are presented.

Dataset S2. Significative kinship estimates between and among Tupiniquim and Guaraní Mbyá individuals. Pairwise IBD estimates for all pairs of individuals inter and intra Tupiniquim and Guaraní Mbyá populations were obtained with SNPRelate R/Bioconductor package (11) using the PLINK method of moments. Pairs of related individuals, filtered with a threshold of k = 0.0625 (equivalent to a 1st cousin), inter and intra populations are presented.

Dataset S3. Admixture date estimates using alternative pairs of parentals. In the same way, as described in Figure S6, dates of admixture events between continental ancestry components (African, European and Native American) were inferred with Rolloff (1) using all pairs Native American and European, Native American and African, and finally African and European populations from dataset vii.

Dataset S4. TRACTS parameters inferred for each model. Based on RFMix (8) output, we have implemented the model-based approach of TRACTS software (23, 24). Here are presented all inferred parameters for the three demographic models tested: Model 1= Single-pulse, Model 2 = Discrete double pulse admixture model and Model 3= Double-pulse admixture model with a continuous migratory flow. Next to the name of each model, in parentheses are the log-likelihoods values.

Dataset S5. ANOVA testing the significance of the differences between means of outgroup F_3. The statistic F_3(Mbuti Pygmy; Y, Z) was calculated for all pairs of Native American individuals in Y modern populations and Z archeological sites, generating multiple estimates of F_3 for each comparison. ANOVA was used to test the significance of the differences between the means of the estimates obtained for each of the Y modern populations when fixing one Z archeological site.

Dataset S6. Summary of admixture graph models statistics reported in Figure S25**-**Figure S28**.** Alternative Tupí expansion hypotheses were modeled and assessed with qpGraph (1), including a growing number of populations and switching the Tupiniquim position. Here are summarized the statistics from several examples of these models.

References (for SI reference citations)

1. Patterson N, et al. (2012) Ancient Admixture in Human History. *Genetics* 192(3):1065–1093.

2. Skoglund P, et al. (2015) Genetic evidence for two founding populations of the Americas. *Nature* 525(7567):104–108.

3. Rasmussen M, et al. (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* 506(7487):225–229.

4. Posth C, et al. (2018) Reconstructing the Deep Population History of Central and South America. *Cell* 175(5):1185–1197.e22.

5. Chang CC, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

6. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.

7. Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* 103(3):338–348.

8. Maples BK, Gravel S, Kenny EE, Bustamante CD (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 93(2):278–288.

9. Browning BL, Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194(2):459–471.

10. Browning SR, et al. (2018) Ancestry-specific recent effective population size in the Americas. *PLoS Genet* 14(5):e1007385.

11. Zheng X, et al. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28(24):3326–3328.

12. Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 17(1):27–32.

13. Delaneau O, Zagury J-F (2012) Haplotype Inference. *Data Production and Analysis in Population Genomics*:177–196.

14. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.

15. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14):1801–1806.

16. Noelli FS (2008) The Tupí Expansion. *The Handbook of South American Archaeology*:659–670.

17. da Cunha MC (1992) *História dos índios no Brasil* (Editora Companhia das Letras).

18. Kirin M, et al. (2010) Genomic runs of homozygosity record population history and

consanguinity. *PLoS One* 5(11):e13996.

19. Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2):275–292.

20. Lemes RB, et al. (2018) Inbreeding estimates in human populations: Applying new approaches to an admixed Brazilian isolate. *PLOS ONE* 13(4):e0196360.

21. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet* 19(4):220–234.

22. Schroeder H, et al. (2018) Origins and genetic legacies of the Caribbean Taino. *Proc Natl Acad Sci U S A* 115(10):2341–2346.

23. Gravel S (2012) Population genetics models of local ancestry. *Genetics* 191(2):607–619.

24. Gravel S, et al. (2013) Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genetics* 9(12):e1004023.

25. Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526–528.

26. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8(1):28–36.

27. Fu Q, et al. (2015) An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524(7564):216–219.