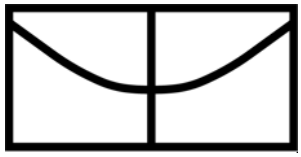
Universidade de Brasília
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Biologia Animal



Estrutura e Ancestralidade Genética de Populações Afro-derivadas Brasileiras

Carolina Carvalho Gontijo

Brasília

2019

Estrutura e Ancestralidade Genética de Populações Afro-derivadas Brasileiras

Tese apresentada ao Programa de Pós-Graduação em Biologia Animal da Universidade de Brasília como requisito parcial para a obtenção do título de Doutora.

Candidata: Carolina Carvalho Gontijo

Orientadora: Prof. Dra. Silviene Fabiana de Oliveira

Brasília

2019

Trabalho desenvolvido conjuntamente no Laboratório de Genética Humana do Instituto de Ciências Biológicas da Universidade de Brasília, Brasília, Brasil, e na *Unidade de Xenética do Instituto de Ciencias Forenses da Universidade de Santiago de Compostela*, Galícia, Espanha, com suporte financeiro da Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Distrito Federal (FAP-DF).

A minha família,

Dedico.

AGRADECIMENTOS

À minha orientadora Silviene Fabiana de Oliveira pelo apoio, confiança e amizade;

A Maria Victoria Lareu, Christopher Phillips e Ángel Carracedo pela supervisão durante o estágio doutoral e por generosamente me receberem em sua equipe para a execução deste projeto;

Aos membros da banca examinadora pela disponibilidade em avaliar esse trabalho;

Às agências de fomento CAPES, CNPq e FAPDF, pelo apoio financeiro;

Aos queridos professores e mestres que contribuiram para a minha formação e acompanharam, guiaram e apoiaram meu crescimento como cientista;

Às equipes do Laboratório de Genética Humana da UnB e da Unidade de Xenética Forense da USC pelo ambiente de trabalho acolhedor e estimulante;

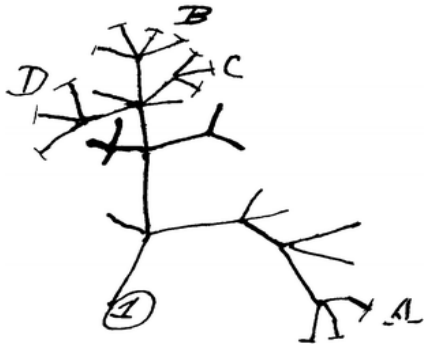
Aos amigos e colegas da UnB e da USC, em especial a Yarimar Ruiz, Ana Freire Aradas, Danel Rey, Carla Santos e Olalla Maroñas que me doaram tempo e conhecimento, e aos amigos de Santiago de Compostela, em especial Marina Martínez, Rodrigo Magañas, Ezequiel Parede e Liliane Buzzi, pela acolhida em um país estranho;

Ao Fábio pelo companheirismo na vida e na academia e pelas longas horas dedicadas ao meu trabalho;

À minha irmã (e consultora para assuntos de história e antropologia) Raquel G e aos grandes amigos que tive a sorte de ter ao meu lado durante os longos anos de doutorado e nos momentos mais difíceis, em especial Erica, Cinthya, Daya, Thiago, Raquel P, Dudu, Elisa, Bia e Rogério;

Agradeço.

I think



ÍNDICE GERAL

ÍNDICE DE TABELAS

Metodologia geral

Capítulo 1

Tabelas suplementares

(disponíveis em https://github.com/ninacalorina/supplementary-material-thesis)

Capítulo 2

Tabelas suplementares

(disponíveis em https://github.com/ninacalorina/supplementary-material-thesis)

Tabel S1. Database. Complete database tested in the exploratory Structure analysis, pairwise Fst estimates and PCA. Populations/samples in bold were excluded from ancestry analysis.

Table S2. Allelic frequencies. Allelic frequencies observed in the quilombos Kalunga and Sacutiaba and the palenque Mulaló.

Table S3. Descriptive parameters. Number of genotypes analysed, Hardy-Weinberg Equilibrium (p-value, s.d.), Observed (Hobs) and expected (Hexp) heterozygosity, number of pairs a given locus is in linkage disequilibrium (LD) for the 34plex panel in the quilombos Kalunga and Sacutiaba and in the palenque Mulaló. Highlighted cells: p-value < 0.01. Highlighted: loci with p-value ± s.d. < 0.01.

Table S4. Pairwise F_{ST}. Pairwise F_{ST} between populations from Africa (AFR), Europe (EUR), Indigenous America (IAM), East Asia (EAS), Central and South Asia (CSA), Northern Africa (NAF), Middle East (MID) and the African derived study populations from Brazil (KAL and SAC) and Colombia (MUL). In dark grey: continental groups and study populations. In red: F_{ST} > 0.10 within continental group. In blue: study populations. In light grey: group formed by EUR, CSA, NAF and MID (mean F_{ST} < 0.0,9). In green: Uygur (CSA) x EAS (mean F_{ST} < 0.10).

Table S5. Ancestry estimates in Mulaló, Kalunga and Sacutiaba. Individual and population ancestry estimates considering Africans (AFR), Europeans (EUR) and Indigenous Americans (IAM) as contributing populations to admixture in the formation of the study populations. Mean and median values, s.d., and minimum and maximum estimates are shown.

Capítulo 3

Table 1. Ancestry estimates in the full range of study populations from the American continent based on PIMA+34plex genotypes analysed with STRUCTURE. Average cluster

Tabelas suplementares
(disponíveis em https://github.com/ninacalorina/supplementary-material-thesis)

Supplementary Table S1A. Genotype data used in all population analyses; comprising 25 of 27 PIMA and 27 of 34plex SNPs. REF: Reference population data. **Table S1B.** HGDP-CEPH population data (genotypes and population-based allele frequency estimates) for tri-allelic SNP rs17287498, not genotyped for American study populations but developed for the final version of the PIMA multiplex.

Supplementary Table S2A. Genomic details of the 28 markers of the PIMA panel. **Table S2B.** PCR primer and SNaPshot extension (EXT) primer details of the final PIMA multiplex design. **Table S2C.** GRCh37 coordinates of 62 SNPs of the PIMA and 34plex panels. The separation in megabases (Mb) of syntenic SNPs is shown and the three closest SNP pairs marked in red.

Supplementary Table S3A. Descriptive population genetics parameters for PIMA+34plex in study and reference populations: Hardy-Weinberg equilibrium (HWE) tests, observed (Hobs) and expected (Hexp) heterozygosity, monomorphic allele rates. **Table S3B.** Pairwise linkage disequilibrium analyses. **Table S3C.** Allele frequencies (25/27 PIMA SNPs; 27/34 34plex SNPs). Allele calls are for SNaPshot-detected nucleotide substitutions.

Avaliação geral das estimativas de ancestralidade

ÍNDICE DE FIGURAS

Metodologia geral

Capítulo 1

Capítulo 2

Capítulo 3

Avaliação geral das estimativas de ancestralidade

ABREVIAÇÕES

AIM: Ancestry Informative Marker
AFR: África
AME: América
ASW: African ancestry from NW USA
BSA: Bovine Serum Albumin
CEP: Comitê de Ética em Pesquisa
CLM: Colômbia
CONEP: Comissão Nacional de Ética em Pesquisa
CSA: Central and South Asia
ddNTP: dideoxinucleotídeo
dNTP: deoxinucleotídeo
EAS: East Asia
EUR: Europa
FCP: Fundação Cultural Palmares
FUNAI: Fundação Nacional do Índio
Hexp: expected heterozygosity
HGDP-CEPH: Human Genome Diversity Project - Center for the Study of Human Polymorphisms
Hobs: observed heterozygosity
HWE: Hardy-Weinberg Equilibrium
IAM: Indigenous American
IBGE: Instituto Brasileiro de Geografia e Estatística
Indel: inserção/deleção
KAL: Kalunga
LD: Linkage Disequilibrium; Desequilíbrio de Ligação
MDS: Multidimensional Scaling
MOC: Mocambo
mtDNA: DNA mitocondrial
MXL: Mexicans in LA USA
PCA: Principal Component Analysis
PCR: Polymerase Chain Reaction
PUR: Puerto Ricans in Puerto Rico
SAC: Riacho de Sacutiaba e Sacutiaba
SAP: Shrimp Alkaline Phosphatase
s.d.: desvio padrão
SEPPIR: Secretaria Nacional de Políticas de Promoção da Igualdade Racial
SNP: Single Nucleotide Polymorphism
STR: Simple Tandem Repeat
Taq: *Thermophilus aquaticus*

RESUMO

A América Latina tem uma história intrincada que está refletida na composição genética de suas populações humanas. Apesar da crescente literatura voltada para a genética de populações do continente, populações não urbanas e não indígenas, que perfazem cerca de 20% das populações caribenhas e sul americanas, estão sub representadas em bancos de dados públicos adequados para análises de estruturação e ancestralidade. Essa deficiência é problemática tanto para a genética forense, que depende de descrições acuradas da variação genética e estruturação populacional, quanto para o entendimento da história do continente e dos eventos demográficos que originaram as populações contemporâneas. Ao longo de todo o continente americano, africanos escravizados fugidos ou libertos e seus descendentes formaram comunidades como forma de resistência contra o sistema escravista ou como consequência de seu colapso. Muitas dessas comunidades ainda existem e se mantiveram, em certa medida, geneticamente isoladas de outras populações. Os quilombos brasileiros e os *palenques* e *comunidades negras* colombianos são exemplos dessas comunidades. Neste trabalho, nosso objetivo principal foi contribuir para o conhecimento a respeito dos quilombos gerando informação sobre sua estrutura e ancestralidade, visando assim melhorar a representatividade de populações rurais em bancos de dados de interesse forense e antropológico. Adicionalmente, buscamos contrastar nossas populações de estudo uma à outra, a suas parentais presumidas a outras populações, além de avaliar a que ponto paralelos históricos estão refletidos na ancestralidade genética. Dois quilombos rurais (Kalunga - GO e Riacho de Sacutiaba e Sacutiaba - BA) foram analisados para os painéis de AIMs *for*InDel, SNPforID 34plex e PIMA e para a combinação PIMA + 34plex. Duas outras populações, o quilombo Mocambo - SE e a *comunidad negra* Mulaló, foram incluídas em determinadas análises para as quais haviam dados disponíveis. Assumimos um modelo de mistura tri-híbrido baseado em plausibilidade histórica e selecionamos populações parentais do HGDP-CEPH seguindo esse critério principal e a quantidade de variação observada dentro e entre os grupos formados por elas. Apresentamos também um novo sistema de AIMs

(PIMA: *Population Informative Multiplex for the Americas*) que se provou eficiente em diferenciar o componente de mistura indígena americano dos de outras populações continentais - especificamente africanos, europeus e leste asiáticos. Nossas análises de ancestralidade genética e estruturação confirmaram que os principais componentes de mistura nos quilombos e na *comunidad negra* acessados são africano e europeu. Apesar dessa similaridade, provavelmente derivada de paralelos históricos e de origem, nossos resultados afirmam que essas populações não são homogêneas. Além de atestar a diversidade existente nas populações humanas da América Latina, nossas observações reiteram a necessidade de descrever populações outras que as urbanas e indígenas para que seja possível construir bancos de dados adequados para a análise de ancestralidade.

Palavras-chave: ancestralidade genética, AIM, quilombo, afro-derivado

ABSTRACT

Latin America has an intricate history which is reflected in the genetic composition of its human populations. In spite of the growing literature on Latin American populations genetics, non-urban and non-indigenous communities, which comprise around 20% of Caribbean and South American populations, are underrepresented in public databases fit for ancestry and structure analysis. That deficiency is problematic for forensic genetics, as it relies on accurate descriptions of genetic variation and population structure, and for the understanding of the continent's history and demographic events that gave rise to its population. All over the continent, communities were founded by runaway or otherwise freed African slaves and their descendants as a way of resisting the slave system or as a consequence of its collapse. Many of those communities exist to this day and to some extent, have remained genetically isolated from neighboring populations. Brazilian *quilombos* and Colombian *palenques* and *comunidades negras* are examples of such. Here, we aimed to contribute to the body of knowledge about quilombos with information on their ancestry and genetic structure, thus improving the representation of rural populations in databases of forensic and anthropological interest. Additionally, we aimed at contrasting our study populations to one another, to their presumptive parentals and to other populations and assess whether historical parallels are reflected in genetic ancestry. Two rural quilombos: Kalunga (Goiás) and Riacho de Sacutiaba e Sacutiaba (Bahia) were analysed for the AIM systems *for*InDel, PIMA and SNP*for*ID 34plex, as well as the combined PIMA + 34plex. The Brazilian quilombo Mocambo (Sergipe) and the Colombian *comunidad negra* Mulaló were included when data was available. We assumed a tri-hibrid admixture model based on historical plausibility and selected parental populations for ancestry analysis from HGDP-CEPH following that main criterion and the amount of variation observed within and between groups. Further, we present a new AIM system (PIMA: Population Informative Multiplex for the Americas), that has proven efficient in differentiating the indigenous American ancestry component from that of other continental populations, namely African, European and East Asian. Our assessment

of genetic ancestry and structure has confirmed that the main admixture components in the *quilombos* and the *comunidad negra* are African and European. Regardless of that similarity that likely stems from parallels in history and origin, our results also state that our study populations are not homogeneous. Beyond attesting the diversity among human populations from Latin America, our observations reiterate the need to describe non urban and non indigenous populations in order to build databases adequate for ancestry analysis.

Keywords: Genetic Ancestry, AIM, Quilombo, Afro-derived, BGA

1. Introdução geral

A genética de populações humanas aplicada à antropologia biológica busca entender a variação genética e a diferenciação entre populações humanas a partir de uma perspectiva evolutiva. Seus princípios são fundamentais para que se tenha uma visão ampla da história da espécie (Hartl e Clark, 2010). No contexto atual, além do interesse antropológico, a genética de populações contribui de forma relevante para áreas tão diversas quanto genômica, biologia evolutiva, medicina e ciências forenses. Sua análise tem como base a comparação de frequências genotípicas de marcadores genéticos entre populações e dentro delas para acessar diferenças e similaridades e, a partir delas, inferir história, parentesco, relações evolutivas e eventos demográficos.

Do ponto de vista da genética de populações, a dispersão humana a partir da África pode ser descrita por uma combinação de diferentes modelos demográficos (Pickrell e Reich, 2014). Tais modelos incluem eventos de perda de variabilidade genética, como eventos fundadores sucessivos (Wang et al., 2007) e substituição populacional, que podem obscurecer a história das populações originais e a relação ancestralidade-origem geográfica; e eventos que introduzem variabilidade por miscigenação e podem reduzir a diferenciação entre populações (Pickrell e Reich, 2014). A variabilidade genética observada em qualquer espécie (ou população) é um balanço entre o surgimento de novidade por mutação (ou fluxo gênico) e sua exclusão, seja por fatores estocásticos ou por seleção natural. A teoria neutra da evolução, proposta por Kimura (1968), dita que uma “fração apreciável” da variação

molecular observada entre espécies é neutra: a diferenciação entre elas ocorreria principalmente de forma aleatória, por deriva genética. No entanto, não há consenso sobre que proporção do genoma humano é neutro como definido por Kimura (Zhang, 2018), especialmente porque a história humana é marcada pela ocupação de ambientes diferentes e exposição a pressões seletivas novas.

Apesar de a história de evolução da nossa espécie ser recente, o isolamento por distância e por fatores culturais, que fazem com que não haja de fato cruzamentos aleatórios na espécie humana, geraram, e ainda geram, diferenças mensuráveis entre populações. Em comparação com outras espécies, essa estruturação é pequena, mas se reflete em diferenciais de frequência gênica e genotípica suficientes para a identificação da origem geográfica de indivíduos e dos fundadores de uma dada população (Phillips, 2015). Um fator que dificulta essa avaliação é a natureza clinal da distribuição da variação genética neutra. Porque pessoas normalmente se relacionam dentro de suas populações e grupos sociais ou com pessoas de populações geograficamente próximas, ela só é brusca em situações específicas (onde existem barreiras geográficas, por exemplo). Além disso, com o aumento da densidade populacional e da migração em longas distâncias e em grandes contingentes nos últimos 500 anos, populações com composições genéticas similares e intermediárias (*i.e.* miscigenadas) às das populações fonte de migração têm se tornando mais comuns. Mais recentemente, crises migratórias geradas por guerras, questões econômicas e políticas e conflitos religiosos devem estar também acentuando esse processo de homogeneização.

Desde as primeiras tentativas de mensurar a variação genética humana, observou-se que a maior parte da variação é intrapopulacional. Lewontin (1972), partindo de um número pequeno de marcadores clássicos, concluiu que cerca de 85% da variação genética é observada dentro de populações humanas e que apenas ~10% se observa entre grupos humanos – cerca de 5% seria observada em diferenças intrapopulacionais decorrentes de subestruturação. Estudos posteriores com maior poder laboratorial e estatístico chegaram a conclusões parecidas, mas obtiveram maiores estimativas de diferenciação interpopulacionais. Rosenberg e colaboradores (2002), analisando 377 STRs em amostras do HGDP-CEPH, concluíram que 93.2-94.1% da variação humana é intrapopulacional e 4.3-3.6% interpopulacional, sendo que 2.5-2.4% seriam derivados de subestruturação populacional.

1.1 Análise de ancestralidade em populações humanas

A inferência de ancestralidade populacional, *i.e.* a avaliação da representação quantitativa de variantes de diferentes origens no genoma dos indivíduos que compõem uma população (Shriver e Kittles, 2004) e da contribuição de grupos parentais na formação de uma população, é um dos alvos da genética de populações e também uma de suas ferramentas. Ela é ferramenta para 1. o entendimento da história e do relacionamento de grupos humanos (Wang et al., 2008; Tang et al., 2007; Martinez et al., 2007; González-Andrade et al., 2007 e Martinez-Marignac et al., 2004); 2. para análises de estruturação populacional como base para estudos de mapeamento gênico por desequilíbrio de ligação (Shriver et

al., 1997); 3. para a definição de amostras controle em estudos de associação (Shriver et al., 1997; Pritchard e Donnelly, 2001); 4. para a genética forense, como forma de restringir o rol de suspeitos em uma investigação (Jobling e Gill, 2004).

A compreensão da história evolutiva de populações humanas é complicada por diferentes fatores: as respostas obtidas com a análise de marcadores genéticos de diferentes naturezas – como clássicos, moleculares e uniparentais (Lum *et al*., 1998; Yang *et al*., 2010) nem sempre geram respostas similares (são portanto complementares, não redundantes). Além disso, as relações entre populações humanas não seguem padrões rígidos e, consequentemente, não geram padrões filogenéticos simples, baseados apenas em distâncias geográficas. Isso resulta de eventos demográficos como a migração, que tende a homogeneizar populações (Nei e Roychoudhury, 1982), eventos de fusão e fissão (Neel e Weiss, 1975), migração e alterações drásticas do tamanho populacional (gargalos populacionais, expansões rápidas e substituições populacionais) causadas, dentre outro fatores, por guerras, catástrofes e inovações tecnológicas. Eventos como estes adicionam complexidade aos processos demográficos que geram a estrutura genética observada em qualquer momento da história de uma população e tornam menos clara a história evolutiva (Thompson, 1979).

Voltando aos trabalhos de Rosenberg e colaboradores (2002 e 2005), a análise de estruturação utilizando o método implementado no *software* Structure (Pritchard et al., 2000) e obteve k ótimo:5, correspondente aos grupos continentais África Subsaariana, Eurásia, Leste da Ásia, América e Oceania. Considerando k:7, Europa, Oriente Médio e Centro-Sul da Ásia foram identificados dentro do grupo

Eurásia. Posteriormente, Li e colaboradores (2008) analisaram 650 mil SNPs na mesma amostra e confirmaram a distribuição da variação genética e os agrupamentos previamente observados. Os dois trabalhos, no entanto, mostraram que a diferenciação de Oriente Médio e Centro-Sul da Ásia do grande grupo Eurasiático não é bem definida. Posteriormente, Rosenberg e colaboradores (2005) expandiram sua análise para 933 STRs e refutaram críticas à amostragem do HGDP-CEPH, que não seria capaz de explicitar a distribuição clinal da diversidade genética humana. Esses autores mostraram que utilizando uma grande quantidade de marcadores, o gradiente genético é correlacionado à distância geográfica e que os agrupamentos previamente observados são válidos e mantidos por pequenas descontinuidades.

Em suma, tais observações apontam 1. a existência de cinco agrupamentos continentais detectáveis por marcadores informativos de ancestralidade dos tipos STR e SNP; 2. que apesar de clinal, a distribuição da variabilidade genética humana apresenta descontinuidades detectáveis e mensuráveis por análises de ancestralidade e estruturação; 3. que o caráter clinal e a ocorrência de processos demográficos que tendem a homogeneizar populações humanas dificulta a análise em níveis mais refinados geograficamente. Assim, qualquer análise geral dos padrões de distribuição de variabilidade e ancestralidade em populações humanas será uma simplificação da realidade.

1.1.1 Marcadores genéticos na análise de ancestralidade

Análises de genética de populações se baseiam em marcadores genéticos de diversas categorias moleculares (classe de mutação) e com diferentes padrões de herança e taxas evolutivas. Marcadores uniparentais estão mais fortemente correlacionados à origem geográfica continental – ou ainda mais específica – da linhagem materna ou paterna de um indivíduo (Phillips, 2015) devido, entre outros fatores, ao padrão de herança do mtDNA e do cromossomo Y e à distribuição geral da variação existente nessas moléculas. Apesar disso, eles não acessam toda a ascendência de um indivíduo (Shriver e Kittles, 2004) nem toda a história de formação de uma população (Bedoya *et al*., 2006).

A análise de grandes conjuntos de marcadores autossômicos ganhou espaço na inferência de ancestralidade ao longo dos últimos anos graças ao barateamento e simplificação das técnicas laboratoriais, ao estabelecimento de grandes bases de dados públicas e ao desenvolvimento de metodologias estatísticas e computacionais sofisticadas. Via de regra, eles não permitem associações tão diretas entre genótipo e origem geográfica, mas informam perfis de miscigenação mais complexos que os marcadores uniparentais, além de requererem bancos de referência menores (Phillips, 2015). Além disso, esses marcadores permitem a quantificação da ancestralidade individual e delineamento de eventos demográficos mais recentes. Metodologias estatísticas para a estimativa de ancestralidade a partir de marcadores autossômicos dependem do conhecimento de genótipos multiloci e frequências

alélicas (Rosenberg et al., 2003) nas populações ditas parentais e naquelas às quais a amostra em análise pertence.

A acurácia das estimativas de ancestralidade a partir do genoma autossômico está diretamente relacionada à quantidade de marcadores analisados: quanto maior o número, mais regiões do genoma são acessadas (Shriver e Kittles, 2004). No entanto, o esforço de genotipagem e computação requeridos pela análise de grandes números de marcadores faz com que seja preferível analisar conjuntos pequenos de marcadores com alto conteúdo informativo (Rosenberg et al. 2003). Determinadas métricas são comumente utilizadas na seleção de marcadores que comporão painéis para ancestralidade. As mais usuais são δ (Shriver et al., 1997; Frudakis et al., 2003); F_{ST} e *In* – ou conteúdo de informatividade para ancestralidade (Rosenberg et al., 2003; Chen et al. 2005).

O δ é uma medida do diferencial de frequências aléĺicas entre duas populações x e y, definida por $\rho_x - \rho_y$ para *loci* binários. Marcadores escolhidos tendo essa métrica como critério apresentam grande poder estatístico para diferenciar as populações x e y (Rosenberg et al., 2003). O índice de fixação F_{ST}, é uma medida de divergência diretamente relacionada à variância das frequências alélicas de um *locus* em uma população: quanto maior o F_{ST} entre duas populações maior a diferença observada entre elas (Holsinger e Weir, 2009). Assim, busca-se selecionar marcadores que mostram F_{ST} alto entre as populações ancestrais e baixos entre as populações que compõem uma parental. Nesse contexto, ele é um parâmetro utilizado como forma de avaliar a coesão de grupos continentais e a adequação deles como populações parentais em análises de ancestralidade. Tanto

o F_{ST} quanto o δ são medidas úteis na triagem inicial de marcadores candidatos, mas o *In* tem sido mais utilizado no contexto forense. Esse índice informa o conteúdo de informação introduzido por cada marcador de um conjunto na inferência de ancestralidade, *i.e.* a probabilidade atrelada a cada marcador de pertencer a uma população específica em comparação a uma população qualquer (Rosenberg et al,. 2003). O *In* é utilizado de forma ranqueada na escolha de marcadores para compor painéis para estimativa de ancestralidade. Assim como o F_{ST} e o δ, o *In* está correlacionado às frequências alélicas e, da mesma forma, alelos com frequências discrepantes ou fixados em populações diferentes geram valores mais altos.

Além da informatividade dos marcadores, a construção de um painel para a estimativa de ancestralidade deve levar em conta a disponibilidade e abrangência de dados em bancos públicos, o balanço da informatividade dos marcadores entre as populações parentais e sua localização genômica (Phillips, 2015). Conjuntos de marcadores com conteúdos informativos desbalanceados entre os pares de populações parentais , *i.e.* com maior conteúdo informativo para uma parental em detrimento das outras, podem gerar vieses nas estimativas de mistura, como exemplificado em Taboada-Echalar e colaboradores (2013). Esse desbalanço é uma das possíveis explicações para o porquê de diferentes painéis de ancestralidade informarem diferentes estimativas de contribuição parental.

Em relação à distribuição dos marcadores, Costas e colaboradores (2005) mostraram haver blocos de baixa diversidade ou baixa taxa de recombinação ao longo de todo o genoma humano. A maior parte dos métodos utilizados para inferir

história demográfica e padrões estruturais e de ancestralidade a partir da distribuição de polimorfismos parte da premissa da neutralidade dos marcadores empregados. Portanto, a seleção de marcadores deve levar em conta sua distribuição no genoma evitando a ligação entre *loci* que compõem um painel e o efeito haplotípico (Phillips 2015) decorrente de eventos evolutivos passados e do efeito carona (ou seleção por ligação), que podem enviesar inferências demográficas e de ancestralidade (Schrider e Kern, 2017).

1.2 A formação da população brasileira

A população brasileira foi formada por um intenso processo de miscigenação entre três grupos parentais principais: indígenas americanos, africanos subsaarianos e europeus, em sua maioria ibéricos. As populações nascentes passaram por diferentes histórias de contato e isolamento, expansão e retração, efeito do fundador, deriva e seleção que geraram a variabilidade hoje observada. Além disso, esses grupos não se dispersaram e contribuíram para o povoamento das cinco regiões brasileiras de forma homogênea. Há que se considerar também migrações mais recentes advindas de outras regiões, como o Oriente Médio, observável especialmente na presença libanesa no país, e Leste Asiático, refletida na grande presença japonesa. Além disso, ainda hoje vivem no Brasil populações isoladas e semi-isoladas, como populações indígenas e quilombos dispersos pelo país.

1.2.1 A população indígena brasileira

A população indígena brasileira à época do início da colonização foi estimada em 3 - 10 milhões de indivíduos (Adhikari et al. 2017; FUNAI, 2019 - http://www.funai.gov.br), em sua maioria Tupi-Guaranis, povos que compartilham semelhanças culturais e linguísticas (Fausto, 2001). Diversos estudos apontam que o povoamento das Américas teve origem na Beríngia entre 12 e 18 mil anos atrás, ainda que não haja consenso (Wang et al., 2007; Rolando-José et al., 2008; Reich et al., 2012; Potter et al., 2015; Llamas et al., 2016; Pinotti et al., 2019, para citar alguns). Lá, a população ancestral beringiana, fundada ~23 mil anos atrás (Moreno-Mayar et al., 2018), provavelmente esteve isolada de seus ancestrais asiáticos por pelo menos 5 mil anos (Rolando-José et al., 2008; Raghavan et al., 2015; Gómez-Carballa et al., 2018). No continente, as populações nascentes passaram por diferentes eventos de contato e isolamento, expansão e retração, fundação, deriva e seleção que geraram a complexidade observada em populações contemporâneas indígenas e miscigenadas.

O contato com os portugueses, em alguns casos pacífico e em outros hostil, levou a uma redução drástica da população indígena estimada atualmente em 817 mil pessoas (cerca de 0,4% da população brasileira) vivendo, a maioria, nas 680 comunidades localizadas em terras indígenas, e muitos em áreas urbanas e rurais fora das reservas (FUNAI, 2019 - http://www.funai.gov.br). Muitos dos grupos étnicos originais já não existem: foram extintos ou incorporados à população miscigenada nascente, mas sua contribuição para o *pool* genético brasileiro consequente da

mistura com africanos e europeus está presente em todo o país (Manta et al., 2013; de Moura et al., 2015).

1.2.2 O Europeu no Brasil

O fluxo de europeus para o Brasil se iniciou em 1492, com a chegada das primeiras caravelas à costa da Bahia. Portugal foi sempre a maior fonte de migração, apesar de em meados do século XIX ter se iniciado, com incentivo do governo brasileiro, um intenso fluxo migratório de toda a Europa com o objetivo de substituir a mão de obra escrava que tinha, então, se tornado ilegal, e associado também à política de branqueamento. Esse fluxo, com origem principalmente italiana, se intensificou e manteve até o início do século XX e teve como destino principal as regiões sul e sudeste (Fausto, 2002).

1.2.3 O Africano no Brasil

Estima-se que 40% de todos os africanos trazidos para as Américas como escravos (3,6 a 10 milhões de indivíduos) tiveram portos brasileiros como destino. O registro histórico sobre o período de comércio e tráfico de escravos é bastante incompleto, em parte pela imprecisão e ausência de parâmetros dos documentos da época, mas também por que em 1850, quando o comércio se tornou ilegal, os portos de saída na África e entrada no Brasil deixaram de ser registrados e passaram a ser constantemente mudados para escapar à fiscalização. De acordo com os registros existentes, a maioria dos africanos escravizados trazidos pelos portugueses eram Bantos e Sudaneses (povos subsaarianos próximos ao deserto,

não necessariamente do Sudão) capturados tanto na costa leste como na oeste da África subsaariana (Salzano e Freire-Maia, 1967). Ao longo dos séculos em que africanos escravizados foram trazidos para as Américas, eles foram capturados e embarcados em diferentes regiões (Melo e Souza, 2006): no início do século XVI, sua principal origem eram portos no Congo e, em menor quantidade, Angola, ambos na costa oeste (Schwartz, 1998; Mello e Souza, 2006). Com o crescimento da influência e dominação holandesa, os portugueses estabeleceram portos e comércio em regiões mais ao sul, em especial Luanda e Benguela, ambos na atual Angola. Durante o século XVIII, o golfo da Guiné, principalmente a baía de Benin e a Costa da Mina, ganhou importância no comércio de escravos, mas durante o século XIX, cerca de um quarto de todos os africanos escravizados trazidos para o Brasil vinham de Angola (Schwartz, 1988).

A abertura e consequente intensificação do comércio atlântico de escravos foram sempre influenciados pela política externa, como mencionado acima, mas também pela política interna no continente africano. Mudanças no cenário interno dependiam grandemente da participação dos portugueses, que faziam alianças com chefes locais de reinos grandes e pequenos (Ferreira, 2006). Assim, diversos povos africanos atuaram como pumbeiros – responsáveis pela captura de escravos – apesar de muitas vezes estarem também sujeitos a serem escravizados. Dentre os muitos grupos que participaram desse processo, estão os Jaga, Ambundo, Imbangala, Ovibundo e Iorubá (Mello e Souza, 2006).

No início do século XVIII, como consequência da intensificação do comércio e de mudanças no cenário político, os escravos eram capturados em regiões mais

internas, próximas aos Grandes Lagos. Os grupos étnicos predominantemente escravizados variaram ao longo do tempo em decorrência da situação política, que definia os lugares e reinos onde se buscavam escravos, e das alianças formadas, que garantiam acesso a certas regiões e controle sobre portos. Congo e Angola, por exemplo, comercializavam principalmente Bantos e Benguela, grupos de diferentes origens (Melo e Souza, 2006). Essas regiões, que são hoje referência no estudo da interação entre europeus e africanos no período colonial, são ainda povoados predominantemente, ainda que não exclusivamente, por povos Bantos. Acredita-se que a maior parte dos africanos trazidos como escravos para as Américas tenha saído de portos nessa região e devem, portanto, ter pertencido a esses povos (Miller, 1987). Os cativos aí embarcados eram em sua maioria dos reinos de Cabinda, Dongo, Congo, Daomé, Kazembe, Luba, Lunda e Lozi (Mello e Souza, 2006).

De acordo com Nina Rodrigues (2004), por que estudos sobre o comércio negreiro para o Brasil abordam prioritariamente o contexto do Porto do Rio de Janeiro – e seu mercado Vacongo – como referência, os grupos Bantos são frequentemente considerados os únicos – ou quase – trazidos para o país. Acordos bilaterais entre os estados brasileiros e portos africanos e o relacionamento entre nações africanas influenciavam diretamente a oferta e origem de pessoas escravizadas. Portanto, determinados grupos étnicos eram mais comuns em determinadas regiões brasileiras. Outro fator que gerou complexidade foi o comércio e migração internos no Brasil: a Bahia, por exemplo, recebia principalmente sudaneses e os vendia para o sul do estado, para Minas Gerais e Goiás (Schwartz,

1998), onde, em consequência, grupos sudaneses estiveram presentes em grande número. Ainda assim, por que os Bantos eram, e ainda são, maioria numérica na África, esses povos também contribuíram com um grande número de indivíduos para as populações da região nordeste e, consequentemente, para seu *pool* gênico.

1.2.4 A presença de outros grupos no Brasil

A partir de 1908 iniciou-se o fluxo migratório advindo do Japão que, apesar de não ter sido tão expressivo quanto o Europeu, resultou na marcante presença desse grupo na população brasileira atual. Atualmente, estima-se que cerca de 800 mil pessoas se identificam como tendo ascendência japonesa apenas no estado de São Paulo (IBGE, 2012).

A migração síria e libanesa para o Brasil, impulsionada por razões políticas, religiosas e econômicas, se iniciou em 1880 e se intensificou na segunda metade do século XIX. Hoje, estima-se que vivam no país cerca de 900 mil pessoas de origem no Oriente Médio, em sua maioria, no Líbano (IBGE, 2010). Existem, no entanto, estimativas de que vivam no país entre 7 e 10 milhões de pessoas de ascendência libanesa (Itamaraty, 2015).

1.2.5 A composição genética da população brasileira

A complexidade da história de povoamento do continente americano e do país se reflete na composição genética da população brasileira, que mostra diferentes proporções dos grupos parentais comumente avaliados nas diferentes regiões do país. A análise de ancestralidade e genética de populações no Brasil foi

inicialmente dirigida a populações ameríndias e, mais tarde, se voltou para o estudo de populações urbanas do nordeste (Toledo, 2016). Hoje, inúmeros trabalhos com marcadores moleculares uniparentais e autossômicos de diferentes categorias moleculares foram publicados (exemplificados abaixo) e mostram um quadro geral coerente com a história conhecida do país, mas incompleto, especialmente em relação a populações não urbanas (Manta et al., 2013a; Adhikari et al., 2017; Gontijo et al., 2018).

Em análises de mistura, os resultados obtidos podem variar conforme as classes de marcadores utilizados devido a suas características moleculares e comportamento evolutivo. Esses resultados se complementam em uma história mais abrangente da população analisada e refletem também eventos demográficos. No Brasil, o estudo de marcadores polimórficos clássicos (grupos sanguíneos e proteínas séricas e eritrocitárias) indicou que a contribuição ameríndia nas populações miscigenadas varia de 55% na população de Alenquer, no Norte do país (Santos e Guerreiro, 1995), a 5% em Santa Catarina, no Sul (Dornelles et al., 1999).

O mesmo padrão foi observado para marcadores do tipo indel. Populações urbanas do norte do país têm componentes de mistura ameríndia (Belém - PA: 29,5% e Manaus - AM: 37,8%; Pereira et al., 2012 e Manta et al., 2013a) mais altos que populações de outras regiões. Nas outras regiões, Manta e colaboradores (2013a) encontraram estimativas variando de 8,9% no Paraná a 18,7% em Alagoas. Com relação à contribuição europeia, variação de 45,9% em Belém (Pereira et al., 2012) a 79,7% em Santa Catarina (Manta et al., 2013a). Já a contribuição africana variou de 11,4% no Espírito Santo (Manta et al., 2013a) a 31,1% no Rio de Janeiro

(Manta et al., 2013b). As cidades do norte Belém e Manaus foram exceção ao quadro geral porque nelas, o componente de mistura indígena americano foi mais alto que o africano, enquanto nas outras regiões do país, o contrário é observado.

Para a população de Belém e da região Sul, um conjunto diferente de Indels informativos de ancestralidade mostrou resultados qualitativamente semelhantes (Santos et al., 2010). Pena e colaboradores (2011) avaliaram 40 indels informativos de ancestralidade e encontraram um delineamento geral similar no que se refere à distribuição do componente indígena americano, mais significativo no norte do país para a região norte, e do componente europeu, mais significativo no sul. Já o componente de mistura africano foi maior no Nordeste (30,3%) que no Sudeste (18,9%). Essas são as duas regiões que receberam o maior contingente de escravos africanos, portanto não é surpreendente que a maior contribuição dessa parental seja observada aí.

A discrepância entre Pena e colaboradores (2011) e os estudos citados anteriormente pode ser decorrente da estratégia de amostragem, visto que a amostra do Sudeste foi coletada de estudantes e trabalhadores de uma instituição de pesquisa - o que poderia enviesar as contribuições europeia e africana. Kehdy e colaboradores (2015) avaliaram 2,2 milhões de SNPs em uma população do nordeste (Salvador), uma do Sudeste (Bambuí) e uma do sul (Pelotas). Suas estimativas mostraram resultado concordante com Pena e colaboradores (2011) no que se refere à maior ancestralidade africana ser observada no estado da Bahia

(50,8%). Já em relação à contribuição europeia, a estimativa mais alta foi observada em Bambuí - MG (78,5%).

Lins e colaboradores (2010) avaliaram 28 SNPs autossômicos em populações urbanas das cinco regiões brasileiras em amostras de testes de paternidade. Esse conjunto de marcadores revelou alta contribuição europeia (variando de 69,5% no Centro-Oeste a 87,7% no Sul), contribuição africana intermediária (7% no Sul a 18,7% no Centro-Oeste) e ameríndia mais baixa (5,2% no Sul a 11,8% no Centro-Oeste, seguido por 10,7% no Norte). O quadro geral de alta mistura europeia em todas as regiões se confirmou, mas discrepâncias na contribuição das outras duas parentais são observadas entre as estimativas de Lins e colaboradores e as citadas nos parágrafos anteriores (derivadas da análise de indels).

Porque SNPs indicativos de ancestralidade, assim como indels indicativos de ancestralidade, são marcadores adequados para a análise de mistura, essa discrepância pode se dever à estratégia de amostragem utilizada, já que há diferenças de acesso a serviços como exames de paternidade entre classes sociais que, no Brasil, estão correlacionadas a raça e/ou ancestralidade (Dalton, 2010). As análises com SNPs autossômicos de Silva e colaboradores (2010), Queiroz e colaboradores (2013), Leite (2011) e Giolo (2013) também mostraram contribuição europeia maior, seguida da africana e da ameríndia, nessa ordem, em Minas Gerais, Distrito Federal e São Paulo. Ruiz-Linares e colaboradores (2014), em um apanhado da América Latina, utilizaram 30 SNPs indicativos de ancestralidade para estimar contribuições parentais. Na população brasileira, esse conjunto de marcadores

revelou um panorama com alta contribuição europeia em todo o país, especialmente alta sul, africana mais alta no Nordeste e Sudeste, e indígena, mais alta no norte.

Estimativas baseadas em marcadores STR autossômicos mostram composições mais homogêneas entre diferentes regiões do país, mas ainda mantêm o quadro global: contribuição indígena maior no Norte (18%) que no Sul (7%) do país; e contribuições europeia e africana similares. O trabalho de Callegari-Jacques e colaboradores (2003) com marcadores do tipo STR, por exemplo, revelou contribuição europeia mais alta na região Sul (81%) e indígena no Norte (18%). Godinho e colaboradores (2008) mostraram resultados similares com dados dos sistema CODIS. Também confirmando o quadro geral, os trabalhos de Ferreira e colaboradores (2006) e Martins e colaboradores (2011) mostraram resultados muito concordantes para São Paulo (EUR 76-79%; AFR 14-18%; AME 6-7%). Francez e colaboradores (2011) encontraram para a região norte (Amapá) estimativas de ancestralidade concordantes com as obtidas pela análise de indels mencionadas acima, com a parental indígena maior que a africana (EUR 46%; AFR 19%; AME 35%).

Marcadores uniparentais mostraram que o processo de miscigenação teve um viés entre as populações que contribuiram para a formação das populações miscigenadas do Brasil. Essencialmente, as patrilinhagens foram europeias e as matrilinhagens, mais diversas (Alves-Silva et al., 2001; Carvalho-Silva et al., 2001; Abé-Sandes et al., 2004; Gratapaglia et al., 2005; Barcelos, 2006; Resque et al., 2016). Marcadores do tipo STR do cromossomo Y indicam alta contribuição masculina europeia e baixa variabilidade entre as regiões brasileiras (Carvalho-Silva

et al., 2006; Pena et al., 2011; Palha et al., 2012). No entanto, devido a sua alta taxa de mutação, essa categoria de marcadores não é adequada para a detecção de diferenciação populacional.

Já haplótipos de SNPs do cromossomo Y têm taxas de mutação menores e são, por isso, mais adequados para a detecção de diferenciação populacional. O trabalho publicado por Resque e colaboradores (2016) mostrou diferenciação entre as populações urbanas das cinco regiões do país em termos de variabilidade e ancestralidade. A região Norte, como esperado, é a que apresenta a maior contribuição masculina da parental indígena (8,4%). Por sua vez, as regiões Sul e Centro-Oeste apresentam maior contribuição europeia, e região Nordeste, apresenta a maior contribuição africana (15,1%). Linhagens de mtDNA, por sua vez, consistentemente indicam maior contribuição indígena e africana que as linhagens paternas em todo o Brasil.

A ancestralidade africana e europeia vêm sendo refinadas com o desenvolvimento de novas metodologias laboratoriais e estatísticas. Kehdy e colaboradores (2015) observaram dois grandes grupos africanos com distribuição diferencial no nordeste (Salvador - BA), sudeste (Bambuí - MG) e sul (Pelotas - RS). Em Salvador, a ancestralidade africana é associada majoritariamente (75%) a populações bantu do leste da África, enquanto em Bambuí e Pelotas, esse grupo corresponde a 36% e 44% da contribuição africana, respectivamente. Nessas duas cidades, a parental africana está mais fortemente relacionada a grupos não bantu do oeste africano (Iorubá e Mandenka). O mesmo trabalho observou um componente

do norte da Europa mais alto em Pelotas - RS (40,2%) que nas outras duas populações (35,8-36,7%).

O trabalho de Resque e colaboradores (2016) dissecou a ancestralidade paterna europeia pela análise de SNPs situados no cromossomo Y. Eles observaram contribuição portuguesa, francesa, italiana, alemã e libanesa como parte da parental europeia. A proporção de cromossomos portugueses variou de 18% no Nordeste a 63% no sul; a francesa, de 0% no Sul e sudeste a 52% no Norte; a italiana, de 1% no Norte a 61% no Nordeste; a alemã, de 0% no Norte a 17% na alemanha e a libanesa, de 0% no Nordeste a 23% no Centro-Oeste.

Outro exemplo é o trabalho de Hünemeier e colaboradores (2007), em que mtDNA e cromossomo Y foram analisados para determinar a origem subcontinental das linhagens maternas e paternas de indivíduos classificados como negros em Porto Alegre (um dos estados que receberam menos africanos escravizados) e do Rio de Janeiro (estado que recebeu um dos maiores contingentes). Seus resultados indicaram que 82% das linhagens maternas no Rio de Janeiro e 69% em Porto Alegre são de origem no Centro-Oeste/Sudeste africano. Já a análise de haplogrupos do cromossomo Y não pôde informar a origem dentro do continente das linhagens paternas que compuseram a amostra.

1.3 Quilombos brasileiros

1.3.1 Conceituação e história

O número de africanos escravizados trazidos para o Brasil foi muito expressivo. De acordo com o censo de 1849, a população de origem africana (incluindo libertos e cativos) correspondia a 43,5% da população do Rio de Janeiro (Karasch, 2000). Durante esse século, quando o fluxo de africanos para o continente se intensificou, a média anual de escravos que desembarcaram no país girou entre 6 e 9 mil (Alencastro, 2008). Entre os anos de 1838 e 1839, esse número chegou a mais de 40 mil (Alencastro, 2008), sendo que destes, 70 a 90% entraram no país pelo Rio de Janeiro (Schwartz, 1998).

Nas Américas, os escravos de origem africana resistiram ao cativeiro de diversas formas, entre elas o suicídio, agressão aos senhores e a fuga e concentração em locais escondidos ou de difícil acesso. Essa última prática originou comunidades isoladas geográfica e culturalmente, embora em graus variáveis, denominadas no Brasil quilombos ou mocambos (Vila Real, 1996; Neme & Andrade, 1987). Em todo o continente foram formadas comunidades miscigenadas relacionadas à presença de africanos escravizados e análogas aos quilombos brasileiros em história de formação e significado na resistência à opressão (Gomes, 2015). Dentre elas, estão os *palenques* e *comunidades negras* colombianos, *cumbes* venezuelanos e *cimarronaje* cubanos e porto riquenhos (Paiva, 2017). No Brasil, muitas dessas comunidades foram formadas como espaços de liberdade e resistência à escravidão antes do ano de 1888, quando a Lei Áurea, promulgada e

sancionada em 13 de maio de 1988 (Lei Imperial n.º 3.353), aboliu a escravatura no país. A lei, no entanto, não instituiu ou previu mecanismos de distribuição de terras e integração social. Por isso, o quilombo continuou a única forma que muitos encontravam para viver em liberdade plena, garantir meios de subsistência, fazer de fato parte de uma comunidade e constituir família, e muitos outros quilombos foram formados após a abolição.

O entendimento do termo quilombo mudou muito desde suas primeiras menções em textos oficiais dos séculos XVIII e XIX, quando aparece em registros e relatórios policiais o medo do “calhambola”. Em 1740, o Conselho Ultramarino, que ditava regras para a colônia brasileira, definiu pela primeira vez em lei o termo quilombo “… como toda habitação de negros fugidos que passem de cinco, em parte desprovida, ainda que não tenham ranchos levantados nem se achem pilões neles.” (Leite, 2009). Posteriormente, a Lei nº 236, de 20 de agosto de 1847, da província do Maranhão, define um conceito para quilombola: “Reputar-se-á escravo quilombado, logo que esteja no interior das matas, vizinho ou distante de qualquer estabelecimento, em reunião de dois ou mais com casa ou rancho.” (Goulart, 1970). Em 1848, a definição legal de quilombo passa a ser “… a reunião em lugar oculto de mais de dois escravos.” (Goulart, 1972).

Esse entendimento do século XVIII de quilombos como comunidades formadas por escravos fugidos e refugiados em locais de difícil acesso é ainda hoje a percepção comum que se tem. A definição atual das chamadas comunidades Remanescentes de Quilombos é "grupos étnico-raciais, segundo critérios de autoatribuição, com trajetória histórica própria, dotados de relações territoriais

específicas, com presunção de ancestralidade negra relacionada com a resistência à opressão histórica sofrida" (Decreto 4.887, artigo 2º, de 20 de novembro de 2003). De acordo com a Associação Brasileira de Antropologia (2012) e O'Dwyer (1995), o termo quilombo não se refere a entidades sociais fixas ou contínuas em relação aos antigos quilombos e tampouco a resquícios arqueológicos ou continuidade material. Os quilombos não são homogêneos, nem necessariamente isolados ou constituídos a partir de movimentos de insurreição.

Existem hoje registros e vestígios da presença de quilombos urbanos e rurais, com diferentes graus de isolamento, em todas as regiões do Brasil (FCP, 2018). O Estado brasileiro reconhece oficialmente a existência de mais de 3000 quilombos (FCP, 2018) habitados por cerca de dois milhões de pessoas (SEPPIR, 2018). Apesar de serem de forma geral considerados comunidades isoladas, os quilombos sempre mantiveram relações comerciais e sociais com populações vizinhas, resultando em miscigenação desde suas fundações (Amorim et al., 2011). Eles possuem histórias únicas, mas formam um grupo coeso em sua ancestralidade africana marcada, cultura e, historicamente, em sua história de luta contra a opressão e pela terra e meios de subsistência de forma livre.

1.3.2 Composição genética de quilombos brasileiros

Muitos trabalhos prévios sobre a genética e a ancestralidade de Quilombos analisaram conjuntos pequenos de marcadores autossômicos, marcadores clássicos e marcadores uniparentais (Salzano e Sans, 2014; Gontijo et al., 2014; Kimura et al., 2016). Essas análises, de forma geral, mostram resultados consistentes com a

história e demografia dos quilombos. A ancestralidade africana é alta, e proporções variáveis de mistura europeia e indígena - observadas comumente nessa ordem (Gontijo et al., 2014; Kimura et al., 2013; Lopes et al., 2011; Santos et al., 2009) -, apesar de haverem exceções a esse padrão (Lopes et al., 2011; Kimura et al., 2013).

Marcadores de linhagem indicam um viés de contribuição masculina/feminina acentuado contra linhagens mitocondriais europeias e a favor de linhagens indígenas (Lopes et al., 2011) e africanas (Kimura et al., 2016; Ribeiro et al., 2011; Abe-Sandes et al., 2004), enquanto as linhagens paternas são predominantemente europeias (Kimura et al., 2016; Kimura et al., 2013; Lopes et al., 2011; Ribeiro et al., 2011; Abe-Sandes et al., 2004). Esse viés é observado não apenas em quilombos, mas em populações de toda a América Latina (como revisado em Adhikari et al., 2017).

Os quilombos Kalunga - GO e Riacho de Sacutiaba e Sacutiaba - BA (daqui em diante referida como Sacutiaba), alvos do presente trabalho, vêm sendo estudados do ponto de vista da genética de populações há quase duas décadas. Marcadores clássicos, moleculares autossômicos e uniparentais foram analisados (Tabela 1, Capítulo 1). O padrão de ancestralidade apresentado acima os descreve adequadamente, e nenhum mtDNA de origem europeia foi encontrado nessas populações. No entanto, como os quilombos não são populações isoladas, é possível que a migração e os casamentos exogâmicos (como definido por Woortmann, 1990) venham a introduzir variação e alterem a diversidade genética nelas observada. Kalunga, por exemplo, que é o mais isolado dos dois, tem taxa de

migração baixa na geração entrevistada e casamentos majoritariamente endogâmicos (Amorim et al., 2011; padrão de casamento definido por Woortmann, 1990). No entanto, com a expansão de fronteiras agrícolas e exploração do turismo ecológico, a migração para áreas adjacentes ao quilombo vem aumentando e poderá impactar a composição genética da população.

2. OBJETIVOS

2.1 Objetivo geral

O objetivo geral deste trabalho é caracterizar a diversidade genética de quilombos brasileiros produzindo conhecimento sobre ancestralidade individual e populacional e estrutura genética de populações rurais de ancestralidade africana da América do Sul, tendo como modelo principal, os quilombos Kalunga (Goiás) e Sacutiaba (Bahia).

2.2 Objetivos específicos

No primeiro capítulo, buscamos estreitar a lacuna que existe no conhecimento sobre a história brasileira e melhorar a representatividade de populações rurais brasileiras em bancos de dados de marcadores de interesse forense. Examinamos os três quilombos rurais Kalunga, Sacutiaba e Mocambo e avaliamos a relação entre sua história conhecida e composição genética estimada a partir de dados do sistema *for*InDel, embasando nossas análises com conhecimento histórico e demográfico, além de dados previamente gerados para as mesmas populações.

No segundo capítulo, nosso objetivo foi acrescentar conhecimento sobre populações afro-derivadas da América do Sul, permitindo a construção de um panorama mais completo das populações do continente. Além disso, buscamos avaliar se os paralelos históricos e de origem compartilhados pelos quilombos

(tendo como modelo Kalunga e Sacutiaba) e a *comunidad negra* (tendo como modelo Mulaló) analisados se refletem em sua ancestralidade genética. O sistema de escolha para esse capítulo foi o SNP*for*ID 34plex.

No terceiro capítulo, apresentamos um novo sistema de marcadores (PIMA: Population Informative Multiplex for the Americas) desenhado com o objetivo de compor um sistema pequeno e informativo para diferenciar o componente de mistura indígena americano dos das parentais africana, européia e, em especial, leste asiática. Os 26 marcadores autossômicos foram selecionados para complementar de forma balanceada o sistema-cerne SNP*for*ID 34plex na detecção e mensuração do componente de mistura indígena americano em populações contemporâneas do continente. O sistema inclui ainda um SNP no cromossomo X e o marcador de sexagem amelogenina. Os dados gerados para esse sistema foram analisados em conjunto com os marcadores do sistema 34plex, compondo um conjunto de 52 marcadores com informatividade balanceada para as quatro parentais mencionadas. Além disso, para avaliar a performance do sistema em uma situação real de pesquisa e acrescentar conhecimento sobre a diversidade das populações humanas do continente americano, descrevemos a ancestralidade e estrutura populacional de 22 populações miscigenadas e indígenas, incluindo os quilombos Kalunga e Sacutiaba e a *comunidad negra* Mulaló.

Ao final, na seção Avaliação Geral das Estimativas de Ancestralidade, avaliamos a concordância entre os sistemas *for*InDel, SNP*for*ID 34plex, PIMA e

PIMA + 34plex comparando as distâncias estatísticas entre as estimativas individuais de mistura nas amostras de Kalunga e Sacutiaba.

3. METODOLOGIA GERAL

3.1 Aspectos éticos

A coleta e uso das amostras de *Kalunga* e Sacutiaba foi aprovada pela CONEP (CEP-FS/UnB 030/2002 e 151/07). Todos os doadores foram informados sobre os objetivos do projeto, o uso das amostras e a confidencialidade dos dados, cumprindo os requisitos exigidos pelas resoluções pertinentes.

3.2 Populações amostradas

As populações alvos do trabalho são quilombos no sentido tradicional da palavra. Estudos prévios da ancestralidade dessas comunidades revelaram um padrão tri-híbrido de mistura entre africanos, europeus e indígenas americanos, nessa ordem de contribuição. Quanto a marcadores de linhagem, o cromossomo Y é majoritariamente europeu e o mtDNA, africano. Nenhuma linhagem materna europeia foi encontrada nessas populações, resultado da reprodução diferencial entre suas parentais (revisado em Gontijo et al., 2014).

Kalunga

Localizado em uma extensa área rural dos municípios de Cavalcante, Monte Alegre e Teresina de Goiás (Figura 1), Kalunga é um dos maiores quilombos do

Brasil, com uma população estimada em 5300 pessoas. A comunidade foi formada por escravos levados para a região para trabalhar na mineração de ouro e que fugiram ou foram abandonados por bandeirantes no século XVII (Soares, 1995). A população é estruturada socialmente em diversas sub-comunidades com graus de isolamento variados tanto em relação umas às outras, quanto em relação ao centro urbano mais próximo, a cidade de Cavalcante. A estrutura demográfica de Kalunga foi descrita em Paiva (2017). O percentual de migrantes é baixo, perfazendo apenas 1,5% da amostra. A proporção de casamentos exogâmicos (como definido por Woortmann, 1990) na geração entrevistada é de 4,7%.

Riacho de Sacutiaba e Sacutiaba

Sacutiaba é uma comunidade localizada no município de Wanderley, no oeste baiano, distante 850 km de Salvador (Figura 1). A população é estimada em 200 indivíduos e forma uma única família organizada em núcleos baseados em relações de parentesco e afinidade. Acredita-se que seus fundadores foram escravos fugidos do norte de Minas Gerais que se instalaram na área do quilombo mais de 200 anos atrás. Em 2004, a Fundação Cultural Palmares (2009 - http://www.palmares.gov.br) reconheceu oficialmente o quilombo. A estrutura demográfica de Sacutiaba foi descrita em Amorim e colaboradores (2011). O percentual de migrantes é de 30% da amostra entrevistada. A proporção de casamentos exogâmicos na mesma geração é de 34,8%.

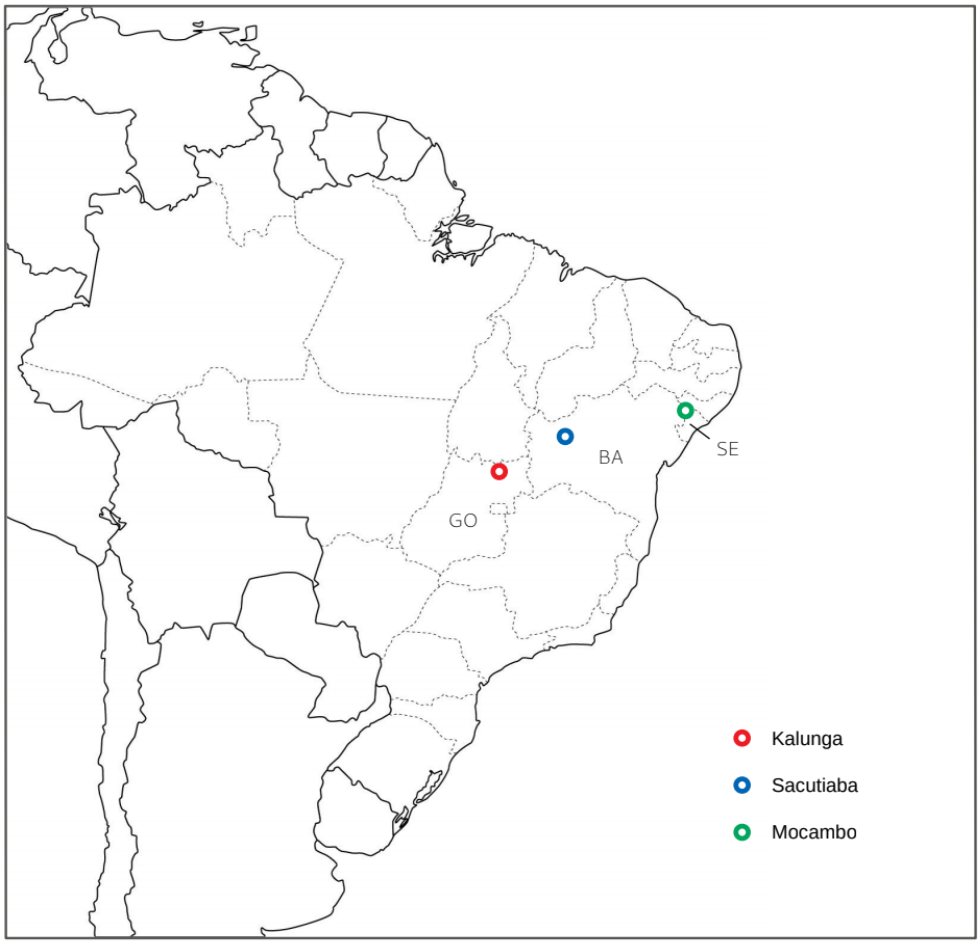


Figura 1. Mapa mostrando a localização dos dois quilombos em estudo: Kalunga – GO e Sacutiaba – BA.

Mocambo

Mocambo foi oficialmente reconhecida como quilombo no ano de 2004 (FCP - http://www.palmares.gov.br) e foi estimada, à época da coleta, em cerca de 500 habitantes. Apesar de a data de fundação do povoado ser incerta, em 1825 já havia um pequeno contingente de escravos habitando o local (no município de Porto da Folha, SE; Figura 1), hoje vizinho à área indígena de Xocó, com quem mantém relações históricas comerciais e de parentesco (Arruti, 2006). A estrutura

demográfica de Mocambo foi descrita em Amorim e colaboradores (2011). O percentual de migrantes é de 16% na amostra, e a taxa de migração foi estimada em 30%. A proporção de casamentos exogâmicos (como definido por Woortmann, 1990) na geração entrevistada é de 40%.

3.3 Análise laboratorial

3.3.1 Coleta e processamento de material biológico

Kalunga

Nos anos de 2001, 2008, 2015 e 2016 equipes do laboratório de Genética da Universidade de Brasília realizaram viagens a essa comunidade. Durante as visitas, foram coletados 5mL de sangue venoso utilizando o sistema de coleta a vácuo com tubos de coleta contendo EDTA como anticoagulante. As amostras foram processadas e armazenadas em freezer a -20ºC no Laboratório de Genética da UnB. Os cerca de 100 voluntários preencheram, no momento da coleta, um questionário que permite acessar informações referentes a duas gerações ascendentes e aos descendentes diretos e inclui dados de relevância demográfica e epidemiológica. 60 amostras foram genotipadas com sucesso nas análises do capítulo 1, 72 nas do capítulo 2 e 69 no capítulo 3.

Riacho de Sacutiaba e Sacutiaba e Mocambo

No ano de 1998, uma equipe do laboratório de Genética da Universidade de Brasília visitou as comunidades. Durante as visitas, foram coletados 5mL de sangue venoso pelo sistema de coleta a vácuo contendo EDTA como anticoagulante. Da mesma forma que as amostras coletadas em Kalunga, essas foram processadas e

armazenadas em freezer a -20ºC no Laboratório de Genética da UnB. Os 30 voluntários de Sacutiaba, no momento da coleta, um questionário que permite acessar informações referentes a duas gerações ascendentes e aos descendentes diretos e inclui dados de relevância demográfica e epidemiológica. Vinte e nove amostras de Sacutiaba foram genotipadas com sucesso nas análises dos três capítulos e 62 de Mocambo foram genotipadas com sucesso no Capítulo 1.

3.1.2 Extração de DNA e quantificação

50µL da fração leucocitária ou 100µL de sangue processado (frações de hemácias e leucócitos) das amostras coletadas foram utilizados para a extração de DNA utilizando o *kit* comercial *blood genomicPrep Mini Spin Kit* da *GE Healthcare*, de acordo com o protocolo do fabricante. O DNA extraído foi armazenado em microtubos devidamente etiquetados com códigos de forma a manter a privacidade sem, no entanto, dissociar as amostras dos dados de identificação individual dos doadores e mantidos em freezer a -20°C no Laboratório de Genética da Universidade de Brasília em um banco de amostras. As amostras foram quantificadas no equipamento NanoView e diluídas a 10 ng/µL em H_2O miliQ autoclavada.

3.1.3 Sistemas de marcadores selecionados

*Sistema for*InDel

Marcadores do tipo inserção-deleção (Indels) são normalmente bialélicos e caracterizados pela inserção ou pela deleção de um fragmento ou par de base.

Indels com mais alelos já foram identificados. Estima-se que representem 15-25% dos polimorfismos humanos (Weber et al., 2002; Mills et al., 2006). Mills e colaboradores (2006) os agrupam em: 1. inserções e deleções de par de base único; 2. expansões de monômeros; 3. expansões de motivos de 2-15 pb; 4. resultantes de eventos de transposição; e 5. sequências aparentemente aleatórias.

O sistema *for*InDel é formado por 46 indels autossômicos informativos de ancestralidade que diferenciam Europeus, Africanos, Leste-Asiáticos e Indígenas Americanos. Esses marcadores apresentam um bom equilíbrio de poder de discriminação entre esses grupos, tornando-o adequado para estudos de ancestralidade de populações brasileiras. Além disso, a inclusão de marcadores discriminatórios de grupos do leste asiáticos permite uma abordagem pouco usual para populações brasileiras, apesar de a contribuição desse grupo parental ser significativa no país.

A Tabela 1 mostra a contribuição de cada população parental (Africana, Europeia, Leste Asiática e Indígena Americana – HGDP-CEPH) estimada a partir do *for*InDel em cada uma delas e em populações de estudo. Como se pode observar, a contribuição estimada de outras populações em cada parental é muito baixa. O poder de discriminação desse sistema fica claro também na Figura 2, que mostra a análise de estruturação populacional das mesmas amostras para k:4-6.

Tabela 1. Contribuição média de cada população continental (Africana - AFR, Europeia - EUR, Leste Asiática – EAS e Indígena Americana - AME) observada em populações dos quatro continentes (África: Angola; Europa: Portugal; Leste da Ásia: Taiwan; Indígena Americano: a partir da análise do sistema *for*InDel (Pereira et al, 2012).

População	AFR	EUR	EAS	AME
HGDP-CEPH AFR	0.969	0.011	0.012	0.008
HGDP-CEPH EUR	0.008	0.963	0.014	0.014
HGDP-CEPH EAS	0.006	0.018	0.952	0.024
HGDP-CEPH AME	0.008	0.041	0.027	0.924
Angola*	0.970	0.011	0.011	0.008
Portugal**	0.018	0.966	0.008	0.008
Taiwan*	0.004	0.003	0.984	0.009
Amazônia (indígenas) *	0.01	0.013	0.032	0.945

* Pereira et al 2012

Metodologicamente, esse sistema tira proveito de características que a classe molecular indel compartilha com SNPs e STRs (Phillips, 2012). Com SNPs compartilha, além de uma maior estabilidade no genoma, o pequeno tamanho do produto amplificado que possibilita a amplificação a partir de DNA degradado. Com STRs, tem em comum a simplicidade metodológica de análise em que o produto amplificado é diretamente genotipado em sequenciador, excluindo a necessidade de passos intermediários.

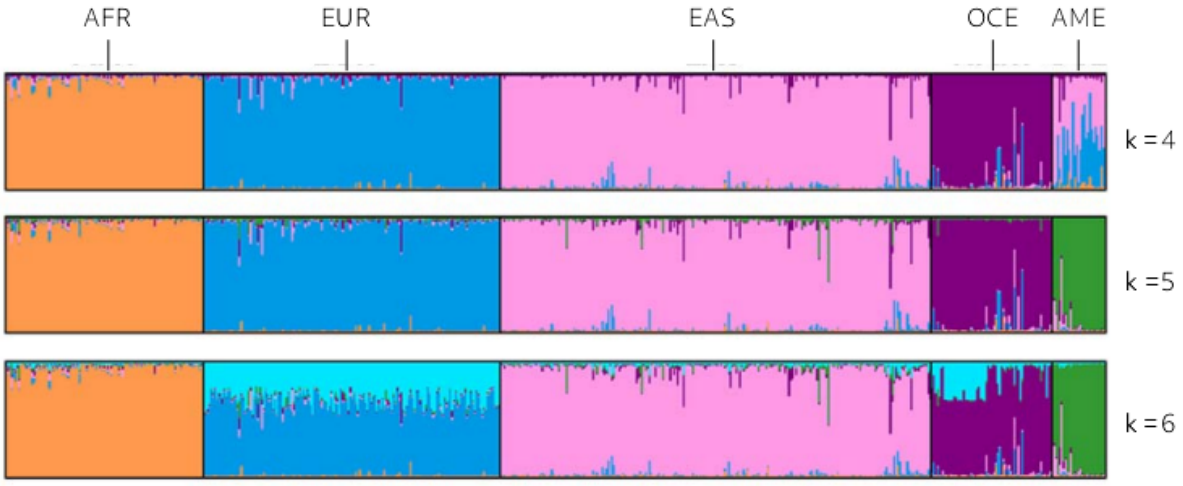


Figura 2. Análise de estruturação populacional de amostras do HGDP-CEPH com k:4-6 feita a partir de dados do sistema *for*InDel.

*Sistema SNP*for*ID 34plex*

O sistema SNP*for*ID 34plex foi desenvolvido por Phillips *et al.* (2007) – atualizado em Fondevila *et al.* (2013) – e é composto por 34 SNPs autossômicos que apresentam alto δ (Figuras 3 e 4) entre populações africanas, européias e asiáticas e são, portanto, AIMs. Todos se enquadram em quatro categorias: 1) marcadores específicos de população; 2) SNPs com δ's maiores que 0,6; 3) SNPs trialélicos com alelos característicos de dada população; e 4) SNPs com um alelo exclusivo de uma população e um exclusivo de outra.

O sistema é capaz de detectar a diferenciação entre populações africanas, europeias, leste asiáticas e indígenas americanas (Figura 3). No entanto, porque foi desenhado para ser o cerne de um conjunto de painéis para avaliação de ancestralidade, a informatividade para componentes eurasiático que não europeu,

indígena americano e oceânico não foi considerada (Phillips, 2015). Os painéis que o complementam são 1. Pacifiplex, composto por 29 AIM SNPs para estimativa de mistura oceânica (Santos et al., 2016); 2. Eurasiaplex, composto por 23 AIM SNPs para diferenciar europeus de sul-asiáticos (Phillips et al., 2012); e 3. PIMA (apresentado no capítulo 3 deste trabalho), composto por 26 AIM SNPs para diferenciar o componente indígena americano.

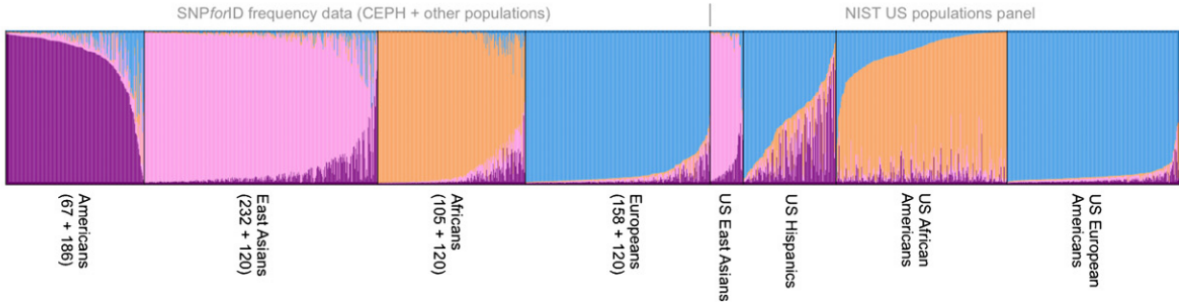


Figura 3. Bar plot mostrando a diferenciação entre africanos, europeus, leste asiáticos e indígenas americanos detectada pela análise de Structure (k:4) utilizando o sistema SNP*for*ID 34plex em amostras do CEPH e do NIST. Modificado de Fondevila et al (2013).

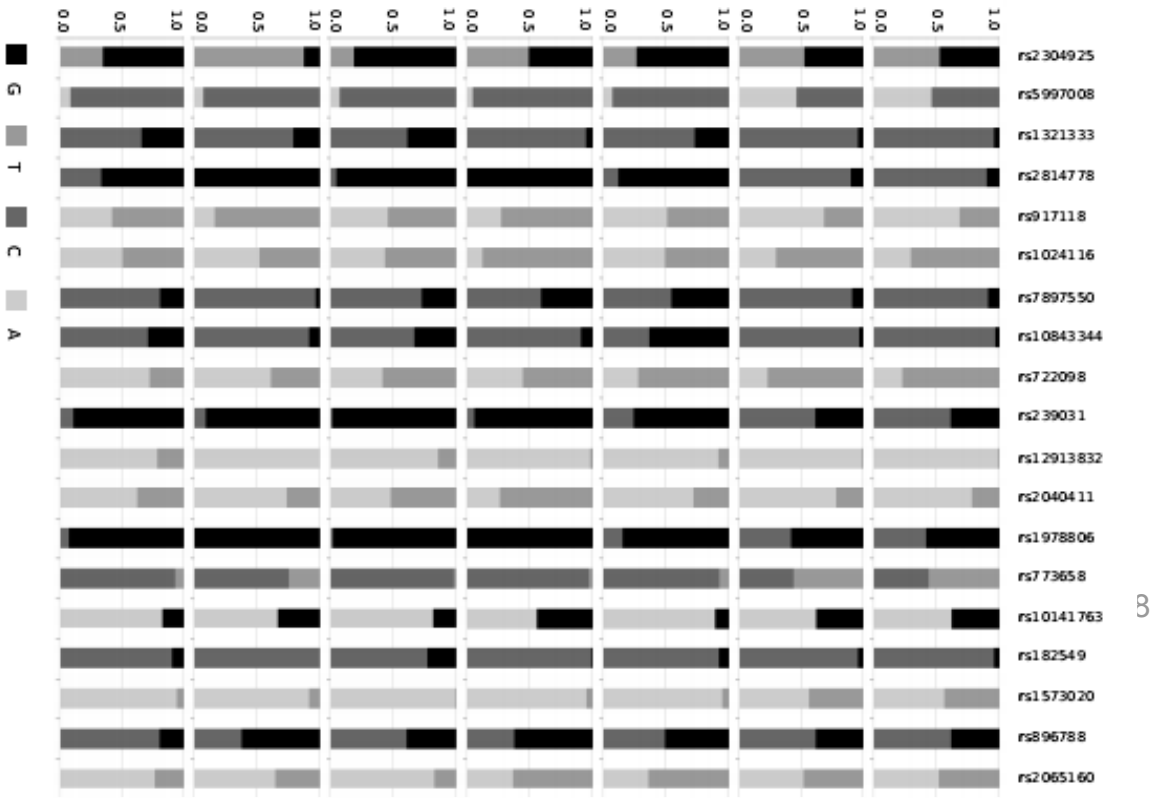


Figura 4. Frequências alélicas dos 34 SNPs em amostras do CEPH compiladas do SNP*for*ID *browser* em populações da África, Europa, América (indígenas), Leste da Ásia, Centro-Sul da Ásia, Oceania e Oriente Médio.

Sistema PIMA

O sistema PIMA, apresentado no capítulo 3 desta tese, foi desenvolvido pelo grupo de pesquisa da Unidade de Genética Forense do Instituto de Medicina Legal na Universidade de Santiago de Compostela, Espanha, com o qual o Laboratório de Genética da Universidade de Brasília tem colaboração estabelecida. O sistema é composto por 26 SNPs autossômicos, o marcador de sexagem amelogenina e um marcador do cromossomo X.

Geralmente, conjuntos de AIMs utilizados em contexto forense para inferência de ancestralidade genômica acumulam valores de divergência maiores para a parental africana, que tem maior diversidade que outras populações, e europeias, porque a maior parte dos estudos na área são conduzidos em países ocidentais (*i.e.* Europa e Estados Unidos). A construção do sistema PIMA teve como objetivo selecionar marcadores com poder para diferenciar a parental indígena americana e complementar o sistema-cerne SNP*for*ID 34plex.

3.1.4 Genotipagem

A análise laboratorial do sistema *for*InDel, descrita em Pereira e colaboradores (2012), se resume a amplificação por PCR em multiplex e genotipagem em eletroforese capilar. Já a análise dos sistemas SNP*for*ID 34plex (Phillips et al., 2007; Fondevila et al., 2013) e PIMA (capítulo 3) inclui reações de purificação e SnaPshot® entre a amplificação e a genotipagem.

Amplificação dos fragmentos de interesse por PCR em multiplex

Os marcadores que compõem os três sistemas analisado foram amplificados por PCR em multiplex utilizando programas específicos e misturas de reagentes em concentrações adequadas a cada conjunto de iniciadores. A amplificação foi verificada por eletroforese em gel de poliacrilamida a 10% corado com nitrato de prata.

***for*InDel:** Os 46 fragmentos de interesse foram amplificados por PCR em multiplex. O volume de reação foi de 10 µL contendo 5 µL de 2x Qiagen Multiplex PCR Master Mix, 10x µL Primer Mix (como especificado em Pereira et al., 2012), 0.75 µL de DNA (entre 0.5 e 5 ng/µl), água miliQ para completar o volume. A amplificação foi feita em 28 ciclos: desnaturação inicial a 95°C por 15 min; desnaturação a 94°C por 30 seg; pareamento a 60°C por 90 seg; extensão a 72°C por 60 seg; extensão final a 72°C por 60 min; manutenção a 4°C.

SNP*for*ID 34-Plex: Os 34 fragmentos de interesse foram amplificados por PCR em multiplex. O volume de reação foi 7.9 µL contendo 0.69 µL de 10x tampão sem $MgCl_2$, 0.63 µL de BSA a µg/µL; 3.2, 1 µL 10x Primer Mix (como especificado em Fondevila et al., 2013), 1,63 µL de $MgCl_2$ a 25mM; 0,4 µL de dNTP mix a 10 mM; 0,1 µL de Taq Gold a 5U/µL; 0,95 de H2O miliQ; 1,5 µL de DNA (entre 1 e 2 ng). A amplificação foi feita em 30 ciclos: desnaturação inicial a 95°C por 15 min; desnaturação a 94°C por 30 seg; pareamento a 60°C por 50 seg; extensão a 65°C por 40 seg; extensão final a 65°C por 15 min; manutenção a 4°C.

PIMA: Os 27 fragmentos de interesse foram amplificados por PCR em multiplex. O volume de reação foi 5 µL contendo 2 µL de 2x Qiagen Multiplex PCR Master Mix, 2 µL 10x Primer Mix, 1 µL de DNA (entre 1 e 10 ng/µl). A amplificação foi feita em 30 ciclos: desnaturação inicial a 95°C por 15 min; desnaturação a 94°C por 30 seg; pareamento a 60°C por 60 seg; extensão a 72°C por 40 seg; extensão final a 72°C por 10 min; manutenção a 4°C. Esse sistema de marcadores está sendo descrito no terceiro capítulo desta tese.

Purificação do produto amplificado

Os produtos da amplificação por PCR dos sistemas PIMA e 34plex foram submetidos a uma etapa de purificação por ExoSAP It com o objetivo de remover iniciadores e dNTP não incorporado às novas moléculas de DNA. A purificação do produto da amplificação do sistema PIMA foi feita utilizando 0,65 µL de ExoSAP It a 1 U/µL para 0.75 µL de produto de PCR em um volume final de 5 µL. A purificação do produto amplificado do sistema 34-Plex foi feita utilizando 1 µL da mesma enzima para 2.5 µL de produto de PCR em um volume final de 3.5 µL. Alternativamente, utilizou-se a enzima ExoStar num volume de 0,6µL a 1 U/uL. O ciclo térmico de purificação foi de 45 minutos a 37ºC e 15 minutos a 85ºC, com manutenção a 4ºC. O produto de PCR do sistema *for*Indel não foi submetido a essa reação.

SnaPshot® (sistemas 34plex e PIMA)

Após a purificação, o produto de PCR dos sistemas PIMA e 34-Plex foi amplificado por SnaPshot®. A análise de SNPs por essa metodologia consiste na extensão de uma única base. Iniciadores se pareiam com a região 3' e com a base

imediatamente *upstream* ao ponto em análise (*i.e.*, o SNP), e possibilitam a extensão da base de interesse utilizando ddNTPs fluorescentes. Posteriormente, a base que foi incorporada é detectada por eletroforese capilar (Sobrino *et al*., 2005). Com o uso de iniciadores adicionados de caudas com tamanhos variados que possibilitam a separação dos fragmentos por eletroforese, é possível montar sistemas em multiplex tão grandes como os aqui analisados..

Amplificação por SnaPshot®: os dois sistemas foram amplificados em 6 µL de volume de reação contendo 4 µL de mix (2,5 µL de SNaPshot® mix; 1,5 µL de mix de iniciadores e 1 µL de H2O MiliQ) e 2 µL de produto de PCR purificado. O programa de termociclagem foi de 30 ciclos para o sistema PIMA e 28 para o sistema 34-plex: 10 segundo a 96ºC; 5 segundos a 55ºC; 30 segundos a 60ºC; manutenção a 4ºC.

Purificação: O produto da reação de SNaPshot® dos dois sistemas foi submetido a uma nova etapa de purificação, também visando a remoção de ddNTP não incorporado às novas moléculas. A purificação do produto de SNaPshot® do sistema PIMA foi feita utilizando 1 µL da enzima SAP para 6 µL de produto amplificado em um volume final de 7 µL. O ciclo térmico de purificação foi de 80 minutos a 37ºC e 15 minutos a 85ºC, com manutenção a 4ºC. A purificação do produto de SNaPshot® do sistema 34-Plex foi feita utilizando 1,3 µL da enzima SAP para 6 µL de produto amplificado em um volume final de 7,3 µL.

Eletroforese capilar

Às placas contendo o produto de PCR do sistema *for*InDel e de purificação com SAP dos demais, foi adicionada uma mistura de reagentes que visa a desnaturação das moléculas de DNA para a eletroforese capilar. Os fragmentos foram analisados em um sequenciador 3130 Genetic Analyzer (Applied Biosystems) usando capilares de 36 cm e polímero POP-4™. A genotipagem foi feita com o software GeneMapper™ 3.2 (Applied Biosystems).

Para a eletroforese do sistema *for*InDel, 0.8 µL do produto de PCR diluído 10x em H_2O MiliQ foi adicionado a 0.3 µL de padrão interno de tamanho 500 LIZ™ e 8.9 µL de formamida Hi-Di™. Para o sistema PIMA, 1 µL do produto de SnaPshot® purificado foi adicionado a 9,5 µL de mix (0.25 µL de padrão interno de tamanho 120 LIZ™ e 10 µL de formamida Hi-Di™). Para o sistema 34-Plex, 1 µL do produto de SnaPshot® purificado foi adicionado a 0.3 µL de padrão interno de tamanho 120 LIZ™ e 8,9 µL de formamida Hi-Di™.

3.4 Análise estatística

3.4.1 Análise descritiva

Frequências alélicas e genotípicas, Equilíbrio de Hardy-Weinberg e heterozigose

O Equilíbrio de Hardy-Weinberg (HWE) é um modelo matemático que descreve as frequências de alelos e genótipos esperadas na ausência de fatores perturbadores da distribuição basal. Desvios no HWE sinalizam a presença de fatores evolutivos ou problemas metodológicos na amostragem ou genotipagem. A

heterozigose é uma das formas de avaliar a razão do desequilíbrio observado em uma população e sugerir seleção, estruturação e miscigenação. Executamos as estimativas de frequências e heterozigose (H_{obs} e H_{esp}) e o teste para verificar adequação ao HWE com o programa Arlequin 3.5 (Excoffier e Lischer, 2010).

Desequilíbrio de ligação

O desequilíbrio de ligação (LD), situação em que a herança de um alelo não é independente da de um outro, pode ser gerado por inúmeros fatores. Dentre eles estão proximidade física entre os *loci*, seleção natural, deriva genética, mistura recente e estruturação populacional (Slatkin, 2008). Porque o efeito da deriva genética sobre o LD é pequeno (Slatkin, 2008) e porque os *loci* aqui analisados são seletivamente neutros e distantes fisicamente um do outro, qualquer LD observado deve ser decorrente de algum dos outros fatores ou efeito de amostragem. Para verificar a ocorrência de LD entre os *loci* aqui analisados, utilizamos o teste exato de Fisher, executando pelo o programa Arlequin 3.5 (Excoffier e Lischer, 2010).

3.4.2 Análise comparativa

Índice de fixação: F_{ST}

O F_{ST} é uma medida de variabilidade diretamente relacionada à variância das frequências alélicas de um *locus* em uma população: quanto maior a variância das estimativas em torno da média, maior o F_{ST}. Isso significa que quanto mais diferentes são as frequências alélicas entre as populações comparadas, maior é o

F_{ST} (Holsinger e Weir 2009). A relação entre F_{ST} e história evolutiva é obscurecida por fatores como:

- diferentes cenários evolutivos podem resultar em estruturas e relações populacionais com F_{ST} similares. F_{ST} baixos, por exemplo, podem ser consequência de divergência recente ou migração;

- marcadores em diferentes regiões do genoma podem informar F_{ST} diferentes para a mesma amostra;

- marcadores de categorias moleculares diferentes informam F_{ST} diferentes para as mesmas populações. Para populações humanas, STRs informam F_{ST} médio de 0.05, enquanto SNPs informam 0.10 (Li et al., 2008).

Para a análise de reconstrução histórica, o F_{ST} informa relações evolutivas e demográficas: quanto maior o F_{ST}, mais distante a relação entre populações. No entanto, ele não diferencia proximidade no valores de F_{ST} decorrente de ancestralidade comum recente ou migração (Holsinger e Weir 2009). Para a genética forense, o F_{ST} reflete a informatividade de um marcador, já que está relacionado à variabilidade em cada nível da estrutura populacional (Phillips 2015).

Aqui, com o objetivo de verificar a diferenciação entre as populações analisadas (capítulos 1 e 2), selecionar parentais adequadas (capítulo 1) e avaliar a informatividade de marcadores (capítulo 2), estimamos o F_{ST} como implementado no *software* Arlequin 3.5 (Excoffier e Lischer, 2010).

3.4.3 Análise de ancestralidade e estruturação populacional

Avaliamos os padrões de estrutura populacional e ancestralidade de forma exploratória utilizando análise de componentes principais (PCA). Em seguida, seguimos duas abordagens distintas: o modelo proposto no programa Structure (Pritchard et al., 2000), que é utilizado para produzir uma estimativa de mistura, e o algoritmo Naïve Bayes, que é utilizado para atribuir origem assumindo ausência de miscigenação.

A PCA é uma metodologia de redução de dimensionalidade que facilita a visualização de dados multivariados (Slatkin 2008). Ela realiza uma rotação do conjunto de dados para um sistema de eixos que elimina a correlação entre cada componente. Desta forma, eixos são ordenados pela quantidade de variância associada, sendo que o primeiro sumariza a maior quantidade de variação observada e cada eixo ortogonal subsequente acumula menos informação que o anterior. Aqui, a PCA foi feita utilizando o *software* Kpop (Mendes e Gontijo, 2017).

Realizamos a análise de mistura e estruturação com o *software* Structure 2.3.4 (Pritchard, et al., 2000). Esse algoritmo parte de dados genotípicos para inferir o agrupamento da amostra em k grupos correspondentes à estruturação existente no conjunto de dados. Cada indivíduo é associado a uma distribuição de probabilidades que corresponde ao grau de afiliação a cada um desses grupos. O valor de k normalmente deve ser informado *a priori*, mas existem métodos heurísticos para inferi-lo a partir da amostra.

Em cada capítulo, adotamos populações parentais e parâmetros adequados, como descrito na metodologia específica. Testamos vários k (2 a n + 2, sendo n o número de populações parentais presumidas na amostra de referência) em múltiplas iterações. Selecionamos os grupos parentais dentre os dados disponíveis na literatura tendo em conta plausibilidade histórica e coesão dos grupos parentais

Finalmente, algumas análises foram complementadas com o resultado da classificação do método Naïve Bayes. Como o nome sugere, o método está baseado na utilização de modelos estatísticos Bayesianos simples e que possuem resoluções exatas. Neste trabalho, utilizamos uma versão do método que, no contexto da genética, corresponde a um modelo muito semelhante ao proposto por Pritchard e colaboradores (2000) na elaboração do Structure. Os modelos diferem apenas em que o Naïve Bayes assume ausência de mistura, enquanto o Structure a modela explicitamente. O resultado de uma execução do Naïve Bayes consiste em uma distribuição para cada indivíduo indicando a probabilidade de possuir uma ancestralidade pura para cada população parental. Isto não deve ser confundido com estimativas de mistura, ainda que existam circunstâncias em que estes valores se aproximam.

A análise de classificação por Naïve Bayes foi realizada utilizando o *Snniper* (http://mathgene.usc.es/snipper/), um *software* que integra dados de populações do CEPH para os sistemas de marcadores *for*InDel e 34plex (aqui utilizados).

CAPÍTULO 1

Sistema *for*InDel: Análise de ancestralidade em populações rurais brasileiras de ascendência africana

Este capítulo consiste no manuscrito original do trabalho publicado com o título “Ancestry analysis in rural Brazilian populations of African descent”, publicado em Forensic Science International: Genetics, Volume 36, September 2018, Pages 160-166).

RESUMO

Comunidade rurais constituem cerca de 20% da população do Caribe e América Latina, mas estão sub-representadas em bancos de dados de marcadores autossômicos. Essa deficiência é problemática para a genética forense, que depende de descrições acuradas da variação genética e da estrutura populacional. As populações brasileiras foram moldadas por um processo de mistura intenso, complexo e heterogêneo que envolveu especialmente indígenas americanos, africanos sub-saarianos e europeus. Esse processo originou populações diversas, dentre elas os quilombos: populações brasileiras com ancestralidade africana marcada e que permaneceram, em certa medida, isoladas geneticamente de outras populações. Aqui, foram analisados dois quilombos rurais: Kalunga - GO e Riacho de Sacutiaba e Sacutiaba - BA; e Mocambo - SE, em conjunto com dados compilados do HGDP-CEPH. Com o objetivo de contribuir para a construção de bancos de dados de marcadores de interesse forense que representem adequadamente a população brasileira, os três quilombos foram analisados, e a relação entre sua história e composição genética, investigada. Um sistema forense composto por 46 marcadores informativos de ancestralidade (AIM) do tipo indel foi escolhido por seu alto poder em diferenciar as principais populações parentais putativas do Brasil. As populações parentais foram selecionadas do banco de dados do HGDP-CEPH disponíveis no *for*InDel *allele frequency browser* com base em padrões históricos aplicáveis às populações em análise e na quantidade de variação observada dentro e entre continentes. Os resultados mostram que os componentes de mistura mais significativos nos Quilombos são o africano e o europeu, em

consonância com análises prévias de marcadores uniparentais e autossômicos nas mesmas populações. PCA, análise de estruturação e estimativas de ancestralidade indicam uma correlação entre o grau de isolamento e a ancestralidade observada: Kalunga é a mais isolada e tem o maior componente de mistura africano (67,3%). Sacutiaba é a menor e mais impactada por migração e tem o maior componente europeu (46,8%). Mocambo é vizinho a uma população ameríndia (Xocó) e, portanto, tem a maior contribuição indígena (12,2%). Os resultados são também consistentes com a história e demografia conhecidas dos quilombos. A heterogeneidade aqui observada ressalta a diversidade genética que o Caribe e a América Latina devem apresentar e reitera a importância de descrevê-las em maior detalhe.

Palavras-chave: Ancestralidade; Indel; AIM; Mistura; Quilombo; Brasil

Ancestry analysis in rural Brazilian populations of African descent

Carolina Carvalho Gontijo*[a,b,c], Fábio Macêdo Mendes[d], Carla A. Santos[a], Maria de Nazaré Klautau-Guimarães[b], Maria Victoria Lareu[a], Ángel Carracedo[a], Christopher Phillips[a], Silviene F Oliveira*[b,c]

[a] Unidade de Xenética, Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain

[b] Human Genetics Laboratory, Institute of Biological Sciences, University of Brasília, Brazil

[c] Animal Biology Graduate Program, Institute of Biological Sciences, University of Brasília, Brazil

[d] University of Brasília at Gama – FGA, University of Brasília, Brasil

ABSTRACT

Rural communities comprise around 20% of Caribbean and South American populations, but are under-represented in autosomal marker databases. That deficiency is problematic for forensic genetics, as it relies on accurate descriptions of genetic variation and population structure. Brazilian populations were shaped by an intense, complex and heterogeneous process of admixture encompassing mainly Amerindians, Sub-Saharan Africans and Europeans. Quilombos are Brazilian populations with significant African descent that have remained genetically isolated to some extent from surrounding populations. In the reported study, we analyzed three rural Quilombo populations: Kalunga; Riacho de Sacutiaba e Sacutiaba; and Mocambo, along with a dataset from the HGDP-CEPH panel. Aiming to contribute to representative genetic databases of forensic interest, we analyse the three rural Quilombos and investigate how their genetic makeup relates to their history. An established forensic test, comprising 46 ancestry-informative (AIM) Indels, was chosen for its high power in differentiating the main contributing populations of Brazil. Parental populations were selected from HGDP-CEPH data available at the *for*InDel allele frequency browser based on historic patterns applicable to the study populations and the amount of variability observed within and between continents. Our results show the main admixture components in the Quilombos are African and European. Those estimates are in accordance with previous analyses for both uniparental and autosomal markers. PCA, structure analysis and ancestry estimates indicate a correlation between the extent of isolation and the degree of admixture in the Quilombos: Kalunga is the most isolated population and accordingly has a higher

African admixture component (67.3%). Sacutiaba is the smallest and most impacted by migration, with the highest European component (46.8%). Mocambo neighbors a Native American population and therefore has the highest Amerindian contribution (12.2%). Our results are consistent with the history and demography of Quilombos. The heterogeneity observed in these populations stresses the genetic diversity that Latin American and Caribbean rural populations can have and reiterates the need to describe them in greater detail.

Keywords: Ancestry; Indel; AIM; Admixture; Quilombo; Brazil

Highlights

- Genetic structure and ancestry of African-Brazilian rural populations are described.
- Data on the 46 AIM Indels is consistent with the Quilombos' history and demography.
- Quilombos are not uniform and their African ancestry surpasses the Brazilian average.
- Rural Brazilian populations are diverse, yet under-represented in forensic databases.

1. Introduction

Brazilian populations were shaped by an intense, complex and heterogeneous process of admixture that encompassed mainly Amerindians, Sub-Saharan Africans and Europeans [1-3]. Those groups have gone through particular demographic processes themselves over Brazil's history. Their movement, dispersion and interaction with indigenous populations were not homogeneous across the vast Brazilian territory, making the demographics of each region unique. Additionally, later migration from different parts of the world must be taken into account, especially from the Middle East and East Asia (late 19th/early 20th century) [4, 5]. Aside from the admixed populations that make up most of Brazil, partially and completely isolated populations persist: Quilombos and Amerindian populations are found throughout the country [6, 7].

Around 40% of all Africans that were enslaved and forcibly brought to the Americas (3.6-10 million people) arrived in Brazil. Historical records of the slave trade and traffic are scarce, but Bantu peoples are believed to have been the main group brought to the Americas [8]. In Brazil, African slaves and those of African descent often resisted captivity and exploitation. One of the most significant forms of resistance was the constitution of Quilombos: communities that were safe havens that were isolated from surrounding populations. Despite being regarded as isolated communities, Quilombos have always maintained commercial and social links with surrounding populations, resulting in the evident admixture that has occurred since their foundation [9]. They have unique histories

and are not homogeneous [1], but form a cohesive group in their ancestry [9]. Currently, over 3000 Quilombos populations or communities have been officially recognized in every Brazilian state [6], except Acre and Roraima [10], comprising over two million people [11].

Many previous studies on the genetics and ancestry of Quilombos have analyzed small sets of classical autosomal markers, or focused on mtDNA and Y-chromosome (as reviewed in [1, 2, 12]). These analyses and those performed with autosomal markers usually show patterns consistent with Quilombos' history and demography and show a high African admixture component, with variable and smaller proportions of European and Amerindian co-ancestry (often in that order of contribution)[2, 13-15]. Lineage markers have indicated a strong sex bias against European mtDNA and towards European Y-chromosome backgrounds [12,16,17].

Like Quilombos, Latin American populations are under-represented in genetic marker databases, evident from the sample composition of databases such as HGDP-CEPH [18] and 1000 Genomes [19]. Larger sets of AIMs, including the panel used for this study, have scant coverage of Native American populations (reviewed in [1, 3]) and rural communities, which comprise 20% of the continent's populations, are not represented. Over the last years, the gap in genetic profiles available for Latin American populations is being addressed, as more studies directed at the continent are being published (reviewed in [1, 3], further exemplified in [20-33]). Incomplete sampling of a continent is problematic for forensic genetics analysis, as accurate descriptions of a region's genetic

variation and structure are important for accurate reporting of forensic DNA data. Choosing ancestry-informative markers with sufficient power and in adequate numbers is also crucial for these analyses [34-39]).

In this study, we aimed to fill a significant gap in knowledge of Brazilian demographics by examining three rural Quilombos and investigating how their genetic makeup relates to their history. We evaluated how consistent the new data is with previous population analyses of Brazil. The three rural Quilombos are Kalunga, Riacho de Sacutiaba e Sacutiaba (herein: Sacutiaba), and Mocambo. They make up a historically, geographically and demographically diverse and well-documented set of populations. We focus on population structure and genomic ancestry, supporting our findings with historic and demographic data.

2. Material and methods

2.1. Sampled populations

Kalunga is one of the largest Quilombos in Brazil, numbering an estimated 5,300. The founding Kalungas were slaves that either escaped captivity or were left behind after gold mining finished in late XVIII century [40]. It is the most isolated of the three populations analyzed, but has seen a recent influx of migrants to its outskirts. Sacutiaba are an extended kinship of 200 people divided into smaller family groups. The community is believed to have been founded by runaway slaves from the neighboring state of Minas Gerais 200 years ago [41]. Mocambo is a small rural community with an estimated population of 500. Xocó

Amerindians inhabit a neighboring territory and the two populations share long-standing relationships and a close history [42]. The founding date is uncertain, but there is record of a small settlement of runaway slaves in the region in 1825.

2.2. Sample collection and ethical approval

Blood was collected from 72 volunteer donors from Kalunga, 30 from Sacutiaba and 81 from Mocambo. Because of their sizes and demographic structures, the final sample size was restricted by endogamic relationships within each population. Further, we applied a questionnaire to the donors and to an expanded sample to assess demographic information about the participants, their offspring and two previous generations. In accordance with Brazilian regulation on ethics in human research, all participants were informed about the confidentiality of the data and the research details. The project was approved by the Research Ethics Committee of the Faculty of Health Sciences at the University of Brasilia (CEP-FS/UnB 030/2002 and 151/07).

2.3. The ancestry-informative Indel panel

An established forensic panel comprising 46 AIM Indels [43] was chosen for its high power in differentiating the main parental populations of Brazil (European, African and Amerindian), making it an appropriate assay for assessing ancestry in Quilombos. Along with the Quilombos' samples, data from HGDP-CEPH populations [18] compiled from the *for*InDel allele frequency

browser (http://spsmart.cesga.es/forindel.php) was included in comparative analyses and ancestry estimates (see Supplementary material for details).

2.4. DNA extraction and Indel genotyping

DNA was extracted with the genomicPrep Mini Spin Kit (GE Healthcare) and quantified with a NanoDrop 1000™. DNA analysis followed the capillary genotyping protocol originally described for the Indel panel [43]. Specific Indel fragments were amplified in a 46-multiplex PCR in 10 µL reactions containing 5 µL 2x Qiagen Multiplex PCR Master Mix, 10x µL Primer Mix (as per Pereira et al., 2012), 0.75 µL DNA (concentration range: 0.5 to 5 ng/µl), miliQ water up to reaction volume. Amplification cycling was: initial denaturation at 95°C for 15 min; 94°C for 30 s, 60°C for 90 s, 72°C for 60 s; final extension at 72°C for 60 min. 0.8 µL of PCR product was added to 0.3 µL of 500 LIZ™ internal size standard and 8.9 µL of Hi-Di™ formamide. Products were detected in a 3130 Genetic Analyzer (Applied Biosystems) using 36 cm capillary arrays and POP-4™ polymer. Allele calling was made with GeneMapper™ 3.2 software (Applied Biosystems).

2.5. Statistical analysis

Population descriptive parameters: Hardy Weinberg equilibrium (HWE) - linkage disequilibrium (LD) and Fst were estimated using ARLEQUIN 3.5 [44] and K-POP 0.1 (available at https://github.com/fabiommendes/kpop) [45], respectively. Allele frequencies were estimated by STRUCTURE 2.3 [46].

Patterns of population structure were assessed by PCA and structure analysis. PCA was made using K-POP, based on individual genotypes. STRUCTURE analyses tested K:2-8, using the following parameters: burn-in of 10,000; MCMC of 100,000; Admixture and LOCPRIOR models; correlated allele frequencies; 8 iterations. From the output, CLUMPAK [47] determined the optimal K (LnPr(X|K)) according to the method of Evanno [48]; and K-POP aided the plotting of cluster membership proportions.

Contributor populations were selected from HGDP-CEPH data available at *for*InDel browser. Following the methodology described below, our initial comparison sample was narrowed down from 907 to 542 individual profiles. The first criterion was the known history of admixture between Amerindians, Europeans and Africans in Brazil. The second criterion was the amount of variability observed within and between continents. F_{ST} estimates were used to ensure cohesive population groupings. Considering F_{ST} ranges commonly observed in human populations for SNPs [49], we established 0.10 as the maximum F_{ST} for within-continent samples and 0.20 as the minimum accepted as descriptive of between-continent population stratification.

Pairwise F_{ST} estimates are shown in Table S1. East Asia was included despite no expectation of a significant contribution to the Quilombos, as this population group has contributed to Brazilian variability in other populations. Middle East, Central and South Asia, three African populations (Biaka, Mbuti and San) and two East Asian populations (Lahu and Daur) were excluded based on their Fst values. The selected admixture contributor set was composed of: African

- Bantu N.E., Yoruba, Mandenka, South African Bantu; European - Basque, French, Sardinian, Tuscan, Bergamo, Orcadian, Russian, Adygei; American - Karitiana, Surui, Colombian, Maya, Pima; East Asian - Cambodian Dai, Han, Hezhen, Miao, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yi, Japanese, Yakut.

Amerindian populations are more structured than other continental populations due to their demography [49-51], so we evaluated this with STRUCTURE analysis (burn-in: 10000; MCMC: 100000; model: Admixture; allelic frequencies: correlated; POPFLAG=0; iterations: 8) and Naïve Bayes classification (Bayes analysis), as implemented in K-POP.

Outlier individuals (those with admixture estimates <80% for the group to which they belong and ≥20% for a second co-ancestry contributor and/or assigned to a different group based on Naïve Bayes analysis) were excluded from our contributor samples. Based on these criteria, only four samples were excluded from the European, four from the East Asian and five from the Amerindian sets.

Ancestry estimations were based on cluster membership proportions identified from STRUCTURE analyses. Assignment of HGDP-CEPH samples to the most likely population of origin used Naïve Bayes classifications in K-POP and extended to Central-South Asian and Middle-Eastern populations and test subjects, in addition to the expected contributor populations. Samples with more than 10% missing data were excluded from Bayes analysis.

3. Results

3.1. Population-descriptive parameters

Table S2 shows basic population parameters of HWE; expected and observed heterozygosities; and allele frequencies, estimated for the study samples. In Kalunga, 8% of the loci were out of HWE, 6% in Sacutiaba and 30% in Mocambo. With Bonferroni correction these values dropped to close to zero in Kalunga and Sacutiaba and to 20% in Mocambo. Table S3 shows LD estimates: 2.5% of the pairs of loci are in LD in Kalunga, 3% in Sacutiaba and 4% in Mocambo.

3.2. Population structure

Fig 1 shows PCA analysis and the STRUCTURE triangle plot depicting the genetic structure seen in study populations and in relation to the contributor population groups. PCA analysis (Fig 1A) indicates three contributor clusters and not four as would be expected. In this analysis, Amerindians and East Asians were not clearly separated. Quilombo individuals fall between Africans and Europeans and despite a large degree of overlap, the Kalunga cluster is placed closer to the African cluster, with some individuals being positioned in that cluster. The Sacutiaba cluster is closer to the European cluster, with some individuals positioned within it. Mocambo occupies a more central position and are more dispersed. The triangle plot (Fig 1B), similarly shows Quilombos to be

intermediate to Africa and Europe and to form rather more tightly grouped clusters than PCA.

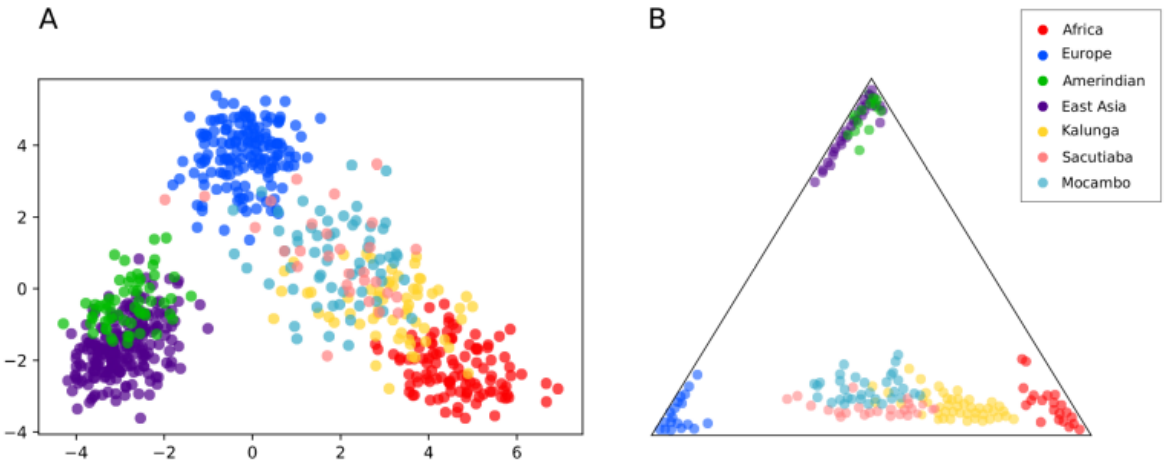


Figure 1. Population structure analyses considering four parental groups (Africa, Europe, America, and East Asia) and the three study populations (Kalunga, Sacutiaba, and Mocambo). **A:** Principal Component Analysis (PCA) done on K-POP 0.1. **B:** Triangle plots generated by STRUCTURE 2.3.

3.3. Ancestry analysis and population affiliations

Ancestry estimates are shown in Fig 2 (pie chart for population estimates), Fig 3 (bar plots for individual estimates) and Fig 4 (violin plots for range of co-ancestry contributions) and in Tables 1 and S4. Bar plots for K:3 to K:5 are shown in Fig 3. The optimum K value was K:3. Kalunga shows higher and more homogeneous African cluster memberships representing a larger overall admixture component, whereas Sacutiaba and Mocambo are more varied in their

admixture patterns with more complexity. The range of individual ancestries observed in each HGDP-CEPH population shows they are not homogeneous. Notably, Yakut are similar in their patterns to Amerindian populations at K:3, but show European and Amerindian ancestry components at K:4.

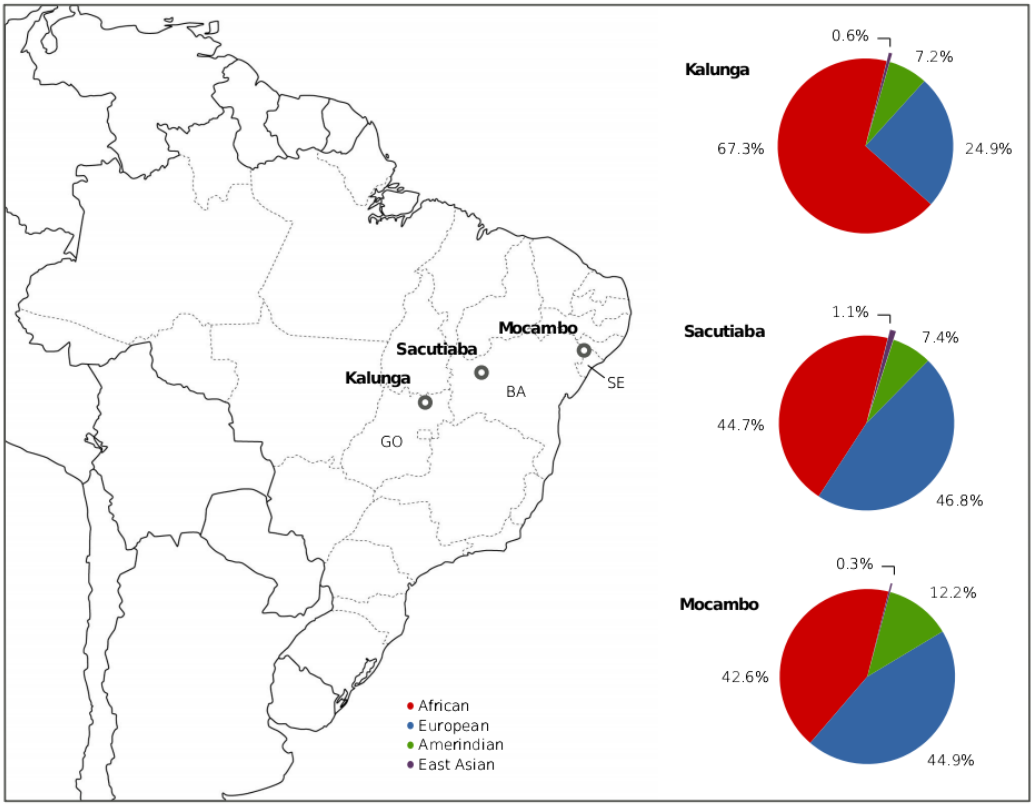


Figure 2. Map of Brazil indicating the location of the three *Quilombos* (Kalunga, Sacutiaba and Mocambo) and pie charts showing their population ancestry estimates. Analysis was done on STRUCTURE 2.3.

Table 1 details population ancestry assignments estimated with the 46 AIM Indel data and previous estimates made with different sets of markers, both lineage (Y-STR and mtDNA) and autosomal (classical markers, Alu insertions, STRs, small sets of AIMs – SNPs and Indels). There is a clear difference in the results obtained by the two types of marker, as expected in highly admixed

populations: lineage markers do not describe the complete complexity in background in each individual, but bring specific information to the analysis that autosomal markers cannot.

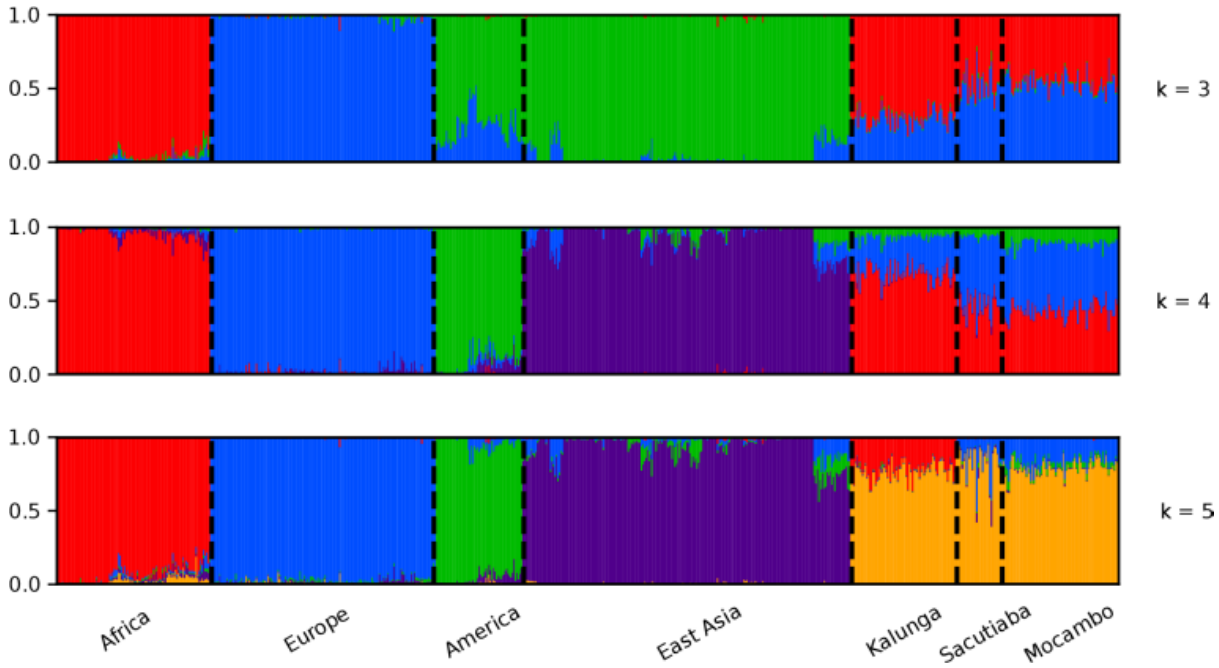


Figure 3. Ancestry analyses for Kalunga, Sacutiaba and Mocambo considering four parental populations (Africa, Europe, America, and East Asia). Bar plots generated on STRUCTURE 2.3. (k:3-5 are shown).

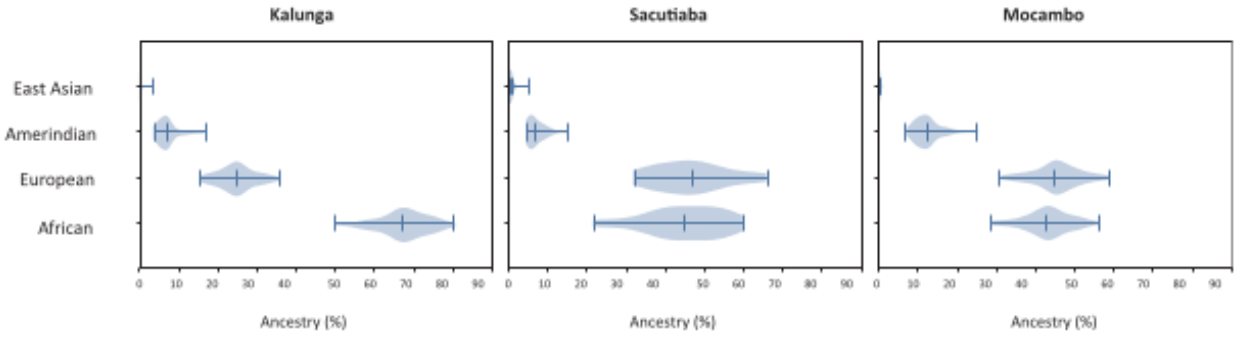


Figure 4. Distribution of individual ancestries in the study populations. Violin plots generated on K-POP 0.1 based on ancestry analyses considering four parental populations (Africa, Europe, America and East Asia).

Table 1. Estimates of population ancestry for Kalunga, Sacutiaba and Mocambo based on the 46 AIM Indels panel and previously published estimates based on lineage (Y-STR and mtDNA) and autosomal (classic proteins, Alu insertions, STRs, small sets of AIMs – SNPs and Indels) sets of genetic markers.

	Marker	Ancestry estimate			
		AFR	**EUR**	**AME**	**EAS**
Kalunga	Y-STRs	27.00	67.00	6.00	–
	mtDNA	86.77	0.00	13.30	–
	Autosomal STRs	57.38	39.29	3.33	–
	SNP AIMs	61.05	29.77	9.18	–
	46 AIM Indel	**67.3**	**24.9**	**7.20**	**0.06**
Sacutiaba	Y-STRs	5.00	95.00	0.00	–
	mtDNA	84.06	0.00	15.4	–
	Classical and *Alu*	72.29	27.71	0.00	–
	Autosomal STRs	35.08	64.31	0.61	–
	SNP AIMs	53.94	43.33	2.73	–
	AIMs	43.00	55.00	0.20	–
	46 AIM Indel	**44.70**	**46.80**	**7.40**	**1.10**
Mocambo	Y-STRs	7.00	86.00	7.00	–
	mtDNA	78.60	0.00	21.40	–
	Classical and *Alu*	32.00	57.57	10.32	–
	Autosomal STRs	46.57	39.81	13.62	–
	SNP AIMs	43.42	37.77	18.81	–
	AIMs	46.00	35.00	19.00	–
	46 AIM Indel	**42.60**	**44.90**	**12.20**	**0.30**

Reference: Y-STRs [17]; mtDNA [57]; classical and Alu insertions [64]; autosomal STRs [65]; SNP AIMs [65]; AIMs [9]; 46 AIM Indel: current work. **AFR**: Africa; **EUR**: Europe; **AME**: Amerindian; **EAS**: East Asia.

Table S4 shows ancestry estimates side by side with population classifications for each individual. Most were assigned to the population with the strongest admixture component in their genomes. Nine individuals were assigned to a non-African population in Kalunga (eight to Europeans and one to

Amerindians), 16 in Sacutiaba (all assigned to be Europeans) and 43 in Mocambo (41 were assigned to be Europeans and two to be Amerindians). The disparity between both analyses reaches 15% in Kalunga, 7% in Sacutiaba and 21% in Mocambo.

Kalunga has the highest overall African contribution, with individual estimates ranging from 49.7-80.2%. Sacutiaba has the highest European contribution, with individuals ranging from 32.1-66%. Mocambo is the population with the highest Amerindian contribution overall (12.2%) and individually – the highest estimated proportion being 25.1% (Figs 3 and 4). Most individuals in our sample have at least 40% African ancestry: all of the Kalunga individuals (where 84% of individuals have at least 60% of African ancestry), 67% in Sacutiaba and 65% in Mocambo.

4. Discussion

Quilombos are a unique group of populations in their history and ancestry. They are part of the Brazilian campesino movements and have particular histories, demographic trajectories and culture. With the goal of investigating how the genetic makeup of three of those rural African-Brazilian populations relates to their history, we used a simple but informative autosomal panel of 46 AIM Indels [43]. This genetic system has a high power to differentiate the main contributing populations to Brazil (namely Europeans, Africans and Amerindians). East Asians were included because of their more recent, but intense influx into Brazil. Our results add to the understanding of the genetic composition of Quilombos and, in

a wider sense, of Latin American rural populations in general, as well as enriching a population database of forensic interest.

The main admixture components to the Quilombos are Africa and Europe, as expected from historical records and corroborated by our data. Those estimates are also in accordance with previous analyses published elsewhere for both uniparental and autosomal markers, as summarized in Table 1. Furthermore, our data indicate a correlation between the extent of population isolation amongst the Quilombos and other populations and the degree of admixture in the Quilombos. Kalunga is the most isolated of them and accordingly has a higher African admixture component (67.3%). Sacutiaba is the smallest community and most impacted by incoming migration [2], with the highest European component (46.8%). Mocambo neighbors a Native American population and therefore has the highest Amerindian contribution (12.2%). The three Quilombo populations of this study share demographic characteristics and an overall genetic ancestry. Nevertheless, they are not homogeneous, as shown by the results.

The predominance of endogamic marriages [52, 53] is further indication of the very low levels of gene flow in and out of the Kalunga population. Sacutiaba shows a genetic composition that is the result of a different admixture process. The community is located in the State of Bahia, closer to other populations and its higher European component reflects very active trade and exchange between these populations. Finally, Mocambo has the highest Amerindian component observed (12.2%), despite a high Amerindian contribution being uncommon amongst Quilombos populations [2] (exceptions were described in northern Brazil

[14]). Mocambo neighbors the Xocó Amerindian population and the relationship between these two populations is old and well documented [42].

The genetic outcome of these inter-population dynamics has been observed in previous studies [9]. The complexity observed in Sacutiaba and Mocambo yielded measurable impact of migration on their genetic compositions and high immigrant proportions (30% in Sacutiaba and 16% in Mocambo). In both populations endogamous marriages are predominant. The European contribution seen by Amorim [9] was higher among immigrants than Quilombolas (especially in Mocambo). These findings match well with the current study, in that migration has brought European variation into the Quilombo communities.

Such high immigration rate could have produced a measure of population substructure in Mocambo. That could be the underlying cause of the HWE deviations detected and the low Hobs, as could be expected when migration is recent and/or continued, especially when the originating and receiving populations have different genetic compositions. That is the case in Mocambo, where 11 of the 46 markers are not in HWE, and have observed heterozygosity (H_{obs}) much lower than expected (H_{exp}). That population's demographic dynamics involves continued influx from the neighboring Xocó and recent migration from surrounding populations with a higher European admixture component [9]. Sacutiaba is a smaller population with higher immigration rates, but did not present the deviations seen in Mocambo. That might be due to the fact that

Bahia, where Sacutiaba is located, is a state with a genetic ancestry comparable to that usually observed in Quilombos [3, 16, 54-56].

The observation of East Asian components in our sample (0.3-1.1%) is low enough to discount the detection of migration or exchange between Quilombos and Asian-Brazilian populations and more likely reflects the difficulty of fully differentiating East Asians and Amerindians with the 46 AIM Indels used.

When comparing our estimates to those of previous studies with autosomal markers (Table 1), highly consistent patterns emerge in which the African component is very important. Classical and protein markers are exceptions, as the loci used lack population-differentiated allele frequency distributions. The ancestry estimates we obtained with 46 AIM Indels can be considered to accurately describe the genetic composition of Kalunga, Sacutiaba and Mocambo. The sex bias in admixture contributions indicated by lineage markers shows a predominantly European male contribution in the three Quilombos populations while the female contribution is mostly African. No European mtDNA was found in Kalunga, Sacutiaba and Mocambo [57].

In the wider context of the Brazilian population, the levels of African co-ancestry we observed in the three populations is only comparable to that seen in other Quilombos and in a few other African-Brazilian populations [3, 16, 55-57]. Salvador, capital of the State of Bahia, is the best studied of such populations and its African component has been estimated to average 49.2% [57]. In contrast, across the whole country, the predominant ancestry is European, the African

component is considerably lower and more variable and the Amerindian ranges from very low to barely detectable [2, 21, 58-62], with some notable exceptions in northern states (as summarized in [3, 13, 25, 31, 63]). As estimated from autosomal markers, the mean African contribution in urban Brazil ranges from 5% in the South to 30% in the Southeast and Northeast [3] – the two regions where slavery was more prominent.

5. Conclusions

The three Brazilian populations with marked African descent studied here represent Quilombos in the traditional sense: rather isolated communities built by former African and African-descendant slaves during the colonial times or shortly after. In spite of the partial isolation and in addition to their already complex origins, Quilombos have maintained bonds with surrounding populations to produce variable genetic backgrounds. Therefore, they are complex admixed populations, which is reflected in their population structure and admixture ratio data.

The genetic data obtained from 46 ancestry informative Indels is consistent with the history and demography of Quilombos and matches well with our previous studies of these populations. They are not homogeneous, but share demographic and genetic characteristics. Kalunga, Sacutiaba and Mocambo have high African co-ancestry levels of 67.3%, 44.7% and 42.6%, respectively, compared to an urban Brazilian population average of 30% in regions where slavery was more significant. That contrast is due to Quilombos' exceptional

history, which itself underlies the diversity seen in Brazilian rural populations and more generally across Latin American regions.

Acknowledgments

CAPES, CNPq and FAPDF supported the research that led to this publication. The authors would like to thank the people from Kalunga, Sacutiaba and Mocambo for their kind cooperation and willingness to participate in the study.

References

[1] F.M. Salzano, M. Sans, Interethnic admixture and the evolution of Latin American populations, Genetics and Molecular Biology, 37 (2014) 151-17

[2] C.C. Gontijo, C.E.G. Amorim, N.M.O. Godinho, R.C.P. Toledo, A. Nunes, W. Silva, M.M.F. Moura, J.C.C. de Oliveira, R.C. Pagotto, M.N. Klautau-Guimarães, S.F. Oliveira, Brazilian Quilombos: A Repository of Amerindian Alleles, American Journal of Human Biology, 26 (2014) 142-150. doi:10.1002/ajhb.22501

[3] R.R. de Moura, A.V.C. Coelho, V.Q. Balbino, S. Crovella, L.A.C. Brandão, Meta-Analysis of Brazilian Genetic Admixture and Comparison with Other Latin America Countries, American Journal of Human Biology 27 (2015) 674–680. doi:10.1002/ajhb.22714

[4] Instituto Brasileiro de Geografia e Estatística – IBGE: http://www.ibge.gov.br (accessed 22 November 2017)

[5] C.B. Moysés, W.M. Tsutsumida, P.E. Raimann, Population data of the 21 autosomal STRs included in the GlobalFiler ® kits in population samples from five Brazilian regions, Forensic Sci Int Genet, 26:e28-e30 (2017). doi:10.1016/j.fsigen.2016.10.017

[6] Fundação Cultural Palmares – FCP: http://www.palmares.gov.br, 2017 (accessed 20 December 2017)

[7] Fundação Nacional do Índio – FUNAI: http://www.funai.gov.br (accessed 20 March 2019)

[8] J.C. Miller, Way of Death. Merchant Capitalism and the Angolan Slave Trade. 1730-1830, Madison, The University of Wisconsin Press, 1988.

[9] C.E.G. Amorim, C.C. Gontijo, G. Falcão-Alencar, N.M.O. Godinho, R.C.P. Toledo, M.A.F. Pedrosa, M.R. Luizon, A.L. Simões, M.N. Klautau-Guimães, S.F. Oliveira, Migration in Afro-Brazilian Rural Communities: Crossing Demographic and Genetic Data, Human Biology, 83 (2011) 509-521. doi:10.3378/027.083.0405

[10] R. S. A, Anjos, A. Cypriano, Quilombolas – tradições e cultura da resistência. Aori Comunicações. São Paulo: Petrobras, 2006.

[11] Secretaria Nacional de Políticas de Promoção da Igualdade Racial – SEPPIR

http://seppir.gov.br (accessed 18 december 2017)

[12] L. Kimura, K. Nunes, L.I. Macedo-Souza, J. Rocha, D. Meyer, R.C. Mingroni-Netto, Inferring paternal history of rural African-derived Brazilian populations from Y chromosomes, Am. J. Hum. Biol. (2016) 1-11. doi:10.1002/ajhb.22930

[13] N.P.C. Santos, E.M. Ribeiro-Rodrigues, A.K.C. Ribeiro-dos-Santos, R. Pereira, L. Gusmão, A. Amorim, J.F. Guerreiro, M.A. Zago, C. Matte, M.H. Hutz, S.E.B. Santos, Assessing Individual Interethnic Admixture and Population Substructure Using a 48–Insertion-Deletion (INSEL) Ancestry-Informative Marker (AIM) Panel, Human Mutation, Vol. 31, No. 2, 184–190 (2009). doi:10.1002/humu.21159

[14] L.G.L. Maciel, E.M.R. Rodrigues, N.P.C. Santos, A.R. Santos, J.F. Guerreiro, S. Santos, Afro-Derived Amazonian Populations: Inferring Continental Ancestry

and Population Substructure, Human Biology, Vol. 83, No. 5, pp. 627-636 (2011). doi: 10.3378/027.083.0504

[15] L. Kimura, E.M. Ribeiro-Rodrigues, M.T.B.M. Auricchio, J.P. Vicente, S.E.B. Santos, R.C. Mingroni-Netto, Genomic Ancestry of Rural African-Derived Populations from Southeastern Brazil, American Journal Of Human Biology 25:35–41 (2013). doi:10.1002/ajhb.22335

[16] K. Abe-Sandes, A. Wilson, J.R. Silva, M.A. ZAGO, Heterogeneity of the Y Chromosome in Afro-Brazilian Populations, Human Biology 76(1) (2004) 77-86. doi:10.1353/hub.2004.0014

[17] G.G.B.L. Ribeiro, K. Abe-Sandes, R.S.S. Barcelos, et al., Who were the male founders of rural Brazilian Afro-derived communities? A proposal based on three populations, Ann. Hum. Biol. 38 237–240 (2011). doi:10.3109/03014460.2010.500471.

[dataset] [18] H.M. Cann, C. Toma, D. Cazes, L. Legrand, M.-F. Morel, V. Piouffre, L. Bodmer, J. Bodmer, W.F. Bonne-Tamir, B.A. Cambon-Thomsen, et al., A human genome diversity cell line panel, Science 296 (5566) (2002) 261–262. doi:10.1126/science.296.5566.261b

[dataset] [19] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68-74. doi:10.1038/nature15393.

[20] T. Hünemeier, C. Carvalho, A.R. Marrero, F.M. Salzano, S.D.J. Pena, M.C. Bortolini, Niger-Congo Speaking Populations and the Formation of the Brazilian Gene Pool: mtDNA and Y-Chromosome Data, American Jounal of Physical Anthropology 133:854–867 (2007). doi:10.1002/ajpa.20604

[21] N.M.O. Godinho, C.C. Gontijo, M.E.C.G. Diniz, et al., Regional patterns of genetic admixture in South America. Forensic Sci. Int.: Genet. Supplement Series 1 329 –330 (2008). doi:10.1016/j.fsigss.2007.10.069

[22] A. Blanco-Verea, J.C. Jaime, M. Brión, A. Carracedo, Y-chromosome lineages in native South American population, Forensic Science International: Genetics 4 187–193 (2010). doi:10.1016/j.fsigen.2009.08.008

[23] A. Ibarra, A. Freire-Aradas, M. Martínez, M. Fondevila, G. Burgos, M. Camacho, H. Ostos, Z. Suarez, A. Carracedo, S. Santos, L. Gusmão, Comparison of the genetic background of different Colombian populations using the SNPforID 52plex identification panel, Int J Legal Med 128(1):19-25 (2013). doi:10.1007/s00414-013-0858-z

[24] T. Heinz, V. Álvarez-Iglesias, J. Pardo-Seco, P. Taboada-Echalar, A. Gómez-Carballa, A. Torres-Balanza, O. Rocabado, A. Carracedo, C. Vullo, A. Salas, Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians, Forensic Science International: Genetics 7 537–542 (2013). doi:10.1016/j.fsigen.2013.05.012

[25] F.S.N. Manta, R. Pereira, R. Vianna, A.R.B. Araújo, D.L.G. Gitaí, D.A. Silva, E.V. Wolfgramm, I.M. Pontes, J.I. Aguiar, M.O. Moraes, E.F. Carvalho, L. Gusmão, Revisiting the Genetic Ancestry of Brazilians Using Autosomal AIM-Indels, PLoS ONE 8(9): e75145 (2013). doi:10.1371/journal.pone.0075145

[26] A. Freire-Aradas, Y. Ruiz, C. Phillips, O. Maroñas, J. Söchtig, A.G. Tato, J. A. Dios, M.C. Cal, V.N. Silbiger, A.D. Luchessi, M.A. Chiurillo, A. Carracedo, M.V. Lareu, Exploring iris colour prediction and ancestry inference in admixed populations of South America, Forensic Science International: Genetics 13 3–9 (2014). doi:10.1016/j.fsigen.2014.06.007

[27] F. Moreno, A. Freire-Aradas, C. Phillips, M. Fondevila, A. Carracedo, M.V. Lareu, SNP variation with latitude: Analysis of the SNPforID 52-plex markers in north, mid-region and south Chilean populations, Forensic Science International: Genetics 10 12–16 (2014). doi:10.1016/j.fsigen.2013.12.009

[28] C. Xavier, J.J. Builes, V. Gomes, J.M. Ospino, J. Aquino, W. Parson, A. Amorim, L. Gusmão, A. Goios, Admixture and Genetic Diversity Distribution Patterns of Non-Recombining Lineages of Native American Ancestry in

Colombian Populations, PLoS One 16; 10(3):e0120155 (2015). doi:10.1371/journal.pone.0120155.

[29] L. Urbano, E.C. Portilla, W. Muñoz, C.H. Sierra-Torres, H. Bolaños, Y. Arboleda, D.P. Aguirre, L. Mendoza, V. Carmona, C.H. Afanador, M. Salgar, L. Gusmão, J.J. Builes, Ancestral genetic composition in a population of South Western Colombian using autosomal AIM-INDELS, Forensic Science International: Genetics Supplement Series 5 (2015) e189–e190. doi:10.1016/j.fsigss.2015.09.076

[30] G. Garavito, B. Martinez, J.J. Builes, D. Aguirre, L. Mendoza, C.H. Afanador, E. Egea, J. Marrugo, Indels markers set and ancestry estimates in a population sample from Atlantic Department of Colombia, Forensic Science International: Genetics Supplement Series 5 e177–e178 (2015). doi:10.1016/j.fsigss.2015.09.071

[31] G.C. Cassiano, E.J.M. Santos, M.H.T. Maia, A.C. Furini, L.M. Storti-Melo, F.M.B. Tomaz, P.C.A. Trindade, M.P. Capobianco, M.A.T. Amador, G.M.R. Viana, M.M. Póvoa, S.E.B. Santos, R.L.D. Machado, Impact of population admixture on the distribution of immune response costimulatory genes polymorphisms in a Brazilian population, Human Immunology (2015). doi:10.1016/j.humimm.2015.09.045

[32] H. Ossa, J. Aquino, R. Pereira, A. Ibarra, R.H. Ossa, L.A. Pérez, J.D. Granda, M.C. Lattig, H. Groot, E.F. Carvalho, L. Gusmão, Outlining the Ancestry Landscape of Colombian Admixed Populations, PLOS ONE 11(10):e0164414 (2016). doi:10.1371/journal.pone.0164414

[33] M. Caputo, M. A. Amador, S. Santos, D. Corach, Potential forensic use of a 33 X-InDel panel in the Argentinean population, Int J Legal Med 131:107–112 (2017). doi:10.1007/s00414-016-1399-z

[34] G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, S. Myers, A Genetic Atlas of Human Admixture History, Science 343 (2014) 747-751. doi:10.1126/science.1243518

[35] H. Tang, S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E.G. Burchard, N.J. Risch, Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans, The American Journal of Human Genetics, 81 (2007) 626-633. doi:10.1086/520769.

[36] C. Tian, D.A. Hinds et al, A genomewide single nucleotide polymorphism panel for Mexican American admixture mapping, Am J Hum Genet 80 (2007) 1014-1023. doi:10.1086/513522.

[37] M.A. Jobling, P. Gill, Encoded evidence: DNA in forensic analysis, Nature Reviews Genetics, 5 (2004) 739–751. doi:10.1038/nrg1455.

[38] M.D. Shriver, R.A. Kittles, Genetic ancestry and the search for personalized genetic histories, Nature Reviews Genetics 5 (2004) 611–618. doi:10.1038/nrg1405.

[39] R. Chakraborty, M.I. Kamboh, R.E. Ferrell, Unique alleles in admixed populations: a strategy for determining hereditary population differences of disease frequencies, Ethn Dis 1 (1991) 245-256.

[40] Vila Real, R.N.S. Cultura e Currículo: Um estudo da escola Kalunga. 1996. Dissertação (Mestrado) - Universidade Federal de Goiás, Goiânia, 1996.

[41] S. Brasileiro, J.A.L. Sampaio, Sacutiaba e Riacho de Sacutiaba: uma comunidade negra rural no oeste baiano, in: Quilombos: Identidade Étnica e Territorialidade, E.C. O'Dwyer, ed. Rio de Janeiro: Fundação Getúlio Vargas and Associação Brasileira de Antropologia, 2002, pp.83–108.

[42] J.M. Arruti, Mocambo: Antropologia e história do processo de formação quilombola, Bauru, SP: Edusc, 2006.

[43] R. Pereira, C. Phillips, N. Pinto, C. Santos, S.E. dos Santos, A. Amorim, A. Carracedo, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) :e29684. doi:10.1371/journal.pone.0029684.

[44] L.G.L. Excoffier, S. Schneider, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol Ecol Resour 10: 564-567 (2010). doi:10.1111/j.1755-0998.2010.02847.x.

[45] F. Mendes, C.C. Gontijo, Kpop: A Python package for population genetics analyses, Forensic Science International: Genetics Supplement Series, 6 (2017) e407-e409. doi:10.1016/j.fsigss.2017.09.159.

[46] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.

[47] N.M. Kopelman, J. Mayzel, M. Jakobsson, N.A. Rosenberg, I.Mayrose, Itay, CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K, Molecular Ecology Resources 15(5) (2015) 1179-1191. doi:10.1111/1755-0998.12387.

[48] G. Evanno, S. Regnaut, J. Gould, Detecting the number of clusters of individuals using the software structure: a simulation study, Mol. Ecol. 14 (2005) 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x.

[49] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M.V. Parra, W. Rojas, C. Duque, N. Mesa, et al., Reconstructing Native American population history, Nature 488 (7411) (2012) 370–374. doi:10.1038/nature11258

[50] A.R. Martin, C.R. Gignoux, R.K. Walters, G.L. Wojcik, B.M. Neale, S.D. Gravel, J. Mark, C.D. Bustamante, E.E. Kenny, Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations, The American Journal of Human Genetics, 100(4) (2017) 635 – 649. doi:10.1016/j.ajhg.2017.03.004.

[51] S. Wang, N. Ray, W. Rojas, M.V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzoti, K. Hill, A.M. Hurtado, et al., Geographic Patterns of Genome Admixture in Latin American Mestizos, PLoS Genet 4(3) (2008) e1000037. doi:10.1371/journal.pgen.1000037

[52] M.E.C.G. Diniz, Fecundidade e Genética em Kalunga: busca de associação entre dados demográficos e marcadores moleculares num remanescente de quilombo brasileiro, Masters Assay, Health Sciences Graduate Program, University of Brasília, Brazil, (2008).

[53] H.I. Novion, Sobrenomes e Demografia em quatro comunidades Remanescentes de Quilombos Brasileiros, Masters Assay, Animal Biology Graduate Program, Institute of Biological Sciences, University of Brasília, Brazil, (2003).

[54] R.R. Lima, A contribuição masculina na formação de três remanescentes de quilombos do Vale do São Francisco (nordeste do Brasil) avaliada por marcadores do cromossomo Y, Animal Biology Graduate Program, Institute of Biological Sciences, University of Brasília, Brazil, 2002.

[55] I.R. Souza, L. Culpi, Valongo, genetic studies on an isolated Afro-Brazilian community. Genet Mol Biol. 28(3) (2005) 402-406. doi:10.1590/S1415-47572005000300012.

[56] T.M.B.M. Machado, Ancestralidade em Salvador-BA. Masters Assay, Biotechnology in Health and Investigative Medicine, Centro de Pesquisa Gonçalo Moniz, FIOCRUZ, Bahia, Brazil, 2008.

[57] S.F. Oliveira, G.G.B.L. Ribeiro, L.B.Ferreira, M.N. Klautau-Guimarães, A.L. Simões, Reconstrucción Histórica de Poblaciones Afro-Descendientes Aisladas de Brasil: el Contraste entre las Contribuciones Masculina y Femenina, in: Sociedad Española de Antropología Física (Org.), Diversidad Biológica y Salud Humana, 1ed, Múrcia: Quaderna Editorial, Spain, 2006, pp. 203-209.

[58] A. Ruiz-Linares, K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, W. Arias, M. Fuentes, M. Pizarro, P. Everardo, F. de Avila, et al., Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. PLoS Genet 10(9) (2014) e1004572. doi:10.1371/journal.pgen.1004572.

[59] T.C. Lins, R.G. Vieira, B.S. Abreu, D. Grattapaglia, R.W. Pereira, Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs, Am. J. Hum. Biol., 22 187–192 (2010). doi:10.1002/ajhb.20976

[60] S.D.J. Pena, G. Di Pietro, M. Fuchshuber-Moraes, J.P. Genro, M.H. Hutz, F.D.S.G. Kehdy, et al., The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected, PLoS ONE 6(2) e17063 (2011). doi:10.1371/journal.pone.0017063

[61] S.R. Giolo, J.M.P. Soler, S.C. Greenway, M.A.A. Almeida, M. de Andrade, J.G. Seidman, C.E. Seidman, J.E.Krieger, A.C. Pereira, Brazilian urban population genetic structure reveals a high degree of admixture, Eur J Hum Genet 20(1) 111-6 (2012). doi:10.1038/ejhg.2011.144.

[62] M.M.S.G. Cardena, A. Ribeiro-dos-Santos, S. Santos, A.J. Mansur, A.C. Pereira, C. Fridman, Assessment of the Relationship between Self-Declared Ethnicity, Mitochondrial Haplogroups and Genomic Ancestry in Brazilian Individuals, PLoS ONE 8(4): e62005 (2013). doi:10.1371/journal.pone.0062005

[63] S.M. Callegari-Jacques, D. Grattapaglia, F.M. Salzano, S.P. Salamoni, S.G. Crossetti, M.E. Ferreira, M.H. Hutz, Historical Genetics: Spatiotemporal Analysis of the Formation of the Brazilian Population, American Journal Of Human Biology, 15:824–834 (2003). doi:10.1002/ajhb.10217

[64] M.A.F. Pedrosa, S.F. Oliveira, Mixtura étnica en poblaciones afro-descendientes semi-aisladas de Brasil, Anales del VIII Congreso de la Asociación Latinoamericana de Antropología Biológica, Caracas, Venezuela (2004).

[65] M.A.F. Pedrosa, Composição Genética de Quatro Populações Remanescentes de Quilombo do Brasil com Base em Microssatélites e Marcadores de Ancestralidade, Masters Assay, Molecular Biology Graduate Program, Institute of Biological Sciences, University of Brasília, Brazil (2006).

[66] C. Santos, C. Phillips, F. Oldoni, J. Amigo, M. Fondevila, R. Pereira, A. Carracedo, M.V Lareu, Completion of a worldwide reference panel of samples for an ancestry informative Indel assay, FSI: Genetics 17 (2015) 75–80. doi:10.1016/j.fsigen.2015.03.011

Figure captions

Figure 1. Two components PCA for reference, comparison and study samples. a) PCA for dispersion of the study set (Mulaló, Kalunga and Sacutiaba) in relation to reference (Africa, Europe and Indigenous America) and comparison (Northern Africa, Central and South Asia, and the Middle East) groups. **b)** PCA for dispersion of the study sample in relation to the reference populations . **c)** PCA for dispersion of the study set in relation to each population that constitute the African parental group.

Figure 2. Structure analysis for reference, comparison and study samples. Structure analysis considering k:3-5 for reference (AFR: Africa, EUR: Europe, IAM: Indigenous American); comparison (EAS: East Asia, CSA: Central and South Asia, NAF: Northern Africa, MID: Middle East), and study sample (Brazilian *quilombos* Kalunga: KAL and Sacutiaba: SAC, and the Colombian *comunidad negra* Mulaló: MUL).

Figure 3. Ancestry analysis. a) Bar plot for k:3 (best k) considering the three parental populations (Africa, Europe and Indigenous America) most likely to have contributed to the gene pools of Mulaló, Kalunga and Sacutiaba. **b)** Pie charts of population ancestry estimates. AFR: Africa, EUR: Europe, IAM: Indigenous America, MUL: Mulaló, KAL: Kalunga, SAC: Sacutiaba.

Figure 4. Distribution of Individual ancestry estimates. Violin plots showing ranges of individual ancestry estimates for each parental group (Africa, Europe, and Indigenous America) in the study populations Mulaló, Kalunga, and Sacutiaba.

Supplementary material

(available at github.com/ninacalorina/supplementary-material-thesis)

Table S1. Pairwise Fst estimates for the Quilombos and the CEPH-HGDP data set (African, European, American, East Asian, Central and South Asian, and Middle Eastern) considered during the selection of the parental populations.

Table S2. Descriptive population parameters estimated for each of the 46 AIM Indels in Kalunga, Sacutiaba and Mocambo. **FOOTNOTE:** HWE: Hardy-Weinberg Equilibrium. P-val: P-value. S.E.: standard error. Hexp: expected heterozygosity. Hobs: observed heterozygosity. In bold: P-values < 0.05. Allele labeling: 1: short allele, 2: long allele, 3: novel mutation on either the long or short allele. Sequence patterns can be found on Pereira et al 2012 [43] and Santos et al 2015 [72].

Table S3. Linkage disequilibrium analysis for each pair of loci in Kalunga, Sacutiaba and Mocambo. **FOOTNOTE:** Upper diagonal: standard error. Lower diagonal: P-Values. In bold: $p < 0.05$ considering the standard errors. Bordered cells: P + S.E. ≥ 0.05.

Table S4. Individual ancestry estimates and population assignment for each sample from Kalunga, Sacutiaba and Mocambo based on data from the 46 AIM Indels panel.

FOOTNOTE: Population assignment was done by Naive Bayes classification on K-POP 0.1. The training sets were compiled from CEPH-HGDP and consisted of: Africa (AFR), Europe (EUR), America (AME), and East Asia (EAS). Two different analyses were performed, one including East Asia (w/ EAS), and another not including EAS (w/o EAS). Ancestry estimates were done on STRUCTURE 2.3 considering the four populations set. In bold: 1) Highest parental contribution in each individual's ancestry and 2) population assignment, when different from the main ancestry component.

CAPÍTULO 2

Ancestralidade e Estrutura Genética em Populações Afro-derivadas da América do Sul

Neste capítulo, os quilombos Kalunga e Sacutiaba e a *comunidad negra* Colombiana Mulaló são analisados para o sistema de AIM SNPs SNP*for*ID 34plex. O trabalho foi redigido na forma de um artigo a ser submetido para publicação no periódico American Journal of Human Biology com o título “Genetic Ancestry and Structure in African-derived populations from South America”. Apresentamos dados novos para o sistema nas populações brasileiras e compilamos dados previamente publicados da população colombiana (SNP*for*ID Browser).

Brasil e Colômbia compartilham uma história de colonização e escravização de populações indígenas americanas e africanas (essas, trazidas para as Américas com esse fim). Em ambos, apesar da introgressão de variabilidade genética proveniente de outras regiões do mundo, a mistura é majoritariamente tri-híbrida entre populações indígenas americanas, européias (especialmente ibéricas) e africanas subsaarianas (Da Silva, 2012). Há que se ressaltar que os africanos levados para a Colômbia não pertenciam necessariamente aos mesmos grupos étnicos que foram levados para o Brasil (Landers et al., 2015). Isso se deveu tanto a mudanças nas relações políticas entre países colonizadores no continente africano, quanto à licença para importar escravos pelo porto de Cartagena de Índias, que se

alternou ao longo do tempo entre portugueses, espanhóis, holandeses, genoveses, franceses e ingleses (Arocha, 1998; Azopardo, 1987).

No Brasil, o movimento quilombista alcançou as primeiras leis de reconhecimento e reparação na década de 80 do século XX. Os entendimentos jurídico e antropológico hoje consideram quilombos como comunidades derivadas diretamente ou não dos quilombos pré-abolição da escravatura (por isso também definidos como históricos). Na Colômbia, os movimentos Afrocolombianos ou de *Negritud* alcançaram leis de reconhecimento e reparação na década de 90, com o *Ato Transitorio 55 de 1991*, incorporado à *Ley 70 de 1993*:

> "*La presente ley tiene por objeto reconocer a las comunidades negras que han venido ocupando tierras baldías en la zonas rurales ribereñas de los ríos de la Cuenca del Pacífico, de acuerdo con sus prácticas tradicionales de producción, el derecho a la propiedad colectiva, de conformidad con lo dispuesto en los artículos siguientes. Así mismo tiene como propósito establecer mecanismos para la protección de la identidad cultural y de los derechos de las comunidades negras de Colombia como grupo étnico, y el fomento de su desarrollo económico y social, con el fin de garantizar que estas comunidades obtengan condiciones reales de igualdad de oportunidades frente al resto de la sociedad colombiana. (art. Transitorio 55 de la Constitución Política – Ley 70 de 1993).*"

que por sua vez, define como *comunidad negra*:

> *"… un conjunto de familias de ascendencia afrocolombiana que posee una cultura propia, comparte una historia, y [que] tiene sus propias tradiciones y costumbres dentro de la relación campo-poblado, que revela y conserva conciencia de identidad que la distingue de otros grupos étnicos… (Art.2. de la Ley 70 de 1993)."*

De acordo com o ICODER (Instituto Colombiano de Desarrollo Rural, 2006; acessado em junho de 2019), existiam no ano de 2006, 60 mil famílias vivendo em

155 *comunidades negras*. O Valle del Cauca, departamento onde se localiza Mulaló, tem 29 comunidades reconhecidas e habitadas por mais de 6 mil famílias (www.etnoterritorios.org, acessado em junho de 2019). Há que se ressaltar a diferença estabelecida na própria *Ley 70 de 1993* entre *palenque* e *comunidad negra*: uma comunidade *palenquera* deve ter continuidade material com *palenques* pré-abolição e seriam, portanto, análogos aos quilombos históricos. Hoje, apenas a comunidade de Palenque de San Basílio é assim reconhecida, enquanto as demais comunidades (análogas a quilombos pós-abolição), incluindo Mulaló, são chamadas *comunidades negras*.

Brasil e Colômbia compartilham dois fatores que justificam a análise paralela de *quilombos* e *comunidades negras*. Primeiramente, uma história colonial que gerou miscigenação entre povos indígenas colonizados, europeus colonizadores e africanos escravizados. Em seguida, a própria existência dos quilombos e *comunidades negras*: populações de ancestralidade africana predominante, relacionadas aos movimentos campesinos, e relacionadas à resistência histórica ao sistema escravista e à estrutura social derivada desse sistema (Da Silva, 2012). Além disso, nos dois países a ambiguidade das relações raciais gerou mitos de igualdade - a chamada "democracia racial" no Brasil e o "*café con leche*" na Colômbia (Da Silva, 2012). Esses mitos, assim como a analogia falaciosa do "*melting pot*" utilizada para descrever populações Latino Americanas, sugerem homogeneidade e mascaram relações sociais desiguais e conflitos fundiários. Assim, o objetivo geral do capítulo é descrever estrutura e ancestralidade genéticas e avaliar semelhanças e diferenças entre os quilombos Kalunga e Sacutiaba e a *comunidad negra* Mulaló para, desta forma, somar ao conhecimento da diversidade das populações humanas do continente sul americano.

RESUMO

A América Latina tem uma história intrincada que se reflete em sua composição genética. Comunidades com ancestralidade africana prevalente, originadas como forma de resistência à escravidão ou como consequência do colapso das estruturas sociais derivadas desse sistema, existem ainda hoje em vários países. Exemplos são os quilombos no Brasil e as *comunidades negras* na Colômbia. Apesar da crescente literatura descrevendo populações miscigenadas da América do Sul, populações não urbanas e não indígenas ainda são sub-representadas em bancos de dados públicos adequados para a análise de ancestralidade. Neste capítulo, buscamos adicionar dados e conhecimento sobre populações afro-derivadas da América do Sul. Focamos nossas análises na descrição de ancestralidade e estrutura, além de contrastar nossas populações de estudo uma à outra, a suas parentais presumidas e a populações de outros continentes. Analisamos o sistema SNP*for*ID 34plex, capaz de diferenciar africanos, leste-asiáticos/indígenas americanos e europeus - as populações que contribuiram majoritariamente para a formação do continente. Geramos dados para os quilombos Kalunga e Sacutiaba e compilamos dados da comunidade negra Mulaló do SNP*for*ID *browser*. Nossos dados mostraram que as três populações tem como principal componente de ancestralidade a parental africana, seguida pela européia e pela indígena americana em proporções compatíveis com suas histórias, localizações geográficas e dinâmica demográfica atual. A análise não revelaou diferenças marcantes entre nossas populações de estudo e as populações africanas que compuseram a amostra de referência, apesar da composição miscigenada das três. Em conclusão, as populações em estudo não são idênticas em ancestralidade africana e diferem substancialmente nas proporções das outras parentais e na distribuição da ancestralidade individual. Nossos resultados reiteram a diversidade do continente, que se revela mesmo dentro de um grupo com paralelos fortes em história de formação e origem.

Palavras-chave: Ancestralidade; AIM; Mistura; América Latina; Quilombo; Comunidade negra; Brasil; Colômbia

Genetic Ancestry and Structure in African-derived populations from South America

Carolina Carvalho Gontijo*[a,b], Ana Freire Aradas[a], Maria Victoria Lareu[a], Ángel Carracedo[a], Christopher Phillips[a], Silviene F Oliveira*[a,b]

[a] Unidade de Xenética, Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain

[b] Human Genetics Laboratory, Institute of Biological Sciences, University of Brasília, Brazil

* Corresponding author: Carolina Carvalho Gontijo

carolinacarvalhogontijo@gmail.com

Laboratório de Genética Humana, Instituto de Ciências Biológicas, Bloco F, 2º andar. Campus Universitário Darcy Ribeiro, Asa Norte, Brasília – DF, Brazil.

CEP 70.910-900

Running title: African ancestry in South American populations

ABSTRACT

Objectives

Latin America has an intricate history, which is reflected on its genetic composition. Communities with prevailing African ancestry (either formed as a way of resisting slavery, or as a consequence of that system) exist to this day. Examples are *quilombos* in Brazil, and *palenques* and *comunidades negras* in Colombia. In spite of the growing literature describing admixed South American populations, non-urban and non-indigenous populations are underrepresented in databases suitable for ancestry analysis. Here, we add to the body of knowledge about African-derived populations from South America, describing ancestry and structure and contrasting our study populations to one another, to their presumptive parentals, and to populations from other continents.

Methods

We analysed three African-derived populations from Colombia and Brazil for the system SNP*for*ID 34plex, capable of differentiating Africans, Europeans and East Asians/Indigenous Americans. Data was generated for the *quilombos* Kalunga and Sacutiaba, and compiled from SNP*for*ID Browser for the *comunidad negra* Mulaló. Statiscal analysis described basic population genetics parameters, ancestry and structure.

Results

The three populations have a major African ancestry, followed by European and Indigenous American components compatible with their history, location, and current demographic dynamics. PCA did not reveal marked differences between our study sample and African populations, regardless of their admixed composition.

Conclusions

Mulaló, Kalunga and Sacutiaba are not identical in African ancestry and differ substantially in structure and contribution from the other parentals. Our results reiterate the continent's diversity, even within a group with such strong parallels in history and origin.

Keywords: Ancestry; African-Derived; Brazil; Colombia; South America

1. Introduction

“Melting pot” has been a common, yet fallacious analogy to describe Latin American populations. The expression erases minorities and implies integration, homogeneity and full admixture, while leaving out the diversity that characterizes contemporary populations from the continent. Examples of that diversity abound in the literature on population genetics (Adhikari et al. 2017; Salzano and Sans, 2014; Ruiz-Linares et al. 2014; Wang et al. 2008) which testify to rich genetic ancestry, structure, and composition, and expose the impropriety of the aforementioned analogy.

The first human populations to set foot in the Americas migrated from East Asia through Beringia, a landmass now mostly submerged. Those first settlers were likely isolated from their own Siberian ancestors for at least 5 kya (Gómez-Carballa et al. 2019) before entering the continent. Archaeological, linguistic and genetic evidence point to an entering time at 15-18 kya ago (Pinotti et al. 2019). In the continent, the rising original populations went through different stories of contact and isolation, expansion and retraction, founder effect, drift and selection leading to the diversity seen in indigenous populations today. Later, during the 15th century, layers of complexity were added to that scenario by the beginning of the Age of Discovery and the Atlantic Slave Trade. In South America, those massive migratory movements introduced variability mostly from Iberian Europe and Sub-Saharan Africa into the continent (Salzano e Freire-Maia, 1967). Recent migration from other European

regions, the Middle East and East Asia, for example, rendered admixture further complex.

Brazil and Colombia were home to the largest ports of entry where enslaved Africans were disembarked in the Americas (Landers et al. 2015) until the 19th century, and are now home to the largest populations of African descent in South America (Da Silva, 2012). In every country where slavery existed, people seeked ways to resist it. A common way for enslaved Africans and their descendents (who were often admixed) to resist captivity was escaping and settling in secluded locations. All over the continent, communities of variable sizes and degrees of isolation were formed, constituting safe havens that share a history of resistance against oppression and fight for freedom and dignity (Gomes, 2015; Da Silva, 2012). Many such populations have remained genetically isolated to some extent from surrounding populations and exist to this day in Brazil (where they are called *mocambos* or *quilombos* - Vila Real, 1996; Neme & Andrade, 1987) and Colombia (where they are called *palenques* or *comunidades negras* - Landers et al. 2015; Da Silva, 2012).

Genetic ancestry estimation, *i.e.* the quantification of variants of different origins in an individual genome (Shriver and Kittles, 2004), is key to: 1. inferring history and relationships among human populations, kinship and demographic events (exemplified in Gontijo et al. 2018; Elhaik 2014; Hellenthal et al. 2014; Wang et al. 2008); 2. describing population structure for gene mapping analysis (Shriver et al. 1997) and guiding case-control studies (Pritchard and Donelly, 2001); 3.

restricting the scope of a search in the forensic genetics context (Jobling e Gill, 2004).

In spite of the growing literature describing admixed South American populations, non-urban and non-indigenous populations are underrepresented in databases of genetic markers adequate for ancestry and structure analysis (Gontijo et al. 2018; Manta et al. 2013). The 1000 Genomes (The 1000 Genomes Project Consortium, 2015) and CEPH (Cann et al. 2002), for instance, include two non-indigenous populations, namely Colombians from Medellín (CLM) and Peruvians from Lima (PEL), both urban. That issue is currently being remedied by studies directed at the continent (exemplified in Gontijo et al. 2018; Ossa et al. 2016; De Moura et al. 2015; Manta et al. 2015; Xavier et al. 2015; Moreno et al. 2014; Freire-Aradas et al. 2014; Heinz et al. 2013) and by databases beginning to compile a wider sampling of the continent (of which SNP*for*ID Browser is an example; Amigo et al. 2008).

Here, we describe ancestry and genetic structure of three African-derived South American populations: the Colombian *comunidad negra* Mulaló and the Brazilian *quilombos* Kalunga and Sacutiaba. Our goal is to add to the body of knowledge about African-derived populations from South America, thus presenting a more accurate description of the continent's variability. Furthermore, we evaluate whether the historical and anthropological parallels shared by the *quilombos* and *comunidades negras* under analysis are reflected on their genetic ancestry.

2. Material and Methods

2.1 Study Populations

Our study populations comprise three rural African-descendant populations from Brazil and Colombia, numbering 144 samples. Published data were compiled from SNP*for*ID Browser (available at http://spsmart.cesga.es/snpforid.php; Amigo et al. 2008) for the Colombian *comunidad negra* Mulaló (n = 42), and new data were generated for the Brazilian *quilombos* Kalunga (n = 72) and Riacho de Sacutiaba e Sacutiaba - herein Sacutiaba (n = 30). Over the years since sample collection, the results of our previous analyses and of the health assessment using the collected samples (blood, urine and stool tests) were communicated to the *quilombos* Kalunga and Sacutiaba, in addition to educational actions regarding health.

The Mulaló (MUL) live in relative isolation in the *Departamento* Valle del Cauca in Colombia, where gold mining and coffee and sugar cane plantations relied heavily on African slave workforce (Landers et al. 2015). Mulaló has its origins in the 16th century in the area of a former slave trading farm, and in 2015 was inhabited by 1800 people (2015 census; DANE, 2015).

Kalunga (KAL) is the largest *quilombo* today in what regards both the land they own (75,233 ha) and population size (estimated 5,300 people), according to the 2010 census (IBGE, 2010) and Fundação Cultural Palmares (FCP, 2017). The founding Kalungas were either freed or escaped slaves that had been brought to

central Brazil to work on gold mining in the late 18th century. Migration rates into Kalunga are very low (Paiva, 2017).

Sacutiaba (SAC) is an extended family of 200 people (Brasileiro and Sampaio, 2002) structured into smaller households. According to the scarce register about SAC, the community was founded by runaway slaves from the neighboring state of Minas Gerais around 200 years ago (Brasileiro and Sampaio, 2002). Sacutiaba shares commercial and social bonds with surrounding populations, and has an estimated immigration rate of 30% (Amorim et al. 2011).

2.2 Sample collection and ethical concerns

Collection and analysis of samples from the Brazilian Kalunga and Sacutiaba followed pertinent ethics guidelines and regulations for human research (project approved by the Research Ethics Committee of the Faculty of Health Sciences at University of Brasilia - CEP-FS/UnB 030/2002 and 151/07). All participants were volunteer and informed about research goals and confidentiality of data, and signed instruments of consent previously to collection.

2.3 Genetic markers

Our study examines a panel comprised by 34 AIM SNPs: SNP*for*ID 34plex (Phillips et al. 2007; revised in Fondevila et al. 2013). This set was designed to be complemented by panels adequate for specific admixture contexts: Pacifiplex (for Oceanic admixture; Santos et al. 2016), Eurasiaplex (for differentiating Europeans from South Asians; Phillips et al. 2013), and PIMA (for detecting the IAM component). By itself, the system is able to detect African, European, and East

Asian/Indigenous American components (Fondevila et al. 2013) - the most numerically relevant contributors to admixture in South America -, and is thus suitable for analysing populations from the continent.

2.4 DNA extraction and SNP genotyping

DNA was extracted from buffy coat using the blood genomicPrep Mini Spin Kit (GE Healthcare) and quantified in a NanoDrop 1000™.

Genotyping was modified from the SNaPshot® protocol described in Fondevila et al. 2013, consisting of multiplex PCR amplification, single-base extension and genotyping by capillary electrophoresis. Specific fragments were amplified by single multiplex reactions in 7 uL containing 1-10 ng DNA; primer mix (0.11-1.46 mM each as per Phillips et al 2007), dNTP mix (10 mM), AmpliTaq™ Gold (0.5 U), PCR buffer (0.690 µL), $MgCl^2$ (25 mM), BSA (2.016 µg), milli Q water up to reaction volume. Amplification conditions were: initial denaturation at 95°C for 15 min; 30 cycles at 95°C for 30 sec, 60°C for 50 sec, 65°C for 40 sec; and a final extension at 65°C for 15 min. For the removal of excess primer and dNTP, a 5 µL reaction volume containing 0.750 µL of PCR product and 0.650 µL of ExoSAP-IT GE Healthcare was incubated for 45 min at 37°C, followed by 15 min at 85°C. Excess primer and dNTP were removed by incubating 0.750 µL of PCR product and 0.650 µL of ExoSAP-IT GE Healthcare for 45 min at 37°C, followed by 15 min at 85°C to inactivate the enzyme. Single base extension (SBE) was performed in a 6 µL reaction volume containing 2.5 µL SNaPshot® mix (AB), 1.5 µL SBE primer mix (0.34-2.98 mM each, as per Phillips et al 2007), and 2.0 µL purified PCR product.

SBE amplification was performed in an AB 9700 thermal cycler following the program: 30 cycles of 96 ºC for 10 s, 55 ºC for 5 s and 60 ºC for 30 s. Excess nucleotide was removed by incubating 6.0 µL of SNaPshot® product with 1.3 µL of SAP (1 U/ml Shrimp Alkaline Phosphatase, GE Healthcare) for 80 min at 37°C, followed by 15 min at 85°C to inactivate the enzyme. SNaPshot® product was prepared for genotyping in a reaction volume of 10.2 µl containing 1 µl purified SNaPshot® product, 0.3 µl 120 LIZ size standard and 8.9 µl Hi-Di™ formamide. Capillary electrophoresis was done on ABI 3130 Genetic Analyzer in 36 cm capillary arrays with POP-4™ polymer. Allele calling was done by GeneMapper™ 3.2. Predefined size windows for each allele were determined from prior analysis.

2.5 Reference populations selection and data compilation

Reference/parental and comparison populations were chosen from The 1000 Genomes data available at SNP*for*ID Browser (Amigo et al. 2008). The first criterion for parental populations was historical plausibility; hence we focused on African (AFR), European (EUR) and Indigenous American (IAM) populations. The second criterion was continent sample homogeneity: After an exploratory Structure analysis, populations and individuals with indication of co-ancestry of 20% or more were excluded from descriptive parameter estimation and from ancestry and related analysis, but maintained for structure, PCA and pairwise F_{ST}. The core of our parental dataset (n = 702) is thus composed of samples from AFR (n = 293), EUR (n = 260) and IAM (n = 149) populations. Our comparison set includes samples from East Asia (EAS; n = 340), Central and South Asia (CSA; n = 281), Northern Africa (NAF; n = 224), and the Middle East (MID; n = 220). Detail can be found on

Supporting information - S1 Table). For assessing structure, the complete dataset (reference, comparison and study) was included. For estimating ancestry, parental and study sets were considered.

2.6 Statistical analysis

Allele frequencies, Hardy-Weinberg Equilibrium (HWE), observed (H_{obs}) and expected heterozygosity (H_{exp}), linkage disequilibrium (LD), and pairwise F_{ST} were estimated using Arlequin 3.5 (Excoffier and Lischer, 2012). Patterns of population structure were assessed by F_{ST}, PCA and STRUCTURE analysis. PCA was done using K-POP (Mendes and Gontijo, 2017). Structure and ancestry were assessed using STRUCTURE v2.3.3 (Pritchard et al. 2000). Analyses tested K:2-8, using the following parameters: burn-in of 10,000; MCMC of 100,000; Admixture model; correlated allele frequencies; 8 iterations; POPFLAG. From the output, CLUMPAK v1.1.2 (Kopelman et al. 2015) determined the optimal K (LnPr(X|K)) according to the method of Evanno (Evanno et al. 2005); and K-POP aided the plotting of cluster membership proportions. Individual ancestry distributions were plotted using Matplotlib 3.0 (Hunter, 2007) on Python 3.7. Ancestry estimations were based on cluster membership proportions identified by STRUCTURE analyses.

3. Results and discussion

African slavery was a pillar of colonial occupation and exploitation of the Americas. Plantations, mining and domestic labor relied heavily on African workforce, and thousands were forcibly brought to the continent with that purpose.

Brazil and Colombia share historical aspects that justify the parallel analysis of *quilombos* and *comunidades negras*. Those populations arose as a form of resistance against the slave system and as a consequence of the collapse of social structures generated by that system (Da Silva, 2012). Today, in both countries, many of those communities persist, and preserve an important part of the continent's history.

Here, to better understand the history and diversity of African-derived South American populations, we have analysed the panel SNP*for*ID 34plex in a set of samples from the Brazilian *quilombos* Kalunga and Sacutiaba and from the Colombian *comunidad negra* Mulaló. Beyond reporting frequencies and basic population genetics parameters, our statistical analysis aimed at ancestry and structure, and at contrasting our study populations to one another, to their presumptive parentals, and to populations from other continents.

3.1. Population-descriptive parameters

Genotypes, along with our reference and comparison datasets, are presented in S1 Table. No data was available for rs3827760 in MUL. Allelic frequencies are shown in S2 Table. Tests for Hardy-Weinberg Equilibrium showed some deviation in the three study populations: One locus is out of HWE in MUL, twelve in KAL, and five in SAC. All loci that do not meet HWE show loss of heterozygosity in the three populations, except for one locus in SAC. Hence, the high proportion of deviation could result from endogamy or underlying population structure - both of which are

reasonable explanations for the populations in question. Our data shows no major LD: 2.94% of the pairs of loci are in LD in MUL, 6.49% in KAL, 4.33% in SAC.

3.2. Patterns of structure and admixture

We assessed patterns of structure seen in our sample and of relationship seen among them and other populations by F_{ST}, PCA and STRUCTURE. Overall, our results show that MUL, KAL and SAC are more closely related to one another and to AFR than to the other populations we tested, as expected given their history, and as previously reported for different sets of markers (Gontijo et al. 2018; Guauque-Olarte et al. 2012; Amorim et al. 2011; Rondón et al. 2008).

Reference/parental and comparison sets of populations

An exploratory STRUCTURE analysis guided our selection process of reference/parental and comparison populations. Pairwise F_{ST} (S4 Table) for all populations included in that STRUCTURE analysis shows each parental population forms a cohesive group with pairwise $F_{ST} < 0.10$, as did our study populations. Based on pairwise F_{ST} and high co-ancestry estimates (AFR < 0.80) provided by the initial STRUCTURE analysis, San and Somalia were excluded from the AFR parental set. Pairwise F_{ST} within the Indigenous American group was higher than within all other continental groups tested, but still smaller than the threshold established as minimum for between-group F_{ST}. Nevertheless, because demographic dynamics in the Americas deemed indigenous populations more structured than those from other continents (Martin et al. 2017; Reich et al. 2012; Wang et al. 2008), and because F_{ST} tends to be inflated between isolated populations (Granot et al. 2016), we considered

our IAM selection adequate. Reinforcing observations from those analyses, PCAs show a compatible picture (Fig 1 a and b), in which AFR, EUR and IAM are clearly clustered apart from one another, as did our structure assessment (Fig 2).

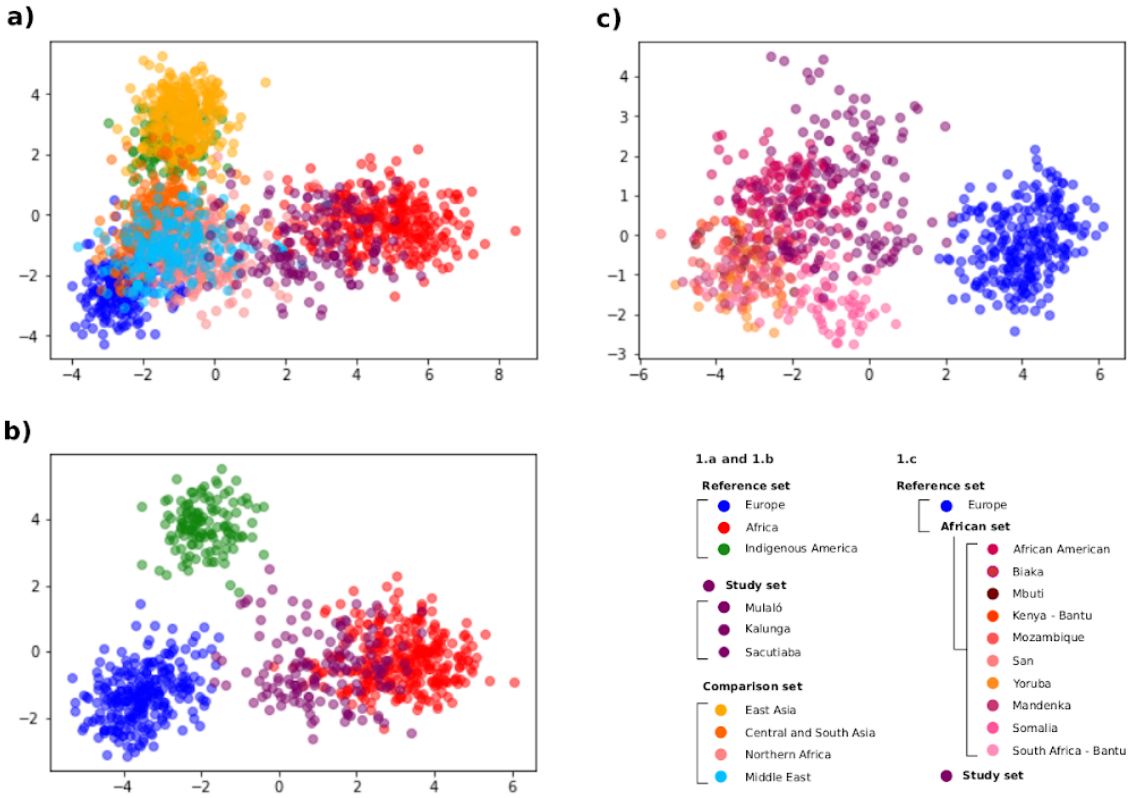


Figure 1. Two components PCA for reference, comparison and study samples. (a) PCA for dispersion of the study set (Mulaló, Kalunga and Sacutiaba) in relation to reference (Africa, Europe and Indigenous America) and comparison (Northern Africa, Central and South Asia, and the Middle East) groups. (b) PCA for dispersion of the study sample in relation to the reference populations. (c) PCA for dispersion of the study set in relation to each population that constitute the African parental group.

Our comparison set is composed of groups from Central and South Asia - CSA, North Africa - NAF; The Middle East - MID; and East Asia - EAS. STRUCTURE, F_{ST} and PCA clearly showed that CSA, NAF and MID cannot be

clearly set apart from each other, nor can they be distinguished from EUR with our set of markers. Agreeing observations have been previously reported by Rosenberg et al (2005; 2002) and Li et al (2008), whose studies found that CSA and MID are not clearly differentiated within Eurasia. Even though Eurasia is vast land mass and isolation by distance might play an important role in differentiation, its mostly continuous geography and its history of conquers and population expansions allowed for nomadic cultures to develop, for migration, and for contact and admixture among human groups. On that matter, an interesting pair of populations is Uygur, a Muslim population from northwestern China, and Adygei, a Muslim group from southeastern Russia. F_{ST} for that pair is < 0.10, even though they are presented as belonging to different continental groups in public databases. Similar results have been reported for that pair (Granot et al. 2016). That might be a consequence of the Muslim expasion that reached as far as Central Asia, leading to admixture with indigenous populations and introducing a culture shared by Uygur and Adigey. Those demographic processes and geographic proximity might have led to a smaller genetic differentiation between them. Alternatively, as proposed by Granot et al. (2016), the low differentiation indicated by F_{ST} might be simply a bias introduced in the estimates by within-population structure.

Regarding the far East of Eurasia, EAS forms a very cohesive cluster (Fig 1a). Nevertheless, PCA shows an almost complete superposition between EAS and IAM - F_{ST} estimates yield agreeing results. That was predictable, as the recent separation and direct origin of those populations make them closer in terms of genetic variability. That observation is reiterated by STRUCTURE analysis (Fig 2). SNP*for*ID 34plex is a

core system designed to differentiate Africans, Europeans and East Asians, and to be complemented by multiplexes designed for specific admixture contexts. In addition, it showed power to detect Indigenous American ancestry, according to the authors (Fondevila et al, 2013). Our data corroborates the system's power to discriminate the Indigenous American from African and European components. Nevertheless, we could not differentiate East Asians from Indigenous Americans in our dataset.

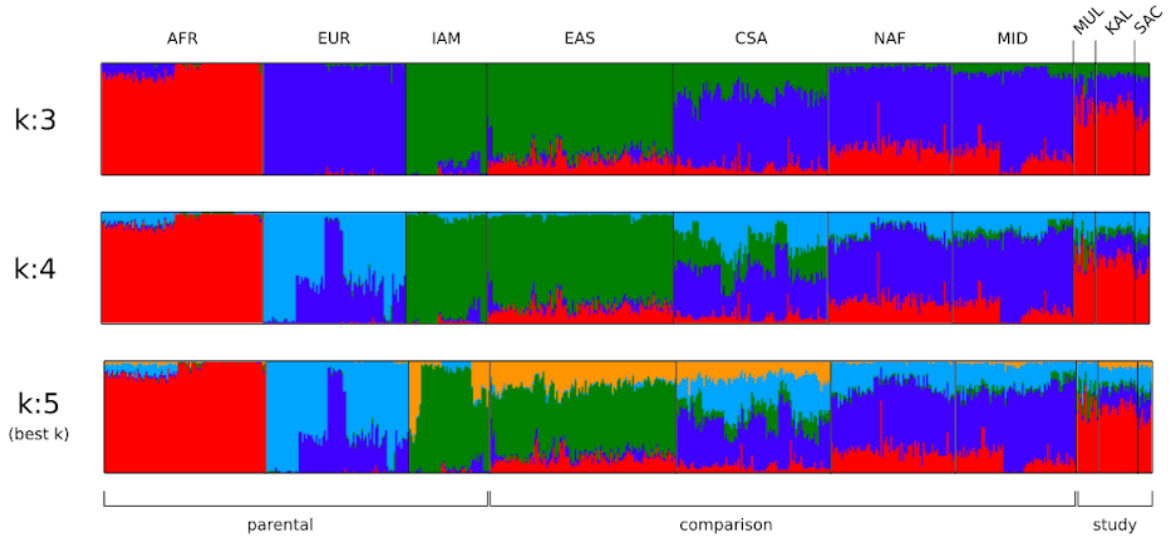


Figure 2. Unsupervised structure analysis for reference, comparison and study samples. Structure analysis considering k:3-5 for reference (AFR: Africa, EUR: Europe, IAM: Indigenous American); comparison (EAS: East Asia, CSA: Central and South Asia, NAF: Northern Africa, MID: Middle East), and study sample (Brazilian *quilombos* Kalunga: KAL and Sacutiaba: SAC, and the Colombian *comunidad negra* Mulaló: MUL)

African-Derived South American populations: ancestry and structure

Mulaló, Kalunga and Sacutiaba form a cohesive group (F_{ST} < 0.10 within the group) genetically closer to one another and to AFR populations than to any of the

other continental groups we tested (F_{ST} < 0.20 with AFR populations). At a glance, they share a strong AFR component, followed by EUR and IAM/EAS for all k tested (Fig 2). That results from admixture between AFR and non-AFR in the constitution of *quilombos* and *comunidades negras*, and is consistent with our PCA (Fig 1b) and previous observations for different sets of markers (Gontijo et al. 2018; Gontijo et al. 2014; Guauque-Olarte et al. 2012; Amorim et al. 2011; Rondón et al. 2008). Both in ancestry estimates and PCA, individuals from MUL, KAL and SAC are intermediate to the three reference/parental samples, but closer to AFR (Fig 1 a and b). In accordance with ancestry analysis (Fig 3) and demographic dynamics usually seen in *quilombos* and *comunidades negras* (Amorim et al. 2011; Ossa et al. 2016), some individuals from MUL are closer to IAM, and some from SAC, to EUR. Over 84% of our sample have AFR estimates above 50%. In MUL, 74% of our sample is above that threshold, and in SAC, 60%. In KAL, the lowest individual AFR estimate was 53.8%.

Mulaló shows the widest range and highest individual AFR estimate (38.9-80.5, s.d. 11.8%), but a population estimate (60.4%) lower than KAL. It also has the highest IAM overall contribution (17.2%; ranging from 7.8%-31.0%, s.d. 6.5%). That suggests social relationships with indigenous populations or populations with indigenous ancestry, either old or current, especially likely given the high IAM ancestry estimates seen in Valle del Cauca (Ossa et al. 2016), where MUL is located. Since MUL is reportedly isolated (Landers et al. 2015; Guauque-Olarte et al. 2010; Rondón et al. 2008), this gene flow is more likely old. Those patterns are evident in Figs 3a and 4. In addition to the widest range of co-ancestry for all three

parentals, MUL has the highest s.d. for the contribution of each parental to individual ancestry (S5 Table), also noticeable in structure analysis (Fig 3). Those observations could indicate recent foundation, substructure or ongoing gene flow. Since MUL has its origins in the 16th century and low migration rates (Guauque-Olarte et al. 2010; Rondón et al. 2008), substructure might be a more suitable explanation.

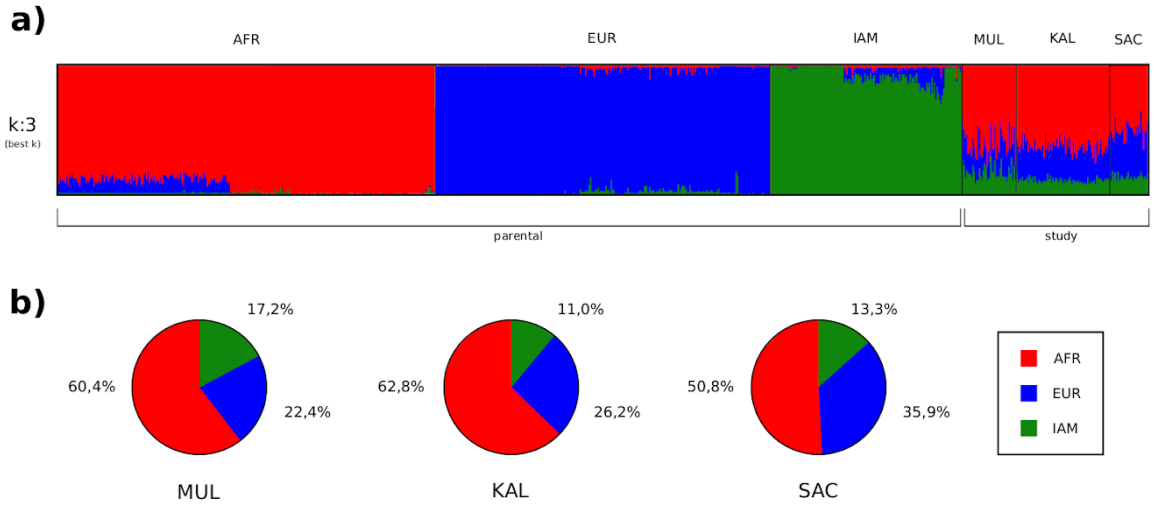


Figure 3. Ancestry analysis. (a) Bar plot for k:3 (best k) considering the three parental populations (Africa, Europe and Indigenous America) most likely to have contributed to the gene pools of Mulaló, Kalunga and Sacutiaba. (b) Pie charts of population ancestry estimates. AFR: Africa, EUR: Europe, IAM: Indigenous America, MUL: Mulaló, KAL: Kalunga, SAC: Sacutiaba.

Kalunga has the highest overall AFR contribution (62.8%). The range of individual ancestry estimates stands out in KAL (53.8 to 73.7%; s.d. 4.3%). This is the most narrow range, indicating more uniformity than what is seen in MUL and SAC (also evident in S5 Table and Figs 3.a and 4), an observation reinforced by the structure assessment in Fig 3a. KAL is the most isolated of the three populations under study, and the near absence of migration into the community (Paiva, 2017)

and relatively large size might have protected it from the effects of introgression and account for the high AFR ancestry. Field work has shown us KAL is a socially structured community, as its 5,300 inhabitants are spread over a wide, inhospitable territory, and isolation by distance could affect the distribution of variability. Nevertheless, KAL seems in fact uniform enough so that substructure is not reflected on ancestry analysis.

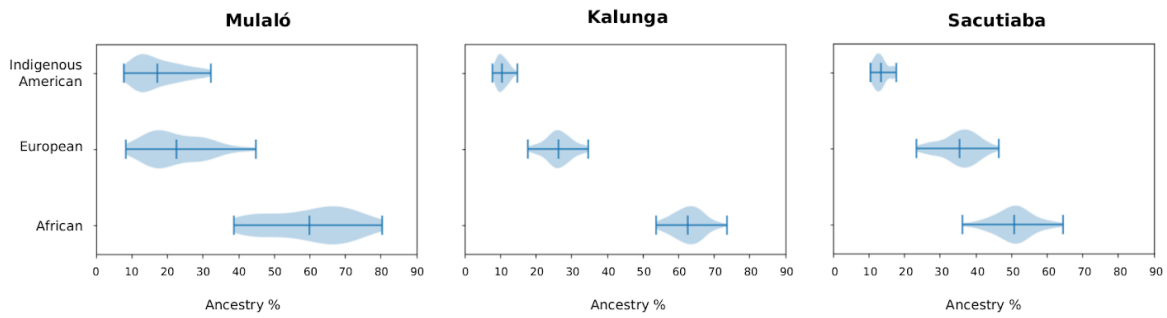


Figure 4. Distribution of Individual ancestry estimates. Violin plots showing ranges of individual ancestry estimates for each parental group (Africa, Europe, and Indigenous America) in the study populations Mulaló, Kalunga, and Sacutiaba

In Sacutiaba, 60% of our sample has AFR ancestry ≥ 50%. This population has the highest EUR contribution (35.9%), with individual admixture estimates ranging from 27.2 to 46.7% (s.d. 5.3%). The estimated migration rate into the community is 30%, and EUR ancestry is higher among immigrants than locals (Amorim et al. 2011). That could alone explain the higher EUR ancestry seen in this community. SAC has a pattern of ancestry distribution closer to KAL, but not as narrow, which is also evidenced by the low s.d. associated to ancestry estimates (S5 Table). While that could be perceived as contradictory, given the 30% imigration rate reported for SAC, some demographic observations (Amorim et al. 2011) might clarify

that pattern: First, SAC has mostly endogamic marriages. Second, the high migration rate includes “travel migration” - when individuals leave the community in search of work and later return to their home community. Third, SAC is located in a state (Bahia) where many populations have comparable ancestry estimates (De Moura et al. 2015; Machado, 2008; Abé-Sandes et al. 2004), hence the source of migration might not be so different as to introduce ancestry variability. Finally, there is no evidence of substructure. All things considered, in spite of immigration, SAC, founded over 200 years ago, is a stable population in terms of ancestry with a relatively homogeneous structure.

African-Derived South American populations on a wider context

Since the most relevant admixture component to the constitution of MUL, KAL and SAC is AFR, a PCA was performed to evaluate the relationship of our study set and AFR and to assess proximity to any specific AFR population (Fig 1C). Each African population in our reference group was considered individually. EUR was included for reference because it is the second most important component in the constitution of MUL, KAL, and SAC - according to ancestry analysis. Here again, the three study populations overlap AFR, though some individuals are closer to EUR. AFR populations could not be clearly set apart from one another by this set of markers, as evidenced by F_{ST} (S4 Table), PCA (Fig 1c) and STRUCTURE (Fig 2). African Americans, who are knowingly admixed with EUR, overlap considerably with KAL, SAC and MUL. That admixture is evident in STRUCTURE (Fig 2) and ancestry analyses (Fig 3a).

For assessing structure and evaluating the study sample on a wider reference scale, STRUCTURE analysis (k:3-5) was performed on our complete dataset (Fig 2). Optimal k is k:5. At k:3, Africans, Europeans and Indigenous Americans/East Asians are set apart as three distinct groups, with IAM and EAS forming a single group. At k:4, a new component appears in EUR that strongly separates Denmark and Orkney Islands (Scotland) from Italy. Both EUR components are shared by the other EUR populations, CSA, NAF and MID, and both appear in our study set. At k:5, a component shared by Indigenous American (IAM) and Asian (EAS and CSA) populations appears. Even at k:5 IAM and EAS were not differentiated, reiterating our observations on F_{ST} and PCA.

Kalunga and Sacutiaba in the Brazilian context

Gontijo et al. (2018) have summarized ancestry estimates for KAL and SAC from previous publications and reported data and estimates for a system of 46 AIM Indels described by Pereira et al. (2012). Their results showed high AFR ancestry estimates (67.3%), followed by intermediate EUR (24.9%) and lower IAM (7.2%) estimates in KAL. In SAC, differences were bigger, especially in EUR estimates (46,8%: 10,9% higher than what we observed). Nevertheless, the overall picture we present is similar to what the 46 AIM Indel panel showed and to the pattern that emerges from their review: high AFR estimates, followed by EUR and IAM, in that order. That is a common pattern in quilombos and some other African-Brazilian populations (examples are seen in De Moura et al. 2015; Machado, 2008; Abé-Sandes et al. 2004).

SNP*for*ID 34plex has been previously employed to estimate ancestry in a sample from São Paulo, Brazil, for a case-control study (Silbiger et al. 2012). São Paulo is an urban admixed population, and as expected, the highest ancestry estimate was EUR (58-59%), followed by AFR (28-31%), IAM (9-13%) and EAS (5%). That is compatible with what is usually seen in Brazilian populations: A predominant EUR ancestry, followed by lower AFR, reaching 30% in states were African slavery was more prominent (De Moura et al. 2015), and very low IAM components, in that order (exemplified in Gontijo et al. 2018; Ruiz-Linares et al. 2014; Manta et al. 2013; Giolo et al. 2012). That well established pattern sets *quilombos* apart from the general Brazilian population.

Mulaló in the Colombian context

A sample from urban Colombia (CLM - The 1000 Genomes, Colombians from Medellín) was also evaluated using SNP*for*ID 34plex (Fondevila et al. 2013). That work found a EUR ancestry of 63.1%, followed by IAM (13%), EAS (13.2%), AFR (6.9%), and Oceanian (3.7%). The contrast in AFR and EUR contributions to the gene pools of urban Medellín and the isolated *comunidad negra* Mulaló is stark: the former is preponderantly EUR, and the latter, AFR. Other research with Colombian populations reinforce that pattern of higher EUR contribution in most densely populated areas.

Ossa et al. (2016) analysed the same 46 AIM Indels panel (Pereira et al. 2012) and found that to be the case for Caribe, Andes and Orinoquía regions (55%, 58%, 53%, respectively), while Amazonia has higher IAM ancestry (65%) and the

Pacific coast has a higher AFR component (63%). Those results are in accordance with Colombian history: Amazonia remained more isolated, and the Pacific relied heavily on African workforce for gold mining during the 16th and 17th centuries, and later received a large migratory contingency of people of African descent after the abolition of slavery in the 19th century (Ossa et al. 2016). According to the same work, Valle del Cauca, the *departamento* where MUL is located, has an overall ancestry estimate of 55.5% EUR, 30.5% IAM, and 14% AFR. Those results set MUL further apart from populations from Valle del Cauca, as a result of its particular history.

3.3 Final remarks

Here, we described three non-urban admixed populations from South America with preponderant African ancestry. Given their history, location and relationship with surrounding populations, we assessed structure and ancestry considering three-way admixture with Africans, Europeans and Indigenous Americans. In common, our study populations share history: They were formed as a response to the African slave system that dominated the Americas up until the 19th century. Their founding peoples were mostly either escaped or in some way freed slaves and their descendants, who were likely admixed with Europeans and Indigenous Americans. Also, those populations were not completely isolated, and maintained social and commercial bonds with other groups. That shared history led to similarities in genetic composition, as they were founded by people with similar genetic backgrounds.

The estimates we present are consonant with the populations known history and demography. The three populations share a strong African ancestry. In that aspect, Mulaló and Kalunga are more alike, and the differences in contribution from other parentals are compatible with their demographic dynamics. Sacutiaba, on its turn, has a more pronounced European component which resonates with its migration rates. Overall, our results add to the knowledge about African-derived populations from South America and stress the diversity seen in the continent, even within a group with such strong parallels in history and origin.

Acknowledgments

CAPES, CNPq and FAPDF supported the research that led to this publication. The authors would like to thank the people from Kalunga and Sacutiaba, who kindly agreed to participate in our work, and field teams who made sample collection possible.

Author Contributions

CCG: designed the study, analyzed the data and drafted the manuscript. CCG, AFA and LPH collected data. MVL, AC and CP provided necessary logistical support. CP and SFO designed the study, edited the manuscript for intellectual content and provided critical comments on the manuscript. SFO also directed implementation and data collection.

Reference

Abe-Sandes, K., Wilson, A., Silva, J.R., Zago, M.A. (2004). Heterogeneity of the Y Chromosome in Afro-Brazilian Populations. *Human Biology* 76(1) 77-86. doi:10.1353/hub.2004.0014

Adhikari, K., Chacón-Duque, J.C., Mendoza-Revilla, J., Fuentes-Guajardo, M., Ruiz-Linares, A. (2017). The Genetic Diversity of the Americas. *Annual Review of Genomics and Human Genetics*, 18(1), 277–296. doi:10.1146/annurev-genom-083115-022331

Amigo, J., Salas, A., Phillips, C., Carracedo, Á. (2008). SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* 9:1–6. - SNP*for*ID Browser (available at http://spsmart.cesga.es/snpforid.php)

Amorim,C.E.G., Gontijo, C.C., Falcão-Alencar, G., Godinho, N.M.O., Toledo, R.C.P., Pedrosa, M.A.F., Luizon, M.R., Simões, A.L., Klautau-Guimães, M.N., Oliveira, S.F. (2011). Migration in Afro-Brazilian Rural Communities: Crossing Demographic and Genetic Data. *Human Biology* 83 509-521. doi:10.3378/027.083.0405

Brasileiro, S., Sampaio, J.A.L. (2002). *Sacutiaba e Riacho de Sacutiaba: uma comunidade negra rural no oeste baiano*, in: Quilombos: Identidade Étnica e Territorialidade, E.C. O'Dwyer (Eds). Rio de Janeiro: Fundação Getúlio Vargas and Associação Brasileira de Antropologia, 2002, pp.83–108.

Cann H.M., Toma C., Cazes D., Legrand L., Morel M.F., Piouffre V., Bodmer L., Bodmer J., Bonne-Tamir W.F., Cambon-Thomsen B.A., et al. (2002). A human genome diversity cell line panel. *Science 296, 261–262.* doi:10.1126/science.296.5566.261b

Cardena, M.M.S.G., Ribeiro-dos-Santos, A., Santos, S., Mansur, A.J., Pereira, A.C., Fridman, C. (2013). Assessment of the Relationship between Self-Declared

Ethnicity, Mitochondrial Haplogroups and Genomic Ancestry in Brazilian Individuals. *PLoS ONE* 8(4): e62005. doi:10.1371/journal.pone.0062005

Da Silva, V.R.R. (2012). *Entre quilombos e palenques: um estudo antropológico sobre políticas públicas de reconhecimento no Brasil e na Colômbia*. Thesis (PhD) - Faculty of Philosophy, Letters and Humanities, Postgraduate Program in Social Sciences, Universidade de São Paulo (USP).

DANE – Dirección Nacional de Estadística, 2015: http://www.dane.gov.co/index.php (accessed June 2019)

De Moura, R.R., Coelho, A.V.C., Balbino, V.Q., Crovella, S., Brandão, L.A.C. (2015). Meta-Analysis of Brazilian Genetic Admixture and Comparison with Other Latin America Countries. *American Journal of Human Biology* 27 674–680. doi:10.1002/ajhb.22714

Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I. S., Maria Calò, C., … Wells, R.S. (2014). Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature Communications*, 5(1). doi:10.1038/ncomms4513

Etnoterritorios - Universidad Javeriana & AECID (Agencia Española de Cooperación Internacional para el Desarrollo), 2006 – www.etnoterritorios.org (accessed June 2019)

Evanno, G., Regnaut, S., Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol.* 14(8):2611–20.

Excoffier, L., Lischer, H.E.L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567.

FCP – Fundação Cultural Palmares: http://www.palmares.gov.br, 2017 (accessed June 2019)

Fondevila, M., Phillips, C., Santos, C., Freire-Aradas, A., Vallone, P.M., Butler, J.M., et al. (2013). Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet* 7(1):63–74. doi: 10.1016/j.fsigen.2012.06.007

Freire-Aradas,A. , Ruiz, Y., Phillips, C., Maroñas, O., Söchtig, J., Tato, A.G., Dios, J. A., Cal,M.C., Silbiger, V.N., Luchessi, A.D., Chiurillo, M.A., Carracedo, A., Lareu, M.V. (2014). Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Science International: Genetics* 13 3–9. doi:10.1016/j.fsigen.2014.0

Garavito, G., Martinez, B., Builes, J.J., Aguirre, D., Mendoza, L., Afanador, C.H., Egea, E., Marrugo, J. (2015). Indels markers set and ancestry estimates in a population sample from Atlantic Department of Colombia. *Forensic Science International: Genetics Supplement Series* 5 e177–e178. doi:10.1016/j.fsigss.2015.09.071

Giolo, S.R., Soler, J.M.P., Greenway, S.C., Almeida, M.A.A., De Andrade, M., Seidman, J.G., Seidman, C.E., Krieger, J.E., Pereira, A.C. (2011). Brazilian urban population genetic structure reveals a high degree of admixture. *Eur J Hum Genet* 20(1) 111-6. doi:10.1038/ejhg.2011.144.

Gomes, F. S. (2015). *Mocambos e Quilombos. Uma história do campesinato negro no Brasil*. São Paulo, SP: Claroenigma

Gómez-Carballa A., Pardo-Seco J., Brandini S., Achilli A., Perego U.A., Coble M.D., et al. (2018). The peopling of South America and the trans-Andean gene flow of the first settlers. *Genome Res.* 28(6):767–79.

Gontijo, C.C., Mendes, F.M., Santos, C.A., Klautau-Guimarães, M.N, Lareu, M.V., Carracedo, A., Phillips, C, Oliveria, S.F. (2018). Ancestry analysis in rural Brazilian populations of African descent. Forensic Sci Int Genet 160–6. doi:10.1016/j.fsigen.2018.06.018

Gontijo, C.C., Amorim, C.E.G., Godinho, N.M.O., Toledo, R.C.P., Nunes, A., Silva, W., Moura, M.M.F., Oliveira, J.C.C., Pagotto, R.C., Klautau-Guimarães, M.N., Oliveira, S.F. (2014). Brazilian Quilombos: A Repository of Amerindian Alleles. *American Journal of Human Biology*, 26 142-150. doi:10.1002/ajhb.22501

Gonzáles-Andrade, F., Sánchez, D., González-Solórzano, J., Gascón, S., Martinez-Jarreta, B. (2007) Sex-specific genetic admixture of mestizos, amerindian kichwas, and afro-ecuadorans from Ecuador. *Human Biology*, v.79, no. 1, pp. 51-77.

Guauque-Olarte, S., Fuentes-Pardo, A.P., Cárdenas-Henao, H., Barreto, G. (2010) Genetic Structure and Diversity of Three Colombian Southwest Afrodescendent Populations Using 8 STR's. *Acta biol. Colomb*., 15 (2) 47-60. ISSN 1900-1649.

Heinz, T., Álvarez-Iglesias, V., Pardo-Seco, J., Taboada-Echalar, P., Gómez-Carballa, A., Torres-Balanza, A., Rocabado, O., Carracedo, A., Vullo, C., Salas, A. (2013). Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. *Forensic Science International: Genetics* 7 537–542. doi:10.1016/j.fsigen.2013.05.012

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, 343(6172), 747–751. doi:10.1126/science.1243518

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering. 9 (3) 90-95. doi: 10.1109/MCSE.2007.55

Ibarra, A., Freire-Aradas, A., Martínez, M., Fondevila, M., Burgos, G., Camacho, M., Ostos, H., Suarez, Z., Carracedo, A., Santos, S., Gusmão, L. (2013). Comparison of the genetic background of different Colombian populations using the SNPforID 52plex identification panel. *Int J Legal Med* 128(1):19-25. doi:10.1007/s00414-013-0858-z

IBGE – Instituto Brasileiro de Geografia e Estatística, 2010: http://www.ibge.gov.br (accessed June 2019)

Jobling M.A., Gill, P. (2004). *Encoded evidence: DNA in forensic analysis, Nature Reviews Genetics*, 5 739–751. doi:10.1038/nrg1455.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I. (2015). CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15(5) 1179-1191. doi:10.1111/1755-0998.12387.

Landers J., Gómez P., Acuña J.P., Campbell C.J. (2015). *Researching the history of slavery in Colombia and Brazil through ecclesiastical and notarial archives in: From Dust to Digital: Ten Years of the Endangered Archives Programme.* Kominko, M. (Editor). Open Book Publishers. pp. 259-292. https://www.jstor.org/stable/j.ctt15m7nhp.20

Li, J.Z., Absher, D.M., Tang, H.,Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H. M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319 1100–1104.

Lins, T.C., Vieira, R.G., Abreu, B.S., Grattapaglia, D., Pereira, R.W. (2010). Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs *Am. J. Hum. Biol.* 22 187–192. doi:10.1002/ajhb.20976

Machado, T.M.B.M. (2008). Ancestralidade em Salvador-BA. *Masters Assay, Biotechnology in Health and Investigative Medicine*, Centro de Pesquisa Gonçalo Moniz, FIOCRUZ, Bahia, Brazil.

Manta F.S.N., Pereira R., Vianna R., Araújo A.R.B., D.L.G. Gitaí, D.A. Silva, E.V. Wolfgramm, I.M. Pontes, J.I. Aguiar, M.O. Moraes, E.F. Carvalho, L. Gusmão,

Revisiting the Genetic Ancestry of Brazilians Using Autosomal AIM-Indels, PLoS ONE 8(9): e75145 (2013). doi:10.1371/journal.pone.0075145

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S.D., Mark, J., Bustamante, C.D., Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*. 100(4) 635-649. doi:10.1016/j.ajhg.2017.03.004.

Martínez, H., Rodríguez-Larralde, A., Izaguirre, M.H., de Guerra, D.C. (2007) Admixture estimates for Caracas, Venezuela, based on autosomal, Y-chromosome, and mtDNA Markers. *Human Biology*, v. 79, no. 2, pp. 201-213

Martínez-Marignac, V., Bertoni, B., Parra, E.J., Bianchi, N.O. (2004) Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. *Human Biology* v. 76, no 4, pp. 543-557.

Mendes, F.M., Gontijo, C.C. Kpop: A Python package for population genetics analysis. Forensic Sci Int Genet Suppl Ser. 2017,6:e407–9.

Moreno, F., Freire-Aradas, A., Phillips, C., Fondevila, M., Carracedo, A., Lareu, M.V. (2014). SNP variation with latitude: Analysis of the SNPforID 52-plex markers in north, mid-region and south Chilean populations. *Forensic Science International: Genetics* 10 12–16. doi:10.1016/j.fsigen.2013.12.009

Neme, S., Andrade, CO. (1987). Quilombo: forma de resistência. Proposta histórico-arqueológica. In: *Insurreição Negra e Justiça* (Eds. Huber, G., de Souza F.B.). Rio de Janeiro, RJ: OAB.

Ossa, H., Aquino, J., Pereira, R., Ibarra, A., Ossa, R.H., Pérez, L.A., Granda, J.D., Lattig, M.C., Groot, H., Carvalho, E.F., Gusmão, L. (2016). Outlining the Ancestry Landscape of Colombian Admixed Populations. *PLOS ONE* 11(10):e0164414. doi:10.1371/journal.pone.0164414

Paiva, S.G. (2017) *Fatores de risco para doenças cardiovasculares em quilombos contemporâneos do Brasil Central : parâmetros demográficos, socioeconômicos, ancestralidade genética e saúde.* Thesis (PhD) - Postgraduate Program in Anima Biology, Institute of Biology, University of Brasilia, Brazil.

Pedrosa, M.A.F. (2006). Composição Genética de Quatro Populações Remanescentes de Quilombo do Brasil com Base em Microssatélites e Marcadores de Ancestralidade. *Masters Assay, Molecular Biology Graduate Program,* Institute of Biological Sciences, University of Brasília, Brazil.

Pena, S.D.J., Di Pietro, G., Fuchshuber-Moraes, M., Genro, J.P., Hutz, M.H., Kehdy, F.D.S.G., et al. (2011). The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. *PLoS ONE* 6(2) e17063. doi:10.1371/journal.pone.0017063

Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., Casares de Cal, M., Ballard, D., Lareu, M.V., Carracedo, A., The SNPforID Consortium. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics* 1, 273-280.

Phillips, C., Aradas, A. F., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., (…) Lareu, M. V. (2013). Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Science International: Genetics* 7(3), 359–366. doi:10.1016/j.fsigen.2013.02.010

Pinotti T., Bergström A., Geppert M., Bawn M., Ohasi D., Shi W., et al. (2019). Y Chromosome Sequences Reveal a Short Beringian Standstill, Rapid Expansion, and early Population structure of Native American Founders. *Curr Biol*. 29: 1-9.

Pritchard, J.K., Stephens, M., Donnelly, P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–59.

Pritchard, J.K., Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol*. 60(3):227–37.

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. *Nature* 488 (7411) 370–374. doi:10.1038/nature11258

Rondón, F., Osorio J.C., Peña, A.V., Garcés, H.A., Barreto, G. (2008). Diversidad Genética en Poblaciones Humanas en Dos Regiones Colombianas. 39 (2) 52-60 Universidad del Valle Cali, Colombia.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W. (2002). Genetic Structure of Human Populations. *Science*, New Series, Vol. 298, No. 5602, 2381-2385

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*. 1 70.

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., et al. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genet* 10(9) e1004572. doi:10.1371/journal.pgen.1004572.

Salzano, F. M., Sans, M. (2014). Interethnic admixture and the evolution of Latin American populations. *Genetics and Molecular Biology,* 37(1 suppl 1), 151–170. doi:10.1590/s1415-47572014000200003

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A. H., Burchard, E. G., … Lareu, M. V. (2016). Pacifiplex : an ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Science International: Genetics* 20, 71–80.doi:10.1016/j.fsigen.2015.10.003

Shriver, M.D., Kittles, R.A. (2004). Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics 5 (2004) 611–618.* doi:10.1038/nrg1405.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R., Ferrell, R.E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J. Hum Genet*, 60:957-964.

Silbiger, V.N., Hirata, M.H., Luchessi, A.D., Genvigir, F.D.V., Cerda, A., Rodrigues, A.C., et al. (2012). Differentiation of African Components of Ancestry to Stratify Groups in a Case–Control Study of a Brazilian Urban Population. *Genet Test Mol Biomarkers* 16(6):524–30. doi:10.1089/gtmb.2011.0267

Tang H., Choudhry S., Mei R., Morgan M., Rodriguez-Cintron W., Burchard E.G., Risch N.J. (2007). Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics,* 81 626-633. doi:10.1086/520769.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature 526, 68-74.* doi:10.1038/nature15393.

Urbano, L., Portilla, E.C., Muñoz, W., Sierra-Torres, C.H., Bolaños, H., Arboleda, Y., Aguirre, D.P., Mendoza, L., Carmona, V., Afanador, C.H., Salgar, M., Gusmão, L.,Builes, J.J. (2015). Ancestral genetic composition in a population of South Western Colombian using autosomal AIM-INDELS. *Forensic Science International: Genetics Supplement Series* 5 e189–e190. doi:10.1016/j.fsigss.2015.09.076

Vila Real, R.N.S. (1996) *Cultura e Currículo: Um estudo da escola Kalunga. Dissertatio (Masters)* - Universidade Federal de Goiás, Goiânia, GO.

Xavier, C., Builes, J.J., Gomes, V., Ospino, J.M., Aquino, J., Parson, W., Amorim, A., Gusmão, L., Goios, A. (2015). Admixture and Genetic Diversity Distribution Patterns of Non-Recombining Lineages of Native American

Ancestry in Colombian Populations. *PLoS One* 16, 10(3):e0120155. doi:10.1371/journal.pone.0120155.

Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo C., Poletti G., Mazzoti G., Hill K., Hurtado A.M., et al. (2008). Geographic Patterns of Genome Admixture in Latin American Mestizos. *PLoS Genet* 4(3) e1000037. doi:10.1371/journal.pgen.1000037

Figure Legends

Figure 1. Two components (1st and 2nd) PCA for reference, comparison and study samples. a) PCA for dispersion of the study set (Mulaló, Kalunga and Sacutiaba) in relation to reference (Africa, Europe and Indigenous America) and comparison (Northern Africa, Central and South Asia, and the Middle East) groups. **b)** PCA for dispersion of the study sample in relation to the reference populations. **c)** PCA for dispersion of the study set in relation to each population that constitute the African parental group.

Figure 2. Structure analysis for reference, comparison and study samples. Structure analysis considering k:3-5 for reference (AFR: Africa, EUR: Europe, IAM: Indigenous American); comparison (EAS: East Asia, CSA: Central and South Asia, NAF: Northern Africa, MID: Middle East), and study sample (Brazilian *quilombos* Kalunga: KAL and Sacutiaba: SAC, and the Colombian *comunidad negra* Mulaló: MUL).

Figure 3. Ancestry analysis. a) Bar plot for k:3 (best k) considering the three parental populations (Africa, Europe and Indigenous America) most likely to have contributed to the gene pools of Mulaló, Kalunga and Sacutiaba. **b)** Pie charts of population ancestry estimates. AFR: Africa, EUR: Europe, IAM: Indigenous America, MUL: Mulaló, KAL: Kalunga, SAC: Sacutiaba.

Figure 4. Distribution of Individual ancestry estimates. Violin plots showing ranges of individual ancestry estimates for each parental group (Africa, Europe, and Indigenous America) in the study populations Mulaló, Kalunga, and Sacutiaba.

Supporting information

(available at https://github.com/ninacalorina/supplementary-material-thesis)

Table S1. Database. Complete database tested in the exploratory Structure analysis, pairwise Fst estimates and PCA. Populations/samples in bold were excluded from ancestry analysis.

Table S2. Allelic frequencies. Allelic frequencies observed in the quilombos Kalunga and Sacutiaba and the palenque Mulaló.

Table S3. Descriptive parameters. Number of genotypes analysed, Hardy-Weinberg Equilibrium (p-value, s.d.), Observed (Hobs) and expected (Hexp) heterozygosity, number of pairs a given locus is in linkage disequilibrium (LD) for the 34plex panel in the quilombos Kalunga and Sacutiaba and in the palenque Mulaló. Highlighted cells: p-value < 0.01. Highlighted: loci with p-value ± s.d. < 0.01.

Table S4. Pairwise F_{ST}. Pairwise F_{ST} between populations from Africa (AFR), Europe (EUR), Indigenous America (IAM), East Asia (EAS), Central and South Asia (CSA), Northern Africa (NAF), Middle East (MID) and the African derived study populations from Brazil (KAL and SAC) and Colombia (MUL). In dark grey: continental groups and study populations. In red: F_{ST} > 0.10 within continental group. In blue: study populations. In light grey: group formed by EUR, CSA, NAF and MID (mean F_{ST} < 0.0,9). In green: Uygur (CSA) x EAS (mean F_{ST} < 0.10).

Table S5. Ancestry estimates in Mulaló, Kalunga and Sacutiaba. Individual and population ancestry estimates considering Africans (AFR), Europeans (EUR) and Indigenous Americans (IAM) as contributing populations to admixture in the formation of the study populations. Mean and median values, s.d., and minimum and maximum estimates are shown.

CAPÍTULO 3

Sistema PIMA (*Population Informative Multiplex for the Americas)*: Multiplex Informativo de População para as Américas

Neste capítulo, apresentamos o sistema PIMA, um novo painel para a detecção do componente indígena americano em populações miscigenadas contemporâneas. Ele é formado por 26 AIM SNPs autossômicos, além de um SNP no cromossomo X e o marcador de sexagem amelogenina, e foi desenhado para complementar o sistema-cerne 34plex (analisado no capítulo 2). O artigo PIMA: *Population Informative Multiplex for the Americas* apresenta dados novos de diversas populações. Dentre elas, estão os quilombos Kalunga e Sacutiaba, avaliados com outras populações de diferentes composições de ancestralidade e grupos amostrais de marcada ancestralidade africana: a *comunidad negra* Mulaló, na Colômbia, e a amostra ASW, composta por pessoas de ancestralidade africana do Sudoeste dos Estados Unidos. PIMA foi elaborado por uma equipe da *Unidade de Xenética do Instituto de Ciencias Forenses da Universidade de Santiago de Compostela*, Galícia, Espanha. Inicialmente, o sistema foi desenvolvido como parte da tese "*Estudio de marcadores genéticos en poblaciones nativoamericanas y mestizas americanas: aplicaciones forenses, poblacionales y en estudios de asociación*", defendida por Gloria Liliana Porras-Hurtado no ano de 2011. Liliana foi responsável pela seleção

dos marcadores, padronização inicial e genotipagem de populações. Durante meu estágio doutoral, participei da padronização final da análise laboratorial, genotipagem, análise estatística populacional descritiva e comparativa, redação do texto e preparo do material para publicação. Posteriormente, o trabalho expandido e modificado para compor esta tese e está apresentado aqui no formato de artigo definido pelo periódico *Forensic Science International: Genetics*, ao qual foi submetido para publicação.

Resumo

Apresentamos aqui um painel de SNPs autossômicos informativos de ancestralidade desenvolvido como um multiplex pequeno e informativo, capaz de complementar marcadores uniparentais na detecção e mensuração do componente de mistura Indígena Americano (IAM) encontrado em populações contemporâneas. Selecionamos SNPs com: 1. diferenciais de frequência alélica extremos entre Indígenas Americanos e outros grupos que contribuíram para a miscigenação observada em populações americanas atuais (África, Europa e Leste da Ásia); 2. alta informatividade para determinação de ancestralidade (I_n); e 3. distribuição espaçada ao longo do genoma e separada dos SNPs de outros sistemas de pequena escala existentes para o uso forense. O painel resultante, analisado por extensão de base única por SNaPshot e genotipagem em eletroforese capilar, foi denominado PIMA (*Population Informative Multiplex for the Americas*), e é composto por 26 SNPs autossômicos, pelo marcador de sexagem amelogenina e por um SNP no cromossomo X. PIMA complementa o painel-cerne 34plex e juntos proveem uma ferramenta simples e poderosa para a análise de populações do continente americano, incluindo aquelas com histórias complexas de miscigenação.

A comparação dos resultados obtidos combinação dos painéis PIMA e 34plex àqueles obtidos por um painel 5 vezes maior nos permitiu verificar sua eficiência relativa. PIMA+34plex fornecem poder equivalente ao sistema LACE, composto por 314 SNPs, mas requer um esforço de genotipagem menor.

Os perfis de ancestralidade e estrutura genética de 22 populações do continente americano foram estimados utilizando dados de PIMA+34plex, e essas estimativas foram contrastadas com informação obtida de marcadores uniparentais (mtDNA e cromossomo Y) para uma pequena amostra de indivíduos miscigenados da Venezuela. Nossos resultados indicam que o componente de mistura IAM é eficientemente detectado em populações contemporâneas usando o nosso pequeno conjunto de marcadores informativos de ancestralidade, e que esses resultados são

condizentes com a história conhecidas das Américas. O pequeno tamanho e alto poder de discriminação do PIMA, particularmente quando combinado ao 34plex, fazem desse sistema uma ferramenta poderosa e prática para estudos de genética de populações americanas e para a rotina forense.

Palavras-chave: Ancestralidade; SNP; AIM; Indígena Americano; mistura populacional; SNaPshot; HGDP-CEPH

PIMA: Population Informative Multiplex for the Americas

Carolina Carvalho Gontijo[1,2¶], Liliana Porras-Hurtado[1,3¶], Christopher Phillips[1¶*], Ana Freire-Aradas[1], Manuel Fondevila[1], Carla Santos[1], Antonio Salas[1], Julieta Henao[3], Carlos Isaza[3], Leonardo Beltrán[3], Vivian Nogueira Silbiger[4], Adriana Castillo[5], Adriana Ibarra[6], Fabián Moreno Chavez[7], Jens Sochtig[1], Yarimar Ruiz[1], Guillermo Barreto[8], Fernando Rondon[5,8], William Zabala[9], Lisbeth Borjas[9], Silviene F de Oliveira[2], Ángel Carracedo[1], Maria Victoria Lareu[1]

1 Forensic Genetics Unit, University of Santiago de Compostela, Spain.
2 Human Genetics Laboratory, Institute of Biological Sciences, University of Brasília, Brazil.
3 Medical Genetics Laboratory, Human Molecular Genetics Research Group, Technology University of Pereira, Colombia.
4 Department of Clinical and Toxicological Analysis, Health Sciences Center, Federal University of Rio Grande do Norte, Brazil.
5 Medical Genetic Laboratory, Industrial University of Santander (UIS), Colombia.
6 Medical Genetics Laboratory, University of Antioquia, Colombia.
7 Servicio Médico Legal, Ministry of Justice and Human Rights of Chile, Santiago, Chile.
8 Human Molecular Genetics Research Group, University of Valle, Colombia.
9 Molecular Genetics Laboratory, Medical Genetics Unit, University of Zulia, Venezuela.

ABSTRACT

We describe an ancestry-informative autosomal SNP multiplex designed to be a small-scale, flexible panel that can complement uniparental markers in assessing the American variability (i.e. pre-Colombian) found in contemporary indigenous American populations. This study centered on choosing SNPs with the specific characteristics of: 1) extreme allele frequency differences between indigenous Americans and the African, European and East Asian population groups that contribute to present-day population variation in the Americas; 2) high informativeness-for-assignment I_n values; and 3) well-spaced genomic distribution and chromosomal separation from existing small-scale forensic ancestry marker sets. The resulting capillary electrophoresis SNaPshot single base extension test was named: PIMA (Population Informative Multiplex for the Americas), comprising 26 autosomal SNPs, a single X-chromosome SNP plus the amelogenin sex marker adapted for SNaPshot. PIMA complements the established 34plex forensic ancestry panel to provide a powerful and simple tool for the analysis of American populations, including those with admixed histories, commonly encountered in America. Comparing the results obtained with the combined marker panels of PIMA and 34plex to SNP data from a much larger ancestry panel allowed us to gauge their relative efficiency. PIMA+34plex gives equivalent power to the 314-SNP 'LACE' genomic ancestry control panel, while requiring a much smaller genotyping effort. The ancestry profiles and genetic structure of 22 populations spread across the American continent were estimated using PIMA+34plex data, and those estimates were contrasted with information provided by uniparental markers (mtDNA and Y-chromosome loci) for a small set of admixed individuals from Venezuela. Our results indicate that an American genetic component is efficiently detected in contemporary American populations using a small set of ancestry informative SNPs, and these co-ancestry estimates are consistent with the known history and demography of the Americas. The small scale and high population differentiation power of PIMA, particularly when

combined with 34plex, provides a practical and powerful tool for genetic studies of American populations as well as forensic DNA analyses.

Keywords: Ancestry; SNP; AIM; Indigenous American; population admixture;

SNaPshot; HGDP-CEPH

Highlights

- ➔ Compact SNaPshot-based ancestry SNP test called PIMA developed with 26 autosomal SNPs, 1 X-SNP and amelogenin
- ➔ 22 American-continent study populations characterized with 25 PIMA SNPs and 27 SNPs of established 34plex forensic ancestry SNaPshot test
- ➔ PIMA found to be highly efficient at detecting and apportioning co-ancestry in admixed populations from America – even detecting 4-way admixture in São Paulo sample
- ➔ Evaluations of admixed samples typed in common with PIMA+34plex vs a 314-SNP genomics set shows equivalent power to detect American co-ancestry
- ➔ Comparisons of PIMA+34plex with Y/mtDNA patterns in small set of admixed Venezuelans highlights need for powerful autosomal SNPs added to uni-parental data.

1. Introduction

Inference of individual biogeographical ancestry (BGA) plays an important role in a wide range of genetics fields. In medical genetics, it guides case-control studies correcting for substructure effects, important in the evaluation of medical risks [1]. In population genetics, BGA using carefully chosen Ancestry Informative Markers (AIMs) informs the interpretation of human population expansions, movement and interaction. Carefully chosen AIMs [2] particularly improve the estimation of study population's ancestry proportions, assessment of genetic structure and demographic histories. BGA is of increasing relevance to forensic genetics, as an ancestry inference obtained from the DNA sample can help focus a criminal investigation [3], particularly when there are no suspects or matches to national DNA database profiles.

In contemporary American populations, BGA inference is complicated by two factors: the continent's convoluted history and demographics; and the difficulty in fully differentiating the closely related variation found in East Asian and Indigenous American populations: stemming from their separation just 15 thousand years ago (KYA). The American continents now have the most intricate ancestral genetic backgrounds of any contemporary population group defined by continentally-based regional distribution (i.e. Africa, European-Eurasia, East Asia, Oceania, with South Asian-Eurasia often separately defined as a continental region). Demographic movements throughout history have produced high levels of genetic heterogeneity and a complex landscape of population admixture. The peopling of the Americas

happened from north to south, beginning 15-18 KYA [4]. The initial source of migration was Beringia in the far northeast of Siberia, where ancestors of the first migrant groups into NW America were likely isolated from their own Siberian ancestral populations before entering the continent [5]. In the 15th Century, the arrival of European colonizers and the African slave trade led to the introduction and subsequent expansion of new population admixture contributors into the Americas. Before World War II, and more markedly afterwards, migratory waves originating from Europe, East Asia and the Middle East brought a new influx of people - and hence genetic diversity - into the continent [6]. Such a complex history of migration, admixture and founder effects (where only a small number of people undergo rapid expansion in numbers in newly occupied lands) has resulted in the highly admixed American populations seen across the continent today.

Information available from the analysis of uniparental and autosomal markers in American populations provides different perspectives of a population's past. Y-chromosome and mtDNA variation characterize single paternal and maternal lineages [7-10] that may not represent the overall ancestral background of a person if a lineage originates from a deep-rooted ancestry and is therefore potentially unrepresentative of that person's overall ancestry. In contrast, autosomal AIMs offer more scope for understanding the demographic structure of a population and for the analysis of ancestry at the individual level [11-14]. In a continent so diverse and with such a complex history as the Americas, the analysis of both lineage and autosomal markers together becomes necessary for the more complete reconstruction of an individual's ancestral background.

Here, we describe a forensic ancestry test comprising 26 ancestry informative SNPs (AIM-SNPs) that is dedicated to indigenous American population variability, named PIMA: Population Informative Multiplex for the Americas. The panel has SNPs with genomic positions avoiding those previously analyzed in the established 34-SNP ancestry test that forms the core of capillary electrophoresis ancestry analysis in several laboratories (herein 34plex) and was developed to focus on the differentiation of Sub-Saharan Africans, Europeans, and East Asians [15,16]. The goal of the reported study was to bring together a concise, yet highly informative set of AIM-SNPs for differentiating American ancestry from that of other continental populations, which could complement 34plex [15,16].

Because of the extensive degree of population admixture in America outlined above, the term ‘indigenous American’ is largely a theoretical description outside the Amazon region. We used it here to denote populations with ancestry components that are recognizably from pre-Colombian peoples who occupied the American continent prior to the influx of European colonizers, African slaves and, in much smaller numbers, recent East Asian migrants. For brevity, the term ‘American’ is used to denote this distinct ancestry, rather than its more common usage meaning ‘coming from America’.

2. Materials and Methods

2.1. Ethics statement

Collection and analysis of new samples followed standard ethical guidelines. All participants were informed about the research goals and data confidentiality, signing the relevant consent documents. The research was approved by: the Research Ethics Committee of the Faculty of Health Sciences at the University of Brasilia (CEP-FS/UnB 030/2002 and 151/07) for Brazilian population samples; and the Ethics Committee of the University of Santiago de Compostela (details in Supplementary Text S1 of [11]), for samples from Colombia, Guatemala, Chile and Venezuela.

2.2. Populations sampled

A total of 822 DNA study samples were collected with informed consent. Indigenous American samples comprised Colombians: Awa (n=30), Coyaima (n=22), Embera (n=7), Pastos (n=33), Pijao (n=32); Guatemalans: Q'eqchi (n=15); Venezuela: Wayu (n=23). Admixed population samples were collected in Brazil: São Paulo (n=163), Kalunga (n=69), Riacho de Sacutiaba (n=29); Colombia: North West Colombia (n=208), Bucaramanga (n=19), Mulaló (n=34); Chile: North Chile (n=25), South Chile (n=30); Guatemala: Guatemala City (n=20); and Venezuela: Maracaibo (n=29), Caracas (n=34). Four admixed populations from 1000 Genomes (first variant data release, herein 1KG) were included in this set: Puerto Ricans from Puerto Rico (PUR, n=55), Individuals with Mexican ancestry in Los Angeles, USA (MXL, n=66),

Colombians from Medellín (CLM, n=60), and African Americans from SW USA (ASW, n=61).

Fig. 1. Map of American study population and HGDP-CEPH reference population sampling locations.

The reference population set included 851 1KG samples from Africa, Europe, and East Asia. This set was used to assess the relative power of PIMA+34plex SNPs compared to the LACE panel [11] to detect three and four-way admixture components. Reference indigenous American populations without high levels of

admixture were not available from 1KG, so we used indigenous American populations from the CEPH Human Genetic Diversity Panel (HGDP-CEPH) panel comprising Pima and Maya from Mexico; and the Karitiana and Surui populations from Amazonian Brazil (total n=57). The geographic distribution of the study and reference populations originating from the American continent are shown in Fig. 1.

2.3. AIM-SNP selection

The twenty-six PIMA SNPs were selected from the largest set of human SNP genotypes (650,000 loci) analyzed in 52 global populations of the HGDP-CEPH panel by Stanford University [17]. The Stanford SNP genotypes are publicly available from The CEPH Foundation, but we made use of the SPSmart online SNP browsers [18-20], which allow the simultaneous comparison of the Stanford HGDP-CEPH data with genotypes from HapMap, Perlegen and 1KG. This generated a candidate pool of ~600 SNPs from which 32-36 were selected to design a multiplex suitable for a forensic PCR and SNaPshot single base extension test. Emphasis was placed on SNPs that had well differentiated East Asian and American allele frequencies, since these population groups have less divergence than other group comparisons due to more recent shared population histories. A single SNP rs6993205 without prior HGDP-CEPH genotypes was selected from the evaluation of HapMap data, using the SPS browser [19].

2.4. PIMA genotyping protocol

The PIMA multiplex PCR reaction was optimized using QIAGEN® Multiplex PCR kit, adding 2 µl of PCR Master Mix, 2 µl of primer mix (0.125-2.5 µM), and 1-20 ng of DNA in a final volume of 5 µl. Amplification conditions were: initial denaturation at 95 ºC for 15 min; 30 cycles at 94 ºC for 30 s, 60 ºC for 1 min and 72 ºC for 30 s; and a final extension at 72 ºC for 10 min. Excess primers and dNTPs were removed by adding 1 µl of ExoSAP-IT (1 U/µl Exonuclease I and Shrimp Alkaline Phosphatase, GE Healthcare) to 2.5 µl of the PCR product and incubated at 37 ºC for 45 min followed by 85 ºC for 15 min to inactivate the enzyme. The PIMA single base extension (SBE) reaction in a final volume of 3 µl contained 1.25 µl of SNaPshot™ reaction mix (Applied Biosystems: AB), 0.75 µl of SBE primer mix (0.5-4 µM) and 1 µl of purified PCR product. The SBE primer mix was diluted in 160 mM ammonium sulphate to avoid non-specific hybridization amongst the primers. The SBE reaction was performed in an AB 9700 thermal cycler with the following cycle program: 30 cycles of 96 ºC for 10 s, 55 ºC for 5 s and 60 ºC for 30 s. Excess nucleotides were removed by addition of 1 µl SAP (1 U/ µl Shrimp Alkaline Phosphatase, GE Healthcare) to the total volume of the extension products and incubated at 37 ºC for 80 min and 85 ºC for 15 min. A combination of 1 µl of sample, 9.5 µl LIZ 120 size standard plus HiDi formamide at a ratio of 1:33.3 (both AB) was analyzed by capillary electrophoresis using an AB 3130 Genetic Analyzer with POP4 and analyzed with GeneMapper v4.0. Predefined size windows for each allele were determined from prior analysis of a minimum of 20 test samples.

2.5. Analysis of mtDNA and Y chromosome STRs

To compare the estimated co-ancestry proportions in admixed individuals obtained from panels of autosomal SNPs with those from uniparental markers, a small sample of individuals from Caracas, Venezuela were genotyped (n=8) that had been previously analyzed for mitochondrial DNA [21] and Y chromosome STRs. Although small, this sample provided an insight into differences in co-ancestry proportion estimates that can arise from admixture sex bias.

2.6. Statistical analyses

Allele frequencies were downloaded from SPSmart [18-20] (http://spsmart.cesga.es). Hardy-Weinberg Equilibrium (HWE), observed (Hobs) and expected heterozygosity (Hexp), linkage disequilibrium (LD), pairwise Fst, pairwise difference within and between populations, and Nei's genetic distances were estimated using Arlequin 3.5 [22]. Genetic cluster analysis was made using Structure v2.3.3 [23] using the following parameters: burn-in length of 200,000; MCMC of 200,000; admixture model; correlated allele frequencies; POPFLAG. Five independent runs were analyzed for each tested K value (ranging from K:2 to K:7). From the output, CLUMPP v1.1.2 [24] was used to obtain the optimal K(LnPr(X|K) according to the method of Evanno [25] and to obtain the average permuted individual and population Q-matrices. Bar plots were constructed from average Q-matrices by DISTRUCT v1.1 [26].

Principal component analysis (PCA) was performed using KPOP [27] (available at https://github.com/fabiommendes/kpop). Four continentally-based reference population samples (1KG: AFR, EUR, EAS; CEPH: AME) were included in all plots for assessing the efficiency of PIMA+34plex SNPs in differentiating the main population groups contributing to admixture in the Americas. Study populations were plotted in four separate PCAs corresponding to groups with shared ancestry of: a) indigenous American populations that had little or no recorded admixture (Awa, Pastos, Pijao, Coyaima, Embera, Wayu, Q'eqchi); b) admixed American populations with a high contribution of European co-ancestry (Colombia: 1KG CLM, NW Colombia, Bucaramanga; Venezuela: Caracas, Maracaibo; Brazil: São Paulo; North and South Chile; 1KG MXL and 1KG PUR); c) admixed American populations with a high contribution of African co-ancestry: (Brazil: Kalunga, Sacutiaba; Colombia: Mulaló; 1KG ASW); and d) urban populations likely to have a high contribution of American co-ancestry (re-analyzing 1KG MXL, Guatemala City, North and South Chile data).

Population assignments using Naive Bayes likelihood ratio tests; estimation in each SNP of Rosenberg's informativeness-for-assignment – a well-established metric that measures population divergence (I_n [2]); and cross validation analysis were performed using the *Snipper* app suite classifier (available at http://mathgene.usc.es/snipper/). Statistical distances and correlation analysis were performed using Matplotlib 3.0 [28] on Python 3.7 and plotted using the library seaborn 0.9.

2.7. Evaluation of ancestry inferences from small-scale forensic AIM panels compared to a 5-fold larger genomics ancestry control panel

The efficiency of PIMA+34plex SNPs for ancestry inference of the four population groups relevant to the Americas was assessed by comparing their genotype data with those obtained from the analysis of the 314-SNP LACE panel [11], taking advantage of seven indigenous American population samples analyzed in common during the PIMA and LACE panel development programs.

It is important to note that in all American populations genotyped in common, seven 34plex SNPs gave problems with SNaPshot tests, with large proportions of incomplete data generated. Therefore, all analyses using PIMA+34plex data combined and reported here, had 27 of the 34plex SNPs, excluding: rs12913832; rs1335873; rs3827760; rs4540055; rs5030240; rs722098; rs881929 genotypes to avoid excessive data gaps.

The LACE panel was developed to detect the American, African and European co-ancestry components in self-declared ‘Latin American’ participants of epidemiological studies and therefore is designed to control population bias in case and control groups [1]. When the LACE AIMs were selected, there was less emphasis on East Asian ancestry differentiation. Nevertheless, a large number of carefully spaced SNPs are compiled in this panel at a five-fold higher genotyping scale, compared to the full 60 SNPs of PIMA+34plex combined, although a smaller set of 52 markers were actually characterized, as noted above.

3. Results and Discussion

3.1. Characteristics of PIMA panel SNPs

From the approximately 600 candidate SNPs identified to be the most informative for differentiating HGDP-CEPH American population variation from that of non-American populations, the PIMA panel assembled 26 autosomal AIM-SNPs, a single X chromosome AIM-SNP and the amelogenin sex marker (to aid assessment of the X-SNP genotype in the same test). The complete grid of PIMA SNP genotypes (and data from the 27/34 34plex SNPs used in all population analyses) is detailed in Supplementary Table S1A. The missing data rate was ~1.7% (1,766 of a total 102,544 individual genotypes). Supplementary Table S1B lists separately the HGDP-CEPH population data for tri-allelic SNP rs17287498 - the one autosomal marker not used for American study populations analyses but developed for the final version of the PIMA multiplex along with X-SNP rs3027749.

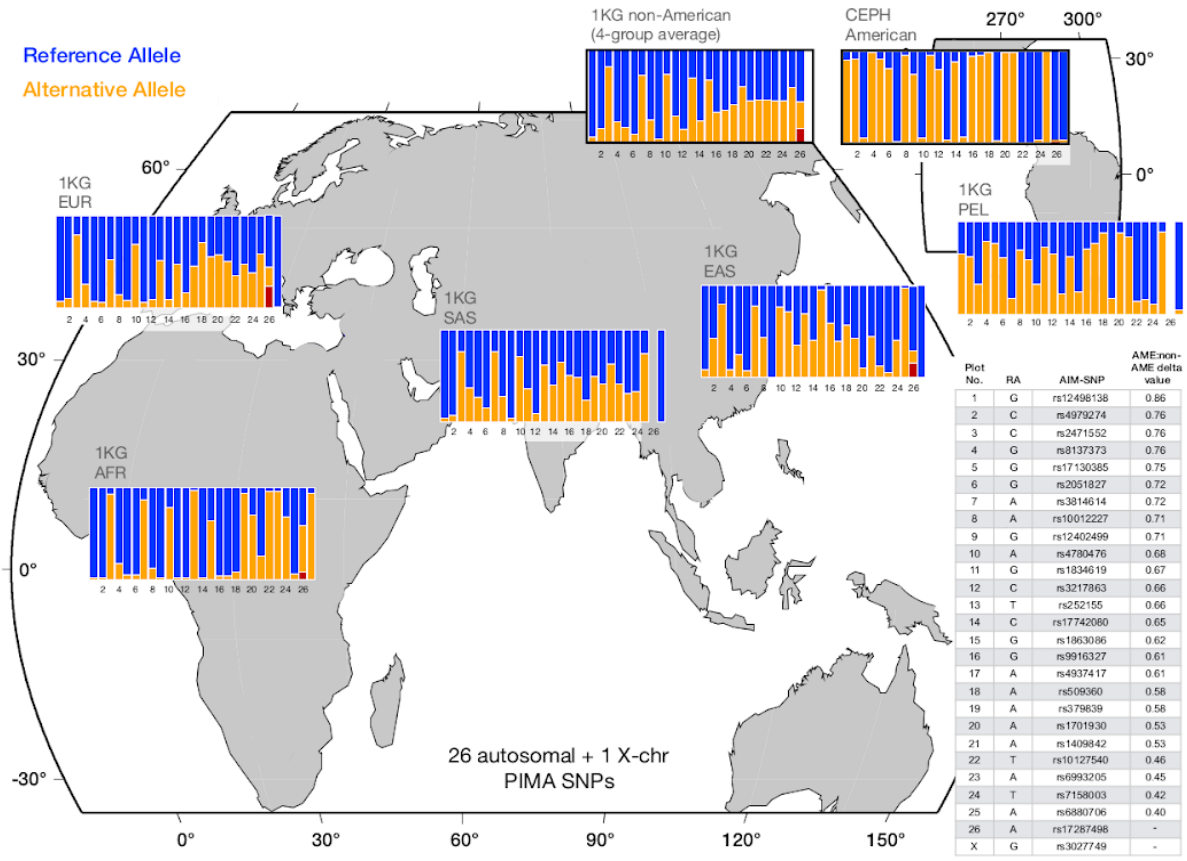
The initial SNaPshot multiplex included three other autosomal SNPs: rs16968965, rs17403380 and rs4792928; but they consistently underperformed, possibly due to interactions within the PCR or extension (EXT) primer sets and were removed. The amelogenin test was adapted for SNaPshot by extending either the A nucleotide on the 3’ side (in the forward RefSeq direction) of the 6-nt sequence absent on the X sequence ([CACTTT/-] A); or the 5’-most C of this sequence present on the Y. Supplementary Table S2A lists the genomic details of the 28 PIMA markers, plus the three removed SNPs. Supplementary Table S2B gives details of the PCR and extension (EXT) primers of the final optimized SNaPshot test.

Supplementary Fig. S1A shows typical SNaPshot peak patterns, and Supplementary Fig. S1B the genotypes obtained for the control DNA 9947A. Although peaks are well separated in the 9947A SNaPshot profile shown, low signals discernible in the rs252155-A (SNP code P20) and rs17130385-T (P8) extension products were not improved further by increasing these SNP's PCR primer mix concentrations. SNaPshot fragment size estimates for the dye sets of the alleles for each PIMA marker (using POP4 polymer electrophoresis) are given in Supplementary Fig. S1C. Both POP4 and POP6 polymer are applicable, but optimum separations were obtained during development with POP4.

Supplementary Table S2C and Supplementary Fig. S2 give the genomic positions of PIMA and 34plex SNPs; indicating a reasonably well spaced distribution that can minimize bias in the estimation of admixture (from the risk of closely sited co-segregating SNPs informative for a particular contributing population). The closest pair of SNPs between panels had a distance of 6.50 megabases (Mb) for rs8137373 (PIMA) and rs2040411 on chromosome 22; much further apart than two 34plex syntenic SNP pairs with ~1 Mb separation.

Worldwide patterns of population variation observed in the PIMA SNPs are detailed in two sets of maps. In Fig. 2, all 27 SNPs are displayed as bar-charts with the reference allele in blue (alternative allele in yellow, plus allele-3 in red in rs17287498), ordered by decreasing American vs non-American delta allele frequency differential values - as a simple system for indicating the most informative SNPs for this differentiation. Fig. 2 summarizes variation in the four 1000 Genomes population groups (AFR, EUR, EAS, SAS – South Asians) and combined

HGDP-CEPH American population frequencies plus 1000 Genomes Peruvians from Lima (PEL). Previous analyses of the AMR admixed American populations in 1KG indicated the highest levels of American co-ancestry in PEL, with ~20-25 individuals having little or no detectable non-American co-ancestry in this sample (detailed in *Snipper*, available at: http://mathgene.usc.es/snipper/forensic_mps_aims.html). Therefore, we regularly scrutinize PEL allele frequencies in SNPs that have not been characterized in other databases (e.g. tri-allelic SNPs and X-SNPs), and included here as a point of reference.



Plot No.	RA	AIM-SNP	AME:non-AME delta value
1	G	rs12498138	0.86
2	C	rs4979274	0.76
3	C	rs2471552	0.76
4	G	rs8137373	0.76
5	G	rs17130385	0.75
6	G	rs2051827	0.72
7	A	rs3814614	0.72
8	A	rs10012227	0.71
9	G	rs12402499	0.71
10	A	rs4780476	0.68
11	G	rs1834619	0.67
12	C	rs3217863	0.66
13	T	rs252155	0.66
14	C	rs17742080	0.65
15	G	rs1863086	0.62
16	G	rs9916327	0.61
17	A	rs4937417	0.61
18	A	rs509360	0.58
19	A	rs379839	0.58
20	A	rs1701930	0.53
21	A	rs1409842	0.53
22	T	rs10127540	0.46
23	A	rs6993205	0.45
24	T	rs7158003	0.42
25	A	rs6880706	0.40
26	A	rs17287498	-
X	G	rs3027749	-

Fig. 2. Summary bar-plots of PIMA SNP variation in the four main population groups of 1000 Genomes (1KG) and HGDP-CEPH indigenous Americans. SNPs are ordered left to right in descending allele frequency differential values (delta) comparing the 1KG 4-group average allele frequencies (non-AME) vs CEPH American (AME) frequencies (not calculated for the

tri-allelic SNP rs17287498 and X-SNP rs3027749). The non-American allele frequencies of rs17287498 were estimated from gnomAD population data (lacking SAS). Additional plots shown for the 1KG 4-group average frequencies and 1KG Peruvians from Lima (PEL). RA: reference allele; EUR: Europeans; AFR: Africans; SAS: South Asian; EAS: East Asian.

The map format of Fig. 2 (adapted from the Santa Cruz genome browser: genome.ucsc.edu) has the advantage of showing the American regions as a separate map. SNP rs17287498 is not listed by 1000 Genomes, so gnomAD data was used (available at: https://gnomad.broadinstitute.org/variant/10-54530788-C-A, South Asian data absent from gnomAD). The 4-group average frequencies from 1000 Genomes are also plotted and juxtapositioned next to the HGDP-CEPH Americans to highlight the most differentiated SNPs (excluding X-SNP rs3027749, which is African-informative). The first (leftmost) 12-14 SNPs show highly contrasting blue vs yellow allele frequency distributions and these loci most efficiently define American variation when present in an individual, whether through their population-of-origin or co-ancestry from admixture. The 26 autosomal PIMA SNPs are individually plotted as pie charts in Supplementary Figs. S3 (in the same delta-based order as Fig. 2), using SNP data from the Santa Cruz Genome Browser maps, where the ancestral (blue) and derived (yellow) alleles are indicated (rs17287498 data is from HGDP-CEPH genotyping completed for this study and given in Supplementary Table S1B). The maps of Supplementary Figs. S3 indicate many PIMA SNPs show a degree of within-American divergence, notably: rs2051827, rs12402499 and rs3217863 (Supplementary Figs. S3, SNPs 6, 9, 12) have considerable variation in their American-indicative allele frequencies amongst

the five HGDP-CEPH populations (1 in Colombia to 0.61 in Karitiana; 0.96 in Karitiana to 0.5 in Maya; 1 in Colombia to 0.37 in Surui, respectively). In contrast, many of the least differentiated SNPs show minimal or no within-American divergence, such as rs10127540 and rs6993205 (Figs. S3, 22, 23). The tri-allelic SNP rs17287498 (Figs. S3, 26) was highly informative for differentiating American and African populations from Eurasia (Europe, Middle East, South Asia), and to a lesser extent East Asia; with rs17287498-A mainly absent from Africa, and four of five HGDP-CEPH American populations monomorphic for the rs17287498-C allele. Overall, the individual SNP maps emphasize that a large proportion of the PIMA SNPs have American-specific alleles, or with similar degrees of informativeness, alleles absent from American populations but present in all or most other population groups worldwide.

Estimations of population genetics parameters: Hardy-Weinberg Equilibrium (HWE) test; observed and expected heterozygosity; pairwise linkage disequilibrium (LD) tests; and allele frequencies (25-SNP and 27-SNP subsets of each panel) are shown in Supplementary Table S3. Deviations from HWE mainly occurred in admixed populations and was most marked when admixture was complex (e.g. three-way co-ancestry in Brazilian populations). Values of pairwise LD also increased with admixture, ranging from ~5% (in Africa) to ~29% (in São Paulo, Brazil). In most populations, pairwise LD was ~10% (mean: 10.61%; median: 9.06%; SD: 3.32%). The high rate of LD in São Paulo, combined with HWE deviations, possibly due to underlying substructure in the population or sample. Awa (indigenous Colombian), Q'eqchi (indigenous Guatemalan) and Guatemala City (urban Guatemalan) were the

three populations with the highest level of monomorphic SNPs (25%, 17%, 15%, respectively), most of which are also monomorphic in either East Asians or Americans (Supplementary Table S3A). These patterns of variation may reflect endogamy characteristic of a small effective population size or recent founder effects. However, while occurrence of endogamy is plausible for the Awa and Q'eqchi, it is much less likely in the urban Guatemala City sample. Pairwise F_{ST}, pairwise differences between and within populations, and Nei's distance (*d*) are shown in Supplementary Fig. S4. Much of the American population structure revealed by these analyses is in accordance with the sample's known history and demographics, further reflected in the co-ancestry estimates detailed below.

3.2. Simple ancestry analysis tests with PIMA and 34plex: PCA, cumulative I_n and ancestry assignment success

Principal component analysis (PCA) of individual and combined PIMA and 34plex panels are shown in Fig. 3A (HGDP-CEPH American populations plus three 1000 Genomes population groups). Noting that 27 of the 34plex SNPs were genotyped for these comparisons, so this panel's data is slightly less informative than in normal forensic use. The PCA plots are matched to charts of the cumulative ancestry informativeness I_n values for each differentiation (i.e. Africans vs non-Africans, etc.) in Fig. 3B.

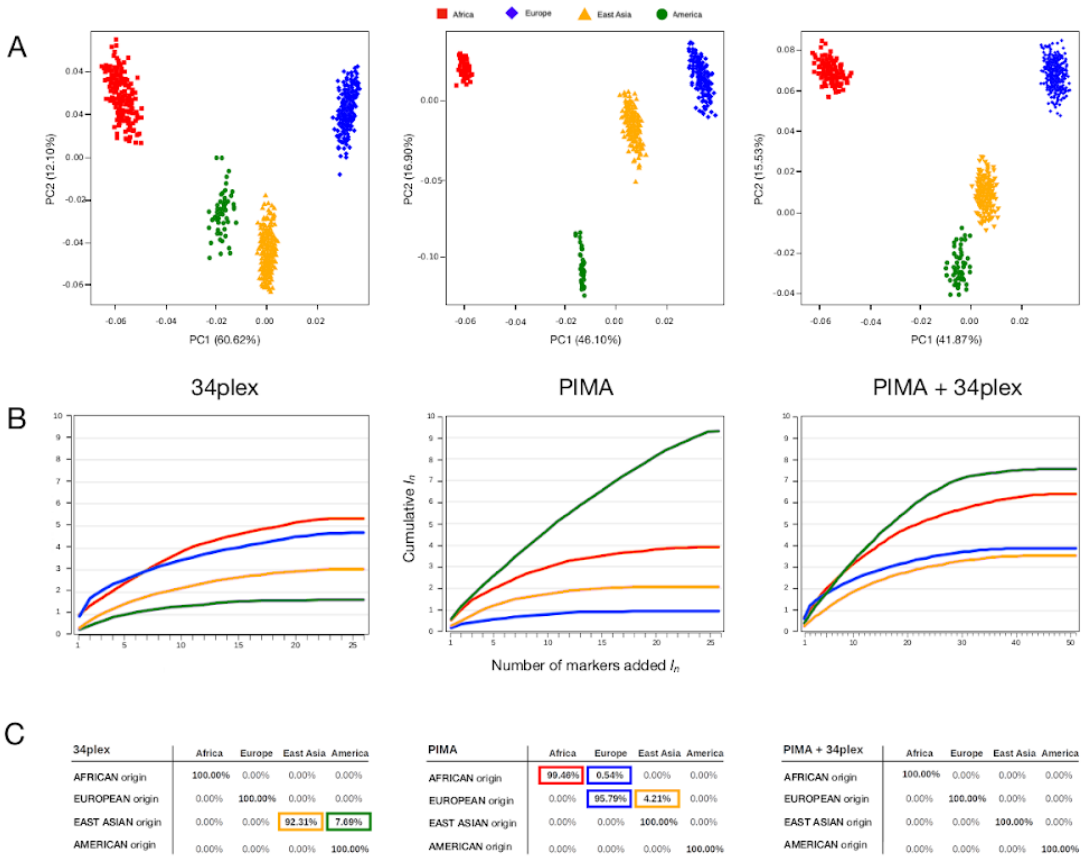


Fig. 3. Evaluation of the population differentiation performance of individual PIMA and 34plex panels and the combined set. Fig. 3A. PCA tests of 3 SNP sets analyzing 1000 Genomes African, European, East Asian; and HGDP-CEPH American genotype data. Fig. 3B. Cumulative I_n charts of PCA tests of 3 SNP sets, adding most powerful loci first. Fig. 3B. Cross validation ancestry assignment success of 3 SNP sets, with error highlighted for the relevant population group.

The PCA patterns of 34plex indicate that despite Africa and Europe being well differentiated, this panel does not separate East Asians and Americans as efficiently as PIMA, and the I_n values from American differentiation are the lowest. PIMA separates Americans from other populations more efficiently, but markedly increases this population group's I_n values; adding the risk of biasing estimation of American co-ancestry. Simultaneous analysis of PIMA and 34plex SNPs (Supplementary Table

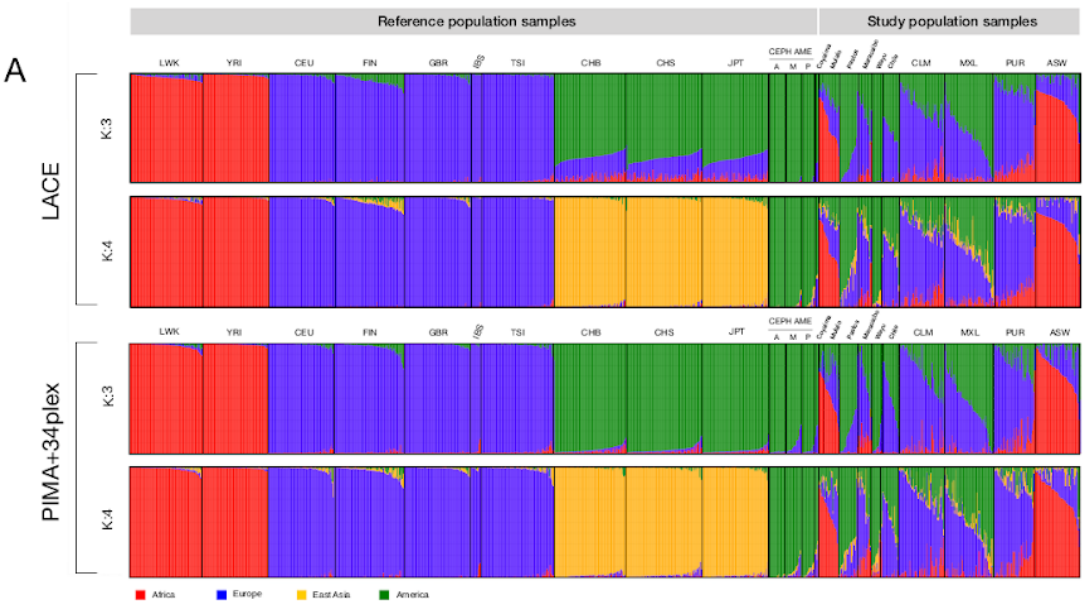
S1), although only marginally changing the separation of clusters in PCA, improves the clusteredness of each population's set of points. The cumulative I_n values also still show an imbalance of American-divergent SNPs overall, but final values of 4, 6 and 7.5 represent improvements on use of one AIM panel. The data in Fig. 3C show classification success rates when each SNP profile is analyzed with the remaining SNP data of each panel (i.e. cross-validated). These success rates indicate 34plex SNPs give almost 8% erroneous American ancestry inference of East Asians, in comparison to the 0.5% African erroneously inferred to be Europeans, and 4.2% Europeans erroneously inferred to be East Asians when using PIMA. Ancestry inference error from cross-validation is lost when both panels are combined, despite the lack of seven 34plex SNPs. In summary, the careful choice of American-informative SNPs of PIMA helps reduce the ancestry inference error observed (as East Asians incorrectly inferred to be American), due to the lack of American-informative SNPs in the 34plex panel. Combining 34plex with PIMA addresses the need to fully differentiate American from East Asian variation - an important prerequisite for the accurate detection of co-ancestry in admixed individuals from the Americas.

3.3. Comparisons of PIMA+34plex with the large-scale LACE ancestry panel

The comparison of the ancestry informativeness of two sets of AIMs can be subjective, but we opted to use STRUCTURE analysis to generate cluster membership proportion estimates for 1266 individuals (with emphasis on admixed

study population samples from the Americas), genotyped in parallel with PIMA+34plex and LACE panel SNPs. Fig 4A shows STRUCTURE cluster plots for K:3 and the optimum K:4 inferred genetic clusters for both AIM panels. Amongst study populations, the majority cluster membership values in LACE informed the ranking of the samples in Fig. 4A (e.g. ASW, Coiyama and Mulalo are ordered by LACE descending African cluster membership proportions, etc.). It is noted that a very small level of joint membership was observed in the reference populations but was mainly confined to 1KG Finnish Europeans and rarely exceeded 10%. Some European cluster joint membership in the Maya (M) and Pima (P) of the HGDP-CEPH American reference data is likely to reflect some admixture in these populations which has been detected in other studies, particularly in Maya [17,29]

The most evident difference between panels is the underestimation with LACE SNPs of the East Asian cluster component in reference population East Asians at K:3. The joint cluster memberships of East Asia (green) and Europe (blue) in the LACE data underlines the paucity of SNPs in this panel that fully differentiate East Asian variation from other groups. In both K:3 and K:4 inferred clusters, the patterns seen in the admixed study population samples are near identical, with just a marginal tendency of a higher proportion of African admixture components in the LACE patterns of CLM, MXL, PUR, ASW. The Indigenous Wayu study population shows the lowest level of non-American co-ancestry overall (less than many HGDP-CEPH Pima samples) and could be included as a reference population for American indicative SNP variation.



B

PIMA + 34plex	Africa	Europe	East Asia	America
Assigned as AFRICAN	100.00%	0.00%	0.00%	0.00%
Assigned as EUROPEAN	0.00%	100.00%	0.00%	0.00%
Assigned as EAST ASIAN	0.00%	0.00%	100.00%	0.00%
Assigned as AMERICAN	0.00%	0.00%	0.00%	100.00%

LACE	Africa	Europe	East Asia	America
Assigned as AFRICAN	100.00%	0.00%	0.00%	0.00%
Assigned as EUROPEAN	0.00%	100.00%	0.00%	0.00%
Assigned as EAST ASIAN	0.00%	0.00%	100.00%	0.00%
Assigned as AMERICAN	0.00%	0.00%	0.00%	100.00%

C

Mean F_{ST}

	LACE	PIMA + 34plex
Africa	0.4179	0.3374
Europe	0.2941	0.2375
East Asia	0.2988	0.3825
America	0.1862	0.2249

Mean I_n

	LACE	PIMA + 34plex
Africa	0.1591	0.1229
Europe	0.0907	0.0748
East Asia	0.1448	0.1453
America	0.0493	0.0683

Fig. 4A. STRUCTURE cluster plots for K:3 and K:4 inferred clusters comparing PIMA+34plex and LACE panels for a comprehensive set of common reference and American study samples. Cluster bar plots ordered in both SNP panels by increasing majority co-ancestry from LACE cluster membership proportions (CLUMPP-merged from 5 runs). HGDP-CEPH populations grouped and marked as: A=Amazonian Karitiana /Surui; M=Maya; P=Pima. **Fig. 4B.** Cross validation ancestry assignment rates for the reference population samples. **Fig. 4C.** Comparisons of mean population differentiation metrics F_{ST} and I_n in each panel.

Cross validation of the reference population samples with the SNPs of each panel showed both achieve 100% classification success for the optimum (K:4) cluster division of Africa, Europe, East Asia and America (Fig. 4B). The mean F_{ST} patterns were comparable on both sets of SNPs (Fig. 4C), although the high values

in East Asia and America for PIMA+34plex reflect the targeting of SNPs with higher than average divergence between these two groups. Informativeness I_n values showed only slight differences between panels and reasonable balance (Fig. 4C), since both panels aimed to keep a balance between the cumulative I_n values amongst the four population groups. The higher I_n value for America with PIMA+34plex SNPs compared to those of LACE, indicates a small panel of carefully chosen AIMs dedicated to a particular population group can successfully achieve high levels of differentiation, despite being limited in scale.

The consistency of individual co-ancestry estimates obtained from STRUCTURE analysis of PIMA+34plex and LACE was evaluated by statistical distance estimates (total variation) and correlation analysis for contributing populations (Supplementary Figs. S5). Correlation coefficients were high ($r^2 > 0.8$) considering three components (K:3 Africa, Europe and America). When the East Asian component at K:4 was included, the correlation coefficients stayed high for Africa, Europe and America ($r^2 > 0.8$), but were low for East Asia (0.0027), as would be expected given the lack of genuine East Asian co-ancestry amongst the study populations. This finding is confirmed by the correlation coefficients obtained when HGDP-CEPH East Asian samples were included in the study sample, obtaining an r^2 value higher than 0.9.

3.4. Patterns of admixture in the full range of American study populations with PIMA+34plex SNPs

Having established a close match between the STRUCTURE cluster patterns with both panels, we interpreted the distribution of cluster membership proportions

for PIMA+34plex at K:4 obtained from separate STRUCTURE analyses of the full range of 22 American study populations (5 runs, with identical reference population data). Table 1 outlines the population Q-matrix generated from the CLUMPP-merged population cluster membership averages of the five STRUCTURE runs. Supplementary Fig. S6 shows the cluster plot of merged data and aligns the study population's average cluster membership proportions under their respective cluster columns. Generally, the patterns of admixture observed in the cluster memberships of the 22 populations in Table 1 appear to be consistent with both their geographic distribution and demographic history.

In view of the wide range of admixture patterns observed in these study populations, an informative approach to visualize the data is to compare populations with similar admixture components and ratios. Therefore, we also made principal component analyses on subsets of the STRUCTURE genotype data from the reference population groups and sets of study populations together, and grouped these into four PCA plots (Fig. 5) comprising: a) indigenous populations; b) admixed populations with high European co-ancestry; c) admixed populations with high African co-ancestry; and d) urban populations with high American co-ancestry. The patterns in these additional PCA tests were consistent with assessments of admixture made in STRUCTURE, as they placed the admixed populations closer to the main contributing population in each case.

Table 1. Ancestry estimates in the full range of study populations from the American continent based on PIMA+34plex genotypes analysed with STRUCTURE. Average cluster membership proportions taken from CLUMPP-based population Q-matrix values with the cluster plot shown in Supplementary Fig. S6. 1000 genomes comprise; CLM, Colombians from Medellín; MXL, Mexican ancestry in LA USA; PUR, Puerto Ricans from Puerto Rico; ASW, African ancestry in SW USA. Highlighted cells indicate three populations with the smallest difference between estimates for the first and second components (<12%).

Population of origin	Inferred ancestry			
	African	European	East Asian	Indigenous American
CLM	0.084	0.693	0.014	0.209
NW Colombia	0.143	0.539	0.005	0.313
Bucaramanga	0.144	0.517	0.007	0.331
Mulaló	0.582	0.241	0.026	0.151
Awa	0.011	0.015	0.052	0.922
Pastos	0.014	0.247	0.012	0.727
Coyaima	0.025	0.184	0.034	0.757
Pijao	0.041	0.405	0.072	0.482
Embera	0.119	0.133	0.035	0.712
Caracas	0.214	0.595	0.005	0.186
Maracaibo	0.209	0.518	0.016	0.256
Wayu	0.074	0.128	0.128	0.67
Guatemala City	0.064	0.298	0.014	0.623
Q'eqchi	0.016	0.133	0.022	0.829
Chile – North	0.034	0.512	0.056	0.399
Chile – South	0.021	0.567	0.025	0.387
MXL	0.052	0.456	0.044	0.448
PUR	0.151	0.711	0.018	0.119
São Paulo	0.279	0.657	0.032	0.032
Kalunga	0.664	0.27	0.005	0.062
Sacutiaba	0.515	0.379	0.027	0.079
ASW	0.739	0.213	0.008	0.04

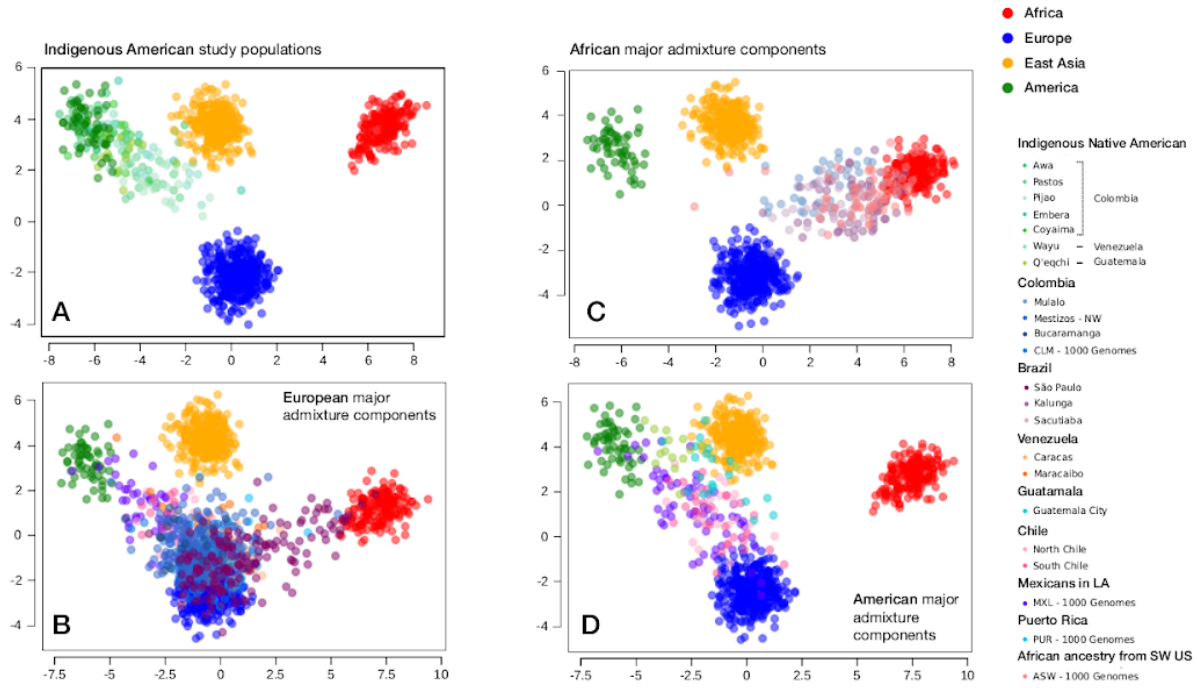


Fig. 5. PCAs of American study population subsets arranged in four plots. (A) Indigenous American study populations (Colombia: Awa, Pastos, Embera, Pijao, Coyaima; Venezuela: Wayu; Guatemala: Q'eqchi). (B) American admixed populations with high European ancestry (Colombia: CLM, NW Colombia, Bucaramanga; Venezuela: Caracas, Maracaibo; Brazil: São Paulo; Chile: North and South; MXL; PUR). (C) American admixed populations with high African ancestry (Brazil: Kalunga, Sacutiaba; Colombia: Mulaló; ASW). (D) Urban populations with high American ancestry (MXL, Guatemala City, and north/south Chile).

Patterns in the eleven urban populations or samples from cosmopolitan regions with high population density (e.g. Puerto Rico) had predominantly European ancestry (82% of individuals >50%, and in descending order of Puerto Rico > CLM-Medellín > São Paulo > Caracas). More than 66% of urban sample individuals had American co-ancestry as the second admixture component with the range 21-45%, while 33% had African co-ancestry as a second component (15-28%). The MXL and Guatemala City samples had high American co-ancestry levels of 45% and 62%, respectively, with European and American components divergent in less than

1% of MXL individuals. Notably, Guatemalans from Guatemala City are very similar in their admixture profiles to Q'eqchi, with individual estimates ranging from 40-80% and 65-90%, respectively. The 30% European and 62% American contribution positions the Guatemala City samples between the European and American clusters in PCA. In Chile, both north and south samples have relatively high European co-ancestry of 51-57%, (American 39-40%) but a much lower African component of <3%, consistent with the lower level of historic African slavery in regions southwest of the Andes divide [30,31].

Not surprisingly, the indigenous American populations that have minimal contact with people from other parts of the country/region in which they were sampled, show very high American ancestry estimates recorded as cluster proportions (in descending order of Awa > Q'eqchi > Coyaima > Pastos > Embera). Pastos, Coyaima and Embera had relatively high European components of 13.3%, 18.4% and 24.7%, respectively. It is noteworthy, that the Colombian Awa is the only indigenous population apart from the HGDP-CEPH Amazonian Karitiana and Surui ('A' columns in Fig. 4A), to show almost completely unadmixed patterns of American ancestry. This pattern is also evident in the PCA (Fig. 5) and pairwise F_{ST} heatmap (Supplementary Fig. S4). Populations with predominant African co-ancestry contributions (Colombian Mulaló, Brazilian *Quilombos* Kalunga and Sacutiaba, ASW) show admixture patterns as variable as those detailed above for American co-ancestry, but equally compatible with their known history and demographics [32-34]. The *Quilombos* are Brazilian populations with significant African descent that have remained genetically isolated to some extent from surrounding populations [32].

Likewise, the Mulaló have lived in relative isolation in west-central Colombia, where gold mining and coffee or sugar cane plantations relied heavily on African slave workers.

Venezuelan Caracas and Maracaibo urban populations both showed three-way admixture patterns with a higher than average European component (>52%), and similar African and American components of ~21% and 18-25%, respectively. The Wayu are the largest indigenous population in Venezuela, and inhabit mostly the Guajira Peninsula, shared with Colombia. They have a history of contact with Europeans and other non-American populations since colonial times, so consequently a minor European component is detectable in their gene pool (>12%).

Unlike the other urban populations analyzed, individuals from São Paulo are widely spread between African and European components in the European admixture PCA (Fig. 5B). Brazil received ~40% of all Africans that were forcibly enslaved and brought to the Americas (3.6-10 million people [33]), and remains a country where social strata are still highly correlated with a person's ancestry and appearance. One outcome from developing PIMA was an enhanced ability to properly gauge the American admixture component in the urban population of São Paulo. Indigenous Brazilian populations were reduced from 2-3 million people to a current estimated ~900,000 following contact with Europeans, with contemporary indigenous groups now mostly located in Northern and Central-Western states. Overall, a three-way admixture pattern describes most of the country's demographic profile, with high levels of European co-ancestry with lower but highly variable African and American co-ancestry (data reviewed in [34]). The São Paulo sample

used here was collected for a case-control study on hypercholesterolemia [35] - underlining the value of simple SNP panels like PIMA to efficiently detect and control ancestry bias in donor samples used in medical genetics. Furthermore, applying the more powerful combination of PIMA+34plex SNPs led to the detection of East Asian co-ancestry in the São Paulo sample, forming a four-way admixture pattern of a predominant European (65%) and African (28%) co-ancestry level overall, but with detectable and consistent East Asian and American admixture components of ~3.5% each. The fact that these two closely related population groups were distinguished despite being present at minimal levels of co-ancestry also highlights the population differentiation power of PIMA, for which it was designed.

3.5. Comparing data from PIMA+34plex and uniparental markers in a small sample of admixed Venezuelans

Uniparental marker data from a small sample (n=8) from Caracas, Venezuela, were compared to ancestry estimates obtained from PIMA+34plex autosomal SNP genotypes. Supplementary Fig. S7 shows the inferred ancestral origin of each individual's mtDNA and Y-chromosome variation and the co-ancestral proportions from their autosomal SNP variation. Although a limited sampling, the overall picture observed matches previous findings of biased mating between European males with females mostly from other ancestries, and most commonly American. In our sample, European Y-chromosome variation was seen in 7 out of 8 individuals and American mtDNA variation in 5 of 8. The total autosomal SNP variation detected a higher European ancestry proportion than seen in mtDNA variation, which itself represented

a higher proportion of American variation; underlining the inferred mainly European male lineages and mainly American female lineages, consistent with sex bias seen in admixture in the Americas. Samples VEN-03 and VEN-18 show a minor American co-ancestry proportion (<10%) and American mtDNA variation, suggesting population admixture was a less recent event in these individuals compared to some of the others.

These patterns were also previously seen in our studies of Brazilian Kalunga and Sacutiaba, where no European mtDNA were found (reviewed in [32]). Studies of other groups of the Colombian population of Antioquia [35,36], found >94% of Y-chromosome lineages were of European origin, while mtDNA was >90% American. In Chile, where autosomal AIMs have shown that ancestry is fairly homogeneous across the country (as shown applying PIMA+34plex in another study [38]), the same sex bias was observed: mtDNA variation was mostly American (82-95%) [30] and Y-chromosome variation mostly European (91.7%) [31]. Overall, such observations emphasize the importance of analyzing both autosomal and uniparental lineage single markers to adequately describe both a population's history and an individual's likely co-ancestry.

4. Concluding remarks

Our results show that the five centuries of contact and admixture among the populations meeting in America has produced complex and varied population structure and co-ancestry patterns amongst all the populations we studied, especially in non-admixed indigenous peoples. Despite a severe reduction in indigenous

American population sizes during this time, American co-ancestry persists in varied degrees in contemporary populations across the continent. This indigenous American gene pool has also survived in places where it might be expected to have been reduced by a strong influx of Europeans - notably in the urban samples we analyzed from Guatemala City, Chile and Mexicans in Los Angeles.

The PIMA panel's capacity to differentiate American and East Asian populations to a level equivalent to, or better than, a much larger genomic ancestry control panel comprising 314 SNPs, exceeded our expectations for the goal of adjusting the shortfall of the 34plex panel that had been identified from initial comparisons of HGDP-CEPH Americans to other populations. A combined PIMA+34plex panel has a more balanced informativeness content and can identify and differentiate American-indicative variants, even when present in relatively minor proportions in admixed individuals. This is substantiated by the detection of all four population group components in the complex admixture patterns of the São Paulo sample, with levels of American and East Asian co-ancestry below 5%. We recognize that inclusion of the seven SNPs in a full 34plex panel, absent due to poor performance when testing study populations, is likely to boost the ancestry differentiation performance of the PIMA+34plex sets still further, notably the most powerful East Asian-informative SNP in 34plex: rs3827760 [16]. Nevertheless, the small scale and high differentiating power of PIMA (particularly in combination with 34plex) makes it an invaluable and practical tool for population genetics, medical studies and forensic DNA analysis.

Acknowledgments

The authors would like to thank the populations who kindly agreed to participate in the study and field teams who made sample collection possible. AFA is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010). CCG was supported by a doctoral scholarship funded by CNPq.

Reference

[1] J.K. Pritchard, P. Donnelly, Case-control studies of association in structured or admixed populations, Theor. Popul. Biol. 60 (2001) 227-237.

[2] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of Genetic Markers for Inference of Ancestry, Am. J. Hum. Biol. 73 (2003) 1402-1422.

[3] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. 18 (2015) 49–65.

[4] T. Pinotti, A. Bergström, M. Geppert, M. Bawn, D. Ohasi, W. Shi, D.R. Lacerda, A. Solli, J. Norstedt, K. Reed, et al., Y Chromosome Sequences Reveal a Short Beringian Standstill, Rapid Expansion and early Population structure of Native American Founders, Curr. Biol. 29 (2019) 149-157.

[5] A. Gómez-Carballa, J. Pardo-Seco, S. Brandini, A. Achilli, U.A. Perego, M.D. Coble, T.M. Diegoli, V. Álvarez-Iglesias, F. Martinón-Torres, A. Olivieri, et al., The peopling of South America and the trans-Andean gene flow of the first settlers, Genome Res. 28 (2018) 767-779.

[6] F.M. Salzano, M.C. Bortolini, The Evolution and Genetics of Latin American Populations. Cambridge Univ Press. 2005; 28.

[7] G.F. Shields, A.M. Schmiechen, B.L. Frazier, A. Redd, M.I. Voevoda, J.K. Reed, R.H. Ward, mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations, Am. J. Hum. Genet. 53 (1993) 549-562.

[8] P.A. Underhill, L. Jin, R. Zemans, P.J. Oefner, L.L. Cavalli-Sforza, A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history, Proc. Natl. Acad. Sci. USA 93 (1996) 196-200.

[9] A. Achilli, U.A. Perego, C.M. Bravi, M.D. Coble, Q.P. Kong, S.R. Woodward, A. Salas, A. Torroni, H.J. Bandelt, The phylogeny of the four pan-American MtDNA haplogroups: Implications for evolutionary and disease studies, PLoS One. 3 (2008) e1764.

[10] C. Phillips, A. Rodriguez, A. Mosquera-Miguel, M. Fondevila, L. Porras-Hurtado, F. Rondon, A. Salas, Á. Carracedo, M.V. Lareu, D9S1120, a simple STR with a common Native American-specific allele: Forensic optimization, locus characterization and allele frequency studies, Forensic Sci Int Genet 3 (2008) 7-13.

[11] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L.U. Figueroa, P. Raska, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet. 8 (2012) e1002554.

[12] A. Salas, A. Acosta, V. Álvarez-Iglesias, M. Cerezo, C. Phillips, M.V. Lareu, Á. Carracedo, The mtDNA ancestry of admixed Colombian populations, Am. J. Hum. Biol. 20 (2008) 584–591.

[13] T.E. King, E.J. Parkin, G. Swinfield, F. Cruciani, R. Scozzari, A. Rosa, S.K. Lim, Y. Xue, C. Tyler-Smith, M.A. Jobling, Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy, Eur. J. Hum. Genet. 15 (2007) 288–293.

[14] G. Bedoya, P. Montoya, J. Garcı, I. Soto, S. Bourgeois, L. Carvajal, D. Labuda, V. Alvarez, J. Ospina, P.W. Hedrick, A. Ruiz-Linares, Admixture dynamics in

Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate, Proc. Natl. Acad. Sci. USA.103 (2006) 7234-7239.

[15] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci Int Genet. 1 (2007) 273-280.

[16] M. Fondevila, C. Phillips, C. Santos, A. Freire Aradas, P.M. Vallone, J.M. Butler, M.V. Lareu, Á. Carracedo, Revision of the SNP*for*ID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies, Forensic Sci. Int. Genet. 7 (2013) 63-74.

[17] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H. M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.

[18] J. Amigo, C. Phillips, M. Lareu, Á. Carracedo, The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project, Int. J. Legal Med. 132 (2018) 435–440.

[19] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinformatics 9 (2008) 1-6.

[20] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.

[21] A. Gómez-Carballa, A. Ignacio-Veiga, V. Álvarez-Iglesias, A. Pastoriza-Mourelle, Y. Ruíz, L. Pineda, Á. Carracedo. A. Salas, A melting pot of multi-continental mtDNA lineages in admixed Venezuelans, Am. J. Phys. Anthropol. 147 (2012) 78-87.

[22] L. Excoffier, G. Laval, S. Schneider, Arlequin (version 3.0): An integrated software package for population genetics data analysis, Evol. Bioinform. Online 1 (2005) 47-50.

[23] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.

[24] M. Jakobsson, N.A. Rosenberg, CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801-1806.

[25] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study, Mol. Ecol. 14 (2005) 2611-2620.

[26] N.A. Rosenberg, DISTRUCT: A program for the graphical display of population structure, Mol. Ecol. Notes 4 (2004) 137-138.

[27] F.M. Mendes, C.C. Gontijo, Kpop: A Python package for population genetics analysis, Forensic Sci. Int. Genet. Suppl. Ser. 6 (2017) e407-409.

[28] J.D. Hunter, Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering. 9 (2007) 90-95.

[29] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381–2385.

[30] A. Gómez-Carballa, F. Moreno, V. Álvarez-Iglesias, F. Martinón-Torres, M. García-Magariños, J.A. Pantoja-Astudillo, E. Aguirre-Morales, P. Bustos, A. Salas, Revealing latitudinal patterns of mitochondrial DNA diversity in Chileans, Forensic Sci. Int. Genet. 20 (2016) 81-88.

[31] U. Toscanini, F. Brisighelli, F. Moreno, J.A. Pantoja-Astudillo, E.A. Morales, P. Bustos, J. Pardo-Seco, A. Salas, Analysis of Y-chromosome STRs in Chile confirms an extensive introgression of European male lineages in urban populations, Forensic Sci. Int. Genet. 21 (2016) 76-80.

[32] C.C. Gontijo, F.M. Mendes, C.A. Santos, M.N. Klautau-Guimarães, M.V. Lareu, Á. Carracedo, C. Phillips, S.F. Oliveira, Ancestry analysis in rural Brazilian populations of African descent, Forensic Sci Int Genet. 36 (2018) 160–166.

[33] J.C. Miller, Way of Death: Merchant Capitalism and the Angolan Slave Trade, 1730–1830, University of Wisconsin Press; 1997.

[34] R. Rodrigues de Moura, A.V.C. Coelho, V. de Queiroz Balbino, S. Crovella, L.A.C. Brandão, Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries, Am. J. Hum. Biol. 27 (2015) 674-680.

[35] V.N. Silbiger, M.H. Hirata, A.D. Luchessi, F.D.V. Genvigir, A. Cerda, A.C. Rodrigues, M.A. Willrich, S.S. Arazi, E.L. Dorea, M.M. Bernik, et al. Differentiation of African Components of Ancestry to Stratify Groups in a Case–Control Study of a Brazilian Urban Population, Genet. Test Mol. Biomarkers 16 (2012) 524-530.

[36] L.G. Carvajal-Carmona, I.D. Soto, N. Pineda, D. Ortíz-Barrientos, C. Duque, J. Ospina-Duque, M. McCarthy, P. Montoya, V.M. Alvarez, G. Bedoya, A. Ruiz-Linares, Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in Northwest Colombia, Am. J .Hum. Genet. 67 (2000) 1287-95.

[37] L.G. Carvajal-Carmona, R. Ophoff, S. Service, J. Hartiala, J. Molina, P. Leon, J. Ospina, G. Bedoya, N. Freimer, A. Ruiz-Linares, Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica, Hum. Genet. 112 (2003) 534-541.

[38] F. Moreno, A. Freire-Aradas, C. Phillips, M. Fondevila, Á. Carracedo, M.V. Lareu, SNP variation with latitude: Analysis of the SNP*for*ID 52-plex markers in north, mid-region and south Chilean populations. Forensic Sci. Int. Genet. 10 (2014) 12-16.

Figure captions

Fig. 1. Map of American study population and HGDP-CEPH reference population sampling locations.

Fig. 2. Summary bar-plots of PIMA SNP variation in the four main population groups of 1000 Genomes (1KG) and HGDP-CEPH indigenous Americans. SNPs are ordered left to right in descending allele frequency differential values (delta) comparing the 1KG 4-group average allele frequencies (non-AME) vs CEPH American (AME) frequencies (not calculated for the tri-allelic SNP rs17287498 and X-SNP rs3027749). The non-American allele frequencies of rs17287498 were estimated from gnomAD population data (lacking SAS). Additional plots shown for the 1KG 4-group average frequencies and 1KG Peruvians from Lima (PEL). RA: reference allele; EUR: Europeans; AFR: Africans; SAS: South Asian; EAS: East Asian.

Fig. 3. Evaluation of the population differentiation performance of individual PIMA and 34plex panels and the combined set. **Fig. 3A.** PCA tests of 3 SNP sets analyzing 1000 Genomes African, European, East Asian; and HGDP-CEPH American genotype data. **Fig. 3B.** Cumulative I_n charts of PCA tests of 3 SNP sets, adding most powerful loci first. **Fig. 3C.** Cross validation ancestry assignment success of 3 SNP sets, with error highlighted for the relevant population group.

Fig. 4A. STRUCTURE cluster plots for K:3 and K:4 inferred clusters comparing PIMA+34plex and LACE panels for a comprehensive set of common reference and American study samples. Cluster bar plots ordered in both SNP panels by increasing majority co-ancestry from LACE cluster membership proportions (CLUMPP-merged from 5 runs). HGDP-CEPH populations grouped and marked as: A=Amazonian Karitiana /Surui; M=Maya; P=Pima. **Fig. 4B.** Cross validation ancestry assignment

rates for the reference population samples. Fig. **4C.** Comparisons of mean population differentiation metrics F_{ST} and I_n in each panel.

Fig. 5. PCAs of American study population subsets arranged in four plots. **(A)** Indigenous American study populations (Colombia: Awa, Pastos, Embera, Pijao, Coyaima; Venezuela: Wayu; Guatemala: Q'eqchi). **(B)** American admixed populations with high European ancestry (Colombia: CLM, NW Colombia, Bucaramanga; Venezuela: Caracas, Maracaibo; Brazil: São Paulo; Chile: North and South; MXL; PUR). **(C)** American admixed populations with high African ancestry (Brazil: Kalunga, Sacutiaba; Colombia: Mulaló; ASW). **(D)** Urban populations with high American ancestry (MXL, Guatemala City, and north/south Chile).

Supplementary tables

(available at https://github.com/ninacalorina/supplementary-material-thesis)

Supplementary Table S1A. Genotype data used in all population analyses; comprising 25 of 27 PIMA and 27 of 34plex SNPs. REF: Reference population data. Table S1B. HGDP-CEPH population data (genotypes and population-based allele frequency estimates) for tri-allelic SNP rs17287498, not genotyped for American study populations but developed for the final version of the PIMA multiplex.

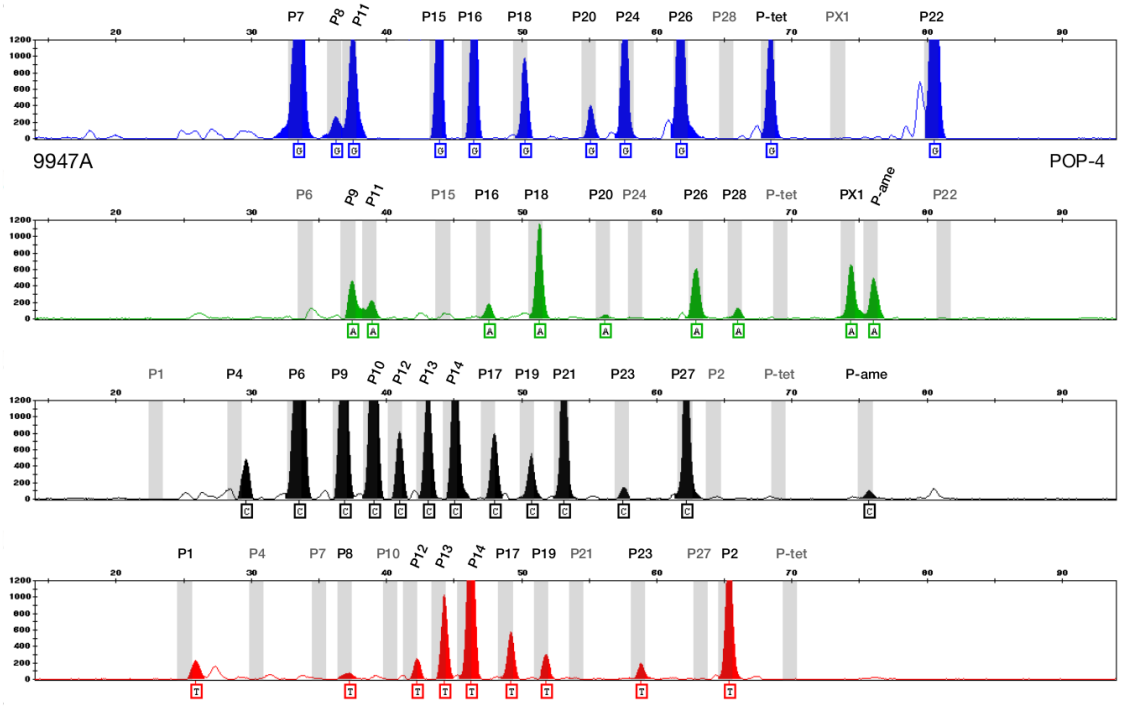
Supplementary Table S2A. Genomic details of the 28 markers of the PIMA panel. Table S2B. PCR primer and SNaPshot extension (EXT) primer details of the final PIMA multiplex design. Table S2C. GRCh37 coordinates of 62 SNPs of the PIMA and 34plex panels. The separation in megabases (Mb) of syntenic SNPs is shown and the three closest SNP pairs marked in red.

Supplementary Table S3A. Descriptive population genetics parameters for PIMA+34plex in study and reference populations: Hardy-Weinberg equilibrium (HWE) tests, observed (Hobs) and expected (Hexp) heterozygosity, monomorphic allele rates. **Table S3B.** Pairwise linkage disequilibrium analyses. **Table S3C.** Allele frequencies (25/27 PIMA SNPs; 27/34 34plex SNPs). Allele calls are for SNaPshot-detected nucleotide substitutions.

Supplementary figures

Supplementary Figure S1.

A



Supplementary Fig. S1A. Typical PIMA SNaPshot profile of 9947A control DNA.

B

P1	rs2471552	TT
P4	rs17742080	CC
P6	rs4979274	CC
P7	rs1863086	GG
P8	rs17130385	GT
P9	rs4780476	AC
P10	rs12498138	CC

P11	rs3814614	AG
P12	rs12402499	CT
P13	rs4937417	CT
P15	rs2051827	GG
P16	rs10127540	AG
P17	rs6993205	CT
P18	rs379839	AG

P19	rs3217863	CT
P20	rs252155	AG
P21	rs10012227	CC
P23	rs8137373	CT
P24	rs1701930	GG
P26	rs1409842	AG
P27	rs9916327	CC

P2	rs509360	TT
P28	rs7158003	AA
P-tet	rs17287498	GG
PX1	rs3027749	AA
P-ame	AMELOGENIN	XY
P22	rs1834619	GG

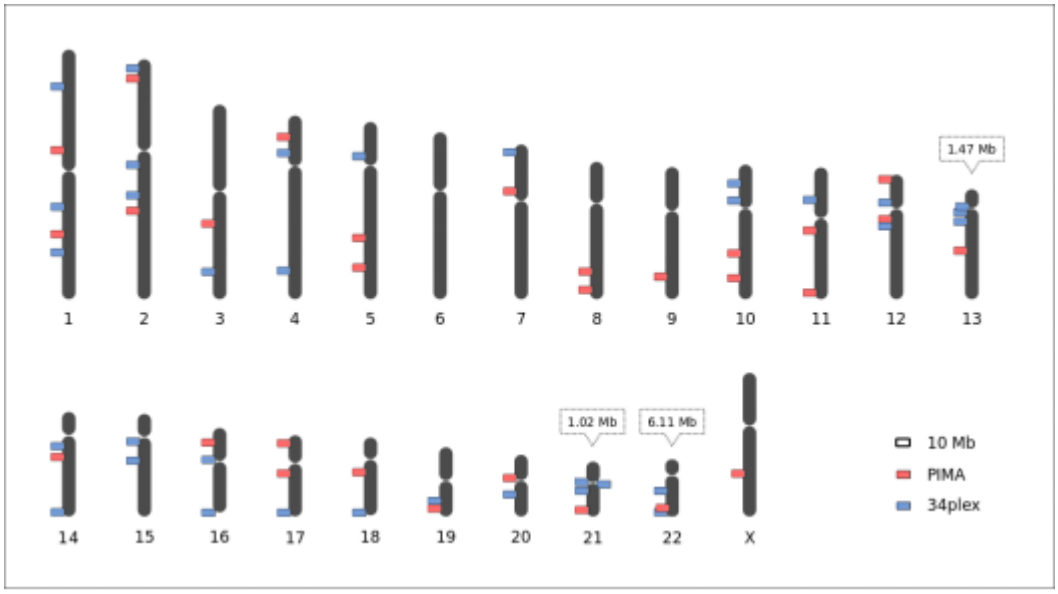
Supplementary Figure S1B 9947A genotypes.

C

Example SNaPshot EXT product size estimates for 9947A control DNA (POP-4)

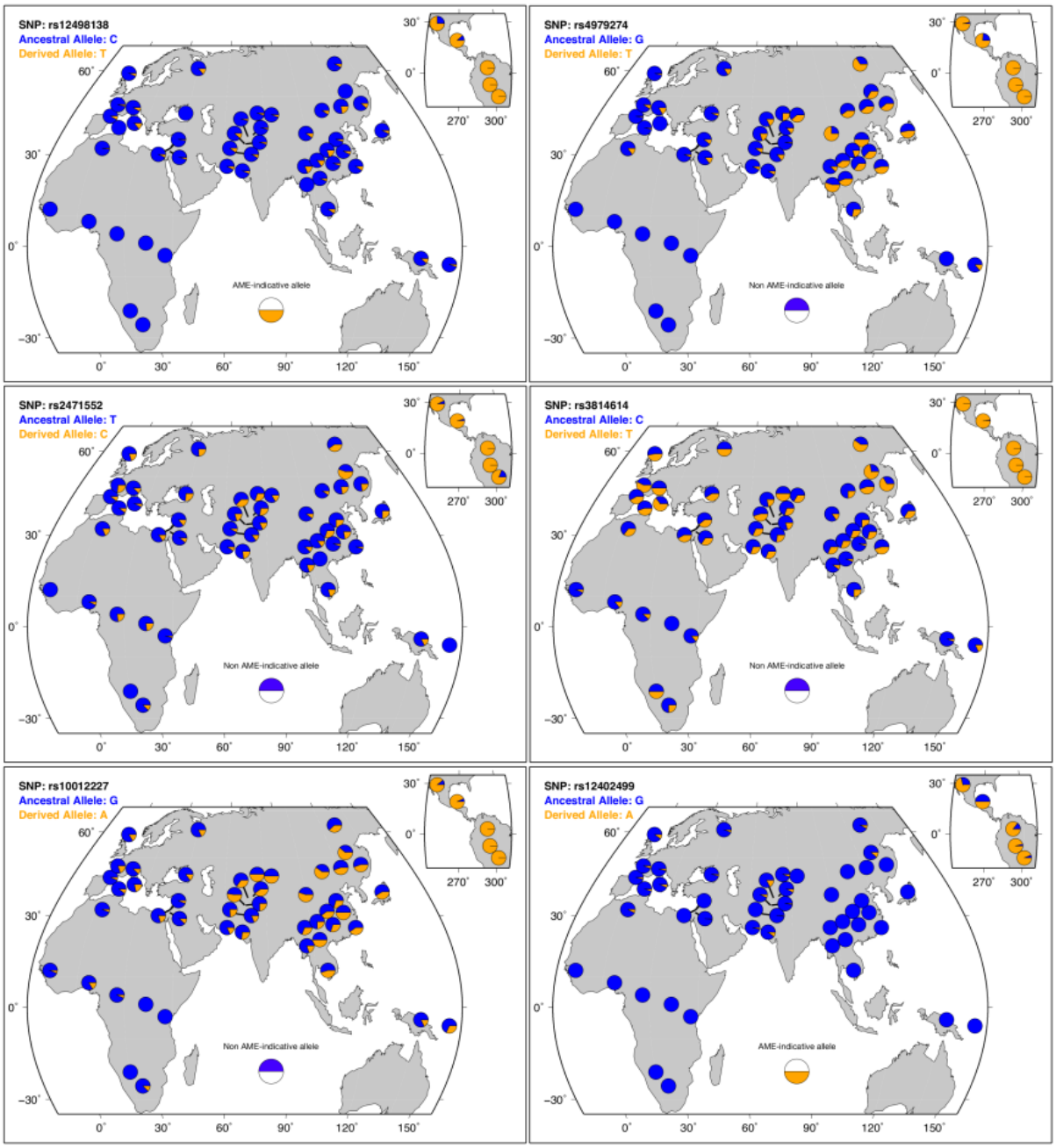
9947A	Internal SNP code for SNaPshot peaks	G	A	C	T
TT	P1			22.93	25.06
CC	P4			28.76	30.35
CC	P6		33.98	33.19	
GG	P7	33.25			35.01
GT	P8	36.10			36.91
CA	P9		37.14	36.57	
CC	P10			38.78	40.25
GA	P11	37.27	38.75		
CT	P12			40.62	41.77
CT	P13			42.74	43.87
CT	P14			44.67	45.73
GG	P15	43.57	44.16		
GA	P16	46.09	47.16		
CT	P17			47.50	48.80
GA	P18	49.91	51.00		
CT	P19			50.36	51.44
GA	P20	54.96	56.01		
CC	P21			52.94	54.02
CT	P23			57.40	58.60
GG	P24	57.73	58.41		
GA	P26	61.77	62.87		
CC	P27			62.08	63.23
AA	P28	65.15	65.78		
TT	P2			64.19	65.06
GG	P-tet	68.20	69.14	69.47	71.25
AA	PX1	73.39	74.12		
CA	P-ame		75.79	75.42	
GG	P22	80.28	81.23		

Supplementary Figure S1C SNaPshot peak size estimates in POP4.

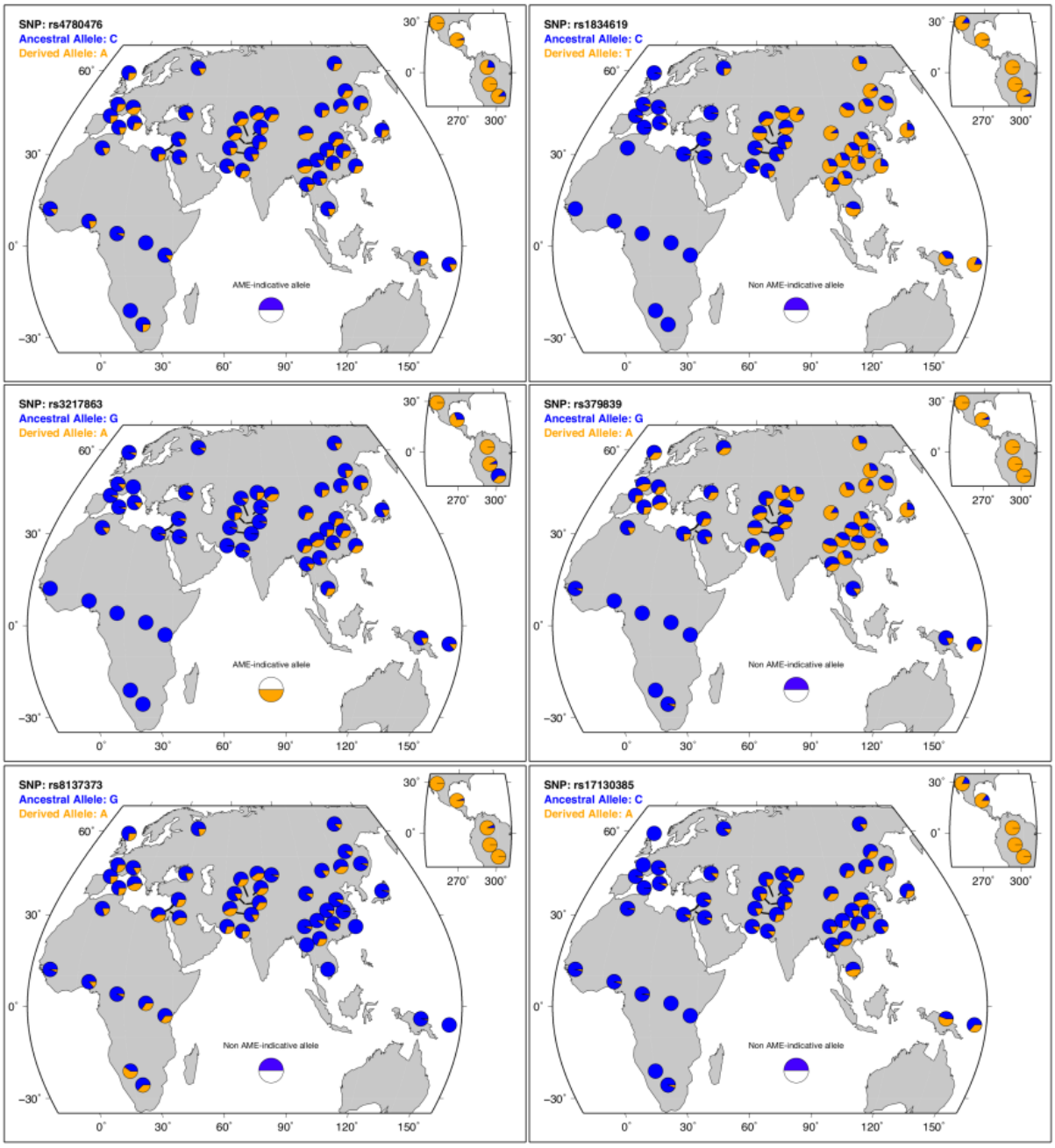


Supplementary Figure S2. Genomic positions of PIMA and 34plex SNPs. Three closest SNP pairs marked with their separation in megabases (Mb). GRCh37 genomic coordinates given in Supplementary Table S2C.

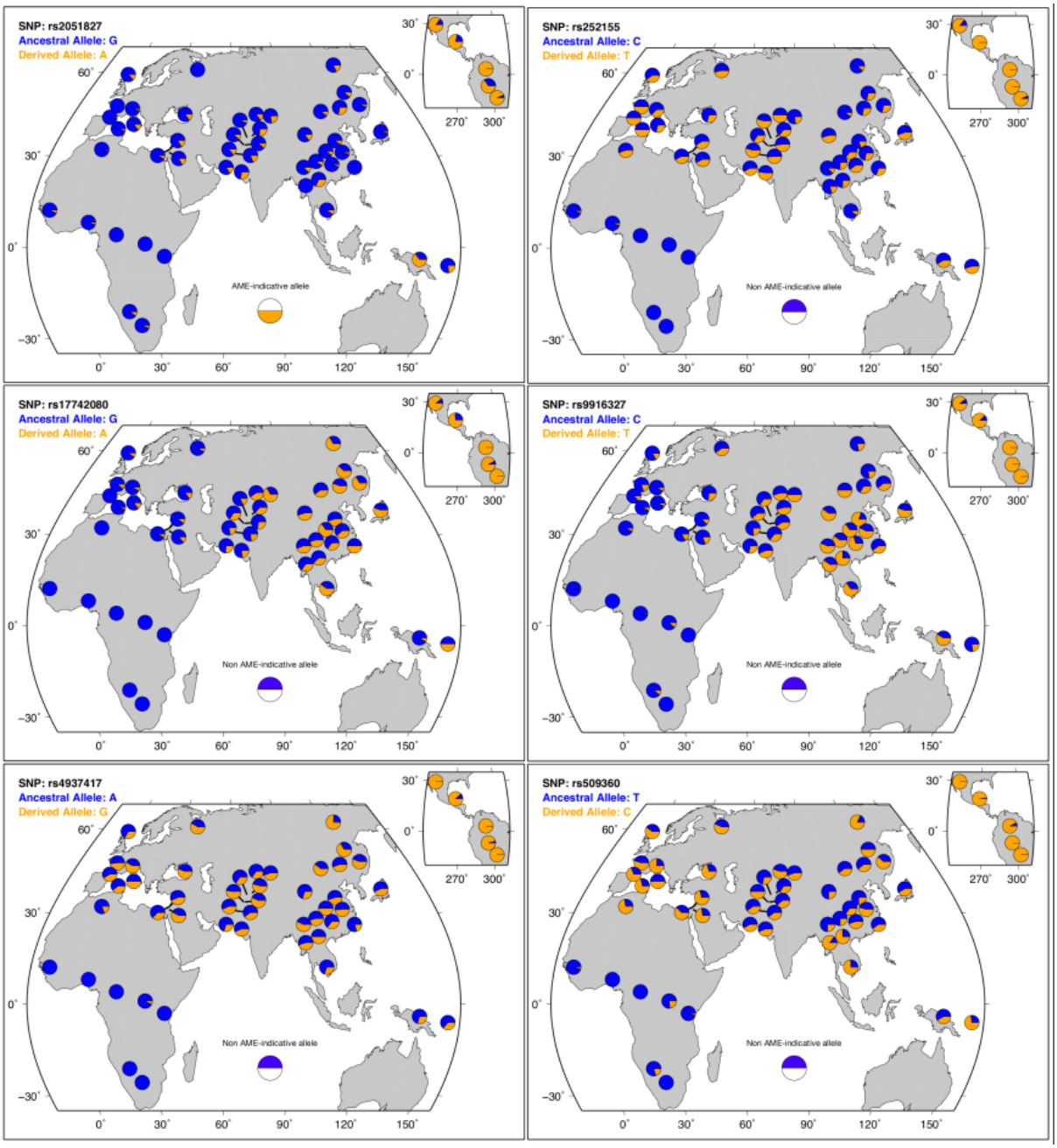
Supplementary Figure 3.



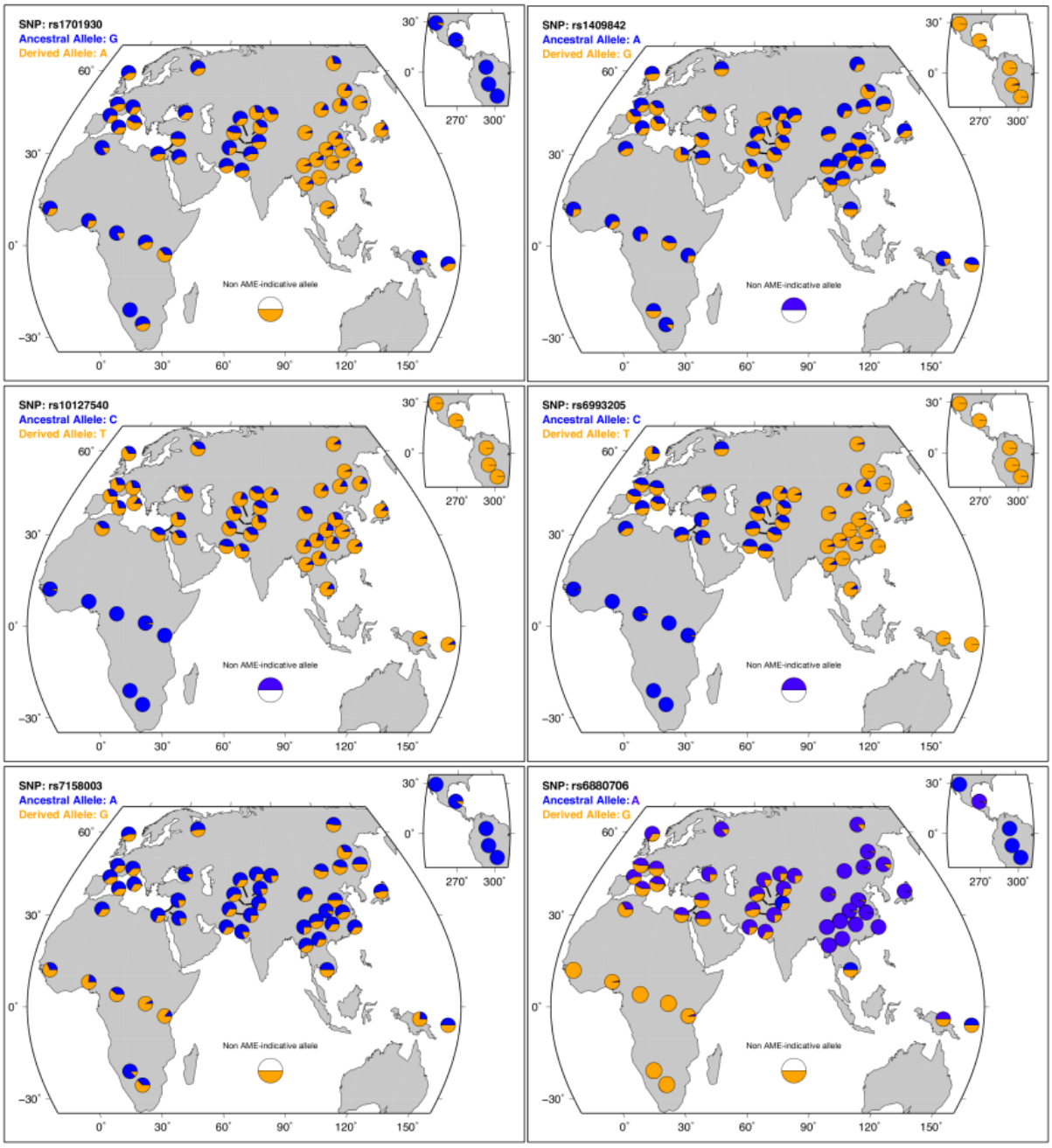
Supplementary Figure S3. Individual PIMA autosomal SNP allele frequency pie charts for 53 HGDP-CEPH populations. **(cont.)**



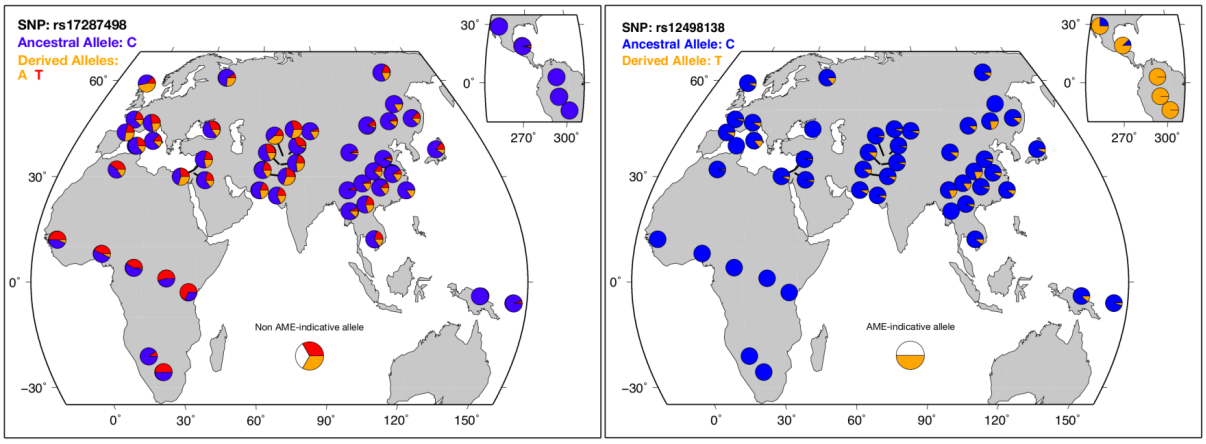
Supplementary Figure S3. Individual PIMA autosomal SNP allele frequency pie charts for 53 HGDP-CEPH populations. **(cont.)**



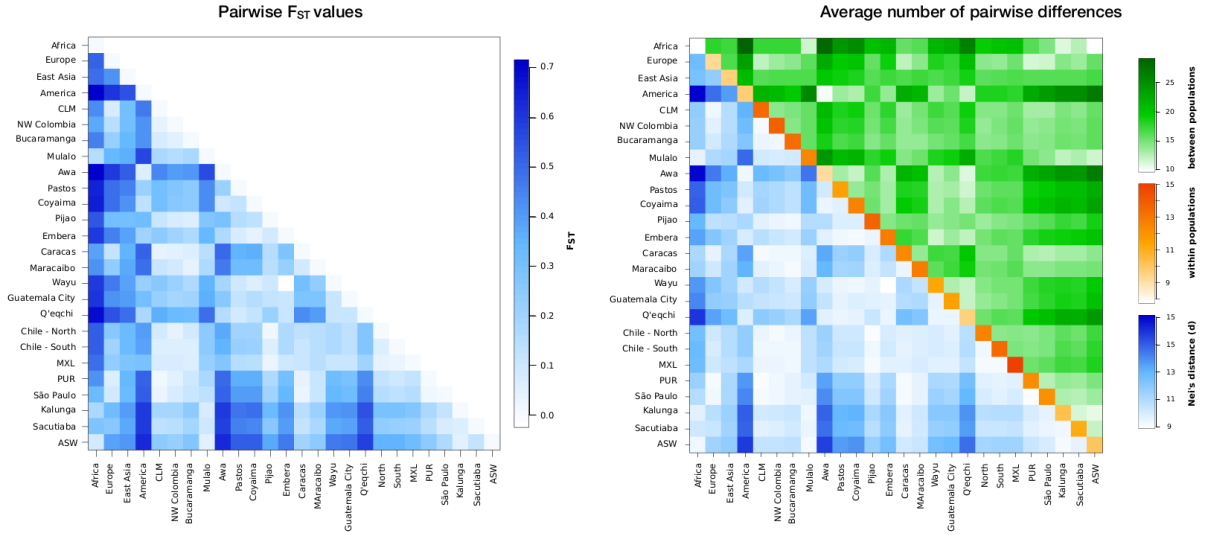
Supplementary Figure S3. Individual PIMA autosomal SNP allele frequency pie charts for 53 HGDP-CEPH populations. **(cont.)**



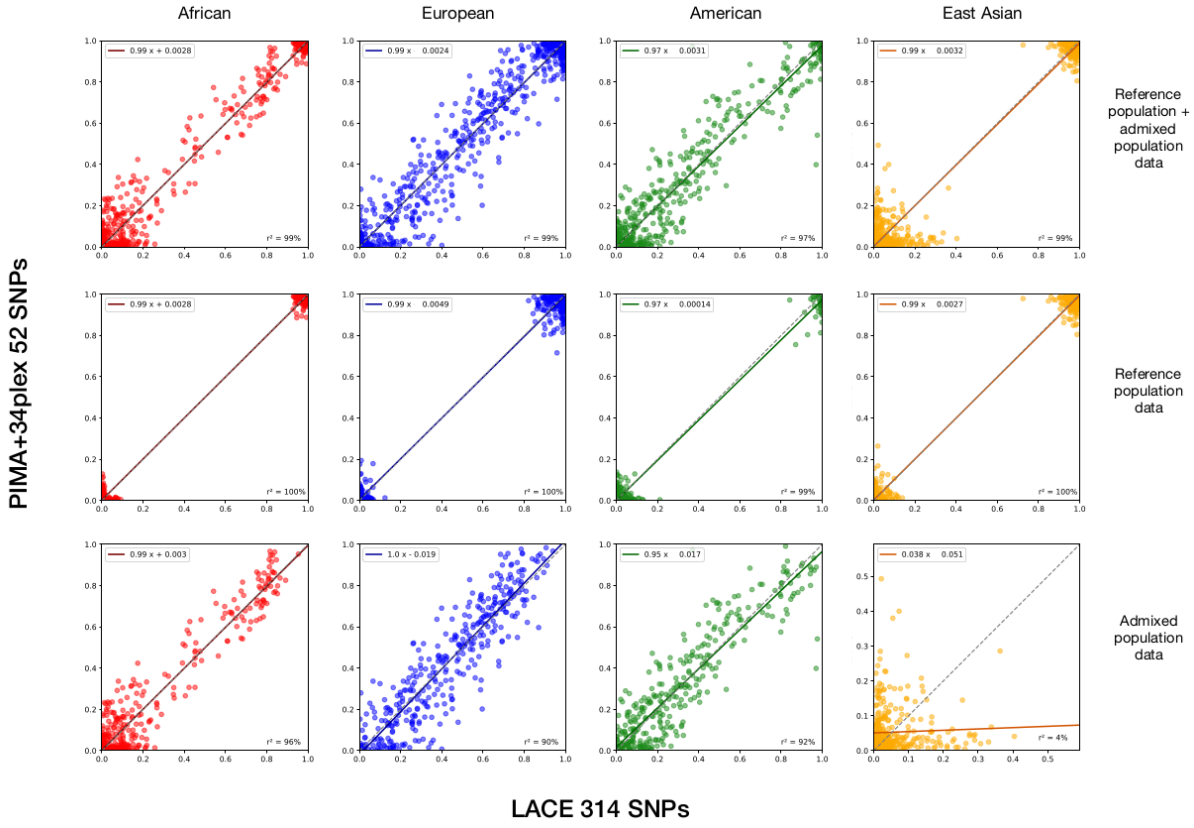
Supplementary Figure S3. Individual PIMA autosomal SNP allele frequency pie charts for 53 HGDP-CEPH populations. **(cont.)**



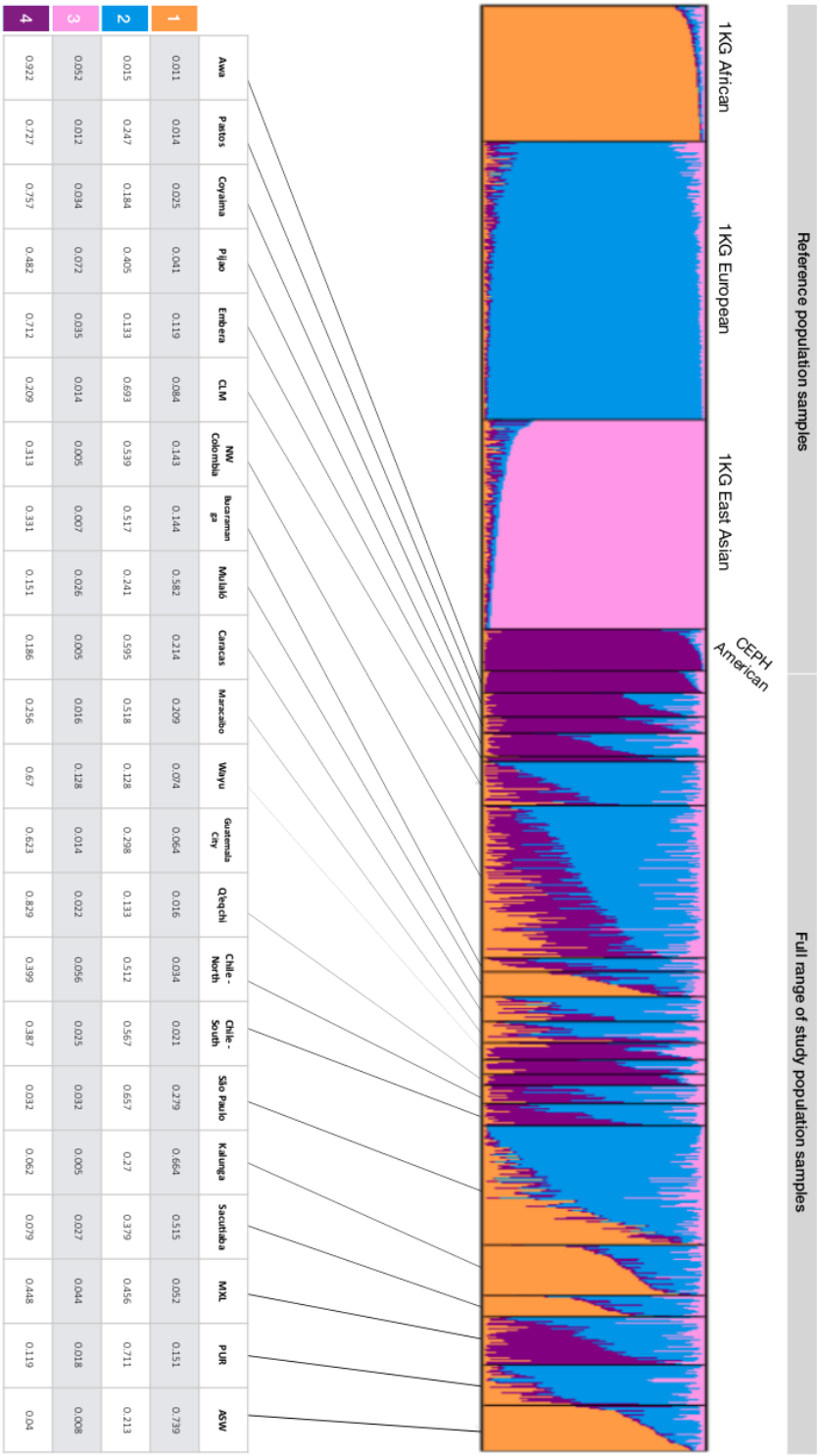
Supplementary Figure S3. Individual PIMA autosomal SNP allele frequency pie charts for 53 HGDP-CEPH populations.



Supplementary Figure S4. Pairwise population comparisons for F_{ST} (left chart), average number of pairwise differences between (right chart, upper right) and within populations (right chart, diagonal) and Nei's distance (right chart, lower left).

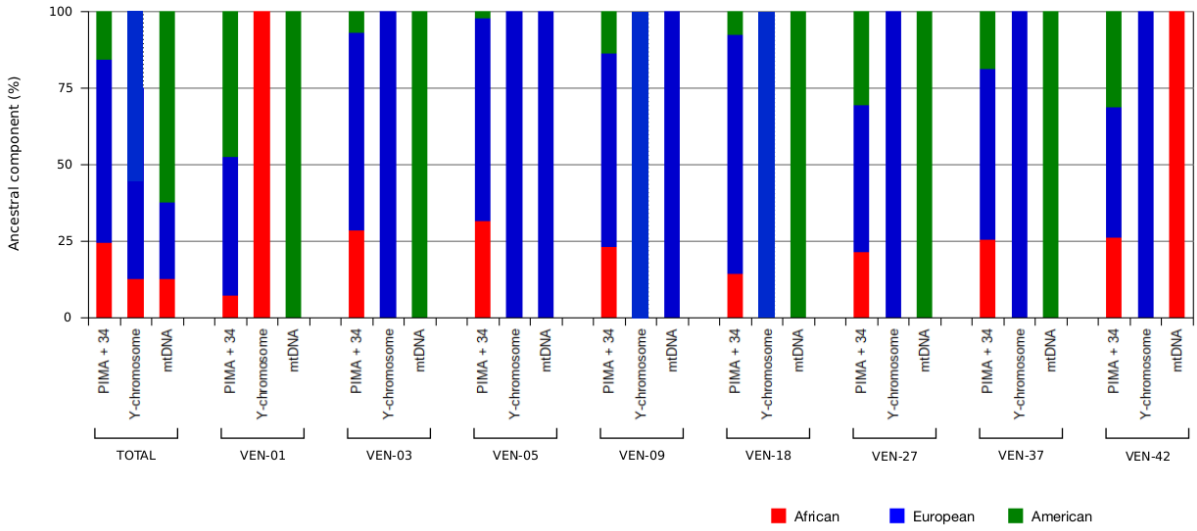


Supplementary Figure S5. Correlation plots for each comparing STRUCTURE cluster membership proportions (K:4 inferred clusters for African, European, American and East Asian co-ancestries) estimated using PIMA+34plex SNPs vs LACE SNPs.



	1	2	3	4
Awa	0.011	0.015	0.052	0.922
Parbos	0.014	0.247	0.012	0.727
Coyaima	0.025	0.184	0.034	0.757
Pijao	0.041	0.405	0.072	0.482
Embera	0.119	0.133	0.035	0.712
CLM	0.084	0.693	0.014	0.209
NW Colombia	0.143	0.539	0.005	0.313
Bucaramanga	0.144	0.517	0.007	0.331
Mulalo	0.582	0.241	0.026	0.151
Caracas	0.214	0.595	0.005	0.186
Maracaibo	0.209	0.518	0.016	0.256
Wayu	0.074	0.128	0.128	0.67
Guatemala City	0.064	0.298	0.014	0.623
Qeqchi	0.016	0.133	0.022	0.829
Chile - North	0.034	0.512	0.056	0.399
Chile - South	0.021	0.567	0.025	0.387
São Paulo	0.279	0.657	0.032	0.032
Kaluinga	0.664	0.27	0.005	0.062
Sacuridaba	0.515	0.379	0.027	0.079
MXL	0.052	0.456	0.044	0.448
PUR	0.151	0.711	0.018	0.119
ASW	0.739	0.213	0.008	0.04

Supplementary Figure S6. Cluster plot from STRUCTURE analysis of the full range of American study population samples - using PIMA+34plex SNPs and same reference genotypes as Fig. 2 run. Individual study samples ordered by ascending American or African co-ancestry (K4/K1 cluster membership proportions).



Supplementary Figure S7. Ancestry assignments of mtDNA and Y-chromosome variation, plus estimated co-ancestry proportions from PIMA+34plex SNPs in 8 admixed individuals from Caracas, Venezuela

4. Avaliação geral das estimativas de ancestralidade

4.1 Apresentação do problema

A estimativa de ancestralidade é uma ferramenta útil na redução do escopo de investigações, na situação forense na construção de amostras para estudos com desenho experimental caso-controle, na genética médica, e um objetivo *per se* para a genética antropológica. Sua acurácia depende do uso de marcadores adequados ao contexto de mistura em análise, mas apesar de existirem métricas que guiam a escolha de marcadores e o desenvolvimento de painéis para estimativa de ancestralidade, não há protocolo estabelecido para a avaliação de seu desempenho. Assim, cabe ao pesquisador selecionar o sistema que lhe pareça adequado tendo em conta estudos prévios, critérios objetivos (como informatividade, balanceamento do conteúdo de informação entre componentes parentais, parentais acessadas pelo sistema e disponibilidade de dados públicos) e fatores práticos (como estrutura laboratorial requerida, e viabilidade financeira).

4.2 Introdução

O objetivo essencial da genética forense é determinar a que indivíduo pertence uma dada amostra biológica (Jobling e Gill, 2004). Ela hoje se vale de metodologias que dependem, via de regra, da comparação de perfis genéticos obtidos da amostra de interesse àqueles obtidos de suspeitos ou depositados em bancos de dados (Kayser e Schneider, 2009). Os sistemas utilizados para esse fim são os desenvolvidos para identificação individual - em especial o CODIS

(Combined DNA Index System). Esses sistemas são compostos por marcadores com: 1. grande variabilidade intrapopulacional, o que lhes confere alto poder de discriminação e reduz a probabilidade de associação espúria entre perfis; e 2. baixa variabilidade interpopulacional - o que os torna potencialmente úteis em qualquer população (Jobling e Gill, 2004). Em populações para as quais não existam bancos de dados de perfis genéticos, em situações em que não existam suspeitos ou em que a degradação do DNA não permita a genotipagem, metodologias comparativas não são úteis.

Nesse contexto, a genética forense busca desenvolver sistemas que relacionem diretamente perfis genéticos a características que possam reduzir o escopo de uma investigação (Frudakis, 2010), incluindo perfil de pigmentação, idade e ancestralidade. O *Forensic DNA Phenotyping* esbarra em barreiras éticas e legais, além de entraves metodológicos impostos pela complexidade da herança e manifestação de traços fenotípicos (Sturm e Larsson, 2009). Ainda assim, esforços têm sido despendidos no desenvolvimento de sistemas com esse fim e alguns deles têm obtido sucesso razoável em populações específicas (por exemplo: Chaitana et al. 2018; Walsh et al. 2017).

O estado incipiente dessa abordagem torna a estimativa de ancestralidade genética e origem geográfica/populacional mais relevante no contexto forense. Sistemas desenvolvidos com esse objetivo agregam características informativas e de análise laboratorial que os tornam úteis também na análise de ancestralidade com finalidades antropológicas e médicas. Abordagens para a avaliação do poder de painéis para a estimativa de ancestralidade incluem a comparação com outros

painéis, com dados de GWAS, a utilização de pares de irmãos, a comparação de estimativas de ancestralidade e de classificação por Naive Bayes, a comparação do *In* acumulado e seu balanceamento, *cross-validation* e PCA (exemplos no Capítulo 3 desta tese; Phillips 2015; Aquino et al. 2015; Galanter et al. 2012; Pereira et al. 2012; Phillips et al. 2007).

Neste trabalho, analisamos três diferentes sistemas de marcadores informativos de ancestralidade (*for*Indel, SNP*for*ID 34plex, PIMA+34plex) em populações Latino Americanas a partir de uma perspectiva predominantemente antropológica, sobretudo nos dois primeiros capítulos, mas também forense, especialmente no terceiro. De acordo com os critérios de seleção de marcadores adotados no desenho dos painéis (descritos em: Capítulo 3 desta tese; Fondevila et al. 2013; Pereira et al. 2012; Phillips et al. 2007), os sistemas são adequados para a análise de populações com o padrões de mistura esperados nas populações de estudo. Aqui, com o objetivo de avaliar a concordância dos sistemas na descrição das populações de estudo, comparamos as estimativas individuais de mistura nas amostras de Kalunga e Sacutiaba.

4.3 Metodologia

Apesar de existirem métricas que guiam a escolha dos marcadores que comporão um painel, não há um protocolo estabelecido para avaliação da acurácia de estimativas de ancestralidade. Aqui, optamos por avaliar a concordância entre os pares de painéis utilizados na estimativa e distribuição de ancestralidade individual quantificando a distância estatística (ou variação total) entre elas. Consideramos na

análise 1. o conjunto total de populações parentais e 2. cada população parental separadamente. Medidas de distância são quantificações de similaridade e dissimilaridade entre entidades estatísticas tais como distribuições de probabilidade ou variáveis independentes (Venturini 2015). Elas servem de base para outras avaliações estatísticas, como métodos de MDS e clusterização. A chamada distância estatística ou variação total avalia a dissimilaridade entre distribuições de probabilidade - aqui, as estimativas de ancestralidade individual para cada um dos quatro sistemas de marcadores. Ela varia de zero a um, sendo zero a situação de identidade e um, a de máxima diferença. Para acessar o quanto as estimativas geradas a partir dos diferentes sistemas variam, utilizamos as medianas das distâncias estimadas para cada indivíduo de forma global e em relação a cada parental (Tabela 1 e Figura Suplementar 1).

As análises foram feitas utilizando o Matplotlib 3.0 (Hunter, 2007) no Python 3.7 e os resultados, plotados utilizando a biblioteca *seaborn* 0.9 (Waskom et al. 2018) no Matplotlib. Apenas as estimativas geradas para Kalunga e Sacutiaba foram comparadas. Mocambo foi excluída da comparação porque sua a análise não gerou resultados laboratoriais satisfatórios para os sistemas SNP*for*ID 34plex e PIMA (e consequentemente para o sistema combinado PIMA+34plex).

4.4 Resultados e discussão

As matrizes de distância estatística geradas (Tabela 1) foram plotadas em *heatmaps* (Figura Suplementar 1). A maior distância global foi observada entre as contribuições estimadas pelos sistemas SNP*for*ID 34plex e PIMA (0.142). Os dois

sistemas foram desenhados para se complementarem no sistema conjunto PIMA+34plex, com o qual ambos mostraram distâncias menores (0.105 e 0.087, respectivamente). A menor distância estatística foi observada entre os sistemas *for*Indel e PIMA+34plex (0.069). Ambos foram desenhados para estimar as contribuições das parentais em análise, e de fato geraram estimativas de co-ancestralidade estatisticamente próximas.

Tabela 1. Distância estatística. Distâncias estatísticas estimadas para todos os componentes de mistura e para cada um deles (africano - AFR, europeu - EUR, indígena americano - IAM e leste asiático - EAS) a partir dos painéis PIMA, SNP*for*ID 34plex, PIMA+34plex e *for*InDel.

		PIMA+34	PIMA	34plex
	PIMA	0.0875	0.0000	0.0000
TOTAL	**34plex**	0.1050	**0.1418**	0.0000
	forInDel	0.0690	0.0830	0.1260
	PIMA	0.0680	0.0000	0.0000
AFR	**34plex**	0.0600	**0.0990**	0.0000
	forInDel	0.0545	0.0625	0.0750
	PIMA	**0.0865**	0.0000	0.0000
EUR	**34plex**	0.0395	0.0610	0.0000
	forInDel	0.0575	0.0610	0.0530
	PIMA	0.0175	0.0000	0.0000
IAM	**34plex**	0.0190	**0.0280**	0.0000
	forInDel	0.0150	0.0250	0.0180
	PIMA	0.0030	0.0000	0.0000
EAS	**34plex**	0.0880	0.0850	0.0000
	forInDel	0.0020	0.0045	**0.0930**
		PIMA+34	**PIMA**	**34plex**

Em negrito: menores distâncias observadas. Em células destacadas: distâncias estatísticas estimadas para o par *for*InDel e PIMA+34plex.

Em relação à parental africana, a maior diferença observada também foi entre PIMA e SNP*for*ID 34plex (0.099). A provável explicação é o balanceamento de cada sistema. O PIMA apresenta menor *In* AFR que os outros sistemas (Capítulo 3 desta tese). Já SNP*for*ID 34plex tem maior *In* acumulado para essa parental que para as demais. Os dois apresentaram distâncias menores em relação a PIMA+34plex (0.068 e 0.060, respectivamente), como era de se esperar, já que compõem esse painel. Da mesma forma, esse par apresentou a maior distância para a parental IAM (0.028) e distâncias menores em relação ao sistema PIMA+34plex (0.017 e 0.019, respectivamente). Aqui também PIMA apresenta maior *In* acumulado.

Já em relação à parental EUR, a maior distância observada foi entre os sistemas PIMA e *for*InDel. O primeiro apresenta *In* acumulado muito mais baixo para essa parental que para as demais (Capítulo 3 desta tese), enquanto o segundo foi desenhado para estimá-la de forma balanceada. Por fim, para a parental EAS, a maior distância foi observada entre *for*InDel e 34plex. Os dois foram desenhados tendo a detecção e mensuração dessa parental como objetivo (Pereira et al. 2012; Philips et al. 2007; Fondevila et al. 2013). No entanto, apenas *for*InDel teve também a diferenciação da parental IAM como objetivo. Assim, a diferença deve ser devida à inadequação de 34plex para detectar e mensurar IAM e diferenciá-la de EAS. 34plex estimou 9% de contribuição EAS nos quilombos, quando não há registro da presença de indivíduos com essa ancestralidade em nenhum dos dois.

A comparação mais relevante aqui, no entanto, é entre o par de sistemas *for*InDel e PIMA+34plex, já que ambos deveriam estimar as contribuições das quatro parentais AFR, EUR, IAM e EAS de forma balanceada. De fato, esse par apresentou

as menores distâncias estatísticas global (0.069) e para cada uma das parentais: AFR (0.055), IAM (0.015) e EAS (0.002). Apenas para a parental EUR ele não apresentou menor distância, mas ela ainda assim foi baixa (0.057). Considerando as distâncias estatísticas entre esse par e a pequena diferença percentual entre as medianas estimadas para as ancestralidades individuais (0.001 a 0.003), os dois sistemas estimam contribuições muito próximas e mostraram desempenho similar na análise de ancestralidade individual e populacional dos quilombos Kalunga e Sacutiaba. Essa proximidade atesta a adequação dos dois sistemas para a análise de populações Latino Americanas em geral.

Material Suplementar

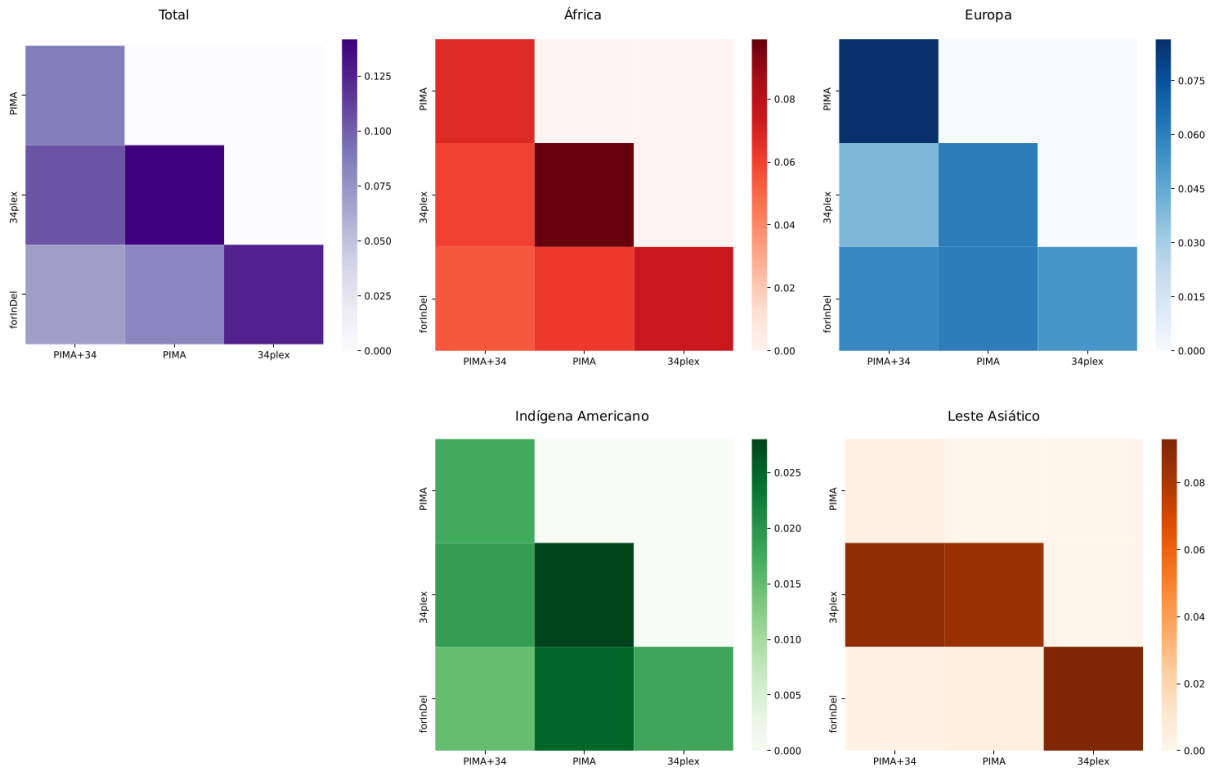


Figura Suplementar 1. *Heatmap* das distâncias estatísticas entre as estimativas de ancestralidade geradas para cada indivíduo das amostras de Kalunga e Sacutiaba a partir de dados dos sistemas *for*Indel, SNP*for*ID 34plex, PIMA e PIMA+34plex. Tons mais escuros indicam distâncias estatísticas maiores.

5. CONCLUSÕES

Com este trabalho, contribuímos para o conhecimento sobre as populações afro-derivadas latinoamericanas, gerando informação sobre ancestralidade e estrutura genéticas e permitindo a construção de um panorama mais completo das populações humanas do continente. Ao longo dos três capítulos que compõem esta tese, duas populações quilombolas rurais (Kalunga - GO e Riacho de Sacutiaba e Sacutiaba - BA) foram analisadas utilizando sistemas de marcadores informativos de ancestralidade. Além dessas populações, o quilombo Mocambo - SE foi analisado para o sistema *for*InDel, no capítulo 1, e a *comunidad negra* Colombiana rural Mulaló, para os sistemas SNP*for*ID 34plex e PIMA nos capítulos 2 e 3.

De forma geral, os resultados das análises de ancestralidade e estrutura genética apresentadas ao longo dos três capítulos confirmaram que os componentes de mistura mais significativos nos quilombos e na *comunidad negra* estudados são o africano e o europeu e que, apesar da ancestralidade compartilhada em decorrência de histórias de formação similares, as populações em estudo não são homogêneas. Além de atestar a diversidade existente no continente, essa observação reforça a necessidade de descrever populações não urbanas e não indígenas para a construção de bancos de dados adequados para a análise de ancestralidade.

No Capítulo 1, apresentamos e discutimos a análise do sistema forInDel nas populações Kalunga, Sacutiaba e Mocambo. Embora os resultados obtidos não tenham fugido ao esperado, com eles refinamos as análises prévias utilizando um

conjunto de marcadores mais poderoso que os anteriormente empregados e contribuímos para a completude de bancos de dados de interesse forense no que se refere a populações rurais brasileiras.

No segundo capítulo, inserimos na análise dados da população afro-colombiana Mulaló, análoga a quilombos pós-abolição. A partir da análise do sistema SNPforID 34plex em Mulaló, Kalunga e Sacutiaba, pudemos construir um panorama das populações afro-derivadas da América do Sul. Pudemos verificar que apesar de as três populações analisadas terem alta contribuição africana em sua composição genética, elas diferem substancialmente nas contribuições das outras parentais e distribuição de ancestralidade. Nossos resultados reiteram a diversidade das populações humanas do continente, mesmo dentro de um grupo com paralelos históricos e de origem tão marcados.

Apresentamos também, no terceiro capítulo, um novo sistema de marcadores (*PIMA: Population Informative Multiplex for the Americas*) que se mostrou eficiente em diferenciar o componente de mistura indígena americano dos componentes africano, europeu e leste asiático, especialmente em conjunto com marcadores do sistema SNP*for*ID 34plex. O painel combinado PIMA+34plex apresenta uma composição balanceada em termos de informatividade para as quatro parentais a que se propõe avaliar (indígena americana, africana, européia e leste asiática). Além disso, mostra performance comparável à de um sistema maior (LACE; Galanter et al. 2012), avaliado como referência a ser alcançada, e gerou resultados condizentes com a história e demografia conhecidas das 22 populações avaliadas.

Por fim, avaliamos a adequação e eficiência dos painéis SNP*for*ID 34plex e PIMA+34plex na estimativa de mistura para as parentais africana, européia, indígena americana e leste asiática em populações com perfil geral de mistura tri-híbrida tendo como modelo as populações Kalunga e Sacutiaba. Nossos resultados mostraram que os dois painéis geram resultados muito próximos, atestando sua adequação e indicando acurácia. Apesar de as populações utilizadas como modelo não terem de fato contribuição do leste asiático em seus *pools* gênicos, a presença de marcadores incluídos nos painéis especificamente para detectar e mensurar esse componente não interferiu de maneira relevante nas estimativas das demais parentais.

6. Referência bibliográfica da parte geral

Abe-Sandes K, Wilson A, Silva JR, Zago MA. (2004). Heterogeneity of the Y Chromosome in Afro-Brazilian Populations. *Human Biology* 76(1) 77-86. doi:10.1353/hub.2004.0014

Adhikari K, Chacón-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, Ruiz-Linares A. (2017). *The Genetic Diversity of the Americas. Annual Review of Genomics and Human Genetics, 18(1), 277–296.* doi:10.1146/annurev-genom-083115-022331

Alencastro, LF. (2000) *O trato dos Viventes: A formação do Brasil no Atlântico Sul*. 1ed. Companhia das Letras, São Paulo, SP.

Alves-Silva J, Santos MS, Guimarães PEM, Ferreira ACS, Bandelt HJ, Pena SDJ, Prado VF. (2000). The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet*, 67, 444–461.

Amorim CEG, Gontijo CC, Falcão-Alencar G, Godinho NMO, Toledo RCP, Pedrosa MAF, Luizon MR, Simões AL, Klautau-Guimães MN, Oliveira SF, Migration in Afro-Brazilian Rural Communities: Crossing Demographic and Genetic Data, *Human Biology*, 83 (2011) 509-521. doi:10.3378/027.083.0405

Anjos RSA, Cypriano A. (2006). *Quilombolas – tradições e cultura da resistência.* Aori Comunicações. São Paulo: Petrobras.

Aquino JG, Jannuzzi J, Carvalho EF, Gusmão L. (2015). Assessing the suitability of different sets of InDels in ancestry estimation. *Forensic Science International: Genetics Supplement Series*, 5, e34–e36. doi:10.1016/j.fsigss.2015.09.014

Arocha, J. (1998). *La inclusión de los afrocolombianos ¿Meta inalcanzable?* En. ICCH. Geografía Humana de Colombia. Los Afrocolombianos. ICANH.

Arruti J. (2006). *Mocambo: Antropologia e história do processo de formação quilombola*. Edusc, Bauru, SP.

Azopardo, IG .(1987) El comercio y mercado de negros esclavos en Cartagena de Indias (1533 - 1850) doi:10.5209/rev_QUCE.1987.v12.1781

Barcelos, RSS. (2005) *Constituição genética de populações urbanas do Centro-Oeste brasileiro (Goiás e Distrito Federal) estimada por marcadores uniparentais Y-específicos e do DNAmt.* Tese de doutorado. Programa de Pós-Graduação em Biologia Animal. Universidade de Brasília. Brasília – DF.

Bedoya G, Montoya P, Garcı J, Soto I, Bourgeois S, Carvajal L, et al. Admixture dynamics in Hispanics : A shift in the nuclear genetic ancestry of a South American population isolate. 2006;103(19).

Butler, JM. Advanced Topics in Forensic DNA Typing: Interpretation. (2015) Library of Congress Cataloging-in-Publication Data.

Callegari-Jacques, SM, Grattapaglia D, Salzano, FM, Salamoni SP, Crossetti SG, Ferreira, ME, Hutz MH. (2003). Historical genetics: Spatiotemporal analysis of the formation of the Brazilian population. *American Journal of Human Biology,* 15(6), 824–834. doi:10.1002/ajhb.10217

Carvalho-Silva DR, Santos FR, Rocha J, Pena SDJ. (2001). The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet* 68:281–286.

Carvalho-Silva DR, Tarazona-Santos E, Rocha J, Pena, SDJ, Santos FR. (2006). Y Chromosome Diversity in Brazilians: Switching Perspectives from Slow to Fast Evolving Markers. Genetica, 126(1-2), 251–260. doi:10.1007/s10709-005-1454-z

Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pospiech E, Kukla-Bartoszek M, Sijen T, de Knijff P, Liu F, Branicki W, Kayser M, Walsh S. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Science International Genetics. doi: 10.1016/j.fsigen.2018.04.004

Chen HD, Chang CH, Hsieh LC, Lee HC. (2005). Divergence and Shannon information in genomes, Phys. Rev. Lett. 94.

Costas J, Salas A, Phillips C Carracedo Á. (2005). Human genome-wide screen of haplotype-like blocks of reduced diversity. *Gene*, 349, 219–225. doi:10.1016/j.gene.2004.12.042

Da Silva, VRR (2012) Entre quilombos e palenques: um estudo antropológico sobre políticas públicas de reconhecimento no Brasil e na Colômbia. Thesis Faculdade de Filosofia, Letras e Ciências Humanas do Curso de Pós-Graduação em Ciências Sociais, Universidade de São Paulo (USP).

De Moura RR, Coelho AVC, Balbino VQ, Crovella S, Brandão LAC, Meta-Analysis of Brazilian Genetic Admixture and Comparison with Other Latin America Countries, American Journal of Human Biology 27 (2015) 674–680. doi:10.1002/ajhb.22714

Excoffier L, Lischer HEL. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564–567.doi:10.1111/j.1755-0998.2010.02847.x

Fausto B. (2001) *História Concisa do Brasil*. São Paulo, SP.

Fausto B. (2002) *História do Brasil*. 10 ed. Editora da Universidade de São Paulo. São Paulo, SP.

FCP, Fundação Cultural Palmares – FCP: http://www.palmares.gov.br, 2017 (accessed 20 December 2018)

Ferreira R. (2006). Biografia, mobilidade e cultura atlântica: a micro-escala do tráfico de escravos em Benguela, séculos XVIII-XIX. Revista *Tempo*, Niterói: Universidade Federal Fluminense, n.20, p. 33-59.

Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, et al. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended

population studies. Forensic Sci Int Genet. 2013;7(1):63–74. doi: 10.1016/j.fsigen.2012.06.007

Francez PAC, Ribeiro-Rodrigues EM, Santos SEB, Allelic frequencies and statistical data obtained from 48 AIM INDEL loci in an admixed population from the Brazilian Amazon, Forensic Sci. Int. Genet. 6 (2011) 132–135.

Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK, A classifier for the SNP-based inference of ancestry, J.Forensic Sci. 48 (2003) 771–782.

Funai (Fundação Nacional do Índio), 2013 - http://www.funai.gov.br (acessado em maio de 2019)

Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, et al. (2012). Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas. Gibson G, editor. PLoS Genet. 8;8(3):e1002554. doi: 10.1371/journal.pgen.1002554

Giolo SR, Soler JMP, Greenway SC, Almeida MAA, de Andrade M, Seidman JG (…) Pereira AC. (2011). *Brazilian urban population genetic structure reveals a high degree of admixture. European Journal of Human Genetics, 20(1), 111–116.*doi:10.1038/ejhg.2011.144

Gomes, F. S. 2015. Mocambos e Quilombos. Uma história do campesinato negro no Brasil. São Paulo: Claroenigma.

Gómez-Carballa A, Pardo-Seco J, Brandini S, Achilli A, Perego UA, Coble MD, et al. The peopling of South America and the trans-Andean gene flow of the first settlers. Genome Res. 2018 Jun;28(6):767–79. doi: 10.1101/gr.234674.118

Gontijo CC, Amorim CEG, Godinho NMO, Toledo RCP, Nunes A, Silva W, Moura MMF, Oliveira JCC, Pagotto RC, Klautau-Guimarães MN, Oliveira SF,

Brazilian Quilombos: A Repository of Amerindian Alleles, American Journal of Human Biology, 26 (2014) 142-150. doi:10.1002/ajhb.22501

Gontijo CC, Mendes FM, Santos CA, Klautau-Guimarães M de N, Lareu MV, Carracedo Á, et al. Ancestry analysis in rural Brazilian populations of African descent. Forensic Sci Int Genet [Internet]. 2018 Sep;36(January):160–6. Available from: https://doi.org/10.1016/j.fsigen.2018.06.018

Gonzáles-Andrade F, Sánchez D, González-Solórzano J, Gascón S, Martinez-Jarreta B. (2007) Sex-specific genetic admixture of mestizos, amerindian kichwas, and afro-ecuadorans from Ecuador. *Human Biology*, v.79, no. 1, pp. 51-77

González-José, R., Bortolini, M. C., Santos, F. R., & Bonatto, S. L. (2008). *The peopling of America: Craniofacial shape variation on a continental scale and its interpretation from an interdisciplinary view. American Journal of Physical Anthropology, 137(2), 175–187.*doi:10.1002/ajpa.20854

Grattapaglia D, Kalupniek S, Guimarães, CS, Ribeiro MA, Diener PS, Soares CN. (2005). *Y-chromosome STR haplotype diversity in Brazilian populations. Forensic Science International, 149(1), 99–107.*doi:10.1016/j.forsciint.2004.06.003

Hartl, DL, Clark AG. (2010). *Principles of population genetics*. 4th ed. Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts.

Holsinger KE, Weir BS. (2009). Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting FST. Nature Reviews Genetics, 10, 639-650. http://dx.doi.org/10.1038/nrg2611

Hünemeier T, Carvalho C, Marrero AR, Salzano FM, Pena SD, Bortolini MC. (2007). Niger-Congo speaking populations and the formation of the Brazilian gene pool: mtDNA and Y-chromosome data. American Journal of Physical Anthropology, 133(2), 854–867. doi:10.1002/ajpa.20604

IBGE, Instituto Instituto Brasileiro de Geografia e Estatística – Censo 2010: http://www.ibge.gov.br/home/estatística/população

Itamaraty (2015) Visita do Ministro das Relações Exterior, Mauro Vieira, ao Irã e ao Líbano – 13 a 16 de dezembro de 2015. *Notas à Imprensa,* 354.

Jobling MA, Gill P, Encoded evidence: DNA in forensic analysis, Nature Reviews Genetics, 5 (2004) 739–751. doi:10.1038/nrg1455.

Karasch M. (2000) *A Vida Dos Escravos No Rio De Janeiro, 1808-1850*. São Paulo, SP, Brazil: Companhia das Letras.

Kehdy FSG, Gouveia, MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL. (2015). *Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences, 112(28), 8696–8701.*doi:10.1073/pnas.1504447112

Kimura L, Nunes K, Macedo-Souza LI, Rocha J, Meyer D, Mingroni-Netto RC, Inferring paternal history of rural African-derived Brazilian populations from Y chromosomes, Am. J. Hum. Biol. (2016) 1-11. doi:10.1002/ajhb.22930

Kimura L, Ribeiro-Rodrigues EM, Auricchio MTBM, Vicente JP, Santos SEB, Mingroni-Netto MC, Genomic Ancestry of Rural African-Derived Populations from Southeastern Brazil, American Journal Of Human Biology 25:35–41 (2013). doi:10.1002/ajhb.22335

Kimura, M. (1968). *Evolutionary Rate at the Molecular Level. Nature, 217(5129),* 624–626.doi:10.1038/217624a0

Landers J, Gómez P, Acuña JP, Campbell CJ. Researching the history of slavery in Colombia and Brazil through ecclesiastical and notarial archives in: From Dust to Digital: Ten Years of the Endangered Archives Programme. Kominko, M. Open Book Publishers. (2015). pp. 259-292 (34 pages). https://www.jstor.org/stable/j.ctt15m7nhp.20

Lareu MV. (2013) Short Tandem Repeats. In: Encyclopedia of Forensic Sciences. Siegal JA, Saukko PJ, Houck MM (editors). Academic Press.

Lewontin RC. (1972) The Apportionment of Human Diversity. In: Dobzhansky T., Hecht M.K., Steere W.C. (eds) Evolutionary Biology. Springer, New York, NY

Li JZ, Absher DM, Tang H, Southwick AM, Casto S, Ramachandran, HM, Cann, GS, Barsh M, Feldman, LL Cavalli-Sforza, RM Myers. (2008). Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.

Lins TC, Vieira RG, Abreu BS, Grattapaglia D, Pereira, RW. (2009). Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. *American Journal of Human Biology, NA–NA.*doi:10.1002/ajhb.20976

Lopes Maciel LG, Rodrigues EMR, Santos NPC, Santos AR, Guerreiro JF, Santos S. (2011). Afro-Derived Amazonian Populations: Inferring Continental Ancestry and Population Substructure. *Human Biology*. Vol. 83, No. 5, pp. 627-636. doi: 10.3378/027.083.0504

Lum JK, Cann RL, Martinson JJ, Jorde LB. (1998). Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *Am J Hum Genet* 63(2):613-24.

Manta FSN, Pereira R, Vianna R, Araújo ARB, Gitaí DLG, Silva DA, Wolfgramm EV, Pontes IM, Aguiar JI, Moraes MO, Carvalho EF, Gusmão L. (2013). Revisiting the Genetic Ancestry of Brazilians Using Autosomal AIM-Indels. *PLoS ONE* 8(9): e75145. doi:10.1371/journal.pone.0075145

Manta F, Caiafa A, Pereira R, Silva D, Amorim A, Carvalho EF, Gusmão L. (2012). Indel markers: Genetic diversity of 38 polymorphisms in Brazilian populations and application in a paternity investigation with post mortem material.

Forensic Science International: Genetics, 6(5), 658–661. doi:10.1016/j.fsigen.2011.12.008

Martínez-Marignac V, Bertoni B, Parra EJ, Bianchi NO. (2004). Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. *Human Biology* v. 76, no 4, pp. 543-557

Martínez, H.; Rodríguez-Larralde, A.; Izaguirre, M.H.; de Guerra, D.C. (2007) Admixture estimates for Caracas, Venezuela, based on autosomal, Y-chromosome, and mtDNA Markers. *Human Biology*, v. 79, no. 2, pp. 201-213

Mello e Souza M. (2006) *Reis Negros no Brasil Escravista: História da Festa de Coroação de Rei Congo*. 1 ed. Editora UFMG, Belo Horizonte, MG – Brasil.

Mendes F, Gontijo CC. (2017). Kpop: A Python package for population genetics analyses. *Forensic Science International: Genetics Supplement Series*, 6 (2017) e407-e409. doi:10.1016/j.fsigss.2017.09.159.

Miller JC. (1997). Way of Death: Merchant Capitalism and the Angolan Slave Trade, 1730–1830 [Internet]. University of Wisconsin Press. Available from: https://books.google.com.br/books?id=flAJ7IH3Cj4C

Mills RE. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research, 16(9), 1182–1190. doi:10.1101/gr.4565806

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, (…) Willerslev E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature.* 553(7687), 203–207. doi:10.1038/nature25173

Neel JV, Weiss KM. (1975). The genetic structure of a tribal population, the Yanomama Indians. XII. Biodemographic studies. *Am J Phys Anthropol* 42(1):25-51.

Nei M, Roychoudhury AK. (1982). Genetic relationship and evolution of human races. *Evolutionary Biology* - vol. 14: 1-59, New York, Appleton - Century - Crafts. Plenum Press. ed. Dobzhanski, T. *et al*.

Neme S, Andrade CO. (1987). Quilombo: forma de resistência. Proposta histórico-arqueológica. In: *Insurreição Negra e Justiça* (eds. G. HUBER e F.B. DE SOUZA). Rio de Janeiro, OAB.

Paiva SG. (2017). *Fatores de risco para doenças cardiovasculares em quilombos contemporâneos do Brasil Central : parâmetros demográficos, socioeconômicos, ancestralidade genética e saúde*. (Tese) Programa de Pós Graduação em Biologia Animal, Instituto de Ciências Biológicas, Universidade de Brasília, Brasil.

Palha TJ, Gusmão L, Ribeiro-Rodrigues E, Guerreiro JF, Ribeiro-Dos-Santos A et al. (2012) Disclosing the genetic structure of Brazil through analysis of male lineages with highly discriminating haplotypes. PLOS ONE 7: e40007. doi:10.1371/journal.pone.0040007.

Pena SDJ, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy FDSG, et al. (2011). The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. *PLoS ONE,* 6(2) e17063. doi:10.1371/journal.pone.0017063

Pena, SDJ, Di Pietro, G, Fuchshuber-Moraes, M, Genro, JP, Hutz, MH, Kehdy, F de SG, Suarez-Kurtz, G. (2011). The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. PLoS ONE, 6(2), e17063. doi:10.1371/journal.pone.0017063

Pereira R, Phillips C, Pinto N, Santos C, Santos SEB dos, Amorim A, et al. (2012). Straightforward Inference of Ancestry and Admixture Proportions through

Ancestry-Informative Insertion Deletion Multiplexing. Kayser M, editor. PLoS One. 7(1):e29684.

Phillips C, Sala A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, Calaza M, Casares de Cal M, Ballard D, Lareu MV, Carracedo A, The SNPforID Consortium. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics* 1, 273-280.

Phillips CP. (2012). Applications of autosomal SNPs and Indels in Forensic Analysis. *Forensic Sci Rev* 24:43.

Phillips C, Aradas AF, Kriegel AK, Fondevila M, Bulbul O, Santos C, (…) Lareu MV. (2013). Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Science International: Genetics*, 7(3), 359–366. doi:10.1016/j.fsigen.2013.02.010

Phillips C. (2015). *Forensic genetic analysis of bio-geographical ancestry. Forensic Science International: Genetics, 18, 49–65.*doi:10.1016/j.fsigen.2015.05.012

Pickrell JK, Reich, D. (2014). *Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics, 30(9), 377–389.* doi:10.1016/j.tig.2014.07.007

Pinotti T, Bergström A, Geppert M, Bawn M, Ohasi D, Shi W, et al. (2019). Y Chromosome Sequences Reveal a Short Beringian Standstill, Rapid Expansion, and early Population structure of Native American Founders. *Current Biology.*

Pritchard JK, Donnelly P. (2001). Case-control studies of association in structured or admixed populations. *Theor Popul Biol*. 60(3):227–37.

Pritchard JK, Stephens M, Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics*. 155: 945–959.

Raghavan, M., Steinrucken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., (…) Malaspinas, A.-S. (2015). *Genomic evidence for the Pleistocene and recent population history of Native Americans. Science, 349(6250), aab3884–aab3884.*doi:10.1126/science.aab3884

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al., Reconstructing Native American population history, Nature 488 (7411) (2012) 370–374. doi:10.1038/nature11258

Resque R, Gusmão L, Geppert M, Roewer L, Palha T, Alvarez L, et al. (2016) Male Lineages in Brazil: Intercontinental Admixture and Stratification of the European Background. PLoS ONE 11(4):e0152573. doi:10.1371/journal.pone.0152573

Ribeiro GGBL, Abe-Sandes K, Barcelos RSS, et al. (2011). Who were the male founders of rural Brazilian Afro-derived communities? A proposal based on three populations, *Ann. Hum. Biol*. 38 237–240. doi:10.3109/03014460.2010.500471.

Rodrigues N. (2004) *Os Africanos no Brasil*. Editora Universidade de Brasília, Brasília, Brasil.

Rosenberg NA, Li LM, Ward R, Pritchard JK. (2003). Informativeness of Genetic Markers for Inference of Ancestry. *The American Journal of Human Genetics*, 73(6), 1402–1422.doi:10.1086/380416

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. (2005). Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1: 70.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. (2002). Genetic Structure of Human Populations. Science, New Series, Vol. 298, No. 5602, pp. 2381-2385

Ruiz-LinaresA, Adhikari K, Acuña-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, (…) Gonzalez-José R. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genetics*. 10(9), e1004572. doi:10.1371/journal.pgen.1004572

Salzano FM, Freire-Maia N. (1967) *Populações Brasileiras.* Companhia Editora Nacional. São Paulo, SP.

Salzano FM, Sans M. (2014). Interethnic admixture and the evolution of Latin American populations. *Genetics and Molecular Biology*. 37:151-17

Santos NPC, Ribeiro-Rodrigues EM, Ribeiro-dos-Santos AKC, Pereira R, Gusmão L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH, Santos SEB. (2009). Assessing Individual Interethnic Admixture and Population Substructure Using a 48–Insertion-Deletion (INSEL) Ancestry-Informative Marker (AIM) Panel. *Human Mutation*, Vol. 31, No. 2, 184–190. doi:10.1002/humu.21159

Santos S, Guerreiro J. (1995). The indigenous contribution to the formation of the population of the Brazilian Amazon Region. Rev Bras Genet 18: 311–315.

Santos C, Phillips C, Fondevila M, Daniel R, van Oorschot RAH, Burchard EG, (…) Lareu MV. (2016). Pacifiplex : an ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Science International: Genetics, 20, 71–80*.doi:10.1016/j.fsigen.2015.10.003

Santos N P C, Ribeiro-Rodrigues EM, Ribeiro-dos-Santos AKC, Pereira R, Gusmão L, Amorim A, (…) Santos, SEB. (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. Human Mutation, 31(2), 184–190. doi:10.1002/humu.21159

Schrider DR, Kern AD. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome*. Molecular Biology and Evolution.* 34(8), 1863–1877. doi:10.1093/molbev/msx154

Schwartz SB. (1988). *Segredos Internos. Engenhos e escravos na sociedade colonial*. São Paulo, SP, Brasil: Companhia das Letras.

SEPPIR, Secretaria Nacional de Políticas de Promoção da Igualdade Racial – SEPPIR. http://seppir.gov.br (accessed 18 december 2018)

Shriver MD, Kittles RA. (2004). Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics.* 5: 611–618. doi:10.1038/nrg1405.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka, R, Ferrell, RE. (1997). Ethinic-affiliation estimation by use of population-specific DNA markers. *Am J. Hum Genet*, 60:957-964.

Slatkin M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6), 477–485. doi:10.1038/nrg2361

Soares AA. (1995). Kalunga: O direito de Existir, Questões Antropológicas e Jurídicas sobre Remanescentes de Quilombo. Goiânia: MinC/Fundação Cultural Palmares.

Sobrino B, Brión M, Carracedo A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Science International. 154:2(181-194).

Taboada-Echalar P, Álvarez-Iglesias V, Heinz T, Vidal-Bralo, L, Gómez-Carballa A, Catelli L, (…) Salas A. (2013). The Genetic Legacy of the Pre-Colonial Period in Contemporary Bolivians. *PLoS ONE*. 8(3). e58980.doi:10.1371/journal.pone.0058980

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ, Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans,

The American Journal of Human Genetics, 81 (2007) 626-633. doi:10.1086/520769.

Thompson EA. Fission models of population variability. (1979). *Genetics* 93: 479-95.

Toledo RCP. (2016). *Fenótipos complexos distintos e componentes genéticos coincidentes: Um estudo sobre hipodontia e histórico familiar de câncer*. Tese de doutorado (Biologia Animal) - Universidade de Brasília.

Venturini GM. (2015). Statistical Distances and Probability Metrics for Multivariate Data, Ensembles and Probability Distributions. Tese de doutorado. Department of Statistics, Universidad Carlos III de Madrid, Leganés, Madrid, Spain.

Vila Real RNS. (1996). *Cultura e Currículo: Um estudo da escola Kalunga. 1996. Dissertação (Mestrado)* - Universidade Federal de Goiás, Goiânia.

Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, Koller J, Kovatsi L, Wollstein A, Branicki W, Liu F, Kayser M. (2017). Global skin colour prediction from DNA. *Human Genetics*. 136(7): p. 847-863.

Wang S, Lewis CM Jr, Jakobsson M, Ramachandran S, Ray N, et a.l (2007). Genetic variation and population structure in Native Americans. *PLoS Genet.* 3(11): e185. doi:10.1371/journal.pgen.0030185

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzoti G, Hill K, Hurtado AM, et al. (2008). Geographic Patterns of Genome Admixture in Latin American Mestizos. *PLoS Genet.* 4(3) e1000037. doi:10.1371/journal.pgen.1000037

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. (2002). Human Diallelic Insertion/Deletion Polymorphisms. *Am J Hum Genet* 71:854–862.

Woortmann K. (1990). Migração, Família e Campesinato. Revista Brasileira de Estudos de População 7:35–53.

Zhang, J. (2018). Neutral Theory and Phenotypic Evolution. *Molecular Biology and Evolution.* 35(6), 1327–1331. doi:10.1093/molbev/msy065.

POSFÁCIO

A coleta do material biológico e dos dados das populações quilombolas utilizados nesta tese foi realizada por equipes do Laboratório de Genética Humana da UnB em expedições chefiadas pelas Professoras Doutoras Silviene Fabiana de Oliveira e Maria de Nazaré Klautau-Guimarães. Para a coleta de dados demográfico e de saúde geral, todos os voluntários foram entrevistados utilizando questionários desenvolvidos especificamente para essa finalidade. A expedição para a comunidade Riacho de Sacutiaba e Sacutiaba ocorreu em 1998. A primeira expedição à comunidade Kalunga aconteceu no ano de 2001. Posteriormente, diversas expedições a essa comunidade foram feitas, das quais pude participar. O questionário utilizado como base para as entrevistas foi expandido, assim como a amostra, e adendos ao projeto inicial incluíram análises de fecundidade e maiores informações sobre saúde geral. O processamento do material biológico, a extração de DNA e o armazenamento foram feitos no Laboratório de Genética da UnB por uma equipe ampla e que variou ao longo das expedições.

A análise laboratorial dos sistemas aqui apresentados foi feita na *Unidade de Xenética do Instituto de Ciencias Forenses da Universidade de Santiago de Compostela,* Galícia, Espanha, durante estágio de doutorado sob supervisão direta de Christopher Phillips, com apoio de Maria Victoria Lareu e Ángel Carracedo. Ali, participaram ativamente Ana Freire-Aradas e Carla Arcanjo Santos. Contribuíram também Yarimar Ruiz Orozco, Olalla Maroñas Amigo e os técnicos e pesquisadores do laboratório Ana Mosquera, Raquel Cruz, Manuel Fondevila e Danel Rey.

O capítulo 1 foi concluído e publicado na forma de artigo no periódico Forensic Science International: Genetics (Gontijo et al., *Ancestry analysis in rural Brazilian populations of African descent*. Forensic Science International: Genetics, Volume 36, September 2018, Pages 160-166). O capítulo 2 está em revisão para submissão (Gontijo et al., *Genetic Ancestry and Structure in African-derived populations from South America*, a ser submetido ao periódico American Journal of Human Biology). O capítulo 3 está em revisão para submissão (Gontijo et al., *PIMA: Population Informative Multiplex for the Americas,* a ser submetido ao periódico Forensic Science International: Genetics).

Um artigo relacionado à tese, mas não incluído, foi publicado: Mendes, F. M., & Gontijo, C. C. (2017). *Kpop: A Python package for population genetics analysis*. Forensic Science International: Genetics Supplement Series, 6, e407–e409. Nesse trabalho, desenvolvemos e apresentamos um *software* que integra ferramentas de manipulação de bancos de dados, algoritmos comuns à análise de genética de populações, métodos de clusterização, categorização e redução de dimensionalidade, além de algumas estatísticas descritivas. O software foi amplamente utilizado na execução desta tese.

Três artigos não relacionados à tese, mas da mesma área de pesquisa, estão em fase de redação ou submissão em colaboração com outros alunos e pesquisadores. O primeiro trata da ressignificação do termo quilombo frente à diversidade histórica e demográfica que se observa nessas populações, e apresenta dados coletados durante nossas visitas a comunidades quilombolas: Paiva et al., *Demography and Migration: Illustrating the Contemporary Meaning of Quilombo*

(Brazilian Afro-Descendant Communities). O segundo trata da distribuição de mutações neutras no gene da Fibrose Cística na população Kalunga e em amostras do HGDP-CEPH (Gontijo et al., *CFTR haplotypes in worldwide urban populations and in Brazilian quilombos*). O terceiro trata da distribuição de SNPs e haplótipos em genes de cor da pele e sua correlação com origem biogeográfica em quilombos e populações do 1000 Genomes: Castro et al., *Analysis of three isolated Afro-descendant Brazilian populations: genetics, biogeographical origin and skin colour*.

Dois trabalhos foram apresentados em evento internacional (27th Congress of the International Society for Forensic Genetics, Seoul, Republic of Korea, 2017): 1. *Ancestry and Structure of African-Brazilian Populations Estimated from 46 AIM-Indels* (Gontijo et al., 2017) e 2. *Kpop: a Python Package for Population Genetics Analysis* (Mendes and Gontijo, 2017). Outros dois foram aceitos, mas retirados por falta de apoio financeiro para participação nos eventos (86th Annual Meeting of The American Association of Physical Anthropologists, 2017, *The distribution of CFTR haplotypes in Brazilian Quilombos as a consequence of history*, Gontijo et al.; The 28th Congress of the International Society for Forensic Genetics, 2019; *PIMA: A Population Informative Multiplex for the Americas*; Gontijo et al.).