ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR
MESTRADO EM BIOLOGIA CELULAR E MOLECULAR

ALINE BRUGNERA FELKL

ANCESTRALIDADE BIOGEOGRÁFICA DE QUATRO GRUPOS POPULACIONAIS DO RIO GRANDE DO SUL (RS), BRASIL, POR SEQUENCIAMENTO MASSIVO PARALELO DE 165 MARCADORES GENÉTICOS

Porto Alegre – RS
2020

ALINE BRUGNERA FELKL

ANCESTRALIDADE BIOGEOGRÁFICA DE QUATRO GRUPOS POPULACIONAIS DO RIO GRANDE DO SUL (RS), BRASIL, POR SEQUENCIAMENTO MASSIVO PARALELO DE 165 MARCADORES GENÉTICOS

Dissertação apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM) da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) como requisito parcial para obtenção do grau de Mestra em Biologia Celular e Molecular.

Orientadora:

Profª. Dra. Clarice Sampaio Alho

Porto Alegre – RS

2020

Ficha Catalográfica

F316a Felkl, Aline Brugnera

Ancestralidade Biogeográfica de quatro grupos populacionais do Rio Grande do Sul (RS), Brasil, por sequenciamento massivo paralelo de 165 marcadores genéticos / Aline Brugnera Felkl. – 2020.
62.
Dissertação (Mestrado) – Programa de Pós-Graduação em Biologia Celular e Molecular, PUCRS.

Orientadora: Profa. Dra. Clarice Sampaio Alho.

1. População brasileira. 2. Genética de populações. 3. Ancestralidade Biogeográfica. 4. Sequenciamento massivo paralelo. 5. Precision ID Ancestry Panel. I. Alho, Clarice Sampaio. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Dedico este trabalho aos meus amados pais,

Flávio e Alice.

AGRADECIMENTOS

À orientadora Clarice Sampaio Alho, minha eterna gratidão. Obrigada pela oportunidade de desenvolver este trabalho e pela inspiração proporcionada, tanto pela área profissional, quanto pela pessoa que és.

Ao coorientador e amigo Eduardo Avila, por todas as discussões e ensinamentos ao longo destes anos, desde o meu período de estágio no laboratório. Guardarei com carinho todos os momentos de aprendizado, e procurarei sempre espelhar-me em tua trajetória.

Ao Prof. Marcio Dorn do Instituto de Informática da UFRGS, por fornecer a infraestrutura e a expertise em bioinformática fundamentais à realização deste trabalho.

Aos amigos e colegas do LGHM | GF que dividiram responsabilidades, fomentaram discussões produtivas e propiciaram inestimáveis momentos de descontração nestes anos de convivência e colaboração.

Aos voluntários que gentilmente doaram suas amostras para a realização deste estudo, bem como aos colegas do Laboratório de Genética Forense (LGHM | GF) do INCT/PUCRS pelo auxílio nas coletas.

À banca examinadora, por cordialmente contribuir com a realização deste trabalho mediante suas valorosas considerações.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa regular de Mestrado.

Aos meus queridos pais, Flávio e Alice, pelo suporte a amor incondicionais.

Ao Stefano, por simplesmente ser meu alicerce ao longo destes anos.

RESUMO

FELKL, A. B. **Ancestralidade biogeográfica de quatro grupos populacionais do Rio Grande do Sul (RS), Brasil, por sequenciamento massivo paralelo de 165 marcadores genéticos**. [Dissertação de Mestrado]. Porto Alegre/RS: PUCRS, Escola de Ciências da Saúde e da Vida, Programa de Pós-Graduação em Biologia Celular e Molecular; 2020; 62p.

A fenotipagem forense pelo DNA (FDP) inclui inferência de ancestralidade e predição de características externamente visíveis (EVC) diretamente de uma amostra de DNA desconhecido como alternativas para fornecer pistas investigativas quando um indivíduo não é identificado pelo método convencional de correspondência genética. Nesse contexto, a aplicação de metodologias de Sequenciamento Massivo Paralelo (MPS), que possibilita a genotipagem simultânea de múltiplas amostras e centenas de marcadores genéticos, vem sendo gradualmente implementada na casuística da genética forense. O *Precision ID Ancestry Panel* (Thermo Fisher Scientific, Waltham, EUA) é um ensaio multiplex forense que consiste em 165 SNPs autossômicos projetados para fornecer informações de ancestralidade biogeográfica. Neste trabalho, 250 indivíduos residentes no Estado do Rio Grande do Sul (RS), classificados em quatro grupos populacionais principais (gaúchos de origem africana, europeia, ameríndia e mista), foram analisados com o painel supracitado no intuito de avaliar a viabilidade desta abordagem em uma população altamente heterogênea. Os parâmetros forenses estimados para cada grupo populacional revelaram que este painel possui SNPs polimórficos e informativos o suficiente para serem utilizados como um instrumento complementar na identificação forense de indivíduos e em exames de parentesco, independentemente da etnia. Nenhum desvio estatisticamente significativo do equilíbrio de Hardy-Weinberg foi observado após a correção de Bonferroni. No entanto, sete pares de lócus exibiram desequilíbrio de ligação (p < 3.70 × 10^{-6}). Comparações interpopulacionais mediante análise de F_{ST}, MDS e STRUCTURE entre os quatro grupos populacionais do RS, e entre estes e 89 populações mundiais de referência, demonstraram que os gaúchos de origem mista e africana apresentam os mais altos índices de miscigenação e estratificação populacional, enquanto os gaúchos de origem europeia e ameríndia exibem uma conformação genético-populacional mais homogênea.

Palavras-chave: População brasileira; Genética de populações; Ancestralidade Biogeográfica; Sequenciamento massivo paralelo; Precision ID Ancestry Panel.

ABSTRACT

FELKL, A. B. **Biogeographic ancestry of four population groups from Rio Grande do Sul (RS), Brazil, by massively parallel sequencing of 165 genetic markers**. [Master's Thesis]. Porto Alegre/RS: PUCRS, School of Health and Life Sciences, Postgraduate Program in Cellular and Molecular Biology; 2020; 62p.

Forensic DNA phenotyping (FDP) includes biogeographic ancestry (BGA) inference and externally visible characteristics (EVCs) prediction directly from an evidential DNA sample as alternatives to provide valuable intelligence when conventional DNA profiling fails to achieve identification. In this context, the application of Massively Parallel Sequencing (MPS) methodologies, which enables simultaneous typing of multiple samples and hundreds of forensic markers, has been gradually implemented in forensic genetic casework. The Precision ID Ancestry Panel (Thermo Fisher Scientific, Waltham, USA) is a forensic multiplex assay consisting of 165 autosomal SNPs designed to provide biogeographic ancestry information. In this work, a sample of 250 individuals from Rio Grande do Sul (RS) State, southern Brazil, apportioned into four main population groups (African-, European-, Amerindian-, and Admixed-derived Gauchos), was evaluated with this panel, to assess the feasibility of this approach in a highly heterogeneous population. Forensic descriptive parameters estimated for each population group revealed that this panel has enough polymorphic and informative SNPs to be used as a supplementary instrument in forensic individual identification and kinship testing regardless of ethnicity. No statistically significant deviation from Hardy-Weinberg equilibrium was observed after Bonferroni correction. However, seven loci pairs displayed linkage disequilibrium in pairwise LD testing ($p < 3.70 \times 10^{-6}$). Interpopulation comparisons by F_{ST} analysis, MDS plot, and STRUCTURE analysis among the four RS population groups apart and along with 89 reference worldwide populations demonstrated that Admixed- and African-derived Gauchos present the highest levels of admixture and population stratification, whereas European- and Amerindian-derived exhibit a more homogeneous genetic conformation.

Keywords: Brazilian population; Population genetics; Biogeographic ancestry; Massively parallel sequencing; Precision ID Ancestry Panel.

LISTA DE ABREVIATURAS E SIGLAS

ADRS (*Admixed-derived Gauchos*):	Gaúcho de origem mista
AFRS (*African-derived Gauchos*):	Gaúcho de origem africana
AFR (*African*):	Africano
AIM (*Ancestry-informative Marker*):	Marcador Informativo de Ancestralidade
AISNP (*Ancestry-informative SNP*):	SNP Informativo de Ancestralidade
AMRS (*Amerindian-derived Gaucho*):	Gaúcho de origem ameríndia
AMOVA (*Analysis of Molecular Variance*):	Análise de Variância Molecular
BGA (*Biogeographic Ancestry*):	Ancestralidade Biogeográfica
BNPG:	Banco Nacional de Perfis Genéticos
CAAE:	Certificado de Apresentação para Apreciação Ética
CODIS:	*Combined DNA Index System*
CE (*Capillary Electrophoresis*):	Eletroforese Capilar
CEP:	Comitê de Ética em Pesquisa
DNA (*Deoxyribonucleic acid*):	Ácido Desoxirribonucleico
EmPCR (*Emulsion PCR*):	PCR de Emulsão
EURS (*European-derived Gaucho*):	Gaúcho de origem europeia
EUR (*European*):	Europeu
EtOH (*Ethanol*):	Etanol
EVC (*Externally Visible Characteristic*):	Característica Externamente Visível
FBI (*Federal Bureau of Investigation*):	Departamento Federal de Investigação
FDP (*Forensic DNA Phenotyping*):	Fenotipagem Forense por DNA
F_{ST} (*Fixation Index*):	Índice de Fixação
He (*Expected Heterozygosity*):	Heterozigosidade esperada
Ho (*Observed Heterozygosity*):	Heterozigosidade observada
HS (*High Sensitivity*):	Alta Sensibilidade
HWE (*Hardy-Weinberg Equilibrium*):	Equilíbrio de Hardy-Weinberg
IBGE:	Instituto Brasileiro de Geografia e Estatística
IBM SPSS (*Software*)	
IISNP (*Individual Identification SNP*):	SNP de Identificação Individual

ISP (*Ion Sphere Particle*)	
LD (*Linkage Disequilibrium*):	Desequilíbrio de Ligação
LISNP (*Lineage-informative SNP*):	SNP Informativo de Linhagem
MCMC (*Markov Chain Monte Carlo*):	Monte Carlo via Cadeias de Markov
MDS (*Multidimensional Scaling*):	Escalonamento Multidimensional
MP (*Match Probability*):	Probabilidade de Combinação
MPS (*Massively Parallel Sequencing*):	Sequenciamento Massivo Paralelo
mtDNA (*Mitochondrial DNA*):	DNA mitocondrial
NAM (*Native American*):	Nativo Americano
NGS (*Next-Generation Sequencing*):	Sequenciamento de Nova Geração
Pb:	Pares de Bases
PCR (*Polymerase Chain Reaction*):	Reação em Cadeia da Polimerase
PD (*Power of Discrimination*):	Poder de Discriminação
PE (*Probability of Exclusion*):	Probabilidade de Exclusão
PIC (*Polymorphism Information Content*):	Conteúdo de Informação Polimórfica
PISNP (*Phenotype-informative SNP*):	SNP Informativo de Fenótipo
PPGBCM:	Programa de Pós-Graduação em Biologia Celular e Molecular
PUCRS:	Pontifícia Universidade Católica do Rio Grande do Sul
RS:	Rio Grande do Sul
SNP (*Single Nucleotide Polymorphism*):	Polimorfismo de Nucleotídeo Único
STR (*Short Tandem Repeat*):	Repetição Curta Consecutiva
STRAF (*STR Allele Frequency - Software*)	
TPI (*Typical Paternity Index*):	Índice de Paternidade Típico
VNTR (*Variable Number of Tandem Repeat):*	Repetição Consecutiva de Número Variado

SUMÁRIO

ARTIGO CIENTÍFICO

Capítulo 1

1. INTRODUÇÃO GERAL

1.1 Aspectos Forenses da Identificação Humana

1.2 Marcadores Genéticos

1.2.1 *STR*

1.2.2 *SNP*

1.3 Fenotipagem Forense por DNA

1.3.1 *Ancestralidade Biogeográfica*

1.4 Sequenciamento Massivo Paralelo

1.4.1 *Precision ID Ancestry Panel*

1.5 A População Brasileira

1.5.1 *Rio Grande do Sul (RS)*

2. OBJETIVOS

1. INTRODUÇÃO GERAL

1.1 *Aspectos Forenses da Identificação Humana*

Na filosofia, "*identidade*" é o que torna uma entidade definível e reconhecível, em termos de possuir um conjunto de qualidades ou características que a distinguem de outras entidades. "*Identificação*", portanto, é o ato de estabelecer essa identidade. A identificação humana e, mais especificamente, os aspectos biológicos da identidade humana, estão fundamentados nas ciências bem definidas e estatisticamente verificáveis da biologia, química e física [1]. Os indicadores biológicos de identidade aproveitam a singularidade composta de nossos corpos para fornecer assinaturas que podem confirmar nossa legitimidade com razoável segurança. A evidência física de identidade é geralmente fornecida pelo exame externo de características, por exemplo, cor da pele, sexo, tatuagens, cicatrizes ou características específicas, como impressões digitais. Por outro lado, o exame interno da evidência física de identidade é obtido a partir de informações médicas e/ou científicas (por exemplo, fraturas cicatrizadas, condições patológicas, grupos sanguíneos, evidências odontológicas ou DNA) [1–3].

As abordagens de identificação mais tradicionais tornam-se insuficientes a depender do estado de conservação dos vestígios e da complexidade das amostras, bem como da ausência de elementos comparativos [4]. Nesse contexto, a análise do DNA emerge como ferramenta-chave na obtenção de dados fundamentais à identificação humana. Essa abordagem, quando realizada de acordo com diretrizes rígidas, é altamente confiável para condenar criminosos e, igualmente importante, exonerar indivíduos inocentes [1,5]. Por exemplo, *The Innocence Project*, um projeto fundado em 1992 nos Estados Unidos, auxiliou, até então, 367 pessoas a comprovarem sua inocência por meio dos testes genéticos, das quais 21 cumpriam pena no corredor da morte. Ademais, foram identificados 162 autores dos crimes em questão, condenados, ainda, por 152 crimes violentos adicionais [6]. O preceito básico dos referidos exames de DNA envolve a caracterização genética da amostra questionada e da(s) amostra(s) de referência, com posterior comparação, aplicação da estatística de dados e geração de um relatório que guarnecerá a investigação [1,5].

1.2 *Marcadores Genéticos*

O conceito de *DNA Profiling*, isto é, o processo de determinar as características do DNA de uma pessoa, foi introduzido pelo geneticista Alec Jeffreys em 1985, após a constatação de que certas regiões do DNA eram altamente variáveis entre indivíduos [7]. Jeffreys, mediante a técnica *Southern Blot*, demonstrou o potencial de regiões polimórficas do tipo VNTR (*Variable Number of Tandem Repeats*) – fragmentos de DNA de seis a cem pares de bases que se repetem centenas de vezes – em propiciar diferenciação individual pelo comprimento das sequências após a ação de uma enzima de restrição [7,8]. A análise dessas regiões polimórficas forneceu uma, alcunhada à época, "impressão digital do DNA" (*DNA fingerprint*) – o que hoje conhecemos por "perfil genético". A técnica de geração de perfis genéticos foi inicialmente aplicada em testes de paternidade no Reino Unido, quando, ainda em 1985, foi utilizada para resolver um caso de imigração [9] e, posteriormente, empregue na resolução de casos criminais e na confirmação identitária de restos mortais [10]. Entretanto, apesar de eficaz, a aplicação forense da análise de marcadores VNTR atrelava-se a críticas limitações: demandava grande quantidade (pelo menos 10 a 25 ng de DNA) e satisfatória qualidade das amostras (DNA relativamente intacto, com fragmentos de até 10.000 pares de bases (pb)), com demorada preparação e tempo de análise, e laboriosa comparação de resultados entre diferentes laboratórios [11].

O advento da PCR (*Polymerase Chain Reaction*), descrita na década de 80 pelo bioquímico Kary Mullis e caracterizada como um procedimento capaz de amplificar regiões específicas do DNA, impactou drasticamente a biologia molecular e, por consequência, a genética forense [12]. A técnica aumentou a sensibilidade do exame genético a ponto de gerar perfis de DNA a partir de um reduzido número de células, com possibilidade de aplicação em amostras degradadas e de investigação de quaisquer polimorfismos do genoma [11,12]. Inovadoras metodologias de extração e quantificação de DNA, desenvolvimento de kits comerciais de tipagem baseados em PCR, e a concepção de equipamentos modernos para detecção de polimorfismos genéticos contribuíram para o grande avanço na área. Junto ao progresso técnico, um rigoroso processo de padronização e controle de qualidade contribuiu para o aperfeiçoamento dos métodos utilizados em genética forense [13,14].

Como mencionado, a capacidade de produzir perfis altamente discriminatórios para aplicação na casuística forense depende de os indivíduos diferirem em nível genético. O genoma humano haploide compreende cerca de 3,2 bilhões de pares de bases, dos quais apenas cerca de 1,5% são expressos em produtos gênicos. A maior porção do genoma caracteriza-se por regiões que não contêm informações genéticas codificantes [15]. A variabilidade genética limita-se, sobremaneira, a essas regiões, devido à restrição funcional ocasionada por pressões seletivas sobre as porções codificantes, e à permissividade a mutações neutras em regiões cujas pressões evolutivas encontram-se relaxadas. Nesse caso, as mutações germinativas são comumente mantidas e transmitidas aos descendentes, ocasionando um aumento significativo na variabilidade genética. Essas regiões, portanto, são bastante apropriadas para fins de identificação na genética forense, pois são alvos potenciais para geração de perfis genéticos [1,11,16]. As vantagens fornecidas pela individualização genética incluem: identificação de vítimas de crimes, acidentes e desastres em massa, identificação de pessoas desaparecidas e de infratores por vestígios deixados em locais de crimes, investigações de paternidade e outros vínculos genéticos, estudos evolutivos e aplicações médicas em geral. O repertório de marcadores genéticos utilizados nas rotinas forenses tem crescido substancialmente [1,5,17]; aqui, serão abordadas duas classes de marcadores: *Short Tandem Repeats* (STR), os mais amplamente utilizados em identificação individual atualmente, e os multifuncionais *Single Nucleotide Polymorphisms* (SNP).

1.2.1 *STR*

O termo DNA satélite refere-se à fração do genoma altamente repetitiva, isto é, sequências de DNA que se repetem milhares ou milhões de vezes. Do ponto de vista forense, regiões repetitivas curtas são particularmente interessantes [18]. Os anteriormente mencionados VNTRs, ou minissatélites, foram os primeiros polimorfismos aplicados na geração de perfis genéticos para a casuística forense. Posteriormente, minissatélites foram substituídos pela análise de fragmentos ainda menores – os microssatélites, repetições curtas consecutivas (STRs) compostas por 2 a 6 pares de bases repetidos em *tandem* [11,19]. A variação genética entre indivíduos nestes sistemas de minissatélites e STRs baseia-se essencialmente no número de elementos repetitivos arranjados em série; todavia, as repetições podem abranger pequenas diferenças nas sequências de nucleotídeos, as quais também contribuem para a variabilidade [16,19]. Embora a estrutura geral de VNTRs e STRs seja a mesma, o último supera as limitações do primeiro e satisfaz os requisitos-chave inerentes a um marcador forense: são robustos e possibilitam a análise bem-sucedida de uma ampla gama de materiais biológicos; os resultados gerados em diferentes laboratórios são facilmente comparados; são altamente discriminatórios, especialmente se analisados diversos *loci* simultaneamente (*multiplexing*); com alta sensibilidade, demandam apenas algumas células para a geração do perfil genético; e há um número satisfatório de STRs evolutivamente neutros no genoma [11,19]. Atualmente, os STRs são o padrão-ouro para estabelecer a identidade de amostras humanas, cuja análise envolve a amplificação do material genômico por PCR e posterior detecção dos perfis genéticos por eletroforese capilar (CE), com base no tamanho dos fragmentos [20]. Há uma miríade de kits de marcadores STRs *multiplex* disponíveis comercialmente, com alta robustez e poder discriminatório, destinados à rotina em genética forense [21].

Em 1995, as análises forenses do DNA viabilizaram a criação do Banco de Dados Nacional de Perfis Genéticos do Reino Unido; em 1997, o FBI (*Federal Bureau of Investigation*, Estados Unidos) fundou o sistema CODIS (*Combined DNA Index System*) para o armazenamento de informações de perfis genéticos no *DNA Database* nacional [11]. Anteriormente composto por 13 marcadores STRs (CSF1PO; D3S1358; D5S818; D7S820; D8S1179; D13S317; D16S539; D18S51; D21S11; FGA; TH01; TPOx; e vWA), o CODIS passou a contar com mais sete (D1S1565; D2S1338; D2S2441; D10S1248; D12S391; D19S433; e D221045), totalizando 20 marcadores genéticos do tipo STR para análise com fins de persecução penal [22]. Atualmente, o CODIS é o sistema de armazenamento de perfis genéticos mais amplamente difundido entre os países, incluindo o Brasil, que iniciou sua implementação junto ao Banco Nacional de Perfis Genéticos (BNPG) a partir de 2010 [23].

1.2.2 *SNP*

A conclusão do Projeto Genoma Humano e o Projeto Internacional HapMap proporcionaram à comunidade científica um repositório de informações de referência sobre o genoma nuclear humano [24,25]. Por exemplo, constatou-se que aproximadamente 85% da variação humana é baseada em polimorfismos de nucleotídeo único (SNPs) [26,27]. Esses polimorfismos são variações de base única que ocorrem em uma posição específica do genoma. SNPs são majoritariamente bialélicos, portanto não fornecem a mesma capacidade informativa por *locus* do que marcadores STRs: estima-se que um conjunto de 50 a 100 SNPs seja requerido para a obtenção do mesmo poder discriminatório dos 10 a 15 STRs de alelos múltiplos rotineiramente empregados em identificação individual [27,28]. Por esse e outros motivos – por exemplo, interpretação de misturas e bancos de perfis genéticos bem estabelecidos e sustentados com dados de STRs –, é improvável, no futuro próximo, que SNPs substituam STRs como marcadores genéticos predominantes na identificação humana [29]. No entanto, SNPs apresentam características que os tornam vantajosos para aplicação forense: taxas de mutação reduzidas em relação aos *loci* STRs (atributo valioso para investigação de paternidade e outros vínculos genéticos); *amplicons* de PCR muito curtos (por exemplo, 50 pb ou menos) para análise de amostras cujo DNA encontra-se altamente degradado; ausência de *stutter artefacts* na geração do perfil (artefatos causados pelo deslizamento da polimerase durante a amplificação dos STRs); possibilidade de genotipagem de centenas a milhares de *loci* por uma gama crescente de tecnologias de alto-rendimento; e maior facilidade de validação – como polimorfismos bialélicos –, devido à viabilidade de estimativas precisas das frequências alélicas, fundamentais para a interpretação de dados de genotipagem forense, analisando um número reduzido de amostras comparado ao necessário para estimativas de frequências alélicas de STRs [28,30,31].

Embora dubitável a substituição dos STRs como a principal classe de marcadores para identificação genética, a introdução das análises envolvendo SNPs ampliou a abrangência da genética forense, possibilitando contribuições significativas: elucidação de linhagens matrilineares (DNA mitocondrial (mtDNA)) [32] e patrilineares (cromossomo Y) [33]; análises de amostras altamente degradadas provenientes de desastres em massa e catástrofes naturais [34]; predição de características externamente visíveis (*Externally Visible Characteristics* (EVCs)) de fontes de DNA humano obtidas de locais de crime [35]; e inferência de ancestralidade biogeográfica (*Biogeographic Ancestry* (BGA)) [36]. A análise dos alvos preditivos mencionados (EVC e BGA) caracteriza o processo de prever o fenótipo de um indivíduo utilizando apenas informações genéticas, e denomina-se Fenotipagem Forense por DNA (*Forensic DNA Phenotyping* (FDP)) [35,37].

Em síntese, SNPs podem ser classificados nas quatro categorias aplicáveis a fins forenses apresentadas na **Tabela 1**.

Tabela 1: Categorias de SNPs e respectivas aplicações em genética forense (adaptada de Budowle e van Daal, 2008 [**28**]).

Categorias de SNPs	Aplicação Forense
SNPs de Identificação Individual (IISNPs)	Individualização; requerem alta heterozigosidade e baixo índice de fixação (F_{ST}) (ou seja, baixa diferenciação populacional).
SNPs Informativos de Ancestralidade (AISNPs)	Estabelecer com elevada acurácia a origem biogeográfica de um indivíduo, com o propósito de agregar informações à investigação; requerem baixa heterozigosidade e altos valores de F_{ST}.
SNPs Informativos de Fenótipo (PISNPs)	Prover com alta acurácia a probabilidade de que um indivíduo possua uma característica fenotípica em particular, como cor da pele, olhos e cabelos, com o propósito de agregar informações à investigação.
SNPs Informativos de Linhagem (LISNPs)	Identificação por análise de parentesco mediante conjuntos de SNPs fortemente ligados que funcionam como haplótipos.

1.3 *Fenotipagem Forense por DNA*

Em geral, a abordagem comparativa de perfis genéticos – independente da classe de marcador utilizada – apenas permite a identificação de indivíduos já conhecidos pelas autoridades investigadoras. Na ausência de bancos de perfis genéticos universais ou amplos o suficiente, a fenotipagem forense por DNA surge como uma alternativa promissora para orientar as investigações policiais. Como mencionado anteriormente, essa abordagem inclui a inferência de ancestralidade biogeográfica e de EVCs diretamente de uma amostra de DNA [37,38]. O propósito é, por exemplo, reduzir o número de potenciais suspeitos em casos criminais e, assim, concentrar e orientar as investigações para encontrar o autor anteriormente desconhecido; ou, ainda, fornecer subsídios para a identificação de pessoas desaparecidas, incluindo vítimas de acidentes e desastres em massa. Deve-se enfatizar que tal abordagem somente deve ser aplicada a amostras desconhecidas pelas autoridades, de maneira semelhante à forma como as declarações de testemunhas oculares são utilizadas hoje; contudo, ao contrário das declarações testemunhais, que parecem abarcar sérias taxas de erro [39], o valor das informações sobre EVCs e BGAs estimadas por análise de DNA pode ser estatisticamente suportado [37,38].

1.3.1 *Ancestralidade Biogeográfica*

A análise da ancestralidade biogeográfica concentra-se na variação populacional encontrada em um indivíduo e possibilita a sinalização da origem de seus ancestrais biológicos a partir de uma determinada região geográfica [36]. Isso posto, a genotipagem de variantes genéticas com frequências alélicas altamente diferenciadas entre populações permite a inferência da ancestralidade genômica em nível populacional e individual. Qualquer classe de marcador molecular pode conter exemplares informativos de ancestralidade (*Ancestry-informative Marker* (AIM)), contanto que satisfaça a condição supracitada [40]. A inferência da BGA possui uma ampla gama de aplicações com propósitos forenses, incluindo (1) obter informações sobre doadores desconhecidos de amostras biológicas encontradas em locais de crime; (2) auxiliar revisões de casos arquivados com dados adicionais sobre perfis vinculados; (3) auxiliar na identificação de ossadas de pessoas desaparecidas ou vítimas de desastres em massa; (4) corroborar relatos de testemunhas oculares sobre a etnia percebida de uma pessoa; (5) confirmar ascendência autodeclarada de doadores coletados para bancos de dados genéticos, como aqueles compilados para STRs, marcadores do cromossomo Y e variação mitocondrial; e (6) ajudar a avaliar combinações atípicas de características físicas devidas à ancestralidade miscigenada do indivíduo [36,41]. Em síntese, a BGA pode indiretamente fornecer informações (ainda que limitadas) sobre a aparência geral de um indivíduo, com potencial valor investigativo.

A inferência da BGA é baseada nas características genéticas que o indivíduo herdou de seus ancestrais biológicos. Quanto mais afastadas as regiões geográficas de origem de duas pessoas, maior a diferenciação genética entre elas. As divergências devem-se a mutações, migrações ao longo da história humana, seleção local, isolamento genético, outros efeitos (inclusive aleatórios) e hereditariedade [36]. Os haplótipos do cromossomo Y e do mtDNA são utilizados para reconstruir a história evolutiva das populações humanas e, portanto, podem inferir parcialmente a BGA de um indivíduo. No entanto, a herança uniparental e a representação limitada no genoma humano não tornam esses marcadores genéticos bons preditores indiretos de fenótipo [28]. O método indireto de avaliação do fenótipo proporcionado pela BGA baseia-se na correlação da expressão fenotípica com certos elementos das estruturas genéticas populacionais humanas, e tem sido explorado com base em SNPs informativos de ancestralidade (*Ancestry-informative SNP* (AISNP)) amplamente distribuídos pelo genoma [40,42-44]. Recomenda-se que a inferência de BGA e a predição de EVCs sejam aplicadas em conjunto na prática forense, contidas em uma estrutura regulatória legal apropriada. A vantagem de se considerar a BGA no processo de fenotipagem por DNA é que esta é capaz de fornecer discernimento de padrões miscigenados detectados por análises comparativas [45].

1.4 *Sequenciamento Massivo Paralelo*

Desde a introdução do método de sequenciamento de Sanger nos anos 70 [46], a tecnologia de sequenciamento de DNA alavancou enormes avanços no âmbito da genética e biologia molecular. O Projeto Genoma Humano foi concluído com sucesso mediante essa tecnologia, bem como projetos genômicos de outras inúmeras espécies. Todavia, as desvantagens do sequenciamento convencional de Sanger – incluindo baixo rendimento, alto custo e dificuldades de operação – limitaram seu uso em análises genômicas mais profundas e complexas [47]. A introdução da tecnologia de sequenciamento massivo paralelo (*Massively Parallel Sequencing* (MPS)) – também denominada *Next-Generation Sequencing* (NGS) para distinguir os novos progressos das tecnologias anteriores – superou amplamente esses obstáculos, potencializando aplicações em diversas áreas das ciências da vida, incluindo a genética forense [48]. Por definição, milhões de moléculas de DNA são sequenciadas simultaneamente em reações químicas autônomas, aumentando substancialmente o rendimento e dispensando a necessidade do método de clonagem de fragmentos frequentemente utilizado no sequenciamento de Sanger. Ainda, as saídas do MPS são detectadas diretamente, sem necessidade de uma etapa de eletroforese [49]. A **Figura 1** ilustra o fluxo de trabalho geral do sequenciamento de alto rendimento (MPS), com a listagem das tecnologias disponíveis para cada etapa a depender do instrumento utilizado.

A análise forense do DNA é confrontada com amostras altamente degradadas e contaminadas, *templates* com baixo número de cópias, necessidade de alta precisão e reprodutibilidade, além de considerações de tempo e custo. Atualmente, a maioria dos exames genéticos forenses emprega métodos de análise de fragmentos baseados em PCR e eletroforese capilar para detectar variação de comprimento em marcadores STR. No entanto, a análise baseada em CE possui limitações, por exemplo, incapacidade de analisar múltiplos polimorfismos genéticos em uma única reação e em um único fluxo de trabalho; genotipagem de baixa resolução dos marcadores atuais; perda de informações genômicas úteis de amostras de DNA degradadas; e baixa resolução em análises de mtDNA e mistura. Essas limitações do sequenciamento de primeira geração levaram cientistas forenses a explorar a utilidade da tecnologia MPS para aplicação na área [47,48], com notório impacto no âmbito da fenotipagem forense por DNA.

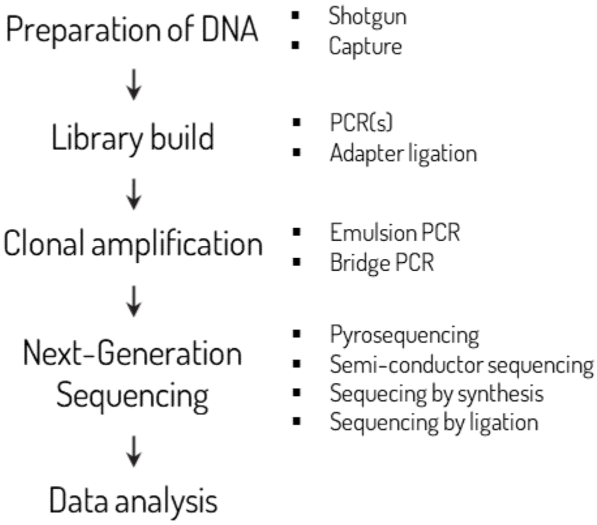


Figura 1: *Workflow* para sequenciamento de alto rendimento. À direita, tecnologias disponíveis para cada etapa a depender da plataforma utilizada. Adaptada de Børsting e Morling, 2015 [48].

1.4.1 *Precision ID Ancestry Panel*

O *Precision ID Ancestry Panel*, comercializado pela Thermo Fisher Scientific Inc. (Waltham, MA, EUA), consiste em 165 SNPs autossômicos que possibilitam inferência de BGA. 55 desses marcadores foram selecionados com base em uma publicação do laboratório do Dr. Kenneth Kidd [50], e 123 são provenientes de uma publicação do Dr. Michael Seldin [51], com 13 SNPs sobrepostos entre os dois painéis constituintes. É relatado que o painel possui cobertura global para grandes populações, incluindo África, Europa, Sudoeste, Sul e Leste da Ásia, Oceania e Américas [52]. Mediante MPS, sobretudo nas plataformas Ion Torrent™ PGM™, Ion S5™ e Ion Chef™ System – baseadas na tecnologia de sequenciamento semicondutor – é possível multiplexar 165 reações de PCR em um tubo com apenas 1 ng de DNA de entrada. Ainda, o tamanho médio dos *amplicons* de cada painel constituinte não excede 130 pb, viabilizando a análise de amostras com DNA degradado [53]. Recentemente, o *Precision ID Ancestry Panel* tem sido aplicado no estudo de diversas populações mundiais com o objetivo de avaliar critérios de validação forense, performance, informatividade, precisão da inferência de BGA, bem como relações genética interpopulacionais e padrões de mestiçagem [54–64].

1.5 *A População Brasileira*

País de proporções continentais habitado por aproximadamente 210 milhões de pessoas [65], o Brasil foi primeiramente povoado por uma ampla variedade de grupos nativos americanos – que compreendiam cerca de 2,5 milhões de pessoas na época da descoberta portuguesa, em 1500. A mestiçagem, predominantemente assimétrica – entre homens europeus e mulheres indígenas – ocorreu imediatamente; no entanto, conflitos e doenças contribuíram para uma redução drástica da população nativa [66,67]. Posteriormente, entre os séculos XVI e XIX, um número estimado de 4 milhões de africanos (sobretudo da atual Guiné, Congo, Angola, Moçambique e Nigéria), foram compelidos a atravessar o Atlântico, como escravos, e direcionados às fazendas de cana-de-açúcar, minas de ouro e diamantes e plantações de café [66,68]. Esse número somente se aproxima ao da imigração europeia que ocorreu nos séculos XIX e XX, majoritariamente oriunda de Portugal, Itália, Espanha e Alemanha. Ainda, em meados do século XX, o país recebeu considerável imigração japonesa e, em menor escala, de outras nacionalidades (Rússia, Polônia e Oriente Médio, por exemplo) [69]. Esses povos encontraram-se e relacionaram-se de maneiras distintas, originando uma população multiétnica altamente miscigenada. Embora a formação biológica do povo brasileiro se deva à contribuição de nativos americanos, europeus e africanos, pode haver uma influência relativa maior de um ou outro grupo a depender da região geográfica [70].

Na atual classificação do Instituto Brasileiro de Geografia e Estatística (IGBE), a respeito do quesito "cor ou raça" – com um critério misto de fenótipo e ancestralidade – encontram-se as seguintes categorias: brancos, pardos, pretos, amarelos e indígenas. No censo de 2010, 47,7% da população (91 milhões) se autodeclararam como brancos; 43,1% (82 milhões) como pardos (multirracial); 7,6% (15 milhões) como pretos; 1,1% (2 milhões) como amarelos e 0,4% (817 mil) como indígenas [71]. A progressão e o direcionamento das colonizações foram deveras diversificados nas cinco regiões brasileiras (Norte, Nordeste, Centro-Oeste, Sudeste e Sul). Este processo demográfico complexo é consequentemente refletido na variação da composição genética das populações atuais [72].

1.5.1 *Rio Grande do Sul (RS)*

O Rio Grande do Sul (RS) é o Estado mais meridional do Brasil, com estimativa atual de aproximadamente 11 milhões de habitantes [65]. Na época em que os primeiros colonizadores europeus chegaram – vindos da ilha dos Açores, em 1740 –, habitavam a região nativos americanos pertencentes a três grandes grupos: (1) Guarani (ramo linguístico Tupi); (2) Kaingang (ramo linguístico Jê); e (3) tribos pampeanas (Charrua, Minuano, Guenoas, etc.; extintos antes das primeiras décadas do século XIX) [73]. A colonização efetiva do RS iniciou apenas no século XVIII, e o controle da região alternava entre os impérios espanhol e português [74]. Dados históricos e genéticos revelaram que uniões assimétricas (entre homens portugueses/espanhóis com mulheres indígenas/africanas) caracterizaram a história demográfica inicial do Rio Grande do Sul [75]. Até 1824, data que marca o início da imigração alemã no RS, predominavam as etnias portuguesa, africana e açoriana, com a presença de nativos no norte do Estado. As levas de imigrantes alemães se sucederam, e aos poucos transformaram o perfil do RS. A partir de 1875, o Estado acolhera uma nova fonte de imigrantes: os italianos, que igualmente participaram da formação da sociedade gaúcha. Ainda, houve, no RS, a inserção de etnias minoritárias diversas. Em suma, a população do RS é em grande parte formada por descendentes de portugueses, alemães, italianos, africanos e indígenas, e em menor parcela por descendentes de espanhóis, poloneses e franceses, dentre outros imigrantes [76]. Atualmente, a população do RS autodeclara-se da seguinte forma quanto à cor ou raça: 83,3% (8,5 milhões) como brancos; 10,5% (1 milhão) como pardos; 5,6% (580 mil) como pretos; 0,3% (35 mil) como amarelos; e 0,3% (33 mil) como indígenas [77].

2. OBJETIVOS

Este estudo tem por objetivo principal analisar os 165 marcadores genéticos informativos de ancestralidade do painel comercial *Precision ID Ancestry Panel* em 250 indivíduos pertencentes aos quatro principais grupos populacionais do Estado do Rio Grande do Sul (RS)[1], Brasil, mediante tecnologia de sequenciamento massivo paralelo. Objetivos específicos incluem:

- Determinar, para cada marcador em cada grupo étnico, as frequências alélicas, genotípicas, e os parâmetros estatísticos forenses associados;
- Verificar a aderência dos marcadores ao Equilíbrio de Hardy-Weinberg e a presença de Desequilíbrio de Ligação;
- Explorar as relações genéticas e estruturas populacionais entre os quatro grupos étnicos e entre estes e 89 populações mundiais;
- Estimar os componentes e proporções de ancestralidade nos quatro grupos populacionais investigados;

[1] Gaúchos de origem africana (AFRS), europeia (EURS), ameríndia (AMRS) e mista (ADRS).

Capítulo 2

3. ARTIGO CIENTÍFICO

ARTIGO CIENTÍFICO

TÍTULO: Ancestry resolution of South Brazilians by forensic 165 ancestry-informative SNPs panel.

AUTORES: Aline Brugnera Felkl[a,c,*], Eduardo Avila[a,b,c], André Zoratto Gastaldo[a,c], Catieli Gobetti Lindholz[a], Márcio Dorn[a,c,d], Clarice Sampaio Alho[a,c].

[a] Forensic Genetics Laboratory, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil.
[b] Technical Scientific Section, Federal Police Department in Rio Grande do Sul State, Porto Alegre, RS, Brazil.
[c] National Institute of Science and Technology – Forensic Science, Porto Alegre, RS, Brazil.
[d] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.
****Corresponding author at:*** Laboratório de Genética Forense, Escola de Ciências da Saúde e da Vida, Pontifícia Universidade Católica do Rio Grande do Sul. Av. Ipiranga, 6681, 90619-900. Prédio 12C, Sala 233. Porto Alegre, RS, Brazil.
E-mail: aline.felkl@edu.pucrs.br.

PERIÓDICO: Forensic Science International: Genetics.

ANO: 2020.

FATOR DE IMPACTO: 4.884 (2018).

CATEGORIA (CAPES): A1.

1. INTRODUCTION

Forensic DNA phenotyping (FDP) includes biogeographic ancestry (BGA) inference and externally visible characteristics (EVCs) prediction directly from an evidential DNA sample as alternatives to provide valuable intelligence when conventional DNA profiling fails to achieve an identification [1]. FDP may reduce the pool of potential suspects and hence guide investigations to find previously unknown perpetrators, as well as helping identify missing persons or mass disaster victims [2]. The indirect method of evaluating physical appearance provided by BGA is based on a set of distinctive, particular features presented by some human groups, which are easily perceivable and recognized as characteristic of such groups. As an example, pigmentation traits are one of the most distinguishing of these physical appearance elements. Different aspects of phenotypic expression can, therefore, be correlated with the different levels of genetic structure observed in human populations, and as such have been widely explored and investigated through techniques based on ancestry-informative markers (AIMs), mainly autosomal single nucleotide polymorphisms (SNPs) [3]. AIMs present marked allele frequency divergences among populations from different geographic regions and are useful for determining an individual's likely biogeographic ancestry or population of origin.

Forensic DNA analysis is often confronted with highly degraded and contaminated samples, requirements for high precision and reproducibility, besides time and cost considerations. In this sense, the advent of Massively Parallel Sequencing (MPS) techniques – used for simultaneous typing of a large number of targeted markers, with high throughput and consequently reduced analysis time – had a hugely positive effect on forensic sciences [4,5]. Soon after, commercial SNP-Panel-based kits for sequencing on high-throughput platforms were introduced to the forensic community. Precision ID Ancestry Panel (formerly HID Ion AmpliSeq™ Ancestry Panel) comprises a set of 165 autosomal ancestry-informative SNPs (AISNPs) previously selected by two laboratories [6,7] and commercially available by Thermo Fisher Scientific (TFS; Waltham, MA, USA) for BGA inference. The average amplicon length is 120–130 bp, projected to successfully allow processing of highly degraded, low input, and other forensic challenging samples.

The Brazilian population is a multicultural and multiethnic nation with a complex demographic history, characterized by intense and heterogeneous admixing processes that encompass three large continental groups – Native Americans (NAM), European (EUR) settlers, and enslaved Sub-Saharan Africans (AFR) [8,9]. The influx of European settlers at the end of the 15th century, mostly coming from the Iberian Peninsula, culminated in both asymmetric mating with Amerindian women and a drastic reduction of the native people due to diseases and conflicts. Soon after, a large contingent of Africans, mostly from Western African territory (Senegal, Gambia, and Guinea-Bissau), was forcedly brought to Brazil as slaves. In the following two centuries, Africans were brought from Angola and Congo; and in the 19th century, the predominant component was from Mozambique [10]. Finally, late migratory movements occurred in the 19th and 20th centuries, with the arrival of Europeans (predominantly Germans, Italians, Portuguese, and Spaniards) and Asian migrants (essentially from Japan and Middle East countries). These peoples met and mated among themselves in different ways, giving rise to a highly admixed multiethnic population [11].

Brazilian territorial occupation followed variable patterns of multidirectional introgression according to social and historical conditions and significantly vary for each distinct geographical region [12]. Heterogeneous processes of migratory flows led to marked divergences in regional ethnical composition, and distinctive proportions of parental populations (NAM, EUR, and AFR) contribution in present-day geopolitical regions are noticeable [13]. Rio Grande do Sul (RS) is the southernmost State of Brazil, with a current estimate of approximately 11 million inhabitants. The history of RS is peculiar since its effective colonization started in the 18th century only. At the time first Europeans arrived, the region was inhabited by Native Americans identified basically with three major groups: (1) Guarani; (2) Kaingang; and (3) Pampean tribes [14]. African contingent established in south Brazil seems to have come mostly from South and East African coasts (current Angola and Mozambique), as well as from the West-Central African region [15]. From the 19th century onwards, large inflows of Germans and Italians gradually transformed the RS profile, shaping its population with one of the highest European ethnic composition of the country.

The present study characterizes the 165 SNPs included in the Precision ID Ancestry Panel (TFS; Waltham, MA, USA) in four main RS State (southern Brazil) population groups (also termed "Gauchos"). We analyzed forensic parameters and conducted population structure analyses among the four population groups apart and along with 89 reference worldwide populations, aiming to scrutinize genetic diversity, similarity levels, ancestry inference, and population stratification of investigated population groups.

2. MATERIALS AND METHODS

2.1 *Ethical Statement*

All samples analyzed in this study were obtained from voluntary donors following informed consent. This work is in accordance with ethical principles stated in World Medical Association's Helsinki Declaration [16] and was approved by the National Research Ethics Committee of CEP/Conep system via Plataforma Brasil, under CAAE number 15620919.3.0000.5336.

2.2 *Samples, DNA extraction, and quantification*

Oral swabs were obtained from 250 unrelated voluntary donors in the metropolitan region of Porto Alegre, Rio Grande do Sul (RS) State, southern Brazil. The population sample comprises 130 women and 120 men, with ages ranging from 18 to 75 years. Subjects provided phenotypic, ethnic, and ancestry information in a self-evaluation form and agreed to the photographic registry. Based on self-declared data and hetero-attribution by multivariate phenotypic evaluation (including eye, skin and hair color, and hair and facial morphology), volunteers were apportioned into four categories: European-derived Gauchos (EURS, *n* = 92), African-derived Gauchos (AFRS, *n* = 62), Amerindian-derived Gauchos (AMRS, *n* = 22, obtained from direct descendants of Guarani and Kaingang population groups from RS), and Admixed-derived Gauchos (ADRS, *n* = 74, characterized by an admixture of two or three parental populations declared by family history and verified by phenotype evaluation).

Genomic DNA from buccal swabs was extracted with a standard phenol-chloroform-isoamyl alcohol protocol. Extracted DNA was quantified using Qubit™ 2.0 Fluorometer with Qubit™ dsDNA High Sensitivity (HS) Assay Kit (TFS; Waltham, MA, USA) according to the manufacturer's recommendations.

2.3 *Library preparation, quantification, and sequencing – Precision ID Ancestry Panel*

Library prep of 132 samples was performed using Ion AmpliSeq™ Library Kit 2.0 (TFS; Waltham, MA, USA) combined with HID-Ion AmpliSeq™ Ancestry Panel (TFS; Waltham, MA, USA). Genomic DNA targets were amplified in a final reaction volume of 20 µL containing 1 µL of template DNA (1 ng), 4 µL of 5x Ion AmpliSeq™ Hi-Fi Mix, 10 µL of 2x Ion AmpliSeq™ primer pool (Ancestry Panel), and 5 µL of nuclease-free water. PCR reaction was performed in a Veriti 96-well Thermal Cycler (TFS; Waltham, MA, USA), under following conditions: enzyme activation at 99°C for 2 min, 21 cycles at 99°C for 15 s and at 60°C for 4 min, and holding at 10°C. PCR amplicons were partially digested with 2 µL FuPa reagent and incubated at 50°C for 10 min, 55°C for 10 min, 60°C for 20 min, and held at 10°C for up to 1 h. Adapters ligation was performed by adding to the 22 µL of digested amplicon: 4 µL of Switch Solution, 0,5 µL of Ion P1 Adapter, 0,5 µL of Barcode X (X was chosen from Ion Xpress™ Barcode Adapters 1-96 Kit or IonCode™ Barcode Adapters 1-384 Kit for different samples), 1 µL of nuclease-free water, 2 µL of DNA ligase and incubated at 22°C for 30 min, 72°C for 10 min, and held at 10°C for up to 1 h. After barcode adapters ligation, libraries were purified with 45 µL of 1.5x Agencourt® AMPure® XP Reagent (Beckman Coulter, FL, USA) and washed two times using freshly prepared 70% ethanol (EtOH), according to manufacturer's instructions.

To assess yield and subsequent normalization, diluted libraries (9 µL at 1:100 dilution) were quantified using a 7500 Real-Time PCR System (TFS; Waltham, MA, USA) with Ion Library TaqMan™ Quantitation Kit (TFS; Waltham, MA, USA). Then multiple libraries diluted to 20 pM were pooled in equivolume for template preparation.

A 25 µL sample of the pooled library was added to the amplification solution to originate template-positive Ion Sphere Particles (ISPs). Emulsion-based clonal amplification (emPCR) was performed on Ion OneTouch™ 2 Instrument (TFS; Waltham, MA, USA) with Ion PGM™ Hi-Q™ View OT2 Kit (TFS; Waltham, MA, USA). Template-positive ISPs were enriched on Ion OneTouch™ Enrichment System (TFS; Waltham, MA, USA). Both emPCR and enrichment were conducted following the manufacturer's protocol (Revision A.0) [17].

Controls and sequencing primers were added to enriched, template-positive ISPs. Sequencing was run on Ion Torrent™ PGM™ Instrument (TFS; Waltham, MA, USA) using an Ion PGM™ Hi-Q™ Sequencing Kit (TFS; Waltham, MA, USA) and an Ion 318™ Chip v2

(TFS; Waltham, MA, USA). A final volume of 30 µL was loaded per chip, according to the manufacturer's instructions (Revision C.0) [18]. Two chips with approximately 65 samples each were used in distinct runs for complete sample set genotyping.

2.4 *Library preparation, quantification, and sequencing – AmpliSeq™ Custom DNA Panel for Illumina®*

Primers were designed by BaseSpace™ DesignStudio™ Sequencing Assay Designer Software (Illumina, CA, USA), using AmpliSeq DNA Hotspot and GRCh38.p2 as reference human genome, at high stringency, a maximum amplicon length of 375 bp, and 100% coverage for the same 165 target SNPs included in the HID-Ion AmpliSeq™ Ancestry Panel.

Genomic DNA of 68 samples was diluted to 10 ng as standard input recommended by the manufacturer's protocol (Document # 1000000036408 v08) [19]. Library preparation was performed using AmpliSeq™ Library PLUS for Illumina® and AmpliSeq™ Custom DNA Panel for Illumina®. Genomic DNA targets were also amplified in a final reaction volume of 20 µL, but with 6 µL of template DNA (10 ng), 4 µL of 5x AmpliSeq™ Hi-Fi Mix, and 10 µL of 2x AmpliSeq™ Custom DNA Panel. PCR reaction was also performed in a Veriti 96-well Thermal Cycler, but under following parameters: enzyme activation at 99°C for 2 min, 18 cycles at 99°C for 15 s and at 60°C for 8 minutes and holding at 10°C. Changes in time and cycles' number considered the 375 bp amplicon length. Amplicons were partially digested similarly to Precision ID Ancestry Panel's library preparation. Indexes I7 and I5 ligation to each sample was conducted using Ampliseq™ CD Indexes Set A for Illumina®, by adding to the 22 µL of digested amplicon: 4 µL of Switch Solution, 2 µL of AmpliSeq CD Indexes, 2 µL of DNA Ligase, and incubated at 22°C for 30 min, 68°C for 5 min, 72°C for 5 min, and held at 10°C for up to 24 h. Libraries were purified with 30 µL AMPure® magnetic beads and washed twice with freshly prepared 70% EtOH. A second amplification step was prepared to guarantee a sufficient library quantity for sequencing on MiSeq® System, as follows: to each library well were added 45 µL of 1x Lib Amp Mix and 5 µL of 10x Lib Amp Primers and incubated at 98°C for 2 min, then 7 cycles of 98°C for 15 min and 64°C for 1 min, and held at 10°C. Subsequently, libraries were subjected to two purification steps using AMPure® magnetic beads and freshly prepared 70% EtOH.

Qubit™ 2.0 Fluorometer and Qubit™ dsDNA HS Assay Kit were used to quantify the libraries. Next, libraries were diluted to starting concentration (2 nM) and pooled with 10 µL of each, afterward denatured with 0.2 N NaOH and diluted to final loading concentration of 9 pM following manufacturer's instructions (Document # 15039740 v10) [20]. Sequencing was performed using the MiSeq® Reagent Kit v2 (500-cycles) on a MiSeq® System instrument (Illumina, CA, USA).

2.5 *Sequencing data analysis*

Ion Torrent™ PGM™: Signal processing (DAT files), base calling, and unmapped and mapped BAM files generation (*Homo sapiens* hg19 as reference genome to perform alignment) were conducted using Torrent Suite™ Software v5.0 (TFS; Waltham, MA, USA). Coverage Analysis v5.0 and Torrent Variant Caller v5.0 plugins were used to calculate the number of mapped reads and perform variant calling, respectively. SNP genotypes were called under standard analysis settings by HID SNP Genotyper v4.3.2 plugin, which allows genotypes filtering at specific locations, given in the hotspot file (here the 165 SNPs that compose Precision ID Ancestry Panel). Minimum coverage was set for six reads per base position, and heterozygote allelic call followed a maximal 70/30 unbalance rate, considering previous studies where the occurrence of allelic unbalance was observed in some genetic markers for HID Ion Ampliseq Precision kits [21,22]. All genotypes and base calls were manually checked by at least two independent reviewers.

MiSeq® System: BaseSpace™ Sequence Hub DNA Amplicon v2.0 App was used to analyze the AmpliSeq™ Custom DNA Panel. Per-sample reads (FASTQ files) were aligned with the BWA algorithm against the reference genome (*Homo sapiens* GRCh38). Variant calling was performed by Pisces Variant Caller at a Depth Filter level of 10 and annotated by Illumina Annotation Engine using RefSeq transcripts. A VCF file containing variants of interest was uploaded to the project for SNP genotypes calling.

Low-pass full genome sequencing: A subset of samples comprising 50 individuals was subjected to full genome sequencing through an external service provider (Gencove Inc., NY, USA). Full sequencing was attained with 1x coverage on an Illumina NextSeq 2000 equipment (Illumina, CA, USA) following library preparation and workflow according to the company's internal procedures, including sequencing protocols and data processing, as described by Wasik and collaborators [23]. Results were provided as data files with different formats and were extracted from provided VCF files using a custom python script. In these files, genetic data is displayed as genotype posterior probabilities, since the bioinformatics pipeline adopted by Gencove includes an imputation step based on the model proposed by Li and Stephens [24], to predict variants located in low coverage regions or undetected during sequencing. A threshold value of 0.98 for the genotype probability was adopted to reduce errors, and the genotype calls rate under the adopted threshold was less than 0.5% (evenly distributed among all 165 SNPs, with no preferential sites for unreliable calls). Genotype calls with reported posterior probability under 0.98 were assigned as missing data.

The conversion of exported SNP genotypes data to downstream software formats was done by PGDSpider v.2.1.1.5 [25]. Allele frequencies of 165 SNPs and corresponding forensic statistical parameters, including observed heterozygosity (Ho), expected heterozygosity (He), polymorphism information content (PIC), match probability (MP), power of discrimination (PD), power of exclusion (PE), and typical paternity index (TPI) were calculated using STR Analysis for Forensics (STRAF) v.1.0.5 [26] online software (available at http://cmpg.unibe.ch/shiny/STRAF/). Random match probability (RMP) calculations were performed with validated, in-house Excel-based workbooks. Exact test of Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium test (LD) were performed using Arlequin v3.5.2.2 [27]. HWE analysis was carried out with 1,000,000 Markov Chain Monte Carlo (MCMC) steps and 1,000,000 dememorization steps. Correction for multiple testing was done according to the method suggested by Bonferroni [28], by dividing the significance level of 0.05 by the number of tests.

2.6 *Data merging and population analyses*

For comprehensive analyses of populations' genetics relationships, our data were combined with genotypic profiles from 89 reference worldwide populations for the 165 AISNPs included in Precision ID Ancestry Panel (TFS; Waltham, MA, USA). 24 populations were extracted from the 1000 Genomes (1kG) Project [29] Phase III and merged with previously published Basques [30] and Chinese Uyghur and Hui [31]. Danes and Somalis' [32] genotypic data were kindly provided by Professor Niels Morling and collaborators. 60 worldwide populations [33] genotyped at Kidd Lab and kindly provided by Professor Kenneth K. Kidd and collaborators also compose reference populations set. Details of populations used in the present study and their abbreviations are listed in **Supplementary Table S1**.

A population differentiation test based on pairwise F_{ST} genetic distances and molecular variance analysis (AMOVA) among our four studied population groups and along with 89 worldwide populations were performed using Arlequin v3.5.2.2. Based on pairwise F_{ST} values, the Multidimensional Scaling (MDS) technique was applied using IBM® SPSS® Statistics v25.0 [34]. Individual ancestry proportions were evaluated using STRUCTURE v.2.3.4 [35], with ten independent runs for each K value, ranging from K = 2 to K = 20. 100,000 burn-in steps followed by 100,000 MCMC repetitions were applied, and 'admixture' and 'correlated allele frequencies' models were considered [36]. Summation and graphical representation of STRUCTURE results were generated using Cluster Markov Packager Across K (CLUMPAK) online server (available at http://clumpak.tau.ac.il/) [37]. To identify the K value that captures the uppermost structure level, we used Structure Harvester v.0.6.94 [38], which implements the Evanno method [39].

3 RESULTS AND DISCUSSION

Three distinct sequencing procedures were adopted to generate genetic profiles of 165 AISNPs in 250 unrelated South Brazilian subjects. A further study evaluating the comparative sequencing performance of these methods is underway. The employed panel comprises 55 autosomal biallelic SNPs from AIM set developed by Kidd group [6,33] and 123 from Seldin's AIM set [7] (13 markers are included in both panels; see **Supplementary Table S2** for SNPs details) and aims to provide biogeographic ancestry information to guide investigative processes. The commercial kit was designed to properly handle degraded DNA samples, with targeted amplicons average size less than 130 bp. Several populations have been investigated using this panel to infer genomic ancestry and population stratification, including Asians (Uyghur and Hui [31], Japanese [40], Chinese Tibetan-Burmese [41], Uyghur and Kazakh [42,43], and other Asian populations [44]), Europeans (Basques [30], Danes [32], and Greenlanders [45]), South Americans (Ecuadorians [46]), Middle Eastern populations (Turks and Iranians [47]) and Africans (Somalis [32]). In the study herein, samples obtained from individuals belonging to the three main ethnicities of Rio Grande do Sul (RS) State, southern Brazil, as well as subjects with multiethnic backgrounds, were firstly investigated to explore genetic relationships and structures within and among them. Subsequently, population stratification analysis and individual ancestry inference were conducted regarding reference populations set.

3.1 *Forensic parameters of 165 SNPs for Rio Grande do Sul (Brazil) main population groups*

The detailed 165 AISNPs genotypes of 250 Brazilian subjects from RS are listed in **Supplementary Table S2**. Observed allele frequencies and forensic parameters estimates of these SNPs, including Ho, He, PIC, MP, PD, PE, and TPI for individual population groups are presented in **Supplementary Table S3-S6**, as well as *p*-values for HWE tests for all loci. Two loci (rs1800414 and rs671) are monomorphic in all four populations investigated. rs3811801 is monomorphic in AFRS, AMRS, and ADRS subsets. rs1871534, rs3916235, and rs7657799 are monomorphic only in Amerindian-derived individuals. Invariable loci rs1800414, rs671, and rs3811801 were also monomorphic for the same alleles in Basques [30], Danes, Somalis [32], Greenlanders [45], and Ecuadorians [46]. Further inquiries at the Ensembl Genome Browser (Release 99) showed that these three markers have the same fixed allele in all European, Native American, and African samples reported to date, while are polymorphic in East Asian populations. Therefore, the lack of genetic variability in these loci should not be extrapolated to the Brazilian population as a whole, as the sampling of this study was conducted in a single Brazilian federative unity (out of 27), particularly the one presenting the lowest rate of Asian ethnic composition, as reported by Brazilian Institute of Geography and Statistics (IBGE) demographic census [48]. These loci are expected to be variable in samples from southeastern Brazil, for instance, given the historical presence of Asian immigrants in this particular region [49].

No statistically significant deviation from HWE was observed after Bonferroni correction ($p > 3.03 \times 10^{-4}$) in any ethnic subset. However, seven loci pairs displayed linkage disequilibrium in pairwise LD testing, even after Bonferroni correction for multiple comparisons ($p < 3.70 \times 10^{-6}$): three pairs in AFRS (two of them also genetically associated in ADRS), three in EURS, and an extra pair in Admixed-derived Gauchos

(**Table 1**). Four out of seven pairs are located on the same chromosome, up to 3.5 cM apart from each other: rs1834619–rs1876482 (Chr. 2), rs260690–rs3827760 (Chr. 2), rs1426654–rs735480 (Chr. 15), and rs3916235–rs4891825 (Chr. 18). The latter pair also had LD statistical significance in Basques [30], Danes, and Somalis [32]. Overall, full recombination and independent inheritance are expected in loci with a genetic distance of over 50 cM [50]. Nevertheless, the aforementioned statistically associated SNPs are located at markedly shorter distances. Besides physical linkage between loci, such non-random associations can be caused by, among other reasons, gene flow among populations with dissimilar allele frequencies, population structure, and small sample sizes [51]. Brazilian populations display varying levels of stratification and complex admixing patterns [52]; therefore, a conjunction of the foregoing factors is presumably inducing the genetic associations observed between seven loci pairs in three RS population groups. LD test *p*-values are detailed in **Supplementary Table S8-S11**. AMRS population presented no significant association among loci after Bonferroni correction.

Table 1: Genetically associated SNP pairs in RS State (Brazil) main population groups. Four out of seven pairs are located on the same chromosome (position based on hg19 genome). *P*-values for linkage disequilibrium tests are also provided.

	Locus #1	**Locus #1 location**	**Locus #2**	**Locus #2 location**	**P-value LD**
AFRS	rs1572018	Chr13: 41715282	rs2166634	Chr10: 118436068	1.76×10^{-06}
	rs1834619	Chr2: 17901485	rs1876482	Chr2: 17362568	7.07×10^{-09}
	rs3916235	Chr18: 67578931	rs4891825	Chr18: 67867663	1.74×10^{-09}
EURS	rs1407434	Chr1: 186149032	rs3827760	Chr2: 109513601	1.03×10^{-06}
	rs260690	Chr2: 109579738	rs3827760	Chr2: 109513601	1.11×10^{-11}
	rs4471745	Chr17: 53568884	rs731257	Chr7: 12669251	2.01×10^{-06}
ADRS	rs1426654	Chr15: 48426484	rs735480	Chr15: 45152371	6.22×10^{-07}
	rs1834619	Chr2: 17901485	rs1876482	Chr2: 17362568	4.22×10^{-08}
	rs3916235	Chr18: 67578931	rs4891825	Chr18: 67867663	1.66×10^{-11}

AFRS = African-derived Gauchos; EURS = European-derived Gauchos; ADRS = Admixed-derived Gauchos.

Observed heterozygosity (Ho) ranges from 0.048 (rs1229984 and rs4471745) to 0.661 (rs1040045) in AFRS, from 0.011 (rs3811801) to 0.576 (rs3784230) in EURS, from 0.045 (rs7251928 and rs7722456) to 0.727 (rs7745461 and rs948028) in AMRS, and from 0.095 (rs1229984) to 0.662 (rs1871428) in ADRS, with average values of 0.361 ± 0.133, 0.274 ± 0.144, 0.355 ± 0.146, and 0.388 ± 0.116, respectively. As expected, the Admixed-derived group (ADRS) has the highest intrapopulational genetic diversity average, followed by the African-derived one. Noteworthy, heterozygosity values indicate greater miscegenation among the South Brazilian Amerindians compared to the European-derived population group. These results reflect the admixed landscape characterizing the Brazilian population and corroborate previous findings regarding their genetic variability in RS population and other Brazilian regions [53–56]. The SNP with the highest discrimination power was rs3916235 (PD = 0.658; MP = 0.342) in AFRS, rs459920 (PD = 0.661; MP = 0.339) in EURS, rs37369 (PD = 0.6653; MP = 0.3347) in AMRS, and rs7554936 (PD = 0.6622; MP = 0.3378) in ADRS. Combined match probability (CMP) was, in the same order as groups above, 2.45×10^{-51}, 8.62×10^{-40}, 1.20×10^{-48}, and 8.82×10^{-56}. In African-derived Gauchos, the SNP with the highest power of exclusion was rs1040045 (PE = 0.3710), while in EURS was rs3784230 (PE = 0.2632). In AMRS population, SNPs with the highest PE were rs948028 and rs7745461, both with a PE value of 0.4717. Combined power of exclusion (CPE) of 165 SNPs included in Precision ID Ancestry Panel was, for AFRS, EURS, AMRS, and ADRS: 99.99999960%, 99.99954437%, 99.99999967%, and 99.99999995%, respectively. CMP and CPE metrics could be regarded as indicators to evaluate the efficiency of genetic markers in forensic individualization. Forensic descriptive parameters of Precision ID Ancestry Panel (TFS; Waltham, MA, USA) estimated for each population group revealed that, although its primary purpose is biogeographic ancestry inference (whereas for an identification tool it is more suitable to use other panels, for instance, the Precision ID Identity Panel [57]), this panel has enough polymorphic and informative SNPs to be used as a supplementary instrument for individual identification in the forensic analytical repertoire.

Moreover, average random match probability (RMP) based on individual genotypic frequencies for all 165 SNPs was calculated for AFRS, EURS, AMRS, and ADRS populations and for RS State as a whole (RSBR). For the latter, an adjusted allele frequencies table was generated considering the relative contribution of each aforementioned group in RS population formation, according to IBGE demographic census [58] **(Supplementary Table S7)**. Results are presented in **Table 2**. A rather significant overlap between ADRS RMPs in the three main ethnic populations ($\mathrm{EURS_{Pop.}}$, $\mathrm{AFRS_{Pop.}}$, and $\mathrm{AMRS_{Pop.}}$) can be observed, corroborating a trihybrid composition to the admixed nature of this population sample. On the other hand, the average probabilities of AFRS, EURS, and AMRS genetic profiles to occur in populations other than their own (and $\mathrm{ADRS_{Pop.}}$, for AFRS and EURS profiles) are at least 25 orders of magnitude lower.

Furthermore, EURS and ADRS are the most frequent genetic profiles found in RSBR population. Wright's *F*-statistics (discussed later) shed light on the above outcomes regarding forensic aspects of the four RS population groups.

Table 2. Average random match probability (RMP) of genetic profiles from each RS State population group in each ethnic population and in RS population as whole (RSBR; adjusted allele frequencies), based on allele frequencies of the 165 SNPs included in Precision ID Ancestry Panel (TFS; Waltham, MA, USA).

	AFRS$_{\text{Prof.}}$	**EURS**$_{\text{Prof.}}$	**AMRS**$_{\text{Prof.}}$	**ADRS**$_{\text{Prof.}}$
AFRS$_{\text{Pop.}}$	3.62E-50 ± 2.03E-49	4.18E-82 ± 2.88E-81	1.16E-88 ± 9.05E-88	8.47E-60 ± 6.42E-59
EURS$_{\text{Pop.}}$	7.89E-75 ± 7.48E-74	7.67E-41 ± 5.40E-40	1.37E-95 ± 1.30E-94	7.67E-54 ± 7.15E-53
AMRS$_{\text{Pop.}}$	1.63E-88 ± 7.39E-88	7.51E-84 ± 3.44E-83	2.36E-48 ± 1.07E-47	4.99E-70 ± 2.29E-69
ADRS$_{\text{Pop.}}$	2.15E-56 ± 1.50E-55	1.56E-48 ± 1.34E-47	3.35E-87 ± 2.85E-86	3.14E-55 ± 2.60E-54
RSBR$_{\text{Pop.}}$	1.27E-72 ± 7.44E-72	6.23E-44 ± 3.07E-43	8.65E-78 ± 3.97E-77	3.05E-49 ± 2.61E-48

AFRS = African-derived Gauchos; EURS = European-derived Gauchos; ADRS = Admixed-derived Gauchos. $^{\text{Prof.}}$ = Profile; $^{\text{Pop.}}$ = Population.

3.2 *Interpopulation genetics analyses*

Based on Precision ID Ancestry Panel (TFS; Waltham, MA, USA), pairwise F_{ST} for RS main population groups ranged from 0.07191 (AFRS and ADRS) to 0.38631 (EURS and AMRS). **Table 3** presents results obtained with pairwise F_{ST} test in investigated population groups. Overall, Amerindian population was found to be the most genetically distinct and structured, with consistently higher observed pairwise F_{ST} values, followed by European, African, and Admixed ethnicities, respectively. Considering the 165 ancestry-informative markers evaluated, there is a remarkable genetic differentiation level among population groups derived from the three main parental populations that bolstered the peopling of Brazil (Europeans, Africans, and Native Americans). Furthermore, the trihybrid multiethnic group displays tighter (and quite similar) genetic relationships with African-derived and European-derived population groups, and more distant (albeit closer than others) with the Amerindian one. These findings contrast with the results of a previous study concerning color and genomic ancestry in Brazilians [59], in which no statistically significant degrees of genetic differentiation was observed among individuals classified as Whites, Intermediates, and Blacks from São Paulo city, southeastern Brazil, by typing of 12 STR loci. The use of forensic STRs for delineating population structure may explain disparities found, as markers with relatively lower mutation rates (SNP, Alu, Indel) are more suitable to provide biogeographic resolution at continental level [60].

Table 3. Pairwise F_{ST} test for RS State (Brazil) main population groups based on 165 SNPs of Precision ID Ancestry Panel (TFS; Waltham, MA, USA). F_{ST} values are presented in lower-left diagonal, while upper-right diagonal exhibits the significance matrix (p = 0.00000).

Population	AFRS	EURS	AMRS	ADRS
AFRS		+	+	+
EURS	0.26051		+	+
AMRS	0.30261	0.38631		+
ADRS	0.07191	0.07702	0.23357	

AFRS = African-derived Gauchos; EURS = European-derived Gauchos; ADRS = Admixed-derived Gauchos.
Significant values were represented by "+" signal.

Furthermore, pairwise F_{ST} values were calculated based on Precision ID Ancestry Panel (TFS; Waltham, MA, USA) SNPs among RS main population groups and 89 reference worldwide populations (see **Supplementary Table S1** for details). Results are displayed as a heatmap in **Supplementary Fig. S1** and pairwise F_{ST} values are detailed in **Supplementary Table S12**. African-derived Gauchos showed higher similarity levels with African Americans (AFRS–ASW: F_{ST} = 0.0162; AFRS–AAM: F_{ST} = 0.0177), followed by Eastern African Somalis (AFRS–SOM: F_{ST} = 0.0372) and Ethiopian Jews (AFRS–ETJ: F_{ST} = 0.0434), and more conspicuous divergence with Native Americans Suruí and Karitiana from Amazon region (AFRS–SUR: F_{ST} = 0.4599; AFRS–KAR: F_{ST} = 0.4366). European-derived Gauchos, on the other hand, presented more genetic proximity with Central and Southern Europe populations (EURS–HGR: F_{ST} = 0.0017; EURS–GRK: F_{ST} = 0.0049; and EURS–TSI: F_{ST} = 0.0055) succeeded by European Americans (EURS–EAM: F_{ST} = 0.0058), and highest differentiation levels with Native American Suruí (EURS–SUR: F_{ST} = 0.5325) and Biaka, pygmies from Central Africa (EURS–BIA: F_{ST} = 0.5309). Brazilians with Amerindian ethnicity from RS State displayed more prominent genetic similarity with Peruvians (AMRS–PEL: F_{ST} = 0.0201), Maya (AMRS–MAY: F_{ST} = 0.0238), and Quechua (AMRS–QUE: F_{ST} = 0.0269), followed by North American Plains Amerindians (AMRS–NPA: F_{ST} = 0.0494), corroborating the admixed nature of AMRS population group. Higher divergence levels were with Biaka pygmies and Western Africans (AMRS–BIA: F_{ST} = 0.5991; AMRS–ESN: F_{ST} = 0.5660; AMRS–YOR:

F_{ST} = 0.5622). Admixed-derived Gauchos, characterized by miscegenation among two or three of Brazilian main ethnic roots (European, African, and Amerindian), revealed higher similarity with Puerto Ricans and Colombians (ADRS–PUR: F_{ST} = 0.0128; ADRS–CLM: F_{ST} = 0.0267), and more evident population differentiation levels with Native Americans from Amazon region (ADRS–SUR: F_{ST} = 0.4005; ADRS–KAR: F_{ST} = 0.3744).

To further investigate the above results regarding interpopulation genetic relationships of RS State main ethnicities and 89 worldwide populations, an MDS plot was drawn based on pairwise F_{ST} values (**Fig. 1**). MDS graph exhibits positive values in Dimension 1 as a characteristic feature for African (AFR) populations. Sub-Saharan African populations are closely clustered at bottom-right edge of the quadrant, while admixed AFR populations have broader dispersion along the axis. AFRS population is relatively close to admixed East African populations (SOM and ETJ) and African Americans (AAM and ASW) in Dimension 1 median positive values. The multiethnic Gaucho group (ADRS) is plotted between the European and African clusters, indicating the substantial presence of these two higher components. This wide-distribution phenomenon is also observed in other Latino populations (PUR, CLM, MLX, and PEL), probably reflecting their admixed nature. Southern and Northern European (EUR) populations are clustered in Dimension 1 negative values. Even with a close disposition, it is feasible to distinguish both EUR regions, besides observing a nearness of Southern EUR with some Middle Eastern populations. European-derived Gauchos subset, as well as European Americans (EAM), are located in the EUR cluster. However, EAM is grouped along with Central and Northern EUR populations, while EURS adjoins Southern EUR ones. Asian populations are spread across both quadrants of Dimension 2, comprising negative values (Middle Eastern populations and Southern Asians), median-positive values (Central, Northern, and Eastern Asians), and higher positive values (Native American (NAM) populations). Amerindian-derived Gaucho group is contiguous to NA cluster, close to MAY, NPA, and QUE populations.

Table 4 shows AMOVA testing results for two datasets: (1) four main population groups from RS State (AFRS, EURS, AMRS, and ADRS) as independent populations; (2) the 93 populations assembled according to geographic location. Considering RS population groups only, among-populations covariance component accounts for an estimate of 18.7% of entire genetic differentiation, whereas 81.3% is due to individual-level divergence. AMOVA results for 93 worldwide populations revealed that 29.8% of genetic differentiation is justified by among-population variance, while 70.2% is due to variation at the individual level. Noteworthy, although there is a significant genetic structuring level among African, European, and Amerindian-derived Gaucho subpopulations, the most comprehensive source of variability is still the individual. Precision ID Ancestry Panel (TFS; Waltham, MA, USA) was designed to identify population genetic structures for ancestry inference purposes; however, it can also access genetic diversity at the individual level. These results support the convenience of using this panel as a supplementary instrument for individual identification in the forensic field.

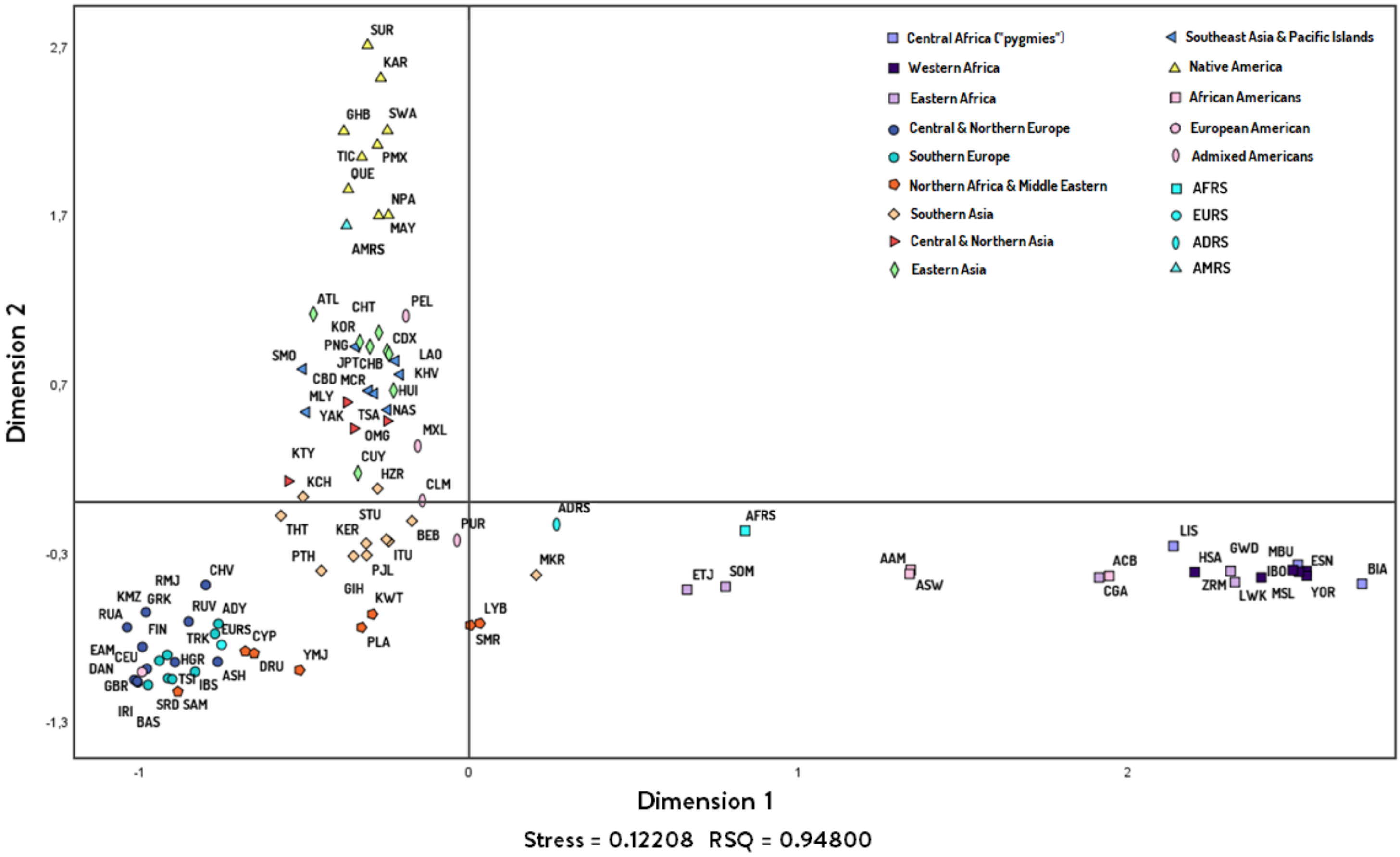


Fig. 1. Genetic distances evaluation among RS State (Brazil) main population groups and 89 worldwide populations, presented as an MDS plot based on pairwise F_{ST} values for 165 SNPs included in Precision ID Ancestry Panel (TFS; Waltham, MA, USA). Genetic distances between all pairs of populations were included, and multi-dimension scaling procedure was applied to reduce dimensionality, from an n-dimensional space to a Cartesian space. Spatial proximity in the plot indicates genetic similarity between populations, while distant populations tend to be located apart from each other.

Table 4. Fixation indexes and AMOVA results for RS State (Brazil) main population groups solely (dataset #1) and along with 89 worldwide populations (dataset #2), based on individual genotypes of Precision ID Ancestry Panel (TFS; Waltham, MA, USA). Statistically significant fixation indexes are highlighted in bold (p ≅ 0.00000).

Dataset	Source of Variation	Relative Variation (%)	Fixation Indexes
#1 Four main population groups from RS State (Brazil)	Among populations	18.66	F_{ST} = **0.18662**
	Among individuals within populations	0.61	F_{IS} = 0.00756
	Within individuals	80.72	F_{IT} = **0.19277**
#2 93 worldwide populations	Among groups	26.85	F_{CT} = **0.26848**
	Among populations within groups	2.94	F_{SC} = **0.04016**
	Among populations	29.79	F_{ST} = **0.29792**
	Among individuals within populations	0.68	F_{IS} = 0.00968
	Within individuals	69.53	F_{IT} = **0.30466**

3.3 *Ancestry inference*

To further characterize the genetic structure of RS State main population groups, ancestry resolution and admixture patterns estimates at populational and individual levels were assessed using Bayesian inference methods, based on 5,000 individuals from 93 worldwide populations. **Fig. 2** presents populational bar charts of estimated cluster membership values from STRUCTURE runs for Brazilian samples alongside 89 reference populations. Estimates are based on individual genotypes for all 165 ancestry-informative SNPs composing Precision ID Ancestry Panel (TFS; Waltham, MA, USA). The optimal number of clusters according to Evanno method is K = 3, although higher K values successfully partitioned the populations into further continental (or even more geographically refined) divisions. When considering runs ranging from K = 5 to 20, Structure Harvester results indicate K = 7 as optimal K number (**Supplementary Fig. S2**). At K = 2 (data not shown) African and non-African ancestry components could be identified. At K = 3, African (blue), European (green), and Native American/Asian (red) ancestry components are discernible.

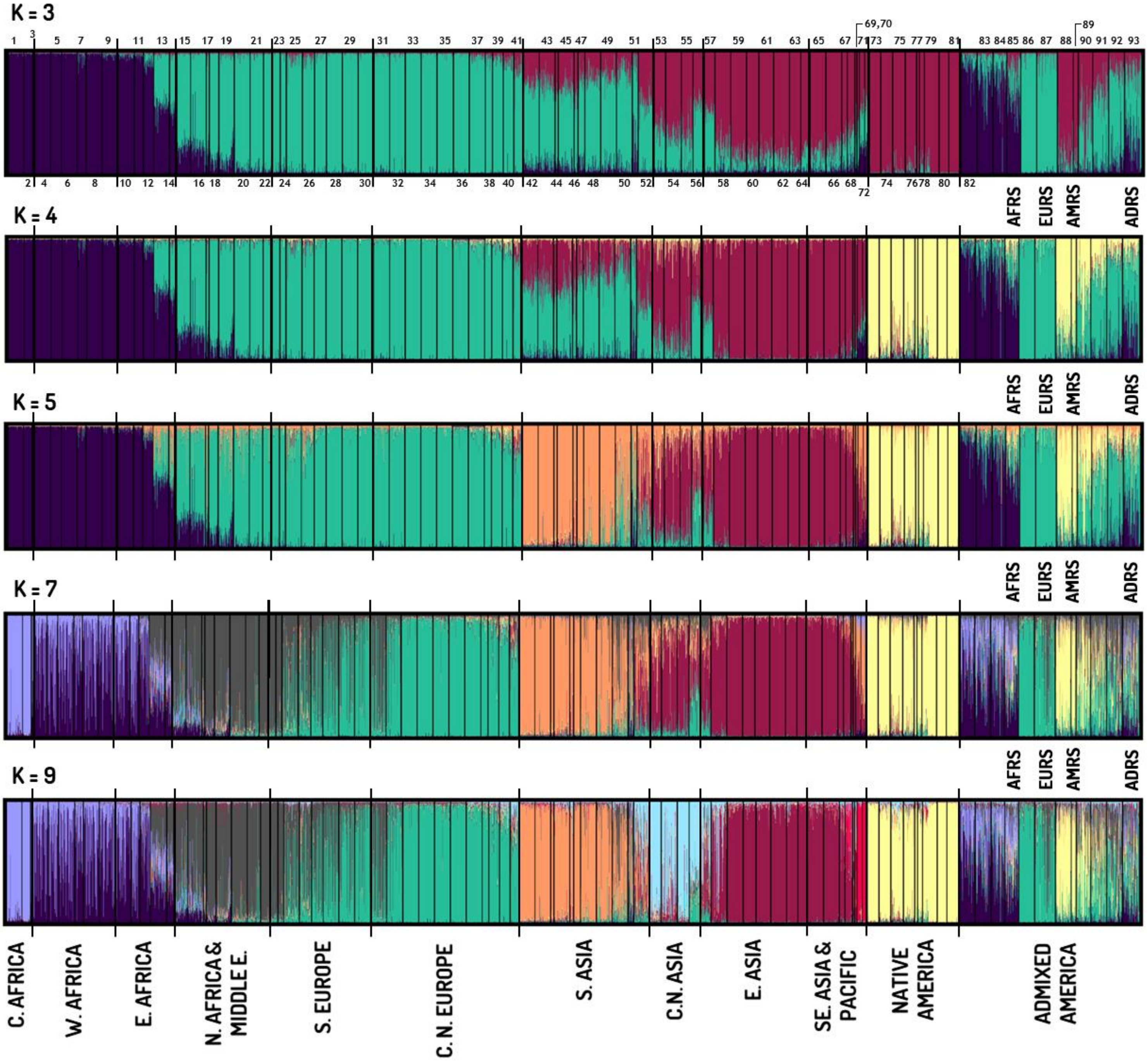


Fig 2. Population structure of RS State (Brazil) main population groups along with 89 worldwide populations, based on 165 SNPs included in Precision ID Ancestry Panel (TFS; Waltham, MA, USA). STRUCTURE plots are presented with cluster (K) number ranging from 3 to 9 (top to bottom; data for K = 6 and K = 8 not shown). The optimal number of clusters was three. Each vertical line stands for an individual, with colors representing the relative proportion of association with each inferred cluster. Populations referring to each number and respective geographic locations are listed in Supplementary Table S1.

At optimal K number of 3, AFRS presents clustering patterns similar to adjoining African-American subpopulations (ACB, AAM, and ASW), although the green (EUR) and red (NAM/Asian) components are more pronounced, suggesting a higher admixing level among the parental populations that originated African-derived Gauchos than in African Americans. Within the AFRS, ancestry proportions range from AFR: 33.7% to 86.1%, EUR: 1.5% to 51.3%, and NAM: 0% to 52.4%. EURS displays a very similar clustering pattern to that of the North American counterpart (EAM) and European populations, with an almost total predominance of European component. Besides, a low NAM/Asian ancestry is also noticeable, and almost none AFR component is perceived. Indeed, within EURS, ancestry proportions vary from AFR: 0% to 1.4%, EUR: 65.9% to 99.5%, and NAM: 0% to 33.2%. AMRS, as the Peruvians (PEL), exhibits an expressive NAM/Asian component. Individually, AMRS ancestry proportions range from 0% to 9.8% (AFR), 0% to 44.2% (EUR), and 55% to 99.4% (NAM). ADRS presents a well-defined admixed pattern, with the three ancestry components clearly discernible. There is a prevalence of EUR composition, followed by AFR and NAM/Asian, respectively, corroborating results obtained with the MDS chart. At individual level, ancestry proportions vary from AFR: 0% to 61.8%, EUR: 19.0% to 89.9%, and NAM: 0% to 43.8%. Average ancestry estimates of RS State population samples (AFRS, EURS, AMRS, and ADRS) were inferred based on both optimal K values and are presented in **Table 5**. Results were extracted from runs with the largest Ln Probability Data [LnP(D)].

Table 5. Ancestry estimates for RS State (Brazil) main population groups for three and seven clusters (K) using 89 worldwide populations as references.

K = 3				
Ancestry:	AFRS	EURS	AMRS	ADRS
African	0.620	0.018	0.026	0.268
European	0.240	0.946	0.089	0.570
Native American	0.140	0.036	0.885	0.162
K = 7				
Ancestry:	AFRS	EURS	AMRS	ADRS
W. African	0.365	0.007	0.017	0.152
C. African	0.255	0.007	0.015	0.110
C. N. European	0.113	0.688	0.066	0.288
SW. Asian/Mediterranean	0.102	0.249	0.057	0.252
S. Asian	0.052	0.021	0.022	0.067
E. Asian	0.023	0.011	0.015	0.023
Native American	0.091	0.017	0.809	0.107

At K = 7, 8, and 9, the Central African, North African/Middle Eastern, Central and North Asia, and Pacific ancestry components become noticeable. Noteworthy, K = 7 distinguishes three Central African ("pygmy") populations from the other Sub-Saharan Africa populations, which now display partial membership to two different clusters. Furthermore, there is a conspicuous transition from North Africa to Southwest Asia, then to Southern Europe, and finally to Northern Europe [61]. Accordingly, Southern Europeans present a "Mediterranean" component and partial assignment to the cluster that is essentially Northern European. The "Mediterranean" component becomes visible in EURS subpopulation, but not in the North American counterpart (EAM), reflecting the unique settlement process of each region.

4 CONCLUSION

In the present study, 250 individuals from RS State, Brazil, apportioned in four main Brazilian population groups were genotyped for 165 AISNPs included in Precision ID Ancestry Panel using massively parallel sequencing technology. Although the main purpose of the Precision ID Ancestry Panel is the biogeographical ancestry inference, forensic effectiveness analysis revealed that the panel could be applied as a supplementary approach in forensic individual identification and kinship testing regardless of ethnicity. However, the use of commercial solutions specifically designed as identification tool applicable to challenging samples, such as the Precision ID Identity Panel [57], is better suited for this purpose. Investigation of genetic differentiation among the four RS population groups shows evidence of a significant genetic structuring degree essentially among the three ethnicities directly derived from parental populations (AFRS, EURS, and AMRS). Therefore, BGA inference could be informative in the context of subpopulations with varying levels of genetic stratification and admixture patterns, although it should be used with caution and, preferably, associated with direct externally visible characteristics predictive markers. Population genetic similarities and divergences among the four RS population groups solely and along with 89 worldwide-distributed reference populations were also investigated. Findings from F_{ST}-based heatmap and MDS plotting demonstrated that Admixed-derived and African-derived Brazilians from RS present the highest levels of admixture and population stratification, being genetically more similar to other admixed populations (respectively, other Latin

American multiethnic populations and African Americans, for instance), whereas European-derived and Amerindian-derived subpopulations exhibit a more homogeneous genetic conformation, similar to their respective parental populations. Finally, the interethnic admixture landscape revealed by the model-based clustering of Structure suggested that AFRS has an essentially trihybrid heritage with larger African ancestry (62.0%) followed by European (24.0%), and a significant Amerindian component (14.0%); EURS has a predominant European ancestry (94.6%), as well as AMRS has a prevailing Amerindian one (88.5%); ADRS has a trihybrid genetic background composed mainly by European ancestry component (57%), followed by African (26.8%), and Amerindian (16.2%).

DECLARATION OF COMPETING INTEREST:

Authors declare they have no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation for the valuable contribution of the volunteers who participated in this study.

FINANCIAL SUPPORT:

The present work was funded by grants provided by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001*, Conselho Nacional de Pesquisa Científica – Brasil (CNPq) - Chamada n° 16/2014 – 465450/2014-8, and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul – Brasil (FAPERGS – 17/2551-0000520-1).

REFERENCES

[1] Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet*, 18: 33–48, 2015.

[2] Kayser, M.; De Knijff, P. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet*, 12 (3): 179–192, 2011.

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

[4] Børsting, C; Morling, N. Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet*, 18: 78–89, 2015.

[5] Bruijns, B.; Tiggelaar, R.; Gardeniers, H. Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, 39(21): 2642–2654, 2018.

[6] Kidd, K. K.; Speed, W.C.; Pakstis, A.J. et al. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet*, 10: 23-32, 2014.

[7] Kosoy, R.; Nassir, R.; Tian, C.; et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat*, 30 (1): 69–78, 2009.

[8] Salzano, F.M.; Sans, M. Interethnic admixture and the evolution of Latin American populations. *Genet Mol Biol*, 37: 151–170, 2014.

[9] Moura, R.R.; Coelho, A.V.; Balbino, V.; Crovella, S.; Brandão, L.A. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am J Hum Biol*, 27(5): 674–680, 2015.

[10] Curtin, P.D. *The Atlantic slave trade: A census*. Madison, WI: University of Wisconsin Press, 1969.

[11] Callegari-Jacques, S.M.; Grattapaglia, D.; Salzano, F.M. et al. Historical genetics: spatiotemporal analysis of the formation of the Brazilian population. *Am J Hum Biol*, 15(6): 824–834, 2003.

[12] IBGE. *Brasil: 500 anos de povoamento*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2007.

[13] Pena, S.D.; Di Pietro, G.; Fuchshuber-Moraes, M. et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One*, 6(2): e17063, 2011.

[14] Marrero, A.R.; Bravi, C.; Stuart, S. et al. Pre- and post-Columbian gene and cultural continuity: the case of the Gaucho from southern Brazil. *Hum Hered*, 64(3): 160–171, 2007.

[15] Gouveia, M.H.; Borda, V.; Leal, T.P. et al. Origins, admixture dynamics and homogenization of the African gene pool in the Americas. *Mol Biol Evol*, 1;37(6): 1647-1656, 2020.

[16] World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, 310(20): 2191–2194, 2013.

[17] Thermo Fisher Scientific. Ion PGM™ Hi-Q™ OT2 Kit. *Revision A.0* (2015). Waltham, MA, USA.

[18] Thermo Fisher Scientific. Ion PGM™ Hi-Q™ Sequencing Kit. *Revision C.0* (2015). Waltham, MA, USA.

[19] Illumina. AmpliSeq for Illumina On-Demand, Custom, and Community Panels. *Document # 1000000036408 v08* (2019). San Diego, CA, USA.

[20] Illumina. MiSeq System: Denature and Dilute Libraries Guide. *Document # 15039740 v10* (2019). San Diego, CA, USA.

[21] Eduardoff M.; Santos C.; de la Puente M. et al. Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™. *Forensic Sci Int Genet*, 17: 110–121, 2015.

[22] Avila E.; Cavalheiro, C.P.; Felkl, A.B. et al. Brazilian forensic casework analysis through MPS applications: Statistical weight-of-evidence and biological nature of criminal samples as an influence factor in quality metrics. *Forensic Sci Int*, 303: 109938, 2019.

[23] Wasik K.; Berisa, T.; Pickrell, J.K. et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomic,* 20;22(1):197, 2021.

[24] Li, N.; Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4): 2213–2233, 2003.

[25] Lischer, H.E.; Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2): 298–299, 2012.

[26] Gouy, A.; Zieger, M. STRAF - A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet*, 30: 148–151, 2017.

[27] Excoffier, L.; Lischer, H.E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 10(3): 564–567, 2010.

[28] Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze.* 8: 3–62, 1936.

[29] 1000 Genomes Project Consortium; Abecasis, G.R.; Auton, A. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422): 56–65, 2012.

[30] García, O.; Ajuriagerra, J.A.; Alday, A. et al. Frequencies of the precision ID ancestry panel markers in Basques using the Ion Torrent PGM™ platform. *Forensic Sci Int Genet*, 31: e1–e4, 2017.

[31] He, G.; Wang, Z.; Wang, M. et al. Forensic ancestry analysis in two Chinese minority populations using massively parallel sequencing of 165 ancestry-informative SNPs. *Electrophoresis*, 39(21): 2732–2742, 2018.

[32] Pereira, V.; Mogensen, H.S.; Børsting, C.; Morling, N. Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. *Forensic Sci Int Genet*, 28: 138–145, 2017.

[33] Pakstis, A.J.; Speed, W.C.; Soundararajan, U. et al. Population relationships based on 170 ancestry SNPs from the combined Kidd and Seldin panels. *Sci Rep*, 9:18874, 2019.

[34] IBM Corp. IBM SPSS Statistics for Windows. Version 25.0, Released 2017. Armonk, NY.

[35] Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959, 2000.

[36] Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 164(4): 1567–1587, 2003.

[37] Kopelman, N.M.; Mayzel, J.; Jakobsson, M.; Rosenberg, N.A.; Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*, 15(5): 1179–1191, 2015.

[38] Earl, D.A.; vonHoldt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour*, 4(2): 359–361, 2011.

[39] Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 14(8): 2611–2620, 2005.

[40] Nakanishi, H.; Pereira, V.; Børsting, C. et al. Analysis of mainland Japanese and Okinawan Japanese populations using the precision ID Ancestry Panel. *Forensic Sci Int Genet*, 33: 106–109, 2018.

[41] Wang, Z.; He, G.; Luo, T. et al. Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities. *Forensic Sci Int Genet*, 34: 141–147, 2018.

[42] Simayijiang, H.; Børsting, C.; Tvedebrink, T.; Morling, N. Analysis of Uyghur and Kazakh populations using the Precision ID Ancestry Panel. *Forensic Sci Int Genet*, 43: 102144, 2019.

[43] Xie, T.; Shen, C.; Liu, C. et al. Ancestry inference and admixture component estimations of Chinese Kazak group based on 165 AIM-SNPs via NGS Platform. *J Hum Genet.* 2020 [Online].

[44] Lee, J.H.; Cho, S.; Kim, M.Y. et al. Genetic resolution of applied biosystems™ precision ID Ancestry panel for seven Asian populations. *Leg Med (Tokyo)*, 34: 41–47, 2018.

[45] Espregueira Themudo, G.; Smidt Mogensen, H.; Børsting, C.; Morling N. Frequencies of HID-ion ampliseq ancestry panel markers among greenlanders. *Forensic Sci Int Genet*, 24: 60–64, 2016.

[46] Santangelo, R.; González-Andrade, F.; Børsting, C.; Torroni, A.; Pereira, V.; Morling, N. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. *Forensic Sci Int Genet*, 31: 29–33, 2017.

[47] Truelsen, D.M.; Farzad, M.S.; Mogensen, H.S. et al. Typing of two Middle Eastern populations with the Precision ID Ancestry Panel. *Forensic Sci Int Genet*, 6: e301–e302, 2017.

[48] IBGE. Sistema IBGE de Recuperação Automática - SIDRA. Tabela 136 - População residente por cor ou raça (2010). Instituto Brasileiro de Geografia e Estatística.

[49] IBGE. Brasil: 500 anos de povoamento (2007). Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística.

[50] Phillips, C.; Ballard, D.; Gill, P.; Court, D.S.; Carracedo, A.; Lareu, M.V. The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Sci Int Genet*, 6(3): 354–365, 2012.

[51] Ardlie, K.G.; Kruglyak, L.; Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3(4): 299–309, 2002.

[52] Saloum de Neves Manta, F.; Pereira, R.; Vianna, R. et al. Revisiting the genetic ancestry of Brazilians using autosomal AIM-Indels. *PLoS One*, 8(9): e75145, 2013.

[53] Parra, F.C.; Amado, R.C.; Lambertucci, J.R.; Rocha, J.; Antunes, C.M.; Pena, S.D. Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci U S A*, 100(1): 177–182, 2003.

[54] Pena, S.D.; Di Pietro, G.; Fuchshuber-Moraes, M. et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One*, 6(2): e17063, 2011.

[55] Muniz, Y.C.; Ferreira, L.B.; Mendes-Junior, C.T.; Wiezel, C.E.; Simões, A.L. Genomic ancestry in urban Afro-Brazilians. *Ann Hum Biol*, 35(1): 104–111, 2008.

[56] Gontijo, C.C.; Mendes, F.M.; Santos, C.A. et al. Ancestry analysis in rural Brazilian populations of African descent. *Forensic Sci Int Genet*, 36: 160–166, 2018.

[57] Avila, E.; Felkl, A.B.; Graebin, P.; Nunes, C.P.; Alho, C.S. Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel. *Forensic Sci Int Genet*, 40: 74–84, 2019.

[58] IBGE. Sistema IBGE de Recuperação Automática - SIDRA. Tabela 136 - População residente por cor ou raça (2010). Instituto Brasileiro de Geografia e Estatística.

[59] Pimenta, J.R.; Zuccherato, L.W.; Debes, A.A. et al. Color and genomic ancestry in Brazilians: a study with forensic microsatellites. *Hum Hered*, 62(4): 190–195, 2006.

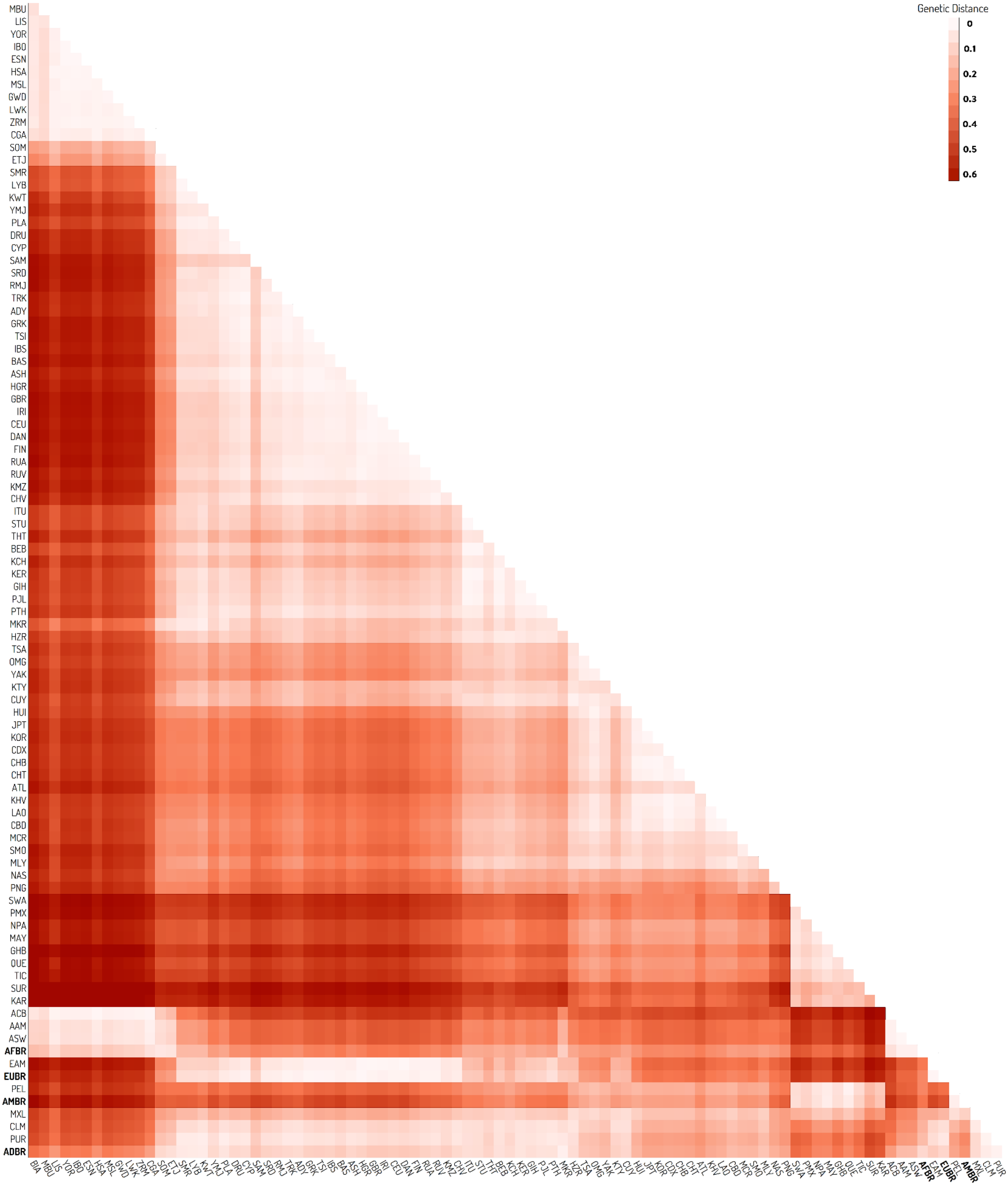
[60] Moriot, A.; Santos, C.; Freire-Aradas, A.; Phillips, C.; Hall, D. Inferring biogeographic ancestry with compound markers of slow and fast evolving polymorphisms. *Eur J Hum Genet*, 26(11): 1697–1707, 2018.

[61] Pakstis, A.J.; Gurkan, C.; Dogan, M. et al. Genetic relationships of European, Mediterranean, and SW Asian populations using a panel of 55 AISNPs. *Eur J Hum Genet*, 27(12): 1885–1893, 2019.

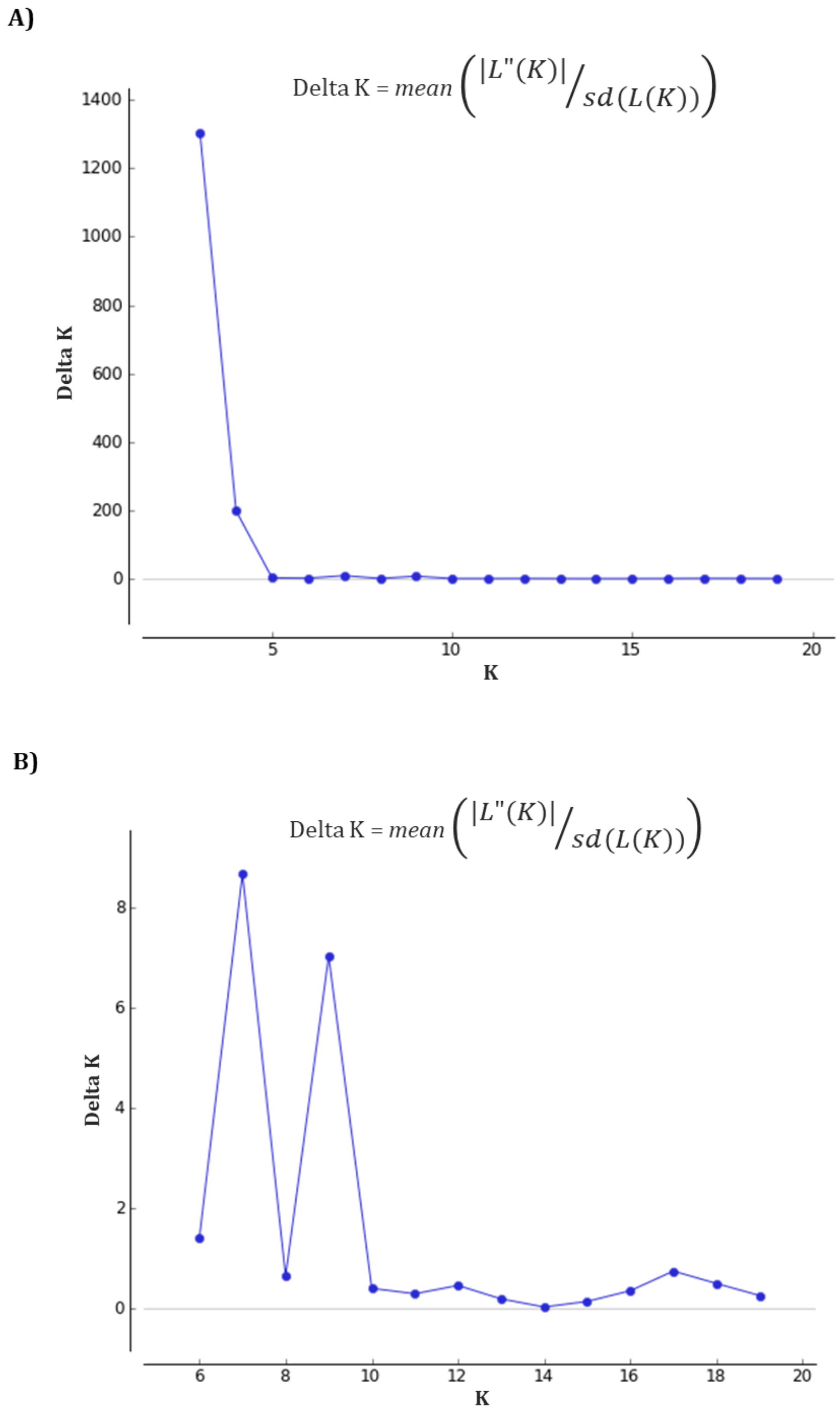
SUPPLEMENTARY MATERIAL

Ancestry resolution of South Brazilians by forensic 165 ancestry-informative SNPs panel

CONTENTS:



Supplementary Fig. S1. Heatmap of pairwise F_{ST} values for 165 ancestry-informative SNPs included in Precision ID Ancestry Panel among RS State (Brazil) main population groups and 89 worldwide populations. Populations' details are presented in Supplementary Table S1.



Supplementary Fig. S2. Evanno's Delta (Δ) K statistics [37] by STRUCTURE HARVESTER software. Results derived from ten independent simulations for each K value on STRUCTURE, based on the full data set of 5,000 individuals from 93 worldwide populations. The modal value of ΔK distribution is selected as the most likely number of K. **(A)** ΔK with K values ranging from 2 to 20. **(B)** ΔK with K values ranging from 5 to 20.

Supplementary Table S1. The 93 populations organized by geographic location.

World Region	Code	Population	Sample size
CENTRAL AFRICA	1 BIA	Biaka, Central African Rep.	69
	2 MBU	Mbuti, Ituri Forest, D.R. Congo	39
	3 LIS	Lisongo	8
WESTERN AFRICA	4 YOR	Yoruba, Benin City, Nigeria	70
	5 IBO	Ibo, Nigeria	48
	6 ESN[a]	Esan in Nigeria	70
	7 HSA	Hausa, Nigeria	39
	8 MSL[a]	Mende in Sierra Leone	70
	9 GWD[a]	Gambian in Western Divisions in the Gambia	70
EASTERN AFRICA	10 LWK[a]	Luhya in Webuye, Kenya	70
	11 ZRM	Zaramo, Tanzania	40
	12 CGA	Chagga, Tanzania	45
	13 SOM[d]	Somali	70
	14 ETJ	Ethiopian Jews	32
NORTHERN AFRICA & MIDDLE EASTERN	15 SMR	Smar, Tunisia	61
	16 LYB	Libyans; from 6 locations in Libya	69
	17 KWT	Kuwaiti	14
	18 YMJ	Yemenite Jews	42
	19 PLA	Palestinian Arabs	68
	20 DRU	Druze, Israel	70
	21 CYP	Turkish Cypriots	59
	22 SAM	Samaritans, Israel	39
SOUTHERN EUROPE	23 SRD	Sardinians	34
	24 RMJ	Roman Jews, Italy	27
	25 TRK	Turkish, Istanbul, Turkey	70
	26 ADY	Adygei	54
	27 GRK	Greeks, Thesaloniki, Greece	52
	28 TSI[a]	Toscani, Italy	70
	29 IBS[a]	Iberian Population in Spain	70
	30 BAS[b]	Basques	70
CENTRAL & NORTHERN EUROPE	31 ASH	Ashkenazi Jews	70
	32 HGR	Hungarians	70
	33 GBR[a]	British in England and Scotland	70
	34 IRI	Irish	70
	35 CEU[a]	Utah Residents (CEPH) with Northern and Western European Ancestry	70
	36 DAN[d]	Danes	70
	37 FIN[a]	Finnish in Finland	70
	38 RUA	Russians, Archangelsk	33
	39 RUV	Russians, Vologda	47
	40 KMZ	Komi Zyriane	47
	41 CHV	Chuvash	42
SOUTHERN ASIA	42 ITU[a]	Indian Telugu from the UK	70
	43 STU[a]	Sri Lankan Tamil from the UK	70
	44 THT	Thoti, India	14
	45 BEB[a]	Bengali in Bangladesh	70
	46 KCH	Kachari, Assam State, India	17

	47 KER	Keralites, India	30
	48 GIH[a]	Gujarati Indian from Houston, Texas	70
	49 PJL[a]	Punjabi from Lahore, Pakistan	70
	50 PTH	Pathans, Pakistan	70
	51 MKR	Negroid Makrani, Pakistan	26
	52 HZR	Hazara, Pakistan	70
CENTRAL & NORTHERN ASIA	53 TSA	Tsaatan	51
	54 OMG	Outer Mongolians	70
	55 YAK	Yakut	51
	56 KTY	Khanty	50
EASTERN ASIA	57 CUY[c]	Chinese Uyghur	47
	58 HUI[c]	Chinese Hui	70
	59 JPT[a]	Japanese in Tokyo, Japan	70
	60 KOR	Koreans	54
	61 CDX[a]	Chinese Dai in Xishuangbanna, China	65
	62 CHB[a]	Han Chinese in Beijing, China	70
	63 CHT	Chinese, Taiwan	50
	64 ATL	Atayal, Taiwan	42
SOUTHEAST ASIA & PACIFIC ISLANDS	65 KHV[a]	Kinh in Ho Chi Minh City, Vietnam	70
	66 LAO	Laotians	70
	67 CBD	Cambodians	24
	68 MCR	Micronesians	34
	69 SMO	Samoans	9
	70 MLY	Malaysians	11
	71 NAS	Nasioi, Bougainville, Solomon Islands	23
	72 PNG	Papuans, New Guinea	22
NATIVE AMERICA	73 SWA	Southwest Amerindians	51
	74 PMX	Pima, Northern Mexico	53
	75 NPA	Plains Amerindians	56
	76 MAY	Maya, Yucatan, Mexico	50
	77 GHB	Guihiba speakers, Colombia	12
	78 QUE	Quechua, Peru	22
	79 TIC	Ticuna, Amazon region, Brazil	65
	80 SUR	Rondonian Surui, Amazon region, Brazil	43
	81 KAR	Karitiana, Amazon region, Brazil	55
ADMIXED AMERICA	82 ACB[a]	African Caribbeans in Barbados	70
	83 AAM	African Americans	70
	84 ASW[a]	Americans of African Ancestry in SW USA	61
	85 AFRS	**African-derived Gauchos**	**62**
	86 EAM	European Americans	70
	87 EURS	**European-derived Gauchos**	**92**
	88 PEL[a]	Peruvians from Lima, Peru	70
	89 AMRS	**Amerindian-derived Gauchos**	**22**
	90 MXL[a]	Mexican Ancestry from Los Angeles USA	64
	91 CLM[a]	Colombians from Medellin, Colombia	70
	92 PUR[a]	Puerto Ricans from Puerto Rico	70
	93 ADRS	**Admixed-derived Gauchos**	**74**
			5,000

[a]Populations from 1000 Genomes Project Phase III. [b][27]. [c][28]. [d][29].

Unformatted populations were provided by Kidd Lab [30].

In bold, Rio Grande do Sul (Brazil) population groups analyzed in this study.

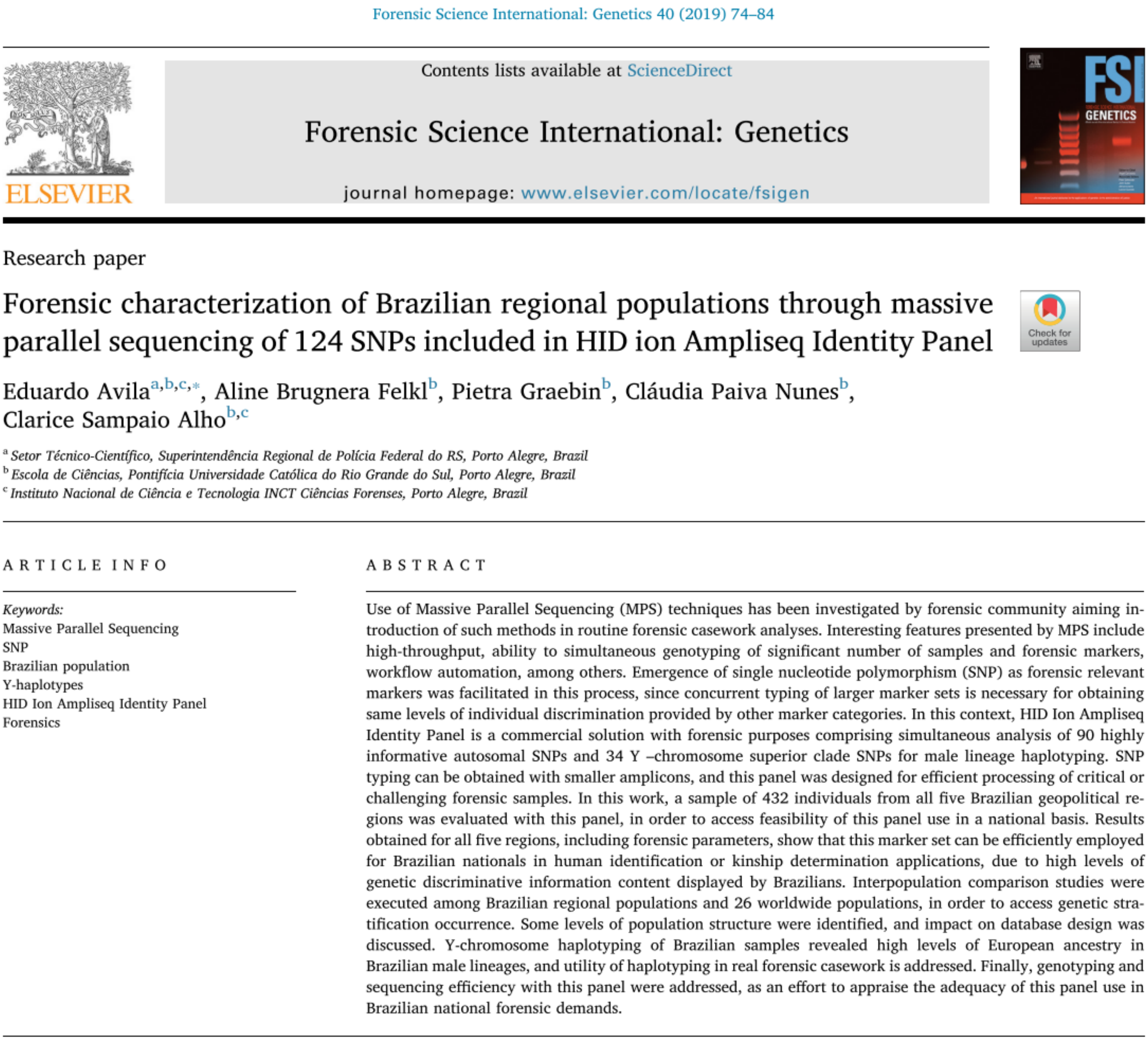
ANEXOS

Anexo A – Artigo “Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel”.

Anexo B – Artigo “Brazilian forensic casework analysis through MPS applications: Statistical weight-of-evidence and biological nature of criminal samples as an influence factor in quality metrics”.

Anexo C – Parecer Consubstanciado do CEP.

ANEXO A



Forensic Science International: Genetics 40 (2019) 74–84

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Forensic characterization of Brazilian regional populations through massive parallel sequencing of 124 SNPs included in HID ion Ampliseq Identity Panel

Eduardo Avila[a,b,c,*], Aline Brugnera Felkl[b], Pietra Graebin[b], Cláudia Paiva Nunes[b], Clarice Sampaio Alho[b,c]

[a] *Setor Técnico-Científico, Superintendência Regional de Polícia Federal do RS, Porto Alegre, Brazil*

[b] *Escola de Ciências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil*

[c] *Instituto Nacional de Ciência e Tecnologia INCT Ciências Forenses, Porto Alegre, Brazil*

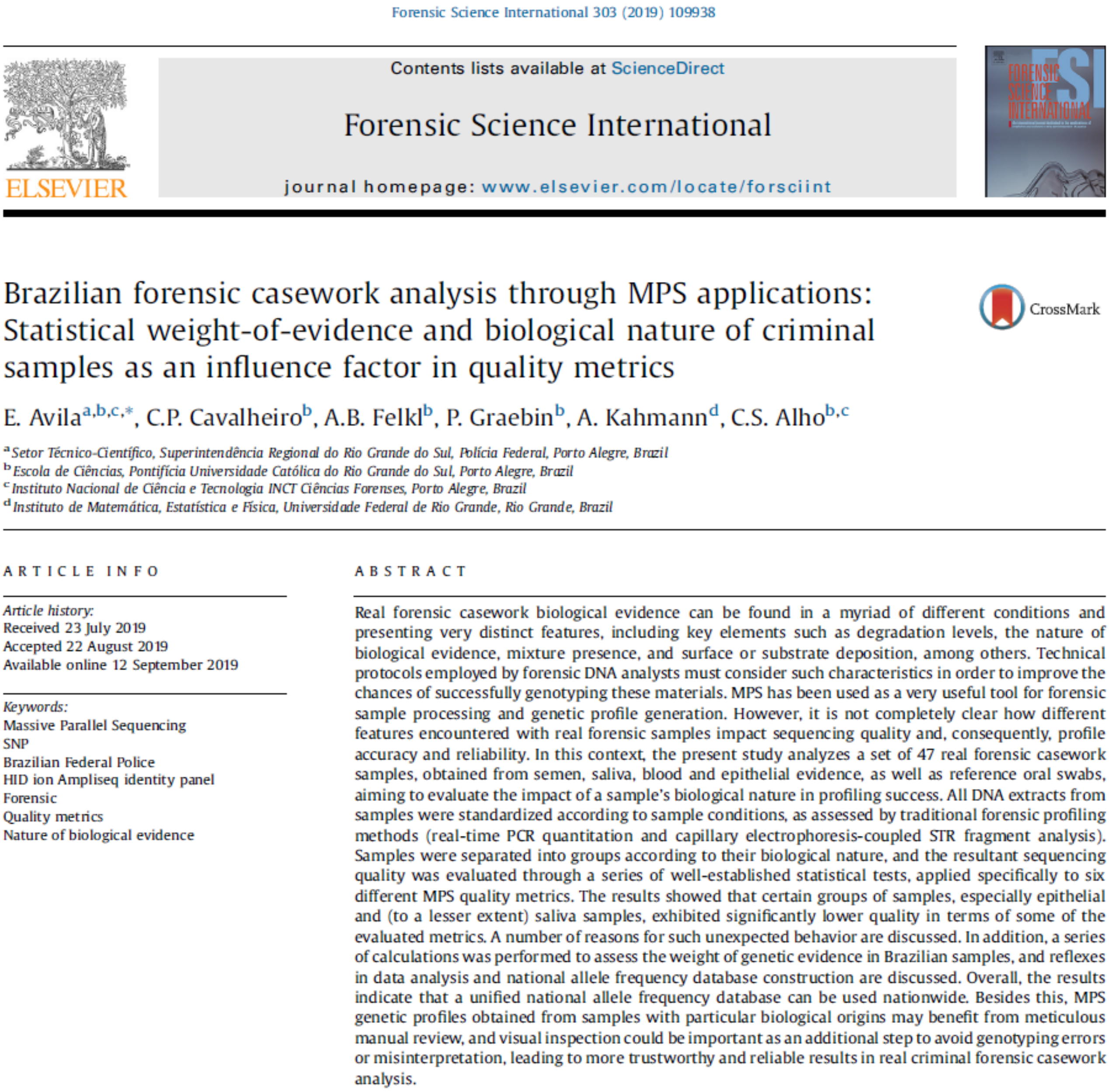
ARTICLE INFO

Keywords:
Massive Parallel Sequencing
SNP
Brazilian population
Y-haplotypes
HID Ion Ampliseq Identity Panel
Forensics

ABSTRACT

Use of Massive Parallel Sequencing (MPS) techniques has been investigated by forensic community aiming introduction of such methods in routine forensic casework analyses. Interesting features presented by MPS include high-throughput, ability to simultaneous genotyping of significant number of samples and forensic markers, workflow automation, among others. Emergence of single nucleotide polymorphism (SNP) as forensic relevant markers was facilitated in this process, since concurrent typing of larger marker sets is necessary for obtaining same levels of individual discrimination provided by other marker categories. In this context, HID Ion Ampliseq Identity Panel is a commercial solution with forensic purposes comprising simultaneous analysis of 90 highly informative autosomal SNPs and 34 Y –chromosome superior clade SNPs for male lineage haplotyping. SNP typing can be obtained with smaller amplicons, and this panel was designed for efficient processing of critical or challenging forensic samples. In this work, a sample of 432 individuals from all five Brazilian geopolitical regions was evaluated with this panel, in order to access feasibility of this panel use in a national basis. Results obtained for all five regions, including forensic parameters, show that this marker set can be efficiently employed for Brazilian nationals in human identification or kinship determination applications, due to high levels of genetic discriminative information content displayed by Brazilians. Interpopulation comparison studies were executed among Brazilian regional populations and 26 worldwide populations, in order to access genetic stratification occurrence. Some levels of population structure were identified, and impact on database design was discussed. Y-chromosome haplotyping of Brazilian samples revealed high levels of European ancestry in Brazilian male lineages, and utility of haplotyping in real forensic casework is addressed. Finally, genotyping and sequencing efficiency with this panel were addressed, as an effort to appraise the adequacy of this panel use in Brazilian national forensic demands.

ANEXO B



Forensic Science International 303 (2019) 109938

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Brazilian forensic casework analysis through MPS applications: Statistical weight-of-evidence and biological nature of criminal samples as an influence factor in quality metrics

E. Avila[a,b,c,*], C.P. Cavalheiro[b], A.B. Felkl[b], P. Graebin[b], A. Kahmann[d], C.S. Alho[b,c]

[a] Setor Técnico-Científico, Superintendência Regional do Rio Grande do Sul, Polícia Federal, Porto Alegre, Brazil
[b] Escola de Ciências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
[c] Instituto Nacional de Ciência e Tecnologia INCT Ciências Forenses, Porto Alegre, Brazil
[d] Instituto de Matemática, Estatística e Física, Universidade Federal de Rio Grande, Rio Grande, Brazil

ARTICLE INFO

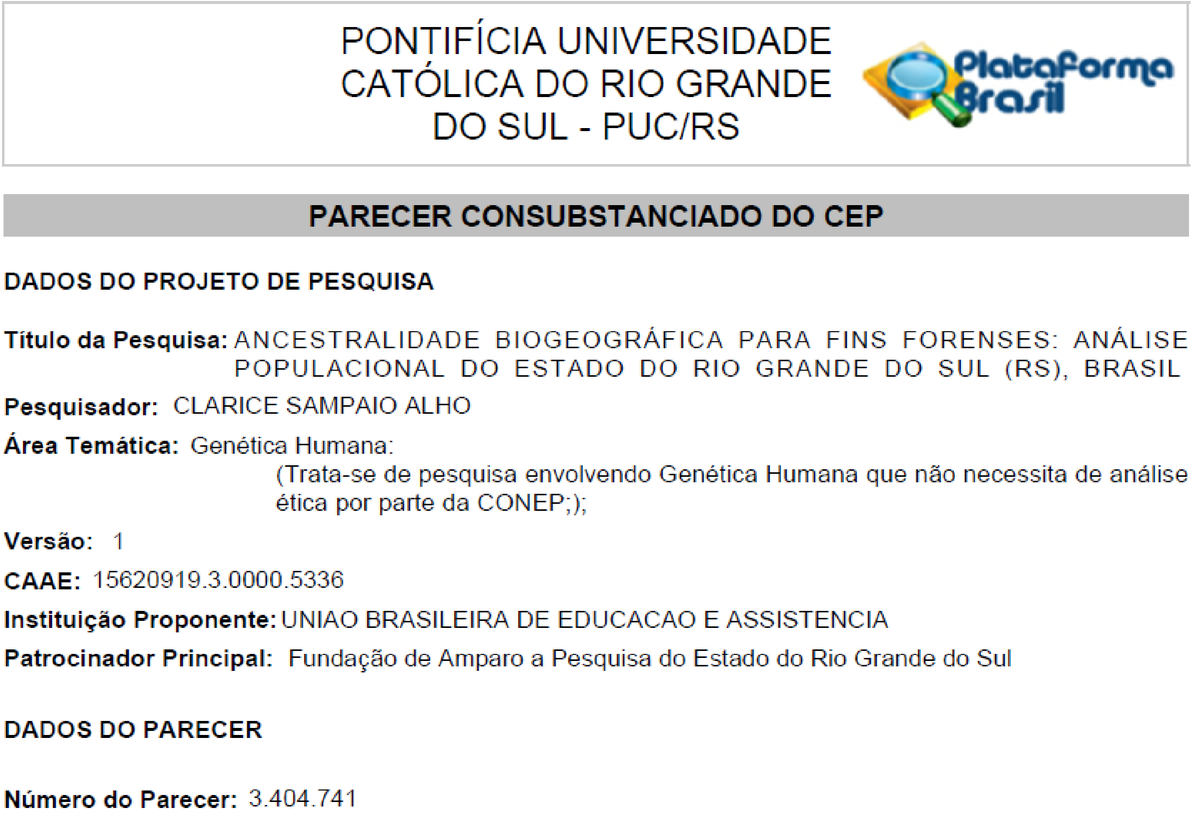
Article history:
Received 23 July 2019
Accepted 22 August 2019
Available online 12 September 2019

Keywords:
Massive Parallel Sequencing
SNP
Brazilian Federal Police
HID ion Ampliseq identity panel
Forensic
Quality metrics
Nature of biological evidence

ABSTRACT

Real forensic casework biological evidence can be found in a myriad of different conditions and presenting very distinct features, including key elements such as degradation levels, the nature of biological evidence, mixture presence, and surface or substrate deposition, among others. Technical protocols employed by forensic DNA analysts must consider such characteristics in order to improve the chances of successfully genotyping these materials. MPS has been used as a very useful tool for forensic sample processing and genetic profile generation. However, it is not completely clear how different features encountered with real forensic samples impact sequencing quality and, consequently, profile accuracy and reliability. In this context, the present study analyzes a set of 47 real forensic casework samples, obtained from semen, saliva, blood and epithelial evidence, as well as reference oral swabs, aiming to evaluate the impact of a sample's biological nature in profiling success. All DNA extracts from samples were standardized according to sample conditions, as assessed by traditional forensic profiling methods (real-time PCR quantitation and capillary electrophoresis-coupled STR fragment analysis). Samples were separated into groups according to their biological nature, and the resultant sequencing quality was evaluated through a series of well-established statistical tests, applied specifically to six different MPS quality metrics. The results showed that certain groups of samples, especially epithelial and (to a lesser extent) saliva samples, exhibited significantly lower quality in terms of some of the evaluated metrics. A number of reasons for such unexpected behavior are discussed. In addition, a series of calculations was performed to assess the weight of genetic evidence in Brazilian samples, and reflexes in data analysis and national allele frequency database construction are discussed. Overall, the results indicate that a unified national allele frequency database can be used nationwide. Besides this, MPS genetic profiles obtained from samples with particular biological origins may benefit from meticulous manual review, and visual inspection could be important as an additional step to avoid genotyping errors or misinterpretation, leading to more trustworthy and reliable results in real criminal forensic casework analysis.

ANEXO C



PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL - PUC/RS

Plataforma Brasil

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: ANCESTRALIDADE BIOGEOGRÁFICA PARA FINS FORENSES: ANÁLISE POPULACIONAL DO ESTADO DO RIO GRANDE DO SUL (RS), BRASIL

Pesquisador: CLARICE SAMPAIO ALHO

Área Temática: Genética Humana:
(Trata-se de pesquisa envolvendo Genética Humana que não necessita de análise ética por parte da CONEP;);

Versão: 1

CAAE: 15620919.3.0000.5336

Instituição Proponente: UNIAO BRASILEIRA DE EDUCACAO E ASSISTENCIA

Patrocinador Principal: Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul

DADOS DO PARECER

Número do Parecer: 3.404.741

MARISTA PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br