

**Construção e validação computacional de um conjunto de
marcadores multi-InDels autossômicos para identificação
humana**

LIVIA CARLA RAMOS

Botucatu

2025

LIVIA CARLA RAMOS

**Construção e validação computacional de um conjunto de
marcadores multi-InDels autossômicos para identificação
humana**

Orientador: Prof. Dr. Erick da Cruz Castelli

Tese apresentada ao Instituto de Biotecnologia,
Universidade Estadual Paulista "Júlio de
Mesquita Filho", Campus de Botucatu, para
obtenção do título de Doutora pelo Programa
de Pós-Graduação em Ciências Biológicas
(Genética).

Botucatu

2025

R175c Ramos, Livia Carla
Construção e validação computacional de um conjunto de marcadores multi-InDels autossômicos para identificação humana / Livia Carla Ramos. -- , 2025
114 f.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu,
Orientador: Erick da Cruz Castelli

1. Genética humana. 2. Biologia forense. 3. Marcadores genéticos. I. Título.

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE LIVIA CARLA RAMOS, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA), DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 18 dias do mês de setembro do ano de 2025, às 13h, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de LIVIA CARLA RAMOS, intitulada **Construção e validação computacional de um conjunto de marcadores multi-InDels autossômicos para identificação humana**, sob orientação do Prof. Dr. Erick da Cruz Castelli. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. ERICK DA CRUZ CASTELLI (Orientador(a) - Participação Virtual) do(a) Departamento de Patologia / Faculdade de Medicina de Botucatu FMB – UNESP, Prof. Dr. CELSO TEIXEIRA MENDES JÚNIOR (Participação Virtual) do(a) Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo (USP), Prof. Dr. MICHEL SATYA NASLAVSKY (Participação Virtual) do(a) Departamento de Genética e Biologia Evolutiva / Instituto de Biociências da Universidade de São Paulo (USP), Prof. Dr. AGUINALDO LUIZ SIMÕES (Participação Virtual) do(a) Departamento de Genética / Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo (USP), Dr^a. MARIANNA MAIA TAULOIS DO ROSÁRIO (Participação Virtual) do(a) Laboratório de Genética Forense / Polícia Científica do Estado Paraná. Após a exposição pela doutoranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final: aprovada . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. ERICK DA CRUZ CASTELLI

Documento assinado digitalmente
 ERICK DA CRUZ CASTELLI
Data: 25/09/2025 07:42:58-0300
Verifique em <https://validar.iti.gov.br>

Dedicatória

Dedico este trabalho à minha mãe, Rita de Cássia, por todo o apoio e por ser meu alicerce em todos os momentos. A ela, que nunca hesitou em me ajudar, incentivar e em cuidar da minha filha para que eu pudesse concluir este trabalho.

Agradecimentos

Agradeço primeiramente a Deus, pela força diária e constante, e por nunca permitir que eu desistisse. Houve dias difíceis, mas, mesmo em meio a tudo isso, sempre mantive a certeza de que conseguiria. Nunca deixei de acreditar na minha capacidade, embora, conciliar os papéis de mãe, esposa, perita criminal e doutoranda, eu duvidei muitas vezes. A caminhada foi longa e árdua.

Sou imensamente grata aos meus pais, familiares e amigos, em especial à minha mãe, Rita de Cássia, ao meu marido, Kleber, e à minha filha, Liz. Vocês estiveram ao meu lado em todos os momentos, com todo amor e apoio.

Agradeço ao meu orientador, Prof. Dr. Erick da Cruz Castelli, por acreditar no meu potencial, mesmo diante das minhas limitações de tempo e compromissos profissionais. Obrigada pelo conhecimento compartilhado e por ter me ajudado a concluir esse ciclo.

Sou grata também ao grupo forense, liderado pelo meu orientador e pelo Prof. Dr. Celso Teixeira Mendes Júnior, e formado por Luiza Rauen Sabbag, Luciellen D. G. Kobachuk, Marianna Maia Taulois do Rosário e João Victor Hamamura Toloí. Obrigada por todo o apoio, troca de saberes e por contribuírem diretamente para a concretização deste projeto.

Às queridas colegas do Laboratório GeMBio, Gabriela Sato Paes, Nayane dos Santos Brito Silva, Viviane Aparecida de Oliveira Ciriaco e Heloísa de Souza Andrade, meu sincero agradecimento pela parceria, pela ajuda constante e por todos os momentos compartilhados ao longo desses anos.

Agradeço ainda à banca examinadora, por dedicar seu tempo, atenção e contribuições valiosas ao projeto.

Meu reconhecimento também à Universidade Estadual Paulista (UNESP), Câmpus de Botucatu/SP, pela oportunidade de crescimento acadêmico e pessoal.

E, por fim, agradeço a mim mesma. Pela coragem, pela persistência, e por não ter desistido de realizar este grande sonho, mesmo diante de tantos obstáculos.

"Tudo parece impossível até que seja feito."

Nelson Mandela

Resumo

Os marcadores genéticos STRs (do inglês, *short tandem repeats*) são considerados padrão-ouro para as análises forenses. Apesar de amplamente utilizados, os STRs apresentam limitações relevantes especialmente em casos complexos, tais como amostras degradadas ou compostas por múltiplos perfis genéticos. Dessa forma, outros marcadores têm despertado interesse na genética forense, com o objetivo de serem complementares aos STRs e superar suas limitações. Os polimorfismos de inserção/deleção, ou InDels, estão amplamente presentes no genoma humano e reúnem algumas características importantes e vantajosas. Eles podem ser amplificados por reação em cadeia da polimerase (PCR) e detectados por eletroforese capilar (CE), tecnologias amplamente presentes nos laboratórios forenses, e podem ser organizados em sistema multiplex. Além de apresentarem baixa taxa de mutação, favorável para análises de vínculo genético e da possibilidade de gerar produtos de amplificação (*amplicons*) menores, facilitando a análise de amostras degradadas. Ademais, os InDels não produzem artefatos de amplificação denominados *stutters*, sendo favorável para casos contendo misturas de perfis genéticos de dois ou mais indivíduos. Entretanto, a maioria dos InDels são bi-alélicos, o que reduz seu poder de discriminação. Para contornar essa limitação, novos marcadores genéticos, como os multi-InDels, têm ganhado destaque, pois, preservam as características individuais dos InDels e ainda proporcionam um aumento do polimorfismo. Os marcadores multi-InDels são classificados de forma generalizada como microhaplótipos, constituídos por dois ou mais *loci* InDels, localizados em um segmento curto de DNA. Atualmente, há painéis de multi-InDels propostos na literatura, todos validados apenas em populações asiáticas. O objetivo deste trabalho foi construir e validar, por meio de abordagens computacionais, um novo conjunto de marcadores multi-InDels autossômicos, que seja capaz identificar os indivíduos da população brasileira e de outras populações do mundo. Previamente, nós avaliamos três conjuntos anteriormente propostos e validados em populações asiáticas, com o objetivo de verificar a sua aplicabilidade em indivíduos brasileiros. Foram utilizados dados públicos de sequenciamento de alta cobertura de genoma completo do projeto *1000 Genomes*, que compreende 26 populações mundiais, e a amostras de brasileiros (Projeto Saúde, Bem-estar e Envelhecimento - SABE), tanto para avaliar os painéis existentes quanto para selecionar novos marcadores. Concluiu-se que a maioria dos marcadores anteriormente propostos é informativa para indivíduos brasileiros, uma vez que apresentam mais de três

haplótipos frequentes com diferentes tamanhos. Entretanto, 64,8% destes marcadores são propensos a erros de amplificação e/ou sequenciamento, devido à presença de regiões repetitivas, de baixa complexidade, homopolimérica ou STRs. Desse modo, neste trabalho, selecionamos e propusemos um novo conjunto de marcadores multi-InDels para identificação humana, organizado em sistema multiplex e baseado na plataforma de CE, composto por 28 marcadores, além de dois marcadores para determinação do sexo. O painel apresentou alto polimorfismo em cinco superpopulações, bem como em indivíduos brasileiros. A probabilidade de *match* combinada (PMC) variou de $5,66 \times 10^{-20}$ a $2,66 \times 10^{-17}$, enquanto o poder de exclusão combinado (PEC) variou de 0,99958 a 0,99998. Além disso, 94,4% dos *amplicons* gerados foram menores que 250 pares de base (pb). Portanto, o painel proposto pode ser uma ferramenta complementar útil para a identificação humana, aplicável para população brasileira e em outras populações ao redor do mundo, especialmente em casos complexos.

Palavras-chave: *genética forense, microhaplótipo, multi-InDel, população brasileira.*

Abstract

Short tandem repeat (STR) markers are considered the gold standard in forensic genetics. Although widely used, STRs present relevant limitations, especially in complex cases involving degraded samples or mixtures of multiple genetic profiles. Therefore, alternative genetic markers have gained interest in forensic applications, aiming to complement STRs and overcome their limitations. Insertion/deletion polymorphisms (InDels) are widely distributed throughout the human genome and exhibit several advantageous characteristics. They can be amplified by polymerase chain reaction (PCR) and detected by capillary electrophoresis (CE), both techniques routinely available in forensic laboratories, and can be organized into multiplex systems. InDels also display low mutation rates, which is favorable for kinship analyses, and can generate short amplification products (amplicons), facilitating the analysis of degraded samples. Moreover, InDels do not produce stutter artifacts, which is advantageous in the analysis of mixtures containing DNA from two or more contributors. However, most InDels are biallelic, which limits their discrimination power. To overcome this limitation, new genetic markers such as multi-InDels have gained prominence, as they preserve the individual characteristics of InDels while providing increased polymorphism. Multi-InDels are generally classified as microhaplotypes, consisting of two or more closely linked InDel loci within a short DNA segment. Currently, all multi-InDel panels described in the literature have been validated only in Asian populations. The aim of this study was to develop and validate, through computational approaches, a novel set of autosomal multi-InDel markers capable of distinguishing individuals from the Brazilian population as well as from other global populations. Previously, we evaluated three panels proposed and validated in Asian populations to assess their applicability to individuals from Brazil. Public high-coverage whole-genome sequencing data from the 1000 Genomes Project, which encompasses 26 worldwide populations, and from Brazilian individuals (Projeto Saúde, Bem-estar e Envelhecimento - SABE) were utilized to both evaluate existing panels and select new markers. Our results showed that most previously proposed markers were informative for the Brazilian population, as they presented more than three frequent haplotypes of different lengths. However, 64.8% of these markers were prone to amplification and/or sequencing

errors due to the presence of repetitive, low-complexity, homopolymeric, or STR-like regions. Thus, in this study, we propose a new multiplex CE-based panel composed of 28 autosomal multi-InDel markers and two sex-determination markers. The panel exhibited high polymorphism across five superpopulations, as well as in Brazilian individuals. The combined match probability (CMP) ranged from 5.66×10^{-20} to 2.66×10^{-17} , while the combined power of exclusion (CPE) ranged from 0.99958 to 0.99998. Additionally, 94.4% of the generated amplicons were shorter than 250 base pairs (bp). Therefore, the proposed panel represents a useful complementary tool for human identification, suitable for the Brazilian population and globally, particularly in complex forensic cases.

Keywords: forensic genetics, microhaplotype, multi-InDel, Brazilian population.

Sumário

1. CAPÍTULO I	14
1.1. REVISÃO DE LITERATURA	15
1.1.1. Genética forense.....	15
1.1.2. Amostras complexas e mistura de perfis de DNA	18
1.1.3. Marcadores genéticos para fins forenses	20
1.1.4. Microhaplótipos	27
1.1.5. População Brasileira	32
1.2. JUSTIFICATIVA	34
1.3. OBJETIVOS	36
1.4. REFERÊNCIAS	37
2. CAPÍTULO II – MANUSCRITO I.....	46
2.1. INTRODUÇÃO.....	48
2.2. MATERIAL E MÉTODOS	49
2.3. RESULTADOS E DISCUSSÕES.....	51
2.4. CONCLUSÃO	54
3. CAPÍTULO III – MANUSCRITO II	60
3.1. <i>ABSTRACT</i>	61
3.2. <i>INTRODUCTION</i>	62
3.3. <i>MATERIALS AND METHODS</i>	63
3.4. <i>RESULTS</i>	66
3.4.1. <i>The final set of multi-InDel markers</i>	66
3.4.2. <i>Allele frequencies</i>	68
3.4.3. <i>Forensic parameters and statistical analysis</i>	70
3.4.4. <i>Hardy-Weinberg equilibrium (HWE) and Linkage disequilibrium (LD)</i>	73
3.4.5. <i>Multiplex Construction</i>	74
3.5. <i>DISCUSSION</i>	76
3.6. <i>CONCLUSION</i>	80
3.7. <i>REFERENCE</i>	81
3.8. <i>APPENDIX A. SUPPLEMENTARY DATA</i>	87
4. CONCLUSÃO.....	104
5. PERSPECTIVAS FUTURAS	106
6. APÊNDICE	107

Lista de Figuras

Capítulo I

Figura 1. Imagem ilustrativa demonstrando o marcador multi-InDel constituído por dois InDels (rs200250620 e rs10591193) (Huang et al., 2014), que formam três haplótipos com tamanhos distintos. 28

Capítulo III

Figure 1. Heatmap of allelic frequencies of 28 multi-InDel markers in Brazil (SABE), Africa (AFR), America without Brazil (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Colors from white to blue represent allele frequencies. The size of each haplotype is on the y-axis. 69

Figure 2. Violin plots of forensic parameters for the 28 multi-InDel markers in different groups. Brazilians (SABE), Africans (AFR), Americans – without Brazil (AMR), East Asians (EAS), Europeans (EUR), and South Asians (SAS). Ae: effective number of alleles; He: expected heterozygosity; Ho: observed heterozygosity; PD: power of discrimination; PE: power of exclusion; TPI: typical paternity index; PIC: polymorphism information content; MP: match probability. 71

Figure 3 Visualization of HWE probabilities of the 28 multi-InDel markers in 26 populations grouped into five superpopulation groups. Brazilian population (SABE), African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The dashed line indicates the corrected significance level after Bonferroni correction for multiple tests ($\alpha_{\text{BONFERRONI}} = 0.00178$). The individual populations are listed in Table 1. 74

Figure 4 Correlation of the 30 markers with their chromosomes and dyes. 75

Figure 5 Distribution of the 30 markers in the multiplex assay. Each box represents the size range of the marker. The sizes are indicated in the x-axis, and the dye in the y-axis. 75

Lista de Tabelas

Capítulo II

Tabela 1. Lista dos marcadores multi-InDel avaliados em uma amostra da população brasileira..... 50

Tabela 2. Probabilidade de *match* combinada (PMC) e o poder de exclusão combinado (PEC) de diferentes conjuntos de marcadores em diferentes populações, incluindo a coorte brasileira SABE.. 52

Capítulo III

Table 1. Populations of the 1000 Genome project and their sample sizes..... 65

Table 2. General information of the 28 multi-InDel markers. 67

Table 3. Comparison of the mean values of forensic parameters across different groups using 28 new multi-InDel markers..... 71

Table 4. Combined match probability (CMP) and cumulative probability of exclusion (CPE) results for different groups using 28 new multi-InDel markers. 72

Table 5. Combined match probability (CMP) and cumulative probability of exclusion (CPE) results for individual populations using 28 new multi-InDel markers. 72

Table 6. Comparison of combined match probability (CMP) and cumulative probability of exclusion (CPE) for different multi-InDels sets..... 78

Supplementary Table S 1. List of allelic frequencies of 28 multi-InDel markers in different groups. 87

Supplementary Table S 2. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the Brazil from SABE cohort..... 92

Supplementary Table S 3. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the African group from the 1000 Genomes Project. 93

Supplementary Table S 4. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the American group from the 1000 Genomes Project..... 94

Supplementary Table S 5. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the East Asian group from the 1000 Genomes Project. 95

Supplementary Table S 6. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the European group from the 1000 Genomes Project. 96

Supplementary Table S 7. Estimated values of A_e , H_e , H_o , PD, PE, TPI, PIC, and MP for the 28 multi-InDel markers in the South Asian group from the 1000 Genomes Project.....	97
Supplementary Table S 8. Estimated values of HWE probabilities for the 28 multi-InDel markers in the Brazil from SABE cohort.....	98
Supplementary Table S 9. Estimated values of HWE probabilities for the 28 multi-InDel markers in the African group from the 1000 Genomes Project (ESN, YRI, GWD, LWK, and MSL populations).	99
Supplementary Table S 10. Estimated values of HWE probabilities for the 28 multi-InDel markers in the American group from the 1000 Genomes Project (MXL, ACB, PUR, CLM, PEL, and ASW populations).	100
Supplementary Table S 11. Estimated values of HWE probabilities for the 28 multi-InDel markers in the East Asian group from the 1000 Genomes Project (CHS, KHV, CDX, CHB, JPT populations).101	
Supplementary Table S 12. Estimated values of HWE probabilities for the 28 multi-InDel markers in the European group from the 1000 Genomes Project (CEU, TSI, GBR, FIN, IBS populations).	102
Supplementary Table S 13. Estimated values of HWE probabilities for the 28 multi-InDel markers in the South Asian group from the 1000 Genomes Project (GIH, BEB, PJJ, STU, ITU populations). 103	

1. Capítulo I

1.1. Revisão de Literatura

1.1.1. Genética forense

A genética forense emergiu no século passado e se desenvolveu de forma gradual ao longo do tempo. Teve início com a descoberta do sistema de grupos sanguíneos ABO por Karl Landsteiner, em 1900, que marcou o início da identificação humana com base em características biológicas e estabeleceu um alicerce importante para a área forense. Contribuições posteriores, como a teoria cromossômica da herança de Thomas H. Morgan, a elucidação da estrutura em dupla-hélice do DNA por James Watson, Francis Crick e Rosalind Franklin, juntamente com a evolução das técnicas de biologia molecular, ampliaram de maneira significativa o conhecimento acerca dos marcadores genéticos, que atualmente são amplamente aplicados no contexto forense (Scheid et al., 2005; Li, 2018; Dias Filho et al., 2020).

O desenvolvimento da genética forense foi marcado pelo uso de diversos tipos de marcadores biológicos, como os grupos sanguíneos (sistema ABO, fator Rh e fatores MNS), HLA (Antígeno Leucocitário Humano), proteínas séricas e marcadores izoenzimáticos eritrocitários (Gutiérrez-Hurtado et al., 2025). A análise de marcadores biológicos em nível molecular teve início nas décadas de 1970 e 1980. O pesquisador Frederick Sanger desenvolveu técnica de sequenciamento de DNA (Sanger et al., 1977) que veio a viabilizar a análise do genoma humano. Em 1984, Alec Jeffreys utilizou-se da técnica de polimorfismo no comprimento de fragmentos de restrição (RFLP, do inglês *restriction fragment length polymorphism*), para avaliar os polimorfismos do tipo minissatélites (VNTR, do inglês *variable number of tandem repeats*), como forma de identificação dos indivíduos, originando o termo DNA *fingerprinting* (Gill et al., 1985). No mesmo período, Kary Mullis desenvolveu a técnica de reação em cadeia da polimerase (PCR), revolucionando a genética forense ao permitir a amplificação de pequenas regiões específicas do DNA genômico a partir de quantidades limitadas de DNA (Butler, 2010).

No início da década de 1990, um novo grupo de marcadores conhecido como microssatélites, ou repetições curtas em tandem (STRs, do inglês *short tandem repeats*),

passou a ser amplamente utilizado para identificação humana. Os STRs são constituídos por unidades de repetição que possuem de 2 a 6 pares de bases (pb) de comprimento (mais frequente 4pb), cujo número de repetições pode variar significativamente entre os indivíduos (Gutiérrez-Hurtado et al., 2025). Essa variabilidade genética confere aos STRs um alto poder discriminatório, essencial para identificação dos indivíduos (Kowalczyk et al., 2018). A partir dos marcadores STR, as análises forenses de DNA se expandiram. Assim, diversos países iniciaram a criação de seus próprios bancos de dados de perfis genéticos. O primeiro deles foi estabelecido no Reino Unido, em abril de 1995, pelo órgão *Forensic Science Service*, abrangendo inicialmente apenas seis *loci* STR utilizados para determinar o perfil genético de indivíduos: FGA, TH01, VWA, D8S1179, D18S51 e D21S11.

Em 1997, o FBI (*Federal Bureau of Investigation*, Estados Unidos) propôs a padronização de 13 *loci* STR autossômicos para compor o sistema CODIS (*Combined DNA Index System*). Os *loci* selecionados foram: CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 e D21S11 (Butler, 2010). Essa padronização permitiu a consolidação, em 1998, de um banco de dados de perfis genéticos dos Estados Unidos, facilitando a comparação e o intercâmbio de informações entre laboratórios forenses.

Em 2017, com o crescimento do banco de dados e com o objetivo de ampliar o poder de discriminação, sete novos *loci* STR foram adicionados ao sistema CODIS: D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 e D22S1045. Com isso, o número total de *loci* passou a ser 20, além do marcador de inserção/deleção (InDel) no gene da amelogenina, utilizado para determinação do sexo (Butler, 2010). Dessa forma, esses marcadores foram estabelecidos como o padrão-ouro na identificação humana. Atualmente, o banco de dados de perfis genéticos dos Estados Unidos, denominado NDIS (*National DNA Index System*), que integra os dados armazenados no CODIS, reúne mais de 25 milhões de perfis genéticos provenientes de laboratórios forenses federais, estaduais e regionais (FBI, 2025).

Diversos países adotaram o sistema CODIS, cuja implementação ocorre mediante acordos governamentais com o FBI. No Brasil, a implantação do sistema CODIS teve início entre os anos de 2009 e 2010, acompanhada da criação da Rede Integrada de Bancos de Perfis Genéticos (RIBPG). Atualmente, a RIBPG é composta por 23 Bancos de Perfis Genéticos

Locais ou BPGs (21 estaduais, 1 do Distrito Federal e 1 da Polícia Federal) (Dias Filho et al., 2020; RIBPG, 2025). Segundo dados divulgados pela própria RIBPG, entre novembro de 2014 e maio de 2025, o Banco Nacional de Perfis Genéticos (BNPG) reunia 254.578 perfis genéticos armazenados, distribuídos por categoria ampla: vestígios, referências criminais, pessoas desaparecidas - referências, restos mortais não identificados e pessoas vivas não identificadas (RIBPG, 2025).

A RIBPG, instituída pelo Decreto nº 7.950/2013, tem como principal objetivo manter, compartilhar e comparar perfis genéticos com a finalidade de auxiliar na investigação criminal e na instrução processual. Além do seu papel no âmbito criminal, os bancos de perfis genéticos também desempenham um papel fundamental na busca e identificação de pessoas desaparecidas. Nesses casos, os perfis obtidos a partir de restos mortais não identificados ou de indivíduos com identidade desconhecida são comparados com perfis genéticos de familiares ou amostras de referência direta de pessoas desaparecidas, como dentes de leite, cordão umbilical, próteses dentárias ou peças de vestuário íntimo (RIBPG, 2025).

As amostras são processadas pelos laboratórios oficiais de genética forense, onde são gerados os perfis genéticos, posteriormente inseridos e comparados internamente nos Bancos de Perfis Genéticos (BPGs). Atualmente, cerca de 90% dos perfis inseridos no BNPG estão relacionados à esfera criminal. Desse total, aproximadamente 76% correspondem a amostras de referências de indivíduos cadastrados criminalmente (condenados, identificados criminalmente, decisões judiciais para a coleta e restos mortais identificados), enquanto cerca de 14,3% (36.456 perfis) referem-se a vestígios biológicos provenientes de locais de crime (RIBPG, 2025).

De acordo com os procedimentos operacionais padrão da perícia criminal, há uma ampla variedade de vestígios que podem ser coletados em locais de crime e encaminhados para análise. Esses vestígios podem ser classificados em diferentes categorias, como: vestígios contendo sangue, espermatozoides ou líquido espermático; vestígios de vestes para a identificação de usuários; vestígios contendo traços de DNA; vestígios contendo material da mucosa oral; vestígios contendo materiais biológicos humanos como ossos, dentes, músculo/cartilagem, etc., e vestígios contendo material fecal (Ministério da Justiça e Segurança Pública, 2024).

1.1.2. Amostras complexas e mistura de perfis de DNA

Os laboratórios de genética forense lidam com vestígios/amostras de elevada complexidade, que frequentemente resultam em perfis genéticos inconclusivos e, portanto, não aptos para serem inseridos nos BPGs. Em grande parte dos casos, trata-se de vestígios com quantidades exíguas de DNA (LCN, do inglês *low copy number*) e com grau elevado de degradação, geralmente decorrente da exposição prolongada a condições ambientais adversas (Butler, 2010).

Fatores ambientais, como umidade e altas temperaturas, favorecem a proliferação de micro-organismos que produzem DNAses, enzimas responsáveis pela degradação do material genético. Esse processo compromete a amplificação e posterior genotipagem do DNA. Vestígios biológicos que contêm quantidades exíguas de DNA, frequentemente degradado, e que apresentam contribuição de dois ou mais indivíduos, situações conhecidas como misturas de perfis genéticos, representam um dos maiores desafios enfrentados na análise forense de DNA (Butler, 2010).

Amostras biológicas que compreendem fluidos corporais ou células de dois ou mais indivíduos em diferentes proporções resultam em misturas de perfis genéticos. Misturas altamente desbalanceadas são frequentemente encontradas em rotinas forenses (Tan et al., 2018). A título de exemplo, em casos de violência sexual, é comum que as amostras contenham uma quantidade predominante de DNA da vítima e uma quantidade reduzida de DNA do agressor. Nessas condições, os marcadores STRs, amplamente utilizados na rotina forense, podem apresentar limitações significativas para a correta identificação e interpretação dos perfis genéticos envolvidos (Buckleton et al., 2005; Liu et al., 2018).

Situações como misturas de dois perfis genéticos em proporções inferiores a 1:10, ou amostras contendo três ou mais indivíduos, representam desafios significativos para a interpretação dos resultados (Liu, 2018; Clayton & Buckleton, 2005). Fatores como o compartilhamento de alelos entre os contribuintes, a ocorrência de alelos suprimidos (*drop-out*), a amplificação preferencial de alelos menores e a presença de artefatos de amplificação denominados *stutters* tornam ainda mais complexa a interpretação dessas misturas (Gill et al., 2015; Bennett et al., 2019).

Essas limitações ficaram evidentes no contexto do Projeto de Backlog de Vestígios de Crimes Sexuais foi proposto e elaborado entre 2018 e 2019 pelo Comitê Gestor da RIBPG. A iniciativa do governo teve como objetivo processar um passivo de mais de 150 mil amostras biológicas relacionadas a crimes sexuais. Como exemplo, o Laboratório de Genética Molecular Forense da Polícia Científica do Paraná (LGMF) processou 2003 vestígios provenientes de aproximadamente 900 vítimas de crimes sexuais ocorridos entre os anos de 2010 e 2021. Esse processamento resultou em 786 perfis genéticos masculinos inseridos no banco de perfis genéticos da RIBPG, os quais propiciaram 350 coincidências observadas, envolvendo 196 vítimas (Kobachuk et al., 2023). A partir de um levantamento de dados realizado para este projeto, verificou-se que, dos 2003 vestígios analisados, 785 apresentaram perfis com mais de dois contribuintes, o que representa mais de 39% das amostras analisadas (dados não publicados). Esse alto índice de casos não resolvidos reforça as limitações dos métodos baseados em STRs para lidar com amostras complexas, evidenciando a necessidade de tecnologias complementares ou alternativas para superar esses obstáculos.

Uma alternativa para superar as limitações seria o uso do sequenciamento massivo paralelo (MPS, do inglês *Massively Parallel Sequencing*), também conhecido como sequenciamento de nova geração (NGS, do inglês *Next-Generation Sequencing*). Essa tecnologia amplia significativamente as possibilidades de análise, permitindo a avaliação simultânea de múltiplos *loci* de interesse (autossômicos, sexuais e mitocondriais), bem como de diferentes categorias de polimorfismos (STRs, InDels, SNPs). Além da identificação individual e de vínculos genéticos (e.g., exames de paternidade), o NGS possibilita análises relacionadas a ancestralidade, inferência fenotípica, avaliações epigenéticas, identificação de fluidos corporais e de espécies, além de possibilitar aplicações forenses em plantas, animais e microrganismos (Yang et al., 2014; Bruijns et al., 2018), desde que os marcadores necessários estejam incluídos na análise. Por outro lado, essa tecnologia ainda apresenta desafios para a implementação na rotina laboratorial. O alto custo dos equipamentos, financiamento limitado, maior complexidade nas análises, conhecimento específico em bioinformática e carência de treinamentos são alguns fatores que limitam seu uso em muitos laboratórios forenses (Gross et al., 2021; Foley & Oldoni, 2023).

Outra alternativa seria o uso de diferentes marcadores genéticos, os polimorfismos de nucleotídeo único (SNPs, do inglês *Single Nucleotide Polymorphisms*) e os polimorfismos de inserção e deleção (InDels, do inglês *Insertion-Deletion*), também conhecidos como DIP (*Deletion-Insertion Polymorphism*) (Pereira et al., 2009; Valle-Silva et al., 2019). Esses marcadores apresentam vantagens importantes em contextos forenses desafiadores e podem ser utilizados como ferramentas complementares aos STRs.

1.1.3. Marcadores genéticos para fins forenses

Os principais marcadores genéticos utilizados na genética forense como alternativa aos STRs são os SNPs e os InDels. Cada marcador apresenta características específicas, com vantagens e limitações relacionadas a taxas de mutação, distribuição genômica, número de alelos, frequências alélicas, poder de discriminação e suscetibilidade à geração de artefatos. Dessa forma, a escolha dos marcadores deve ser criteriosa e adequada ao objetivo da análise a ser realizada (Butler, 2010).

Os STRs consolidaram-se entre os marcadores genéticos mais utilizados em laboratórios forenses, principalmente devido ao seu elevado poder de discriminação. Apesar de VNTRs apresentarem poder de discriminação ainda maior, os STRs apresentam alelos de menor tamanho, o que facilita sua amplificação por PCR e posterior detecção por eletroforese capilar (EC). Além disso, esses marcadores estão amplamente distribuídos ao longo do genoma humano, com ocorrência média um a cada 2.000 pb (Butler, 2010; Butler, 2015; Gutiérrez-Hurtado et al., 2025).

Os STRs selecionados para fins de identificação humana apresentam um conjunto específico de características desejáveis, tais como: capacidade de amplificação por PCR em sistema multiplex, alelos facilmente distinguíveis e analisados por EC, elevado poder de discriminação e ausência de ligação genética com outros *loci* STR incluídos na análise (Goodwin et al., 2010).

Apesar de amplamente utilizados, os STRs apresentam limitações relevantes. Esses marcadores são caracterizados por uma elevada taxa de mutação ($\sim 10^{-3}$), o que resulta em um grande número de alelos por *locus* e, por consequência, em um alto poder de discriminação.

No entanto, essa mesma característica pode gerar inconsistências em análises de vínculo genético, como em testes de paternidade. Esse problema pode ser minimizado pela inclusão de um maior número de *loci* STR no ensaio. Ainda assim, o número de *loci* que podem ser analisados simultaneamente pela metodologia tradicional é limitado tanto pelos intervalos de tamanho dos fragmentos amplificados quanto pela quantidade de fluoróforos que a plataforma de EC comporta (Butler, 2015; Ricci et al., 2007).

Além disso, os STRs são propensos à formação de artefatos conhecidos como produtos de *stutter* ou produtos de “*polymerase slippage*”. O mecanismo de formação dos produtos *stutter* é devido ao pareamento incorreto após o deslizamento das fitas de DNA (*slipped strand mispairing* - SSM) durante a PCR. A enzima DNA polimerase se dissocia temporariamente da fita molde, o qual pode formar uma alça (*loop*) composta por uma ou mais unidades repetitivas, ocasionando um desalinho no momento do reparamento e retomada da síntese (Bright et al., 2014). Por consequência, gera-se uma pequena proporção de produtos de PCR com tamanhos diferentes dos alelos reais do indivíduo, que são indistinguíveis de alelos verdadeiros presentes na população (Gill et al., 2015). A presença desses artefatos traz grande dificuldade ao processo de identificação e interpretação de mistura de diferentes perfis de indivíduos, em especial quando há uma amostra sub representada, visto que alelos verdadeiros podem facilmente se confundir com artefatos de amplificação (Gill et al., 2006; Gill et al., 2015).

Comumente, a taxa de produtos de *stutter* gerados em uma reação de PCR fica em torno de 2 a 15% dependendo do *locus*. Em geral, essa taxa é influenciada pelo tamanho da unidade de repetição, pois repetições dinucleotídicas possuem uma taxa maior do que trinucleotídicas, que por sua vez é superior a tetranucleotídicas, e assim sucessivamente. O tamanho dos alelos no *locus* também influenciam a taxa de formação dos *stutters*, pois alelos com maior número de unidades repetidas tendem a gerar mais produtos *stutter* (Rodrigues et al., 2022). Além disso, os pares de base adenina e timina (A–T) possui influência na quantidade de *stutter*. Unidades de repetição com alto conteúdo de A–T produzem maiores quantidades de *stutter* quando comparado aos pares de base citosina e guanina (C–G) (Brookes et al., 2012).

Os avanços na tecnologia de MPS, aliados aos consórcios nacionais e internacionais de sequenciamento do genoma humano, têm proporcionado uma contribuição fundamental para a identificação de novos polimorfismos. Por exemplo, no Projeto 1000 Genomas (1kGP) fase

III, foram identificados aproximadamente 84,7 milhões de SNPs, 3,6 milhões de InDels, e cerca de 60 mil variantes estruturais (SVs, do inglês *Structural Variants*), a partir da análise de 2.504 indivíduos provenientes de 26 populações. As populações foram classificadas em 5 superpopulações: África (AFR), Leste Asiático (EAS), Europa (EUR), Sul da Ásia (SAS) e Américas (AMR) (Auton et al., 2015). Em 2022, o projeto 1kGP (fase IV) foi expandido para 3.202 amostras, incluindo 602 trios. Os pesquisadores utilizaram sequenciamento de genoma completo (WGS) de alta cobertura e identificaram cerca de 111 milhões de SNVs, 14 milhões de InDels e 173 mil SVs, demonstrando um ganho significativo de sensibilidade e precisão na identificação das variantes genéticas em comparação com a fase anterior do projeto, que utilizava dados de baixa cobertura (Byrska-Bishop et al., 2022). Quando comparado ao polimorfismo do tipo STR, um estudo recente de Shi e colaboradores (2023) identificaram aproximadamente 366 mil variantes STRs no genoma humano, utilizando dados de alta cobertura provenientes do Projeto 1kGP e do Projeto NyuWa, o qual inclui 3.983 amostras da população chinesa. Considerando apenas os dados do 1kGP, foi possível identificar em torno de 276 mil STRs (Shi et al., 2023).

Os polimorfismos do tipo SNVs (do inglês, *Single Nucleotide Variants*) representam a forma mais comum de variação genética no genoma humano, caracterizado por alterações que envolvem a substituição de um único nucleotídeo em uma sequência de DNA, independentemente de sua frequência na população. Já os SNPs, consistem no subgrupo de SNVs que apresentam uma frequência alélica igual ou superior a 1% na população (Brookes, 1999).

Os SNPs apresentam uma taxa de mutação menor ($\sim 10^{-8}$) em comparação aos STRs ($\sim 10^{-3}$ a 10^{-5}). Além disso, os produtos de amplificação (*amplicons*) gerados para se analisar os SNPs são geralmente menores, o que os torna mais adequados para a análise de amostras degradadas ou em casos envolvendo desastres em massa (DVI - *disaster victim identification*). Os SNPs também se destacam pela ausência de artefatos do tipo *stutters*, comumente observados na amplificação e na análise dos STRs, bem como o alto potencial de automação e capacidade de ser inseridos em sistemas multiplex (Kidd et al., 2006; Butler, 2010; Canturk et al., 2014; Kidd & Speed 2015).

Recentemente, a comunidade forense classificou os SNPs em quatro tipos: SNPs informativos de identidade (iiSNPs), usados para identificação individual; SNPs de linhagem, que fornecem dados para testes de vínculo genético e estudos evolutivos; SNPs informativos de ancestralidade (aiSNPs), usados para inferir a origem biogeográfica do indivíduo; e SNPs informativos de fenótipo (piSNPs), o qual refere-se à previsão das características visíveis como pele, cabelo, cor dos olhos, características faciais, entre outros, conhecidas como Características Externamente Visíveis (EVCs) (Gutiérrez-Hurtado et al., 2025).

Apesar dos SNPs apresentarem diversas vantagens, ainda existem limitações importantes quanto ao seu uso na área forense. Os SNPs são pouco informativos (geralmente são bi-alélicos), o que torna necessário a análise de grande quantidade de marcadores para se obter um poder de discriminação equivalente ao obtido com os STRs. Estima-se que para obter o mesmo poder de discriminação utilizando 10 *loci* STR, sejam necessários analisar em torno de 50 a 80 *loci* SNP (Goodwin et al., 2010). Por esses marcadores serem geralmente bi-alélicos, são incapazes de detectar de forma confiável uma amostra contendo misturas de perfis (Kidd & Speed, 2015). Uma das alternativas seria o uso de SNPs multi-alélicos como os microhaplótipos, SNPs tri e tetra-alélicos, os quais demonstraram capacidade em analisar amostras contendo misturas de perfis e DNA degradado (Westen et al., 2009; Phillips et al., 2015; Bennett et al., 2019).

A detecção de SNPs pode ser realizada por diferentes metodologias, tais como o mini-sequenciamento SNaPshot® (Applied Biosystems), sondas de hibridização TaqMan®, *microarrays* e MPS (Gutiérrez-Hurtado et al., 2025). A metodologia de *microarray* é capaz de analisar centenas de marcadores, a custos mais baixos, com aplicações em testes de vínculo genético. Entretanto, são menos eficazes frente a amostras complexas em contextos forenses (Matute et al., 2025).

O uso da metodologia MPS possibilita análise simultânea de um grande número de SNPs, bem como outros tipos de variantes. Um exemplo é o kit ForenSeq DNA Signature Prep (Verogen, San Diego, CA, EUA), o qual permite a análise simultânea de 27 STRs autossômicos, 24 Y-STRs, 7 X-STRs, 94 iiSNPs, 22 piSNPs e 56 aiSNPs (Kiesler et al., 2023). Apesar do uso desta tecnologia aumentar o poder de discriminação, a abrangência de marcadores e melhorar as análises de amostras complexas, ela representa um desafio para

muitos laboratórios forenses no Brasil, principalmente devido ao custo elevado, necessidade de infraestrutura computacional robusta, além de preocupações éticas, legais e de privacidade quanto ao uso dos dados obtidos.

Os polimorfismos do tipo InDel são polimorfismos de comprimento, gerados em decorrência da inserção ou deleção de um segmento de DNA, variando de um a centenas de nucleotídeos (Mills et al., 2006). Apresentam alta estabilidade genética devido à sua baixa taxa de mutação, sendo, por esse motivo, considerados marcadores de evolução lenta (Bastos-Rodrigues et al., 2006). Os InDels estão amplamente distribuídos ao longo do genoma humano, sendo classificados como a segunda forma mais abundante de variação genética. Desse modo, os InDels têm ganhado crescente reconhecimento por sua contribuição significativa para a diversidade genômica humana (Bhangale et al., 2005).

Os estudos de Weber et al. (2002) e Mills et al. (2006) contribuíram fundamentalmente para o conhecimento sobre os InDels, destacando sua abundância no genoma e sua facilidade de análise. Weber et al. (2006) identificaram e caracterizaram aproximadamente 2.000 InDels no genoma humano, além de terem mensurado as suas frequências alélicas em diferentes populações. Mills e colaboradores (2006) conduziram o primeiro mapa de identificação dos InDels e relatou mais de 415.000 InDels no genoma humano, classificando-os em cinco categorias: (1) inserção/deleção de uma base, (2) expansões monoméricas de uma base, (3) expansões múltiplas de unidades repetitivas de 2-15 pb, muitas relacionadas a STRs, (4) inserção de transposons, e (5) InDels contendo sequências de DNA randômicas. A última categoria de inserções e deleções, caracterizada por conter sequências de DNA randômicas, representa aproximadamente 41% de todos os InDels identificados. Esses polimorfismos apresentam uma ampla variação de tamanho, 2 pb até 10 kilobases (kb), sendo que a maioria com comprimento inferior a 100 pb, o que os torna compatíveis com a detecção por EC (Pereira et al., 2009).

Outro estudo a partir de dados de sequenciamento de 179 indivíduos provenientes de três populações distintas (Nigéria - YRI, Utah – CEU e China/Japão – CHB/JPT) conseguiram identificar cerca de 1,6 milhão de InDels, os quais foram classificados em 5 categorias: sequências de homopolímeros (HR); repetições em tandem (TR), regiões *hotspots*, próximas a regiões repetitivas (PH) e regiões não repetitivas (NR), subdividido em regiões com alteração

de cópias (NR-CCC), com duplicação local, e sem alteração de cópias (NR, non-CCC), como os InDels complexos (Montgomery et al., 2013). Conforme Montgomery et al. (2013), o deslizamento da polimerase é responsável por cerca de 75% de todos os InDels, enquanto os casos restantes são, principalmente, deleções simples em regiões com sequências complexas.

Na fase IV do projeto 1kGP, foram identificados cerca de 14 milhões InDels, representando um aumento de aproximadamente 4,05 vezes em relação às chamadas das variantes realizado na fase III. Aproximadamente 10 milhões de “novos” InDels foram identificados, destes 60% estavam localizados em regiões de homopolímeros e repetições em tandem (Byrska-Bishop et al., 2022).

Atualmente, os InDels têm despertado crescente interesse na área forense, devido ao seu alto potencial, uma vez que reúnem diversas características consideradas vantajosas, tais como: a) abundância e ampla distribuição no genoma, em cromossomos autossômicos e sexuais (Weber et al., 2002); b) apresentam taxas de mutações mais baixas, estimado em 0,53 a $1,5 \times 10^{-9}$ (Kloosterman et al., 2015) quando comparado aos SNPs, resultando em uma melhor estabilidade no genoma, favorecendo assim análises envolvendo vínculo genético (Bastos-Rodrigues et al., 2006; Pereira et al., 2009; Gao et al., 2015); c) InDels menores produzem *amplicons* menores, aumentando a eficiência da amplificação em amostras degradadas e com baixa quantidade de DNA, como em situações de desastres em massa, cadáveres em avançado estado de decomposição ou carbonizados (Pereira et al., 2009; Francez et al., 2012; Pereira & Gusmão, 2012); d) não geram *stutters*, o que favorece a clareza na análise e interpretação de amostras com mistura de perfis genéticos (Larue et al., 2014); e) são capazes de serem inseridos em sistema multiplex (Pereira et al., 2009); f) a genotipagem pode ser realizada, de forma automatizada, utilizando uma única amplificação por PCR seguida da detecção dos fragmentos amplificados por EC, metodologias padronizadas nos laboratórios de genética forense (Fondevila et al., 2012).

Diante do potencial dos InDels como marcadores genéticos complementares, Pereira e colaboradores (2009) desenvolveram um ensaio multiplex composto por 38 InDels autossômicos bi-alélicos denominado 38 indel-plex. A amplificação é passível de ser realizada a partir de uma única PCR multiplex seguida por detecção em EC. Os perfis completos foram obtidos utilizando 0,3ng de DNA e os *amplicons* detectados apresentaram

tamanho inferior a 160pb, indicando uma alternativa para identificação humana em casos envolvendo amostras degradadas.

Posteriormente, foi desenvolvido o primeiro kit comercial, o *Investigator® DIPplex Kit* (Qiagen), composto por 30 InDels bi-alélicos distribuídos em 19 cromossomos. O ensaio demonstrou alta sensibilidade, sendo capaz de gerar perfis completos com apenas 62 pg de DNA sendo uma alternativa eficaz para identificação humana (Larue et al., 2012). O *Investigator® DIPplex Kit* foi aplicado em diversas populações, incluindo Europa, Ásia, África e América. No Brasil, Palha e colaboradores (2015) aplicaram o *Investigator® DIPplex Kit* em 519 indivíduos oriundos de seis estados brasileiros, obtendo um elevado poder de discriminação ($PMC = 3,4 \times 10^{-13}$). No entanto, observaram um poder de exclusão relativamente baixo ($PEC = 0,9973$), indicando que, embora o kit seja eficaz para identificação individual, seu uso em testes de vínculo genético deve ser complementar aos STRs.

Assim como os SNPs, outra vantagem bastante promissora dos InDels consiste no fato de que esses polimorfismos apresentam variações significativas nas frequências alélicas entre diferentes populações, o que os torna úteis como marcadores informativos de ancestralidade (AIMs, do inglês *Ancestry-Informative Markers*) (Bastos-Rodrigues et al., 2006; Francez et al., 2012; Valle-Silva et al., 2019). Bastos-Rodrigues et al. (2006) propuseram um painel composto por 40 InDels bi-alélicos capazes de fornecer informações sobre a estrutura genética populacional, permitindo a distinção de cinco grandes regiões geográficas: América, África Subsaariana, Leste Asiático, Oceania e um grupo abrangendo Europa, Oriente Médio e Ásia Central.

Outro painel desenvolvido por Pereira et al. (2012), composto por 46 InDels bi-alélicos informativos de ancestralidade, demonstrou ser eficiente na estimativa da ancestralidade individual para quatro grupos ancestrais: africano, europeu, leste-asiático e nativo-americano. Os marcadores foram otimizados em um sistema multiplex com detecção por EC, representando uma alternativa mais simples e acessível em comparação aos métodos tradicionais baseados em AIM-SNPs, que geralmente requerem protocolos mais complexos.

Como se pode observar, os painéis descritos até o momento são compostos exclusivamente por InDels bi-alélicos. Esses marcadores exigem a análise de um número

maior de *loci* para alcançar um poder de discriminação equivalente ao dos STRs (Schneider, 2012; Huang et al., 2014), assim como os SNPs (Goodwin et al., 2010). No entanto, novos tipos de marcadores nomeados como microhaplótipos vêm sendo investigados e têm despertado crescente interesse na genética forense, com o objetivo de superar essa limitação.

1.1.4. Microhaplótipos

O conceito de microhaplótipo foi introduzido por Kidd e colaboradores (2013) e refere-se originalmente a regiões do DNA que contêm dois ou mais SNPs dentro de uma distância molecular inferior a 200 pb, formando três ou mais combinações alélicas distintas (haplótipos). Além de manter as propriedades individuais das variantes, um microhaplótipo proporciona uma maior informatividade quando comparado a um único *locus* bi-alélico (Kidd et al., 2013; Huang et al., 2014).

Atualmente, diversos estudos têm explorado a associação de diferentes variantes, denominado de forma generalizada, como microhaplótipos. Entre esses, destacam-se combinações como SNP-STR (Tan et al., 2018; Wei et al., 2018), DIP-SNP (Liu et al., 2018), DIP-STR (Castella et al., 2013; Cereda, 2014) e polimorfismos InDels interligados, também chamados de multi-InDels (Huang et al., 2014; Qu et al., 2020). Esses novos marcadores têm se mostrado promissores em diversas aplicações forenses, como a identificação individual, inferência de ancestralidade e de linhagem familiar, além de oferecerem uma alternativa para análises de amostras degradadas, com DNA exíguo e contendo misturas de perfis genéticos (Kidd et al., 2013; Oldoni & Podini, 2019; Zhao et al., 2018; Qu et al., 2020; Sun et al., 2022).

O nosso trabalho tem como foco o marcador denominado como multi-InDel, definido como um haplótipo formado por dois ou mais InDels com uma distância inferior a 200pb (Figura 1). Este tipo de marcador preserva as características inerentes dos InDels, porém amplia o nível de informatividade (Huang et al., 2014).

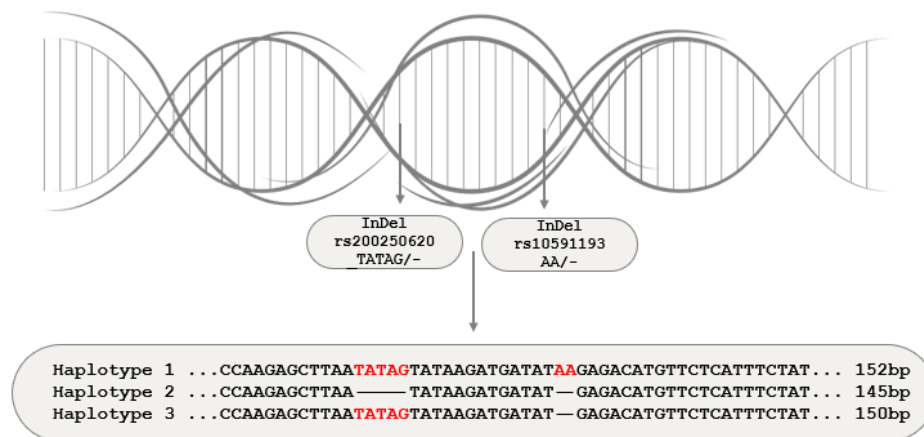


Figura 1. Imagem ilustrativa demonstrando o marcador multi-InDel constituído por dois InDels (rs200250620 e rs10591193) (Huang et al., 2014), capazes de formar três haplótipos com tamanhos distintos.

Com o intuito de facilitar a compreensão sobre os marcadores multi-InDel, apresenta-se a seguir uma linha do tempo com as principais informações relacionadas a esses marcadores, reunindo descobertas e avanços relevantes que evidenciam seu potencial na genética forense.

Há cerca de 11 anos, foi desenvolvido o primeiro painel composto por marcadores multi-InDel, proposto por Huang et al. (2014). Os autores desenvolveram um sistema contendo 20 marcadores multi-InDel, totalizando 43 *loci* de InDels, combinados em um ensaio multiplex com detecção por EC. A genotipagem foi realizada em 150 indivíduos chineses Han, resultando na identificação de 63 *amplicons* específicos, com tamanhos variando entre 80 e 170 pb. Os resultados de PMC e de PEC foram de $5,7 \times 10^{-12}$ e 0,9989, respectivamente. Esses valores são comparáveis aos obtidos com o painel 38 indel-plex (38 InDels bi-alélicos), que apresentou PMC na ordem de 10^{-14} a 10^{-15} (Pereira et al., 2009).

Posteriormente, Song et al. (2015) realizaram a validação do painel de multi-InDels proposto por Huang et al. (2014), avaliando sua especificidade frente a diferentes espécies, sensibilidade, capacidade de detecção em amostras degradadas e análise de misturas, também aplicado em indivíduos chineses Han. O ensaio demonstrou alta sensibilidade, sendo capaz de gerar perfis confiáveis com apenas 0,3 ng de DNA. Além disso, os 20 multi-InDels puderam ser genotipados com sucesso em amostras degradadas artificialmente. Em testes de mistura, o painel foi capaz de identificar todos os alelos de dois indivíduos não relacionados

em proporção de até 3:7, ou seja, mistura na proporção de 0,3ng e 0,7ng de DNA controle padrão (9947 A e 9948). Gao e colaboradores (2015), demonstrou que o painel desenvolvido por Huang et al. 2014 também é eficaz para estudos forenses de parentesco, especialmente em situações as quais mutações em *loci* STR dificultam a interpretação dos resultados.

Alguns pesquisadores começaram a explorar marcadores multi-InDels localizados no cromossomo X (ChrX). Fan et al. (2015) desenvolveram o primeiro painel composto por 13 multi-InDels, englobando 28 *loci* InDel no ChrX. O painel apresentou alto polimorfismo na população chinesa Han, demonstrando seu potencial para aplicações forenses, especialmente na identificação individual e em testes de parentesco, incluindo casos de vínculo familiar complexos que possam se beneficiar da análise de cromossomos X. O mesmo grupo de pesquisadores avaliou um painel composto por 10 *loci* InDel no ChrX para inferência da ancestralidade de três populações chinesas (Uigur, Tibetano e Han), o qual obteve capacidade em distinguir a ancestralidade em três grandes regiões continentais (África, Leste-Asiático e Europa) (Fan et al., 2016).

Em seguida, Sun et al. (2016 e 2019) publicaram um painel composto por 12 marcadores multi-InDels autossômicos voltado também para inferência de ancestralidade. O painel demonstrou-se eficaz para inferir a ancestralidade de indivíduos oriundos das regiões Norte, Sul e Sudeste da Ásia (nível subcontinental), apresentando potencial para aplicações forenses em contextos populacionais asiáticos.

Em 2020, Qu et al. desenvolveram um novo painel composto por 18 marcadores multi-InDels, totalizando 54 *loci* InDels autossômicos. Esses marcadores foram organizados em sistema multiplex com detecção por EC, gerando *amplicons* entre 107 e 326 pb. O painel foi aplicado em 200 indivíduos não relacionados da província de Sichuan (China). Os resultados de PMC e PEC foram de $2,77 \times 10^{-12}$ e 0,9984, respectivamente. Todos os perfis foram detectados em proporções variando entre 4:1 a 1:4, ou seja, mistura em proporções de 0,2ng e 0,8ng de DNA controle (F312 e M308), e em amostras com quantidades superiores a 0,125 ng de DNA geraram perfis completos. Os autores também demonstraram sua utilidade frente a identificação individual e como ferramenta complementar aos marcadores STRs amplamente utilizados.

Outro painel foi proposto contendo 18 multi-InDels autossômicos em sistema multiplex e com detecção em EC, abrangendo 36 *loci* InDels (Li et al., 2021). O painel foi aplicado em 192 amostras provenientes da população Han de Hunan (China), nas quais foram detectados *amplicons* de até 125 pb e com uma sensibilidade de 0,125ng de DNA. Os resultados publicados demonstraram PMC de $3,23 \times 10^{-12}$ e PEC de 0,9989, sugerindo um painel independente para identificação humana e que pode ser utilizado como suporte para casos envolvendo parentescos (Li et al., 2021).

Em 2022, Liu et al. propuseram um novo painel composto por 20 multi-InDels, totalizando 41 *loci* InDels. O painel foi empregado em 200 indivíduos não relacionados oriundos da população de Han Chinesa. Os pesquisadores detectaram *amplicons* entre 110 e 289 pb, obtendo uma sensibilidade de 0,125ng de DNA. Os resultados publicados demonstraram PMC de 7×10^{-11} e PEC de 0,992, valores inferiores aos obtidos por Li et al. (2021).

Diante do potencial desses marcadores, Yao et al. (2022) conduziram uma busca por marcadores multi-InDels e InDels multi-alélicos utilizando dados de sequenciamento de alta cobertura do 1kGP. Foram filtrados apenas os InDels classificados como sequências de DNA randômicas, uma vez que, ao contrário de outras categorias resultantes de múltiplas ocorrências de ganho ou perda de unidades repetitivas em decorrência do deslizamento da DNA polimerase, esses InDels podem refletir eventos mutacionais raros e únicos ao longo da história (Montgomery et al., 2013). Embora sejam mutações mais raras, é possível que outros eventos ocorram coincidentemente na mesma posição, originando os InDels multi-alélicos com três ou mais alelos. Além disso, uma segunda mutação pode surgir nas proximidades do *locus* InDel, levando à formação dos multi-InDels. No estudo, foram identificados 3.599 InDels multi-alélicos e 6.375 multi-InDels, sendo a maioria das variantes observadas em mais de uma superpopulação (AFR, EAS, EUR, SAS e AMR) (Yao et al., 2022).

No mesmo ano, Sun e colaboradores (2022) exploraram o uso de técnicas de *machine learning* e diversas abordagens para otimizar a estratificação populacional e inferência de ancestralidade em populações asiáticas, utilizando um conjunto de 12 marcadores multi-InDels previamente selecionados (Sun et al., 2016; Sun et al. 2019). De maneira bastante particular, as populações asiáticas, assim como o Brasil, também apresentam o desafio da

miscigenação, resultado de frequentes trocas culturais, movimentos populacionais dentro do continente, origens ancestrais complexas e de interações entre diferentes grupos étnicos.

A partir de 2023, pesquisadores passaram a adotar uma nova abordagem, integrando marcadores multi-InDel com potencial para identificação individual e inferência de ancestralidade em um único ensaio multiplex, seguido por detecção por EC. Lan et al. (2023) desenvolveram um ensaio multiplex composto por 41 marcadores multi-InDel, abrangendo 82 *loci* InDel, aplicado às populações Manchu e Mongol, na China. Nessas populações, o estudo obteve valores de PMC entre $3,68 \times 10^{-27}$ e $2,15 \times 10^{-26}$, e PEC entre 0,999999418 e 0,999999845, desempenho comparável ao de 26 a 27 *loci* STR. Além disso, os resultados também permitiram inferências sobre a ancestralidade genética dos dois grupos analisados.

O mesmo painel foi aplicado à população Zhuang da província de Yunnan (China), obtendo resultados promissores, com um PMC de $8,0671 \times 10^{-26}$ e um PEC de 0,9999995959. Além disso, os dados revelaram que o grupo Zhuang de Yunnan apresenta proximidade genética com cinco populações do Leste Asiático (Mei et al., 2023).

Outra abordagem mais recente envolve a combinação de marcadores multi-InDels com InDels multi-alélicos, visando aumentar ainda mais a informatividade dos marcadores e possibilitar, simultaneamente, a identificação individual, a análise de vínculo genético e a inferência de ancestralidade. Inicialmente, Yao e colaboradores (2023) desenvolveram um outro painel composto somente por 31 InDels multi-alélicos, denominado como FA-HID32plex, e aplicaram na população Han Chinesa, obtendo resultados de PMC de $1,5 \times 10^{-19}$ e PEC de 0,999920.

Logo em seguida, em março de 2024, o mesmo grupo de pesquisadores desenvolveu um novo painel, denominado MMIDplex, composto por 17 InDels multi-alélicos e 22 multi-InDels. Quando aplicado à população Han chinesa, o painel apresentou um PMC de $1,3 \times 10^{-23}$ e um PEC de 0,99999017. A análise dos parâmetros forenses desses marcadores foi ampliada para cinco superpopulações, utilizando dados de sequenciamento de alta cobertura do 1kGP (AFR, n = 661, AMR, n = 347, EUR, n = 503, EAS, n = 504 e SAS, n = 489) nas quais foram observados valores de PMC na ordem de 10^{-19} e PEC na ordem de 10^{-4} (Yao et al., 2024). Contudo, além de combinar dois tipos de marcadores (InDels multi-alélicos e multi-InDels), este painel apresentou limitações na inferência de ancestralidade,

particularmente na distinção entre populações das Américas e do Sul da Ásia. **Até o momento, este foi o único estudo a avaliar os parâmetros forenses dos marcadores multi-InDels em diferentes populações mundiais, anteriormente analisados apenas em populações asiáticas.** Ressalta-se que a população brasileira, uma população amplamente reconhecida por seu alto grau de miscigenação, não está representada no 1kGP.

Diante do exposto, destaca-se que os painéis de multi-InDels representam uma abordagem recente, promissora e em constante evolução na literatura científica. Diversos estudos evidenciam o seu potencial uso para diferentes aplicabilidades no âmbito forense, reunindo cada vez mais conhecimento sobre esses marcadores.

1.1.5. População Brasileira

O Brasil é o maior país da América Latina e possui a maior população recentemente miscigenada do mundo, com auge de miscigenação entre os séculos XVIII e XIX (Nunes et al., 2025). A população brasileira é o reflexo de cinco séculos de miscigenação entre três diferentes grupos ancestrais, ameríndios, europeus e africanos, resultando em uma população geneticamente heterogênea (Pena et al., 2011; Pena et al., 2020). Estima-se que antes da chegada dos colonizadores, aproximadamente 10 milhões de indígenas habitavam o território brasileiro, falando mais de 1.000 línguas (Nunes et al., 2025). A imigração europeia teve início com os portugueses, sendo que, nos três primeiros séculos de colonização, cerca de 500 mil indivíduos vieram de Portugal. Paralelamente, o tráfico de escravos teve início na segunda metade do século XV e perdurou até 1850, trazendo cerca de 3,5 milhões de africanos para o Brasil (Santos et al., 2009). Em sua história recente, o Brasil recebeu muitos imigrantes de outras partes da Europa, do Oriente Médio e até mesmo do Leste Asiático, como italianos, espanhóis, japoneses, alemães, libaneses, sírios e outros (Pena et al., 2011; Pena et al., 2020).

A formação histórica e demográfica contribuiu para a ampla diversidade genética da população brasileira, sendo observada em estudos populacionais recentes. Um ótimo exemplo é o estudo genômico da coorte denominada SABE (Saúde, Bem-Estar e Envelhecimento), representativa do censo de idosos da cidade de São Paulo, que analisou o genoma completo (WGS – *whole-genome sequencing*) de alta cobertura de 1.171 indivíduos não aparentados

com 60 anos ou mais. Foram identificadas cerca de 76 milhões de variantes genéticas, incluindo SNVs e InDels, das quais aproximadamente 2 milhões não estavam presentes em bancos genômicos internacionais (Naslavsky et al., 2022). Mais recentemente, Nunes et al. (2025) avaliou 2.723 sequências genômicas completas de alta cobertura da população brasileira, incluindo indivíduos de comunidades urbanas, rurais e ribeirinhas, e confirmaram a enorme diversidade genômica dos brasileiros, identificando aproximadamente 9 milhões de variantes até então desconhecidas.

Nesse contexto, é possível identificar novos alelos (haplótipos) e frequências alélicas distintas dos marcadores multi-InDels na população brasileira, considerando que a diversidade genética das populações é fortemente influenciada pelas características histórico-demográficas, além de populações miscigenadas tender a abranger variantes específicas (Naslavsky et al., 2022; Nunes et al., 2025).

1.2. Justificativa

Atualmente, a análise de DNA tornou-se uma evidência amplamente aceita, indispensável e integrada à rotina das investigações forenses. Por meio do uso de marcadores genéticos, é possível estabelecer a ligação de suspeitos a locais de crime, ou mesmo conectar diferentes locais entre si, com base em vestígios biológicos coletados, tais como saliva em pontas de cigarro, células epiteliais em superfícies (*touch DNA*) ou fios de cabelo e pelos encontrados em roupas. No entanto, a evidência genética deve ser sempre interpretada em conjunto com outros tipos de evidências, sendo todas apresentadas ao sistema judiciário de maneira imparcial (Bond, 2007; Jobling & Gill, 2004).

A rotina dos laboratórios de genética forense é constantemente desafiada por situações complexas, especialmente quando envolve análises de amostras com quantidade exígua de DNA, elevado grau de degradação ou com contribuição de dois ou mais indivíduos. O desafio torna-se ainda mais complexo quando todas as condições ocorrem simultaneamente. Com o avanço das pesquisas, melhorias significativas foram alcançadas em diferentes âmbitos, incluindo extração de DNA das amostras, descobertas de novos marcadores genéticos, e desenvolvimento de plataformas de detecção e ferramentas analíticas. No entanto, a análise dessas amostras críticas continuam sendo um processo que exige considerável tempo e esforço (Tao et al., 2018). Dessa forma, destaca-se a importância do desenvolvimento de novos painéis genéticos para identificação humana, que atuem como ferramentas complementares aos painéis tradicionalmente adotados como padrão-ouro e que sejam alternativas eficazes para a análise de amostras complexas.

Conforme evidenciado na literatura, os marcadores multi-InDels têm demonstrado resultados atuais e promissores diante de diversos desafios enfrentados na área forense. No entanto, até o momento, **nenhum dos painéis desenvolvidos foi avaliado na população brasileira, e poucos marcadores foram explorados em populações geneticamente distintas das asiáticas.** Nesse contexto, torna-se essencial avaliar os marcadores previamente propostos em amostras de indivíduos brasileiros, além de buscar, com o auxílio de ferramentas computacionais, novos marcadores mais adequados, a fim de desenvolver alternativas eficazes

e compatíveis com a diversidade genética da nossa população, assim como a de outras populações globais.

1.3. Objetivos

Geral:

Construir e validar, por meio de abordagens computacionais, um painel de marcadores genéticos composto por haplótipos de InDels autossômicos denominado multi-InDels, com o objetivo de identificar os indivíduos de forma eficaz, sendo aplicáveis tanto na população brasileira quanto em outras populações do mundo.

Específicos:

1 - Avaliar a aplicabilidade de três conjuntos de marcadores multi-InDel previamente propostos quanto ao seu potencial uso para identificação humana na população brasileira, utilizando dados públicos de sequenciamento de alta cobertura (30x) de 1.171 indivíduos do projeto Saúde, Bem-estar e Envelhecimento (SABE);

2 - Analisar as regiões genômicas dos marcadores multi-InDel previamente propostos e estimar, na população brasileira, a quantidade de alelos (haplótipos), frequência alélica, a informatividade forense, o número efetivo de alelos e aderência ao equilíbrio de Hardy-Weinberg;

3 - Realizar a seleção, por meio de abordagens computacionais, de novos marcadores multi-InDel, a partir de dados públicos de sequenciamento de alta cobertura (30x) do Projeto 1000 Genomas e desenhar primers específicos para sua amplificação;

4 - Analisar as regiões genômicas dos novos marcadores multi-InDel selecionados e estimar a quantidade de alelos (haplótipos), frequências alélicas, informatividade forense, número efetivo de alelos e aderência ao equilíbrio de Hardy-Weinberg, utilizando dados públicos do projeto 1000 Genomas e do projeto SABE;

5 - Proposição de um novo conjunto de marcadores multi-InDel, organizados em sistema multiplex, com aplicabilidade na população brasileira e em diferentes populações mundiais.

1.4. Referências

AUTON, A. et al. A global reference for human genetic variation. **Nature**, v. 526, n. 7571, p. 68–74, out. 2015.

BASTOS-RODRIGUES, L.; PIMENTA, J.R.; PENA, S.D. The genetic structure of human populations studied through short insertion-deletion polymorphisms. **Ann Hum Genet.** v. 70, p. 658–665, 2006.

BENNETT, L. et al. Mixture deconvolution by massively parallel sequencing of microhaplotypes. **International Journal of Legal Medicine**, v. 133, n. 3, p. 719–729, maio 2019.

BHANGALE, T.R.; RIEDER, M.J.; LIVINGSTON, R.J.; NICKERSON, D.A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. **Hum Mol Genet.** v. 14, p. 59–69, 2005.

BOND, J. W. Value of DNA evidence in detecting crime. **Journal of Forensic Sciences**, v. 52, n. 1, p. 128–136, jan. 2007.

BRIGHT, J.A. et al. Modelling PowerPlex® Y stutter and artefacts. **Forensic Science International: Genetics**, v. 11, p. 126 – 136, 2014.

BYRSKA-BISHOP, M.; EVANI, U.S.; ZHAO, X.; BASILE, A.O.; ABEL, H.J.; REGIER, A.A., et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. **Cell**, v. 185, p. 3426–3440.e19, 2022.

BROOKES, A. J. The essence of SNPs. **Gene**, v. 234, n. 2, p. 177–186, 8 jul. 1999.

BROOKES, C. et al. Characterising stutter in forensic STR multiplexes. **Forensic Science International: Genetics**, v. 6, n. 1, p. 58 – 63, 2012.

BRUIJNS, B.; TIGGELAAR, R.; GARDENIERS, H. Massively parallel sequencing techniques for forensics: A review. **Electrophoresis**, v. 39, n. 21, p. 2642, nov. 2018.

BUCKLETON, J. S.; TRIGGS, C. M.; WALSH, S. J. (EDS.). **Forensic DNA evidence interpretation**. Boca Raton: CRC Press, 2005.

BUTLER, J. M.; BUTLER, J. M. **Fundamentals of forensic DNA typing**. Amsterdam; Boston: Academic Press/Elsevier, 2010.

BUTLER, J. M. The future of forensic DNA analysis. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1674, p. 20140252, 5 ago. 2015.

CANTURK, K. M. et al. Current status of the use of single-nucleotide polymorphisms in forensic practices. **Genetic Testing and Molecular Biomarkers**, v. 18, n. 7, p. 455–460, jul. 2014.

CASTELLA, V.; GERVAIX, J.; HALL, D. DIP – STR : Highly Sensitive Markers for the Analysis of Unbalanced Genomic Mixtures. **Human Mutation**, v. 34, n. 4, p. 644–654, abr. 2013.

CEREDA, G. An investigation of the potential of DIP-STR markers for DNA mixture analyses. **Forensic Science International**, v. 11, p. 229-240, abr. 2014.

CLAYTON, T. AND BUCKLETON, J. Mixtures. Chapter 7 in Forensic DNA Evidence Interpretation (Eds.:Buckleton, J., Triggs, C.M., Walsh, S.J.), **CRC Press**, p. 217-274, 2005.

DIAS FILHO, C. R.; RODRIGUES, E. L.; MALAGHINI, M.; FRANCEZ, P. A. C.; GARRIDO, R. G. Introdução à Genética Forense. Campinas: Millennium Editora, 2020.

FAN, G.Y.; YE, Y.; LUO, H.; HOU, Y. Use of multi-InDels as novel markers to analyze 13 X-chromosome haplotype loci for forensic purposes. **Electrophoresis**. v. 36, p. 2931-2938, 2015.

FAN, G.Y.; YE, Y.; HOU, Y.P. Detecting a hierarchical genetic population structure via Multi-InDel markers on the X chromosome. **Sci Rep**. v. 18, p. 6:32178, Aug. 2016.

FEDERAL BUREAU OF INVESTIGATION (FBI). CODIS-NDIS Statistics. Disponível em: <<https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints/codis/codis-ndis-statistics>>. Acesso em: 10 abr. 2025.

FRANCEZ, P. A. DA C.; RIBEIRO-RODRIGUES, E. M.; DOS SANTOS, S. E. B. Allelic frequencies and statistical data obtained from 48 AIM INDEL loci in an admixed population

from the Brazilian Amazon. **Forensic Science International: Genetics**, v. 6, n. 1, p. 132–135, jan. 2012.

FRANCEZ, P. A. C., RIBEIRO-RODRIGUES, E. M., VELASCO, A. M., SANTOS, S. E. Insertion– deletion polymorphisms—utilization on forensic analysis. **International Journal of Legal Medicine**, v. 126, p. 491-496, 2012.

FOLEY, M.; OLDONI, F. A global snapshot of current opinions of next-generation sequencing technologies usage in forensics. **Forensic Science International: Genetics**, V. 63, 102819, mar 2023.

FONDEVILA, M.; PHILLIPS, C.; SANTOS, C.; PEREIRA, R.; GUSMÃO, L.; CARRACEDO, A.; BUTLER, J.M.; LAREU, M.V.; VALLONE, P.M. Forensic performance of two insertion-deletion marker assays. **Int J Legal Med.** v. 126, n.5, p. 725-37, set. 2012.

GAO, T.Z. The application of multi-InDel as supplementary in paternity cases with STR mutation. **Forensic Science International: Genetics Supplement Series**, v.5, p. e218-e219, 2015.

GILL, P.; JEFFREYS, A. J.; WERRETT, D. J. Forensic application of DNA ‘fingerprints’. **Nature**, v. 318, n. 6046, p. 577–579, dez. 1985.

GILL, P. et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. **Forensic Science International**, v. 160, n. 2-3 p. 90-101, jul. 2006.

GILL, P. et al. Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches—Twenty years of research and development. **Forensic Science International: Genetics**, v. 18, p. 100–117, set. 2015.

GOODWIN, W.; LINACRE, A.; HADI, S. **An introduction to forensic genetics**. John Wiley & Sons, 2010.

GUTIÉRREZ-HURTADO, I.A.; GARCÍA-ACÉVES, M.E.; PUGA-CARRILLO, Y.; GUARDADO-ESTRADA, M.; BECERRA-LOAIZA, D.S. et al. Past, Present and Future Perspectives of Forensic Genetics. **Biomolecules**. v.15, n. 5, p. 713, Mai. 2025.

GROSS, T.E.; FLECKHAUS, J.; SCHNEIDER, P.M. Progress in the implementation of massively parallel sequencing for forensic genetics: results of a European-wide survey among professional users. **Int J Legal Med**, v. 135, n. 4, p. 1425-1432, Jul 2021.

HUANG, J. et al. A novel method for the analysis of 20 multi-InDel polymorphisms and its forensic application: Nucleic acids. **Electrophoresis**, v. 35, n. 4, p. 487–493, fev. 2014.

JOBLING, M. A.; GILL, P. Encoded evidence: DNA in forensic analysis. **Nature Reviews Genetics**, v. 5, n. 10, p. 739–751, 1 out. 2004.

KIDD, K. K. et al. Developing a SNP panel for forensic identification of individuals. **Forensic Science International**, v. 164, n. 1, p. 20–32, dez. 2006.

KIDD, K. K. et al. Microhaplotype loci are a powerful new type of forensic marker. **Forensic Science International: Genetics Supplement Series**, v. 4, n. 1, p. e123–e124, 1 jan. 2013.

KIDD, K.K. ; SPEED, W.C. Criteria for selecting microhaplotypes: mixture detection and deconvolution. **Investig Genet**. v. 6, n. 1, 2015.

KIESLER, K.M.; BORSUK, L.A.; STEFFEN, C.R.; VALLONE, P.M.; GETTINGS, K.B. US Population Data for 94 Identity-Informative SNP Loci. **Genes (Basel)**, v.14, n.5, p.1071, mai. 2023.

KLOOSTERMAN, W.P. ; FRANCIOLI, L.C. ; HORMOZDIARI, F. ; MARSCHALL, T.; HEHIR-KWA, J.Y. ; ABDELLAOUI, A.; WENDL, M.C. Characteristics of de novo structural changes in the human genome. **Genome Res**. v.25, n.6, p.792-801, Jun. 2015.

KOBACHUK, L.D.G., CASTRO TEIXEIRA, C., EIDT FORNARI, P., MARINA SCHELLIN BECKER, C., DE MORAES BRAGA, A.C., ANDRESSA MACHADO TRENTIN LUZ, J., BRODZINSKI, J., MAIA TAULOIS DO ROSARIO, M., MAIA, R., MARTINS STUHLER, C., LEONARDO SILVA RODRIGUES, F. E MALAGHINI, M. Processamento de backlog de vestígios de crimes sexuais no Estado do Paraná. **Revista Brasileira de Criminalística**, v. 12, n.2, p. 55–60, 2023.

KOWALCZYK, M. et al. Molecular markers used in forensic genetics. **Medicine, Science, and the Law**, v. 58, n. 4, p. 201–209, out. 2018.

LAN, Q.; CAI, M.; LEI, F.; SHEN, C.; ZHU, B. Systematically exploring the performance of a self-developed Multi-InDel system in forensic identification, ancestry inference and genetic structure analysis of Chinese Manchu and Mongolian groups. **Forensic Sci Int.** v.346, p.111637, Mai. 2023.

LARUE, B.L.; GE, J.; KING, J.L.; BUDOWLE, B. A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification. **Int J Legal Med.** v. 126, n. 4, p. 533-40, jul. 2012.

LARUE, B. L. et al. Characterization of 114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification. **Legal Medicine (Tokyo, Japan)**, v. 16, n. 1, p. 26–32, jan. 2014.

LI, C. Forensic genetics. **Forensic Sciences Research**, v. 3, n. 2, p. 103–104, 2018.

LI, J. et al. An 18 Multi-InDels panel for analysis of highly degraded forensic biological samples. **Electrophoresis**, v. 42, n. 9–10, p. 1143–1152, may. 2021.

LIU, Z., et al. A set of 14 DIP-SNP markers to detect unbalanced DNA mixtures, **Biochemical and Biophysical Research Communications**, 497, p. 591-596, 2018.

LIU, J.; ZHANG, X.; ZHANG, X.; LI, W.; GAO, L.; LI, J.; WANG, J.; LIU, Z.; LIU, Y.; YAN, J.; ZHANG, G. A new set of 20 Multi-InDel markers for forensic application. **Electrophoresis**. v. 43, n. 11, p.1193-1202, Jun. 2022.

MATUTE, S. P.; IYAVOO, S. Implementation of NGS and SNP microarrays in routine forensic practice: opportunities and barriers. **BMC Genomics**. v.26, n.1, p. 541, mai 2025.

MILLS, R. E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. **Genome Res**, v.16, n. 9, p. 1182-90, set. 2006.

MINISTERIO DA JUSTIÇA E SEGURANÇA PÚBLICA. Procedimentos Operacionais Padrão - Perícia Criminal: genética forense. Disponível em: <<https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/analise-e-pesquisa/pop/pops-pericia-criminal-2024-genetica-forense-vol-4-pdf.pdf/view>>. Acesso em 11 de abr. 2025.

MEI, S.; YI, S.; CAI, M.; ZHANG, Y.; CUI, W.; XU, H.; LAN, Q.; ZHU, B. Exploring the forensic effectiveness and population genetic differentiation by self-constructed 41 multi-InDel panel in Yunnan Zhuang group. **Gene**. v.860, p.147180. Abr. 2023.

MONTGOMERY, S.B.; GOODE, D.L.; KVIKSTAD, E.; ALBERS, C.A.; ZHANG, Z. D.; MU, X.J. et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. **Genome Res**. v. 23, 749–61, 2013.

NASLAVSKY, M.S.; SCLIAR, M.O.; YAMAMOTO, G. L et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. **Nat Commun**. v. 13, p. 1004, 2022.

NUNES, K. et al. Admixture's impact on Brazilian population evolution and health. **Science**. v. 388(6748): ead13564, 2025.

OLDONI, F.; PODINI, D. Forensic molecular biomarkers for mixture analysis. **Forensic Science International. Genetics**, v. 41, p. 107–119, jul. 2019.

PALHA, T.J.B.F.; RODRIGUES, E.M.R.; CAVALCANTE, G.C.; MARRERO, A.; DE SOUZA, I.R.; SEKI UEHARA, C.J.; et al. Population genetic analysis of insertion-deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit. **Forensic Sci Int Genet**. v. 19, p.10-14, nov. 2015.

PENA, S. D. J. et al. The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. **PLoS ONE**, v. 6, n. 2, p. e17063, 16 fev. 2011.

PENA, S. D. J.; SANTOS, F. R.; TARAZONA-SANTOS, E. Genetic admixture in Brazil. **American Journal of Medical Genetics Part C: Seminars in Medical Genetics**, v. 184, n. 4, p. 928–938, dez. 2020.

PEREIRA, R. et al. A new multiplex for human identification using insertion/deletion polymorphisms. **Electrophoresis**, v. 30, n. 21, p. 3682–3690, nov. 2009.

PEREIRA, R. et al. Insertion/deletion polymorphisms: A multiplex assay and forensic applications. **Forensic Science International: Genetics Supplement Series**, v. 2, p. 513-515, 2009.

PEREIRA, R.; PHILLIPS, C.; PINTO, N.; SANTOS, C.; DOS SANTOS, S.E.; AMORIM, A.; CARRACEDO, Á.; GUSMÃO, L. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. **PLoS One**. v. 7, n. 1, p.e29684, 2012.

PEREIRA, R.; GUSMÃO, L. Capillary electrophoresis of 38 noncoding biallelic mini-InDels for degraded samples and as complementary tool in paternity testing. **Methods in Molecular Biology (Clifton, N.J.)**, v. 830, p. 141–157, 2012.

PHILLIPS C, AMIGO J, CARRACEDO Á, LAREU MV. Tetra-allelic SNPs: informative forensic markers compiled from public wholegenome sequence data. **Forensic Sci Int Genet**. v. 19, p.100–106, 2015.

REDE INTEGRADA DE BANCOS DE PERFIS GENÉTICOS (RIBPG). XXII RELATÓRIO DA REDE INTEGRADA DE BANCOS DE PERFIS GENÉTICO. Disponível em: <<https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/ribpg/relatorio/xxii-relatorio-da-rede-integrada-de-bancos-de-perfis-geneticos-maio-2025.pdf/view>>. Acesso em: 09 julho de 2025.

QU, S. et al. Multi-InDel: A Microhaplotype Marker Can Be Typed Using Capillary Electrophoresis Platforms. **Frontiers in Genetics**, v. 11, p. 567082, 23 out. 2020.

RICCI, U. et al. A Single Mutation in the FGA Locus Responsible for False Homozygosities and Discrepancies Between Commercial Kits in an Unusual Paternity Test Case. **Journal of Forensic Sciences**, v. 52, n. 2, p. 393–396, 2007. RODRIGUES, E. L., et al. Genética forense: manual de interpretação de resultados e estatística. 1. ed. Campinas, SP: **Millennium**, 2022. 432 p.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–5467, dez. 1977.

SANTOS, N.P.; RIBEIRO-RODRIGUES, E.M.; RIBEIRO-DOS-SANTOS, A.K.; PEREIRA, R.; GUSMÃO, L.; AMORIM, A.; GUERREIRO, J.F.; ZAGO, M.A.; MATTE, C.; HUTZ, M.H.; SANTOS, S.E. Assessing individual interethnic admixture and population substructure

using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. **Hum Mutat.** v. 31, n.2, p.184-90, Fev. 2010.

SCHEID, N. M. J., FERRARI, N., DELIZOICOV, D. A construção coletiva do conhecimento científico sobre a estrutura do DNA. **Ciencia & Educação**, v. 11, n 2, p. 223-233, 2005.

SCHNEIDER, P. M. Beyond STRs: The Role of Diallelic Markers in Forensic Genetics. **Transfusion Medicine and Hemotherapy: Offizielles Organ Der Deutschen Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie**, v. 39, n. 3, p. 176–180, jun. 2012.

SHI, Y.; NIU, Y.; ZHANG, P.; LUO, H.; LIU, S.; ZHANG, S.; WANG, J.; LI, Y.; LIU, X.; SONG, T.; XU, T.; HE, S. Characterization of genome-wide STR variation in 6487 human genomes. **Nat Commun.** v. 14, n.1, p. 2092, abr. 2023.

SONG, F.; LUO, H.B.; LI, J.N.; HOU, Y.P. Validation of a multiplex system with 20 multi-Indels for forensic purposes. **Forensic Science International: Genetics Supplement Series**, v.5, p.e48-e49, 2015.

SUN, K. et al. Multi-InDel Analysis for Ancestry Inference of Sub-Populations in China. **Scientific Reports**, v. 6, n. 1, p. 39797, dez. 2016.

SUN, K. et al. Evaluation of 12 Multi-InDel markers for forensic ancestry prediction in Asian populations. **Forensic Sci Int Genet.** v.43, p.102155, nov. 2019.

SUN, K. et al. Application of machine learning for ancestry inference using multi-InDel markers. **Forensic Science International: Genetics**, v. 59, 1 jul. 2022.

TAN, Y. et al. Two-person DNA mixture interpretation based on a novel set of SNP-STR markers. **Forensic Science International: Genetics**, v. 37, p. 37–45, 2018.

TAO, R. et al. Separation/extraction, detection, and interpretation of DNA mixtures in forensic science (review). **International Journal of Legal Medicine**, v. 132, n. 5, p. 1247–1261, set. 2018.

VALLE-SILVA, G. D. et al. Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing. **Forensic Science International: Genetics**, v. 40, p. 201–209, maio 2019.

WEBER, J. L. et al. Human Diallelic Insertion/Deletion Polymorphisms. **Am. J. Hum. Genet.**, v. 71, p. 854-862, 2002.

WEI, T. et al. A novel multiplex assay of SNP-STR markers for forensic purpose. **PLoS ONE**, v. 13, n. 7, jul. 2018.

YANG, Y.; XIE, B.; YAN, J. Application of Next-generation Sequencing Technology in Forensic Science. **Genomics, Proteomics & Bioinformatics**, v. 12, n. 5, p. 190–197, out. 2014.

WESTEN, A.A.; MATAI, A.S.; LAROS, J.F.J.; MEILAND, H.C.; JASPER, M.; DE LEEUW, W.J.F.; DE KNIJFF, P.; SIJEN, T. Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. **Forensic Sci Int Genet.** v. 3, p. 233–241, 2009.

YAO, Y.; SUN, K.; YANG, Q.; ZHOU, Z.; SHAO, C.; QIAN, X.; TANG, Q.; XIE, J. Assessing Autosomal InDel Loci With Multiple Insertions or Deletions of Random DNA Sequences in Human Genome. **Front Genet.** v. 12, 809815, Fev. 2022.

YAO, Y.; SUN, K.; YANG, Q.; ZHOU, Z.; QIAN, J.; LI, Z.; SHAO, C.; QIAN, X.; TANG, Q.; XIE, J. Development of a multiplex panel with 31 multi-allelic InDels for forensic DNA typing. **Int J Legal Med.** v. 137, n. 1, p. 1-12, Jan. 2023.

YAO, Y.; JI, Q.; LI, Z.; ZHOU, Z.; QIAN, J.; JI, L.; SUN, K.; XIE, J. Development of a 39 MM-InDel multiplex assay for the forensic application. **Electrophoresis.** v. 45, p.505-516. Mar. 2024.

ZHAO, X. Construction and forensic genetic characterization of 11 autosomal haplotypes consisting of 22 tri-allelic InDels. **Forensic Science International**, v. 34, p. 71-80, fev. 2018.

2. Capítulo II – Manuscrito I

****O Manuscrito I foi publicado na revista *Forensic Science International: Genetics*, porém não está no formato Open Access. Dessa forma, o conteúdo será apresentado aqui na forma de resumo expandido, em português.**

Citação: Ramos, L. C. et al. Assessing the performance of multi-InDel panels for human identification among admixed Brazilians. *Forensic Science International: Genetics* 74, 103161 (2025). doi: 10.1016/j.fsigen.2024.103161. PMID: 39418769.

**Título: Assessing the performance of multi-InDel panels for human identification
among admixed Brazilians**

Livia Carla Ramos^{a,b,c}, Luciellen D. G. Kobachuk^{d,e}, Douglas Mendes Nadur^a, Luiza Rauen Sabbag^e, Marianna Maia Taulois do Rosário^d, Michel S. Naslavsky^f, Celso Teixeira Mendes Junior^{g*}, Erick C. Castelli^{b,h*}

^a São Paulo State University (UNESP), Institute of Biosciences, Botucatu, São Paulo, Brazil.

^b Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit - Unipex, School of Medicine, São Paulo State University - UNESP, Botucatu, São Paulo, Brazil.

^c Superintendência da Polícia Técnico-Científica, Núcleo de Perícias Criminalísticas de Bauru, São Paulo, Brazil.

^d Polícia Científica do Paraná, Seção de Genética Molecular Forense, Curitiba, Paraná, Brazil.

^e Department of Genetics, School of Medicine, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil.

^f Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo (USP), São Paulo, São Paulo, Brazil.

^g Department of Chemistry, Forensic and Genomics Research Laboratory, Faculty of Philosophy, Science and Letters at Ribeirão Preto, University of São Paulo, 14040-901, Ribeirão Preto-SP, Brazil.

^h São Paulo State University (UNESP), Department of Pathology, School of Medicine, Botucatu, São Paulo, Brazil;

2.1. Introdução

Os laboratórios de genética forense lidam frequentemente com amostras biológicas altamente complexas, como aquelas contendo material genético degradado, misturas de perfis genéticos desbalanceados de dois ou mais contribuintes, baixa quantidade de DNA ou mesmo a combinação desses fatores [1]. Os polimorfismos STRs (do inglês, *short tandem repeats*) são considerados padrão-ouro para identificação humana. No entanto, os STRs apresentam limitações, especialmente em análises de amostras contendo misturas desbalanceadas de diferentes perfis genéticos. A amplificação por PCR (do inglês, *polymerase chain reaction*) dos *loci* STR pode resultar em erros caracterizados pela perda ou ganho de unidades repetitivas durante a síntese da nova fita de DNA, devido a um erro inato da enzima DNA polimerase. Como consequência, podem ser gerados picos de artefatos denominados *stutters*, que muitas vezes são indistinguíveis dos alelos verdadeiros de mesmo tamanho [2-3].

Conjuntos alternativos de marcadores genéticos, incluindo os SNPs (do inglês, *single nucleotide polymorphisms*) e os InDels (do inglês, *Insertion/deletion polymorphisms*), também denominados como polimorfismos de deleção/inserção (DIP), foram recentemente introduzidos como forma de complementar as análises de DNA [4-5]. Entretanto, tendo em vista o poder de discriminação limitado desses marcadores por serem em sua maioria bi-alelicos, foram propostos os marcadores denominados como biomarcadores compostos. Kidd e colaboradores propuseram os microhaplótipos definidos como fragmentos de DNA contendo dois ou mais SNPs localizados a uma distância molecular inferior a 200 pares de bases (pb), resultando em três ou mais combinações alélicas (haplótipos) de SNPs [6]. Atualmente, diversos estudos investigam a associação entre diferentes tipos de variantes, incluindo os haplótipos de InDels (multi-InDels) [7].

Os InDels estão amplamente distribuídos no genoma humano, representam o segundo tipo de polimorfismo de DNA mais abundante, e combinam características desejáveis dos SNPs e dos STRs [8-9]. Eles podem ser genotipados por PCR e eletroforese capilar (CE), as mesmas plataformas amplamente utilizadas para os STRs, além de permitirem a análise em ensaios multiplex. Também apresentam taxas de mutação mais baixas em comparação aos *loci* STR e podem gerar produtos de amplificação (*amplicons*) menores, assim como os SNPs. No entanto, a maioria dos InDels é bi-alelico e, portanto, menos informativa do que os STRs,

exigindo um número maior de *loci* para alcançar o mesmo nível de poder de discriminação [10-11]. Nesse contexto, os multi-InDels, definidos como haplótipos compostos por dois ou mais InDels localizados em um segmento curto de DNA, foram sugeridos como uma forma de superar a limitação de informatividade dos InDels, mantendo as características individuais desses polimorfismos [7].

Atualmente, os painéis multi-InDels propostos foram desenvolvidos e avaliados somente em populações asiáticas. A população brasileira é altamente heterogênea, resultado de séculos de miscigenação entre ameríndios, europeus, africanos e, mais recentemente, imigrantes de diversas origens [12]. Em razão da alta miscinegação, a população brasileira representa relevância para avaliar o desempenho de qualquer painel de marcadores para identificação humana. Portanto, este estudo teve como objetivo avaliar a aplicabilidade de três painéis de multi-InDels, desenvolvidos para populações asiáticas, para identificação humana no Brasil.

2.2. Material e Métodos

Foram selecionados três painéis de marcadores multi-InDels publicados entre 2014 e 2021, com aplicação para identificação humana. O primeiro painel (Painel 1) foi proposto por Huang et al. (2014) [7], contendo 20 marcadores multi-InDels que abrangem 43 *loci* InDel. O segundo painel (Painel 2) foi proposto por Qu et al. (2020) [13], com 18 multi-InDels que compreendem 54 *loci* InDel. O terceiro painel (Painel 3) foi proposto por Li et al. (2021) [14], composto por 18 multi-InDels que incluem 36 *loci* InDel.

No total, os três painéis abrangem 56 marcadores. No entanto, os marcadores D1P21 (Painel 1) e mh01zl002 (Painel 2) incluem os mesmos InDels, assim como os marcadores D9p23 (Painel 1) e D9MIM120 (Painel 3). Portanto, realizamos as análises considerando apenas 54 marcadores multi-InDels (Tabela 1).

Tabela 1. Lista dos marcadores multi-InDel avaliados em uma amostra da população brasileira.

No.	Chr.	Marcadores	SNPids	Amplicon (hg38)	Painel
1	1	D1P21*	rs67213950/rs55796544; rs71075445	99729244-99729371*	1
	1	mh01zl002*	rs67213950/rs55796544; rs67810269; rs71075445	99729275-99729381*	2
2	2	D2q36.3	rs200250620; rs10591193	229049600-229049753	1
3	3	D3p24.2	rs5847129; rs3073117	21734265-21734424	1
4	4	D4q22.1	rs33959224; rs10644346	90547094-90547226	1
5	6	D6q22.31	rs35354644; rs71797128	119137216-119137355	1
6	6	D6q25.1	rs57440464; rs71017565	152991066-152991207	1
7	7	D7p22.2	rs199716537; rs143204446	6471088-6471181	1
8	8	D8q12.3	rs141772481; rs200867035	66297219-66297346	1
9	8	D8q21.2	rs72068320; rs139730154	86457387-86457515	1
10	9	D9p23**	rs56168866; rs150820709	9246974-9247073**	1
	9	D9MIM120**	rs150820709; rs56168866	9246972-9247090**	3
11	11	D11q25	rs138543073; rs10669740	132609930-132610075	1
12	12	D12q24.31	rs62919960; rs62863861	125515853-125516025	1
13	13	D13q12	rs34660708; rs71709719	23391786-23391904	1
14	15	D15q25.2	rs36163330; rs60732398	86642556-86642664	1
15	16	D16p12.1	rs71272469; rs59157676; rs57953638	25853271-25853421	1
16	19	D19p13.2	rs71179859; rs71179860	8592113-8592230	1
17	20	D20p12.3	rs71183098; rs368791646; rs147797679	8548168-8548271	1
18	20	D20q11	rs11478629; rs4011949	34977017-34977184	1
19	21	D21q22.13	rs113244099; rs10582534	38296165-38296321	1
20	21	D21q22.3	rs67796568; rs201015636	45479051-45479219	1
21	1	mh01zl001	rs368828322; rs372567620; rs148361309	2098014-2098173	2
22	2	mh02zl001	rs142363578; rs144117237; rs148016741	30758890-30759106	2
23	2	mh02zl003	rs575990766; rs141442566; rs66617012	211296748-211296989	2
24	3	mh03zl001	rs34404453; rs149171688; rs34483288	73829814-73829939	2
25	3	mh03zl002	rs200679094; rs370444413; rs74604190	87303486-87303608	2
26	3	mh03zl003	rs80013016; rs372207681; rs111507567	163952714-163952912	2
27	4	mh04zl001	rs11282557; rs58595156; rs372089291	18389655-18389812	2
28	4	mh04zl003	rs66502037; rs77222977; rs71871946	186203031-186203187	2
29	7	mh07zl001	rs71053237; rs72447238; rs71053238	57255119-57255300	2
30	10	mh10zl001	rs145059123; rs539040996; rs34860860	7098216-7098439	2
31	10	mh10zl002	rs143378119; rs370632025; rs571358799	111822772-111822881	2
32	13	mh13zl001	rs58303500; rs34156563; rs67523118	24867756-24867979	2
33	18	mh18zl001	rs377195018; rs59925455; rs201823781	64005360-64005521	2
34	19	mh19zl001	rs138906215; rs35679623; rs11283323	490332-490490	2
35	19	mh19zl002	rs543507020; rs559033587; rs36127315	55527748-55527950	2
36	21	mh21zl001	rs562677243; rs576365249; rs373980639	19888920-19889086	2
37	21	mh21zl002	rs147144385; rs200218606; rs138205093	41012753-41012900	2
38	2	D2MIS16	rs200111128; rs72106284	190864783-190864875	3
39	2	D2MIM7	rs34271698; rs568009789	219363230-219363330	3
40	3	D3MIM9	rs10662420; rs145695868	143606952-143607064	3
41	4	D4MIM97	rs149811386; rs34874736	10119977-10120104	3
42	5	D5MIS23	rs58834048; rs10624797	83313873-83313974	3
43	6	D6MIS18	rs34510803; rs35277561	11671527-11671634	3
44	7	D7MIS24	rs71766343; rs56207509	30789143-30789236	3
45	8	D8MIS25	rs10660220; rs33917446	124738714-124738795	3
46	8	D8MIM57	rs373473214; rs71567919	116757215-116757309	3
47	9	D9MIM25	rs72491658; rs576969871	29262751-29262838	3
48	9	D9MIM121	rs10583220; rs5898502	76720079-76720194	3
49	12	D12MIS26	rs10685722; rs10685723	105045750-105045868	3
50	13	D13MIS27	rs34991318; rs3831371	24884030-24884157	3
51	13	D13MIL20	rs200068613; rs66616647	68389682-68389770	3
52	14	D14MIL25	rs10560257; rs141541522	98791592-98791717	3
53	15	D15MIM39	rs531528413; rs572157415	85762965-85763067	3
54	19	D19MIL37	rs147256339; rs148264147	20576043-20576161	3

Nota: (*/**) marcadores compostos pelos mesmos InDels. Painel 1: Huang et al. (2014) [7]; Painel 2: Qu et al. (2020) [13]; Painel 3: Li et al. (2021) [14].

Foram avaliados todos os 54 marcadores multi-InDels, utilizando dados públicos de sequenciamento do genoma completo de alta cobertura (30x) de 1.171 indivíduos idosos brasileiros. Esses indivíduos foram genotipados pelo estudo longitudinal SABE – Saúde, Bem-estar e Envelhecimento, realizado na cidade de São Paulo (EGA nº EGAD00001008640) [15].

Inicialmente, foi compilado um arquivo BED listando todas as regiões genômicas propostas pelos três estudos [7; 13; 14], considerando 2000 pares de base (pb) adicionais em cada extremidade. Posteriormente, recuperamos os dados de sequenciamento a partir dos arquivos BAM originais utilizando o Samtools. Utilizamos a ferramenta GATK *HaplotypeCaller* 4.2.6.1 para chamada das variantes e o WhatsHap 1.4 e Shapeit4 para fazer o faseamento dos genótipos, resultando em um VCF final contendo todas as variantes e haplótipos.

Em seguida, a partir do VCF, exportamos a sequência completa de ambos os cromossomos de cada indivíduo, usando a ferramenta *vcfx* fasta (www.castelli-lab.net/apps/vcfx), considerando as coordenadas genômicas de cada *amplicon*. Aplicamos um *script* em Perl para detectar o tamanho de cada sequência, o qual foi usado para designar cada haplótipo, registrando-os em um arquivo TXT no formato STRAF [16]. O *software* online STRAF 2.1.5 foi utilizado para calcular as frequências alélicas e os parâmetros forenses [16]. As análises de equilíbrio de Hardy-Weinberg (HWE) foram realizadas com o *software* Arlequin 3.5.2.2 [17]. O número efetivo de alelos (A_e) foi calculado com base na fórmula descrita por Kidd & Speed (2015) [18]. Utilizamos o programa IGV (do inglês, *Integrative Genome Viewer*) para avaliar manualmente cada região quanto à presença de sequências repetitivas, sequências de baixa complexidade, presença de STRs ou outras condições que possam tornar a região propensa à erros de amplificação.

Por fim, utilizamos os dados públicos de sequenciamento da fase III do Projeto 1000 Genomas para verificar a presença de polimorfismos abrangendo as regiões de cada *primer* proposto pelos painéis 1, 2 e 3.

2.3. Resultados e Discussões

As análises das regiões genômicas destacaram que 35 dos 54 marcadores (64,8%) estão localizados em regiões repetitivas, com presença de homopolímeros de adenina ou timina, ou

apresentam um STR, sendo, portanto, propensas a erros de amplificação. Algumas das regiões apresentam mais de um dos problemas citados anteriormente. Dessa forma, não foi possível detectar de forma confiável os genótipos de quatro desses marcadores, indicados como números 23, 24, 27 e 31 na Tabela 1 (mh02zl001, mh02zl003, mh03zl003 e mh10zl001). Sequências de baixa complexidade são propensas a erros de amplificação, em especial à produção de *amplicons* com ausência de alguns nucleotídeos em segmentos repetitivos [11;19], assemelhando-se aos produtos de *stutter* observados em STRs.

No Brasil, 21 dos marcadores propostos (38,9%) apresentam variantes na região de ligação dos *primers*, sendo que 10 (18,5%) deles possuem polimorfismos frequentes. Esses polimorfismos podem levar à perda de alelos (*drop-out*) devido à falha na hibridização e amplificação da sequência-alvo [1]. Nesses casos, recomenda-se redesenhar os *primers* previamente propostos.

A média e mediana de *Ae* para 50 *loci* foram 2,52 e 2,44, respectivamente, com nove marcadores apresentando *Ae* superior a 3. Nove *loci* apresentaram desvios significativos do equilíbrio de Hardy-Weinberg (HWE) após correção de Bonferroni para múltiplos testes ($p < 0,001$).

A probabilidade de *match* combinada (PMC) e o poder de exclusão combinado (PEC) obtidos em cada painel estão detalhados na Tabela 2. O painel 1, proposto por Huang et al., foi o mais informativo para os indivíduos brasileiros, com PMC de $4,92 \times 10^{-14}$ e PEC de 0,9991. No entanto, esse painel é menos informativo quando comparado ao STR CODIS e painéis maiores compostos por SNPs (Tabela 2). Além disso, mais de 50% dos marcadores do Painel 1 são propensos a erros de amplificação.

Tabela 2. Probabilidade de *match* combinada (PMC) e o poder de exclusão combinado (PEC) de diferentes conjuntos de marcadores em diferentes populações, incluindo a coorte brasileira SABE.

Tipo de marcadores	Conjunto de marcadores	Populações	PMC	PEC	Referência
Multi-InDel	20 Multi-InDel (Painel 1)	Brasil	$4,92 \times 10^{-14}$	0,9991	Presente Estudo
	14 Multi-InDel (Painel 2) *	Brasil	$2,07 \times 10^{-8}$	0,9790	Presente Estudo
	18 Multi-InDel (Painel 3)	Brasil	$7,33 \times 10^{-12}$	0,9973	Presente Estudo
	19 Multi-InDel (Painel 4)	Brasil	$1,43 \times 10^{-11}$	0,9964	Presente Estudo
	50 Multi-InDel (Painel 5)	Brasil	$1,67 \times 10^{-31}$	0,9999999032	Presente Estudo
	20 Multi-InDel (Painel 1)	Ásia	$5,7 \times 10^{-12}$	0,9989	[7]
	18 Multi-InDel (Painel 2)	Ásia	$2,77 \times 10^{-12}$	0,9984	[13]

	18 Multi-InDel (Painel 3)	Ásia	$3,23 \times 10^{-12}$	0,9989	[14]
	11 Multi-InDel	Ásia	$2,65 \times 10^{-9}$	0,996176	[11]
	Investigator® DIPplex kit (30 bi-allelic InDels)	Brasil	$3,4 \times 10^{-13}$	0,9973	[4]
	Investigator® DIPplex kit (30 bi-allelic InDels)	Japão	$2,67 \times 10^{-11}$	-	[20]
InDel	Investigator® DIPplex kit (30 bi-allelic InDels)	Polônia	$7,98 \times 10^{-14}$	-	[21]
	Investigator® DIPplex kit (30 bi-allelic InDels)	Taiwan	$1,22 \times 10^{-11}$	-	[21]
	38 InDels-plex	Angola, Moçambique, Portugal, Macau e Taiwan	$1,7 \times 10^{-14}$ a $3,6 \times 10^{-15}$	-	[22]
	CODIS 13	Caucasianos, Afro- americanos, Hispânicos e Asiáticos	$5,02 \times 10^{-16}$	-	[23]
	CODIS 20	Caucasianos, Afro- americanos, Hispânicos e Asiáticos	$9,35 \times 10^{-24}$	-	[23]
STR	20 STRs	Croácia	$3,71 \times 10^{-24}$	-	[24]
	21 STRs - GlobalFiler®	China	$7,69 \times 10^{-26}$	0.999999997	[25]
	21 STRs - GlobalFiler®	Brasil	2×10^{-27}	0.9999999990	[26]
SNP	SNPforID 52-plex (51 SNPs)	Brasil	$6,48 \times 10^{-21}$	0.9997	[27]

*Quatro dos 18 marcadores multi-InDels originais não foram avaliados nos dados de NGS porque coincidem com regiões repetitivas ou com regiões que contêm homopolímeros e/ou STRs.

Por fim, dois novos conjuntos foram organizados a partir dos marcadores analisados neste estudo. O painel 4 proposto é composto por 19 marcadores localizados em regiões não propensas a erros de amplificação e o painel 5 combina todos os 50 multi-InDels analisados (Tabela 2). Apesar do painel 4 eliminar possíveis erros de amplificação, sua capacidade informativa foi inferior ao dos demais painéis. Os resultados obtidos com o Painel 5 (união de todos os marcadores dos painéis 1, 2 e 3) resultaram em um aumento considerável nos parâmetros forenses. No entanto, sua avaliação por métodos baseados em EC é desafiadora ou até mesmo inviável, exigindo abordagens baseadas em sequenciamento de nova geração (NGS) [28-29].

Portanto, é necessário identificar novos marcadores com maior número de alelos distintos e frequentes, que não sejam propensos a erros de amplificação e que possam ser aplicados em diferentes populações.

2.4. Conclusão

Foi avaliada a aplicabilidade de três painéis de multi-InDels para o seu potencial uso na identificação humana no Brasil. Os resultados mostraram que a maioria dos marcadores propostos é informativa para indivíduos brasileiros, pois apresentam mais de três haplótipos frequentes. Entre os painéis testados, o de Huang et al. (painel 1) é o mais promissor para o Brasil, com valores de PMC e PEC de $4,92 \times 10^{-14}$ e 0,9991, respectivamente. No entanto, esses valores são inferiores aos obtidos com os STRs CODIS e com painéis maiores de SNPs. Além disso, parte dos marcadores são propensos a erros de amplificação devido à presença de regiões repetitivas. Portanto, novas tentativas de busca de marcadores multi-InDels polimórficos ainda são necessárias para desenvolver um painel de multi-InDels adequado para populações ao redor do mundo.

2.5. Referências

- [1] J.M. Butler. Fundamentals of forensic DNA typing. Amsterdam; Boston: Academic Press/Elsevier, 2010.
- [2] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, *Forensic Sci. Int* 160 (2006) 90–101, <https://doi.org/10.1016/j.forsciint.2006.04.009>.
- [3] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-Twenty years of research and development, *Forensic Sci. Int Genet* 18 (2015) 100–117, <https://doi.org/10.1016/j.fsigen.2015.03.014>.
- [4] T.J.B. Ferreira Palha, E.M. Ribeiro Rodrigues, G.C. Cavalcante, A. Marrero, I.R. de Souza, C.J. Seki Uehara, et al., Population genetic analysis of insertion-deletion polymorphisms in a

Brazilian population using the Investigator DIPplex kit, *Forensic Sci. Int. Genet.* 19 (2015) 10–14, <https://doi.org/10.1016/j.fsigen.2015.03.015>.

[5] G.D. Valle-Silva, F.D.N. Souza, L. Marcorin, et al., Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing, *Forensic Sci. Int. Genet.* 40 (2019) 201–209, <https://doi.org/10.1016/j.fsigen.2019.03.003>.

[6] K.K. Kidd, A.J. Pakstis, W.C. Speed, et al., Microhaplotype loci are a powerful new type of forensic marker, *Forensic Sci. Int. Genet. Suppl. Ser.* 4 (2013) e123–e124, <https://doi.org/10.1016/j.fsigss.2013.10.063>.

[7] J. Huang, H. Luo, W. Wei, Y. Hou. A novel method for the analysis of 20 multi-Indel polymorphisms and its forensic application. *Electrophoresis.* 35 (2014) 487–493. <https://doi.org/10.1002/elps.201300346>.

[8] J.L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, G. Marth, Human diallelic insertion/deletion polymorphisms, *Am. J. Hum. Genet.* 71 (2002) 854–862, <https://doi.org/10.1086/342727>.

[9] R.E. Mills, C.T. Luttig, C.E. Larkins, A. Beauchamp, C. Tsui, W.S. Pittard, S. E. Devine, An initial map of insertion and deletion (INDEL) variation in the human genome, *Genome Res* 16 (2006) 1182–1190, <https://doi.org/10.1101/gr.4565806>.

[10] T.Z. Gao, L.B. Yun, W. He, Y. Gu, Y.P. Hou, The application of multi-InDel as supplementary in paternity cases with STR mutation, *Foren. Sci. Intern. Genet. Suppl. Ser.* 5 (2015) e218–e219, <https://doi.org/10.1016/j.fsigss.2015.09.087>.

[11] X.H. Zhao, X.G. Chen, Y.C. Zhao, S. Zhang, Z.H. Gao, Y.W. Yang, et al., Construction and forensic genetic characterization of 11 autosomal haplotypes consisting of 22 tri-allelic indels, *Foren. Sci. Int. Genet.* 34 (2018) 71–80, <https://doi.org/10.1016/j.fsigen.2018.02.001>.

[12] S.D.J. Pena, G. Di Pietro, M. Fuchshuber-Moraes, J.P. Genro, M.H. Hutz, The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected, *PLoS One* 6 (2011) e17063, <https://doi.org/10.1371/journal.pone.0017063>.

- [13] S. Qu, M. Lv, J. Xue, J. Zhu, L. Wang, H. Jian, et al. Multi-Indel: a microhaplotype marker can be typed using capillary electrophoresis platforms. *Front Genet.* 11 (2020) 567082. <https://doi.org/10.3389/fgene.2020.567082>.
- [14] J. Li, L. Lin, B. Jiang, C. Wang, M.M.J. Zeye, D. Wen, et al. An 18 Multi-InDels panel for analysis of highly degraded forensic biological samples. *Electrophoresis* 42 (2021) 1143–52. <https://doi.org/10.1002/elps.202000245>.
- [15] M. S. Naslavsky, M. O. Scliar, G. L. Yamamoto, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. *Nat Commun* **13** (2022) 1004. <https://doi.org/10.1038/s41467-022-28648-3>.
- [16] A. Gouy, M. Zieger. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet.* 30 (2017) 148-151. <https://doi.org/10.1016/j.fsigen.2017.07.007>.
- [17] L. Excoffier, H. E. Lischer. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10 (2010) 564-7. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>.
- [18] K.K. Kidd, W.C Speed. Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investig Genet.* 6 (2015). <https://doi.org/10.1186/s13323-014-0018-3>
- [19] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, R. Durbin. Dindel: accurate indel calls from short-read data. *Genome Res.* 21 (2011) 961-73. <https://doi.org/10.1101/gr.112326.110>.
- [20] M. Nunotani, T. Shiozaki, N. Sato, S. Kamei, K. Takatsu, T. Hayashi, et al., Analysis of 30 insertion-deletion polymorphisms in the Japanese population using the Investigator DIPplex(R) kit, *Leg. Med. Tokyo (Tokyo)* 17 (2015) 467–470. <https://doi.org/10.1016/j.legalmed.2015.10.002>.

- [21] W. Pepinski, M. Abreu-Glowacka, M. Koralewska-Kordel, E. Michalak, K. Kordel, A. Niemcunowicz-Janica, et al., Population genetics of 30 INDELs in populations of Poland and Taiwan, *Mol. Biol. Rep.* 40 (2013) 4333–4338. <https://doi.org/10.1007/s11033-013-2521-7>.
- [22] R. Pereira, C. Phillips, C. Alves, A. Amorim, A. Carracedo, L. Gusmão. A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis*. 30 (2009) 3682-90. <https://doi.org/10.1002/elps.200900274>.
- [23] J. M. Butler, C. R. Hill, M. D. Coble. Variability of New STR Loci and Kits in US Population Groups, 2012. <https://www.promega.com.br/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/> (accessed 01 Apr. 2024).
- [24] L. Barbaric, P. Ozretic, I. Horjan, M. Korolija, G. Mrsic, Forensic evaluation of the 20 STR loci in the population of Croatia, *Forensic Sci. Int. Genet.* 28 (2017) e49–e50. <https://doi.org/10.1016/j.fsigen.2017.03.011>.
- [25] L. Wu, B. Pei, P. Ran, X. Song, Population genetic analysis of Xiamen Han population on 21 short tandem repeat loci, *Leg. Med. Tokyo (Tokyo)* 26 (2017) 41–44. <https://doi.org/10.1016/j.legalmed.2017.03.002>.
- [26] C. B. Moysés, W. M. Tsutsumida, P. E. Raimann, C. H. da Motta, et al. Population data of the 21 autosomal STRs included in the GlobalFiler® kits in population samples from five Brazilian regions. *Forensic Sci Int Genet.* 26 (2017) e28–e30. <https://doi.org/10.1016/j.fsigen.2016.10.017>.
- [27] G.D. Valle-Silva, F. D. N. Souza, L. Marcorin, et al. Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing. *Forensic Sci Int Genet.* 40 (2019) 201-209. <https://doi.org/10.1016/j.fsigen.2019.03.003>
- [28] Y. Yang, B. Xie, J. Yan. Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics.* 12 (2014) 190-7. <https://doi.org/10.1016/j.gpb.2014.09.001>.

- [29] B. Bruijns, R. Tiggelaar, H. Gardeniers, Massively parallel sequencing techniques for forensics: a review, *Electrophoresis* (2018). <https://doi.org/10.1002/elps.201800082>.

3. *Capítulo III – Manuscrito II*

Title: A new set of multi-InDel markers with potential for human identification across different populations

Livia Carla Ramos ^{a,b,c}, Luiza Rauen Sabbag^d, Joao Victor Hamamura Toloi ^{a,b}, Luciellen D. G. Kobachuk ^{d,e}, Erick C. Castelli ^{b,f*}

^a São Paulo State University (UNESP), Institute of Biosciences, Botucatu, São Paulo, Brazil.

^b Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit - Unipex, School of Medicine, São Paulo State University - UNESP, Botucatu, São Paulo, Brazil.

^c Superintendência da Polícia Técnico-Científica, Núcleo de Perícias Criminalísticas de Bauru, São Paulo, Brazil.

^d Department of Genetics, School of Medicine, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil.

^e Polícia Científica do Paraná, Seção de Genética Molecular Forense, Curitiba, Paraná, Brazil.

^f São Paulo State University (UNESP), Department of Pathology, School of Medicine, Botucatu, São Paulo, Brazil;

3.1. Abstract

Multi-InDels (Insertion/Deletion) or InDel microhaplotypes consist of two or more InDels within a small genomic region forming three or more alleles with different sizes, increasing the informativity of a single InDel locus. Multi-InDel panels are becoming ideal complementary markers to STRs in forensic science for human identification, paternity testing, and ancestry inference due to their advantages over other markers, including short amplicon sizes, low mutation rates, the absence of stutter artifacts, and the ability to be amplified and detected via Polymerase Chain Reaction (PCR) and capillary electrophoresis (CE). However, developing a panel that is efficient in different populations remains a major challenge. In this study, we proposed a multiplex assay for human identification based on a CE platform, comprising 28 multi-InDel and two sex identification markers. We selected the markers in two steps. First, we selected candidates using the high-coverage whole-genome sequencing data from the 1000 Genomes Project (30x), with two or more InDels in a short segment, and producing more than 3 alleles with different sizes. After, we tested these candidates using 30x whole-genome sequencing data from 1,171 Brazilians, to select the best candidates for Brazil and other populations, with 94.4% of the amplicons shorter than 250bp. Using the 1000 Genomes dataset, we evaluated the polymorphism and forensic parameters of the panel across different populations worldwide. The panel showed high polymorphism among Europeans, Africans, East and South Asians, and Americans. The combined match probability (CMP) and combined power of exclusion (CPE) were 5.66×10^{-20} to 2.66×10^{-17} , and 0.99958 to 0.99998, respectively. Therefore, this system has the potential as a complementary tool for individual identification across worldwide populations, especially for challenging samples.

Keywords: *forensic genetics, microhaplotype, multi-InDel, capillary electrophoresis*

3.2. Introduction

Multi-InDels consist of two or more InDels (Insertion/Deletion polymorphisms) within a small genomic region, forming microhaplotypes, which can increase the informativeness of individual InDels [1]. They have some combined advantages over other genetic markers: similar to STRs (short tandem repeat), they can be detected by capillary electrophoresis (CE) platforms, which is present in most forensic DNA laboratories [1-2]; similar to SNPs, they can be amplified in short amplicons, which is more suitable for degraded samples [3]; the allelic mutation rate is low when compared to STRs, assuring a more stable inheritance in paternity tests [4-6]; multi-InDels that do not include repetitive sequences result in enhanced genotyping accuracy and avoid artifacts called stutters, which is a benefit when analyzing DNA mixtures [7;1]; and they may be used to infer individual ancestry when presenting high inter-population variability [8].

Multi-InDel are becoming ideal complementary markers in forensic sciences for human identification, paternity tests, and ancestry inference [8-10]. Nonetheless, most panels have only been proposed and analyzed in populations from Asia. Recently, we analyzed the applicability in Brazil of three previously proposed multi-InDel panels, demonstrating that, although most markers are informative, with more than three frequent haplotypes, 64.8% of them can potentially present amplification issues due to the presence of repetitive sequences, homopolymers, low-complexity sequences, and STR-like sequences [11].

InDels with random DNA sequences seem to rarely occur in the human genome, present a low mutation rate, and usually imply a unique mutation event across human history [12]. Considering the rarity of InDels with random DNA sequences and the variability of allele frequencies across populations, the development of a robust and broadly applicable multiplex panel of multi-InDel markers remains a major challenge in forensic genetics. In response to this need, we propose a novel multiplex set comprising 28 multi-InDels for human identification and two markers for sex determination, that could be assembled for use in a CE-based platform. These markers were selected evaluating high-coverage (30x) whole-genome sequencing data from the 1000 Genomes Project and the SABE Brazilian cohort [13-14]. Finally, we evaluated the polymorphism and forensic parameters of the panel across different populations worldwide.

3.3. Materials and methods

The first step in the marker selection process was retrieving adjacent InDels forming microhaplotypes from the 1000 Genomes phase 3 public VCF files (available at <https://www.internationalgenome.org>). The criteria for the InDel selection were: they must be located on an autosome, a pair must present a physical distance of 50 bp or less, all the InDels must present a minor allele frequency higher than 0.05, and the region should not present low-complexity sequences (e.g., STRs and homopolymers). Since the 1000 Genomes phase 3 data is phased, in the next step, for each microhaplotype candidate, we generated complete sequences (one per chromosome) considering the adjacent InDels and 10 nucleotides upstream and downstream them, using the *vcfx* fasta program (www.castelli-lab.net/apps/vcfx). With these data, we calculated the size of each haplotype (two per individual).

The 1000 Genomes phase 3 project has 2,504 individuals from 26 populations from Africa, Europe, America, East Asia, and South Asia [13]. The frequency of each haplotype among these five groups was calculated, and only markers with three or more valid alleles (different sizes) that present a minimum allele frequency of 0.05 were further considered.

After filtering out candidates as described above, we manually inspected each candidate marker using IGV (Integrative Genomics Viewer), searching for repetitive and low-complexity regions. Therefore, we removed from the candidate list multi-InDels in regions presenting STRs, homopolymers, or any sort of repetitive sequence, maintaining only markers related to random insertion/deletion sequences positioned within a random sequence region.

Next, we designed primers for each multi-InDel using Primer3 and Primer-BLAST, aiming for amplicons ranging from 50 to 150bp. Only primer binding sites without any known SNP or InDel with a minor allele frequency of 0.5% were selected. We also used Primer-BLAST to check for nonspecific hybridization in other genome regions. Potential interactions between primers were examined using AutoDimer software [15]. Some candidates were disregarded because we could not design a specific pair of primers producing an amplicon with less than 200 nucleotides. The outcome was 150 potential candidate markers.

After the initial selection process, we validated the presence of the microhaplotypes by evaluating the more recent high-coverage whole-genome sequencing data (30x) from the 1000 Genomes dataset, which comprises 3,202 samples (1,598 males and 1,604 females), including 602 complete trios, from 26 populations (Table 1) [16]. In addition, we also analyzed the same set of markers in Brazilians from the high-coverage whole-genome sequencing data (30x) cohort known as SABE, with 1,171 elderly Brazilians from the city of São Paulo, EGA number EGAD00001008640 [14]. To do that, we compiled a BED file listing the genomic regions of the 150 remaining candidates, considering an additional 2,000 nucleotides in each direction. Then, we retrieved the sequencing data from the original BAM files using Samtools. To call variants, we used GATK HaplotypeCaller 4.2.6.1 in the GVCF mode, using default parameters, combining all the GVCF files into a multi-sample GVCF using GATK GenomicsDBImport (each chromosome separately). After, we called the genotypes using GATK GenotypeGVCFs.

We used WhatsHap 1.4 to phase variants directly from the sequencing reads. It should be noted that WhatsHap does not phase multi-allelic variants, and some of the variants in our dataset are indeed multi-allelic. To maximize the number of phased variants, including the multi-allelic ones, we created a VCF for each sample, removing alternative alleles that do not exist for that sample and applying WhatsHap afterward. After, we combined all phased VCFs in a multi-sample VCF file. This strategy directly phased the multi-allelic variants, except when the sample was heterozygous for two alternative alleles in a given locus. We inferred the complete haplotypes using shapeit4, which considers the phasing sets detected by WhatsHap. The outcome is a phased VCF, one per chromosome.

We exported the complete sequences (both chromosomes) of each multi-InDel for every individual using *vcfx* fasta (www.castelli-lab.net/apps/vcfx) considering the genomic coordinates of the amplicons (the coordinates of the primer pair). Subsequently, we used a local Perl script to detect each sequence's size, which was used to designate the haplotype, and report them in a TXT file in the STRAF format [17]. The STRAF 2.1.5 online software was used to calculate allele frequencies and the forensic parameters for five different superpopulations and the admixed Brazilians [17]. Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) analyses (r^2) were performed with the Arlequin 3.5.2.2 software [18].

The effective number of alleles (A_e) was calculated based on the formula described by Kidd and Speed (2015) [19]. The data was processed using the Rstudio IDE (Ocean Storm version) for R language v4.4.1 [20-21]. Visual representations were generated using the ggplot2 v3.5.1 package from the tidyverse metapackage [22].

Table 1. Populations of the 1000 Genome project and their sample sizes.

Population symbol	Sample Size	Population name	Population	Superpopulation Group
ESN	149	ESN	Esan, Nigeria	African
GWD	178	GWD	Gambian in Western Divisions in the Gambia	African
LWK	99	LWK	Luhya in Webuye, Kenya	African
MSL	99	MSL	Mende in Sierra Leone	African
YRI	178	YRI	Yoruba in Ibadan, Nigeria	African
ACB	116	ACB	African Caribbeans in Barbados	America
ASW	74	ASW	Americans of African Ancestry in SW US	America
CLM	132	CLM	Colombians from Medellin, Colombia	America
MXL	97	MXL	Mexican Ancestry from Los Angeles USA, California	America
PEL	122	PEL	Peruvians from Lima, Peru	America
PUR	139	PUR	Puerto Ricans from Puerto Rico	America
CDX	93	CDX	Chinese Dai in Xishuangbanna, China	East Asia
CHB	103	CHB	Han Chinese in Beijing, China	East Asia
CHS	163	CHS	Southern Han Chinese	East Asia
JPT	104	JPT	Japanese in Tokyo, Japan	East Asia
KHV	122	KHV	Kinh in Ho Chi Minh City, Vietnam	East Asia
CEU	179	CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	Europe
FIN	99	FIN	Finnish, Finland	Europe
GBR	91	GBR	British in England and Scotland	Europe
IBS	157	IBS	Iberian Population in Spain	Europe
TSI	107	TSI	Toscans in Italy	Europe
BEB	131	BEB	Bengali from Bangladesh	South Asia
ITU	107	ITU	Indian Telugu in the UK	South Asia
GIH	103	GIH	Gujarati Indian from Houston, Texas	South Asia

PJL	146	PJL	Punjabi from Lahore, Paskistan	South Asia
STU	114	STU	Sri Lankan Tamil in the UK	South Asia

3.4. Results

3.4.1. The final set of multi-InDel markers

We screened multi-InDel candidates from the 1000 Genomes phase 3 data, detecting 4,018 potential candidates. After filtering out candidates that do not present at least 3 frequent alleles and present repetitive sequences, we designed primers for 150 candidates. Then, we genotyped all these candidates in the 1000 Genomes and SABE/Brazil high-coverage whole-genome sequencing data, detecting the haplotypes using a combination of read-aware and statistical phasing. After looking at allele frequencies with STRAF and LD between multi-InDels with Arlequin, we chose 28 independent multi-Indel markers that cover 60 InDel locations across 21 autosomal chromosomes, leaving out chromosome 16 (Table 2). Only multi-InDels with allele frequencies higher than 0.05 for at least three different sizes were included. In addition, we also included two genetic markers for sex identification: amelogenin and SRY loci. The amelogenin primer sets are designed by Sullivan et al., producing AMELX/AMELY amplicons of 106/112 bp [23] and SRY gene was detected using specific primers that generate a 197 bp PCR product [24].

We have withheld the positions of the InDels and the primer sequences due to the patentable nature of this study. A specific nomenclature system for multi-InDel markers has not been standardized. Because they are microhaplotypes, we named them as suggested by Kidd (2016) [25]. The nomenclature begins with the preface “mh” or “MH”, followed by the chromosome number represented with two digits, a laboratory designation consisting of 2 to 4 characters unique to each Lab, and a specific number is chosen by the Lab to be unique to the Lab and chromosome combination [25].

Table 2. General information of the 28 multi-InDel markers.

Markers	Chr.	N° of InDels	Deletion/Insertion	Amplicon size in bp (Alleles)	N° of Alleles (freq. > 0.05)
MH01GB001	1	2	G / GTATC T / TC	113,114,115,116,117,119	3
MH02GB001	2	2	T / TGA A / AT	160,161,162,163	3
MH03GB001	3	2	G / GCTCACAGCTAATA G / GC	71,84,85,97	3
MH04GB001	4	2	T / TTGC ACT / ACTCT	181,182,183,185,186,188,189	3
MH05GB001	5	2	T / TG T / TTAAG	133,136,137	3
MH06GB001	6	2	A / AT A / AGTTGAC	120,126,127	3
MH07GB001	7	2	A / AT A / ACTC	93,95,96	3
MH08GB001	8	2	G / GA G / GATAATTTGTTC	191,202,203	3
MH08GB002	8	2	A / AT G / GTT	119,120,121	3
MH08GB003	8	2	T / TTAG T / TATAG	91,92,93,96	3
MH09GB001	9	2	G / GTAGA A / ATCT	76,77,80	3
MH09GB002	9	2	A / AG G / GCTCC	126,129,130,133,134	4
MH10GB001	10	2	T / TTTCCCAATA A / AC	113,114,122,123	3
MH11GB001	11	2	G / GTATTGGAGAC T / TG	137,138,148	3
MH11GB002	11	3	T / TG A / ATAGAT T / TA	172,173,174,179	4
MH12GB001	12	2	C / CCTT T / TC	129,131,132	3
MH12GB002	12	2	A / ATTTGCTCT C / CTATAT	242,245,250	3
MH13GB001	13	2	C / CATTT A / AT	125,129,130	3
MH14GB001	14	2	C / CAA T / TA	83,84,85	3
MH15GB001	15	2	G / GATC C / CA	175,178,179	3
MH17GB001	17	2	C / CTG A / AG	161,162,163,164	3

MH17GB002	17	3	C / CA G / GA A / ACTGCCCTGGGCGCC	86,87,100,101,102	4
MH18GB001	18	2	C / CTACTAAGGCGCACACAGG G / GC	54,72,73	3
MH19GB001	19	3	A / AGGACGAGCAGGATT - / CAGGATTGGAAGCG G / GC	146,188,189,190,202,203	4
MH20GB001	20	2	A / AGTT C / CCAGA	231,232,235	3
MH20GB002	20	2	A / AGAGT T / TC	189,193,194	3
MH21GB001	21	2	C / CTA C / CAGGATCTCTTTTA	91,93,104	3
MH22GB001	22	3	A / ACATACCTAAT A / AAAATAT T / TCAGCATGA	289,297,299,301,306,307	4

Note: Alleles with frequencies > 0.05 shown in bold.

3.4.2. Allele frequencies

We calculated the allele frequencies of the 28 markers across five different biogeographic regions, Africa (AFR), America (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS), as well as among admixed Brazilian from the SABE cohort (Naslavsky et al., 2022). For the 1000 Genomes data, we detected 107 specific amplicons with an average length of 149 bp (ranging from 54 to 307 bp), of which 85.05% (91 alleles) were shorter than 200 bp and 94.4% (101 alleles) were shorter than 250 bp. For Brazilians, we detected 100 specific amplicons with an average length of 148 bp (54–307 bp), of which 86% (86 alleles) were shorter than 200 bp and 96% were shorter than 250 bp. Only one marker, MH22GB001, exhibited alleles exceeding 250 bp. Altogether, at least three haplotypes (alleles) of varying sizes with frequencies greater than 0.05 were identified per marker when considering all groups combined (Table 2). Figure 1 illustrates the alleles of the 28 multi-InDel markers and their frequencies, represented by colors across different groups, while Supplementary Table S1 provides a detailed description of all frequencies.

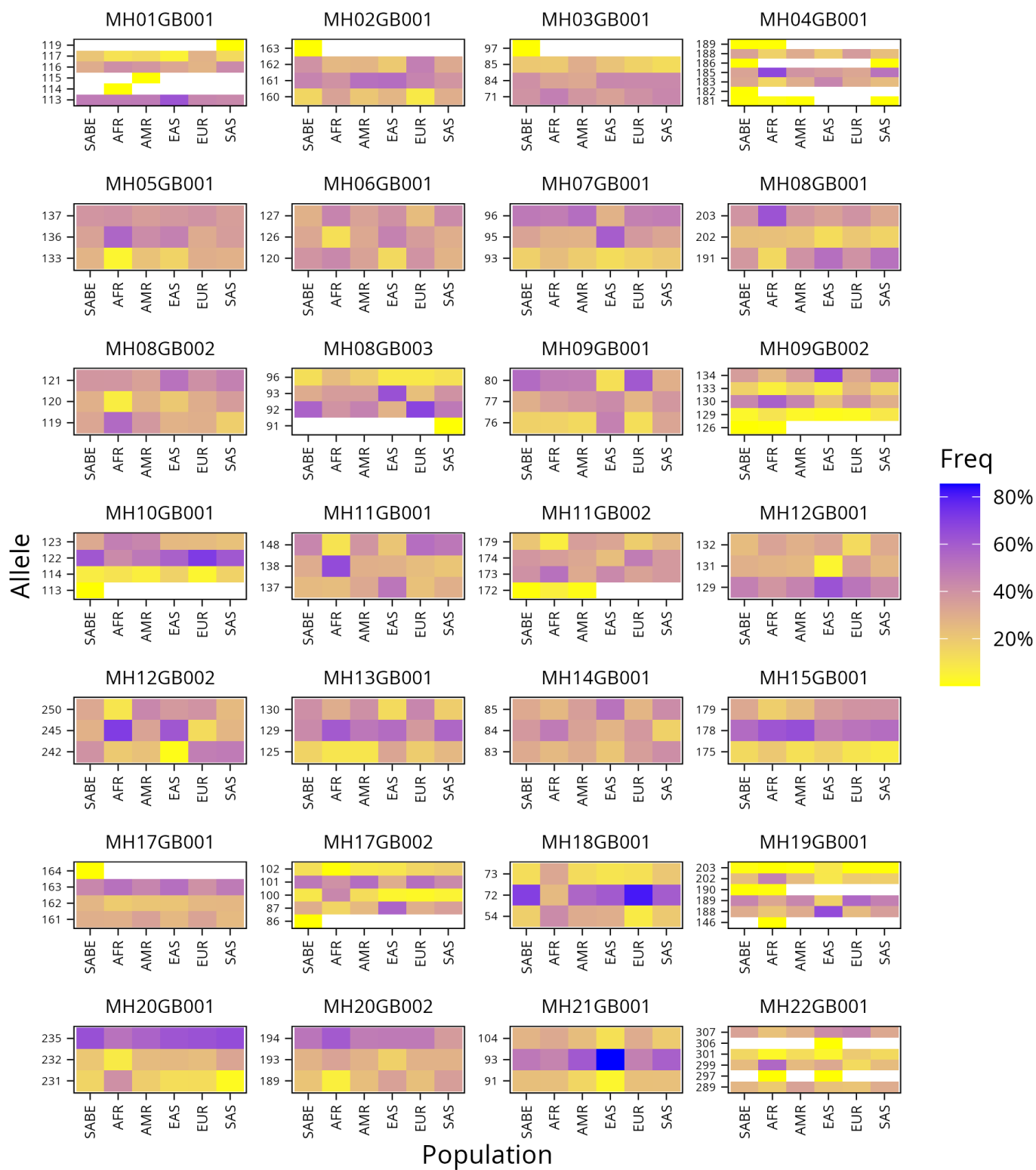


Figure 1. Heatmap of allelic frequencies of 28 multi-InDel markers in Brazil (SABE), Africa (AFR), America without Brazil (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Colors from white to blue represent allele frequencies. The size (bp) of each haplotype is on the y-axis.

3.4.3. Forensic parameters and statistical analysis

We estimated the forensic parameters for the 28 multi-InDel markers across all groups (Figure 2, Table 3, Supplementary Table S2–S7). The median and mean values of effective number of alleles (A_e) in the Brazilian sample (SABE) were 2.79 and 2.71, respectively. This was followed by AMR, with values of 2.76 and 2.71; SAS, with 2.67 and 2.68; EUR, with 2.65 and 2.58; AFR, with 2.46 for both median and mean; and EAS, with 2.44 and 2.37 (Figure 2 and Table 3).

The mean values of expected heterozygosity (H_e) and observed heterozygosity (H_o) across all groups exceeded 0.56, while the median power of discrimination (PD) in these groups was above 0.73. The highest H_e and PD values were observed in the SABE and AMR groups (Figure 2 and Table 3).

We also calculated the combined match probability (CMP) and cumulative probability of exclusion (CPE) for each group and population (Tables 4 and 5). The values of CMP ranged from 5.66×10^{-20} (AMR) to 2.66×10^{-17} (EAS). The values of CPE ranged from 0.99958 (EAS) to 0.99998 (SABE). In the Brazilian sample (SABE), the values of CMP and CPE were 1.03×10^{-19} and 0.99998, respectively.

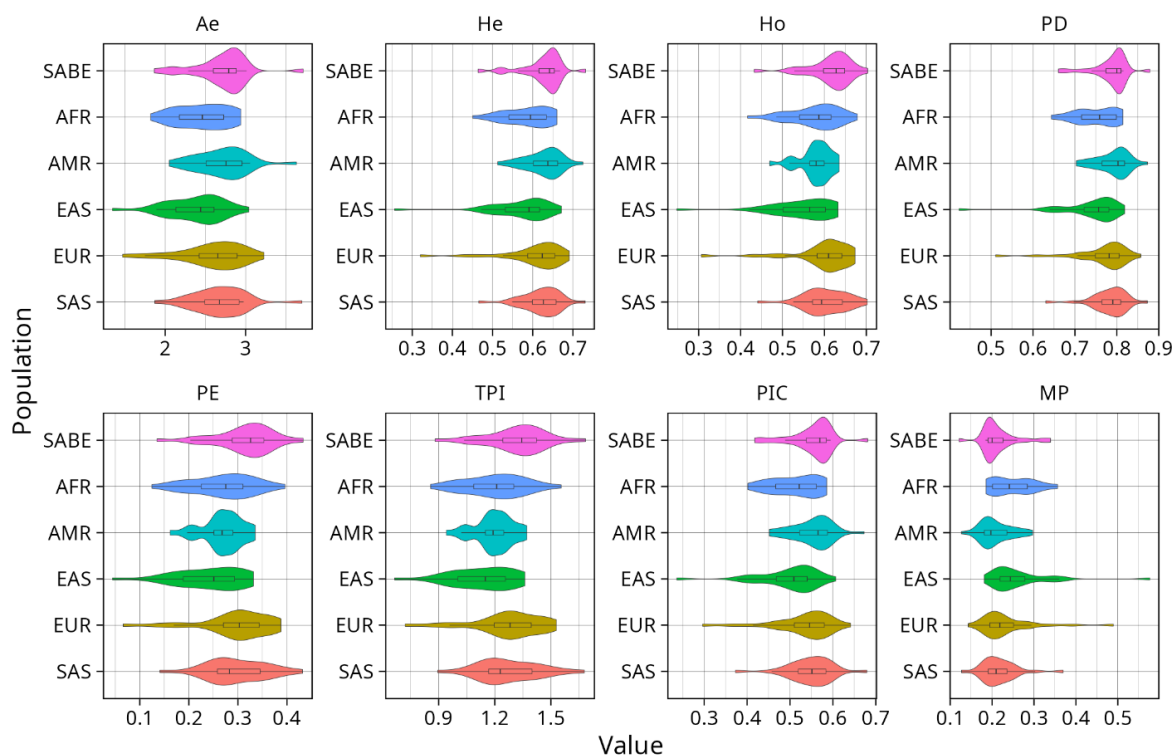


Figure 2. Violin plots of forensic parameters for the 28 multi-InDel markers in different groups. Brazilians (SABE), Africans (AFR), Americans – without Brazil (AMR), East Asians (EAS), Europeans (EUR), and South Asians (SAS). Ae: effective number of alleles; He: expected heterozygosity; Ho: observed heterozygosity; PD: power of discrimination; PE: power of exclusion; TPI: typical paternity index; PIC: polymorphism information content; MP: match probability.

Table 3. Comparison of the mean values of forensic parameters across different groups using 28 new multi-InDel markers.

Group	Ae*	He	Ho	PD	PE	TPI	PIC	MP
SABE/Brazil	2.71	0.6249	0.6135	0.7862	0.3124	1.3166	0.5536	0.2138
Africans	2.46	0.5861	0.5766	0.7497	0.2696	1.2033	0.5119	0.2503
Americans *	2.71	0.6262	0.5786	0.7912	0.2685	1.1960	0.5528	0.2088
East Asians	2.37	0.5677	0.5449	0.7355	0.2394	1.1271	0.4957	0.2645
Europeans	2.58	0.6008	0.5930	0.7631	0.2921	1.2631	0.5293	0.2369
South Asians	2.68	0.6207	0.6008	0.7822	0.2966	1.2740	0.5470	0.2178

* Americans without Brazil

Ae: effective number of alleles; He: expected heterozygosity; Ho: observed heterozygosity; PD: power of discrimination; PE: power of exclusion; TPI: typical paternity index; PIC: polymorphism information content; MP: match probability.

Table 4. Combined match probability (CMP) and cumulative probability of exclusion (CPE) results for different groups using 28 new multi-InDel markers.

Group	CMP	CPE
SABE/Brazilians	1.03×10^{-19}	0.99998
Africans	8.45×10^{-18}	0.99986
Americans *	5.66×10^{-20}	0.99985
East Asians	2.66×10^{-17}	0.99958
Europeans	1.15×10^{-18}	0.99995
South Asians	1.70×10^{-19}	0.99995

* Americans without Brazil

Table 5. Combined match probability (CMP) and cumulative probability of exclusion (CPE) results for individual populations using 28 new multi-InDel markers.

Group	Population	CMP	CPE
SABE/Brazilians	BRA	1.03×10^{-19}	0.999980
Africans	ESN	1.02×10^{-17}	0.999906
	GWD	1.36×10^{-17}	0.999890
	LWK	3.06×10^{-17}	0.999890
	MSL	2.58×10^{-17}	0.999958
	YRI	1.74×10^{-17}	0.999752
Americans *	ACB	5.94×10^{-18}	0.999942
	ASW	9.16×10^{-19}	0.999969
	CLM	3.99×10^{-19}	0.999922
	MXL	3.87×10^{-18}	0.999869
	PEL	4.99×10^{-17}	0.999102
	PUR	2.81×10^{-19}	0.999957
East Asians	CDX	8.69×10^{-17}	0.999355
	CHB	8.77×10^{-17}	0.999862
	CHS	4.45×10^{-17}	0.999740
	JPT	1.94×10^{-17}	0.999429
	KHV	2.37×10^{-16}	0.999523
Europeans	CEU	1.27×10^{-18}	0.999926
	FIN	5.36×10^{-18}	0.999946
	GBR	3.09×10^{-18}	0.999987
	IBS	3.30×10^{-18}	0.999967
	TSI	1.82×10^{-18}	0.999936
South Asians	BEB	3.13×10^{-19}	0.999967
	ITU	7.08×10^{-19}	0.999952
	GIH	5.76×10^{-19}	0.999978
	PJL	3.22×10^{-19}	0.999944
	STU	3.74×10^{-19}	0.999960

* Americans without Brazil

3.4.4. Hardy-Weinberg equilibrium (HWE) and Linkage disequilibrium (LD)

In our study, Hardy-Weinberg equilibrium (HWE) testing in the SABE group revealed that 27 of the 28 markers showed no significant deviation. The threshold p -value for the HWE test was set at 0.00178 after the Bonferroni correction (i.e., $\alpha_{BONFERRONI} = 0.05 / 28$), and only MH19GB001 marker exhibited a significant deviation (Supplementary Table S8, Figure 3).

In America (without Brazil), only two markers deviate: MH05GB001 in Peru and MH22GB001 among African Americans and Caribbeans (Figure 3, Supplementary Table S5).

We found that deviations in the MH22GB001 marker occurred primarily in the African and East Asian populations, as most populations within these groups showed consistent deviations (Figure 3, Supplementary Table S4, and Table S6).

We calculated LD among all 28 markers, considering all samples from the 1000 Genomes dataset. The r^2 values of LD between the multi-InDels markers in the same chromosome were investigated, and no significant LD was observed. Similarly, in the SABE group, no significant LD ($p < 0.05$) was detected among any marker pairs on the same chromosome.

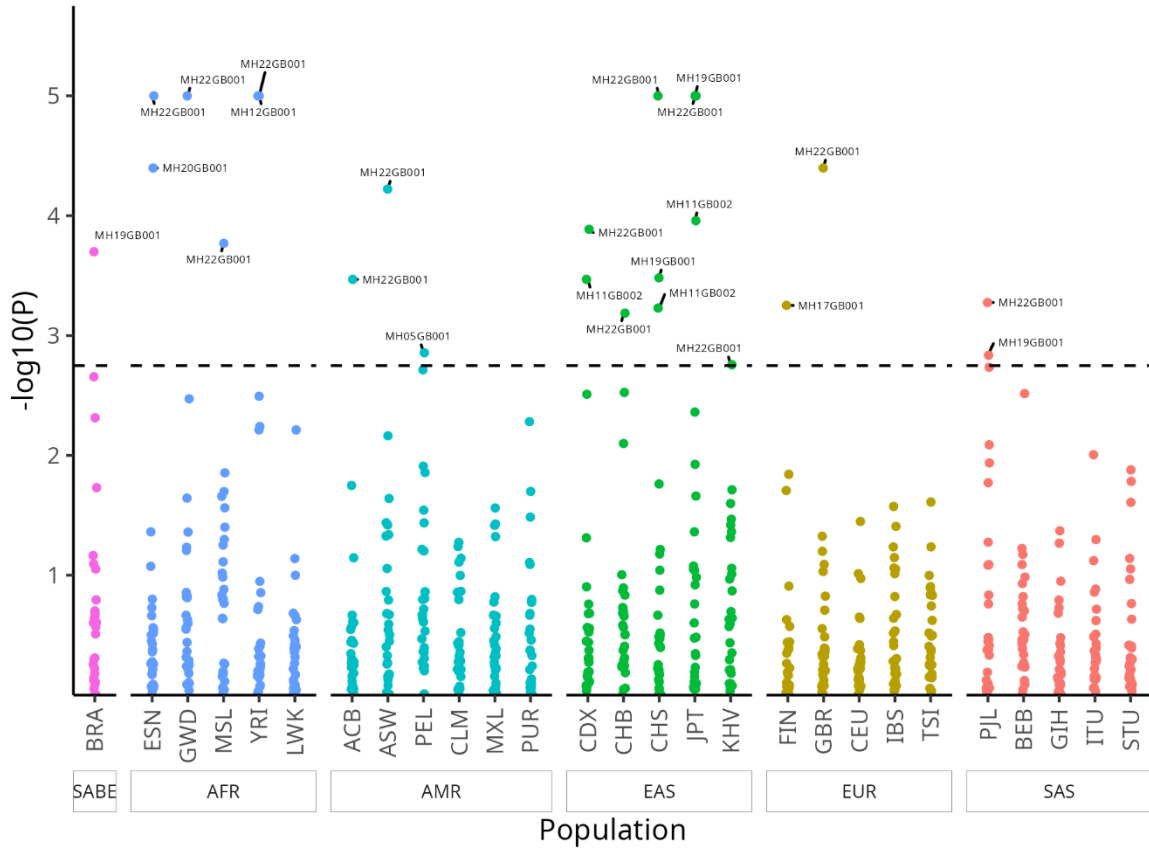


Figure 3 Visualization of HWE probabilities of the 28 multi-InDel markers in 26 populations grouped into five superpopulation groups. Brazilian population (SABE), African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The dashed line indicates the corrected significance level after Bonferroni correction for multiple tests ($\alpha_{BONFERRONI} = 0.00178$). The individual populations are listed in Table 1.

3.4.5. Multiplex Construction

We propose a multiplex assay combining 28 markers together with the Amelogenin (AMEL) and SRY loci, which are used as genetic markers for sex identification [26]. In total, 30 markers could be labeled using five fluorescent dyes (6-FAM, VIC, NED, TAZ, and SID) to distinguish amplified sample products. To ensure an optimal distribution, we analyzed the

allele size range across all markers, considering both the smallest and largest alleles. This strategy was implemented to minimize potential overlaps during analysis (Figures 4 and 5).

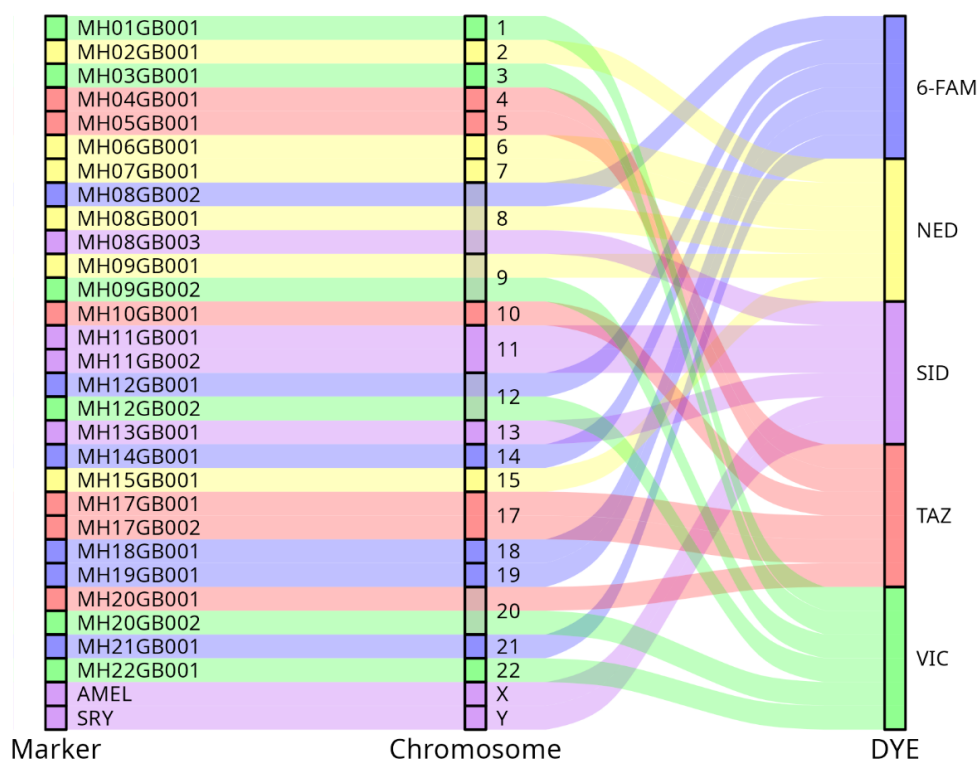


Figure 4 Correlation of the 30 markers with their chromosomes and dyes.

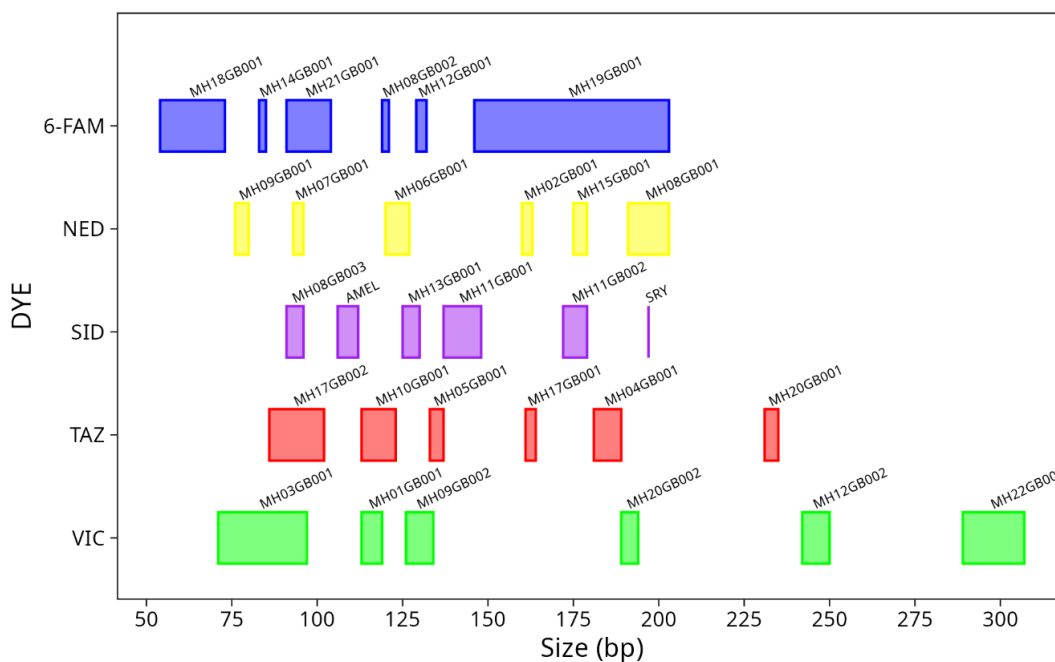


Figure 5 Distribution of the 30 markers in the multiplex assay. Each box represents the size range of the marker. The sizes are indicated in the x-axis, and the dye in the y-axis.

3.5. Discussion

InDel markers have been explored as potential tools in forensic DNA analysis. However, their power of discrimination (PD) and power of exclusion (PE) tend to be relatively low, primarily due to the low genetic diversity of biallelic markers [10]. To overcome this limitation, multi-InDel markers have been proposed, as they increase polymorphism levels while retaining the advantages of InDels, such as a low mutation rate, shorter amplicons, and compatibility with CE platforms [10; 27-28]. Thus, the forensic genetics community keeps trying to develop optimized multi-InDel panels for different populations.

Furthermore, understanding the different classes of InDels and their origins is essential for the development of such panels. Previous studies have noted the relationship between the genomic sequence of InDels and the mechanism of InDel formation. The generation of InDels can be attributed primarily to slipped strand mispairing and other mechanisms, including cellular DNA repair against secondary structure, defective mismatch repair, double-strand DNA breaks, and unequal meiotic recombination [12; 29]. Most InDels consisted of local tandem duplications and mononucleotide homopolymer structures, indicating local changes in copy number. InDels classified with random DNA sequences appear to be rare in the human genome [12; 30-31].

Corroborating this, our group analyzed three previous multi-InDels panels for human identification applied to the admixed Brazilian sample. It demonstrated that most markers are located within repetitive and low complexity sequences, homopolymers, or STR-like. In addition, some markers analyzed were biallelic in Brazilian individuals [11]. For this reason, it is essential to evaluate them across different populations, as biallelic multi-InDel markers retain the primary limitation of single InDel loci, their limited number of alleles, which leads to reduced biostatistical power per locus [27].

In this study, we aimed to develop a multiplex assay for human identification composed of 28 multi-InDel markers and two sex identification markers (Table 1). These new markers should not be located in repetitive and low-complexity sequences, avoiding potential amplification errors, such as the formation of stutter artifacts observed in STRs. Stutters are a major issue when analyzing STRs because they can mimic real alleles in mixed DNA samples, and are quite frequent, from 1.56% to 8.87% when using the GlobalFiler™ Kit [32], eventually reaching more than 15% [33]. The stutter percentages vary depending on the STR locus and

even the length of the allele within the locus: it increases when the allele size increases [32-33]. It is worth noting that multi-InDel markers not within repetitive and homopolymeric regions do not produce stutter artifacts associated with PCR amplification. This characteristic enhances their potential for forensic applications, especially in cases involving DNA mixtures from two or more individuals [28].

DNA degradation can occur when samples are exposed to environmental stressors such as light, humidity, elevated temperatures, and microbial contamination, including both bacterial and fungal agents. STR multiplexes produce complete profiles in most high-quality DNA samples. However, problematic samples containing degraded DNA, often characterized by minimal quantities of genetic material, are more challenging to analyze. In these cases, the higher molecular weight markers (> 250 bp) fail to bring unambiguous results or drop out completely [34-35]. Initially, we limited amplicon sizes to less than 150 bp to address the challenge of detecting degraded samples. However, some primers had to be redesigned due to the presence of frequent SNPs in the primer binding sites. Nevertheless, considering all groups analyzed (SABE, AFR, AMR, EAS, EUR, and SAS), a total of 107 specific amplicons were identified, with an average length of 149 bp (ranging from 54 to 307 bp). Most of these amplicons (85,05%, or 91 alleles) were shorter than 200 bp, and 94.4% were shorter than 250 bp. Only one marker, MH22GB001, exhibited alleles exceeding 250 bp (Figures 1 and 6). Therefore, the proposed panel may be useful for the analysis of degraded samples because only one marker exceeds 250bp.

As previously mentioned, multi-InDel panels offer a potential advantage for the analysis of degraded DNA. Li et al. (2021) [36] proposed a panel composed of 18 multi-InDels with amplicon sizes no greater than 125 bp. The authors evaluated the panel using highly degraded samples and generated complete profiles with as little as 62.5 pg of input DNA. However, most of its markers are in low-complexity, repetitive, and homopolymeric regions, which may present challenges for analysis [11]. In comparison, the GlobalFiler™ Kit and Identifiler Plus Kit consistently produced full profiles with input amounts of 125 pg or greater, with a length range of approximately 100–400 bp [32;35].

Different from STRs, multi-InDel are useful for ancestry inference when significant differences in allele frequencies are observed across populations [36-37]. Consequently, multiplex assays designed for specific populations may exhibit reduced genetic diversity in

other groups, which can restrict their overall applicability. In our study, we selected markers suitable for human identification and, therefore, all markers have at least three haplotypes (alleles) of varying sizes and with frequencies greater than 0.05 in all biogeographic regions (Figure 1 and Supplementary Table S1), guaranteeing high polymorphism everywhere. However, because of how we selected the markers, they are not very useful for ancestry inferences.

We estimated the CMP and CPE for each group and population (Tables 4 and 5). The values of CMP ranged from 5.66×10^{-20} (AMR) to 2.66×10^{-17} (EAS), and CPE ranged from 0.99958 (EAS) to 0.99998 (SABE). When compared to other multi-InDel panels designed only for human identification (10;27-28;38], this new multiplex assay showed greater discrimination power (Table 6).

Moreover, this new panel outperformed the Investigator® DIPplex Kit, which reported a CMP of 3.4×10^{-13} and a CPE of 0.9973 when applied in a sample of 519 individuals from six Brazilian states (Pará, Amazonas, Distrito Federal-Brasília, São Paulo, Rio de Janeiro e Santa Catarina) [39]. However, the new panel has not exceeded the results obtained with the GlobalFiler™ Kit (21 STRs), which was applied to 1,107 Brazilian individuals from five regions (North, Northeast, Central-West, Southeast, and South), with a CMP of 2×10^{-27} [40]. Similarly, GlobalFiler™ Kit (24 STRs) achieved a CMP of 1.30×10^{-22} and the PowerPlex® Fusion (24 STRs) achieved a CMP of 2.35×10^{-24} , both applied to a population from Mexico [41]. Nevertheless, this new multi-InDels panel might be a useful tool in forensic cases of individual identification, especially in those that require analyses of degraded samples or mixed samples, complementing conventional STR analysis.

Table 6. Comparison of combined match probability (CMP) and cumulative probability of exclusion (CPE) for different multi-InDels sets.

Number of markers	Population sample	CMP	CPE	Reference
28	Brazil	1.03×10^{-19}	0.99998	This Study
28	Americans	5.66×10^{-20}	0.99985	This Study
28	East Asia	2.66×10^{-17}	0.99958	This Study
20	Asia	5.7×10^{-12}	0.9989	Huang et al., 2014
18	Asia	2.77×10^{-12}	0.9984	Qu et al., 2020
18	Asia	3.23×10^{-12}	0.9989	Li et al., 2021
20	Asia	7×10^{-11}	0.992	Liu et al., 2022

The HWE is an important fundamental principle of population genetics and has been one of the key mathematical principles to understand the characteristics of genetic variation in population over multiple generations [42]. Additionally, HWE test have long been used as a metric for evaluating genotype quality, as technical artifacts leading to incorrect genotype calls often can be identified as deviations from HWE. We observed that the MH22GB001 marker, with alleles exceeding 250 bp, deviated from equilibrium primarily in the African and East Asian groups (Figure 3, Supplementary Table S4 and Table S6). These findings may indicate a potential genotyping error or issues with genotype phasing, highlighting the need for caution in future analyses.

Furthermore, in datasets comprising individuals from diverse ancestral backgrounds, HWE can be violated even without genotyping error, complicating the use of HWE testing to assess genotype data quality [43]. Accordingly, we conducted a more detailed analysis of p -values for HWE within individual populations to obtain more reliable and accurate results (Figure 3, Supplementary Tables S8–S13).

It is worth noting that, although it results in the largest amplicons observed in the present dataset, which may negatively impact genotyping using short-read NGS technologies, we manually checked the called genotypes and found no errors. Nevertheless, since our major goal is to implement this assay in a different genotyping platform (i.e., capillary electrophoresis) we opted to retain this marker in our multiplex assay, while emphasizing the need for caution in its interpretation in future analyzes.

We designed our multiplex assay and conducted all analysis in the present manuscript keeping in mind one pair of primers for each multi-InDel marker. Since CE platforms are currently widespread in forensic laboratories, this multiplex can be implemented using the same equipment and infrastructure already in place. Additionally, the five dyes used to label the amplified sample products (6-FAM, VIC, NED, TAZ, and SID) may be the same as those used in the GlobalFiler™ Kit [44], which is widely employed in routine forensic laboratory workflows (Figures 6 and 7). The sixth dye (LIZ) can be used to label size standards for the reproducible sizing of fragment analysis data. In addition, this panel can also be analyzed using short-read sequencing, as was done here. NGS could even increase the informativeness

because multi-InDels might present isoalleles due to SNPs. We have not addressed this in this article because our current goal is to develop a new panel for CE. However, we have observed that many of the proposed markers do present isoalleles. Finally, the selection and computational analyses conducted in this study were based on high-coverage whole-genome sequencing data from the 1000 Genomes Project (30x) and a census-based cohort of Brazilians (SABE). At present, the data obtained will support the laboratory testing.

Recently, the synthesis of fluorescent dye-labeled primers was requested to proceed with the experimental validation steps. Validation studies, including allele frequencies, species specificity, sensitivity, mixture analysis, and artificially degraded samples, are currently being conducted. Thus, we will obtain more consistent results regarding the potential of these markers.

3.6. Conclusion

In this study, we selected and proposed a multiplex assay for human identification suitable for use in a PCR-CE workflow, comprising 28 multi-InDel markers and two sex identification markers. Our multi-InDel panel have shorter fragments than conventional STR markers, with 94.4% of amplicons shorter than 250bp, offering greater potential for forensic applications, particularly when analyzing degraded DNA samples. Additionally, these multi-InDels do not generate stutter artifacts typically associated with PCR amplification, as they are not located in repetitive sequences, homopolymers, or low-complexity regions. This feature enhances their possible usefulness in cases involving DNA mixtures. The panel showed high polymorphism in five different biogeographic regions and for admixed Brazilians. The CMP and CPE ranged from 5.66×10^{-20} to 2.66×10^{-17} , and 0.99958 to 0.99998, respectively. Therefore, this system might be helpful as a complementary tool for individual identification across worldwide populations, especially for challenging samples.

Data Availability Statement

The raw sequencing data used in this study are available in public repositories: 1000 Genomes, available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections, the Brazilian/SABE

cohort, EGAD00001008640. Some of the scripts used in this study are available on <https://github.com/erickcastelli/supporting>.

Acknowledgements

This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP/ Brasil) (#2022/13924–5).

3.7. Reference

- [1] S. Qu, M. Lv, J. Xue, J. Zhu, L. Wang, H. Jian, et al. Multi-Indel: a microhaplotype marker can be typed using capillary electrophoresis platforms. *Front Genet.* 11 (2020) 567082. <https://doi.org/10.3389/fgene.2020.567082>.
- [2] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, S. E. Devine. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16 (2006) 1182-90. <https://doi.org/10.1101/gr.4565806>.
- [3] J. Li, L. Lin, B. Jiang, C. Wang, M.M.J. Zeye, D. Wen, et al. An 18 Multi-InDels panel for analysis of highly degraded forensic biological samples. *Electrophoresis* 42 (2021) 1143–52. <https://doi.org/10.1002/elps.202000245>.
- [4] M.W. Nachman, S.L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156 (2000) 297-304, <https://doi:10.1093/genetics/156.1.297>.
- [5] J.M. Mullaney, R.E. Mills, W.S. Pittard, S.E. Devine. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet.* 19 (2010), R131-6, <https://doi:10.1093/hmg/ddq400>.
- [6] T. Z. Gao, L. B. Yun, W. He, Y. Gu, Y. P. Hou. The application of multi-InDel as supplementary in paternity cases with STR mutation. *Foren. Sci. Intern. Genet. Suppl. Ser.* 5 (2015) e218–e219. <https://doi.org/10.1016/j.fsigss.2015.09.087>

- [7] F. Song, H.B. Luo, J.N. Li, Y.P. Hou. Validation of a multiplex system with 20 multi-Indels for forensic purposes. *Forensic Science International: Genetics Supplement Series* 5 (2015) e48-e49, <https://doi.org/10.1016/j.fsigss.2015.09.019>.
- [8] K. Sun, Y. Ye, T. Luo, Y. Hou. Multi-InDel analysis for ancestry inference of sub-populations in China. *Sci. Rep.* 6 (2016) 39797. <https://doi.org/10.1038/srep39797>.
- [9] J. Huang, H. Luo, W. Wei, Y. Hou. A novel method for the analysis of 20 multi-Indel polymorphisms and its forensic application. *Electrophoresis*. 35 (2014) 487–493. <https://doi.org/10.1002/elps.201300346>.
- [10] Liu, J.; Zhang, X.; Zhang, X.; Li, W.; Gao, L.; Li, J.; Wang, J.; Liu, Z.; Liu, Y.; Yan, J.; Zhang, G. A new set of 20 Multi-InDel markers for forensic application. *Electrophoresis*. 43 (2022), 1193-1202. doi: 10.1002/elps.202100361.
- [11] L.C. Ramos, L.D.G. Kobachuk, D.M. Nadur, L.R. Sabbag, M.M.T.D. Rosário, M.S. Naslavsky, C.T. Mendes-Junior, E.C. Castelli. Assessing the performance of multi-InDel panels for human identification among admixed Brazilians. *Forensic Sci Int Genet.* 74 (2025), 103161, <https://doi:10.1016/j.fsigen.2024.103161>.
- [12] S.B. Montgomery, D.L. Goode, E. Kvikstad, C.A. Albers, Z.D. Zhang, X.J. Mu, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 23 (2013), 749–61, <https://doi:10.1101/gr.148718.112>.
- [13] 1000 Genomes Project Consortium; A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis. A global reference for human genetic variation. *Nature*. 526 (2015), 68-74, <https://doi: 10.1038/nature15393>.
- [14] M. S. Naslavsky, M. O. Scliar, G. L. Yamamoto, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. *Nat Commun* 13 (2022) 1004. <https://doi.org/10.1038/s41467-022-28648-3>.

- [15] P.M.Vallone, J.M. Butler. AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques*. 37 (2004), 226–31, <https://doi.org/10.2144/04372ST03>.
- [16] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O., Basile, H.J. Abel, A.A. Regier, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*. 185 (2022), 3426–3440.e19, <https://doi.org/doi:10.1016/j.cell.2022.08.004>.
- [17] A. Gouy, M. Zieger. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet*. 30 (2017) 148-151. <https://doi.org/10.1016/j.fsigen.2017.07.007>.
- [18] L. Excoffier, H. E. Lischer. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10 (2010) 564-7. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>.
- [19] K.K. Kidd, W.C Speed. Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investig Genet*. 6 (2015). <https://doi.org/10.1186/s13323-014-0018-3>.
- [20] R (language): TEAM, R. Core. R development core team. *RA Lang Environ Stat Comput*, 55 (2013), 275-286.
- [21] RStudio: TEAM, RStudio et al. RStudio: integrated development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com>, 42 (2015), 14.
- [22] H. Wickham, D. Vaughan, M. Girlich. tidy: Tidy Messy Data. (2024). R package version 1.3.1.
- [23] Sullivan KM, Mannucci A, Kimpton CP, Gill P. A rapid and quantitative DNA sex test: Fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *BioTechniques* 15 (1993), 640-641. PMID: 8251166.
- [24] Tozzo P, Giuliadori A, Corato S, Ponzano E, Rodriguez D et al. Deletion of amelogenin Y-locus in forensics: Literature revision and description of a novel method for sex confirmation. *J Forensic Leg Med*. 20 (2013), 387-391, <https://doi.org/10.1016/j.jflm.2013.03.012>.

- [25] K. K. Kidd. Proposed nomenclature for microhaplotypes. *Hum. Genom.* 10 (2016), 16, <http://doi:10.1186/s40246-016-0078-y>.
- [26] E. Butler, R. Li. Genetic Markers for Sex Identification in Forensic DNA Analysis. *J Forensic Investigation.* 2 (2014), 10.
- [27] J. Huang, H. Luo, W. Wei, Y. Hou. A novel method for the analysis of 20 multi-Indel polymorphisms and its forensic application. *Electrophoresis.* 35 (2014) 487–493. <https://doi.org/10.1002/elps.201300346>.
- [28] S. Qu, M. Lv, J. Xue, J. Zhu, L. Wang, H. Jian, et al. Multi-Indel: a microhaplotype marker can be typed using capillary electrophoresis platforms. *Front Genet.* 11 (2020) 567082. <https://doi.org/10.3389/fgene.2020.567082>.
- [29] Y. Yao, K. Sun, Q. Yang, Z. Zhou, C. Shao, X. Qian, Q. Tang, J. Xie. Assessing Autosomal InDel Loci With Multiple Insertions or Deletions of Random DNA Sequences in Human Genome. *Front Genet.* 12 (2022),809-815, [http:// doi: 10.3389/fgene.2021.809815](http://doi:10.3389/fgene.2021.809815).
- [30] M.S. Taylor, C. P. Ponting, R. R. Copley. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* 14 (2004), 555–566, [https:// doi. org/ 10. 1101/gr. 19778 04](https://doi.org/10.1101/gr.1977804).
- [31] S. Zhang, Q. Zhu, X. Chen, Y. Zhao, X. Zhao, Y. Yang, Z. Gao, T. Fang, Y. Wang, J. Zhang. Forensic applicability of multi-allelic InDels with mononucleotide homopolymer structures. *Electrophoresis.* 39 (2018), 2136-2143. [https://doi: 10.1002/elps.201700468](https://doi:10.1002/elps.201700468).
- [32] M.J. Ludeman, C. Zhong, J.J. Mulero. et al. Developmental validation of GlobalFiler™ PCR amplification kit: a 6-dye multiplex assay designed for amplification of casework samples. *Int J Legal Med.* 132 (2018), 1555–1573, <https://doi.org/10.1007/s00414-018-1817-5>.
- [33] Butler J. M. *Advanced Topics in Forensic DNA Typing: Interpretation.* Elsevier – Academic Press. 2014.

- [34] P.M. Schneider, K. Bender, W. R. Mayr, W. Parson, B. Hoste, R. Decorte, J. Cordonnier, et al. STR analysis of artificially degraded DNA-results of a collaborative European exercise. *Forensic Sci Int.* 139 (2004), 123-34, [https://doi: 10.1016/j.forsciint.2003.10.002](https://doi.org/10.1016/j.forsciint.2003.10.002).
- [35] P. Grubwieser, R. Mühlmann, B. Berger, H. Niederstätter, M. Pavlic, W. Parson. A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA. *Int J Legal Med.* 120 (2006), 115-20, [https://doi: 10.1007/s00414-005-0013-6](https://doi.org/10.1007/s00414-005-0013-6). Epub 2005 Jul 13. PMID: 16012830.
- [36] Sun, K. et al. Multi-InDel Analysis for Ancestry Inference of Sub-Populations in China. *Scientific Reports.* 6 (2016), 39797.
- [37] S. Mei, S. Yi, M. Cai, Y. Zhang, W. Cui, H. Xu, Q. Lan, B. Zhu. Exploring the forensic effectiveness and population genetic differentiation by self-constructed 41 multi-InDel panel in Yunnan Zhuang group. *Gene.* 860 (2023), 147180, <https://doi.org/10.1016/j.gene.2023.147180>.
- [38] J. Li, L. Lin, B. Jiang, C. Wang, M.M.J. Zeye, D. Wen, et al. An 18 Multi-InDels panel for analysis of highly degraded forensic biological samples. *Electrophoresis* 42 (2021) 1143–52. <https://doi.org/10.1002/elps.202000245>.
- [39] T.J.B. Ferreira Palha, E.M. Ribeiro Rodrigues, G.C. Cavalcante, A. Marrero, I.R. de Souza, C.J. Seki Uehara, et al., Population genetic analysis of insertion-deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit, *Forensic Sci. Int. Genet.* 19 (2015) 10–14. <https://doi.org/10.1016/j.fsigen.2015.03.015>.
- [40] C. B. Moysés, W. M. Tsutsumida, P. E. Raimann, C. H. da Motta, et al. Population data of the 21 autosomal STRs included in the GlobalFiler® kits in population samples from five Brazilian regions. *Forensic Sci Int Genet.* 26 (2017) e28-e30. <https://doi.org/10.1016/j.fsigen.2016.10.017>.
- [41] Ramos-González, B., Aguilar-Velázquez, J.A., Chávez-Briones, M.L., Rangel-Villalobos, H. et al. Population data of 24 STRs in Mexican-Mestizo population from

Monterrey, Nuevo Leon (Northeast, Mexico) based on Powerplex[®] Fusion and GlobalFiler[®] kits. Forensic Science International: Genetics, 21 (2016), e15 - e17.

[42] A. W. Edwards. G. H. Hardy (1908) and Hardy-Weinberg equilibrium. Genetics. 179 (2008), 1143-50, <https://doi.org/10.1534/genetics.104.92940>.

[43] A.M. Kwong, T.W. Blackwell, J. LeFaive, M. de Andrade, J. Barnard, K.C. Barnes, et al. Robust, flexible, and scalable tests for Hardy-Weinberg equilibrium across diverse ancestries. Genetics. 218 (2021), iyab044, <https://doi.org/10.1093/genetics/iyab044>.

[44] GlobalFiler[™] and GlobalFiler[™] IQC PCR Amplification Kits User's Guide, ThermoFisher Scientific, 2019.

3.8. Appendix A. Supplementary data

Supplementary Table S 1. List of allelic frequencies of 28 multi-InDel markers in different groups.

Marker	Group	Amplicon size in bp (Frequency)
MH01GB001	SABE	113 (0.4849); 116 (0.3036); 117 (0.2115)
	AFR	113 (0.4829); 114 (0.0007); 116 (0.4110); 117 (0.1052)
	AMR	113 (0.4868); 115 (0.0007); 116 (0.3831); 117 (0.1294)
	EAS	113 (0.6282); 116 (0.3256); 117 (0.0462)
	EUR	113 (0.4455); 116 (0.2733); 117 (0.2812)
	SAS	113 (0.4359); 116 (0.4285); 117 (0.1348); 119 (0.0008)
MH02GB001	SABE	160 (0.1492); 161 (0.4502); 162 (0.3999); 163 (0.0008)
	AFR	160 (0.3406); 161 (0.3990); 162 (0.2603)
	AMR	160 (0.2059); 161 (0.5243); 162 (0.2699)
	EAS	160 (0.2675); 161 (0.5350); 162 (0.1974)
	EUR	160 (0.0790); 161 (0.4502); 162 (0.4708)
	SAS	160 (0.2970); 161 (0.3885); 162 (0.3145)
MH03GB001	SABE	71 (0.3958); 84 (0.4048); 85 (0.1986); 97 (0.0008)
	AFR	71 (0.4694); 84 (0.3357); 85 (0.1949);
	AMR	71 (0.3846); 84 (0.3176); 85 (0.2978)
	EAS	71 (0.3436); 84 (0.4368); 85 (0.2197)
	EUR	71 (0.4036); 84 (0.4321); 85 (0.1643)
	SAS	71 (0.4384); 84 (0.4334); 85 (0.1281)
MH04GB001	SABE	181 (0.0011); 182 (0.0004); 183 (0.3225); 185 (0.3350); 186 (0.0004); 188 (0.3403); 189 (0.0004)
	AFR	181 (0.0100); 183 (0.1508); 185 (0.6714); 188 (0.1664); 189 (0.0014)
	AMR	181 (0.0007); 183 (0.3007); 185 (0.3926); 188 (0.3059)
	EAS	183 (0.4470); 185 (0.3641); 188 (0.1889)
	EUR	183 (0.3065); 185 (0.3262); 188 (0.3673)
	SAS	181 (0.0008); 183 (0.2454); 185 (0.5141); 186 (0.0017); 188 (0.2379)
MH05GB001	SABE	133 (0.2719); 136 (0.3395); 137 (0.3886)
	AFR	133 (0.0377); 136 (0.5626); 137 (0.3997)
	AMR	133 (0.2213); 136 (0.4213); 137 (0.3574)
	EAS	133 (0.1607); 136 (0.4615); 137 (0.3778)
	EUR	133 (0.2954); 136 (0.3081); 137 (0.3965)
	SAS	133 (0.2845); 136 (0.3594); 137 (0.3561)
MH06GB001	SABE	120 (0.3927); 126 (0.3221); 127 (0.2851)
	AFR	120 (0.4353); 126 (0.1152); 127 (0.4495)
	AMR	120 (0.3412); 126 (0.3169); 127 (0.3419)

	EAS	120 (0.1393); 126 (0.4581); 127 (0.4026)
	EUR	120 (0.3949); 126 (0.3626); 127 (0.2425)
	SAS	120 (0.2787); 126 (0.2962); 127 (0.4251)
MH07GB001	SABE	93 (0.1718); 95 (0.3361); 96 (0.4921)
	AFR	93 (0.2397); 95 (0.2852); 96 (0.4751)
	AMR	93 (0.1846); 95 (0.2816); 96 (0.5338)
	EAS	93 (0.1274); 95 (0.5923); 96 (0.2803)
	EUR	93 (0.1635); 95 (0.3728); 96 (0.4637)
	SAS	93 (0.1963); 95 (0.3278); 96 (0.4759)
MH08GB001	SABE	191 (0.3773); 202 (0.2323); 203 (0.3905)
	AFR	191 (0.1430); 202 (0.2283); 203 (0.6287)
	AMR	191 (0.4007); 202 (0.2140); 203 (0.3853)
	EAS	191 (0.5282); 202 (0.1265); 203 (0.3453)
	EUR	191 (0.4005); 202 (0.2030); 203 (0.3965)
	SAS	191 (0.5191); 202 (0.1614); 203 (0.3195)
MH08GB002	SABE	119 (0.3255); 120 (0.2897); 121 (0.3848)
	AFR	119 (0.5477); 120 (0.0661); 121 (0.3862)
	AMR	119 (0.3787); 120 (0.2779); 121 (0.3434)
	EAS	119 (0.2812); 120 (0.2009); 121 (0.5179)
	EUR	119 (0.2938); 120 (0.2994); 121 (0.4068)
	SAS	119 (0.1780); 120 (0.3611); 121 (0.4609)
MH08GB003	SABE	92 (0.5733); 93 (0.3066); 96 (0.1201)
	AFR	92 (0.3947); 93 (0.3606); 96 (0.2447)
	AMR	92 (0.4574); 93 (0.3618); 96 (0.1809)
	EAS	92 (0.2812); 93 (0.6359); 96 (0.0829)
	EUR	92 (0.6833); 93 (0.2330); 96 (0.0837)
	SAS	91 (0.0017); 92 (0.4925); 93 (0.3977); 96 (0.1082)
MH09GB001	SABE	76 (0.1654); 77 (0.2919); 80 (0.5427)
	AFR	76 (0.1629); 77 (0.3549); 80 (0.4822)
	AMR	76 (0.1382); 77 (0.3875); 80 (0.4743)
	EAS	76 (0.4607); 77 (0.4239); 80 (0.1154)
	EUR	76 (0.1216); 77 (0.2694); 80 (0.6090)
	SAS	76 (0.3295); 77 (0.3802); 80 (0.2903)
MH09GB002	SABE	126 (0.0004); 129 (0.0310); 130 (0.4418); 133 (0.1673); 134 (0.3595)
	AFR	126 (0.0036); 129 (0.0996); 130 (0.5825); 133 (0.0583); 134 (0.2560)
	AMR	129 (0.0279); 130 (0.4500); 133 (0.1500); 134 (0.3721)
	EAS	129 (0.0137); 130 (0.2350); 133 (0.0667); 134 (0.6846)
	EUR	129 (0.0197); 130 (0.3910); 133 (0.2694); 134 (0.3199)
	SAS	129 (0.0824); 130 (0.2903); 133 (0.1581); 134 (0.4692)

MH10GB001	SABE	113 (0.0004); 114 (0.0695); 122 (0.6148); 123 (0.3153)
	AFR	114 (0.1024); 122 (0.4282); 123 (0.4694)
	AMR	114 (0.0603); 122 (0.4963); 123 (0.4434)
	EAS	114 (0.1650); 122 (0.5803); 123 (0.2547)
	EUR	114 (0.0379); 122 (0.7133); 123 (0.2488)
	SAS	114 (0.1622); 122 (0.6115); 123 (0.2263)
MH11GB001	SABE	137 (0.2526); 138 (0.3040); 148 (0.4434)
	AFR	137 (0.2418); 138 (0.6579); 148 (0.1003)
	AMR	137 (0.3191); 138 (0.2912); 148 (0.3897)
	EAS	137 (0.5034); 138 (0.2855); 148 (0.2111)
	EUR	137 (0.2267); 138 (0.2417); 148 (0.5316)
	SAS	137 (0.2903); 138 (0.2121); 148 (0.4975)
MH11GB002	SABE	172 (0.0038); 173 (0.4029); 174 (0.3826); 179 (0.2107)
	AFR	172 (0.0583); 173 (0.5213); 174 (0.3585); 179 (0.0619)
	AMR	172 (0.0110); 173 (0.3096); 174 (0.3228); 179 (0.3566)
	EAS	173 (0.4333); 174 (0.2376); 179 (0.3291)
	EUR	173 (0.3397); 174 (0.4779); 179 (0.1825)
	SAS	173 (0.3735); 174 (0.3735); 179 (0.2529)
MH12GB001	SABE	129 (0.4645); 131 (0.2897); 132 (0.2458)
	AFR	129 (0.3890); 131 (0.2731); 132 (0.3378)
	AMR	129 (0.4529); 131 (0.2588); 132 (0.2882)
	EAS	129 (0.6316); 131 (0.0376); 132 (0.3308)
	EUR	129 (0.5063); 131 (0.3578); 132 (0.1359)
	SAS	129 (0.4251); 131 (0.2654); 132 (0.3095)
MH12GB002	SABE	242 (0.3984); 245 (0.2829); 250 (0.3187)
	AFR	242 (0.1970); 245 (0.7084); 250 (0.0946)
	AMR	242 (0.2257); 245 (0.3346); 250 (0.4397)
	EAS	242 (0.0120); 245 (0.6154); 250 (0.3726)
	EUR	242 (0.4739); 245 (0.1311); 250 (0.3949)
	SAS	242 (0.4834); 245 (0.2662); 250 (0.2504)
MH13GB001	SABE	125 (0.1582); 129 (0.4298); 130 (0.4120)
	AFR	125 (0.0960); 129 (0.6038); 130 (0.3001)
	AMR	125 (0.0926); 129 (0.4993); 130 (0.4081)
	EAS	125 (0.3274); 129 (0.5393); 130 (0.1333)
	EUR	125 (0.1856); 129 (0.3720); 130 (0.4423)
	SAS	125 (0.2604); 129 (0.5591); 130 (0.1805)
MH14GB001	SABE	83 (0.3116); 84 (0.3758); 85 (0.3127)
	AFR	83 (0.2568); 84 (0.4844); 85 (0.2589)
	AMR	83 (0.3066); 84 (0.3397); 85 (0.3537)

	EAS	83 (0.2111); 84 (0.2735); 85 (0.5154)
	EUR	83 (0.3547); 84 (0.3760); 85 (0.2694)
	SAS	83 (0.4126); 84 (0.1656); 85 (0.4218)
MH15GB001	SABE	175 (0.1386); 178 (0.5431); 179 (0.3184)
	AFR	175 (0.1984); 178 (0.6216); 179 (0.1799)
	AMR	175 (0.1140); 178 (0.6456); 179 (0.2404)
	EAS	175 (0.1641); 178 (0.4735); 179 (0.3624)
	EUR	175 (0.0972); 178 (0.5087); 179 (0.3942)
	SAS	175 (0.0682); 178 (0.5358); 179 (0.3960)
MH17GB001	SABE	161 (0.2931); 162 (0.2681); 163 (0.4384); 164 (0.0004)
	AFR	161 (0.2888); 162 (0.1899); 163 (0.5213)
	AMR	161 (0.3426); 162 (0.2103); 163 (0.4471)
	EAS	161 (0.2573); 162 (0.2137); 163 (0.5291)
	EUR	161 (0.3349); 162 (0.2607); 163 (0.4044)
	SAS	161 (0.2521); 162 (0.2621); 163 (0.4859)
MH17GB002	SABE	86 (0.0004); 87 (0.2995); 100 (0.0872); 101 (0.4958); 102 (0.1171)
	AFR	87 (0.1543); 100 (0.4381); 101 (0.4047); 102 (0.0028)
	AMR	87 (0.2544); 100 (0.1257); 101 (0.5404); 102 (0.0794)
	EAS	87 (0.5675); 100 (0.0368); 101 (0.3009); 102 (0.0949)
	EUR	87 (0.2970); 100 (0.0245); 101 (0.5276); 102 (0.1509)
	SAS	87 (0.3519); 100 (0.0433); 101 (0.4301); 102 (0.1747)
MH18GB001	SABE	54 (0.1752); 72 (0.6998); 73 (0.1250)
	AFR	54 (0.4253); 72 (0.2525); 73 (0.3222)
	AMR	54 (0.3074); 72 (0.5618); 73 (0.1309)
	EAS	54 (0.2957); 72 (0.6051); 73 (0.0991)
	EUR	54 (0.0750); 72 (0.8136); 73 (0.1114)
	SAS	54 (0.1963); 72 (0.5990); 73 (0.2047)
MH19GB001	SABE	188 (0.3044); 189 (0.4468); 190 (0.0011); 202 (0.2406); 203 (0.0072)
	AFR	146 (0.0007); 188 (0.2127); 189 (0.3201); 190 (0.0050); 202 (0.4609); 203 (0.0007)
	AMR	188 (0.3051); 189 (0.4412); 202 (0.2485); 203 (0.0051)
	EAS	188 (0.6573); 189 (0.1675); 202 (0.0889); 203 (0.0863)
	EUR	188 (0.2512); 189 (0.5577); 202 (0.1880); 203 (0.0032)
	SAS	188 (0.3561); 189 (0.4642); 202 (0.1764); 203 (0.0033)
MH20GB001	SABE	231 (0.1560); 232 (0.2028); 235 (0.6412)
	AFR	231 (0.4061); 232 (0.0768); 235 (0.5171)
	AMR	231 (0.1801); 232 (0.2478); 235 (0.5721)
	EAS	231 (0.1231); 232 (0.2556); 235 (0.6214)
	EUR	231 (0.1256); 232 (0.2417); 235 (0.6327)
	SAS	231 (0.0183); 232 (0.3286); 235 (0.6531)

MH20GB002	SABE	189 (0.2137);193 (0.2791);194 (0.5072)
	AFR	189(0.0583); 193 (0.3378); 194 (0.6038)
	AMR	189 (0.2368); 193 (0.2816); 194 (0.4816)
	EAS	189 (0.3436); 193 (0.1692); 194 (0.4872)
	EUR	189 (0.2322); 193 (0.2804); 194 (0.4874)
	SAS	189 (0.3536); 193 (0.2795); 194 (0.3669)
MH21GB001	SABE	91 (0.2326); 93 (0.4943); 104 (0.2730)
	AFR	91 (0.2333); 93 (0.4523); 104 (0.3144)
	AMR	91 (0.1500); 93 (0.6096); 104 (0.2404)
	EAS	91 (0.0342); 93 (0.8547); 104 (0.1111)
	EUR	91 (0.2283); 93 (0.4668); 104 (0.3049)
	SAS	91 (0.2296); 93 (0.5824); 104 (0.1880)
MH22GB001	SABE	289 (0.2666); 299 (0.2462); 301 (0.1469); 307 (0.3403)
	AFR	289 (0.1927); 297 (0.0007); 299 (0.5541); 301 (0.0277); 307 (0.2248)
	AMR	289 (0.3404); 299 (0.2507); 301 (0.1235); 307 (0.2853)
	EAS	289 (0.2265); 297 (0.0017); 299 (0.3145); 301 (0.0350); 306 (0.0009); 307 (0.4214)
	EUR	289 (0.2141); 299 (0.1359); 301 (0.1943); 307 (0.4558)
	SAS	289 (0.3037); 299 (0.2438); 301 (0.1364); 307 (0.3161)

Note: Three alleles were observed: two with frequencies > 0.05, and one with a frequency < 0.05 (shown in bold).

Supplementary Table S 2. Estimated values of *Ae*, *He*, *Ho*, *PD*, *PE*, *TPI*, *PIC*, and *MP* for the 28 multi-InDel markers in the Brazil from SABE cohort.

Marker	<i>Ae</i>	<i>He</i>	<i>Ho</i>	<i>PD</i>	<i>PE</i>	<i>TPI</i>	<i>PIC</i>	<i>MP</i>
MH01GB001	2.69	0.6282	0.6110	0.7907	0.3043	1.2854	0.5553	0.2093
MH02GB001	2.60	0.6154	0.5944	0.7742	0.2842	1.2328	0.5342	0.2258
MH03GB001	2.78	0.6403	0.6307	0.7956	0.3293	1.3538	0.5634	0.2044
MH04GB001	3.01	0.6683	0.6798	0.8135	0.3976	1.5613	0.5946	0.1865
MH05GB001	2.94	0.6601	0.6571	0.8096	0.3651	1.4581	0.5856	0.1904
MH06GB001	2.95	0.6609	0.6367	0.8158	0.3373	1.3763	0.5867	0.1842
MH07GB001	2.60	0.6156	0.5989	0.7811	0.2896	1.2467	0.5397	0.2189
MH08GB001	2.87	0.6515	0.6261	0.8094	0.3234	1.3374	0.5760	0.1906
MH08GB002	2.96	0.6623	0.6579	0.8123	0.3661	1.4614	0.5880	0.1877
MH08GB003	2.29	0.5631	0.5461	0.7397	0.2311	1.1015	0.4889	0.2603
MH09GB001	2.46	0.5932	0.5785	0.7648	0.2659	1.1864	0.5220	0.2352
MH09GB002	2.83	0.6468	0.6397	0.8055	0.3413	1.3878	0.5773	0.1945
MH10GB001	2.07	0.5179	0.5279	0.6821	0.2131	1.0592	0.4380	0.3179
MH11GB001	2.83	0.6474	0.6367	0.8042	0.3373	1.3763	0.5740	0.1958
MH11GB002	2.83	0.6471	0.6480	0.7990	0.3525	1.4206	0.5719	0.2010
MH12GB001	2.78	0.6401	0.6201	0.8011	0.3157	1.3161	0.5675	0.1989
MH12GB002	2.94	0.6599	0.6495	0.8106	0.3546	1.4267	0.5858	0.1894
MH13GB001	2.64	0.6208	0.6118	0.7741	0.3053	1.2879	0.5401	0.2259
MH14GB001	2.98	0.6642	0.6631	0.8133	0.3736	1.4843	0.5900	0.1867
MH15GB001	2.41	0.5848	0.5642	0.7570	0.2501	1.1473	0.5095	0.2430
MH17GB001	2.86	0.6502	0.6503	0.8044	0.3556	1.4298	0.5770	0.1956
MH17GB002	2.80	0.6434	0.6299	0.8155	0.3283	1.3510	0.5845	0.1845
MH18GB001	1.87	0.4642	0.4328	0.6606	0.1351	0.8815	0.4177	0.3394
MH19GB001	2.86	0.6501	0.6397	0.8070	0.3413	1.3878	0.5790	0.1930
MH20GB001	2.10	0.5236	0.5174	0.7165	0.2032	1.0360	0.4675	0.2835
MH20GB002	2.63	0.6194	0.6186	0.7844	0.3138	1.3109	0.5485	0.2156
MH21GB001	2.68	0.6272	0.6080	0.7924	0.3006	1.2755	0.5560	0.2076
MH22GB001	3.72	0.7312	0.7032	0.8788	0.4332	1.6845	0.6811	0.1212

Ae: effective number of alleles; *He*: expected heterozygosity; *Ho*: observed heterozygosity; *PD*: power of discrimination; *PE*: power of exclusion; *TPI*: typical paternity index; *PIC*: polymorphism information content; *MP*: match probability.

Supplementary Table S 3. Estimated values of A_e , H_e , H_o , PD , PE , TPI , PIC , and MP for the 28 multi-InDel markers in the African group from the 1000 Genomes Project.

Marker	A_e	H_e	H_o	PD	PE	TPI	PIC	PM
MH01GB001	2.42	0.5871	0.6159	0.7294	0.3104	1.3019	0.4990	0.2706
MH02GB001	2.92	0.6574	0.6515	0.8083	0.3573	1.4347	0.5827	0.1917
MH03GB001	2.70	0.6294	0.6401	0.7830	0.3418	1.3893	0.5540	0.2170
MH04GB001	1.99	0.4990	0.4950	0.7028	0.1832	0.9901	0.4518	0.2972
MH05GB001	2.09	0.5227	0.5676	0.6491	0.2538	1.1563	0.4198	0.3509
MH06GB001	2.47	0.5956	0.6373	0.7262	0.3380	1.3784	0.5082	0.2738
MH07GB001	2.74	0.6359	0.6060	0.7998	0.2981	1.2690	0.5635	0.2002
MH08GB001	2.14	0.5325	0.5391	0.7202	0.2241	1.0849	0.4726	0.2798
MH08GB002	2.21	0.5469	0.5676	0.6939	0.2538	1.1563	0.4532	0.3061
MH08GB003	2.89	0.6548	0.6785	0.7984	0.3958	1.5553	0.5795	0.2016
MH09GB001	2.60	0.6154	0.6117	0.7764	0.3051	1.2875	0.5374	0.2236
MH09GB002	2.39	0.5822	0.5846	0.7581	0.2728	1.2038	0.5265	0.2419
MH10GB001	2.41	0.5863	0.6159	0.7247	0.3104	1.3019	0.4966	0.2753
MH11GB001	1.99	0.4990	0.4851	0.6893	0.1748	0.9710	0.4381	0.3107
MH11GB002	2.45	0.5929	0.5861	0.7597	0.2745	1.2079	0.5168	0.2403
MH12GB001	2.94	0.6604	0.6401	0.8144	0.3418	1.3893	0.5858	0.1856
MH12GB002	1.82	0.4507	0.4168	0.6438	0.1245	0.8573	0.4018	0.3562
MH13GB001	2.16	0.5365	0.5064	0.7166	0.1932	1.0130	0.4620	0.2834
MH14GB001	2.72	0.6329	0.5875	0.8004	0.2761	1.2121	0.5612	0.1996
MH15GB001	2.18	0.5422	0.5405	0.7322	0.2255	1.0882	0.4838	0.2678
MH17GB001	2.56	0.6092	0.5718	0.7812	0.2584	1.1678	0.5378	0.2188
MH17GB002	2.63	0.6209	0.5960	0.7815	0.2861	1.2377	0.5406	0.2185
MH18GB001	2.87	0.6520	0.6060	0.8127	0.2981	1.2690	0.5777	0.1873
MH19GB001	2.78	0.6404	0.6159	0.8034	0.3104	1.3019	0.5679	0.1966
MH20GB001	2.28	0.5622	0.5050	0.7136	0.1919	1.0101	0.4685	0.2864
MH20GB002	2.07	0.5182	0.4865	0.6935	0.1759	0.9737	0.4314	0.3065
MH21GB001	2.79	0.6426	0.6259	0.8031	0.3231	1.3365	0.5687	0.1969
MH22GB001	2.53	0.6050	0.5647	0.7752	0.2507	1.1487	0.5464	0.2248

A_e : effective number of alleles; H_e : expected heterozygosity; H_o : observed heterozygosity; PD : power of discrimination; PE : power of exclusion; TPI : typical paternity index; PIC : polymorphism information content; MP : match probability.

Supplementary Table S 4. Estimated values of *Ae*, *He*, *Ho*, *PD*, *PE*, *TPI*, *PIC*, and *MP* for the 28 multi-InDel markers in the American group from the 1000 Genomes Project.

Marker	<i>Ae</i>	<i>He</i>	<i>Ho</i>	<i>PD</i>	<i>PE</i>	<i>TPI</i>	<i>PIC</i>	<i>PM</i>
MH01GB001	2.50	0.6000	0.6015	0.7619	0.2926	1.2546	0.5172	0.2381
MH02GB001	2.56	0.6104	0.5985	0.7799	0.2891	1.2454	0.5404	0.2201
MH03GB001	2.96	0.6630	0.6221	0.8198	0.3182	1.3230	0.5886	0.1802
MH04GB001	2.96	0.6623	0.5779	0.8242	0.2652	1.1847	0.5882	0.1758
MH05GB001	2.82	0.6463	0.6088	0.8037	0.3016	1.2782	0.5706	0.1963
MH06GB001	3.00	0.6668	0.6250	0.8212	0.3220	1.3333	0.5922	0.1788
MH07GB001	2.51	0.6021	0.5765	0.7770	0.2636	1.1806	0.5316	0.2230
MH08GB001	2.82	0.6457	0.5794	0.8113	0.2669	1.1888	0.5692	0.1887
MH08GB002	2.95	0.6619	0.5971	0.8209	0.2874	1.2409	0.5873	0.1791
MH08GB003	2.68	0.6277	0.5912	0.7918	0.2804	1.2230	0.5502	0.2082
MH09GB001	2.54	0.6063	0.5985	0.7653	0.2891	1.2454	0.5239	0.2347
MH09GB002	2.75	0.6363	0.5662	0.8058	0.2522	1.1525	0.5638	0.1942
MH10GB001	2.24	0.5538	0.5103	0.7143	0.1967	1.0210	0.4534	0.2857
MH11GB001	2.95	0.6620	0.5985	0.8181	0.2891	1.2454	0.5876	0.1819
MH11GB002	3.06	0.6732	0.5618	0.8324	0.2475	1.1409	0.6018	0.1676
MH12GB001	2.82	0.6452	0.6132	0.8036	0.3071	1.2928	0.5721	0.1964
MH12GB002	2.81	0.6442	0.5735	0.8113	0.2603	1.1724	0.5694	0.1887
MH13GB001	2.36	0.5760	0.5618	0.7335	0.2475	1.1409	0.4855	0.2665
MH14GB001	2.99	0.6660	0.6353	0.8196	0.3354	1.3710	0.5914	0.1804
MH15GB001	2.05	0.5128	0.4691	0.7036	0.1619	0.9418	0.4519	0.2964
MH17GB001	2.77	0.6390	0.5985	0.8021	0.2891	1.2454	0.5635	0.1979
MH17GB002	2.64	0.6215	0.5765	0.8008	0.2636	1.1806	0.5673	0.1992
MH18GB001	2.34	0.5732	0.5162	0.7493	0.2021	1.0334	0.4992	0.2507
MH19GB001	2.86	0.6509	0.5809	0.8131	0.2686	1.1930	0.5786	0.1869
MH20GB001	2.37	0.5793	0.5235	0.7601	0.2089	1.0494	0.5135	0.2399
MH20GB002	2.72	0.6331	0.5721	0.8022	0.2587	1.1684	0.5610	0.1978
MH21GB001	2.21	0.5485	0.5324	0.7344	0.2174	1.0692	0.4858	0.2656
MH22GB001	3.63	0.7251	0.6353	0.8732	0.3354	1.3710	0.6730	0.1268

Ae: effective number of alleles; *He*: expected heterozygosity; *Ho*: observed heterozygosity; *PD*: power of discrimination; *PE*: power of exclusion; *TPI*: typical paternity index; *PIC*: polymorphism information content; *MP*: match probability.

Supplementary Table S 5. Estimated values of *Ae*, *He*, *Ho*, *PD*, *PE*, *TPI*, *PIC*, and *MP* for the 28 multi-InDel markers in the East Asian group from the 1000 Genomes Project.

Marker	<i>Ae</i>	<i>He</i>	<i>Ho</i>	<i>PD</i>	<i>PE</i>	<i>TPI</i>	<i>PIC</i>	<i>PM</i>
MH01GB001	1.99	0.4976	0.4889	0.6666	0.1780	0.9783	0.4114	0.3334
MH02GB001	2.52	0.6037	0.5846	0.7737	0.2728	1.2037	0.5343	0.2263
MH03GB001	2.80	0.6435	0.6325	0.8016	0.3317	1.3605	0.5681	0.1984
MH04GB001	2.72	0.6325	0.6222	0.7858	0.3184	1.3235	0.5552	0.2142
MH05GB001	2.62	0.6190	0.6120	0.7816	0.3055	1.2885	0.5393	0.2184
MH06GB001	2.56	0.6092	0.6034	0.7677	0.2950	1.2608	0.5262	0.2323
MH07GB001	2.24	0.5548	0.5385	0.7341	0.2234	1.0833	0.4853	0.2659
MH08GB001	2.41	0.5863	0.5675	0.7555	0.2537	1.1561	0.5065	0.2445
MH08GB002	2.58	0.6128	0.5983	0.7811	0.2888	1.2447	0.5419	0.2189
MH08GB003	2.04	0.5101	0.5026	0.6917	0.1898	1.0052	0.4391	0.3083
MH09GB001	2.47	0.5952	0.6017	0.7430	0.2929	1.2554	0.5080	0.2570
MH09GB002	1.89	0.4718	0.4205	0.6563	0.1269	0.8628	0.4148	0.3437
MH10GB001	2.33	0.5716	0.5624	0.7518	0.2482	1.1426	0.5056	0.2482
MH11GB001	2.64	0.6210	0.5350	0.7939	0.2200	1.0754	0.5493	0.2061
MH11GB002	2.84	0.6480	0.6256	0.8035	0.3228	1.3356	0.5734	0.1965
MH12GB001	1.96	0.4906	0.4684	0.6401	0.1613	0.9405	0.4015	0.3599
MH12GB002	1.93	0.4827	0.4821	0.6267	0.1723	0.9653	0.3770	0.3733
MH13GB001	2.41	0.5847	0.5538	0.7540	0.2391	1.1207	0.5077	0.2460
MH14GB001	2.60	0.6155	0.5846	0.7806	0.2728	1.2037	0.5449	0.2194
MH15GB001	2.61	0.6181	0.6103	0.7807	0.3034	1.2829	0.5395	0.2193
MH17GB001	2.55	0.6088	0.5778	0.7825	0.2651	1.1842	0.5396	0.2175
MH17GB002	2.36	0.5775	0.5436	0.7576	0.2286	1.0955	0.5102	0.2424
MH18GB001	2.16	0.5370	0.5162	0.7205	0.2021	1.0336	0.4636	0.2795
MH19GB001	2.10	0.5250	0.4615	0.7241	0.1560	0.9286	0.4861	0.2759
MH20GB001	2.14	0.5339	0.4974	0.7227	0.1853	0.9949	0.4693	0.2773
MH20GB002	2.60	0.6165	0.5897	0.7756	0.2788	1.2188	0.5396	0.2244
MH21GB001	1.34	0.2562	0.2479	0.4236	0.0442	0.6648	0.2362	0.5764
MH22GB001	3.04	0.6716	0.6291	0.8186	0.3272	1.3479	0.6067	0.1814

Ae: effective number of alleles; *He*: expected heterozygosity; *Ho*: observed heterozygosity; *PD*: power of discrimination; *PE*: power of exclusion; *TPI*: typical paternity index; *PIC*: polymorphism information content; *MP*: match probability.

Supplementary Table S 6. Estimated values of *Ae*, *He*, *Ho*, *PD*, *PE*, *TPI*, *PIC*, and *MP* for the 28 multi-InDel markers in the European group from the 1000 Genomes Project.

Marker	<i>Ae</i>	<i>He</i>	<i>Ho</i>	<i>PD</i>	<i>PE</i>	<i>TPI</i>	<i>PIC</i>	<i>PM</i>
MH01GB001	2.84	0.6483	0.6398	0.8034	0.3414	1.3882	0.5749	0.1966
MH02GB001	2.32	0.5699	0.5940	0.7063	0.2837	1.2315	0.4743	0.2937
MH03GB001	2.66	0.6239	0.6098	0.7812	0.3028	1.2814	0.5437	0.2188
MH04GB001	2.98	0.6653	0.6619	0.8136	0.3719	1.4790	0.5907	0.1864
MH05GB001	2.95	0.6611	0.6556	0.8122	0.3630	1.4518	0.5867	0.1878
MH06GB001	2.89	0.6543	0.6603	0.8050	0.3696	1.4721	0.5790	0.1950
MH07GB001	2.63	0.6198	0.6209	0.7782	0.3167	1.3188	0.5406	0.2218
MH08GB001	2.79	0.6417	0.6161	0.7964	0.3107	1.3025	0.5646	0.2036
MH08GB002	2.93	0.6591	0.6635	0.8080	0.3741	1.4859	0.5848	0.1920
MH08GB003	1.89	0.4722	0.4787	0.6618	0.1695	0.9591	0.4139	0.3382
MH09GB001	2.18	0.5422	0.5719	0.7173	0.2585	1.1679	0.4748	0.2827
MH09GB002	3.05	0.6724	0.6540	0.8246	0.3608	1.4452	0.6033	0.1754
MH10GB001	1.75	0.4282	0.4297	0.6082	0.1330	0.8767	0.3633	0.3918
MH11GB001	2.55	0.6081	0.6177	0.7766	0.3127	1.3079	0.5395	0.2234
MH11GB002	2.65	0.6235	0.6240	0.7817	0.3207	1.3298	0.5474	0.2183
MH12GB001	2.48	0.5976	0.6003	0.7534	0.2913	1.2510	0.5173	0.2466
MH12GB002	2.51	0.6027	0.5861	0.7650	0.2745	1.2080	0.5191	0.2350
MH13GB001	2.71	0.6320	0.6288	0.7900	0.3268	1.3468	0.5543	0.2100
MH14GB001	2.94	0.6608	0.6730	0.8069	0.3877	1.5290	0.5860	0.1931
MH15GB001	2.36	0.5769	0.5703	0.7349	0.2567	1.1636	0.4882	0.2651
MH17GB001	2.91	0.6568	0.6003	0.8180	0.2913	1.2510	0.5822	0.1820
MH17GB002	2.56	0.6105	0.6051	0.7815	0.2970	1.2660	0.5437	0.2185
MH18GB001	1.47	0.3203	0.3065	0.5108	0.0662	0.7210	0.2960	0.4892
MH19GB001	2.44	0.5910	0.5640	0.7684	0.2499	1.1467	0.5249	0.2316
MH20GB001	2.11	0.5259	0.5055	0.7175	0.1924	1.0112	0.4642	0.2825
MH20GB002	2.70	0.6304	0.6098	0.7954	0.3028	1.2814	0.5585	0.2046
MH21GB001	2.75	0.6375	0.6461	0.7923	0.3499	1.4129	0.5641	0.2077
MH22GB001	3.23	0.6908	0.6098	0.8571	0.3028	1.2814	0.6413	0.1429

Ae: effective number of alleles; *He*: expected heterozygosity; *Ho*: observed heterozygosity; *PD*: power of discrimination; *PE*: power of exclusion; *TPI*: typical paternity index; *PIC*: polymorphism information content; *MP*: match probability.

Supplementary Table S 7. Estimated values of *Ae*, *He*, *Ho*, *PD*, *PE*, *TPI*, *PIC*, and *MP* for the 28 multi-InDel markers in the South Asian group from the 1000 Genomes Project.

Marker	<i>Ae</i>	<i>He</i>	<i>Ho</i>	<i>PD</i>	<i>PE</i>	<i>TPI</i>	<i>PIC</i>	<i>PM</i>
MH01GB001	2.55	0.6087	0.5824	0.7705	0.2702	1.1972	0.5249	0.2295
MH02GB001	2.96	0.6625	0.6789	0.8072	0.3963	1.5570	0.5880	0.1928
MH03GB001	2.52	0.6040	0.6023	0.7539	0.2937	1.2573	0.5188	0.2461
MH04GB001	2.62	0.6193	0.6473	0.7811	0.3515	1.4175	0.5502	0.2189
MH05GB001	2.97	0.6636	0.6672	0.8118	0.3794	1.5025	0.5889	0.1882
MH06GB001	2.89	0.6544	0.6256	0.8114	0.3228	1.3356	0.5805	0.1886
MH07GB001	2.68	0.6281	0.5957	0.7916	0.2857	1.2366	0.5531	0.2084
MH08GB001	2.52	0.6029	0.5824	0.7696	0.2702	1.1972	0.5280	0.2304
MH08GB002	2.67	0.6260	0.6423	0.7756	0.3447	1.3977	0.5484	0.2244
MH08GB003	2.42	0.5881	0.5774	0.7498	0.2646	1.1831	0.5015	0.2502
MH09GB001	2.96	0.6632	0.6323	0.8173	0.3314	1.3597	0.5886	0.1827
MH09GB002	2.97	0.6643	0.6589	0.8290	0.3676	1.4659	0.6070	0.1710
MH10GB001	2.22	0.5490	0.5258	0.7369	0.2111	1.0544	0.4879	0.2631
MH11GB001	2.65	0.6237	0.5907	0.7898	0.2798	1.2215	0.5516	0.2102
MH11GB002	2.92	0.6575	0.6473	0.8086	0.3515	1.4175	0.5823	0.1914
MH12GB001	2.88	0.6536	0.6439	0.8071	0.3469	1.4042	0.5795	0.1929
MH12GB002	2.72	0.6333	0.6073	0.7959	0.2998	1.2733	0.5615	0.2041
MH13GB001	2.42	0.5875	0.5491	0.7669	0.2342	1.1089	0.5199	0.2331
MH14GB001	2.66	0.6249	0.5874	0.7883	0.2760	1.2117	0.5447	0.2117
MH15GB001	2.23	0.5519	0.5607	0.6970	0.2464	1.1383	0.4573	0.3030
MH17GB001	2.72	0.6322	0.5691	0.8003	0.2554	1.1602	0.5606	0.1997
MH17GB002	2.93	0.6593	0.6306	0.8210	0.3292	1.3536	0.5928	0.1790
MH18GB001	2.28	0.5612	0.5657	0.7453	0.2518	1.1513	0.4998	0.2547
MH19GB001	2.68	0.6271	0.5724	0.7911	0.2590	1.1693	0.5506	0.2089
MH20GB001	1.87	0.4656	0.4409	0.6304	0.1408	0.8943	0.3727	0.3696
MH20GB002	2.96	0.6628	0.5907	0.8241	0.2798	1.2215	0.5880	0.1759
MH21GB001	2.34	0.5733	0.5474	0.7562	0.2325	1.1048	0.5093	0.2438
MH22GB001	3.70	0.7304	0.7022	0.8728	0.4316	1.6788	0.6792	0.1272

Ae: effective number of alleles; *He*: expected heterozygosity; *Ho*: observed heterozygosity; *PD*: power of discrimination; *PE*: power of exclusion; *TPI*: typical paternity index; *PIC*: polymorphism information content; *MP*: match probability.

Supplementary Table S 8. Estimated values of HWE probabilities for the 28 multi-InDel markers in the Brazil from SABE cohort.

Marker	<i>p</i> -HWE	s.d.
MH01GB001	0.2283	0.00020
MH02GB001	0.2115	0.00017
MH03GB001	0.6601	0.00019
MH04GB001	0.5650	0.00015
MH05GB001	0.1918	0.00016
MH06GB001	0.3097	0.00020
MH07GB001	0.4992	0.00022
MH08GB001	0.0859	0.00012
MH08GB002	0.7553	0.00018
MH08GB003	0.0856	0.00014
MH09GB001	0.4928	0.00023
MH09GB002	0.7760	0.00016
MH10GB001	0.1622	0.00014
MH11GB001	0.6020	0.00023
MH11GB002	0.8822	0.00013
MH12GB001	0.2435	0.00018
MH12GB002	0.0195	0.00006
MH13GB001	0.2425	0.00024
MH14GB001	0.9670	0.00008
MH15GB001	0.2668	0.00019
MH17GB001	0.8240	0.00014
MH17GB002	0.2135	0.00014
MH18GB001	0.0024	0.00002
MH19GB001	0.0004	0.00001
MH20GB001	0.7199	0.00019
MH20GB002	0.9939	0.00004
MH21GB001	0.0688	0.00010
MH22GB001	0.0047	0.00003

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction).

Supplementary Table S 9. Estimated values of HWE probabilities for the 28 multi-InDel markers in the African group from the 1000 Genomes Project (ESN, YRI, GWD, LWK, and MSL populations).

	ESN		YRI		GWD		LWK		MSL	
Marker	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.
MH01GB001	0.35351	0.00021	0.18275	0.00017	0.57441	0.00023	0.21890	0.00018	0.10374	0.00013
MH02GB001	0.98898	0.00005	0.39037	0.00022	0.81526	0.00017	0.09998	0.00013	0.54158	0.00022
MH03GB001	0.60839	0.00020	0.55583	0.00022	0.36307	0.00020	0.20754	0.00016	0.15848	0.00018
MH04GB001	0.43511	0.00024	0.82888	0.00017	0.65828	0.00023	0.82388	0.00018	0.15819	0.00016
MH05GB001	0.86994	0.00014	0.00321	0.00003	0.05837	0.00011	0.34872	0.00022	0.63411	0.00020
MH06GB001	0.81214	0.00018	0.11244	0.00013	0.28140	0.00018	0.23468	0.00019	0.17126	0.00017
MH07GB001	0.55202	0.00020	0.57765	0.00023	0.51490	0.00022	0.48140	0.00021	0.05596	0.00010
MH08GB001	0.37567	0.00021	0.47209	0.00023	0.78618	0.00017	0.38203	0.00022	0.93506	0.00012
MH08GB002	0.54536	0.00024	0.62771	0.00022	0.15404	0.00016	0.87879	0.00015	0.09583	0.00013
MH08GB003	0.41854	0.00023	0.68100	0.00019	0.60105	0.00021	0.32078	0.00020	0.55209	0.00022
MH09GB001	0.15853	0.00016	0.00613	0.00004	0.04361	0.00010	0.28911	0.00023	0.14580	0.00016
MH09GB002	0.91008	0.00013	0.64407	0.00021	0.27551	0.00019	0.67457	0.00022	0.23015	0.00019
MH10GB001	0.31460	0.00022	0.71486	0.00019	0.25397	0.00020	0.91871	0.00012	0.66201	0.00021
MH11GB001	0.65782	0.00022	0.96902	0.00007	0.91225	0.00013	0.78200	0.00019	0.86682	0.00015
MH11GB002	0.84511	0.00016	0.60566	0.00021	0.55796	0.00022	0.96469	0.00008	0.99346	0.00004
MH12GB001	0.18669	0.00018	0.00000	0.00000	0.15151	0.00016	0.74990	0.00020	0.07726	0.00011
MH12GB002	0.84813	0.00016	0.19324	0.00018	0.02272	0.00006	0.07254	0.00011	0.02194	0.00007
MH13GB001	0.30323	0.00020	0.41133	0.00022	0.43192	0.00021	0.40873	0.00023	0.69589	0.00020
MH14GB001	0.99194	0.00004	0.84444	0.00017	0.79792	0.00017	0.89706	0.00014	0.05029	0.00009
MH15GB001	0.04342	0.00009	0.36892	0.00022	0.06197	0.00011	0.39340	0.00022	0.76833	0.00019
MH17GB001	0.67509	0.00021	0.13975	0.00015	0.25017	0.00018	0.60434	0.00022	0.03969	0.00009
MH17GB002	0.08415	0.00013	0.54582	0.00022	0.48251	0.00023	0.90452	0.00013	0.02738	0.00007
MH18GB001	0.21755	0.00018	0.47717	0.00020	0.14753	0.00017	0.89650	0.00013	0.90501	0.00013
MH19GB001	0.28369	0.00018	0.92598	0.00012	0.13929	0.00016	0.41543	0.00022	0.13103	0.00016
MH20GB001	0.00004	0.00000	0.00574	0.00003	0.00337	0.00003	0.43766	0.00022	0.02003	0.00007
MH20GB002	0.27419	0.00020	0.47587	0.00024	0.21598	0.00018	0.54287	0.00022	0.22866	0.00018
MH21GB001	0.52732	0.00021	0.68945	0.00021	0.23379	0.00019	0.00613	0.00004	0.01398	0.00005
MH22GB001	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.39895	0.00022	0.00017	0.00001

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction). Populations: ESN – Esan; YRI - Yoruba in Ibadan; GWD - Gambian in Western Divisions in the Gambia; LWK - Luhya in Webuye, Kenya; MSL-Mende in Sierra Leone.

Supplementary Table S 10. Estimated values of HWE probabilities for the 28 multi-InDel markers in the American group from the 1000 Genomes Project (MXL, ACB, PUR, CLM, PEL, and ASW populations).

Marker	MXL		ACB		PUR		CLM		PEL		ASW	
	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.
MH01GB001	0.35277	0.00021	0.49669	0.00019	0.30428	0.00021	0.51937	0.00023	0.62667	0.00022	0.53719	0.00022
MH02GB001	0.15128	0.00015	0.89747	0.00014	0.86529	0.00015	0.63447	0.00020	0.15654	0.00017	0.74239	0.00020
MH03GB001	0.62111	0.00022	0.24860	0.00020	0.46869	0.00023	0.10034	0.00012	0.24530	0.00020	0.46797	0.00021
MH04GB001	0.79298	0.00018	0.34920	0.00024	0.03271	0.00008	0.30114	0.00021	0.13728	0.00015	0.21289	0.00018
MH05GB001	0.91810	0.00013	0.76720	0.00018	0.16838	0.00016	0.44015	0.00022	0.00139	0.00002	0.20900	0.00018
MH06GB001	0.59845	0.00021	0.24463	0.00019	0.98935	0.00005	0.05764	0.00010	0.97905	0.00006	0.04703	0.00010
MH07GB001	0.57079	0.00022	0.45078	0.00023	0.78735	0.00019	0.89627	0.00014	0.01230	0.00005	0.13647	0.00015
MH08GB001	0.03832	0.00008	0.58723	0.00021	0.80405	0.00018	0.60659	0.00020	0.40158	0.00023	0.67240	0.00021
MH08GB002	0.96956	0.00008	0.99405	0.00003	0.07836	0.00013	0.52124	0.00024	0.43326	0.00022	0.31405	0.00020
MH08GB003	0.53282	0.00021	0.65711	0.00021	0.91986	0.00012	0.88022	0.00014	0.57077	0.00022	0.95141	0.00009
MH09GB001	0.46953	0.00024	0.52773	0.00023	0.42146	0.00023	0.38675	0.00021	0.22911	0.00018	0.28603	0.00021
MH09GB002	0.48103	0.00022	0.58156	0.00021	0.85356	0.00016	0.84307	0.00017	0.19157	0.00015	0.03798	0.00009
MH10GB001	0.21520	0.00018	0.21540	0.00019	0.57163	0.00020	0.13861	0.00016	0.06080	0.00010	0.35608	0.00021
MH11GB001	0.27712	0.00020	0.36872	0.00023	0.01998	0.00007	0.36755	0.00020	0.33917	0.00019	0.16092	0.00017
MH11GB002	0.03734	0.00009	0.52128	0.00021	0.87390	0.00015	0.07782	0.00011	0.21433	0.00017	0.69980	0.00020
MH12GB001	0.86857	0.00015	0.36097	0.00021	0.34717	0.00020	0.61691	0.00022	0.15926	0.00016	0.55067	0.00022
MH12GB002	0.41061	0.00023	0.69375	0.00020	0.73397	0.00020	0.91821	0.00013	0.02862	0.00008	0.08778	0.00013
MH13GB001	0.04755	0.00009	0.07152	0.00011	0.94782	0.00010	0.70083	0.00020	0.43616	0.00022	0.02286	0.00006
MH14GB001	0.69737	0.00020	0.53900	0.00021	0.61279	0.00021	0.47505	0.00022	0.59553	0.00022	0.68603	0.00020
MH15GB001	0.21916	0.00018	0.89168	0.00014	0.20874	0.00018	0.52732	0.00022	0.13767	0.00015	0.66745	0.00021
MH17GB001	0.02744	0.00007	0.82564	0.00017	0.16472	0.00016	0.13621	0.00015	0.29234	0.00022	0.03650	0.00009
MH17GB002	0.95610	0.00009	0.85052	0.00016	0.15940	0.00016	0.07211	0.00011	0.03657	0.00008	0.82844	0.00016
MH18GB001	0.32601	0.00021	0.27226	0.00020	0.21522	0.00019	0.15955	0.00016	0.49322	0.00023	0.25595	0.00019
MH19GB001	0.38462	0.00023	0.86309	0.00015	0.76281	0.00018	0.05313	0.00009	0.00193	0.00002	0.04591	0.00010
MH20GB001	0.85606	0.00016	0.01781	0.00006	0.28151	0.00020	0.37599	0.00022	0.06269	0.00011	0.00686	0.00004
MH20GB002	0.25035	0.00019	0.28343	0.00020	0.00523	0.00003	0.93088	0.00012	0.53126	0.00021	0.98002	0.00006
MH21GB001	0.16745	0.00015	0.56027	0.00020	0.42607	0.00022	0.60975	0.00020	0.21297	0.00018	0.39591	0.00022
MH22GB001	0.44709	0.00021	0.00034	0.00001	0.08121	0.00013	0.35880	0.00023	0.01391	0.00005	0.00006	0.00000

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction). Populations: MXL - Mexican Ancestry from Los Angeles USA; ACB- African Caribbeans in Barbados; PUR - Puerto Ricans from Puerto Rico; CLM - Colombians from Medellin; PEL - Peruvians from Lima, Peru; and ASW - Americans of African Ancestry in SW US.

Supplementary Table S 11. Estimated values of HWE probabilities for the 28 multi-InDel markers in the East Asian group from the 1000 Genomes Project (CHS, KHV, CDX, CHB, JPT populations).

Marker	CHS		KHV		CDX		CHB		JPT	
	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.
MH01GB001	0.40003	0.00022	0.36565	0.00020	0.42291	0.00022	0.09893	0.00013	0.77316	0.00018
MH02GB001	0.99452	0.00003	0.02518	0.00007	0.49350	0.00022	0.39155	0.00023	0.46749	0.00023
MH03GB001	0.98935	0.00004	0.08747	0.00012	0.64290	0.00022	0.42292	0.00023	0.88795	0.00014
MH04GB001	0.31434	0.00019	0.80115	0.00018	0.63691	0.00022	0.56002	0.00021	0.32889	0.00021
MH05GB001	0.87296	0.00014	0.23316	0.00019	0.12476	0.00014	0.58152	0.00021	0.64362	0.00020
MH06GB001	0.21535	0.00018	0.92819	0.00012	0.83274	0.00017	0.87708	0.00015	0.88669	0.00014
MH07GB001	0.09077	0.00013	0.84661	0.00016	0.36306	0.00022	0.19158	0.00018	0.08901	0.00012
MH08GB001	0.74229	0.00020	0.61790	0.00021	0.20921	0.00020	0.58080	0.00022	0.85959	0.00016
MH08GB002	0.33679	0.00022	0.80301	0.00019	0.29794	0.00020	0.89606	0.00014	0.82042	0.00016
MH08GB003	0.61297	0.00021	0.04329	0.00009	0.35519	0.00021	0.51902	0.00023	0.70182	0.00021
MH09GB001	0.06650	0.00011	0.03405	0.00008	0.27290	0.00018	0.31414	0.00019	0.01187	0.00005
MH09GB002	0.67256	0.00020	0.65797	0.00023	0.77215	0.00018	0.31068	0.00020	0.58669	0.00021
MH10GB001	0.77574	0.00018	0.01936	0.00006	0.17436	0.00016	0.14646	0.00016	0.10405	0.00014
MH11GB001	0.32641	0.00021	0.50703	0.00022	0.91656	0.00012	0.56789	0.00021	0.24989	0.00018
MH11GB002	0.00059	0.00001	0.09782	0.00014	0.00034	0.00001	0.65046	0.00020	0.00011	0.00000
MH12GB001	0.39461	0.00023	0.13502	0.00015	0.68999	0.00020	0.13485	0.00015	0.17375	0.00017
MH12GB002	0.65153	0.00020	0.04838	0.00009	0.83711	0.00016	0.48854	0.00022	0.02183	0.00007
MH13GB001	0.01731	0.00006	0.10764	0.00014	0.70378	0.00020	0.20427	0.00016	0.12012	0.00015
MH14GB001	0.37065	0.00020	0.03806	0.00009	0.48786	0.00022	0.27086	0.00020	0.09124	0.00013
MH15GB001	0.55794	0.00021	0.80952	0.00017	0.52743	0.00022	0.00298	0.00003	0.88791	0.00014
MH17GB001	0.59960	0.00020	0.61290	0.00021	0.78308	0.00018	0.18587	0.00017	0.04337	0.00009
MH17GB002	0.30321	0.00019	0.26754	0.00019	0.20784	0.00017	0.21917	0.00021	0.08377	0.00014
MH18GB001	0.89550	0.00014	0.83661	0.00017	0.04869	0.00010	0.20747	0.00018	0.45101	0.00023
MH19GB001	0.00033	0.00001	0.22732	0.00018	0.53359	0.00023	0.26077	0.00019	0.00000	0.00000
MH20GB001	0.13323	0.00015	0.25298	0.00019	0.27864	0.00020	0.00795	0.00004	0.33420	0.00021
MH20GB002	0.06093	0.00011	0.20113	0.00019	0.27716	0.00020	0.12756	0.00014	0.00435	0.00003
MH21GB001	1.00000	0.00000	0.44512	0.00022	0.00309	0.00002	1.00000	0.00000	0.79622	0.00019
MH22GB001	0.00000	0.00000	0.00175	0.00002	0.00013	0.00000	0.00065	0.00001	0.00000	0.00000

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction). Populations: CHS - Southern Han Chinese, KHV - Kinh in Ho Chi Minh City, CDX - Chinese Dai in Xishuangbanna, CHB - Han Chinese in Beijing, JPT- Japanese in Tokyo.

Supplementary Table S 12. Estimated values of HWE probabilities for the 28 multi-InDel markers in the European group from the 1000 Genomes Project (CEU, TSI, GBR, FIN, IBS populations).

Marker	CEU		TSI		GBR		FIN		IBS	
	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.
MH01GB001	0.37966	0.00020	0.10038	0.00013	0.27887	0.00019	0.37849	0.00022	0.83197	0.00016
MH02GB001	0.57484	0.00023	0.31714	0.00021	0.84786	0.00018	0.23451	0.00020	0.29117	0.00020
MH03GB001	0.51893	0.00022	0.66385	0.00021	0.86160	0.00014	0.59461	0.00023	0.03909	0.00009
MH04GB001	0.84377	0.00015	0.88953	0.00014	0.61555	0.00023	0.44788	0.00021	0.15050	0.00016
MH05GB001	0.40927	0.00022	0.12798	0.00014	0.75054	0.00019	0.26686	0.00019	0.22848	0.00019
MH06GB001	0.76657	0.00018	0.30332	0.00020	0.82590	0.00016	0.12329	0.00014	0.35783	0.00021
MH07GB001	0.60089	0.00020	0.02449	0.00007	0.06315	0.00011	0.36337	0.00022	0.05799	0.00010
MH08GB001	0.73864	0.00019	0.87819	0.00015	0.67638	0.00020	1.00000	0.00000	0.09692	0.00012
MH08GB002	0.48905	0.00022	0.55589	0.00023	0.45598	0.00023	0.82616	0.00017	0.80701	0.00018
MH08GB003	0.72101	0.00019	0.14129	0.00016	0.40869	0.00022	0.89147	0.00014	0.59480	0.00023
MH09GB001	0.70644	0.00023	0.23812	0.00017	0.14932	0.00016	0.35964	0.00022	0.07117	0.00010
MH09GB002	0.59542	0.00022	0.92478	0.00012	0.94092	0.00011	0.40948	0.00021	0.52285	0.00021
MH10GB001	0.62753	0.00020	0.70561	0.00020	0.47535	0.00022	0.79743	0.00018	0.90316	0.00014
MH11GB001	0.54599	0.00025	1.00000	0.00000	0.43527	0.00021	0.89112	0.00014	0.91194	0.00012
MH11GB002	0.22926	0.00018	0.48350	0.00023	0.08137	0.00011	0.84705	0.00017	0.53369	0.00022
MH12GB001	0.03557	0.00008	0.14743	0.00016	0.97204	0.00007	0.68373	0.00020	0.30356	0.00021
MH12GB002	0.60210	0.00021	0.12449	0.00015	0.55200	0.00023	0.54362	0.00024	0.59488	0.00021
MH13GB001	0.56090	0.00023	0.56939	0.00023	0.19567	0.00017	0.86516	0.00015	0.75950	0.00018
MH14GB001	0.53151	0.00023	0.71657	0.00020	0.62420	0.00022	0.64756	0.00021	0.21253	0.00018
MH15GB001	0.42402	0.00024	0.43226	0.00021	0.43367	0.00022	0.36319	0.00022	0.08875	0.00012
MH17GB001	0.09713	0.00013	0.70774	0.00022	0.48156	0.00023	0.00056	0.00001	0.49533	0.00021
MH17GB002	0.99604	0.00003	0.55921	0.00023	0.32647	0.00020	0.98358	0.00006	0.60462	0.00022
MH18GB001	0.22325	0.00019	0.14579	0.00016	0.04721	0.00009	0.86563	0.00016	0.66800	0.00022
MH19GB001	0.90586	0.00013	0.05798	0.00011	0.40585	0.00023	0.01963	0.00006	0.52220	0.00022
MH20GB001	0.61164	0.00022	0.39252	0.00022	0.09300	0.00013	0.78433	0.00018	0.38643	0.00020
MH20GB002	0.83955	0.00017	0.32133	0.00021	0.83201	0.00017	0.42100	0.00022	0.65944	0.00019
MH21GB001	0.74621	0.00019	0.41669	0.00023	0.62228	0.00021	0.95321	0.00009	0.08692	0.00012
MH22GB001	0.10640	0.00013	0.18038	0.00017	0.00004	0.00000	0.01439	0.00005	0.02667	0.00007

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction). Populations: CEU - Utah Residents (CEPH) with Northern and Western European Ancestry, TSI - Toscani in Italia, GBR - British in England and Scotland, FIN - Finnish, IBS - Iberian Population in Spain.

Supplementary Table S 13. Estimated values of HWE probabilities for the 28 multi-InDel markers in the South Asian group from the 1000 Genomes Project (GIH, BEB, PJJ, STU, ITU populations).

Marker	GIH		BEB		PJJ		STU		ITU	
	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.	<i>p</i> -HWE	s.d.
MH01GB001	0.38647	0.00023	0.22337	0.00018	0.01156	0.00005	0.48699	0.00022	0.13818	0.00015
MH02GB001	0.60770	0.00022	0.36974	0.00023	0.35970	0.00024	0.50089	0.00021	0.13031	0.00016
MH03GB001	0.66649	0.00021	0.11729	0.00014	0.41886	0.00022	0.61693	0.00022	0.69122	0.00021
MH04GB001	0.70937	0.00019	0.34308	0.00021	0.14623	0.00016	0.60828	0.00021	0.23903	0.00018
MH05GB001	0.33147	0.00021	0.53621	0.00022	0.46427	0.00021	0.73531	0.00019	0.00986	0.00004
MH06GB001	0.67560	0.00021	0.18970	0.00017	0.80206	0.00019	0.17227	0.00018	0.73449	0.00021
MH07GB001	0.43703	0.00022	0.10358	0.00013	0.32998	0.00019	0.87333	0.00015	0.53293	0.00023
MH08GB001	0.04258	0.00008	0.81319	0.00017	0.64458	0.00021	0.10828	0.00014	0.86749	0.00015
MH08GB002	0.47249	0.00022	0.55405	0.00021	0.38540	0.00022	0.57425	0.00020	1.00000	0.00000
MH08GB003	0.11212	0.00014	0.05981	0.00010	0.05301	0.00010	0.72117	0.00020	0.87540	0.00014
MH09GB001	0.93320	0.00011	0.17319	0.00016	0.88549	0.00014	0.92605	0.00011	0.19189	0.00018
MH09GB002	0.63817	0.00021	0.97733	0.00007	0.08158	0.00012	0.01648	0.00005	0.40604	0.00022
MH10GB001	0.93359	0.00011	0.06701	0.00010	0.42122	0.00021	0.40384	0.00022	0.07553	0.00012
MH11GB001	0.73556	0.00020	0.34428	0.00021	0.87027	0.00015	0.38371	0.00022	0.42813	0.00022
MH11GB002	0.82006	0.00016	0.31164	0.00022	0.92092	0.00012	0.07241	0.00011	0.46518	0.00022
MH12GB001	0.20721	0.00018	0.75650	0.00019	0.17358	0.00017	0.81269	0.00017	0.92904	0.00012
MH12GB002	0.79719	0.00017	0.40930	0.00022	0.93371	0.00011	0.67704	0.00021	0.31071	0.00022
MH13GB001	0.96309	0.00008	0.54948	0.00024	0.08233	0.00014	0.54176	0.00022	0.60088	0.00023
MH14GB001	0.99033	0.00004	0.58380	0.00021	0.85764	0.00016	0.02464	0.00007	0.48276	0.00022
MH15GB001	0.36930	0.00022	0.79534	0.00018	0.34525	0.00022	0.62258	0.00021	0.05042	0.00010
MH17GB001	0.50464	0.00021	0.15066	0.00017	0.00816	0.00004	0.68639	0.00019	0.37260	0.00021
MH17GB002	0.18706	0.00016	0.19803	0.00018	0.01690	0.00006	0.01321	0.00005	0.73897	0.00019
MH18GB001	0.82451	0.00017	0.57530	0.00021	0.75223	0.00020	0.81086	0.00019	0.69393	0.00021
MH19GB001	0.95369	0.00009	0.29718	0.00023	0.00146	0.00002	0.39046	0.00023	0.70901	0.00020
MH20GB001	0.16105	0.00017	0.90059	0.00013	0.86420	0.00015	0.79753	0.00017	0.52879	0.00023
MH20GB002	0.78277	0.00017	0.00305	0.00003	0.00184	0.00002	0.23244	0.00018	0.32579	0.00020
MH21GB001	0.05399	0.00010	0.46001	0.00022	0.87276	0.00016	0.08854	0.00013	0.71343	0.00021
MH22GB001	0.52935	0.00024	0.08140	0.00011	0.00053	0.00001	0.84740	0.00017	0.50324	0.00021

Notes. HWE: Hardy-Weinberg equilibrium; s.d: standard deviation ($p < 0.00178$ after Bonferroni correction). Populations: GIH - Gujarati Indian from Houston, Texas, BEB - Bengali from Bangladesh, PJJ - Punjabi from Lahore, STU - Sri Lankan Tamil, ITU - Indian Telugu.

4. Conclusão

Conforme evidenciado na literatura, os marcadores multi-InDels têm demonstrado resultados atuais e promissores para diversas aplicações na área forense. No entanto, até o momento, nenhum dos painéis desenvolvido havia sido avaliado na população brasileira. Desse modo, este é o primeiro estudo que avaliou a aplicabilidade desses marcadores em nossa população.

Até então, todos os painéis compostos apenas por marcadores multi-InDels haviam sido avaliados exclusivamente em populações asiáticas. Por esse motivo, realizamos uma análise prévia de três painéis desenvolvidos por Huang et al. (2014), Qu et al. (2020) e Li et al. (2021), com o objetivo de avaliar sua aplicabilidade em indivíduos brasileiros.

Nós avaliamos 54 marcadores propostos anteriormente e constatamos que a maioria são informativos na população brasileira, apresentando três ou mais haplótipos frequentes, com diferentes tamanhos. Entretanto, 64,8% dos marcadores são propensos a erros de amplificação e/ou sequenciamento, devido à presença em regiões repetitivas, de baixa complexidade, homopolímeros e regiões associadas a STRs.

Dentre os painéis analisados, o proposto por Huang et al. demonstrou ser o mais promissor para aplicação no Brasil, com probabilidade de *match* combinada de $4,92 \times 10^{-14}$ e poder de exclusão combinado de 0,9991. Ainda assim, esses valores foram inferiores aos obtidos com os STRs CODIS ou com painéis de SNPs mais abrangentes.

Nesse contexto, selecionamos e propusemos um novo conjunto de marcadores, organizados em ensaio multiplex, composto por 28 marcadores multi-InDel e dois marcadores para determinação do sexo. Após as análises, nós detectamos 107 alelos específicos, sendo que 94,4% possuem tamanhos inferiores a 250pb, o que pode conferir maior potencial nas análises de amostras degradadas.

Além disso, os multi-InDels não geram artefatos do tipo *stutter*, comumente observados na amplificação por PCR de marcadores STRs, uma vez que não estão localizados

em regiões repetitivas. Essa característica pode aumentar sua utilidade em casos envolvendo amostras contendo misturas de diferentes perfis genéticos.

Todos os marcadores mostraram-se altamente polimórficos nas cinco superpopulações de diferentes ancestralidades, bem como em indivíduos brasileiros, apresentando pelo menos três haplótipos distintos. A probabilidade de *match* combinada variou entre $5,66 \times 10^{-20}$ e $2,66 \times 10^{-17}$, enquanto o poder de exclusão combinado variou entre 0,99958 e 0,99998.

Portanto, esse novo painel pode representar uma ferramenta complementar promissora para a identificação humana em diversas populações ao redor do mundo, especialmente em casos envolvendo amostras complexas.

5. Perspectivas Futuras

O presente estudo é realizado em colaboração com a Universidade de São Paulo (USP) e o Laboratório de Genética Molecular Forense da Polícia Científica do Paraná (LGMF).

Após a seleção e análise computacional dos marcadores utilizando dados públicos de sequenciamento de alta cobertura (30x) do projeto 1000 Genomas e do projeto SABE, realizamos a validação experimental preliminar de todos os primers desenhados por meio da técnica de eletroforese em gel de poliacrilamida, obtendo resultados promissores.

Recentemente, foi solicitada a síntese dos primers marcados com fluoróforos para prosseguir com as etapas experimentais de validação utilizando a plataforma de eletroforese capilar. Atualmente, está em andamento a coleta de amostras de aproximadamente 550 indivíduos voluntários não relacionados e 10 trios aparentados, provenientes de cinco cidades do estado do Paraná: Curitiba, Cascavel, Guarapuava, Londrina e Maringá.

A partir dessas amostras coletadas, serão aplicados o novo conjunto de marcadores multi-InDel e calculados as frequências alélicas, os parâmetros forenses, a aderência ao equilíbrio de Hardy-Weinberg, bem como a avaliação da sensibilidade e especificidade do ensaio. Ainda, será verificado o desempenho do painel frente a amostras degradadas artificialmente, amostras contendo material genético de dois ou mais indivíduos e amostras provenientes de casos reais.

Por essa razão, consideramos desnecessária a análise do desempenho deste painel em amostras degradadas e contendo mistura de perfis por meio de simulações computacionais, uma vez que serão abordadas de forma prática e direta. Com essa abordagem, esperamos obter resultados ainda mais robustos quanto à aplicabilidade dos marcadores propostos na nossa população.

Paralelamente, novos marcadores multi-InDel, capazes de inferir a ancestralidade dos indivíduos, estão sendo investigados no genoma humano com o objetivo de serem incorporados ao painel, ampliando sua utilidade em diferentes aplicações forenses.

Por fim, esperamos que o painel desenvolvido neste estudo seja submetido à patente, viabilizando sua futura aplicação prática nos laboratórios de genética forense.

6. *Apêndice*

Análises Estatísticas

Os cálculos das frequências alélicas baseiam-se no Princípio de Hardy-Weinberg (HW), um modelo matemático que demonstra que as frequências genotípicas de uma população permanecem constantes ao longo do tempo, desde que não haja forças evolutivas atuando sobre ela.

No modelo de HW, a relação matemática entre as frequências alélicas e genotípicas de um marcador bi-alélico se dá conforme a equação:

$$p^2 + 2pq + q^2 = 1$$

Onde p^2 , $2pq$ e q^2 são as frequências dos genótipos **AA**, **Aa** e **aa** nos zigotos em qualquer geração, **p** e **q** são as frequências alélicas de **A** e **a** nos gametas da geração precedente e $p + q = 1$. A equação constitui a base do Princípio de HW (Beiguelman, 2008).

Para que o princípio seja demonstrado, a população deve obedecer às seguintes premissas:

- 1- A população é infinita;
- 2- Existem o mesmo número de homens e mulheres na população;
- 3- A população está em panmixia, ou seja, todos casam e os casamentos são aleatórios. Os casamentos consanguíneos podem existir, desde que também ocorram aleatoriamente;
- 4- Todos os casais da população são igualmente férteis e geram o mesmo número de filhos;
- 5- Não há sobreposição de gerações da população;
- 6- Os genes não sofrem mutação;
- 7- A população não está sob pressão de seleção natural, porque todos os indivíduos são igualmente viáveis, não existindo fatores que aumentem ou diminuam a sobrevivência dos indivíduos com um determinado genótipo;

8- A população não recebe nem emite um fluxo gênico capaz de alterar a sua composição gênica original, porque ela não sofre miscigenação com população imigrante que apresenta frequências gênicas diferentes da dela, e nem emigração diferencial.

Se uma população humana atendesse às premissas, ela apresentaria uma estabilidade genética que permaneceria inalterada ao longo do tempo. Em outras palavras, mostrariam fixidez genética, ou seja, inércia evolutiva, pois, não estariam sujeitas a fatores evolutivos capazes de alterar as frequências gênicas (mutações, seleção natural, fluxo gênico de populações de migrantes e deriva genética) e fatores que aumentam a homozigose (endocruzamento, casamento consanguíneos em humanos e subdivisão da população em grandes isolados) (Beiguelman, 2008).

Esta população é denominada ideal e, evidentemente, suas premissas não são plenamente atendidas por nenhuma população real. Aliás, é justamente a violação dessas premissas que permite explicar o processo evolutivo dos seres vivos. Apesar de nenhuma população obedecer a todas as premissas estabelecidas, na prática tem demonstrado em numerosas populações humanas, a distribuição dos genótipos conforme a Lei de HW (Beiguelman, 2008).

Neste estudo, para as análises das frequências alélicas e parâmetros forenses foi utilizado o *software online* STRAF 2.1.5 (Gouy & Zieger, 2017). Para a análise da aderência ao equilíbrio de Hardy-Weinberg (HWE) e o desequilíbrio de ligação (LD) foi utilizado o *software* Arlequin 3.5.2.2 (Excoffier & Lischer, 2010) e o cálculo do número efetivo de alelos (A_e) foi realizado conforme a fórmula descrita por Kidd & Speed (2015).

Desse modo, apresenta-se a descrição detalhada de cada parâmetro analisado:

a) Heterozigosidade Esperada (H_e) ou Diversidade Genética (GD) é calculado utilizando a seguinte fórmula:

$$H_{\text{exp}} = GD = \frac{n}{n-1} \left(1 - \sum_{i=1}^n (p_i)^2 \right)$$

Onde o n é o número de cópias gênicas amostradas (ou seja, o tamanho da amostra para amostras haplóides, e o dobro do tamanho da amostra para amostras diplóides), e p_i é a frequência do alelo na população.

Interpretação – GD ou H_e representa a probabilidade de que um indivíduo, selecionado aleatoriamente na população, seja heterozigoto em um determinado *locus*.

Por exemplo, um valor de GD igual a 0,6 indica que há 60% de probabilidade de um indivíduo qualquer ser heterozigoto nesse *locus*. O valor de GD depende diretamente da diversidade genética presente no *locus*, a qual, por sua vez, é influenciada pelas frequências alélicas da população.

Caso existam dois alelos em uma população: a diversidade genética será maior em uma população onde as frequências alélicas são 0,5 e 0,5 do que em outra onde são 0,1 e 0,9, pois é possível formar menos heterozigotos quando os alelos são raros (Gouy & Zieger, 2017).

b) Heterozigosidade Observada (H_o) é definida como a razão dos genótipos heterozigotos, em um determinado *locus*, pelo total de indivíduos analisados em uma amostra populacional.

Quando o H_o é maior que o H_e , isso indica um excesso de heterozigotos em relação ao que é previsto pelo modelo de Hardy-Weinberg. Por outro lado, se o H_o for menor que o H_e , há um excesso de homozigotos na população.

c) Probabilidade de Match (PM) é definida como a probabilidade de que dois indivíduos não relacionados compartilhem aleatoriamente o mesmo genótipo.

Fórmula:

$$PM = \sum_i (G_i)^2,$$

Onde G_i representa a frequência do genótipo i em um determinado *locus* na população.

O cálculo para vários *loci* seria o produto de PM de cada *locus*.

d) Poder de discriminação (PD) é definido como a probabilidade de discriminação entre dois indivíduos não relacionados.

Fórmula:

$$PD = 1 - PM$$

Interpretação: O PD é calculado como 1 - PM. Ao invés de considerar a probabilidade de coincidência (*match*) entre dois indivíduos, o foco está na probabilidade de não coincidência, ou seja, na probabilidade de discriminação entre eles (Gouy & Zieger, 2017). Quando menor for o valor de PM, maior será o valor de PD.

e) Poder de exclusão (PE) pode ser calculado utilizando a seguinte fórmula:

$$PE = h^2 (1 - 2hH^2)$$

Onde h é heterozigose observada e H é a homozigose observada em uma população.

Interpretação: o poder de exclusão depende das proporções observadas de indivíduos heterozigotos e homozigotos em uma população. Essas proporções, quando multiplicadas conforme mostrado na equação, fornecem a probabilidade média de exclusão de um determinado genótipo em um teste de paternidade. Intuitivamente, quanto mais diversa for a população, maior será esse valor, uma vez que a diversidade está relacionada à heterozigosidade (Gouy & Zieger, 2017).

Cálculo para vários loci:

$$P_{E \text{ combinado}} = 1 - \prod_{l=1}^L (1 - P_{El})$$

Onde P_{El} é o poder de exclusão de cada *locus* e L é o número de *locus*.

f) **Índice de paternidade típico (TPI)** é calculado conforme a fórmula:

$$TPI = \frac{1}{2H}$$

Onde H é a frequência de homozigotos.

Interpretação: os valores de TPI não estão na faixa de 0 a 1 e não podem ser interpretados como uma probabilidade. Esse valor representa uma razão de verossimilhança (*odds ratio*), indicando quantas vezes é mais provável observar-se o perfil genotípico em questão caso um suposto pai seja, de fato, o pai biológico em relação a um homem escolhido aleatoriamente da população (Gouy & Zieger, 2017).

g) **Conteúdo Informativo de Polimorfismo (PIC)** é calculado utilizando a seguinte fórmula:

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i^2 \right)^2 + \sum_{i=1}^n p_i^4$$

Onde p_i é a frequência de cada alelo distinto e n é o número de alelos distintos.

Interpretação: O valor de PIC pode ser interpretado como: a) probabilidade de que os alelos materno e paterno de uma criança sejam dedutíveis; b) a probabilidade de ser possível deduzir qual alelo foi transmitido por um dos genitores à criança (Gouy & Zieger, 2017).

O valor PIC calculado por *locus* e informa o quanto o polimorfismo é informativo. Valores de $PIC \geq 0,5$ indicam alto valor informativo; $0,5 > PIC \geq 0,25$ possui um conteúdo informativo médio e valores $PIC < 0,25$ possuem fraco conteúdo informativo.

h) Número efetivo de alelos (A_e) é recíproco da homozigose esperada.

Fórmula:

$$1 / \sum p_i^2$$

Onde p_i é igual à frequência do alelo i e a soma é realizada sobre todos os alelos presentes no *locus*.

Interpretação: O potencial de detecção de misturas genéticas é ampliado quando os alelos apresentam frequências mais equilibradas. O uso do número efetivo de alelos permite padronizar a comparação entre *loci*, e quanto maior esse valor, maior a probabilidade de identificar a presença de uma mistura em uma amostra (Kidd & Speed, 2015).

i) Desequilíbrio de Ligação - DL

A ligação é o fenômeno em que alelos adjacentes situados em um mesmo segmento cromossômico tendem a ser herdados juntos durante a meiose, com uma frequência maior do que o esperado ao acaso. O conjunto desses alelos ligados é denominado como haplótipo.

Ao se analisar dois *locus* (A e B), cada um com dois alelos com frequências p_A , p_a , p_B e p_b , respectivamente, espera-se que, em equilíbrio, a frequência do haplótipo AB seja igual ao produto das frequências alélicas $p_A.p_B$. As frequências esperadas dos demais haplótipos (Ab, aB e ab) também podem ser estimadas pelo produto das frequências dos alelos correspondentes (Hartl e Clark, 1997).

O desequilíbrio de ligação (DL) é definido como a ocorrência conjunta de determinadas combinações alélicas em diferentes *loci* com uma frequência superior à esperada, sob independência.

O cálculo do coeficiente de DL (D) é a medida da diferença existente entre os produtos das frequências dos haplótipos em acoplamento e o produto das frequências dos haplótipos em repulsão (Lewontin & Kojima, 1960), conforme a seguinte fórmula:

$$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$$

Assim, quando os *locus* estão em equilíbrio obtém-se um valor de $D = 0$ e quando os *locus* são considerados em desequilíbrio obtém-se um valor de $D \neq 0$. Por tanto, trata-se de

uma medida altamente influenciada pelas frequências alélicas, pois, quando há excesso de haplótipos em acoplamento observam-se desvios positivos ($D = \text{positivo}$) e o excesso de haplótipos em repulsão geram desvios negativos ($D = \text{negativo}$).

Devido a ocorrência da recombinação, uma proporção de haplótipos em acoplamento é convertida em haplótipos de repulsão e vice-versa, reduzindo o DL a cada geração. O DL decairá, tendendo a 0, se nenhum outro processo evolutivo, além da recombinação, estiver agindo na população. Quando a taxa de recombinação é baixa, o DL reduz de forma lenta, por outro lado, quando atingir seu valor máximo ($r = 0,5$, taxa de recombinação máxima entre *locus* em diferentes cromossomos), o DL decai rapidamente. Ainda assim, o equilíbrio só será alcançado após algumas gerações (Hartl e Clark, 1997).

O DL em qualquer geração futura pode ser calculado pela fórmula:

$$DL = (1 - r)^t DLo$$

Onde: DLo = DL inicial; t = número de gerações com cruzamento ao acaso; r = taxa de recombinação.

A grande parte dos casos, o DL é explicado pela proximidade física entre os *loci* no genoma. *Loci* que estão próximos entre si tendem a sofrer menos recombinação durante a meiose, o que reduz o embaralhamento dos alelos entre as gerações. Por outro lado, *loci* localizados em cromossomos diferentes ou em regiões distantes de um mesmo cromossomo geralmente não apresentam DL significativo, pois seus alelos são recombinados de forma independente a cada geração (Gouy & Zieger, 2017).

Outra medida é denominada como D' foi introduzida por Lewontin (1964), e trata-se de uma medida normalizada e padronizada, podendo variar no intervalo entre -1 a 1, sendo calculado conforme as fórmulas:

$$D' = D/D_{max} \text{ (quando } D \text{ for positivo)}$$

$$D' = D/D_{min} \text{ (quando } D \text{ for negativo)}$$

Onde: $D_{máx} = \min. (p_{Apb}; p_{aPB})$; $D_{min} = \min. (p_{APB}; p_{apb})$.

O uso de $D_{\text{máx}}$ ou D_{min} está condicionado ao sinal encontrado na mensuração do D , se este for positivo utiliza-se o $D_{\text{máx}}$ e se for negativo utiliza-se o D_{min} .

Posteriormente, Hill e Robertson (1968) introduziu a medida r^2 para medir a correlação entre alelos de *locus* distintos, conforme fórmula:

$$r^2 = D^2 / [p_A (1 - p_A) p_B (1 - p_B)]$$

A medida r^2 é uma das mais utilizadas para *locus* bi-alélicos, podendo variar entre 0 e 1. O valor de $r^2 = 1$, indica associação preferencial máxima entre alelos de pares de *locus*, pode-se dizer que dois dos quatro haplótipos estão ausentes na amostra, o que caracteriza uma situação de perfeito DL.

Outra alternativa para avaliar o DL entre dois *locus* é o uso de testes de significância estatística. A aplicação desses testes permite verificar se há ou não uma associação estatisticamente significativa entre alelos de diferentes *locus*, possibilitando a inferência sobre o DL (Slatkin, 1994). Um dos testes mais utilizados é o Teste Exato de Fisher, embora seja considerado altamente conservador (Zapata & Álvarez, 1997).